creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.0 KOREA

You are free to :

- Share — copy and redistribute the material in any medium or format

Under the follwing terms :

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

NonCommercial — You may not use the material for commercial purposes.

NoDerivatives — If you remix, transform, or build upon the material, you may not distribute the modified material.

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

This is a human-readable summary of (and not a substitute for) the license.

Disclaimer

dCollection

A DISSERTATION FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

Adaptive evolution of polyploid ginseng (*Panax ginseng*) revealed by genome annotation and comparative transcriptomes

BY

MURUKARTHICK JAYAKODI

FEBRUARY, 2018

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Adaptive evolution of polyploid ginseng (*Panax ginseng*) revealed by genome annotation and comparative transcriptomes

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 DOCTOR OF PHILOSOPHY
BY THE COMMITTEE MEMBERS

FEBRUARY, 2018

CHAIRMAN _____
Doil Choi, Ph.D.

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

MEMBER _____
Choonkyun Jung, Ph.D.

MEMBER _____
Beom-Seok Park, Ph.D.

MEMBER _____
Senthil Natesan, Ph.D.

Adaptive evolution of polyploid ginseng (*Panax ginseng*) revealed by genome annotation and comparative transcriptomes

MURUKARTHICK JAYAKODI

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

GENERAL ABSTRACT

Panax ginseng C. A. Meyer, reputed as the king of medicinal herbs, has slow growth, long generation time, low seed production, and complicated genome structure that hamper its study. Furthermore, the knowledge of molecular responses to various abiotic stresses is still limited in *P. ginseng*. To facilitate its functional genomics, metabolomics and breeding in *P. ginseng*, I have performed four independent studies or chapters. With advent of sequencing technologies, the ginseng genome project was initiated in 2011 and the first draft genome assembly was completed in 2016. In first chapter, I annotated a total of 59,352 protein coding genes in tetraploid *P. ginseng*. Of them, 97% of the genes got functional descriptions. A total of 3, 588 transcription factors and 851 transcription regulators were identified and grouped into 94 families. Functional and evolutionary analyses suggest that production of pharmacologically important dammarane type ginsenosides originated in *Panax* and are produced largely in shoot tissues and transported to roots; that

newly evolved *P. ginseng* fatty acid desaturases increase freezing tolerance; and that unprecedented retention of chlorophyll a/b binding protein genes enables efficient photosynthesis under low light. Furthermore, eleven novel candidates UDP-glucuronosyltransferase (UGTs) were identified through integrated transcriptome and metabolome data.

Heat and light stress poses an important threat to the growth and sustainable production of ginseng. Efforts have been made to study the effects of high temperature on ginseng physiology, but knowledge of the molecular responses to heat stress is still limited. Thus, in the second chapter, I have compared the transcriptomes (RNA-Seq) of two ginseng cultivars, Chungpoong (CP) and Yunpoong (YP), which are sensitive and resistant to heat stress, respectively, after 1- and 3-week heat treatments. Differential gene expression (DEG) and gene ontology (GO) enrichment along with profiled chlorophyll contents were performed. CP is more sensitive to heat stress than YP, and exhibited a lower chlorophyll content than YP. Moreover, heat stress reduced the chlorophyll content more rapidly in CP compared to YP. A total of 329 heat-responsive genes were identified. Intriguingly, genes encoding chlorophyll ab binding (CAB) proteins, WRKY transcription factors, and fatty acid desaturase (FAD) were predominantly responsive during heat stress and appeared to inhibit photosynthesis. In addition, a genome-wide scan of photosynthetic and sugar metabolic genes revealed reduced transcript levels for *ribulose 1,5-bisphosphate carboxylase/oxygenase* (*RuBisCO*) under heat stress, especially in CP, possibly attributable to elevated levels of soluble sugars.

Long noncoding RNAs (lncRNAs) have been implicated with diverse biological roles including genome regulation, various developmental processes and diseases. In the third chapter, through a systematic pipeline

using ~104 billion sequencing RNA reads from various tissues, stages of growth and abiotic stress treatments of *P. ginseng*, I catalogued 19,495 and identified more than 100 candidate lncRNAs involved in abiotic stress responses to drought, salt, cold, heat and methyl jasmonate (MeJA) and 2,607 involved in specialized unknown function in specific tissues and growth stages of *P. ginseng*. Further, transposons might have been the contributor for the functional potential of lncRNAs in ginseng.

Having realized the importance of this plant to humans, an integrated omics resource becomes indispensable to facilitate genomic research, molecular breeding and pharmacological study of this herb. In the fourth chapter, using the draft genome, transcriptome, and functional annotation datasets of *P. ginseng*, I developed the Ginseng Genome Database http://ginsengdb.snu.ac.kr/, the first open-access platform to provide comprehensive *P. ginseng* genomic resources. The current version of this database provides the latest draft genome sequence along with the structural and functional annotations of genes and digital expression of genes based on transcriptome data from different tissues, growth stages and treatments. In addition, tools for visualization and analysis of genomic data are provided. All data in the database were manually curated and integrated within a user-friendly query page. Overall, this study will enable us to develop new cultivars carrying resistant to biotic/abiotic stresses, tolerant to direct sun light, and improving medicinal values of ginseng either through genomics-assisted breeding or metabolic engineering.

Key words: *Panax ginseng,* genome annotation, long noncoding RNAs, adaptation, database

Student number: 2014-30835

CONTENTS

CHAPTER 2. Comparative transcriptome analysis of heat stress responsiveness between two contrasting ginseng cultivars

CHAPTER 3. Genome-wide screening of transcriptomes revealed the landscape of long noncoding RNAs in ginseng (*Panax ginseng*)

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CP	Chunpoong
CS	Cheongsun
SH	Sunhyang
SU	Sunun
SRA	Sequence Read Archive
GO	Gene Ontology
FPKM	fragments per kilobase of exon per million fragments mapped
lncRNA	Long noncoding RNA
CPC	Coding Potential Calculator
EST	expressed sequence tag
MP	mate-paired
PE	paired-end
SQE	squalene epoxidase
OSC	oxidosqualene cyclase
FAD	fatty acid desaturase
CAB	chlorophyll a/b binding

GENERAL INTRODUCTION

Ginseng (called Asian/Korean ginseng: *Panax ginseng* C. A. Meyer) belongs to the family Araliaceae. It is regarded as "the king of herbal medicinal plant," whose roots have been used in traditional medicine for thousands of years [1]. Currently, ginseng has become one of the most important agricultural commodities in Asia including Korea and China, whose distribution markets, together with *P. quinquefolius* (American ginseng), was estimated at over 2 billion USD as of 2009 [2]. Plant species in the genus *Panax* are shady perennials and classified either diploid or tetraploid based on its chromosome number3. Most diploid ginseng species such as *P. notoginseng*, *P. vietnamensis*, *P. bipinnatifidus*, *P. stipuleanatus* and *P. pseudoginseng* grow in high altitude limited regions in warm area without freezing winter from Eastern Himalayas onward through Southern China to north and central highlands of Vietnam. However, tetraploid species including *P. ginseng* and *P. quinquefolius* grow broadly in the northern hemisphere, in North Eastern Asia and North America, respectively and have a capability of overwintering.

Notable therapeutic effects of *P. ginseng* on life-threatening diseases such as neurodegenerative [4,5] and cardiovascular diseases [6], diabetes [7] and cancer [8,9] have been well documented. Such medicinal efficacies are often attributed to its various unique saponins called ginsenosides, which are glycosylated triterpenes and classified as either dammarane- (exclusively biosynthesized in the genus *Panax*) or oleanane-type ginsenosides based on the skeletal structure of aglycones. Ginsenosides are accumulated in roots, leaves, stems, flower buds and berries with varying quantity according to tissues [10,11], age [10,12], and environmental condition [13,14]. Similarly,

the ginsenoside accumulation pattern also varies between ginseng cultivars [15]. Understanding such variation in the ginsenoside accumulation is essential for finding the responsible genes and the increased production of ginsenosides through genetic or metabolic approach. Due to lack of multifaceted omics resources, the underlying regulatory mechanisms of that variation remain elusive. Furthermore, limited genomic resources and lack of genetic population caused by slow-growth (~4 years per generation), sensitivity to environmental stresses, and very low seed yield (40 seeds) per generation have hampered genetic mapping and following molecular genetic improvements. Therefore, I have done four independent studies that provide deep insights into ginseng genome.

In the first chapter, the whole genome annotation including protein coding gene prediction, functional annotations, genome-wide noncoding RNA prediction and gene family characterization have been done. Intriguingly, genes encoding enzymes involved in ginsenoside biosynthetic pathway were characterized along with expression profiling. Using integrated multi-omic approach candidate genes for high accumulation of ginsenosides were identified. Finally, a comparative gene family analysis identified genes responsible for adaptation to shade and cold environments.

In the second chapter, as many elite ginseng cultivars have been developed, and ginseng cultivation has become well established during the last century. However, heat stress poses an important threat to the growth and sustainable production of ginseng. Therefore, a comprehensive comparative transcriptome analysis was conducted using two contrasting ginseng cultivars, Chungpoong (CP) and Yunpoong (YP), which are sensitive and resistant to heat stress, respectively, after 1- and 3-week heat treatments. This study reveals candidate loci/gene targets for breeding and

functional studies related to developing high-temperature tolerant ginseng varieties.

The major biological processes such as regulation of shady nature, long life span and environmental adaptation in ginseng need to be largely explored in the perspectives of both coding and noncoding RNAs to increase the production and availability for human use. Therefore, in the third chapter, a comprehensive set of lncRNAs were identified and characterized using ginseng draft genome sequence and RNA-seq datasets from 39 samples used for ginseng genome annotation. Total lncRNAs were classified as intergenic, intronic and anti-sense lncRNAs according their genomic proximity. The dynamic expression profiling has detected the functional potential lncRNAs involved in abiotic stress, secondary metabolite biosynthesis, and environmental adaptation.

In the fourth chapter, I built a dynamic database that integrates a draft genome sequence, transcriptome profiles, and annotation datasets of ginseng. This Ginseng Genome Database is now publicly available (http://ginsengdb.snu.ac.kr/) for the use of scientific community around the globe for exploring the vast possibilities. This user-friendly database will serve as a hub for mining gene sequences and their digital expression data of samples from various tissues, developmental stages, and treatments. Our database interface will facilitate the easy retrieval of gene families and associated functional annotations using InterPro, KEGG, BLAST and Gene Ontology (GO) databases. To expedite metabolomics in ginseng, we have made a separate section that categorizes the genes associated with various metabolic pathways including the ginsenoside biosynthesis pathway. In addition, we have included robust tools such as BLAST and genome browser (JBrowse) for survey and visualization of ginseng genomic features.

REFERENCES

1. Wang J, Gao WY, Zhang J, Zuo BM, Zhang LM, Huang LQ: Advances in study of ginsenoside biosynthesis pathway in *Panax ginseng* C. A. Meyer. *Acta Physiol Plant* 2012, 34(2):397-403.

2. Saito H, Yoshida Y, Takagi K: Effect of *Panax Ginseng* root on exhaustive exercise in mice. *Jpn J Pharmacol* 1974, 24(1):119-127.

3. Peng D, Wang H, Qu C, Xie L, Wicks SM, Xie J: Ginsenoside Re: Its chemistry, metabolism and pharmacokinetics. *Chin Med* 2012, 7:2.

4. Attele AS, Wu JA, Yuan CS: Ginseng pharmacology: multiple constituents and multiple actions. *Biochem Pharmacol* 1999, 58(11):1685-1693.

5. Shang W, Yang Y, Zhou L, Jiang B, Jin H, Chen M: Ginsenoside Rb1 stimulates glucose uptake through insulin-like signaling pathway in 3T3-L1 adipocytes. *J Endocrinol* 2008, 198(3):561-569.

6. Radad K, Gille G, Liu L, Rausch WD: Use of ginseng in medicine with emphasis on neurodegenerative disorders. *J Pharmacol Sci* 2006, 100(3):175-186.

7. Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park M, Choi D, Lim YP, Kwon SJ, Park BS *et al*: Major repeat components covering one-third of the ginseng (*Panax ginseng* C.A. Meyer) genome and evidence for allotetraploidy. *Plant J* 2014, 77(6):906-916.

