A THESIS FOR THE DEGREE OF MASTER OF SCIENCE

Transcriptome Profiling and Comparative Analysis of *Panax ginseng* Adventitious Roots

BY

MURUKARTHICK JAYAKODI

FEBRUARY, 2014

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE
THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Transcriptome Profiling and Comparative Analysis of Panax ginseng Adventitious Roots

UNDER THE DIRECTION OF DR. TAE-JIN YANG
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF SEOUL NATIONAL UNIVERSITY

BY
MURUKARTHICK JAYAKODI

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE

APPROVED AS A QUALIFIED DISSERTATION OF
MURUKARTHICK JAYAKODI
FOR THE DEGREE OF MASTER OF SCIENCE
BY THE COMMITTEE MEMBERS

FEBRUARY, 2014

CHAIRMAN ______________________________
Hak-Soo Seo, Ph.D.

VICE-CHAIRMAN ______________________________
Tae-Jin Yang, Ph.D.

MEMBER ______________________________
Chan-Seok Shin, Ph.D.

Transcriptome Profiling and Comparative Analysis of *Panax ginseng* Adventitious Roots

MURUKARTHICK JAYAKODI

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

GENERAL ABSTRACT

Panax ginseng (C. A. Meyer) is a traditional medicinal plant famous for its strong therapeutic effects. However, genomic resources for *P. ginseng* are still very limited. In this study, we performed *de novo* assembly of transcriptomes from adventitious roots of two *P. ginseng* cultivars, Chunpoong (CP) and Cheongsun (CS). The assemblies were generated from ~85 and ~77 million high-quality Illumina HiSeq reads from CP and CS cultivars, respectively. A total of 35,527 and 27,716 transcripts were obtained from the CP and CS assemblies, respectively. Annotation of the transcriptomes showed that approximately 90% of the transcripts had significant matches in TAIR databases. We identified candidate genes involved in ginsenoside biosynthesis: 10 transcripts for farnesyl diphosphate synthase to protopanaxatriol synthase and 21 transcripts for UDP-glycosyltransferase. A large number of transcripts (17%) with different GO designations were uniquely detected in adventitious roots compared to normal ginseng roots. In addition, 10,213 and 7,928 cDNA SSRs were identified as potential molecular markers in CP and CS, respectively. Our assembly of ginseng transcriptomes demonstrates the successful application of genomics approaches to large complex genomes. In addition, we have predicted the long noncoding(lncRNA) based on our CP RNA-Seq data. A total of 11,270 lncRNA were identified. Among them, some were precursors of small RNAs such as microRNAs and siRNAs. The assembly and comparative analysis data have been deposited to our newly created adventitious root transcriptome database (http://im-crop.snu.ac.kr/transdb/index.php) for public use. Further, To better understand our ginseng genome (*P. ginseng*), we used Chunpoong (CP) adventitious root RNA-Seq data to identify lncRNAs of *P. gisneng*. we found 11,270 long noncoding RNAs which had multiexonic structures. A total of 433 lncRNAs showed significant similarity against publicly available lncRNA database of Arabidopsis, Maize, lncRNA database and Rfam. My study provides a preliminary source for future studies of lncRNA content and function in ginseng.

Key words: Transcriptome, next generation sequencing, *De novo* assembly, *Panax ginseng*

Student number: 2012-22610

CONTENTS

CHAPTER 1. Transcriptome Profiling and Comparative Analysis of *Panax ginseng* Adventitious Roots

CHAPTER 2. Computational prediction of long noncoding RNAs (lncRNAs) in *P. ginseng*

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CP:	Chunpoong
CS:	Cheongsun
SRA:	Sequence Read Archive
GO:	Gene Ontology
ART:	Adventitious Root Transcriptome
RPKM:	Reads Per Kilobase per Million
SSR:	Simple Sequence Repeat
lncRNA:	Long noncoding RNA
CPC:	Coding Potential Calculator

CHAPTER 1

Transcriptome Profiling and Comparative Analysis of *Panax ginseng* Adventitious Roots

Abstract

Panax ginseng (C. A. Meyer) is a traditional medicinal plant famous for its strong therapeutic effects. However, genomic resources for *P. ginseng* are still very limited. In this study, we performed *de novo* assembly of transcriptomes from adventitious roots of two *P. ginseng* cultivars, Chunpoong (CP) and Cheongsun (CS). The assemblies were generated from ~85 and ~77 million high-quality Illumina HiSeq reads from CP and CS cultivars, respectively. A total of 35,527 and 27,716 transcripts were obtained from the CP and CS assemblies, respectively. Annotation of the transcriptomes showed that approximately 90% of the transcripts had significant matches in TAIR databases. We identified candidate genes involved in ginsenoside biosynthesis: 10 transcripts for farnesyl diphosphate synthase to protopanaxatriol synthase and 21 transcripts for UDP-glycosyltransferase. A large number of transcripts (17%) with different GO designations were uniquely detected in adventitious roots compared to normal ginseng roots. In addition, 10,213 and 7,928 cDNA SSRs were identified as potential molecular markers in CP and CS, respectively. Our assembly of ginseng transcriptomes demonstrates the successful application of genomics approaches to large complex genomes. In addition, we have predicted the long noncoding(lncRNA) based on our CP RNA-Seq data. A total of 11,270 lncRNA were identified. Among them, some were precursors of small RNAs such as microRNAs and siRNAs. The assembly and comparative analysis data have been deposited to our newly created adventitious root transcriptome database (http://im-crop.snu.ac.kr/transdb/index.php) for public use.

Key words: Transcriptome, next generation sequencing, *De novo* assembly, *Panax ginseng,* lncRNA

Introduction

Panax ginseng (ginseng), one of 17 species in the *Panax* genus of the Araliaceae family, has been widely used as a source of medicine in eastern Asia and North America[1], [2]. Ginseng serves as an adaptogen, with effects on immune system stimulation [3], [4], anti-cancer activity [5], and anti-hyperlipidemic effects [6]. *P. ginseng* is a deciduous perennial with red or orange berries and yellowish-brown roots [7], [8], [9]. The thick roots of *P. ginseng* contain the medicinally active triterpene glycosides or saponins, commonly referred to as ginsenosides. Despite growing demand in the pharmacological industry, very limited genomic information is available for *P. ginseng*. It has been reported to be a paleo-tetraploid with an estimated genome size of ~3.2 Gb [10]. To date, ten varieties have been bred by the Korea Ginseng Corporation and registered as commercial cultivars with the Korea Seed and Variety Service. Among them, 'Chunpoong' is a very pure inbred line with relatively low heterozygosity, high yield, and superior quality [11]. Genetic study in ginseng has been challenging because of its long generation time (4 years/generation), the small numbers of seeds it sets (40 seeds/plant), and the difficulty of maintaining ginseng in the field. As an alternative, various tissue culture methods, including callus, hairy root, and adventitious root culture systems, have been adapted for mass production of ginseng. Among these, adventitious root culture has been a promising alternative for production of ginsenoside because the total saponin contents of the adventitious roots are comparable to those of field-grown roots and higher than those of callus and hairy roots[12]. Moreover, mass production of adventitious roots is well-established through a balloon-type bubble bioreactor (BTBB) system [12].

A few expressed sequence tag (EST) libraries have been generated and 17,114 EST sequences are present in the dbEST database at NCBI (release 130101; 01 January 2013). Most of the ESTs have been generated with the aim of identifying genes involved in ginsenoside biosynthesis and developing molecular markers [13], [14], [15], [16]. In addition, several efforts have been made to construct BAC libraries and BAC-end sequences for physical mapping, positional cloning and sequencing of the genome [17], [18]. However, sequencing and assembling such a large and complex genome is costly and requires high computational power. Fortunately, next generation sequencing (NGS) technologies have revolutionized genome analysis and made it possible to sequence cDNA (RNA-Seq) and examine cellular transcriptomes along with high-throughput gene expression analysis [19], [20]. To date, a few studies have applied NGS technology to transcriptome analysis of *Panax* species including *P. notoginseng* [21], *P. quinquefolius* [22] and *P.* ginseng [23]. These studies used the 454 sequencing platform mainly to identify ginsenoside biosynthetic genes in the normal root transcriptome. Gene discovery and comparative gene expression profiling was very limited in the previous studies.

Here, we used the Illumina sequencing platform for large-scale transcriptome analysis and present *de novo* adventitious root transcriptome assemblies for Chunpoong (CP), which is the oldest elite cultivar in Korea, and Cheongsun (CS), which is a superior cultivar for adventitious root production. We assembled CP and CS transcriptomes from the millions of short sequence reads generated by Illumina paired-end transcriptome sequencing. After annotation, we conducted gene expression profiling, as well as identification of candidate genes involved in ginsenoside biosynthesis. This work provides the first transcriptome profiles of *in vitro*-grown adventitious roots of two ginseng cultivars. It also describes an advanced method for transcriptome assembly and validation in non-model plant species and for the study of genes related to secondary metabolites, which can be affected greatly by small modifications of environment conditions.

Materials and Methods

Plant material and RNA isolation

Stratified seeds of Korean ginseng cultivars Chunpoong (CP) and Cheongsun (CS) from KGC (Korea ginseng Corporation, Daejeon, Korea) were sterilized by immersion in 70% ethanol for 1 min and 2% sodium hypochlorite for 10 min. After each step, seeds were gently washed with distilled water three times. All procedures were aseptically performed in a laminar hood. To induce adventitious roots, cotyledons separated from sterilized stratified seeds were cultured on solid Schenk & Hilderbrant (SH) media containing 2.0 mg L^{-1} indole butyric acid (IBA), 3% sucrose, and 0.23% Gelrite. After 1 month, induced adventitious roots were separated from cotyledon explants and cultured again for secondary growth on the same medium. Then, the roots were transferred to 30 mL liquid SH medium supplemented with 3.0 mg L^{-1} IBA and 5% sucrose and maintained on rotary shaker (100 rpm) at 25℃ in the dark. For further mass production, 12 g fresh adventitious roots in suspension culture were inoculated into a 2-L airlift balloon type bioreactor (Biopia, Korea) containing 1 L of the same SH medium as used for liquid suspension culture (Fig. 1). The medium was replaced with fresh medium after 2 weeks, and 4 weeks later, 12 g adventitious roots were sub-cultured into a new bioreactor. After 10 d of cultivation, the sub-cultured adventitious roots were used for total RNA extraction with the Plant RNeasy mini Kit (QIAGEN, Germany) according to manufacturer's instructions. Approximately 2 ug total RNA from each cultivar were used for sequencing on the Illumina platform after the quality and quantity were checked using spectrophotometry.

Illumina sequencing and quality control

Paired-end (PE) reads with an average of 101 bp were generated for CP and CS using the Illumina Hiseq2000 platform. The library construction and sequencing was performed by the National Instrumentation Center and Environmental Management (NICEM), Seoul National University, South Korea. The sequence data generated in this study have been deposited at NCBI in the Short Read Archive (SRA) with accession number SRA061905. The sequencing reads underwent various stringent quality controls such as filtering of high-quality reads and removal of reads with an adaptor or primer-contaminated sequence using the NGS QC Toolkit [24].

De novo assembly

All *de novo* assemblies were performed on a server with 48 cores and 512 GB random access memory (RAM). Publicly available transcriptome and genome assemblers were used to assemble the PE reads. Among the transcriptome assemblers, the open source program, Oases [25] (version: 0.2.06; http://www.ebi.ac.uk/~zerbino/oases), which uploads a preliminary assembly produced by Velvet, was validated for k-mer optimization. Various assembly parameters were also examined to yield statistically as well as biologically significant results. In addition, other publicly available transcriptome assemblers were used to determine the best assembler for the CP data set. This included Trinity [26] (package: trinityrnaseq_r2012-04-27; http://trinityrnaseq.sourceforge.net), SOAPdenovo-Trans (version: 1.01; http://soap.genomics.org.cn/SOAPdenovo-Trans.html) and genome assemblers that also had been used for *de novo* transcriptome assembly, such as ABySS [27] (version: 1.3.3; http://www.bcgsc.ca/platform/bioinfo/software/abyss) and commercially available CLC Genomics workbench [28] (version: 5.1). The data for the Cheongsun (CS) cultivar were assembled using the assembler that was identified as the best from the CP cultivar assembly.

Functional annotation and analysis

The assembled CP and CS transcript sequences were annotated by sequence comparison with well annotated protein databases. All assembled transcripts were searched against the NCBI non-redundant protein (nr) database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) using BLASTX with an E-value cut-off of 1E-05. In addition, CP and CS transcripts were searched against the Uniprot (TrEMBL and Swissprot) (ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz) and TAIR (The Arabidopsis Information Resource; ftp://ftp.arabidopsis.org/home/tair/Proteins/TAIR10_protein_lists/TAIR10_pep_20101214) databases using the BLASTX search with cutoff E-values of 1E-05 and 1E-10. Transcripts were functionally classified following the Gene Ontology scheme (http://www.geneontology.org). The Blast2GO program [29] was used to determine the molecular function, biological process, and cellular component categories associated with the best BLASTX hit in the nr database for the corresponding CP and CS transcripts.

