

Transcriptome analysis using de novo RNA-seq to compare ginseng roots cultivated in different environments

Byung Wook Yang¹ · Young Tae Hahm¹

Received: 12 March 2017 / Accepted: 18 September 2017 / Published online: 22 September 2017
© Springer Science+Business Media B.V. 2017

Abstract Ginseng (*Panax ginseng* C.A. Meyer), one of the most widely used medicinal plants in traditional oriental medicine, is used for the treatment of various diseases. Ginseng is typically classified according to its cultivation environment, such as field-cultivated ginseng (FCG) and mountain-cultivated ginseng (MCG). However, little is known about differences in gene expression in ginseng roots between FCG and MCG. In order to investigate the whole transcriptome landscape of ginseng, we employed high-throughput sequencing technologies using the Illumina HiSeqTM2500 system, and generated a large amount of transcriptome data from ginseng roots. Approximately 77 and 87 million high-quality reads were produced via FCG and MCG root transcriptome analysis, respectively, and we obtained 256,032 assembled unigenes with an average length of 1171 bp by de novo assembly methods. Functional annotation of the unigenes was performed using sequence similarity comparisons against the following databases: the non-redundant nucleotide database, the InterPro domains database, the Gene Ontology Consortium database, and the Kyoto Encyclopedia of Genes and Genomes pathway database. A total of 4207 unigenes were assigned to specific metabolic pathways, and all of the known enzymes involved in starch and sucrose metabolism pathways were also identified in the KEGG library. This study indicated that alpha-glucan phosphorylase 1, putative pectinesterase/pectinesterase

inhibitor 17, beta-amylase, and alpha-glucan phosphorylase isozyme H might be important factors involved in starch and sucrose metabolism between FCG and MCG in different environments.

Keywords *Panax ginseng* C.A. Meyer · De novo RNA sequencing · Transcriptome · Gene expression

Abbreviations

<i>P. ginseng</i>	<i>Panax ginseng</i> C.A. Meyer
FCG	Field-cultivated ginseng
MCG	Mountain-cultivated ginseng
NGS	Next-generation sequencing
EST	Expressed sequence tag
CDS	Coding sequence
GO	Gene Ontology
DEG	Differentially expressed genes
KEGG	Kyoto Encyclopedia of Genes and Genomes

Introduction

Panax ginseng C. A. Meyer is a widely used medicinal plant with multiple clinical and pharmacological effects related to cancer, diabetes, and cardiovascular diseases. It also promotes immune and central nervous system function, as well as relieving stress (Nam 2002; O'Hara et al. 1998; Yun et al. 2001). Ginseng is classified as either field-cultivated or mountain-cultivated. Field-cultivated ginseng (FCG) is grown in an open field with a sun shield to help control the amount of total sunlight. In addition, fertilizer and pesticides are provided during FCG cultivation. Generally, FCG is cultivated for 4–6 years and then harvested. After more than 6 years of cultivation, ginseng roots become lignified and cannot be used as a medicinal herb. In contrast,

Electronic supplementary material The online version of this article (doi:10.1007/s10725-017-0328-6) contains supplementary material, which is available to authorized users.

✉ Young Tae Hahm
ythahm@cau.ac.kr

¹ Department of Systems Biotechnology, Chung-Ang University, Anseong 17546, South Korea

mountain-cultivated ginseng (MCG) matures in a deep mountain environment within thick forests with less exposure to direct sunlight. MCG can be cultivated either by directly sowing ginseng seed along with a soil covering to produce long rhizomes or as seedlings. Planted MCG has additional fine roots to allow for the uptake of minerals and water in the forest. The growth rate of MCG is much slower than that of FCG; this may be attributable to the shortage of nutrients in the mountain soil. Therefore, MCG is generally harvested after at least 10 years of growth (Jung 2005; Choi et al. 2007). These differences in cultivation environments and growth periods may lead to differences in biochemical and pharmacological activities between FCG and MCG.

Genetic studies on ginseng are challenging due to its long generation time, the small number of seeds it sets, and the difficulty of maintaining ginseng in the field. Recent advances in massively parallel cDNA sequencing (RNA-Seq) have provided a cost-effective way to obtain large amounts of transcriptome data from many organisms and tissue types (Birol et al. 2009; Choi et al. 2005; Trapnell et al. 2010). Next-generation sequencing (NGS) techniques are the most widely used methods for de novo sequencing and analysis of transcriptomes in non-model organisms. Although transcriptome sequencing and the assembly of large and complex plant genomes, such as that of *P. ginseng*, remain difficult, NGS technologies make it affordable to sequence cDNA (RNA-Seq) and examine cellular transcriptomes along with performing high-throughput gene expression analysis (Wang et al. 2009). Since their inception, NGS technologies have experienced a rapid improvement in throughput, read length, and accuracy. Here, we used the

Illumina sequencing platform for large-scale transcriptome analysis. We assembled FCG and MCG transcriptomes from millions of short sequence reads generated by Illumina paired-end transcriptome sequencing. After annotation, we conducted gene expression profiling, as well as the differential identification of relevant genes.