8. Waminal NE, Park HM, Ryu KB, Kim JH, Yang TJ, Kim HH: Karyotype analysis of *Panax ginseng* C.A.Meyer, 1843 (Araliaceae) based on rDNA loci and DAPI band distribution. *Comp Cytogenet* 2012, 6(4):425-441.

9. Jayakodi M, Lee SC, Park HS, Jang W, Lee YS, Choi BS, Nah GJ, Kim DS, Natesan S, Sun C *et al*: Transcriptome profiling and comparative analysis of *Panax ginseng* adventitious roots. *J Ginseng Res* 2014, 38(4):278-288.

10. Jayakodi M, Lee SC, Lee YS, Park HS, Kim NH, Jang W, Lee HO, Joh HJ, Yang TJ: Comprehensive analysis of *Panax ginseng* root transcriptomes. *BMC Plant Biol* 2015, 15:138.

11. Lee Y, Park HS, Lee DK., Jayakodi M, Kim NH, Koo HJ, Lee SC, Kim YJ, Kwon SW, Yang TJ. Integrated transcriptomic and metabolomic analysis of five *Panax ginseng* cultivars reveals the dynamics of ginsenoside biosynthesis. *Front Plant Sci* 2017, 8:1048.

12. Li C, Zhu Y, Guo X, Sun C, Luo H, Song J, Li Y, Wang L, Qian J, Chen S: Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng* C. A. Meyer. *BMC Genomics* 2013, 14:245.

13. Kim K, Lee SC, Lee J, Lee HO, Joh HJ, Kim NH, Park HS, Yang TJ: Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. *PloS one* 2015, 10(6):e0117159.

14. Kim K, Lee SC, Lee J, Yu Y, Yang K, Choi BS, Koh HJ, Waminal NE, Choi HI, Kim NH, *et al*: Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. *Sci Rep* 2015, 5:15655.

15. Kim K, Nguyen VB, Dong JZ, Wang Y, Park JY, Lee SC, Yang TJ: Evolution of the Araliaceae family inferred from complete chloroplast genomes and 45S nrDNAs of 10 *Panax*-related species. *Sci Rep* 2017, 7:4917.

CHPTER 1

Ginseng genome annotation and genes involved in ginsenoside biosynthetic pathway

Abstract

Ginseng (*Panax ginseng*) is reputed as the king of medicinal herb owing to its high therapeutic values for thousand years. However, a slow growth under shade, longer than 4 years of growing for one generation to bear tens of seeds, duplicated genomic structure and huge repeats have hampered the genetic and genomic research for this plant. Ginseng is remained as mysterious plants for its function. Recently, the draft genome assembly was made using paired-end (PE) reads covering 746 Gbp (206x) and mate-pair (MP) reads covering 365 Gbp (101x) from Chunpoong. These reads were assembled into 9,845 scaffolds covering 2.98 Gbp with N50 of 569 Kbp and longest scaffold of 3.6 Mbp. Here, I annotated 59,352 protein coding genes of which 38% of them contained alternative splicing (AS) transcripts. Homology based functional annotation revealed 97% of protein coding genes attained functional descriptions. A total of 3,588, transcription factor and 851 transcription regulators were annotated in ginseng. Further, genes involved in ginsenosides biosynthesis were identified. Additionally, I clarified the explicit biosynthesis of ginsenosides in the genus *Panax* and expanded our understanding in varying accumulation pattern through integrated transcriptome and metabolome analyses. Functional and evolutionary analyses suggest that newly evolved *P. ginseng* fatty acid desaturases increase freezing tolerance; and that unprecedented retention of chlorophyll a/b binding protein genes enables efficient photosynthesis under low light. These results will enable us to develop new cultivars carrying resistant to biotic/abiotic stresses, tolerant to direct sun light, and improving medicinal

values of ginseng either through genomics-assisted breeding or metabolic engineering.

Keywords: *Panax ginseng*, ginsenosides, evolution, metabolic network, adaptation

Introduction

Roots of Asian/Korean ginseng have been used for thousands of years, today being an important Asian agricultural commodity with markets (together with *P. quinquefolius*, American ginseng) estimated at over 2 billion USD [1]. *Panax* species are shade-requiring perennials [2]. Most diploid *Panax* such as *P. notoginseng*, *P. vietnamensis, P. bipinnatifidus, P. stipuleanatus* and *P. pseudoginseng* grow at high altitudes in warm freeze-free areas from the Eastern Himalayas through Southern China to north and central highlands of Vietnam. Tetraploid *P. ginseng* and *P. quinquefolius* overwinter and are broadly distributed in northeastern Asia and North America, respectively.

Therapeutic effects of *P. ginseng* on neurodegenerative disorders [3, 4], cardiovascular diseases [5], diabetes [6] and cancer [7, 8] are often attributed to unique saponins called ginsenosides, glycosylated triterpenes classified as either dammarane- (*Panax*-specific) or oleanane-type based on aglycone skeletal structure. Ginsenosides are accumulated in roots, leaves, stems, flower buds and berries, in quantities varying with tissue [9, 10] age [9, 11] environment [12, 13] and cultivar [14]. Limited genomic resources and genetic populations due to slow-growth (~4 years/generation), sensitivity to environmental stresses, and low seed yield (40/generation) hamper developmental and genetic studies and breeding.

Here I report the genome annotation of *P. ginseng* cultivar (cv.) Chunpoong (ChP). Investigation of the *P. ginseng* genome provides new insights into adaptation and clarifying the origin and regulation of ginsenoside accumulation. These discoveries provide a foundation for improving therapeutic effects, understanding shade plant biology, and empowering Araliaceae genomic studies.

Results and discussion

Genome assembly

Paired-end (PE) reads covering 746 Gbp (206x) and mate-pair (MP) reads covering 365 Gbp (101x) from ChP were assembled into 9,845 scaffolds covering 2.98 Gbp with N50 of 569 Kbp and longest scaffold of 3.6 Mbp. The predicted *P. ginseng* genome size ranged from 3.3 to 3.6 Gbp through flow cytometry and *k*-mer frequency, slightly bigger than the reported 3.12 Gbp [15]. Assembly accuracy and completeness was indicated by: correct mapping of 90% of reads from four MP libraries; alignment to 13 finished bacterial artificial chromosome (BAC) sequences [16, 17] showing 99% homology with perfect contiguity; and Benchmarking Universal Single-Copy Orthologs (BUSCO_v2) analysis finding 1339 (93%) of 1440 conserved orthologous angiosperm genes assembled completely.

Genome structural annotation

Integrated pipeline for genome annotation (IPGA) strategy was developed and applied for *P. ginseng* gene structure annotation (**Figure 1-1**). In this strategy, the repeat unmasked draft genome sequences were used for mapping of transcriptome, protein and ESTs mapping and consensus gene model constructions. IPGA combined the evidences from the mapping of RNA-seq transcripts, ESTs and proteins to the reference draft genome as

well as independent gene prediction based on the RNA-seq mapping. The EvidenceModeler [18] (EVM) program combines these evidences and predicted gene models to build consensus gene structures.

Evidence based gene predictions

A total of 39 RNA-seq samples of ChP cultivar were generated from various tissues, different stage of plants and samples with abiotic stress treatments and all data were used for gene annotation. Each sample was mapped to the reference genome using HISAT[19]. Then, the reference-guided genome assembly was performed separately for each sample using StringTie [20]. All the individual assemblies were merged using the cuffmerg utility in Cufflinks [21]. From the reference transcriptome, the putative coding regions were predicted using TransDecoder (https://transdecoder.github.io/). A custom Python script was used to select the most appropriate representative gene structure from each loci as reference based transcriptome assembly included a set of unigenes from single loci as isoforms or spliced variants. These genome based coding models prevented the reflection of merged gene models by merged transcriptome assembly in EVM analysis. In addition, *de novo* as well as reference-guided transcriptome assembly was also made using Trinity [22]. Both the *de novo* and reference-guided transcriptome assembly by Trinity were aligned to the ginseng draft genome sequences to make alignment assemblies using PASA [23] with Cufflinks transcript structures (gtf file). Additionally, a total of 17,773 ESTs from NCBI dbEST was also aligned to reference genome using GMAP [24] to increase the accuracy of gene prediction. Since PASA and GMAP produced

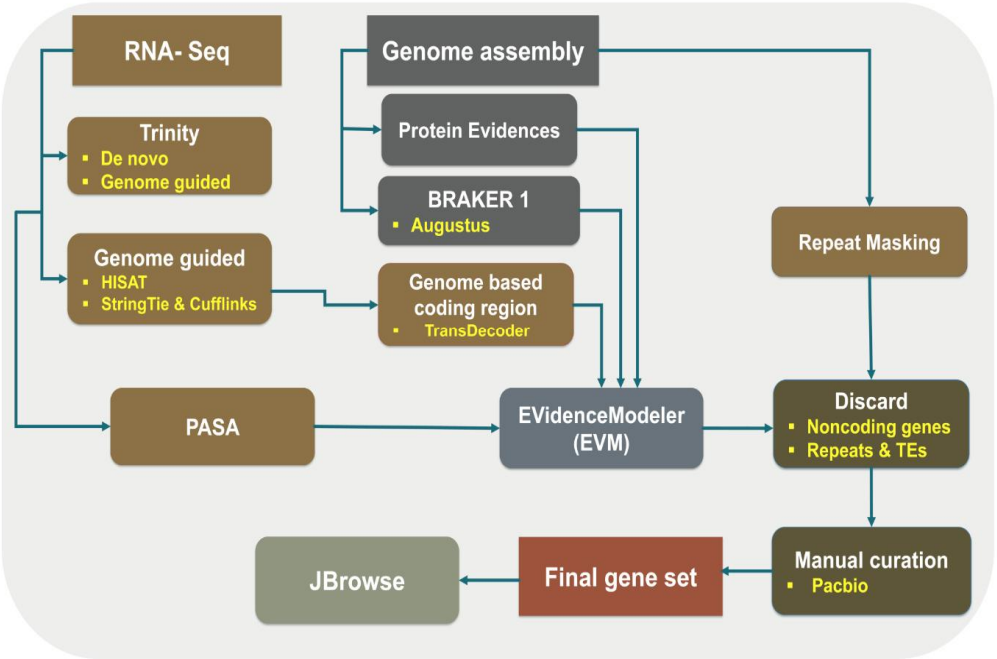

Figure 1-1. **Integrated pipeline for genome annotation (IPGA).** The pipeline figure shows the methodology and programs used for *P. ginseng* genome annotation. To annotate the genic regions in *P. ginseng* draft genome, RNA-seq data generated from *P. ginseng* cv. ChP were assembled by both *de novo* and reference based methods and then resulting gene information was merged with *ab initio* and protein evidences to make consensus gene models. Further curation of gene model was done using PacBio transcriptome and then the curated gene set was used to construct *P. ginseng* genome browser.

the high quality spliced alignments, the maximum weight value was given for these output file (gff3 file) during the EVM weighted consensus gene structure prediction. The protein sequences of tomato (*Solanum*

lycopersicum), potato (*Solanum tuberosum*), grape (*Vitis vinifera*), Arabidopsis (*Arabidopsis thaliana*), pepper (*Capsicum annum*), and coffee (*Coffea canephora*) were aligned to the draft genome using Spaln2 [25, 26]. Spliced alignments of protein improve the precision of eukaryotic genes and their structure prediction. Spaln can rapidly locate the genomic regions corresponding to the target protein sequences and perform spliced alignment in a single job. The Splan2 output was also considered as reliable evidence and used for constructing consensus gene model using EVM.

Ab initio gene predictions

BRAKER 1 [27], which uses the advantages RNA-seq spliced alignments and GeneMark-ET [28] and AUGUSTUS [29] tools, prediction was used as *de novo* annotation for EVM. For BRAKER 1, all the 39 RNA-seq alignment was merged and utilized for unsupervised training. After training, an *ab initio* gene models was created by GeneMark-ET. Subsequently, RNA-seq supported gene structures were selected for automated training of AUGUSTUS. Then, the more accurate gene models were predicted by AUGUSTUS based on the spliced alignment evidences. The BRAKER1 output file was reformatted using EVM utilities for EVM prediction.