Expression profiling

Trimmed raw reads were mapped onto their assembled transcripts to quantify transcript abundance using the CLC Genomics Workbench (version 5.1). The number of reads and reads per million (rpm) were determined using the CLC mapping program. Further, RPKM (reads per kilobase per million) for each transcript and average RPKM were determined [30]. In addition, the CP and CS transcripts were compared based on RPKM values. The assembled transcripts of CP were used as references to determine the RPKM of transcripts in both cultivars.

Identification of candidate transcripts involved in ginsenoside biosynthesis

P. ginseng gene sequences that were reported to be involved in the biosynthesis of ginsenosides were collected from GenBank. The amino acid sequences of these genes were used as queries to search for homologous sequences in the CP and CS assembled transcript datasets using the TBLASTN program. Candidate transcripts were identified based on E-value, bit score, alignment length and further validation using BLASTP.

SSR identification

SSRs in CP and CS transcripts were identified using a Perl script referred to as the MIcroSAtellite identification tool (MISA, http://pgrc.ipk-gatersleben.de/misa/). The following criteria for SSR identification were used in the MISA script: mono-nucleotides repeated more than ten times, di-nucleotides repeated more than six times, and tri-, tetra-, penta- or hexa-nucleotides repeated more than five times.

Adventitious root transcriptome (ART) database creation

The ART database was created to serve as a public web resource for ginseng transcriptome data. This database was designed using PHP (v4.3.9) and MySQL (v4.1.20). The front-end language PHP was connected with back-end MySQL by the Apache web server. The annotation, expression and marker data are stored in MySQL as tables, and this database is currently hosted on a CentOS (v5.8) Linux operating system. This database can be accessed at http://im-crop.snu.ac.kr/transdb/index.php.

Results

Adventitious root growth and morphology of two ginseng cultivars

We obtained adventitious roots from the cotyledons of CP and CS cultivars. Although the same culture conditions were used for both cultivars, the adventitious root induction rate of CS was 9% higher than that of CP (Fig. 1-1E). In addition, the cultivars showed different adventitious root morphology during proliferation in bioreactor culture. CP adventitious roots appeared to be dark yellow, callus-like clumps (Fig. 1-1D), while those of CS showed typical adventitious root morphology and were bright yellow (Fig. 1-1C). This indicates that CS is better suited for adventitious root induction and growth under these conditions.

De novo assembly and validation of Illumina paired-end sequences

We generated a total of 90,242,024 and 82,011,294 raw reads from CP and CS, respectively (Table 1-1). After trimming the low-quality reads with Phred quality scores ≤ 25 and removing primer/adaptor sequences, we obtained 85,335,736 (94.5%) and 77,583,736 (94.6%) high-quality reads with an average read length of 99 bp in CP and CS, respectively (Table 1-1).

The average quality score for each base of the sequence reads increased significantly after filtering.

To obtain high quality assemblies, we tested several algorithms for *de novo* assembly with different options. We used several criteria to determine desirable assembly: number of reads used in assembly, total length of transcriptome, average contig length, N50, and annotation by BLASTX against the TAIR protein database.

Using Velvet followed by Oases, we compared assembly results with randomly selected *k*-mer lengths of 31, 39, 41, 49, 51, 59, 61, 69, 71 and 79. The best assembly was obtained at *k*=69, as it resulted in the highest total length (~138 Mbp), the largest N50 length (1,092 bp), the largest average contig length (19,999 bp) and a significant number of TAIR hits (74.79%). In addition to Oases assembly, we also used Trinity (*k*=25 as a fixed option), SOAP-Trans, ABySS, and the CLC Genomics workbench with default parameters. We also compared the assembly results by mapping all raw reads onto each assembly in order to determine the read usage. We obtained the best assembly results from Oases and Trinity, as they showed the largest assembled transcriptome sizes, numbers of mapped reads, average contig lengths and numbers of TAIR hits (Blastx). Since these two assemblers use *k*-mer based approaches for assembly, it is common to have mis-assemblies in both datasets. Hence, we analyzed the read depth for the Oases and Trinity datasets to identify mis-assemblies. We determined the percentage read depth by dividing the number of mapped reads for each assembled transcript by the effective length (without 'N') of the transcript. We removed the transcripts that had <1% read-depth in both datasets because those were considered not to be correctly assembled from paired reads. For further evaluation of the accuracy of the datasets, we compared both against 108 *P.*

ginseng full-length gene sequences retrieved from GenBank. We used BLAST searches to check whether the start and stop codons of the full-length gene sequence were present in both datasets. Large numbers of full-length sequences (including untranslated regions) were found in the Trinity dataset with 95-100% identity. We found that many truncated transcripts (without the start and stop codons) were included in Oases dataset. The extracted dataset sequences were also successfully mapped onto our ongoing *P. ginseng* draft genome sequence assembly using the BLAST algorithm. The Trinity dataset showed more hits and a higher percentage of identity than the Oases dataset, demonstrating that Trinity was the best assembler for our transcriptome assembly. Using Trinity, we obtained 35,527 CP transcripts with an average length of 1,978 bp and 27,716 CS transcripts with an average length of 1,980 bp (Table 1-1). The lengths of the assembled transcripts ranged from 400 to 15,980 bp with large numbers of transcripts in range of 1000 to 2000 bp in CP as well as CS. We identified sets of 14,051 and 11,209 non-redundant transcripts from the CP and CS assembled datasets, respectively, by selecting only the longest sequence among isoforms that included alternatively spliced forms predicted by the Trinity assembler. We used the total assembled transcripts including isoforms for further analysis because it was difficult to select the optimal representative nr dataset among various isoforms without a *P. ginseng* reference sequence.

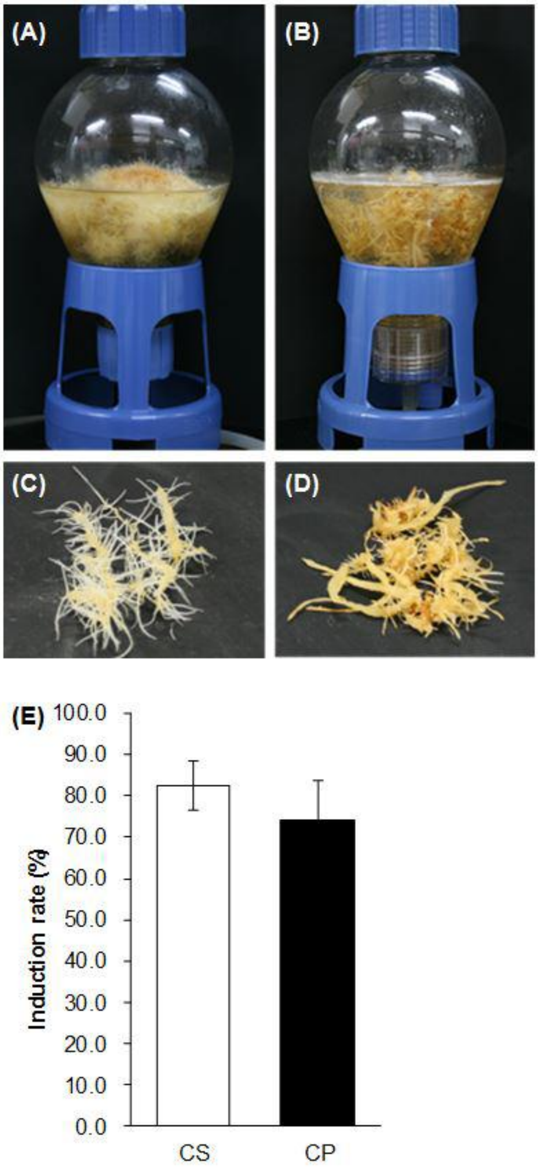


Figure 1-1. Adventitious roots of CP and CS 2 weeks after inoculation in 2-L balloon-type bubble bioreactors (BTBBs). Adventitious roots of CS showed bright yellow, root hair-like morphology (A, C), whereas those of CP were dark yellow, with callus-like morphology (B, D). (E) The rate of adventitious root induction from cotyledons. Data are means with standard deviation from four independent experiments.

Table 1-1. Summary statistics of Chunpoong (CS) and Cheongsun (CS) sequencing, assembly and validation.

Illumina sequencing data		
Parameters	CP	CS
Number of raw reads	90,242,024	82,011,294
Number of trimmed reads	85,335,736	77,583,736
Trinity assembly		
Total number of transcripts	35,527	27,716
Total transcriptome size (bp)	70,295,564	54,892,571
Non-redundant set (without isoforms)	14,051	11,209
Small transcript length (bp)	410	411
Large transcript length (bp)	15,918	15,980
Average transcript length (bp)	1,978	1,980
N50 length (bp)	2,274	2,277
Validation by BLASTX similarity searches (E-value cutoff of 1E-05)		
Nr (NCBI)	33,718 (94%)	26,513 (95%)
TAIR (Arabidopsis)	32,996 (92%)	25,996 (93%)
Swiss-prot	27,745 (78%)	21,923 (79%)

Functional annotation and classification

For further validation and annotation of assembled transcripts, sequence similarity searches were conducted against TAIR and Uniprot (SwissProt and TrEMBL) protein databases using the

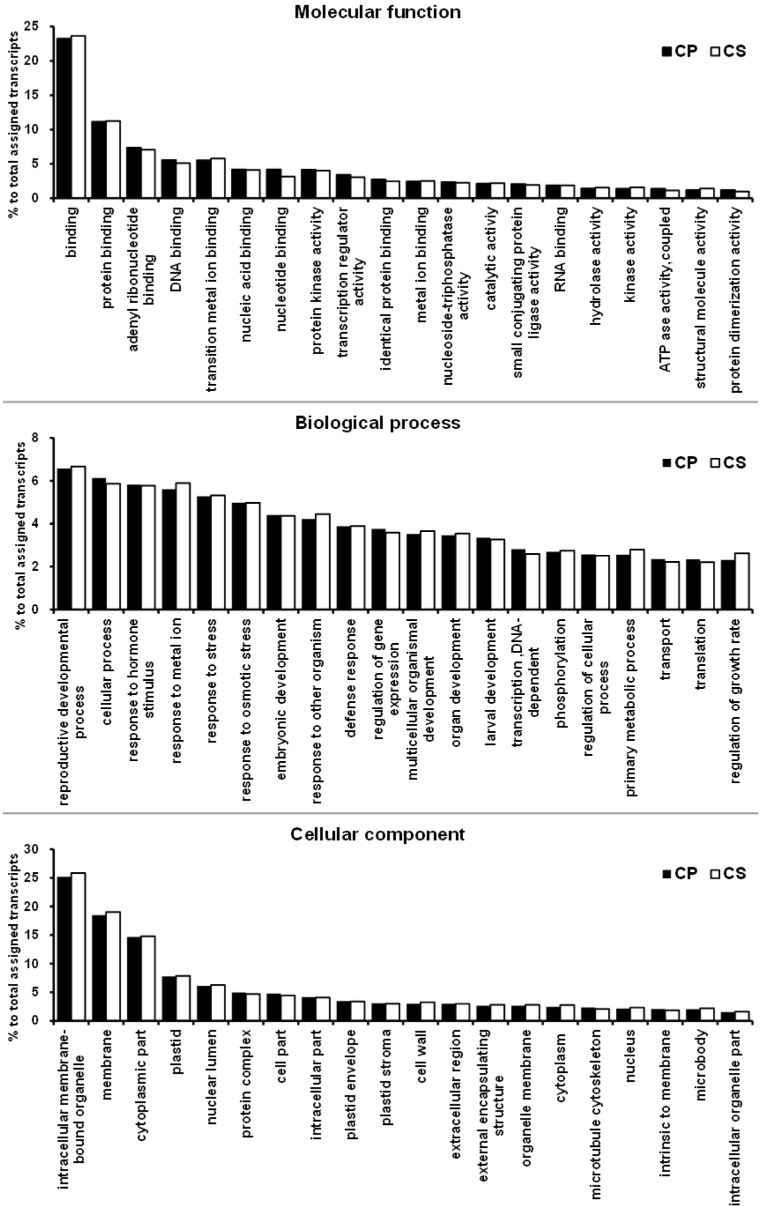


Figure 1-2. GO analysis of transcripts expressed in CP and CS adventitious roots. Terms in the molecular function (A), biological process (B), and cellular component (C) categories are shown. A total of 26,423 CP and 21,096 CS transcripts were assigned to GO terms using Blast2GO with whole assembled transcripts of CP (35,527) and CS (27,716).

To classify the predicted functions of the transcripts, Gene ontology (GO) terms were assigned to CP and CS transcripts using Blast2GO, based on their similarity to nr database proteins. A total of 26,423 (74.37%) CP transcripts were assigned to GO classes. Of those, assignments to the cellular component class ranked the highest (22,706, 63.91%), followed by biological process (22,215, 62.53%) and molecular function (21,560, 60.68%). In CS, a total of 21,096 (76.11%) transcripts were assigned at least one GO term, and among them, 17,512 (63.18%), 17,249 (62.23%), and 18,178 (65.58%) were assigned at least one GO term in the biological process, molecular function, and cellular component category, respectively. Binding was the most abundant GO Slim within the molecular function category (Fig. 1-2A). Reproductive development, cellular process, and stress response were the most abundant among various biological processes (Fig. 1-2B). Intracellular membrane-bound organelle and membrane were the most highly represented GO terms in the cellular component category (Fig. 1-2C).