Materials and methods

Plant materials

The experiment was conducted using *P. ginseng* C.A. Meyer planted in different environments, including field and mountain cultivation. The FCG roots were 6 years old (Fig. 1a) and the MCG were 10 years old (Fig. 1b). Actively growing ginseng plants were harvested from a field in Anseong (N37°19'48.36"E127°39'35.67") and a mountain in Geochang (N35°49'36.34"E127°50'16.81"), Korea. The main roots were collected separately from individual plants and were cut into small pieces, followed immediately by storage in liquid nitrogen until further processing.

RNA preparation

Total RNA was isolated from roots using the RNeasy Plus Mini kit (Qiagen, Valencia, CA, USA). Quality control was performed on the samples using RNA 6000 Nano LabChips with a Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA, USA), and the obtained concentrations were assessed using a NanoDrop ND-1000 spectrophotometer (NanoDrop

Fig. 1 *Panax ginseng* C.A. Meyer cultivated in different environments. **a** Field-cultivated ginseng (FCG), **b** Mountain-cultivated ginseng (MCG)



Technologies, Wilmington, DE, USA) before additional processing. The RNA samples were treated with TURBO DNase (Ambion, Austin, TX, USA) prior to cDNA synthesis at a concentration of 1.5 units/ μ g total RNA.

RNA sequencing

cDNA library preparation was performed at Theragen Etx Bio (TBI), Suwon, South Korea. Using oligo(dT), mRNA was separated from 2 μ g of total RNA. The library was progressed with paired-end 100 bp, and preparation of the library was performed using the TruSeq RNA Sample Prep Kit from Illumina. Isolated mRNA was fragmented and single-stranded cDNA was then synthesized using random hexamer priming. Using this cDNA as a template, the second strand was synthesized to generate double-stranded cDNA. After end repair to generate a blunt-end, A-tailing and adapter ligation were performed consecutively. Amplification was performed using polymerase chain reaction (PCR). The final product was evaluated using a BioAnalyzer 2100. After quantifying the constructed library using the KAPA library quantification kit, the library underwent cluster generation and sequence decoding using an Illumina HiSeqTM2500 system.

De novo assembly and clustering

Prior to the assembly process used to obtain the sequence of transcripts, sequence filtering was conducted. Sequence filtering is a preliminary process that removes reads or adaptor sequences with the incorrect sequence, which are created during the sequence interpretation process, and reduces errors in assembly caused by low-quality sequences. First, to remove low-quality sequences, reads with 10% or more bases designated as N from among the total sequence information and reads with 20% or more bases below Q20 were removed. Reads with an average quality of Q20 or lower were also removed. After removing low-quality sequences, bases below Q20 present on both ends of the sequences were all removed from the remaining sequences. This removal was performed to increase the quality of the assembled sequence, since the quality on both ends drops noticeably over time due to the characteristics of mRNA. (Martin and Wang 2011) The overall filtering process was conducted using a program developed internally. Sequence assembly was conducted by combining the sequence data from all samples into a single sample and then the Trinity program (Grabherr et al. 2011; Haas et al. 2013) was used. Trinity is a representative RNA Assembler that conducts RNA-seq de novo assembly. In Trinity, sequence assembly is conducted using the de Bruijn graph (DBG) algorithm. The Trinity sequence assembly process of is divided into three stages; Inchworm, Chrysalis, and Butterfly. In the first stage,

Inchworm, all sequences are connected with each other and are divided into subgroups for efficient use of memory. Here, each sequence is divided into pieces with lengths of 25 bp, and through the process of combining the pieces using a 24-bp overlap, a contig is formed. In the Chrysalis stage, the contig formed in Inchworm is used to conduct clustering for the read. In addition, a perfect de Bruijn graph is again prepared for each divided group. In the Butterfly stage, the graph is interpreted to predict the transcript sequence. In this study, sequence assembly was conducted using the default value of Trinity. The sequence of the transcriptome assembled using Trinity is based on a transcript which includes isoforms instead of genes and, thus, overlaps in sequence exist. Chimeric transcripts, created during the process of sequence assembly, are included (Yang and Smith 2013). A method for clustering the transcript sequences assembled in Trinity was used to remove the overlaps, as well as the chimeric sequences. Clustering of the assembled transcriptome was conducted using the TGICL program (Perte et al. 2003). By comparing the given sequences 1:1, TGICL calculates the similarity among sequences, and conducts sequence clustering based on this similarity. Then, using CAP3 (Huang and Madan 1999), the program reassembles the sequence in each group to create a sequence representative of the group. The standard of sequence similarity for clustering was set at 0.94. To gain a preliminary understanding of the function of the unigenes, the area of protein expression (coding sequences [CDS]) was predicted using TransDecoder (Haas et al. 2013). TransDecoder is a program in Trinity used to find the CDS by predicting possible CDS from among the given transcript sequences. TransDecoder works as follows: First, using the given sequences, it predicts all possible CDS. Then, for the CDS predicted, it uses the GeneID (Blanco et al. 2007) program to verify the possibility, and then selects sequences with log-likelihood scores greater than 0. From among the selected potential protein sequences, the program chooses the one with the highest score. In cases where there are multiple open reading frames (ORFs) that do not overlap, all CDS are selected.