Integration of evidences with EVM

Gene models from mapped transcript data and protein data and *ab initio* gene models were combined into consensus gene model using EVM. All the gene models and alignment evidences were prepared in gff3 format to run EVM. The evidence weight was set intuitively for combining diverse evidence types into accurate gene structure annotation. The input genome sequences and predicted gene models were partitioned into smaller overlapping chunks as to reduce the memory requirements. Then, EVM was executed on each of

the data partitions and the output corresponding to the partitioned data were joined into single output. Finally, the raw output generated by EVM was converted to the standard GFF3 format. The EVM gene models showing longer and shorter gene length and high number of exons containing genes were validated with NCBI Nr database and in-house python script. Further, the non-protein coding genes and transposon genes were discarded. Finally, the consensus gene models were refined using PacBio data.

Filtration of non-protein coding genes

All the gene models predicted by EVM were analyzed to identify putative long noncoding RNA (lncRNA) genes. In this process, genes with ≥200 nucleotide length and ORF (open reading frame) ≤ 100 amino acids were discarded. From the filtered genes, homology search with Swiss-Prot protein databases with E-value cutoff of 1E-03 and Pfam domain search were also conducted to remove putative protein coding genes. Additionally, to eliminate other classes of non-coding RNAs including (transfer (t) RNAs, small nuclear (sn) RNAs and small nucleolar (sno) RNAs), a housekeeping RNA database was made using tRNA sequences from the genomic tRNA database (http://gtrnadb.ucsc.edu/), rRNAs from the silva database (http://www.arb-silva.de/no_cache/download/archive/current/Exports/), and other ncRNAs (snRNAs, snoRNAs, 7SL/SRP) downloaded from NONCODE (http://noncode.org/). The remaining genes were aligned to the housekeeping database with an E-value cutoff of 1E-10 and other classes of noncoding genes were also discarded. Finally, the coding potential was accessed using coding potential calculator (CPC) [30] and CPAT [31]. Genes with CPC score ≤-1.0 and CPAT score <0.39 were considered as putative lncRNA genes and thus, they were removed.

Filtration of transposon genes

The repeat masked information was used to remove genes predicted from repeat masked region. An in-house Python script was used to identify genes loci that span over 40% of the repeat masked region of the draft genome and removed these gene models because these were not perceived as true protein coding genes. TEs also contain complete ORF gene for their transposition. In order to remove TE genes, Pfam domain search was performed using HMMER [32]. A list of TE domains were collected from the literatures as well as manual inspection. Finally, TE domain containing genes in the EVM gene models were also identified and discarded.

Curation of gene models using PacBio

A total of 184,171 high quality contigs were generated using PacBio Iso-Seq method. These contigs were aligned to the draft genome using GMAP and complete ORF was predicted using TransDecoder. Approximately, 73% of the PacBio data were consisted full-length ORF structure. Using full-length structure, the split gene models (i.e. single gene fragmented into two or more genes) and merged gene (i.e. two or more combined genes separated into one or more) models in the EVM consensus genes were curated. A total of 3,797 erratic gene models were refined and 1,395 new models were attained which were not predicted by EVM. In addition, genes containing internal stop codon were also refined using PacBio data. Finally, a total of 59,352 (IPGA v1.1) putative protein coding genes were obtained using IPGA pipeline. The CDS feature was similar to other plant species (**Figure 1-2**). In particular, the average intron size was increased (**Table 1-1**). A large intron size of 63 kbp was identified in the final gene set (Pg_S1988.16). This large intron gene structure was perfectly supported by the PacBio data. Before this curation, the large intron split the one gene as two gene models predicted by EVM.

This indicated that the need of PacBio data to obtain comprehensive gene sets for newly sequenced genome of non-model eukaryotes.

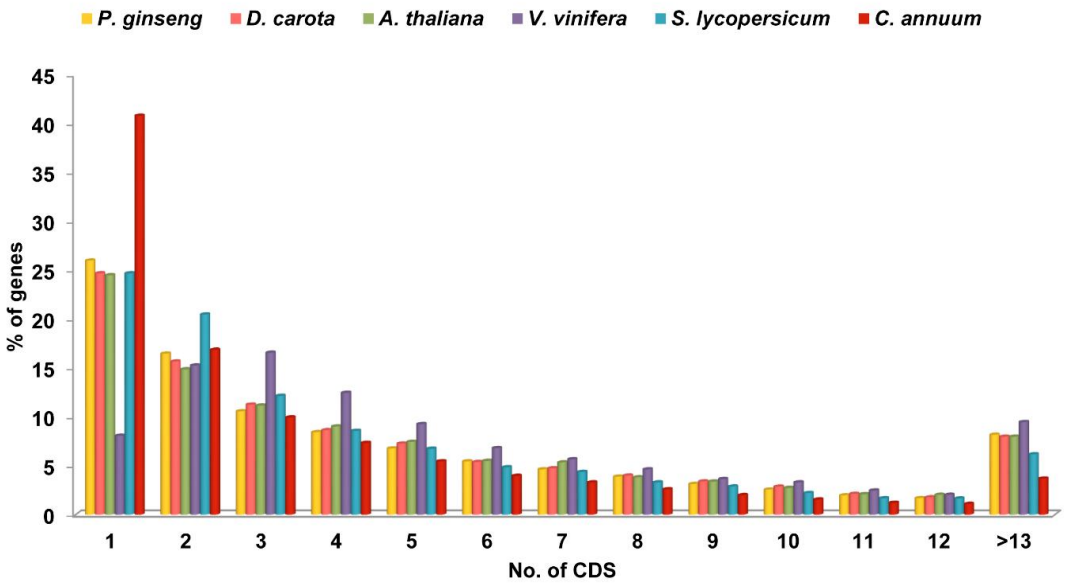

Figure 1-2. Number of coding exons (CDS) comparison between plant species. A Python script was used to calculate number of CDS for each gene in the above species based on genome annotation file (gff).

Table 1-1. Comparative gene metrics of *P. ginseng* gene models

Species	Protein coding genes	Total CDS length (bp)	Avg. CDS length (bp)	Avg. intron length (bp)	Max. gene length (bp)
P. ginseng	59,352	66,481,566	1,120	777	93,383
S. lycopersicum	34,725	35,971,085	1,035	541	244,621
D. carota	30,824	37,019,955	1,201	478	71,177
C. annuum	34,899	35,244,093	1,009	541	62,694
A. thaliana	27,206	33,015,750	1,212	164	71,498
V. vinifera	26,346	29,958,339	1,137	969	31,257

Alternative splicing (AS)

The alternative splicing transcripts for the final curated protein coding genes was identified using reference-based assembly by PacBio and Illumina sequencing data. Then, the reference-guided transcripts and annotated protein coding genes were compared to identify novel isoforms using cufflink utility. Further, those novel isoforms were used to find the specific splicing events (i.e. skipping exon, mutually exclusive exons, alternative 5' or 3' splice-site, retained intron and alternative first and last exon) using SUPPA [33]. Approximately, 38% (22,384) of the genes have alternative splicing forms wherein genes mostly contain one splicing forms (**Figure 1-3**) and retained intron types (**Figure 1-4**). In addition, GO enrichment analysis was performed for the genes containing AS.

Functional annotation of protein coding genes

Gene functions were assigned using various annotation methods including INTERPRO, Gene Ontology (GO), KEGG and homology search. Overall,

the functional description was assigned to a total of 97% of the *P. ginseng* genes (**Table 1-2**). This reflects the accuracy of our gene prediction strategy.

INTERPROSCAN

Motifs and domains of genes were identified using InterProScan [34, 35] version 5.13. InterPro combines various protein resources including TIGRFAM, ProDOM, Hamap, SMART,

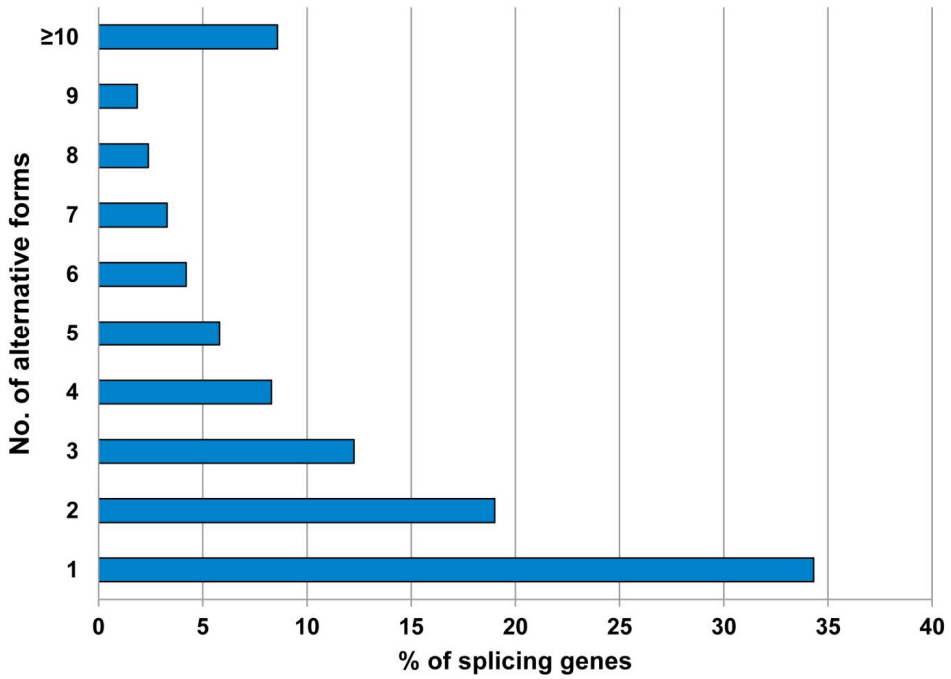

Figure 1-3. **Number of alternative splicing transcripts.** The x-axis represents percentage of genes from total number of genes containing isoforms and y-axis represents the number of splicing transcripts.

ProSiteProfiles, SUPERFAMILY, PRINTS, PANTHER, Gene3D, PIRSF, Pfam and Coils. Eighty four percentages of the genes have been assigned at least one protein domain. The top 10 Interpro domains are plotted in **figure**

1-5. In addition, this analysis was performed for other species including Arabidopsis, tomato, carrot, grapes and pepper for effective comparative analysis.

Gene Ontology (GO) annotation

Local BLASTX with an E-value cutoff of 1E-05 was performed against NCBI Nr protein database. This homology information was used for GO annotation using BLAST2GO [36]. This program categorizes the GO terms into molecular function, biological process and cellular component. Eighty two percentages of genes were assigned to at least one GO term. In total,

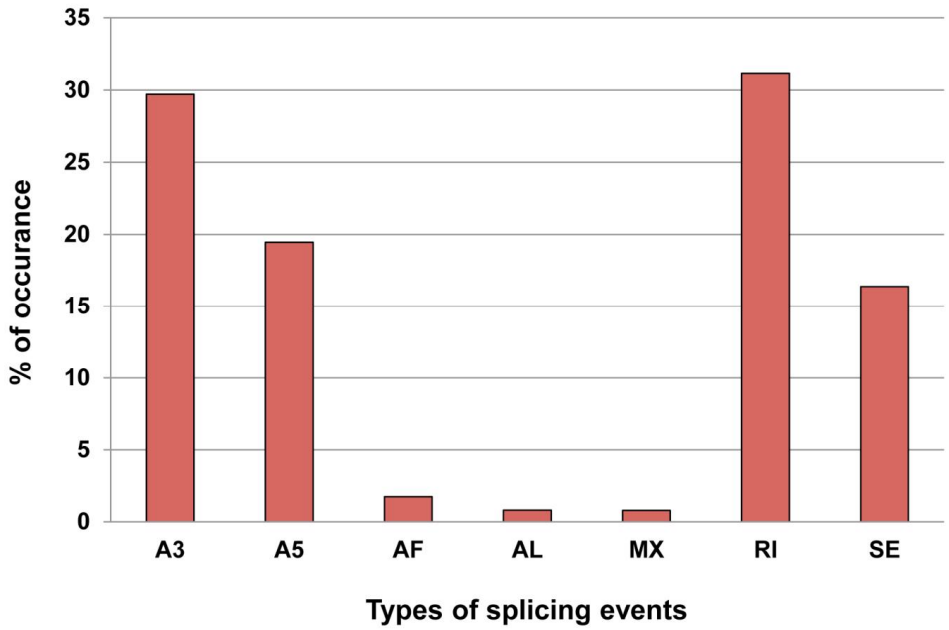

Figure 1-4. **Alternative splicing (AS) events in *P. ginseng*.** The x-axis shows the types of splicing events including alternative 3' splice-site (A3), alternative 5' splice-site (A5), alternative first exon (AF), alternative last exon (AL), mutually exclusive exons (MX), retained intron (RI) and

skipping exon (SE) and y-axis represents the percentage of occurrence of corresponding AS events in *P. ginseng*.