Gene expression profiling

Transcript quantification, also called digital gene expression, is an efficient approach for gene expression profiling [31], [32]. We mapped all of the CP and CS reads onto their respective assembled transcripts in order to determine the RPKM. For the CP transcripts, the RPKM ranged from 0.16 to 4609 with an average of 15.93, and the RPKM for CS ranged from 0.22 to 4118 with an average of 19.90. This indicates that both CP and CS transcripts showed wide range of expression levels from very low to strong expression. However, over 97% of transcripts were in the RPKM range < 100 (Fig. 1-3A), of which 1,244 (3.5%) and 585 (2.1%) had RPKM values below 1.0 for CP and CS, respectively. We compared the expression patterns of the transcripts in CP and CS cultivars based on RPKM value differences for each transcript in the cultivar datasets. As shown in Fig. 1-3B and 1-3C, the transcripts showed very similar expression patterns in both cultivars, which was also confirmed by the high correlation co-efficiency value of more than 0.9 between both datasets.

To identify highly expressed transcripts and their putative functions, we selected the 100 most abundant transcripts based on their RPKM values in the CP and CS datasets and investigated the biological processes in which those transcripts might be involved. Although many transcripts (15 in CP and 23 in CS) could not be assigned to known biological process, most (52 in CP and 51 in CS) were involved in stress response and protein metabolism, including pathogenesis-related proteins, antioxidant enzymes, heat-shock proteins, and metallothionein-like protein in the stress response category and translation- and protein degradation-related proteins in the protein metabolism category (Fig. 1-4). After those, transcripts related to lipid metabolism, such as fatty acid desaturases and lipid transfer proteins, were most abundant.

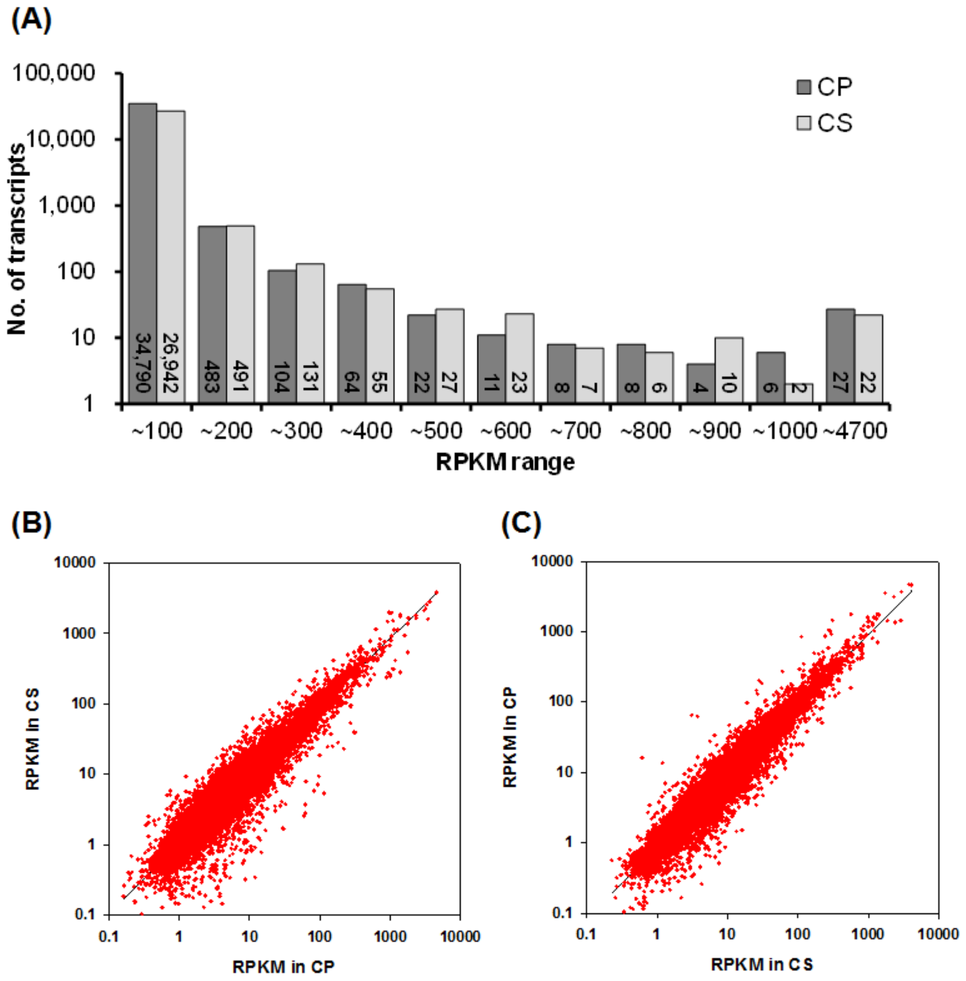


Figure 1-3. RPKM values of transcripts expressed in CP and CS adventitious roots. RPKM values were calculated after reads of both cultivars were mapped onto the respective assembled transcripts. (A) Distribution of RPKM values. Number of transcripts belonging to the RPKM range defined in the X-axis is shown in the bottom of each bar. Y axis is in log-scale. Transcripts showing RPKM values of less than 1.0 were included. (B) RPKM value comparison calculated by mapping reads of both cultivars onto CP transcripts as references. (C) RPKM value comparison calculated by mapping reads of both cultivars onto CS transcripts as references. Red spots

indicate transcripts expressed in both cultivars. Regression lines (r^2=0.92 in (B) and r^2=0.94 in (C) at 99% confidential level) were calculated and depicted using Sigmaplot s/w. X and Y axes are in log-scale.

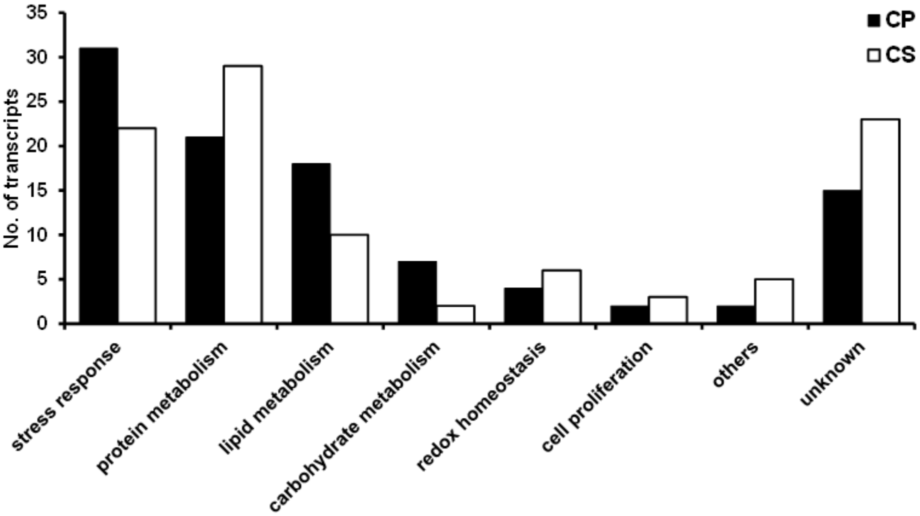


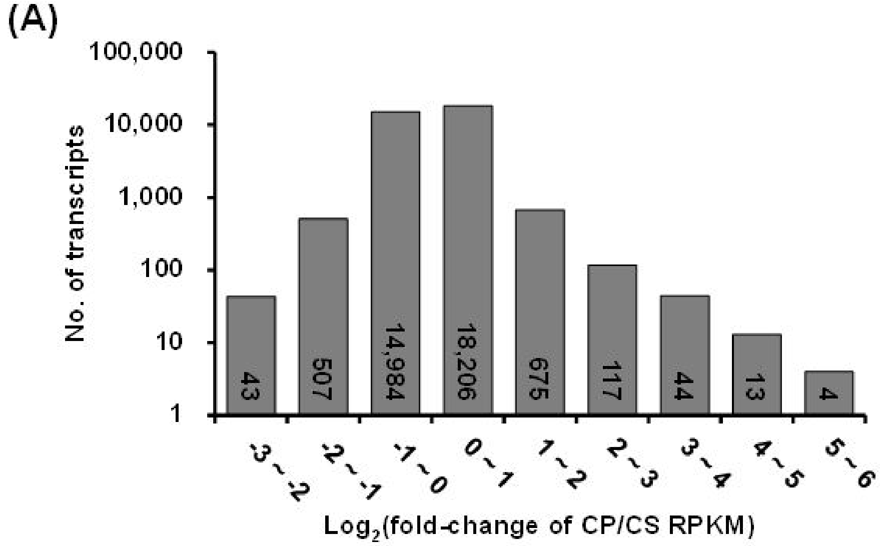
Figure 1-4. Functional distribution of the 100 most highly expressed transcripts in CP and CS adventitious roots. The top 100 transcripts were chosen on the basis of their expression level determined by RPKM values and then biological processes in which those transcripts are putatively involved were predicted based on GO assignment and BLASTX searches.

Transcripts showing biased expression between CP and CS adventitious roots

To identify genes with differential expression between CP and CS cultivars, the raw sequencing reads were mapped to 35,527 CP transcripts, followed by calculation of RPKM values using the CLC mapping program. We removed transcripts with RPKM values less than 1.0 in both cultivars.

The remaining 34,594 were used to calculate fold-change values for expression between both cultivars, by dividing the RPKM of the CP transcript by that of the CS transcript. The distribution of fold-change values is summarized in Fig. 1-5. Among all of the transcripts analyzed, more than 95% did not show expression differences greater than 2-fold between the two cultivars. A total of 1,403 (4.06%) showed more than 2-fold differences in expression, of which 853 (2.47%) and 550 (1.59%) were more highly expressed in CP and CS cultivars, respectively.

The transcripts showing the highest fold differences between the two cultivars were selected for biological process evaluation. The 100 CP transcripts showing the greatest bias in expression (the top 100 CP-biased transcripts) were 6.3 to 41.5-fold increased relative to their expression in CS, whereas the top 100 CS-biased transcripts showed 3.1 to 7.6-fold increases relative to that in CP. As found for the top 100 most abundant transcripts (Fig. 1-4), most transcripts were not assigned to known biological process categories (Fig. 1-6). Among the CP- and CS-biased transcripts, 22 and 35, respectively, were designated as unknown transcripts. Another 19 CP and 7 CS transcripts encoded proteins related to stress response, such as pathogenesis-related proteins and peroxidase. In addition, some transcripts involved in terpenoid biosynthesis, protein metabolism, redox homeostasis, cell proliferation, and transportation were present in both datasets, although their proportions were different depending on cultivar type. The top 100 lists also included 25 CP-unique and 9 CS-unique transcripts. The 25 CP-unique transcripts were putatively involved in photosynthesis, auxin response, circadian rhythms, and thiamine biosynthesis, whereas the 9 CS-unique transcripts were related to suberin biosynthesis, transcription, and transposons.



(B)

	in CP	in CS
Not changed (%)	33,190 (95.94)	
2-fold up (%)	853 (2.47)	550 (1.59)
4-fold up (%)	178 (0.51)	43 (0.12)
8-fold up (%)	61 (0.18)	0 (0.00)

Figure 1-5. Number of transcripts expressed differently between CP and CS adventitious roots. (A) Distribution of fold-change values. Number of CP transcripts in the indicated fold-change range is shown inside the bars. (B) Number of CP transcripts expressed more than 2-, 4-, and 8-fold differently between the adventitious roots of both cultivars. Percentage of the total CP transcripts (34,594) is shown in parentheses.

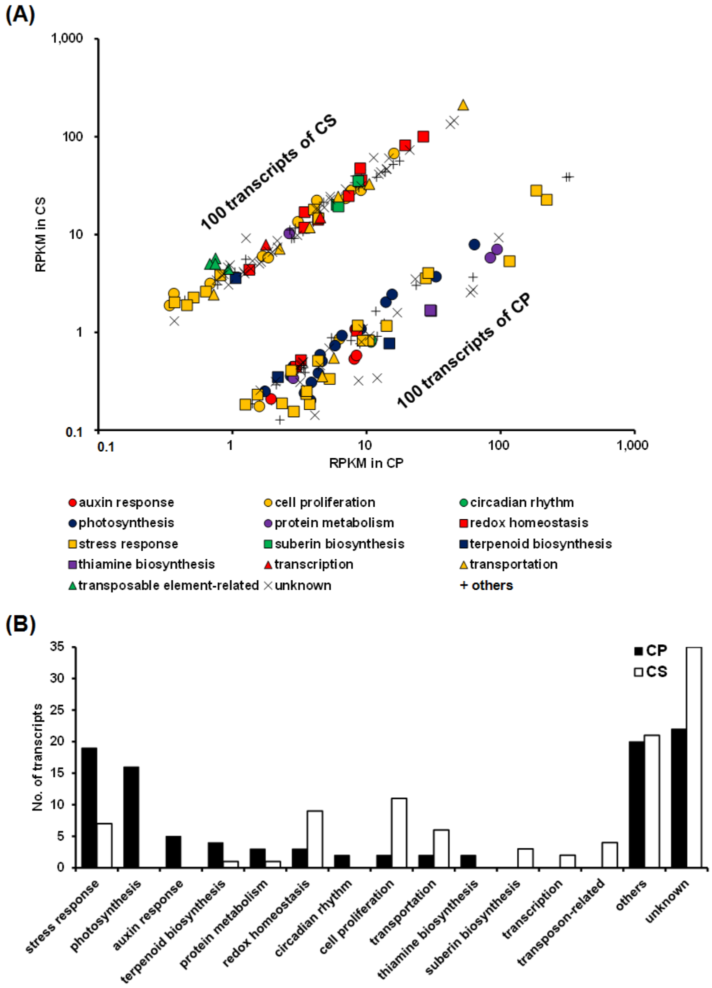


Figure 1-6. Functional distribution of 100 transcripts expressed differently between CP and CS adventitious roots. (A) Top 100 transcripts biased toward CP or CS were chosen among those expressed differently with fold-change of more than 2 between the two cultivars. X and Y axes are in log-scale. (B) Biological processes in which the transcripts are putatively involved were predicted based on GO assignment and BLASTX searches. Among biological processes examined, those that showed high differences in number of transcripts assigned were used for this graph; the remaining terms and unknown biological process are denoted by “others” and “unknown”.