Analysis of homogeneity and gene function

To forecast the functions of the assembled unigenes and the CDS predicted based on these, search methods using the homogeneity of the DNA and the protein sequences were used. Blast was used to search the homogeneity of the DNA sequence, while InterProScan was used to search the homogeneity of the protein sequence. To search for homogeneity based on DNA sequence, the NCBI Blast 2.2.28+ and NCBI non-redundant (nr) databases (2013-07-17) were used. Among the search methods available in Blast, Blastx was used to search all possible protein sequences for the unigene sequences from the nr database.

This search was conducted to predict function. During the search process, the significance of homogeneity was set as an E-value $< 1e-5$. Searches using InterProScan for the predicted protein sequences were conducted as the second method of forecasting the function of unigenes. In contrast to Blast, which searches similar sequences using the partial homogeneity of the sequences, InterProScan forecasts function using the hidden Markov model to search for similarities at the domain level, a unit of protein function. This study used InterProScan v5 to search six protein databases, ProDom, PfamA, Panther, SMART, SuperFamily, and Gene3d, using the standard E-value $< 1e-5$.

Functional annotation of *P. ginseng* unigenes

BLASTX (Altschul et al. 1997) searches of *P. ginseng* unigenes against the NCBI non-redundant database were conducted with a cut off E-value of $1e-5$. The results were imported into the BLAST2GO program for functional annotation and GO analysis (Götz et al. 2008). Briefly, the mapping and annotation steps were carried out to retrieve and assign GO terms to the unigenes, and annotation results were summarized using plant GOSlim mapping. Directed acyclic graphs for three categories, biological process (BP), molecular function (MF), and cellular component (CC), were drawn for visualization of the hierarchical structures of the GO terms and summary statistics.

Expression profiling

Analysis of the expression level of each set of samples was conducted using the RSEM program (Li and Dewey 2011). RSEM is a tool for measuring the expression level of genes. Without the information provided by a reference genome, the program is designed to calculate the expression level of the transcriptome. Using Bowtie, RSEM aligns the read on the transcriptome, and then measures the expression level of the transcriptome using a directed graph model. Using the expression level for each group, a search for differentially expressed genes (DEGs) was conducted. For DEG analysis, the DEGES/DESeq method of the TCC program (Sun et al. 2013) was used. This method is an analysis method based on the DESeq (Anders and Huber 2010) method which conducts DEG analysis using negative-binomial distribution. By repetitively conducting the normalization process between two samples initially while considering DEG, the method finds DEGs between the two samples more accurately (Kadota et al. 2012). Using this tool, we repeated the normalization process three times. The standard value for selecting DEG was set at $q < 0.05$.

Pathway assignment with KEGG

Pathway assignments were carried out according to KEGG mapping (Kanehisa and Goto 2000). Enzyme commission (EC) numbers were assigned to unique sequences that had BLASTX scores with cutoff values of $E < 1e-5$, as determined upon searching the protein databases. The sequences were mapped to the KEGG biochemical pathways according to the EC distribution in the pathway database.

Results and discussion

De novo assembly and validation of Illumina paired-end sequences

We generated a total of 84,503,190 and 95,039,120 raw reads from FCG and MCG, respectively. After trimming the low-quality reads with Phred quality scores of Q30 and removing primer/adaptor sequences, we obtained 77,167,789 (91.2%) and 86,896,217 (91.4%) high-quality reads, with an average read length of 100 bp, in FCG and MCG, respectively (Table 1). To obtain high-quality assemblies, we tested several algorithms for de novo assembly using different options. We also compared the assembly results by mapping all raw reads onto each assembly, in order to determine read usage. We obtained the best assembly results from Trinity, as these results showed the largest assembled transcriptome sizes, the highest numbers of mapped reads, longer average contig lengths, and the greatest numbers of BLASTX and InterPro hits. Using Trinity, we obtained 256,032 unigenes with an average length of 1,171 bp. The length of the assembled transcripts ranged from 200 to 2900 bp, with a length of unigenes in 299,860,972 bp (Table 2). We identified sets of non-redundant transcripts from the assembled datasets by selecting only the longest sequence from among the isoforms, which included alternatively spliced forms predicted by the Trinity assembler.