44,553(75%), 38,466 (64%), 46,670 (78%) genes have at least one GO tem in the categories of biological process, molecular function and cellular component respectively. The top GO terms in each category are plotted in **figure 1-6**.

Homology search

Gene functions were also assigned according to the best hit alignment using BLASTP with E-value 1E-05 to tomato, arabidopsis and NCBI Nr protein databases. The high percentages of hits were found using NCBI Nr (95% of genes) followed by tomato (92% of genes) followed by arabidopsis (89% of genes) (**Table 1-2**).

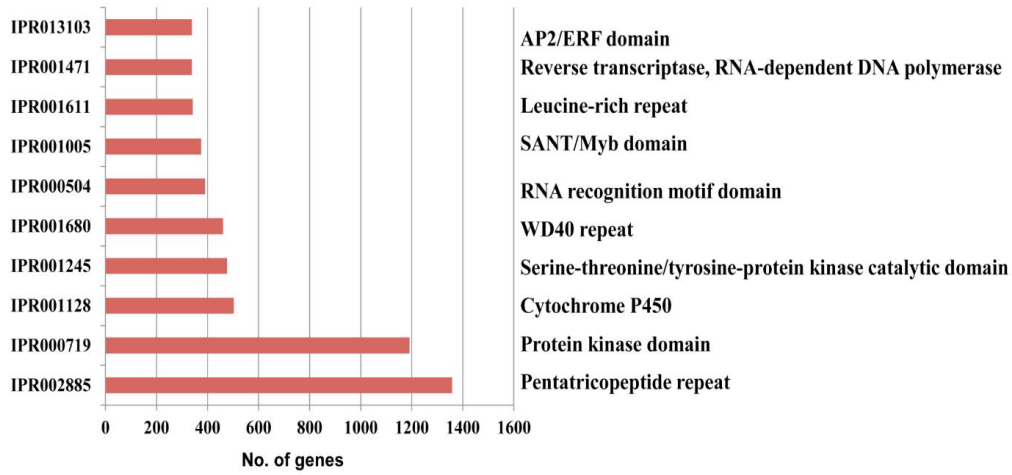

Figure 1-5. Top 10 INTERPRO domains in the IPGA gene set version 1.1.

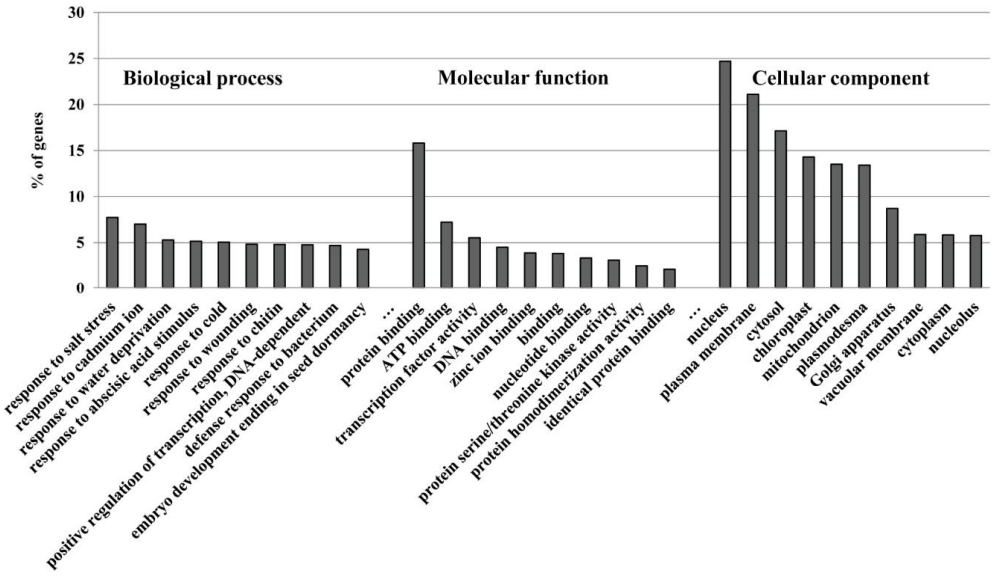

Figure 1-6. Gene Ontology (GO) annotation. Top GO terms were shown in this figure. GO terms were categorized into biological process, molecular function and cellular component.

KEGG annotation

Genes involved in metabolic pathways were identified using KAAS [37] (KEGG automatic Annotation Server). To run KAAS job, the reference gene set was selected from *Arabidopsis thaliana*, *Citrus sinensis*, *Glycine max*, *Vitis vinifera*, *Solanum lycopersicum* manually. BBH (bi-directional best hit) method was used to assign KEGG orthologs to *P. ginseng* gene set. A total of twenty-one percentage of genes was assigned to 337 metabolic pathways.

Table 1-2. Functional annotations of protein coding genes.

Database	Program	% of annotations
Tomato proteins	BLASTP	92
Arabidopsis proteins	BLASTP	89
NCBI NR protein	BLASTP	95
Gene Ontology (GO)	BLAST2GO	82
KEGG	KAAS	21
Protein domains	INTERPRO	84
Total		97

Transcription factor, transcriptional regulator, and protein kinase

Transcription factors (TFs), transcriptional regulator (TRs), and protein kinase (PK) play important roles in many biological processes for plant developments and environmental responses. TF, TR, and PK genes of the *P. ginseng* genome were identified and classified based on conserved domain structure analyzed by iTAK 1.6b standalone [38] with default parameters. In the *P. ginseng* genome, 3,588 TF genes in 66 families and 851 TR genes in 25 families were identified, which accounted for 6.05% and 1.43% of total annotated genes (59,352), respectively.

Small RNA (sRNA) annotation

Previously, the small RNA sequencing data was generated from six-year old flower buds, leaves and lateral roots[39]. The raw data from those libraries were pooled which contain large number of sRNAs with range of 18-20 bp. These reads were filtered using UEA tool-kit[40] to remove adapters, low complexity sequences, rRNA/tRNA sequences and reads with outside the range of 18-26 bp. In addition, a non-redundant set (17,737) was made from

genome aligned sRNAs for subsequent analysis. Then, de novo miRNAs were predicted using mireap v0.2 (https://sourceforge.net/projects/mireap/).

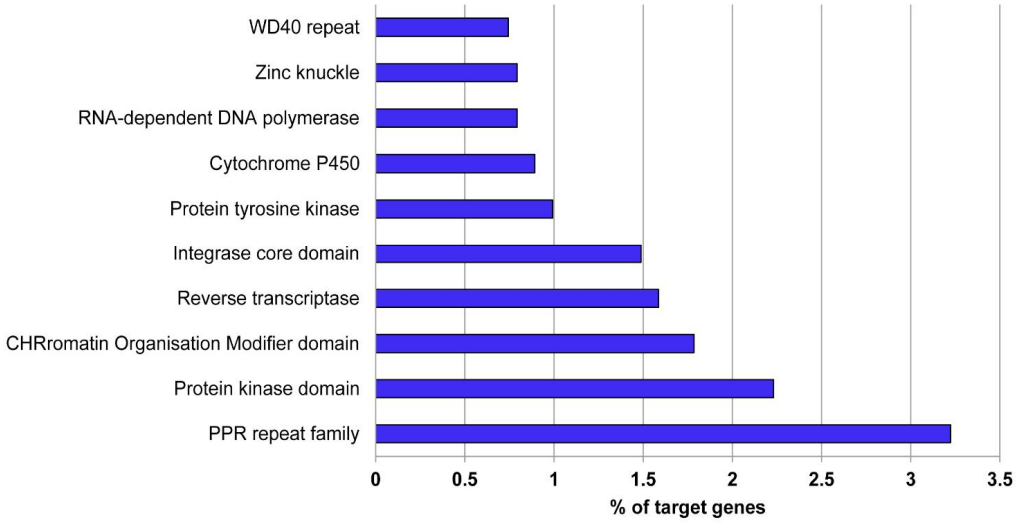

Figure 1-7. **Protein domain (Pfam) based grouping of genes targeted by miRNA.**

A total of 865 miRNAs were found from 817 genomic loci and most of them are in the size of 20 (nt). The predicted miRNAs with match to miRBase (v21) (http://www.mirbase.org/) are referred to as conserved miRNAs (451 miRNAs). The target prediction was performed for conserved miRNAs using psRNATarget [41]. Further, the target genes were grouped according to their protein domain (**Figure 1-7**). Intriguingly, many members of PPR repeat gene family was targeted by miRNAs which is similar to arabidopsis and populous [42, 43]. Evolutionary, the PPR gene family has expanded dramatically in plants [44, 45]. A posttranscriptional silencing may control rapid expansion of gene families as to minimize the detrimental dosage effects [46, 47]. Notably, the PPR gene family also expanded in *P. ginseng*

and it is speculated that many of those PPR genes might be regulated by miRNAs. Other than PPR, cytochrome P450 family was also targeted by miRNAs like other plant species [48, 49].

Ginsenoside biosynthesis

Ginsenosides, the major pharmacologically active compounds of ginseng, are triterpene saponins, of which more than 150 have been isolated from *Panax* plants. Ginsenosides are biosynthesized through cyclization, hydroxylation and glycosylation of 2,3-oxidosqualene that is synthesized via mevalonate (MVA) and 2-C-methyl-D-erythritol-4-phosphate (MEP) pathways. In most plants, 2,3-oxidosqualene is subsequently cyclized into cycloartenol, α-, β-amyrin, or lupeol, to be further converted to phytosterols and pentacyclic triterpenoids [50]. In *P. ginseng*, an additional cyclic compound, dammaranediol, can be biosynthesized by a specific cyclase then oxidized through a set of cytochrome P450 enzymes to form the major dammarane-type sapogenins [protopanaxadiol (PPD)/protopanaxatriol (PPT)], while the minor oleanane-type algycone (oleanolic acid) is biosynthesized from β-amyrin. These precursors are further glycosylated via several UDP-glycosyltransferases (UGTs) to synthesize various types of ginsenosides (**Figure 1-8).**

The genes encoding enzymes involved in this pathway were identified using KEGG annotation as well as BLASTP search against the datasets of KEGG as well as MetaCyc. Twelve squalene epoxidase (SQE) genes were identified in *P. ginseng*, twice as many as in other plants (**Table 1-3**), suggesting increased ginsenoside precursor production. Twenty *P. ginseng* oxidosqualene cyclase (OSC) genes were found in the biosynthesis of dammarane-/oleanane-type ginsenosides [dammarenediol synthase (DDS), β-amyrin synthase (β-AS)] and sterols (lanosterol synthase (LSS),

cycloartenol synthase (CAS)]. Phylogenetic analysis of OSC families found DDS to be specific to *P. ginseng* (**Figure 1-9**), suggesting that DDS and production of dammarane type ginsenosides originated in *Panax*.

Table 1-3. Downstream genes involved in ginsenosides biosynthesis comparison with relative plant species.