Identification of candidate genes involved in ginsenoside biosynthesis

Ginsenosides are the most important phytochemicals in ginseng and are known to be synthesized through the mevalonic acid (MVA) pathway [33]. We focused on downstream enzymes from farnesyl diphosphate synthase (FPS) to UDP-glycosyltransferase (UGT) in the MVA pathway (Fig. 1-7A). In previous studies, 17 genes for the 7 downstream enzymes (FDS to protopanaxatriol synthase) have been reported in *P.* ginseng [34], [35], [36], [37], [38], [39], [40], [41]. We used amino acid sequences of the 17 genes as queries for TBLASTN searches against transcript datasets of the CP cultivar, resulting in identification of 10 genes encoding the 7 downstream enzymes. Of them, a single transcript for FDS was identified with 15 isoforms in the CP dataset. Squalene synthase (SQS), dammarenediol synthase (DDS), β-amyrin synthase (β-AS), protopanaxadiol synthase (CYP716A47), and protopanaxatriol synthase (CYP716A53v2) were also identified to be encoded by single transcripts with several isoforms. Exceptionally, four transcripts were identified for squalene epoxidase (SQE). Although we identified the isoforms using a reliable algorithm (Trinity assembler), the forthcoming *P. ginseng* genome sequence will provide more solid information about them. Based on our analysis, we considered the isoforms likely to originate from a single gene.

To investigate the expression levels of the transcripts, the RPKM values of isoforms from the same transcripts were averaged and compared (Fig. 1-7B). All showed similar expression levels between CP and CS cultivars, with transcripts encoding cytochrome P450 for protopanaxatriol synthase showing the highest expression in both cultivars.

Three UGT proteins, SvUGT74M1, MtUGT73K1, and MtUGT71G1, were used as queries for TBLASTN searches, because *UGT* genes for ginsenoside biosynthesis had not been identified in *P. ginseng*. Three UGT proteins were reported to function in triterpene saponin biosynthesis in *Medicago truncatula* and *Saponaria vaccaria* [42], [43]. Among the transcript hits, a total of 42 isoforms were selected based on their Blast bit score of more than 200. Of those, 21 transcripts remained after removing redundant isoforms based on their similarity at the amino acid level. Phylogenetic analysis of the deduced protein sequences of the 21 transcripts revealed that some transcripts were closely grouped with the three reported UGT proteins (Fig. 1-8A). In particular, three transcripts, CP_comp126017_c1_seq1, CP_comp142900_c0_seq2, and CP_comp82124_c0_seq2, showed much higher similarity to the reported UGT proteins than did the other transcripts. Overall, the expression patterns of the 21 transcripts were similar between CP and CS, with the exception of CP_comp144124_c0_seq10, which showed 2-fold higher expression in CP (2.53 in CP vs. 1.26 in CS) (Fig. 1-8B).

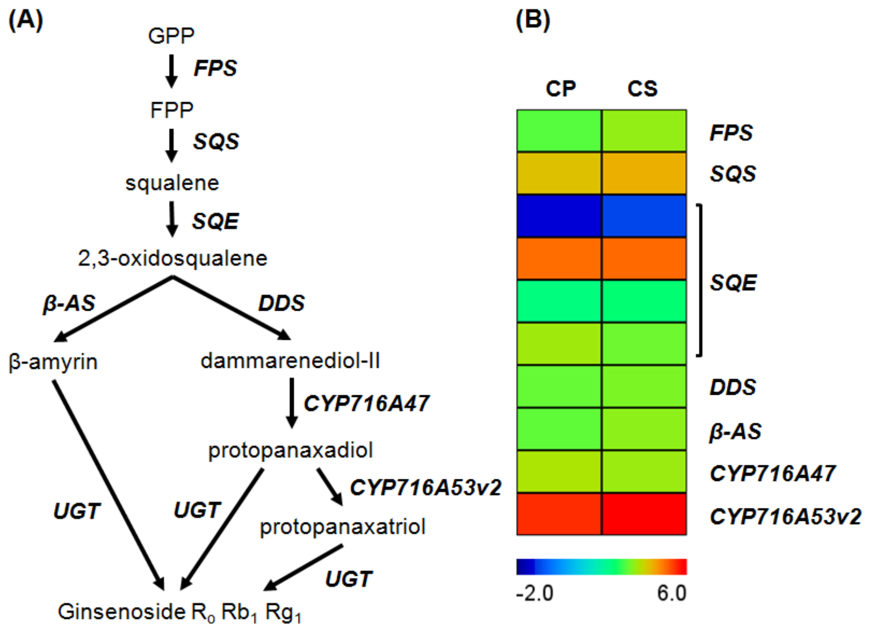


Figure 1-7. Putative ginsenoside biosynthesis pathway and expression level of ginsenoside biosynthesis genes found in transcript datasets of CP and CS adventitious roots. (A) Candidate genes identified in this study are shown in bold. GPP, geranyl diphosphate; FPS, farnesyl diphosphate synthase, FPP, farnesyl diphosphate; SQS, squalene synthase; SQE, squalene epoxidase; β-AS, beta-amyrin synthase; DDS, dammarenediol synthase; CYP716A47, cytochrome P450 for protopanaxadiol synthase; CYP716A53v2, cytochrome P450 for protopanaxatriol synthase; UGT, UDP glycosyltransferase. (B) Expression levels of the candidate genes in both cultivars. Several isoforms of the candidate genes were identified, and RPKM values were averaged and represented with the color scale shown below. The expression map was generated by using MeV s/w (http://www.tm4.org/mev/) with $\log_2$ (RPKM) values.

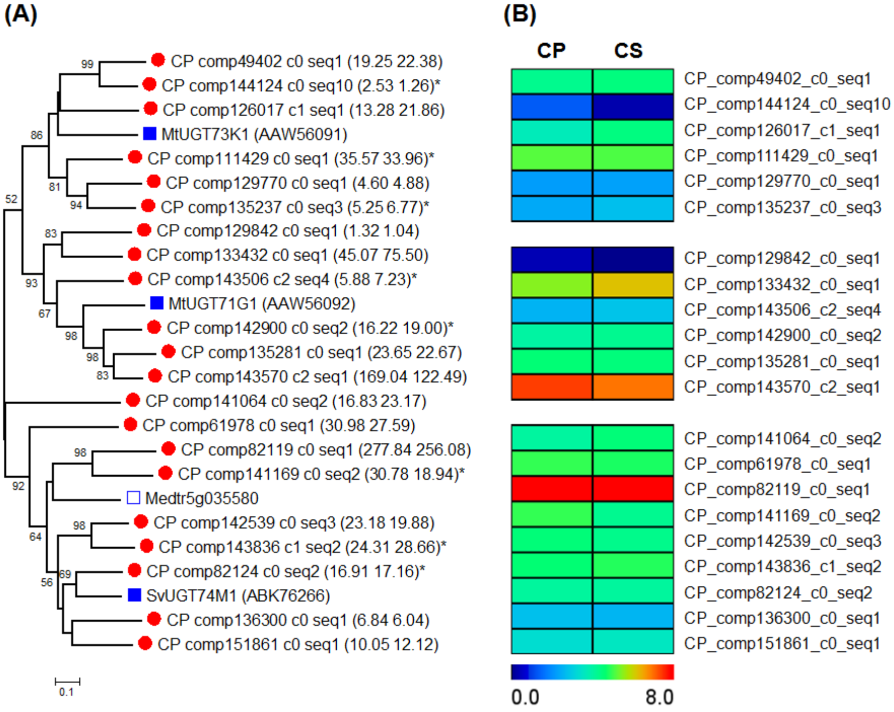


Figure 1-8. Phylogenetic analysis of putative UGT proteins and comparison of their expression levels. (A) Twenty-one transcripts predicted to encode proteins with high similarity to UGT proteins involved in ginsenoside biosynthesis, SvUGT74M1 (ABK76266), MtUGT73K1 (AAW56091), and MtUGT71G1 (AAW56092), were identified in CP transcript dataset. The deduced amino acid sequences were aligned using ClustalW and the phylogenetic tree was generated using Poisson correction and the neighbor-joining (NJ) method in MEGA5. Bootstrap values calculated for 1000 replicates are shown on the branches; the values less than 50% are not shown. RPKM values of transcripts are shown in parentheses next to transcript name. Asterisk indicates that the transcript is a representative one selected among their isoforms and its RPKM is an average value calculated from RPKM values of its isoforms. A *Medicago truncatula* homologue of

SvUGT74M1, Medtr5g035580, was identified in *M. truncatula* genome database (http://medicago.jcvi.org/cgi-bin/medicago/overview.cgi) and then included in the tree. (B) Expression comparison of the UGT genes in both cultivars. The RPKM values of transcripts closely grouped in the phylogenetic tree were compared and depicted by the expression map using MeV s/w (http://www.tm4.org/mev/) with $\log_2$ (RPKM) values.

Comparative analysis of the transcriptomes of adventitious and normal roots

To investigate the transcript expression differences between adventitious roots and primary roots, we compared 35,527 CP reference transcripts with 38,966 transcripts from 11-year-old ginseng primary roots, after assembly of 454 reads from the NCBI SRA database (accession no. SRX017443) [44]. When their sequence similarity was analyzed, 6,057 (17.0%) transcripts in adventitious roots and 6,354 (16.3%) in primary roots were found to be uniquely expressed. Of the 62,082 total transcripts, 29,470 (83.0%) from adventitious roots and 32,612 (83.7%) from primary roots, were commonly expressed. GO analysis of unique transcripts was performed to characterize their functional category. As shown in Fig. 1-9, more transcripts from adventitious roots were assigned GO terms than from normal roots. Overall, the proportion of GO assignment in adventitious root transcriptomes was 2-fold higher than that of normal roots although the most frequent GO terms such as binding, response to other organism, and nuclear lumen were generally similar between both datasets. In particular, 11 out of 20 GO terms for biological process had more transcripts in adventitious roots than in normal roots. Terms such as response to metal ion, transcription, multicellular organismal development, and reproductive developmental process showed more than 8-fold higher proportions than in normal roots. On the other hand, only two biological process terms, regulation of growth rate and response to stress, accounted for higher proportions in normal roots than in adventitious roots.

Identification of cDNA-derived SSR markers

EST-SSR markers are useful for genetic diversity analysis, marker assisted selection, and genetic mapping. We identified a total of 10,213 SSRs in 8,347 (23.49%) CP transcripts, of which 1,523 transcripts contained more than one SSR (Table 1-2). In addition, 464 SSRs were found in compound form (Table 2). The largest fraction of SSR motifs were di-nucleotide (37.73%) followed by mono-nucleotide (30.97%) and tri-nucleotide (25.72%) SSRs. We also identified tetra (396), penta (86) and hexa-nucleotide (87) SSR motifs. Similarly, a total of 7,928 potential SSRs were identified in 6,479 (23.37 %) CS transcripts, with 1,179 transcripts having more than one SSR and 376 SSRs in compound form. The proportion of repeat types in CP and CS cultivars followed similar patterns and the identified SSRs provide a cost-effective method for development of functional markers in ginseng.