Table 1 Summary statistics of FCG and MCG sequencing by Illumina HiSeq™2500

Sequencing data	FCG	MCG
Number of total bases	8,460,319,000	9,503,912,000
Bases (> Q30)	7,722,370,924 (91.3%)	8,694,299,685 (91.5%)
Trimmed bases	7,709,231,588 (91.1%)	8,689,325,255 (91.4%)
Number of total reads	84,603,190	95,039,120
Reads (> Q30)	77,167,789 (91.2%)	86,896,217 (91.4%)
Trimmed reads	77,358,824 (91.4%)	87,212,990 (91.8%)

FCG field-cultivated ginseng, MCG mountain-cultivated ginseng

Table 2 Trinity assembler unigene assembly statistics

Length of unigene (bp)	229,860,972
Average length of unigene (bp)	1171
Number of unigenes	256,032
N50 of unigenes (bp)	1851

Gene cluster analysis and identification of novel transcripts from ginseng RNA-Seq sequences

To construct high quality ginseng transcriptome data, we generated a substantial number of pair-end reads using an Illumina HiSeqTM2500. Using Trinity, over 90% of the reads were successfully mapped and assembled. A novel bioinformatics protocol for processing large amounts of transcriptome data was built. As RNA sequences were obtained from two different samples, we defined a new concept, the unigene cluster (UC), which contains overlapped unigene sequences originating from multiple samples. Utilizing the current annotation, 256,032 unigene clusters (UCs) with a total length of 299 Mb were identified. For validation and annotation of the assembled transcripts, sequence similarity searches were conducted against the protein databases using BLASTX and InterproScan. Significant hits were identified based on an E-value threshold of $1e-5$. The results indicated that more than 57.7% of the FCG and MCG transcripts showed significant similarity to proteins in the database. Approximately 108,057 (45.6%) of the FCG and MCG transcripts had significant matches with InterPro proteins. We also compared the FCG and MCG transcripts against the NCBI nr protein database using BLASTX and found that a total of 147,397 (57.6%) transcripts had significant hits.

Functional annotation and classification

GO classification separates the genes involved in different biological processes, molecular functions, and cellular components. The biological processes include broad biological events, such as mitosis or purine metabolism, which are accomplished by the ordered assembly of molecular functions. The molecular functions describe the task performed by individual gene products, examples of which include transcription factors and DNA helicases. The cellular component classification type involves subcellular structures, locations, and macromolecular complexes, such as the nucleus, telomeres, and origin recognition complex (Alvord et al. 2007; Dennis Jr et al. 2003). To classify the predicted functions of the identified transcripts, GO terms were assigned to transcripts using Blast2GO, based on their similarity to the nr database. A total of 3310 (78.6%) transcripts were assigned to GO classes. Among these, assignments to the biological process class ranked highest (2216; 51%), followed by the

cellular component (819; 39%) and molecular function (275; 10%) classes (Fig. 2).

Gene expression profiling

In order to determine the FPKM, we mapped all of the FCG and MCG reads onto assembled transcripts. For the FCG transcripts, the FPKM ranged from 0.01 to 7551.11, with an average of 9.15, and the FPKM for MCG ranged from 0.01 to 5602.44, with an average of 7.36. This indicates that both the FCG and MCG transcripts showed a wide range of expression levels (see supplementary data). We compared the expression patterns of the transcripts in FCG and MCG based on the differences in FPKM value for each transcript in the datasets. The transcripts showed very similar expression patterns in both FCG and MCG, which was also confirmed by the high correlation co-efficiency value of more than 0.8 between datasets. To identify highly expressed transcripts and their putative functions, we selected the 100 most abundant transcripts based on their FPKM values in the FCG and MCG datasets, and categorized the biological processes, cellular components, and molecular functions in which those transcripts might be involved. In the biological process category, transcripts from FCG were only expressed in response to radiation, modification-dependent macromolecule catabolic processes, modification-dependent protein catabolic processes, macromolecular complex assembly, response to salt stress, and response to osmotic stress. In contrast, MCG transcripts were only expressed in ncRNA metabolic processes, protein localization, nitrogen compound biosynthetic processes, macromolecular complex subunit organization, and intracellular transport (Fig. 3a). In the cellular component category, transcripts from FCG were only expressed in external encapsulating structures, cell walls, apoplasts, large ribosomal subunit, and cytosolic large ribosomal subunits (Fig. 3b). In the molecular function category, transcripts from MCG were only expressed in the RNA binding term (Fig. 3c).