Genes	***P. ginseng***	***D. carota***	***S. lycopersicum***	***V. vinifera***	***A. thaliana***
FPPS	2	1	2	1	2
SQS	4	1	3	1	2
SQE	12	5	2	5	6
DDS	4	0	0	0	0
β-AS	8	1	3	3	1
CAS	6	0	1	6	1
LSS	2	0	0	0	1

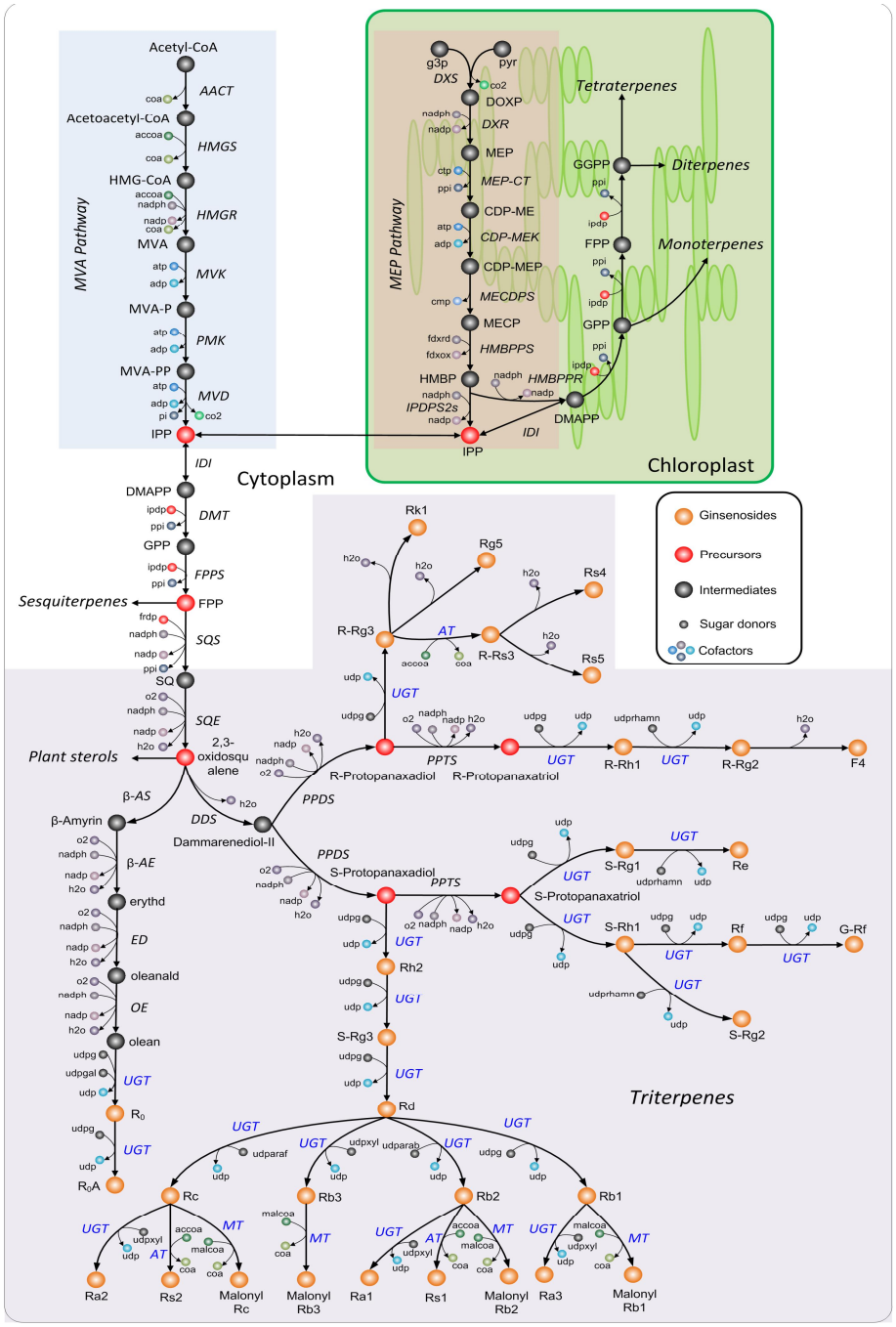

Figure 1-8. Overview of the ginsenoside biosynthetic pathway in *P. ginseng*. The blue colored uridine 5'-diphospho-glucuronosyltransferases

(UDP-glucuronosyltransferase, UGTs) are an unknown enzymes involved in the glycosylation of ginsenosides.

Of 383 *P. ginseng* cytochrome P450 genes, two candidate protopanaxadiol synthase (PPDS) and two protopanaxatriol synthase (PPTS) genes were identified by homology search against curated PPDS [51] and PPTS, respectively. Plant secondary product glycosyltransferase (PSPG) motif containing UGTs are reported to be involved in the glycosylation of plant secondary metabolites [52, 53]. Therefore, the UGTs were retrieved from the InterPro annotation using the ProSitePatterns of PS00375 and IPR002213. In the last glycosylation step, 226 UGTs were annotated and eleven identified as candidate UGTs associated with elevated expression pattern upon methyl jasmonic acid (MeJA) treatment (**Figure 1-10**), which is well-known elicitor for inducing secondary metabolites [51, 54]. These candidate UGTs could be involved in synthesis of PPD-type ginsenosides, as MeJA triggers mainly PPD-type [10].

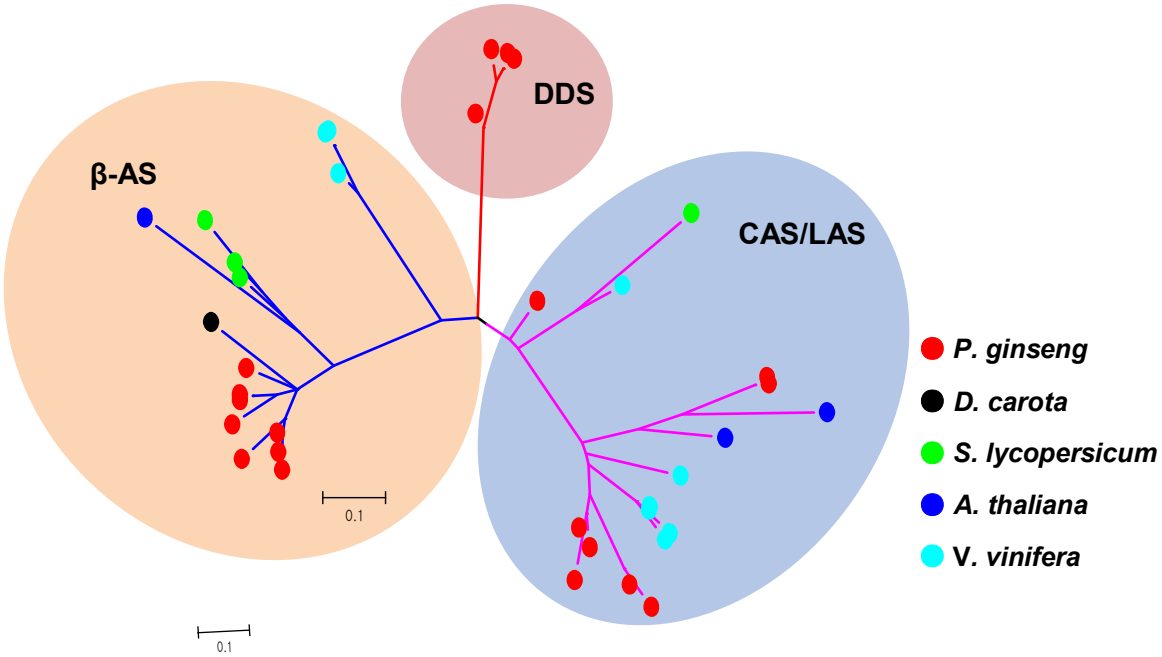

Figure 1-9. A phylogenetic analysis of OSC gene families.

The high ginsenoside contents for which older (above 4~6 years) *P. ginseng* roots are harvested might reflect transportation from shoot tissues rather than active biosynthesis. Downstream genes (SQE, DDS, PPDS and PPTS) in the ginsenoside biosynthetic pathway showed higher expression in leaves (one- and five-year old) than roots (one- and six-year old main body roots, lateral roots and rhizomes) (**Figure 1-11**). It is also expected that genes in the same biosynthetic pathway are usually co-expressed. Hence, in order to identify the key gene that limit the ginsenoside accumulation, a co-expression analysis was performed using Pearson correlation coefficients (PCC) across all RNA-seq data of ChP. This result showed that three highly expressed DDS genes among 20 OSC are co-regulated with several SQE

genes, and disrupting function of either DDS or SQE affects production of ginsenosides [55, 56]. This implies that DDS and SQE may be rate limiting enzymes and that ginsenoside production co-evolved with these key enzymes. Further, I identified bHLH and Tify transcription factors associated with regulation of ginsenoside accumulation**.** While many CPY450 and UGTs are not yet characterized with respect to different types of ginsenosides (**Figure 1-8**), dynamic changes in expression of various genes providing a foundation for *in silico* analysis and ultimately empirical metabolic engineering.

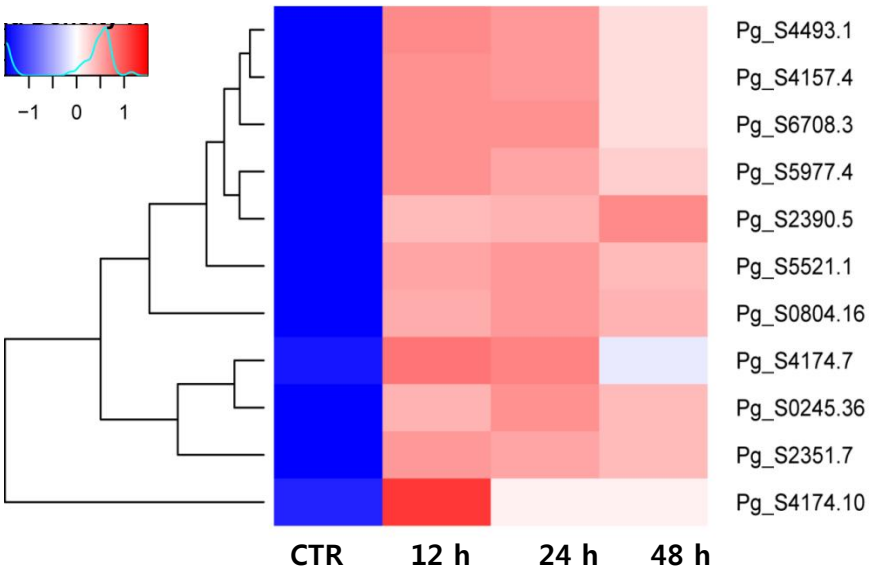

Figure 1-10. DEG analysis using MeJA treated RNA-seq data from cv. CS.

Gene expression analysis

High-quality filtered RNA-seq reads from each sample were aligned to the genes (CDS) to estimate the abundance or digital gene expression using RSEM [57] (v1.2.4) with the parameter of minimum and maximum fragment

length of 200 and 300 respectively. RSEM calculates the number of paired reads or fragments mapped to genes as FPKM (Fragments Per Kilobase per Million). Eighty-six percentage of genes acquired at least ≥1 FPKM value in any one of the above RNA-seq samples. For visualization of gene expression in heatmap, normalized FPKM values obtained by Trimmed Mean of M-values (TMM) [58] was used. Heatmap was generated using heatmap.2 function provided by the R-package.