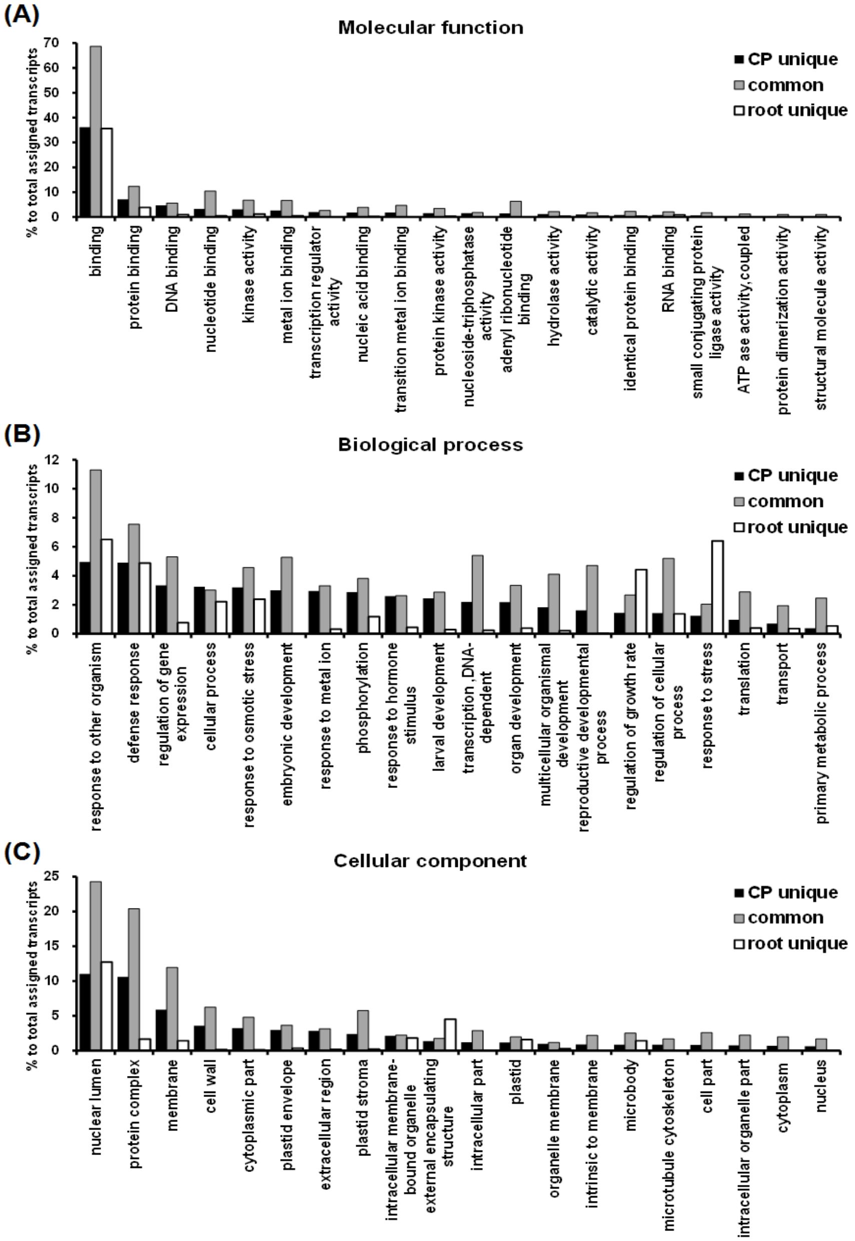


Figure 1-9. GO analysis of transcripts common and unique to CP adventitious roots and normal roots. Terms are presented based on molecular function (A), biological process (B), and cellular component (C). A total of 3,241 CP-unique and 2,378 root-unique transcripts were assigned GO terms

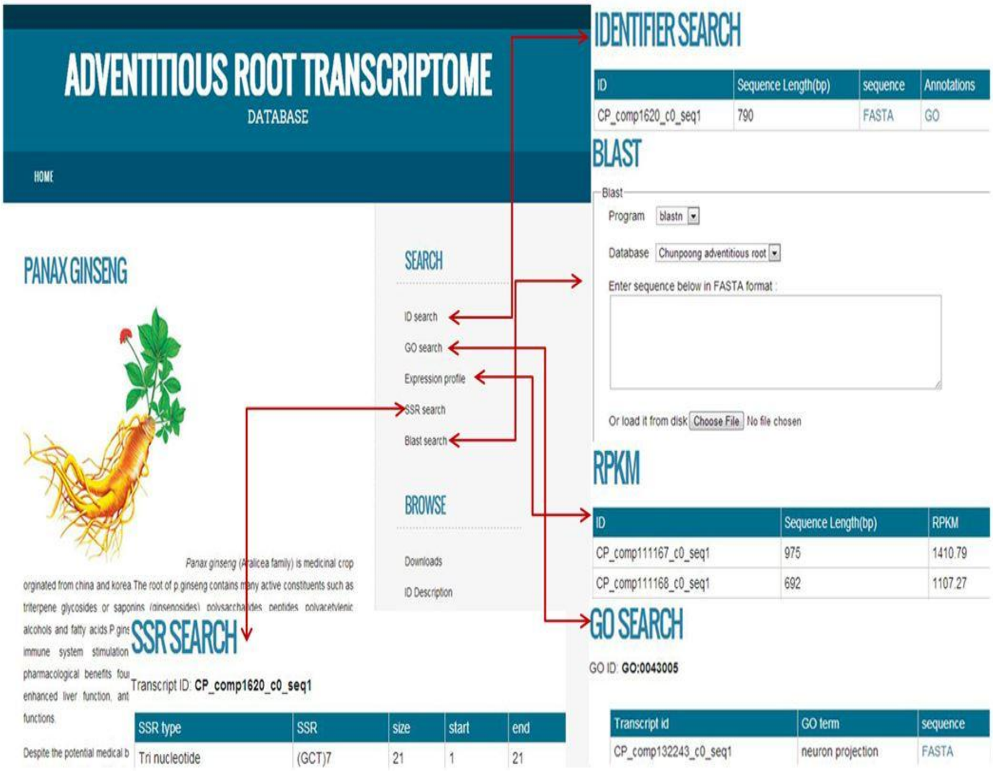
using Blast2GO with transcripts shown in supporting information. GO assignment of transcripts common to CP adventitious roots and normal roots was analyzed for 23,675 common transcripts from the CP dataset.

Table 1-2. Summary of SSRs identified in CP and CS transcripts

SSR - Mining	CP	CS
Total number of sequences examined	35,527	27,716
Total number of identified SSRs	10,213	7,928
Number of SSR-containing sequences	8,347 (23.49 %)	6,479 (23.37 %)
No. of sequences containing more than 1 SSR	1,523	1,179
Number of SSRs present in compound formation	464	376
Distribution of different repeat types		
Mono-nucleotide	3,163 (30.97 %)	2,386 (30.09 %)
Di-nucleotide	3,854 (37.73 %)	3,053 (38.50 %)
Tri-nucleotide	2,627 (25.72 %)	2,064 (26.03 %)
Tetra-nucleotide	396 (3.8 %)	294 (3.7 %)
Penta-nucleotide	86 (0.84 %)	61 (0.76 %)
Hexa-nucleotide	87 (0.85 %)	70 (0.88 %)

Adventitious root transcriptome database

We developed an open-access web database called the adventitious root transcriptome (ART) database to provide a platform for exploring



Panax ginseng (Aralicea family) is medicinal crop orginated from china and korea The root of p.ginseng contains many active constituents such as triterpene glycosides or saponins (ginsenosides) polysaccharides peptides polyacetylenic alcohols and fatty acids P.gins immune system stimulation pharmacological benefits four enhanced liver function, ant functions.

Despite the potential medical b

ID	Sequence Length(bp)	sequence	Annotations
CP_comp1620_c0_seq1	790	FASTA	GO

Blast

Program: blastn

Database: Chunpoong adventitious root

Enter sequence below in FASTA format :

Or load it from disk Choose File No file chosen

ID	Sequence Length(bp)	RPKM
CP_comp111167_c0_seq1	975	1410.79
CP_comp111168_c0_seq1	692	1107.27

GO ID: **GO:0043005**

Transcript id	GO term	sequence
CP_comp132243_c0_seq1	neuron projection	FASTA

Transcript ID: **CP_comp1620_c0_seq1**

SSR type	SSR	size	start	end
Tri nucleotide	(GCT)7	21	1	21

Figure 1-10. Adventitious root transcriptome (ART) database developed in this study. The database is publicly available at http://im-crop.snu.ac.kr/transdb/index.php and provides information of CP and CS transcriptome mentioned in this study.

adventitious root transcriptome data of *P. ginseng* (Fig. 1-10). This database is publicly available at http://im-crop.snu.ac.kr/transdb/index.php. The webpages provide information about CP and CS ID descriptors [45]. Users can query the database by transcript ID and other functional annotation ID.

Individual transcript sequences and their annotations can be accessed through ID search. Expression value (RPKM) can be queried by either transcript ID or ranges between minimum and maximum values, which returns the transcript accessions and their sequences related to the user-defined RPKM value ranges. Furthermore, we included markers that were identified in CP and CS non-redundant assembled transcripts. SSR marker information can be queried by transcript ID, which returns output such as type of SSR identified (mono- to hexa-), and SSR sequence, size, and start and end positions. The database uses the NCBI BLAST algorithm (version:2.2.15) for sequence-based searches. From BLAST searches, users can match nucleotide or protein sequence(s) against the CP or CS adventitious root transcriptome data reported in this study at user-defined parameters. We expect that this database will expedite functional genomics in *P. ginseng* and be helpful for gene identification and marker development. Furthermore, we plan to update the database frequently with transcriptome data from other ginseng cultivars.

Discussion

Assembly of Illumina transcriptome sequences

Transcriptome profiling using NGS technology, so-called RNA-Seq, is one of the most efficient tools for gene discovery and various functional studies. Illumina transcriptome sequencing and assembly have been successfully used for several non-model organisms [46], [47], [48], [49], [50], but transcriptome assembly has many challenges, including mis-assembled or chimeric contigs (i.e. assembled contigs containing reads from different transcripts [51]). Due to differences in time points, tissues and other biotic and abiotic factors, the assembled sequences of a species may not necessarily match well with reference sequences from the same or closely related species [52], [53], [54], [55], [56]. Here, we describe a method to choose the best assembly result for both biologically and computationally meaningful results. We used metrics such as reads used in assembly, average length [57] and number of annotated proteome hits (TAIR), as indicators of assembly quality.

Our results show that the quality of a *de novo* transcriptome assembly is not highly dependent on the user-defined single *k*-mer length or multiple *k*-mer length [58], because we found the best assembly in Oases at *k*-mer 69 (user-defined: substantially higher length) and in Trinity at *k*-mer 25 (fixed: lower length) based on the assembly indicators. Due to algorithmic differences between these two assemblers, they produced almost the same quantity of contigs with different proportions of accuracy. Thus, validation of *de novo* transcriptome assembly is highly challenging and there is no standard method or criteria to identify mis- or chimeric assembly. Fortunately, we had 108 published full-length *P. ginseng* gene sequences, as well as our draft genome sequence, available for precise validation. We found that the accuracy was greater in the assembled transcripts with Trinity compared to with Oases. Our BLASTX annotation against the NCBI nr protein database yielded more than 90% hits in the CP and CS assembly sets. This is similar to a previous ginseng EST study in which 90% of the ESTs had hits with the nr database [59]. This high percentage is probably due to the removal of mis-assembled sequences and the high frequency of long sequences (approximately 1.9 kb average length) in our assembled transcripts. Our work shows that it is possible to obtain reliable transcriptome sequences in non-model species by performing re-sequencing and read-depth analysis.

Comparative analysis of transcriptomes in adventitious roots

Increased pharmacological efficiency is a main goal for genomic studies of ginseng [10]. Under field conditions, a normal ginseng root is affected by biotic and abiotic factors and becomes vulnerable to many diseases. To analyze differences between the cultivars, the effects caused by environmental factors need to be controlled as much as possible because cultivar-specific characteristics could be masked by environmental variation. The adventitious root culture system described herein provided useful material for comparative analysis of the transcriptomes of two cultivars because the adventitious roots were cultured under more controlled environmental conditions than can be obtained in the field. Although tissue culture represents a stress condition for plants due to the high concentration of plant growth regulators like auxin or the lack of proper nutrients for growth, the observed differences between cultivars mostly represent the unique characteristics of each cultivar in a tightly controlled environment. From our experience, adventitious roots are easy to handle and their transcriptomes are highly reproducible.

As ginseng research requires highly reliable reference sequences for functional genomics, we have created a reference sequence from CP, a highly desirable cultivar because of its superior quality [60]. Various root transcriptome studies have been reported using NGS technologies [61], [62], [63] but not for the adventitious root transcriptome. Our successful comparative analysis of the transcriptomes of two cultivars using adventitious roots promotes systematic approaches for functional genomics and metabolomics in medicinal plants.

To investigate the expression level of the assembled transcripts, we determined the RPKM value for each transcript. According to GO categorization of the resulting 100 most abundant transcripts, stress response-related transcripts were the most highly expressed category in the adventitious roots (Fig. 1-3 and 1-4). Moreover, many transcripts showing more than 2-fold differences between CP and CS cultivars also belonged to the stress-response biological process category (Fig. 1-6). Consistent with our results, stress proteins have also been reported to be highly expressed in calli of other plant species [64], [65], [66]. Furthermore, proteomic analysis of ginseng hairy roots revealed that stress response-related proteins are the mostly highly expressed category [67]. In normal ginseng plants, stress-responsive proteins are induced to high levels upon exposure to abiotic and biotic stresses [68], [69], [70], [71], [72]. Therefore, our results strongly suggest that *in vitro* culture conditions represent a stress to adventitious roots of ginseng plants, with stress response-related transcripts induced to protect cells from harmful conditions.

As would be expected, the majority of transcripts showed similar expression patterns between these two Korean cultivars that were bred by selection from Korean landraces [73], [74]. On the other hand, about 5% of transcripts showed differences in expression of more than 2-fold between the two cultivars. When we investigated the top 100 differentially expressed transcripts, certain transcripts were found to be unique to each cultivar, for example retrotransposon-related transcripts were among the most abundant transcripts expressed only in CS (Fig. 1-6). Consistent with this, transposable elements (TEs) have also been reported to be activated in tissue culture conditions [75], [76], [77], [78], [79]. Tissue culture induces an array of mutations such as somaclonal variation [80] and also disturbs cellular epigenetic controls [75], [76]. Considering that the activity of TEs is suppressed by DNA hypermethylation [81], [82], [83], our data imply that DNA methylation status was decreased sufficiently to activate retrotransposons in CS adventitious roots. Even though the same tissue culture conditions were utilized for CP adventitious roots, no TE-related transcripts were found in the top 100 CP-biased transcripts.