Pathway assignment by KEGG

Analysis of pathway assignment was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG). First, 4207 unique genes were compared against the KEGG database using BLASTX with an E-value cutoff of $< 1e-5$. A total of 4207 unigenes were assigned to specific metabolic pathways, and all of the known enzymes involved in the starch and sucrose metabolism pathways were also identified in the KEGG library for *Arabidopsis* (Fig. 4). This result indicated that alpha-glucan phosphorylase 1, putative pectinesterase/pectinesterase inhibitor 17, beta-amylase, and alpha-glucan phosphorylase isozyme H might be important factors involved in variation in starch and sucrose

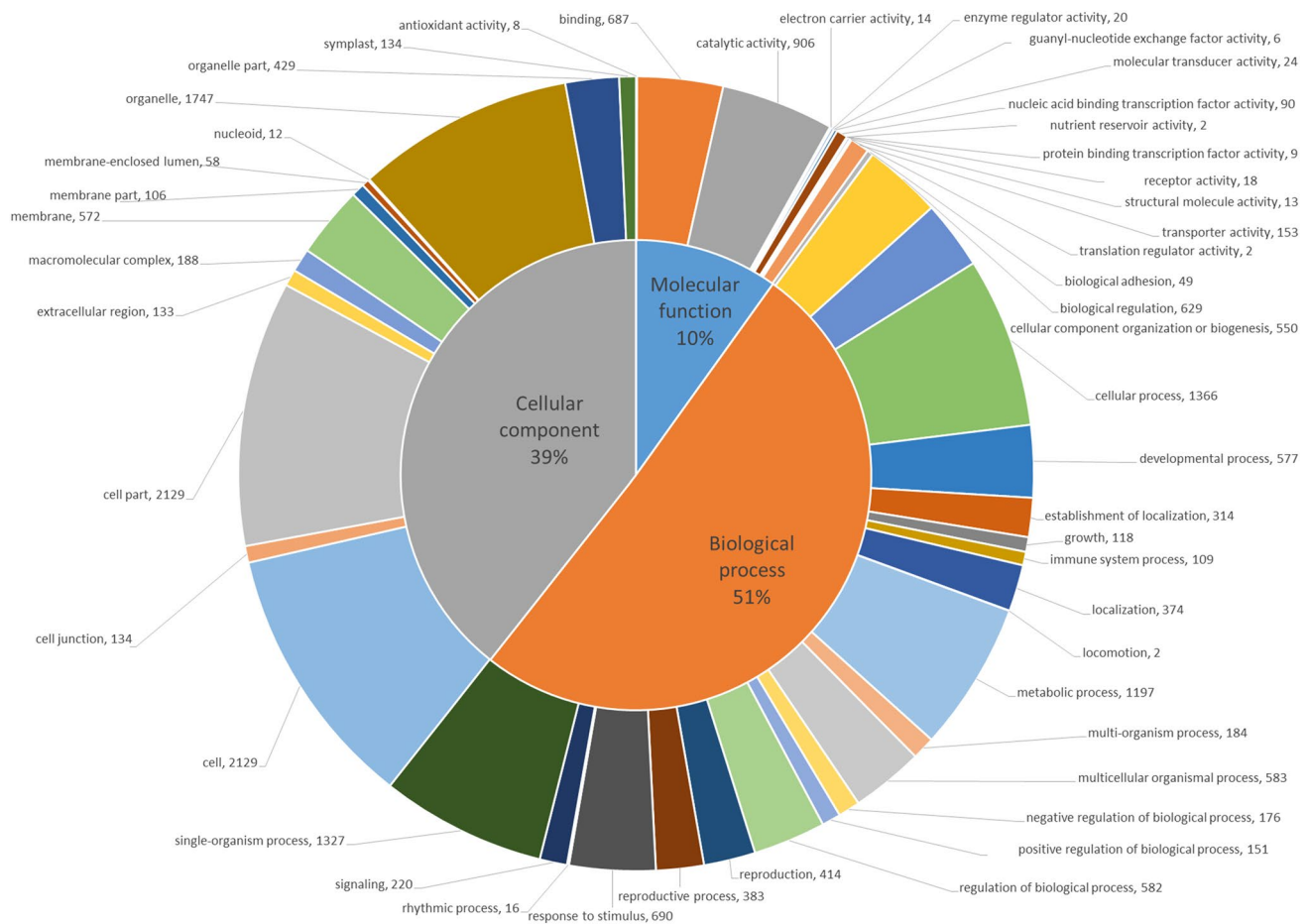


Fig. 2 Functional classification of *P. ginseng* unigenes based on Gene Ontology (GO) annotation. The BLAST2GO program was used to obtain GO terms for the *P. ginseng* unigenes. Shown are the level 2

GO distributions of cellular component, molecular function, and biological process categories

metabolism between FCG and MCG in different environments (Table 3).

Conclusion

Molecular approaches mainly based on transcriptome analyses revealed that gene expression changes are associated with broad variation in plant growth characteristics. Beside the transcriptional changes strictly related to primary and secondary metabolism, other major alterations were found in host tissues and included the activation of related genes in plants. In this study, we sought to investigate the whole transcriptome landscape of ginseng. Growth of MCG is much slower than FCG, although MCG can survive for far longer (several decades). However, little information is available as to why MCG grows slowly and is so long-lived. Furthermore, variations in the transcriptional characteristics of FCG and MCG have not been clearly demonstrated.