Table 1-4. Enriched GO biological terms among DE genes in response to abiotic stress

GO-ID	Term	Category	FDR
GO:0006636	unsaturated fatty acid biosynthetic process	P	1.05E-40
GO:0019344	cysteine biosynthetic process	P	4.79E-38
GO:0010114	response to red light	P	2.55E-36
GO:0010207	photosystem II assembly	P	3.04E-34
GO:0009637	response to blue light	P	1.48E-33
GO:0010218	response to far red light	P	2.67E-32
GO:0010155	regulation of proton transport	P	1.97E-28
GO:0009768	photosynthesis, light harvesting in photosystem I	P	4.39E-27
GO:0009744	response to sucrose stimulus	P	1.90E-26
GO:0009773	photosynthetic electron transport in photosystem I	P	2.35E-26
GO:0010196	nonphotochemical quenching	P	9.74E-26
GO:0006098	pentose-phosphate shunt	P	5.95E-24
GO:0009750	response to fructose stimulus	P	8.86E-24
GO:0042742	defense response to bacterium	P	1.10E-20
GO:0006364	rRNA processing	P	1.44E-20
GO:0000023	maltose metabolic process	P	1.68E-19
GO:0080167	response to karrikin	P	1.13E-18
GO:0009409	response to cold	P	1.25E-18
GO:0015995	chlorophyll biosynthetic process	P	5.56E-18
GO:0035304	regulation of protein dephosphorylation	P	6.17E-17

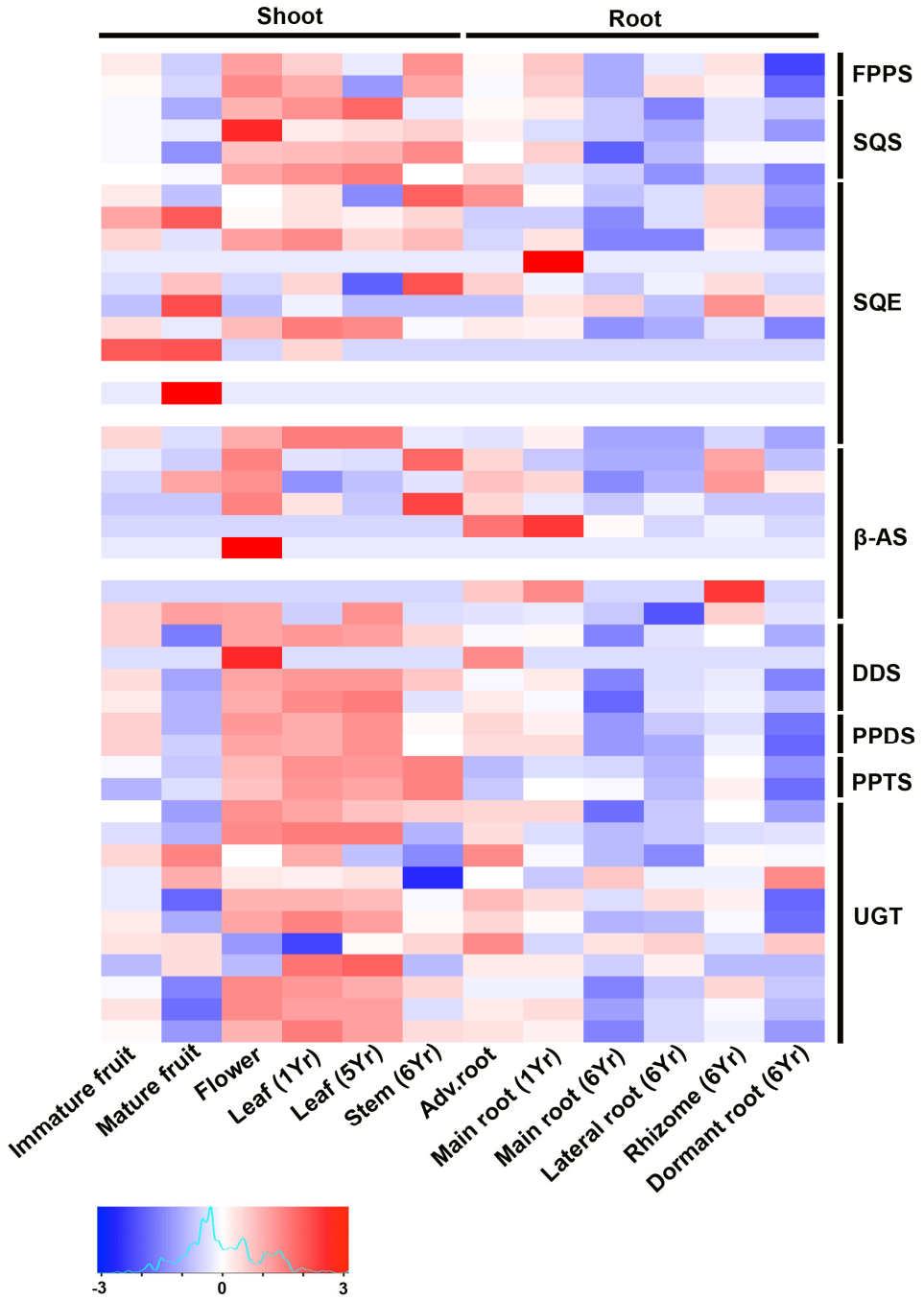

Figure 1-11. Heatmap shows TMM normalized expression values of putative downstream genes involved in ginsenosides biosynthesis.

Differential expressed gene (DEG) analysis

Differentially expressed (DE) genes were identified with two/three biological replicates stress treated samples using the bioconductor package edgeR [59] with ≥ 2 folds and false discovery rate (FDR) adjusted *p*-value of 0.01. The sequencing library differences (due to different Illumina platform) between replicated RNA-seq samples were normalized using Trimmed Mean of M values (TMM) [58]. This condition was applied each sets of condition such as control and drought, control and salt, control and cold RNA-seq data separately. In detail, 703, 152 and 23 genes were shown different expression in response to drought, cold and salt respectively (**Figure 1-12**). Similarly, DEG analysis was also performed between non heat-treated leaves and heat treated (1week and 3 week) leaves of three replicates. In total, 1,409 genes were identified as DEG after 1 and 3 weeks of heat treatment (**Figure 1-12**). Altogether, 1,880 genes were found to be differentially expressed (DE) and the numbers of DE genes including up- or down-regulated genes are represented in **figure 1-12**. In addition, the DE genes were grouped based on their Pfam domain in order to identify gene families that are influenced by environmental stress in *P. ginseng*. Majorly, genes for fatty acid desaturase, cytochrome P450, chlorophyll-ab-binding were shown over two-fold expression in response to abiotic stresses (**Figure 1-13**). Similarly, the DE transcription factors family was also plotted **Figure 1-14**. Genes of WRKY transcriptions factors found to be largely influenced by abiotic stress among transcription factors in *P. ginseng*.

GO enrichment of the target DE genes

GO enrichment analysis was performed for the total DE genes (1,880) using Fisher's Exact Test with multiple testing correction of FRD with cutoff 0.05. The result indicated that light and temperature responsive genes were

significantly regulated (**Table 1-4**). In addition, pentose-phosphate shunt, defense response to bacterium, maltose metabolic process and cold responsive genes were also enriched in the list of significant differential expressed genes.

Gene family analysis

To assess the repertoire of biological and metabolic process in *P. ginseng*, the genes were further grouped into families in according to the Pfam and InterPro domain signatures annotated by InterproScan. The over- and under-represented Pfam domains were identified using in-house

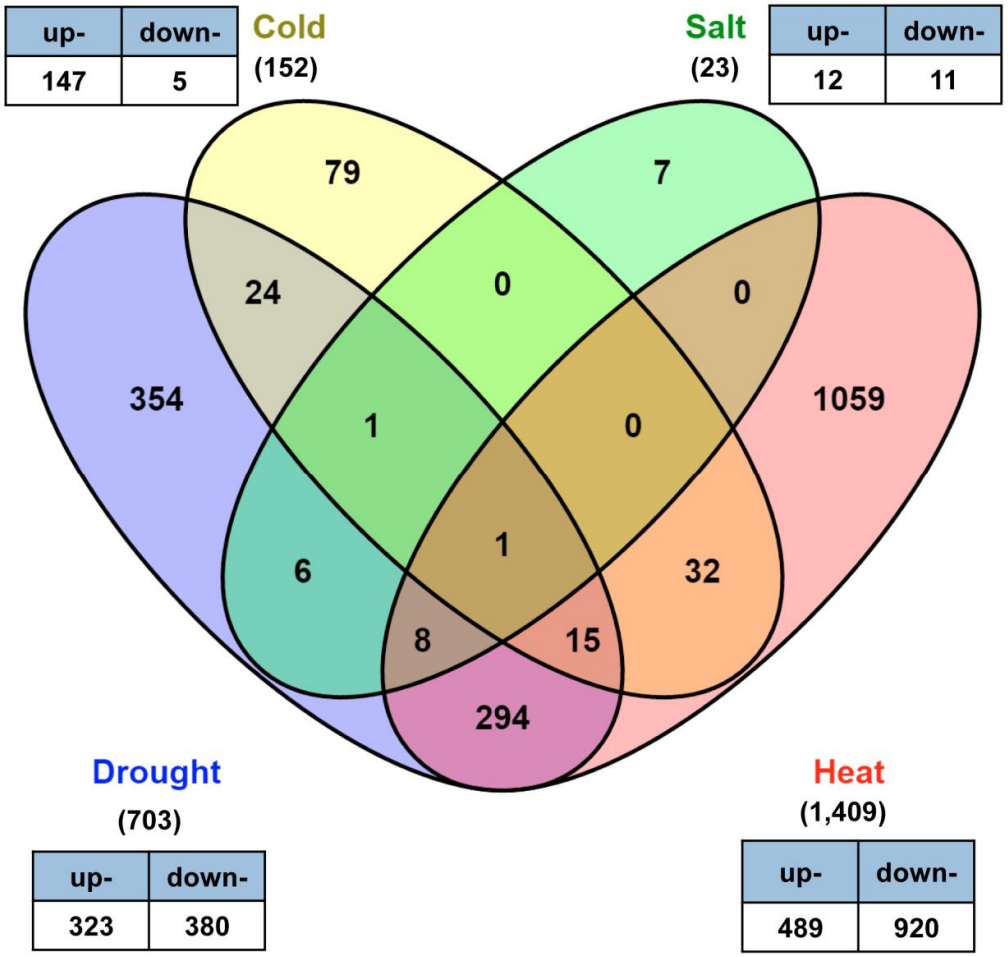

Figure 1-12. **The number of differentially expressed (DE) genes among drought, salt, cold and stress samples.** A Venn diagram intersects the number of DE genes among abiotic stress along the number of up- and down-regulated genes.

program using Bonferroni correction for multiple testing. Remarkably, a highly abundant proteins domain was identified in *P. ginseng* as compared to other plant genomes. These included domains that are predicted to be

involved in fatty acid synthesis (Fatty acid desaturase), photosynthesis (Chlorophyll a-b binding protein).

Fatty acid desaturase (FAD)

FADs are enzymes that synthesize mono- and poly unsaturated fatty acids such as oleic acid, linoleic acid. FADs in *P. ginseng* were identified using the Pfam domains including PF00487,

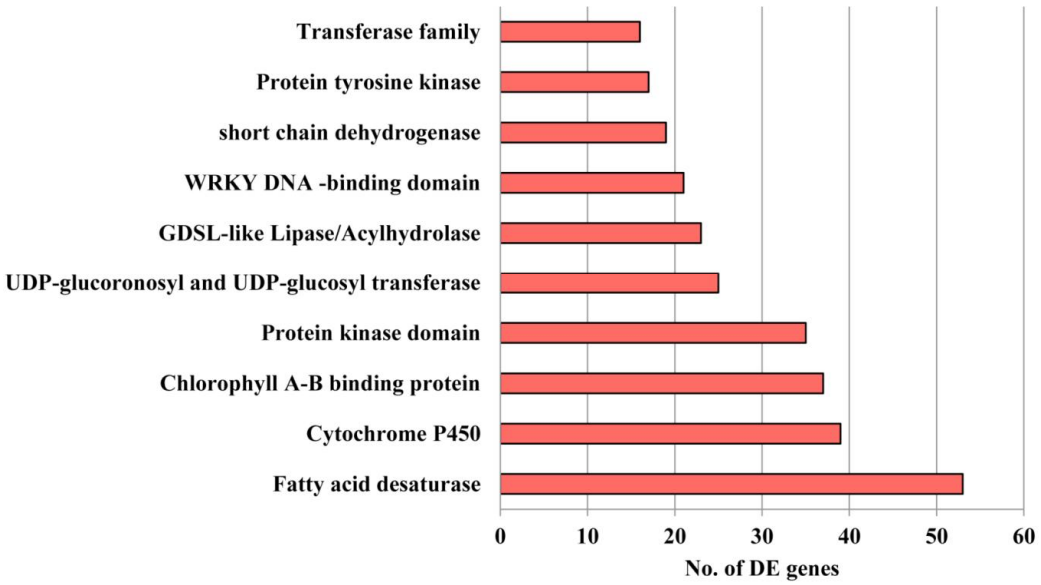

Figure 1-13. **Number of differentially expressed (DE) genes for the major gene families.** The total DE genes were grouped based on their Pfam domain annotated by InterPro scan and plotted.