Auxin-response transcripts were identified among the top 100 CP-biased transcripts, but not in the top 100 CS-biased transcripts. Conversely, cell proliferation was a more abundant category among top CS-biased transcripts, compared to the CP dataset (Fig. 1-6). The SH media used in this study included the auxin analog IBA. Therefore, the increased expression of transcripts related to auxin response and cell proliferation might be a consequence of exposure to IBA. Nevertheless, the difference in gene expression between the two cultivars implies that each cultivar responds to the tissue culture conditions in a unique manner. In fact, during the induction and cultivation of adventitious roots, CP embryo cells were not easily transformed into adventitious roots, instead generating callus-like clumps, whereas CS embryo cells quickly generated adventitious roots, followed by rapid proliferation (Fig. 1-1). Furthermore, both cultivars show unique morphological and physiological traits in field growth conditions. In particular, the rate of seed maturation and germination is lower in CP than in CS (Korea seed & variety service, http://www.seed.go.kr). Presumably, the differences in gene expression and growth responses during adventitious root culture result from genetic background differences of the cultivars.

Genes related to ginsenoside biosynthesis

Ten transcripts encoding enzymes involved in ginsenoside biosynthesis were also identified through similarity searches with reported genes. Most of the ginsenoside biosynthesis transcripts were highly expressed in both cultivars, including transcripts for DDS or β-AS, which catalyze the rate-limiting step for ginsenoside biosynthesis [84], [85] (Fig. 1-7). This implies that the content and composition of ginsenosides may not be different between the cultivars under *in vitro* culture conditions. In addition, 21 transcripts related to UGT proteins were identified in the datasets of both cultivars (Fig. 1-8). Among those, three transcripts were closely related to MtUGT73K1, MtUGT71G1, and SvUGT74M1, which function in triterpene saponin biosynthesis. Therefore, these transcripts most likely encode UGTs involved in the last step of ginsenoside biosynthesis in *P. ginseng.* Simultaneous analysis of metabolite profiles and the transcriptome may promote in-depth understanding of the ginsenoside biosynthesis pathway.

Comparative analysis of transcriptomes between roots and adventitious roots

Through comparative analysis with the transcriptome of normal ginseng roots, more than 6,000 transcripts were identified to be unique to adventitious roots or normal roots, whose functional differences were characterized using GO analysis (Fig. 1-9). Although almost the same numbers of unique transcripts were analyzed for each tissue, transcripts unique to adventitious roots were more abundant for each individual GO term compared to those of normal roots. This indicates that a broader range of transcripts might be actively expressed in adventitious roots than in normal roots. In fact, more of the abundant transcripts in the adventitious root dataset were involved in transcription, cell proliferation, reproductive developmental processes and multicellular organismal development.

Conclusion

In this study, we have generated a gene catalog for ginseng adventitious roots via *de novo* transcriptome assembly, which served as a useful resource for gene discovery in the ginsenoside pathway and for SSR marker development. In addition, we established an evaluation process to enhance assembly quality. To the best of our knowledge, this is the first report precisely categorizing the adventitious root transcriptome of *P. ginseng*. The approach we used to obtain the final transcriptome can be adopted for transcriptome assembly of other non-model species. Our work also reveals that adventitious roots are advantageous for transcriptome profiling analysis for genes related to secondary metabolites. If metabolite profiling is conducted along with transcriptome analysis, we may obtain more knowledge about complex metabolic pathways. In this work, we also developed an open web database for access and retrieval of our analyzed data. We anticipate that this study will take ginseng research to the next level, facilitating identification of additional ginsenoside genes and functional markers, as well as promoting understanding and engineering of complex metabolic pathways.

REFERENCES

1. Wen J, Zimmer EA (1996) Phylogeny and biogeography of *Panax* L. (the ginseng genus, Araliaceae): Inferences from ITS sequences of nuclear ribosomal DNA. Mol Phylogenet Evol 6: 167-177.
2. Hu SY (1976) The genus *Panax* (ginseng) in Chinese medicine. Econ. Bot 30: 11-28.
3. Lee FC (1992) Facts about ginseng: The elixir of life. Elizabeth, NJ: Hollym International Corporation press.
4. Liu J, Wang S, Liu H, Yang L, Nan G (1995) Stimulatory effect of saponin from *Panax ginseng* on immune function of lymphocytes in the elderly. Mech Ageing Dev 83: 43-53.
5. Shin HR, Kim JY, Yun TK, Morgan G, Vainio H (2000) The cancer preventive potential of *Panax ginseng*: a review of human and experimental evidence. Cancer Causes Control 11: 565-576.
6. Kim SH, Park KS (2003) Effects of *Panax ginseng* extract on lipid metabolism in humans. Pharmacol Res 48: 511-513.
7. Duke J (2000) The Green Pharmacy Herbal Handbook: Your Comprehensive Reference to the Best Herbs for Healing. Emmaus. PA: Rodale press. p115.
8. Blumenthal M (2003) The ABC Clinical Guide to Herbs. New York: Theime press. 211 p.
9. Weiss R (1988) Herbal Medicine. Gothenburg, Sweden: Beaconsfield Publishers LTD, 176 p.
10. Chen SL, Sun YZ, Xu J, Luo HM, Sun C, et al. (2010) Strategies of the study on herb genome program. Acta Pharm Sin 45: 807-812.

11. Kwon WS, Chung CM, Kim YT, Lee MG, Choi KT (1998) Breeding process and characteristics of KG101, a superior line of *Panax ginseng* C.A. Meyer. Korean J Ginseng Sci. 22: 11–17.

12. Juan W, Shuli M, Wenyuan G, Liming Z, Luqi H (2013) Cluster analysis of ginseng tissue cultures, dynamic change of growth, total saponins, specific oxygen uptake rate in bioreactor and immuno-regulative effect of ginseng adventitious root. Industrial Crops and Products 41: 57-63.

13. . Jung JD, Park HW, Hahn Y, Hur CG, In DS, et al. (2003) Discovery of genes for ginsenoside biosynthesis by analysis of ginseng expressed sequence tags. Plant Cell Rep 22: 224–230.

14. Sathiyamoorthy S, In JG, Gayathri S, Kim YJ, Yang DC (2010) Generation and gene ontology based analysis of expressed sequence tags (EST) from a *Panax ginseng* C. A. Meyer roots. Mol Biol Rep 37**:** 3465-3472.

15. Choi HI, Kim NH, Kim JH, Choi BS, Ahn IO, et al. (2011) Development of Reproducible EST-derived SSR Markers and Assessment of Genetic Diversity in *Panax ginseng* Cultivars and Related Species. J Ginseng Res 35: 399-412.

16. Bang KH, Lee JW, Kim YC, Jo IH, Seo AY, et al (2011) Development of an ISSR-Derived SCAR Marker in Korean Ginseng Cultivars (*Panax ginseng* C. A. Meyer). J Ginseng Res 35: 52-59.

17. Hong CP, Lee SJ, Park JY, Plaha P, Park YS, et al. (2004) Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences. Mol Genet Genomics 276: 709–716.

18. Choi HI (2013) Genome structure and evolution of *Panax ginseng* C.A. Meyer. Ph.D. thesis, Seoul National University. Republic of Korea.

19. Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu. Rev. Genomics Hum. Genet 10: 135-51.

20. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.

21. Luo H, Sun C, Sun Y, Wu Q, Li Y, et al. (2011) Analysis of the transcriptome of *Panax notoginseng* root uncovers putative triterpene saponin-biosynthetic genes and genetic markers. BMC Genomics 12(suppl 5): S5.

22. Sun C, Li Y, Wu Q, Luo H, Sun Y, et al. (2010) *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11**:** 262.

23. Li C, Zhu Y, Guo X, Sun C, Luo H, et al. (2013) Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng* C. A. Meyer. BMC Genomics 14: 245.

24. Patel RK, Jain M (2012) NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. PLoS ONE **7**(2): e30619.

25. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) *Oases*: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086-1092.

26. Grabherr MG, Haas BJ, Yassour M (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 15: 644-52.

27. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, et al. (2009) De novo transcriptome assembly with ABySS. Bioinformatics 25: 2872-2877.

28. Franchini P, Van der Merwe M, Roodt-Wilding R (2011) Transcriptome characterization of the South African abalone Haliotis midae using sequencing-by-synthesis. BMC Research Notes 4: 59.

29. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

30. Velculescu VE, Kinzler KW (2007) Gene expression analysis goes digital. Nature Biotechnology 25: 878 - 880.

31. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5: 621-628.

32. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al.(2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-1349.

33. Haralampidis K, Trojanowska M, Osbourn AE (2002) Biosynthesis of triterpenoid saponins in plants. Adv Biochem Eng Biotechnol 75: 31-49.

34. Kushiro T, Shibuya M, Ebizuka Y (1998) Beta-amyrin synthase--cloning of oxidosqualene cyclase that catalyzes the formation of the most popular triterpene among higher plants. Eur J Biochem 256: 238-244.

35. Lee MH, Jeong JH, Seo JW (2004) Enhanced triterpene and phytosterol biosynthesis in *Panax ginseng* overexpressing squalene synthase gene. Plant Cell Physiol 8: 976-984.

36. Han JY, Kwon YS, Yang DC, Jung YR, Choi YE (2006) Expression and RNA interference-induced silencing of the dammarenediol synthase gene in *Panax ginseng*. Plant Cell Physiol 47: 1653-1662.

37. Tansakul P, Shibuya M, Kushiro T, Ebizuka Y (2006) Dammarenediol-II synthase, the first dedicated enzyme for ginsenoside biosynthesis in *Panax ginseng*. FEBS Lett 580: 5143-5149.

38. Han JY, In JG, Kwon YS, Choi YE (2010) Regulation of ginsenoside and phytosterol biosynthesis by RNA interferences of squalene epoxidase gene in *Panax ginseng*. Phytochemistry 71: 36-46.

39. Han JY, Kim HJ, Kwon YS, Choi YE (2011), The Cyt P450 enzyme CYP716A47 catalyzes the formation of protopanaxadiol from dammarenediol-II during ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Physiol 52: 2062-2073.

40. Kim TD, Han JY, Huh GH, Choi YE (2011) Expression and functional characterization of three squalene synthase genes associated with saponin biosynthesis in *Panax ginseng*. Plant Cell Physiol 1: 125-137.

41. Han JY, Hwang HS, Choi SW, Kim HJ, Choi YE (2012) Cytochrome P450 CYP716A53v2 catalyzes the formation of protopanaxatriol from protopanaxadiol during ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Physiol 53:1535-1545.

42. Achnine L, Huhman DV, Farag MA, Sumner LW, Blount JW, et al. (2005) Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume Medicago truncatula. Plant J 41: 875-887.

43. Meesapyodsuk D, Balsevich J, Reed DW, Covello PS (2007) Saponin biosynthesis in Saponaria vaccaria. cDNAs encoding beta-amyrin

synthase and a triterpene carboxylic acid glucosyltransferase. Plant Physiol 143: 959-969.

44. Chen S, Luo H, Li Y, Sun Y, Wu Q, et al. (2011) 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in Panax ginseng. Plant Cell Rep 30: 1593-1601.

45. Garg R, Patel RK, Jhanwar S, Priya P, Bhattacharjee A, et al. (2011) Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. Plant Physiol 156:1661-1678.

46. Garg R, Patel RK, Tyagi AK, Jain M (2011) *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res 18: 53-63.

47. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, et al. (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci USA 106: 12353-12358.

48. Li D, Deng Z, Qin B, Liu X, Men Z (2012) *De novo* assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis Muell. Arg*.). BMC Genomics 13: 192.

49. Xu DL, Long H, Liang JJ, Zhang J, Chen X, et al. (2012) *De novo* assembly and characterization of the root transcriptome of *Aegilops variabilis* during an interaction with the cereal cyst nematode. BMC Genomics 13: 133.

50. Zhang J, Liang S, Duan J, Wang J, Chen S, et al. (2012) *De novo* assembly and Characterisation of the Transcriptome during seed development, and generation of genic-SSR markers in Peanut (*Arachis hypogaea* L.). BMC Genomics 13: 90.

51. Vijay N, Poelstra JW, Künstner A, Wolf JBW (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol Eco 22: 620-634.

52. Thanh NT, Murthy HN, Yu KW, Hahn EJ, Paek KY (2005) Methyl jasmonate elicitation enhanced synthesis of ginsenoside by cell suspension cultures of *Panax ginseng* in 5-l balloon type bubble bioreactors. Appl Microbiol Biotechnol 67: 197-201.

53. De Boer TE, Birlutiu A, Bochdanovits Z, Timmermans MJ, Dijkstra TM, et al. (2011) Transcriptional plasticity of a soil arthropod across different ecological conditions. Molecular Ecology 20: 1144-1154.

54. Muller L, Hutter S, Stamboliyska R, Saminadin-Peter S, Stephan W, et al. (2011) Population transcriptomics of *Drosophila melanogaster* females. BMC Genomics 12: 81.