We employed high-throughput sequencing technologies using the Illumina HiSeq™2500 system, and generated a large amount of transcriptome data from ginseng roots. Approximately 77 and 87 million high-quality reads were produced from the FCG and MCG roots transcriptome analyses, respectively, and we obtained 256,032 assembled unigenes with an average length of 1171 bp using de novo assembly methods. Functional annotation for the unigenes was performed using sequence similarity comparisons against the following databases: the non-redundant nucleotide database, the InterPro domains database, the Gene Ontology Consortium database, and the KEGG pathway database. A total of 4207 unigenes were assigned to specific metabolic pathways, and all of the known enzymes involved in the starch and sucrose metabolism pathways were also identified in the KEGG library from Arabidopsis. Arabidopsis is a particularly good species for this sort of forward genetic approach to the elucidation of pathways of starch turnover: it is a diploid, inbreeding species with a short life cycle and very pronounced daily changes in starch content

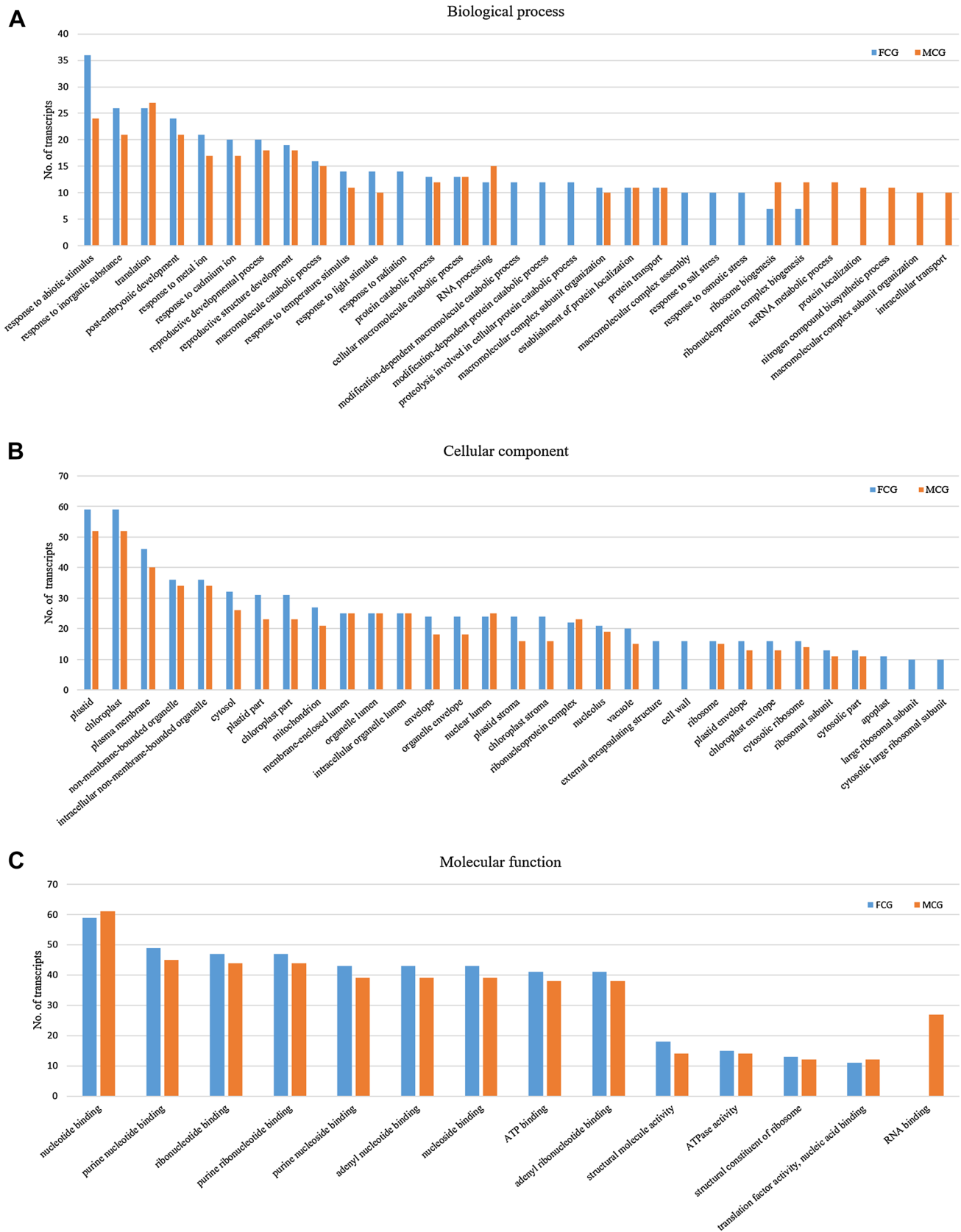


Fig. 3 Functional distribution of the 100 most highly expressed transcripts from FCG and MCG in the categories of **a** biological process, **b** cellular component, and **c** molecular function