Table 1-5. The number of members in FAD gene family in plant genomes.

Species	No. of genes	Genome size
P. ginseng	85	~3.6 Gbp
A. thaliana	17	125 Mbp
S. lycopersicum	12	760 Mbp
O. sativa	10	206 Mbp
P. trichocarpa	20	~500 Mbp
D. carota	63	~480 Mbp
P. notoginseng	55	~2110 Mb

PF11960 and PF03405 annotated by InterproScan. Intriguingly, a large number of (85) FAD gene was found in *P. ginseng*, which is three times higher than model annual plants of *A. thaliana*, *S. lycopersicum*, *O. sativa* (**Table 1-15**). Notably, the many members of this gene family have arisen through tandem duplication. For comparative analysis, the FAD genes were collected from *P. notoginseng*, *A. thaliana* (TAIR10), *S. lycopersicum* (ITAG2.4), *O. sativa* (v7.0), *Populus trichocarpa*, and *D. carota*. Besides *P. ginseng*, the number of FADs was almost two time higher in biennial *D. carota* [60, 61] as compared to other annual plants. This indicated that over-wintering crop plants (i.e. biennial and perennial) might contain more FADs than annual plants. The digital gene expression of FADs was investigated between abiotic stress RNA-seq samples in *P. ginseng*. Overall, a high expression pattern was observed under cold stress as compared other stresses. In the list of DE genes, a total of fifty-three genes in FAD family showed significant differential expression pattern in response to drought, cold and heat stress conditions. Among them, forty-five genes under heat, three genes upon drought, two under salt stress were identified.

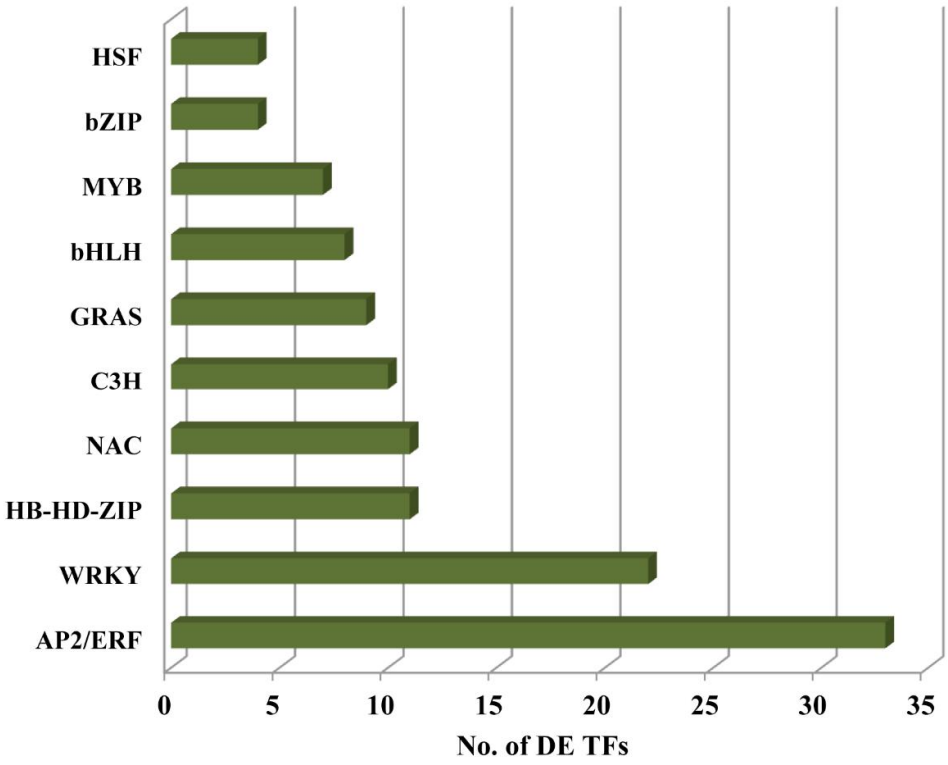

Figure 1-14. **Number of differentially expressed (DE) transcription factors (TF) between abiotic stresses including drought, salt, cold and heat.**

Further, a phylogenetic analysis was performed between *P. ginseng*, *P. notoginseng*, *A. thaliana*, *S. lycopersicum*, *O. sativa*, *P. trichocarpa* and *D. carota*. Notably, two separate groups specific to *P. ginseng* was identified (**Figure 1-15**). Among them, a separate group containing twenty-six FAD genes was related to putative acetylene and a group of nine genes was identified as newly evolved delta-12-FADS-like genes. Acetylenes are diverged forms of fatty acids, which are derived from further desaturation of fatty acid molecules. These compounds are found in restricted taxa including[62] *Asteraceae* (sunflower), *Apiaceae* (carrot) and *Araliacea*

(*Hedera helix, P. ginseng*). Among sub groups of FAD genes in *P. ginseng*, the acetylene related FADs showed high expression pattern under cold and salt stress as comparted to normal conditions (**Figure 1-16**). Similarly, delta-12-FADS-like members also showed higher expression under cold condition. In addition, specific clades were found for FAD2, FAD5 and FAD family protein genes. This suggested that the diversification of FAD structures was happened in the evolution of *P. ginseng*. Furthermore, gene presence/absence analysis showed specific expansion in *P. ginseng* when compared to *P. vietnamensis* with strong expression against cold stress (**Figure 1-16**).

Cold acclimation is a multigenic trait which includes biochemical and physiological changes such as changes in proline, sugars and inorganic acids and membraned lipid composition[63, 64]. A well-characterized phenomenon demonstrates that temperature modulates the membranes fluidity, which is major site of freezing injury [65-67]. It is also known that the role of FADs in cold acclimation in various plant species [68-71]. In addition, the divergent FAD genes have been associated with synthesis of divergent fatty acid structures that play major role against biotic/abiotic stresses [72]. Therefore, the expansion of FAD genes with diverse FAD structures are expected in biennial and perennial crop plants from temperate region and this expansion with diverse forms is necessary for those plants to tolerate chilling and to survive at low temperatures. As compared to diploid ginseng, the polyploid ginseng species such as *P. ginseng* and *P. quinquefolius* have been commonly found in the habitat of South Korea and North America respectively, where freezing temperature prevails in the winter. This suggested that polyploidization of ginseng species might have led chilling tolerance.

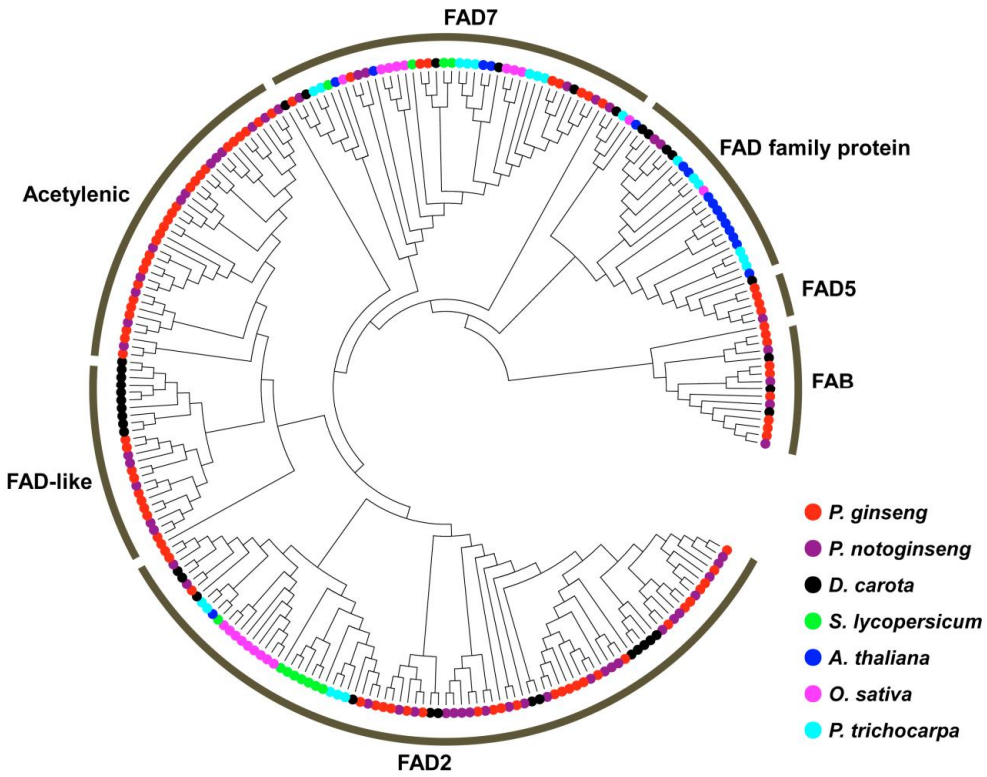

Figure 1-15. A phylogenetic relationship of FAD genes. The red, purple, black, blue, green, pink and cyan color indicates the corresponding genes in the FAD family from Korean ginseng, Chinese ginseng, carrot, Arabidopsis, tomato, rice and poplar. The outer circle indicates the grouping of subgroup FADs including acetylenic, FAD-like, FAD2, FAD5, FAD7, FAD family protein and FAB.

Chlorophyll ab binding (CAB)

Ginseng has been grown under canopy or artificial shade, however the reason behind this process is largely unexplored. It is obvious that the ginseng plant should have acquired a novel mechanism to ensure an efficient photosynthesis under low light conditions. The light-harvesting chlorophyll a-b binding proteins (LHCPs or CAB) are the key components of the

photosynthesis antennae complexes which transfer the light energy to the reaction centers of photosystem I (PS I) and photosystem II (PS II) where the light energy is converted to form chemical bond energy (i.e. NADPH and ATP). Intriguingly, *P. ginseng* genome contains more CAB genes (49 genes

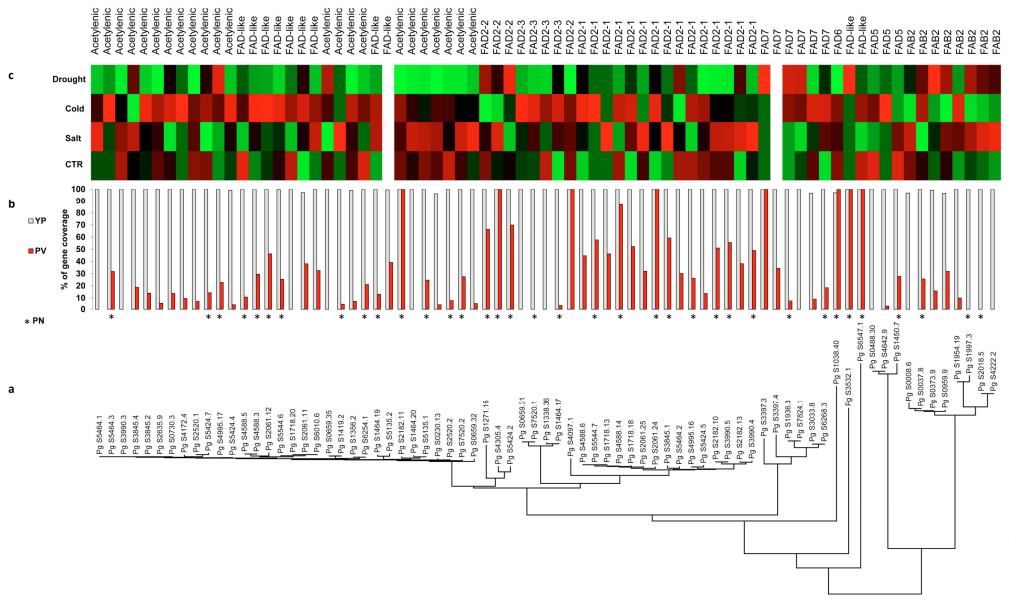

Figure 1-16. Classification and expression of FAD genes. (a) Phylogenetic analysis of FAD genes. **(b)** Mapping coverage for coding (CDS) genes using 10x coverage WGS reads from tetraploid *P. ginseng* cv. YuP (white bars) and diploid *P. vietnamensis* (red bars). Orthologous FADs in diploid *P. notoginseng* (PN) denoted as * under the bar graphs. **(c)** TMM normalized expression of each gene under control (CTR), drought, salt and cold stress conditions. The FAD subclass is represented at the top to show expansion of specific sub classes and its role in abiotic stress responses.

identified by Pfam domain PF00504) than any plant species to date (**Table 1-6**). Equivalently, a total of 53 genes (including pseudogenes) were also

identified in the genome of brown algae [73] (*Ectocarpus siliculosus*) and that expansion was attributed to adapt to variable or dim light conditions. This suggested that the expansion of CAB genes might be associated with low light adaptations. Similarly, it is speculated that the large number of CAB genes in *P. ginseng* might enable the ginseng plant to adapt to an environment with low light conditions.