55. Van LH, Kliebenstein DJ, West MA, Kim K, Van PR, et al. (2007) Natural Variation among Arabidopsis thaliana Accessions for Transcriptome Response to Exogenous Salicylic Acid. The Plant Cell Online 19: 2099-2110.

56. Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution. Molecular Ecology 15: 1197-1211.

57. Mark Y, Daniel E (2012) A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13: 329-342.

58. Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. Genome Re 20: 1432-1440.

59. Hong-Il C, Nam-Hoon K, Junki L, Beom Soon C, Kyung Do K, et al. (2012) Evolutionary relationship of *Panax ginseng* and *P. quinquefolius* inferred from sequencing and comparative analysis of expressed sequence tags. Genetic Resources and Crop Evolution 10: 1007.

60. Hongtao W, Hua S, Woo-Saeng, K, Haizhu J, Deok-Chun Y (2009) Molecular identification of the Korean ginseng cultivar "Chunpoong" using the mitochondrial nad7 intron 4 region. Mitochondrial DNA 20: 41-45.

61. Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, et al. (2010) *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). BMC Genomics 11:726.

62. Daniel RR, Felipe HB, Natasha TH, Louise EJ, Daniel PS (2010) Tomato root transcriptome response to a nitrogen-enriched soil patch. BMC Plant Biology 10:75.

63. Sun C, Li Y, Wu Q, Luo H, Sun Y, et al. (2010) *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11:262.

64. Zhao F, Chen L, Perl A, Chen S, Ma H (2011) Proteomic changes in grape embryogenic callus in response to *Agrobacterium tumefaciens*-mediated transformation. Plant Sci 181: 485-495.

65. Ge XX, Chai LJ, Liu Z, Wu XM, Deng XX, et al. (2012) Transcriptional profiling of genes involved in embryogenic, non-embryogenic calluses and somatic embryogenesis of Valencia sweet orange by SSH-based microarray. Planta 236: 1107-1124.

66. Arican E, Albayrak G, Gozukirmizi N (2008) Calli cultures from *Abies equi-trojani* (Aschers et Sinten) and changes in antioxidant defense system enzymes. J Environ Biol 29: 841-844.

67. . Kim SI, Kim JY, Kim EA, Kwon KH, Kim KW, et al. (2003) Proteome analysis of hairy root from *Panax ginseng* C. A. Meyer using peptide fingerprinting, internal sequencing and expressed sequence tag data. Proteomics 3: 2379–2392.

68. Lee OR, Sathiyaraj G, Kim YJ, In JG, Kwon WS, et al. (2011) Defense Genes Induced by Pathogens and Abiotic Stresses in *Panax ginseng* C. A. Meyer. J Ginseng Res 1: 1-11.

69. Kim YJ, Lee JH, Lee OR, Sun Shim JS, Jung SK, et al. (2010) Isolation and Characterization of a Type II Peroxiredoxin Gene from *Panax ginseng* C. A. Meyer. J Ginseng Res 4: 296-303.

70. Sun H, Kim MK, Pulla RK, Kim YJ, Yang DC (2010) Isolation and expression analysis of a novel major latex-like protein (MLP151) gene from *Panax ginseng*. Mol Biol Rep 37: 2215-2222.

71. Sathiyaraj G, Lee OR, Parvin S, Khorolragchaa A, Kim YJ, et al. (2011) Transcript profiling of antioxidant genes during biotic and abiotic stresses in *Panax ginseng* C. A. Meyer. Mol Biol Rep 38: 2761–2769.

72. Sathiyaraj G, Srinivasan S, Subramanium S, Kim YJ, Kim YJ, et al. (2010) Polygalacturonase inhibiting protein: isolation, developmental regulation and pathogen related expression in *Panax ginseng* C.A. Meyer. Mol Biol Rep 37**:** 3445–3454.

73. Choi HI, Kim NH, Kim JH, Choi BS, Ahn IO, et al. (2011) Development of reproducible EST-derived SSR markers and assessment of genetic diversity in *Panax ginseng* cultivars and related species. J Ginseng Res 35: 399-412.

74. Kim NH, Choi HI, Ahn IO, Yang TJ (2012) EST-SSR marker sets for practical authentication of all nine registered ginseng cultivars in Korea. J Ginseng Res 36: 298-307.

75. Kikuchi K, Terauchit K, Wada M, Hirano HY (2003) The plant MITE mPing is mobilized in anther culture. Nature 421: 167–170.

76. Hirochika H (1993) Activation of tobacco retrotransposons during tissue culture. EMBO J 12: 2521–2528.

77. Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci USA 93: 7783–7788.

78. Liu ZL, Han FP, Tan M, Shan XH, Dong YZ, et al. (2004) Activation of a rice endogenous retrotransposon *Tos17* in tissue culture is accompanied by cytosine demethylation and causes heritable alteration in methylation pattern of flanking genomic regions. Theor Appl Genet 109: 200-209.

79. Wu R, Guo WL, Wang XR, Wang XL, Zhuang TT, et al. (2009) Unintended consequence of plant transformation: biolistic transformation caused transpositional activation of an endogenous retrotransposon *Tos17* in rice ssp. japonica cv. Matsumae. Plant Cell Rep 28: 1043-51.

80. Evans DA (1989) Somaclonal variation-genetic basis and breeding applications. Trend Genet 5: 46–50.

81. Grandbastien MA (1998) Activation of plant retrotransposons under stress conditions. Trend Plant Sci 3: 181–187.

82. Kubis SE, Castilho AM, Vershinin AV, Heslop-Harrison JS (2003) Retroelements, transposons and methylation status in the genome of oil

palm (*Elaeis guineensis*) and the relationship to somaclonal variation. Plant Mol Biol 52: 69–79.

83. Tanurdzic M, Vaughn MW, Jiang H, Lee TJ, Slotkin RK, et al. (2008) Epigenomic consequences of immortalized plant cell suspension culture. PLoS Biol 6: 2880–2895.

84. Kushiro T, Ohno Y, Shibuya M, Ebizuka Y (1997) *In vitro* conversion of 2,3-oxidosqualene into dammarenediol by *Panax ginseng* microsomes. Biol Pharm Bull 20: 292-294.

85. Hu W, Liu N, Tian Y, Zhang L (2013) Molecular Cloning, Expression, Purification, and Functional Characterization of Dammarenediol Synthase from *Panax ginseng*. Biomed Res Int 2013:285740.

CHAPTER 2

Computational prediction of long noncoding RNAs (lncRNAs) in *P. ginseng*

Abstract

Very few long noncoding RNAs (lncRNAs) have been found in plants, many of which have unknown biological roles. To better understand our ginseng genome (*P. ginseng*), we used Chunpoong (CP) adventitious root RNA-Seq data to identify lncRNAs of *P. gisneng*. we found 11,270 long noncoding RNAs which had multiexonic structures. A total of 433 lncRNAs showed significant similarity against publicly available lncRNA database of Arabidopsis, Maize, lncRNA database and Rfam. Our predicted lncRNAs were classified as small RNA precursors. 196, 549, and 105 lncRNAs were classified as miRNA, shRNA and siRNA respectively. Around 2% of repetitive masking was observed from repeat masking analysis. Our study provides a preliminary source for future studies of lncRNA content and function in ginseng.

keywords

lncRNA, CPC, miRNA, Chunpoong (CP)

Introduction

Noncoding RNAs(ncRNAs) are transcripts that are not translated to proteins but act as functional RNAs. Noncoding RNAs can include small RNAs, generally under 15-25 base pairs(bp) in length, and longer RNAs (>200 bp) referred as long noncoding RNAs (lncRNAs). lncRNAs are mainly transcribed by RNA polymerase II (Pol II) and are polyadenylated, spliced, and mostly localized in the nucleus[1]. LncRNAs can be classified as natural antisense transcripts (NATs), long intronic noncoding RNAs and long intergenic noncoding RNAs(lincRNAs) according to their characteristics. lncRNAs play roles in a wide range of biological processes including in developmental regulations and stress responses [2], plant reproductive developments [3] and response to pathogen invasion [4][5]. Even though lncRNAs are involved in various regulatory roles, the systematic identification of lncRNAs is limited to a few plant species [5-9]. A comparatively small amount of lncRNAs have been identified in plants. One of the first was discovered in *Medicago truncatula*. Further investigation revealed two roles for the RNA: as a peptide-encoding mRNA, and as a ncRNA molecule with a functional secondary structure [10][11]. In addition, in *Arabidopsis thaliana*, evidence suggests that lncRNAs transcribed from the flowering locus C (FLC) are necessary for vernalization. One study found an antisense transcript, COOLAIR, to FLC that blocks transcription of the sense transcript [12]. Another study reported an intronic lncRNA, COLDAIR, originating within FLC that recruits PRC2 to epigenetically silence the locus [13].

In Arapdopsis, 6480 intergenic transcripts can be classified as lincRNAs by using a tiling array-based strategy, among which 2708 lincRNAs was detected by RNA sequencing experiments [4]. Interestingly, a subset of lincRNA genes shows organ-specific expression, whereas others are responsive to biotic and/or abiotic stresses. More interestingly, Wu et al. identified a number of lncRNAs as Endogenous Target Mimics (eTM) for microRNAs (miRNAs) in both Arabidopsis and rice, in which the eTMs of several miRNAs, such as miR160,miR166, miR156, miR159 and miR172, can effectively inhibit the functions of their corresponding miRNAs, and the eTMs of miR160 and miR166 play a role in regulation of plant development [6]. Recently, Xin et al. applied computational analysis and experimental approach identifying 125 putative wheat stress responsive lncRNAs, which are not conserved among plant species [5]. Among them, two lncRNAs were identified as signal recognition particle (SRP) 7S RNA variants, and three were characterized as U3 snoRNAs. Furthermore, the wheat lncRNAs also showed tissue dependent expression patterns like the lncRNAs in Arabidopsis [5], suggesting that the highly tissue-specific expression pattern might be a general trait of lncRNAs in plant development. In addition to Arabidopsis and wheat, Zhang et al. have analyzed global patterns of allelic gene expression in developing maize endosperms from reciprocal crosses between inbreds B73 and Mo17, and found that 38 lncRNAs expressed in the endosperm are imprinted. Among them, 25 are maternally expressed transcripts, whereas 13 are paternally expressed transcripts, and transcribed in either sense or antisense orientation from intronic regions of normal protein-coding genes or from intergenic regions [7]. Subsequently, Boerner et al. identify the potential lncRNAs using the maize full length cDNA sequences. The results showed the noncoding transcription appears to be widespread in the maize genome, and these ncRNAs were predicted to originate from both genic and intergenic loci. Computational predictions indicated that they may function to regulate expression of other genes through multiple RNA mediated mechanisms [8]. The growing reports of lncRNA identifications in different species indicate that lncRNAs ubiquitously exist in the plant kingdom with conserved roles. More recently, a computational approach for comprehensive identification of lincRNAs from rice using 40 existing rice RNA-Seq data sets were developed. Genome wide screening identified 2063 lincRNAs in rice, and most of them have a reproductive process preferred expressing pattern. Further functional analyze showed a set of lincRNAs could induce reproductive deficiencies. These studies would provide new insight into the involvement of lncRNAs in the reproductive development of rice. All together suggest that there are a large number of lncRNAs exist in various plant species, which might play a role in regulating the plant development and stress response. Our ginseng adventitious root transcriptome analysis showed high level of expression in transcripts related to stress. Therefore, we also suspected that the lncRNA can play major role in ginseng plant. In this study, we used computational method to predict lncRNA from our Chunpoong (CP) adventitious root RNA-Seq data.

Materials and methods

Data sources

Adventitious roots from the cotyledons of CP were obtained and Paired-end (PE) reads with length of 101 bp were generated using the Illumina Hiseq2000 platform. The library construction and sequencing was performed by the National Instrumentation Center and Environmental Management (NICEM), Seoul National University, South Korea. The sequence data generated in this study have been deposited at NCBI in the Short Read Archive (SRA) with accession number SRA061905.

The Coding Potential Calculator (CPC)

Coding Potential Calculator(CPC) program was downloaded and used for prediction lncRNAs[14]. The output data was analyzed, and a list of the transcript IDs described as ‘‘noncoding’’ and ‘‘weakly noncoding’’was created.

Small RNA Database Creation

Small RNA sequences were downloaded as FASTA files from miRBase (http://www.mirbase.org/). Small RNAs were classified into 3 different databases like miRNA, shRNA and siRNA. BlastN search were performed with an E-value of 0.01.

Results and Discussion

Computational prediction

We first developed a pipeline by reviewing various study for transcripts from RNA-seq and poly(A)-site data sets. Because our focus was on lncRNAs, we chose not to consider information helpful for predicting protein-coding transcripts (such as sequence conservation, homology with known genes, codon usage, or coding potential), reasoning that by avoiding the consideration of this information we could use our accuracy for identifying previously annotated mRNAs to indicate accuracy for identifying lncRNAs. Using TopHat, an alignment program that maps RNA-seq reads to putative exon junctions as well as genomic sequence [15], we mapped morethan 80 million reads CP adventitious root RNA-seq data set. We used cufflinks program [16] to generate reference based assembly set. A total of 86,796 contigs were assembled. To identify lncRNAs, our whole contigs were filtered to remove those that overlapped the protein-coding genes using blastx search against swiss-prot protein database. We obtained 30,960 contigs after removal of protein coding contigs. Further, we removed the contigs that have sequence length of below 200 bp and contigs with single exon. The coding potential was evaluated, removing those with scores ≥ -1.0 when using coding potential calculation (CPC) program[14]. Finally, we have identified a total of 11,270 (Fig.2-1) lncRNAs in *P. ginseng*. Since the ginseng genome has not been anchored on chromosomes, we could not classify based on the genomic location.