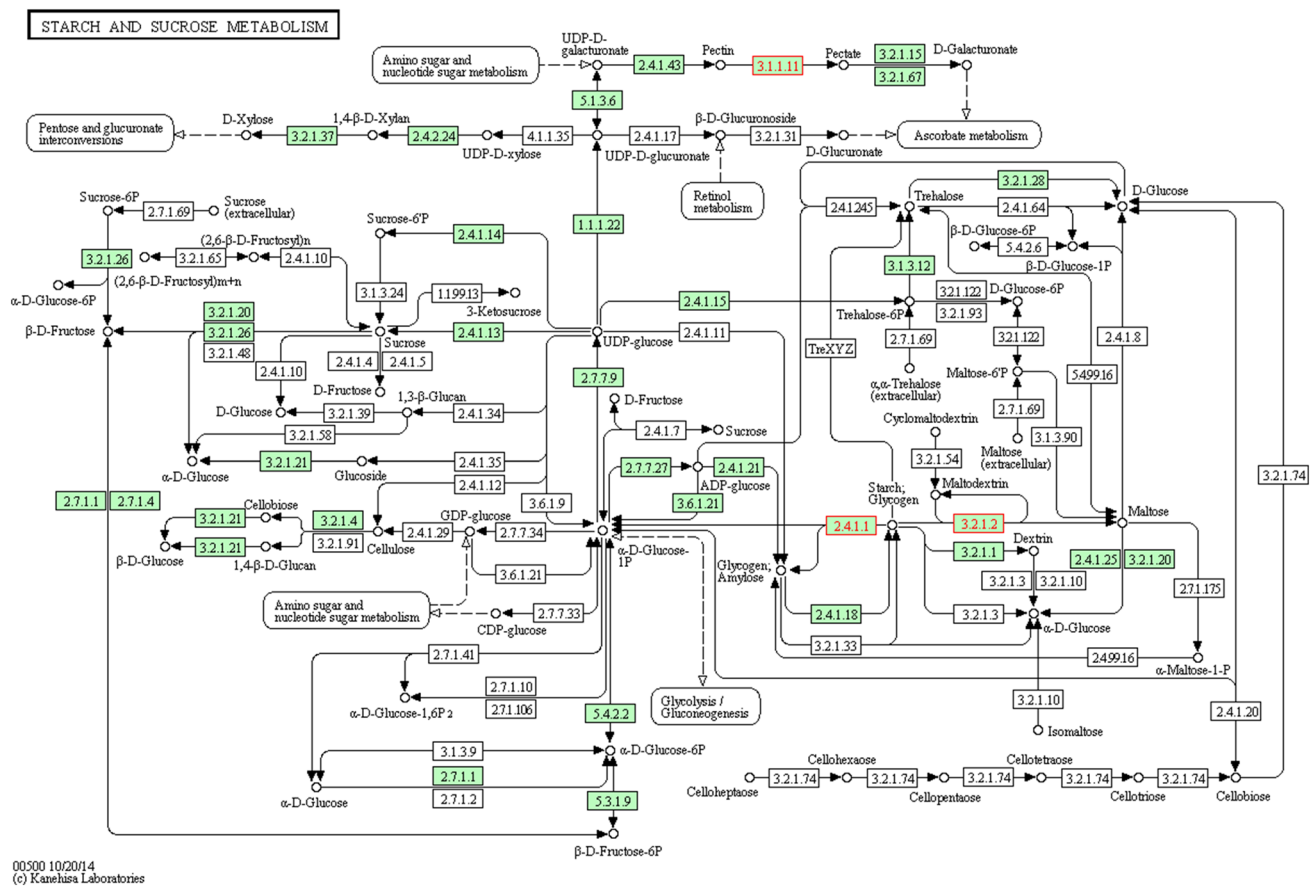


Fig. 4 Pathway assignments based on the Kyoto Encyclopedia of Genes and Genomes (KEGG)

Table 3 Identified genes putatively related to the starch and sucrose metabolism pathway

Gene	EC number	Gene name ^a	Arabidopsis gene ^b
Starch and sucrose metabolism pathway			
Alpha-glucan phosphorylase 1	2.4.1.1	PHS1	AT3G29320
Putative pectinesterase/pectinesterase inhibitor 17	3.1.1.11	PME17	AT2G45220; AT3G60730; AT4G00190
Beta-amylase	3.2.1.2	CT-BMY	AT4G17090
Alpha-glucan phosphorylase isozyme H	2.7.9.3	PHS2	AT3G46970

^aFollowing the *Arabidopsis thaliana* gene nomenclature

^bGene list from the KEGG pathway database

in most plants. Also, publication of the *Arabidopsis* genome sequence has greatly enhanced speed and ease of gene identification. Carbon (C) and nitrogen (N) metabolism are the basis for life on Earth. The production, balance, and trade-offs of C and N metabolism are critical to all plant growth, yield, and local adaptation (Coruzzi and Bush 2001; Gutiérrez et al. 2007). In plants, there is a critical balance between tissues that are producing energy (sources) and those using it (sinks), as the identities and locations of these vary with time and developmental stage. While a great deal of research

has focused on the key genes and proteins involved in these processes (Kim et al. 2000; Takahashi et al. 2009; Wang et al. 1993), relatively little is known about natural variation within a species in different cultivation environments. This study indicated that alpha-glucan phosphorylase 1, putative pectinesterase/pectinesterase inhibitor 17, beta-amylase, and alpha-glucan phosphorylase isozyme H might be important factors involved in starch and sucrose metabolism variation between FCG and MCG in different environments. Starch is of importance both as a carbon storage reserve in plants and

as a biotechnological product. In most plants, sucrose is the major transport form and starch the major storage form of carbohydrates. The conversion of sucrose to starch in storage tissues has been the subject of intense research for many years. Starch represents a cornerstone for human and animal nutrition and a feedstock for many industrial applications, including bioethanol production. We examined the regulation of starch and sucrose metabolism of ginseng plants cultivated in different environments and provided new insight, but further research into these processes is needed. For example, it remains unclear whether the mechanisms discussed herein play a role in biosynthesis, and thus whether they are variables affecting the influence of different cultivation environments.

Acknowledgements We would like to thank Dr. Woo Jun Sul of the Department of Systems Biotechnology at Chung-Ang University for their constructive comments on the initial manuscript. We would also like to thank Dr. Junsu Ko of Theragen BiO Institute at TheragenEx for RNA-Seq. library construction.

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
- Alvord G, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA (2007) The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* 8:183
- Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:106
- Birol I et al (2009) De novo transcriptome assembly with ABySS. *Bioinformatics* 25:2872–2877. doi:10.1093/bioinformatics/btp367
- Blanco E, Parra G, Guigó R (2007) Using geneid to identify genes. *Curr Protoc Bioinform*. doi:10.1002/0471250953.bi0403s18
- Choi D-W, Jung J, Im Ha Y, Park H-W, In DS, Chung H-J, Liu JR (2005) Analysis of transcripts in methyl jasmonate-treated ginseng hairy roots to identify genes involved in the biosynthesis of ginsenosides and other secondary metabolites. *Plant Cell Rep* 23:557–566
- Choi Y-E, Kim Y-S, Yi M-J, Park W-G, Yi J-S, Chun S-R, Han S-S, Lee S-J (2007) Physiological and chemical characteristics of field- and mountain-cultivated ginseng roots. *J Plant Biol* 50:198–205
- Coruzzi G, Bush DR (2001) Nitrogen and carbon nutrient and metabolite signaling in plants. *Plant Physiol* 125:61–64
- Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 4:P3
- Götz S et al (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 36:3420–3435
- Grabherr MG et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652
- Gutiérrez RA, Lejay LV, Dean A, Chiaromonte F, Shasha DE, Coruzzi GM (2007) Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in *Arabidopsis*. *Genome Biol* 8:R7
- Haas BJ et al (2013) De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat Protoc* 8:1494–1512
- Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Res* 9:868–877
- Jung C-H, Seog H-M, Choi I-W, Cho H-Y (2005) Antioxidant activities of cultivated and wild Korean ginseng leaves. *Food Chem* 92:535–540
- Kadota K, Nishiyama T, Shimizu K (2012) A normalization strategy for comparing tag count data. *Algorithms Mol Biol* 7:5. doi:10.1186/1748-7188-7-5
- Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30
- Kim J-Y, Mahé A, Brangeon J, Prioul J-L (2000) A maize vacuolar invertase, IVR2, is induced by water stress organ/tissue specificity diurnal modulation of expression. *Plant Physiol* 124:71–84
- Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi:10.1186/1471-2105-12-323
- Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nat Rev Genetics* 12:671–682
- Nam KY (2002) Clinical applications and efficacy of Korean ginseng (*Panax ginseng* C.A. Meyer). *J Ginseng Res* 26:111–131
- O'Hara M, Kiefer D, Farrell K, Kemper K (1998) A review of 12 commonly used medicinal herbs. *Arch Family Med* 7:523–536
- Perteu G et al (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19:651–652
- Sun J, Nishiyama T, Shimizu K, Kadota K (2013) TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics* 14:219. doi:10.1186/1471-2105-14-219
- Takahashi H et al (2009) Pleiotropic modulation of carbon and nitrogen metabolism in *Arabidopsis* plants overexpressing the NAD kinase2 gene. *Plant Physiol* 151:100–113
- Trapnell C et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28:511–515. doi:10.1038/nbt.1621
- Wang J-L, Turgeon R, Carr JP, Berry JO (1993) Carbon sink-to-source transition is coordinated with establishment of cell-specific gene expression in a C4 plant. *Plant Cell* 5:289–296
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63
- Yang Y, Smith SA (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genom* 14:328. doi:10.1186/1471-2164-14-328
- Yun TK, Lee YS, Lee YH, Kim SI, Yun HY (2001) Anticarcinogenic effect of *Panax ginseng* C.A. Meyer and identification of active compounds. *J Korean Med Sci* 16:6–18