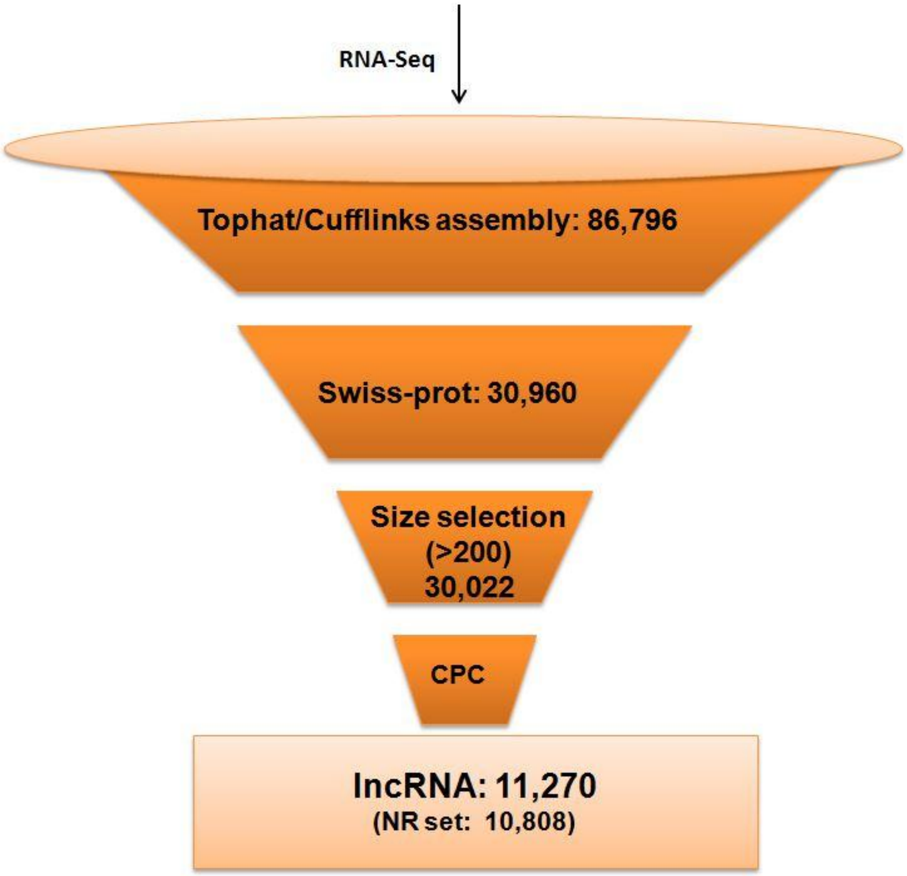


Figure 2-1. lncRNA prediction pipeline

Homology based functional search

Even though lncRNAs are less conserved among species, we tried to find any relationship with available sequences. We have collected the publicly available lncRNA sequences from Arabidopsis, maize, lncRNA database and Rfam database.

Table 2-1. Homology search against publicly available lncRNA sequences

Source	Available database sequence	No. of hits
Arabidopsis	1340	184
Maize	2492	150
lncRNA database	224	95
Rfam	225	48

We found a small number of hit sequences from the homology search (Table 2-1). we used those sequence for further functional classification. Among them, we selected top 10 sequences based on the FPKM value. As we expected, we found a contig "lncRNA_ginseng.54120.1" that play role in preventing stress response (Table 2-2).

Table 2-2. Functional summary of top 10 lncRNAs based on FPKM.

No	lncRNA ID	Top FPKM	lncRNA db Function
1	lncRNA_ginseng.42810.1	59.33	unknown
2	lncRNA_ginseng.46853.1	57.65	It been reported to interact with a number of chromatin binding protein/complexes
3	lncRNA_ginseng.42281.1	54.51	unknown
4	lncRNA_ginseng.35907.1	34.38	Interact and regulate chromobox 7 (CBX7), a component of the Polycomb Repressor Complex 1 (PRC1), and is also up-regulated in prostate cancer.
5	lncRNA_ginseng.58192.1	31.22	Associates with the RNA binding protein
6	lncRNA_ginseng.54120.1	29.9	Fundamental role in preventing **stress response** and apoptosis of the host cell
7	lncRNA_ginseng.64793.1	22.41	PRC2 chromatin modification complex
8	lncRNA_ginseng.14907.1	20.45	no function
9	lncRNA_ginseng.67169.1	20.18	Xist and RepA RNAs bind to and recruit PRC2, which catalyzes the deposition of the repressive chromatin mark H3K27me3 on the Xist
10	lncRNA_ginseng.31130.1	20.09	It may involve distinct regulatory mechanisms in different cell types .

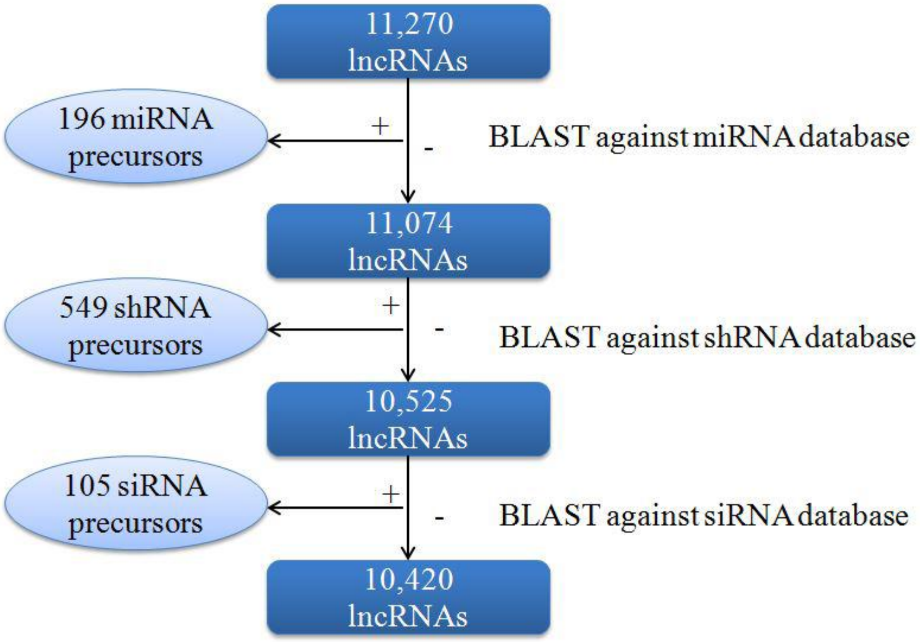


Figure 2-2. Classification of lncRNAs based on small RNA precursor potential

Classification of ginseng lncRNAs as small RNA precursors

In plants, small RNAs 20–25 nucleotides in length are an important class of noncoding RNA for the regulation of gene expression, and can originate from longer transcripts that are processed by endonucleases like Dicer. These small RNAs can influence gene expression at both the transcriptional and posttranscriptional level, and are produced via distinct pathways in plants [17]. One anticipated fate of the lncRNA candidates would be to serve as precursor molecules that are processed into small RNAs. The 11,270 lncRNA candidates were characterized for small RNA precursor potential (Fig 2-2) based upon homology with known small RNA sequences from miRBase. Three separate databases were created for these categories to align with candidate lncRNAs (Fig 2-2). The 11,270 lncRNA candidates identified with CPC were sequentially aligned to the small RNA databases and classified according to the results (Fig 2-2). 196 of the lncRNAs had homology with a miRNA sequence. In total, 549 ncRNAs were classified as shRNA precursors, and 105 as siRNA precursors. The remaining 10,420 ncRNAs were classified as lncRNAs that are likely to function as longer molecules. A total of 850 transcripts contained a small RNA sequences in ginseng even thought it does not have very closely related genome annotation. This may reflect that small RNAs are important and abundant regulatory molecules in plants. While it may also be indicative of biased datasets, with over-representation of these types of ncRNAs, only miRNAs are known to be predominantly dependent upon pol II transcription in plants. Thus, we anticipate that any bias in the current dataset would underrepresent small RNA precursors due to the exclusion of non-polyadenylated molecules.

Table 2-3. Repetitive element content of lncRNAs in ginseng.

	lncRNA
Sequences	11,270
Total length (bp)	7,810,304 (bp)
GC level	37.51 %
bases masked	174288 (2.23%)
	# No. of elements
SINE	0
LINE	0
LTR elements	85 (0.19%)
DNA transposons	47 (0.07%)
Unclassified	8 (0.02%)
Small RNA	2
Satellites	0
Simple repeats	3,086 (1.61%)
Low complexity	551 (0.34%)
	# bases masked
Total interspersed repeats	22,311 bp (0.29%)

Repetitive Element content of lncRNAs

We also performed repetitive element content analysis using (Table 2-3) Repeat Masker (www.repeatmasker.org). We found very low of 2 % repetitive content in ginseng lncRNAs set (Table 2-3). We identified as few as 88 retro elements of which 3 elements were SINEs and 85 elements were LTR elements. In addition, we identified a total of 47 DNA transposons. Generally, transposons are believed to be sources of ncRNAs that are important mediators of gene silencing [18], and the DNA transposon-related sequences detected in this study may reflect this type of gene silencing mechanism in the ginseng genome.

Conclusion

Although more lncRNAs will undoubtedly be found in ginseng in the future, this initial study will help for genome characterization and its evolution and detailed genome annotation. This study can now contribute to a starting point to predict lncRNA in ginseng which may lead to identify the growth related and epigenetic regulators in ginseng.

REFERENCES

1. Wierzbicki AT (2012) The role of long non-coding RNA in transcriptional gene silencing. Curr. Opin. Plant Biol. 15:517–522.

2. Kim ED, Sung S (2012) Long noncoding RNA: unveiling hidden layer of gene regulatory networks. Trends Plant Sci. 17:16–21.

3. Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 33:76–79.

4. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell 24:4333–4345.

5. Xin M, Wang Y, Yao Y, Song N, Hu Z, Qin D, Xie C, Peng H, Ni N, Sun Q (2011) Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. BMC Plant Biol. 11:61.

6. Wu HJ, Wang ZM, Wang M, Wang XJ (2013) Widespread long noncoding RNAs as endogenous target mimics for MicroRNAs in plants. Plant Physiol. 161:1875–1884.

7. Zhang M, Zhao H, Xie S, Chen J, Xu Y, Wang K, Guan H, Hu X, Jiao Y, Song W, Lai J (2011). Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm. Proc. Natl. Acad. Sci. U.S.A.108:20042–20047.

8. Boerner S, McGinnis KM (2012) Computational identification and functional predictions of long noncoding RNA in Zea mays. PLoS One 7: e43047.

9. Liu C, Muchhal US, Raghothama KG (1997) Differential expression of TPS11, a phosphate starvation-induced gene in tomato. Plant Mol. Biol. 33:867–874.

10. Campalans A, Kondorosi A, Crespi M (2004) Enod40, a short open reading frame-containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula. Plant Cell 16: 1047–1059.

11. Bardou F, Merchan F, Ariel F, Crespi M (2011) Dual RNAs in plants. Biochimie 93: 1950–1954.

12. Swiezewski S, Liu F, Magusin A, Dean C (2009) Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature 462: 799–802.

13. Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76–79.

14. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, et al. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35: W345–349.

15. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.

16. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van 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. Nat Biotechnol 28: 511–515.

17. Chen X (2009) Small RNAs and their roles in plant development. Annu Rev Cell Dev Biol 25: 21–44.

18. Zaratiegui M, Irvine DV, Martienssen RA (2007) Noncoding RNAs and gene silencing. Cell 128: 763–776.

ABSTRACT IN KOREAN

Panax ginseng(인삼)은 전통적으로 약리효과가 뛰어나 질병 예방 및 치료 목적으로 다양하게 사용되어 오고 있다. 그러나 이에 대한 분자유전학적 연구는 매우 미비한 실정이다. 본 논문은 서로 다른 두 지역의 인삼 부정근의 전사체를 *de novo* assembly 방법을 이용하여 분석하였다. 분석을 통하여 지놈 구조가 매우 복잡한 것으로 예상되는 인삼의 전사체 90% 이상이 annotation되었으며, 특히 ginsenoside 생합성 과정에 관여하는 유전자를 동정하였다. 이 후 일반 인삼뿌리와 부정근의 전사체를 기능별로 분석해 본 결과 두 샘플에서 발현되는 유전자의 약 17% 정도가 서로 다른 기능을 갖는 것으로 밝혀졌으며, cDNA SSR분석 결과는 지역 간 분자 유전학적 마커를 개발하는데 이용될 수 있을 것으로 기대된다. 본 논문의 후반부에는 인삼 특이 long noncoding RNA 분석을 실시하였는데 이와 같은 연구 결과는 이후 인삼의 약리작용에 관여하는 유전자의 발현 및 기능 조절 연구에 기초 데이터로 활용될 수 있을 것으로 판단된다.