It is also reported that the members in CAB gene family play major roles in plant adaptation to various environmental stresses [74-76]. Approximately, seventy percentages of the members in CAB family (37 genes) were identified in the list of DE genes. From the total forty-nine CAB genes, twenty-six genes showed altered expression during drought and heat, two genes during drought, heat and salt condition, and eight genes specific to heat stress. Overall, one CAB gene (Pg_S6185.8) was found to be regulated in all the above abiotic stress conditions. Further, a phylogenetic analysis was performed for CAB genes between *P. ginseng, P. notoginseng*, *A.*

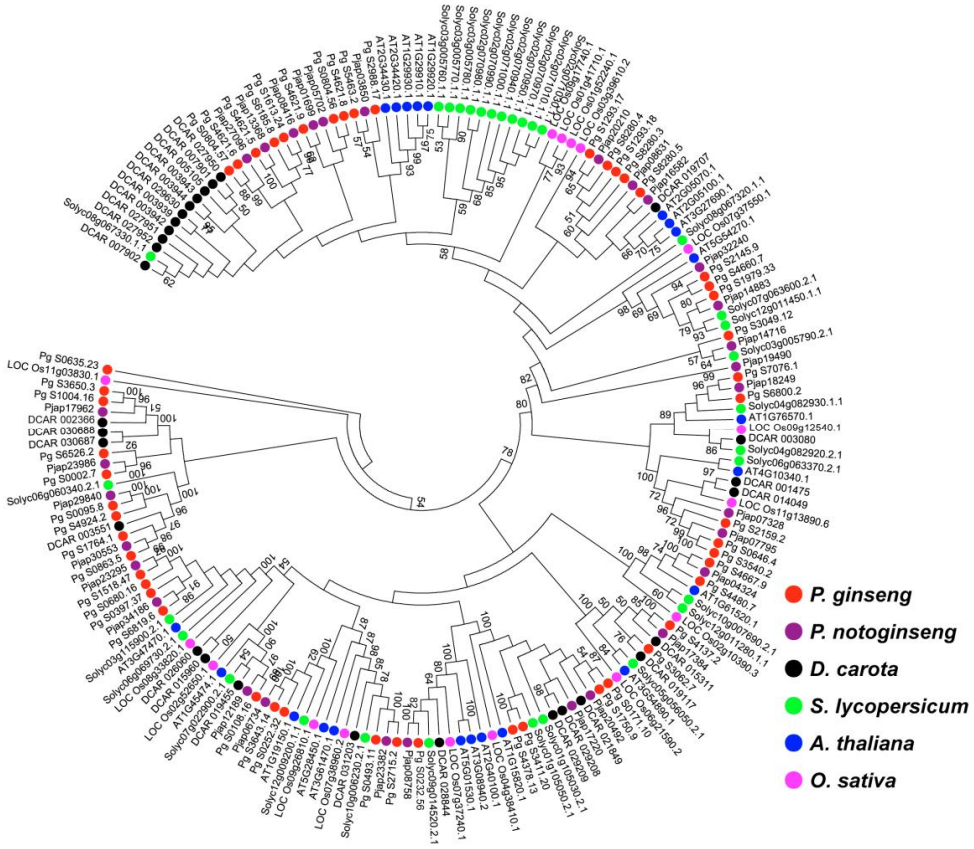

Figure 1-17. **A phylogenetic relationship of CAB family genes.** The red, purple, black, blue, green and pink color indicates the corresponding genes in the CAB family from Korean ginseng, Chinese ginseng, carrot, arabidopsis, tomato and rice respectively.

Table 1-6. The number of genes in CAB family in eukaryote genomes.

Species	No. of genes	Genome size
P. ginseng	49	~3.6 Gbp
A. thaliana	22	125 Mbp
O. sativa	15	206 Mbp
S. lycopersicum	31	760 Mbp
D. carota	32	~480 Mbp
E. siliculosus	53	214 Mbp
P. notoginseng	30	~2110 Mb

thaliana, *S. lycopersicum*, *O. sativa* and *D. carota*. From phylogenetic analysis, it was observed that a member of CAB gene in *P. ginseng* have 4:1 (four copies in *P. ginseng*: one copy in other species), and 4:2 and 4:3 relationship with other species (**Figure 1-17**). Intriguingly, estimation of presence/absence of orthologous gene copies in *P. vietnamensis* revealed the abundance of CAB genes both shade plants, tetraploid and diploid ginseng species (**Figure 1-18**).

Conclusion

The genome sequence clarifies the evolution, shade adaptation, and medicinal properties of *P. ginseng*. Two Araliaceae-specific WGDs played key roles in environmental (shade and freezing) adaptation and medicinal importance (dammarane type ginsenoside production), the former also providing information that might apply to improvement of other cultigens. The widespread importance of collecting and cataloguing crop relatives is especially urgent in *Panax*, in which extant diploid relatives are at risk of extinction from global warming, progenitors of cultivated tetraploids are already extinct, and wild tetraploids are endangered by over-harvesting [1, 2].

Materials and methods

De novo sequencing, assembly and quality evaluation

DNA from leaves of 4-year old ChP, an elite Korean cultivar, was used for sequencing and assembly. The ChP was cultivated in a ginseng experimental field of research farm (College of Agriculture and Life Science, Seoul National University, Suwon, South Korea) and used for

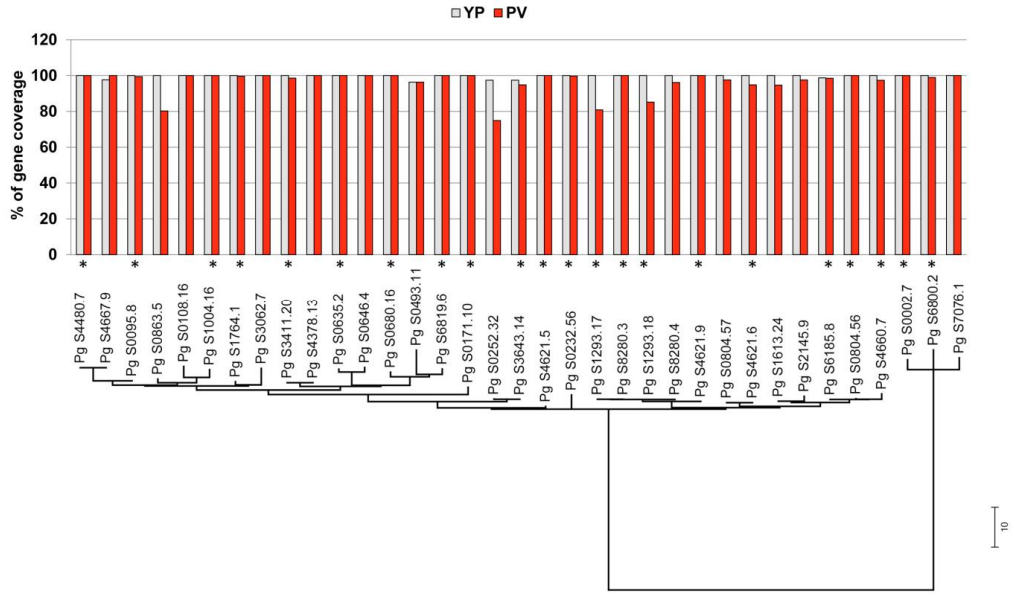

Figure 1-18. Classification and estimation of CAB orthologs gene copies. The figures shows two layers of sub figures wherein the bottom figure represents the phylogenetic tree of CAB orthologs and the middle shows their corresponding coding (CDS) gene coverage using 10x WGS reads from *P. ginseng* cv. YuP (white bar) and diploid *Panax* species, *P. vietnamensis* (red bar). In addition, orthologous CAB in *P. notoginseng* (PN) which is also a diploid was denoted as * under the bar graph.

isolation of genomic DNA and total RNA. To reduce heterogeneity, we used DNA from three individuals. Whole genome shotgun reads of ChP were generated using Illumina platform (HiSeq2000 and MiSeq) at National Instrumentation Center for Environmental Management (NICEM), Macrogen Co. (Seoul, Korea), and LabGenomics Co. (Pankyo, Korea). The five paired-end (PE) libraries (with 200 to 600 bp insert sizes) were sequenced into 746 Gbp for primary assembly and the 365 Gbp was sequenced from four mate pair libraries with 1.5 kb, 3 kb, 5 kb, and 10 kb insert for scaffolding. First, low quality reads and duplicated reads were eliminated using SOAPfilter 2.0, an utility in SOAPdenovo package [77] with default parameter. Furthermore, low frequency reads were eliminated based on *k*-mer frequency by SOAPec 2.0 with KmerFreq_HA 2.0 and Corrector_HA 2.0, which cannot support for initial contig assembly. Genome size estimation was conducted by flow cytometry and 23 bp *k*-mer frequency analysis with JELLYFISH [78]. Taken together, the genome size of *P. ginseng* was estimated to range between 3.3 and 3.6. The *k*-mer frequency-based genome size, 3.6 Gbp, was used for further analysis and discussion for the genome composition. The genome assembly was mainly conducted using SOAPdenovo2. The contigs containing length over 1 kb and filtered mate-pair reads were used for scaffolding with SSPACE followed by error correction by in-house Perl scripts (Phyzen, Korea).

Transcriptome sequencing and analysis

Total RNAs from each sample were extracted using RNeasy Plant kits (QIAGEN, Germany) and/or Hybrid-R kits (GeneAll, Korea) according to the manufacturer's instructions, and used for construction of 300-bp PE libraries using an Illumina TruSeq RNA sample preparation kit according to the manufacturer's instructions. These libraries were pooled and sequenced

by Illumina HiSeq2000 and NextSeq500 platforms. The resulting RNA-Seq reads were mapped to the *P. ginseng* draft genome and assembled using HISAT [19] and StringTie [20] respectively. *De novo* assembly was performed using Trinity [22] to obtain full-length transcripts. All RNA-Seq samples were normalized using Trimmed Mean of M values [79] (TMM). Analysis of differential gene expression was performed using edgeR [59] with false discovery rate (FDR) adjusted *p*-value of 0.01. Transcriptomes of 22 ChP samples including normal tissues and abiotic stress-treated samples were also analyzed using 26 SMRT cells with P6-C4 chemistry of the PacBio RSII platform. Generated sequences were classified and clustered by the PacBio Iso-Seq analysis procedure (ver. 0.1) with default parameters (www.pacb.com) to generate high-quality (HQ) consensus isoform sequences (99% consensus accuracy based on Quiver). The HQ sequences were further processed to remove PCR chimeras and redundant sequences by cd-hit-est (http://weizhongli-lab.org/cd-hit/) and final HQ non-redundant (nr) isoform sequences were obtained based on genome positional coordinates.

Genome annotation

The IPGA pipeline was used for genome annotation, incorporating evidence from protein and RNA-Seq mapping and *ab initio* gene prediction to determine consensus gene models by EVM [18], that were curated using PacBio transcript sequences. The alternative splicing transcripts for the final curated protein coding genes were identified using reference-based assembly generated by PacBio and Illumina sequencing data. Then, the reference-guided transcripts and annotated protein coding genes were compared to identify novel isoforms using cufflink utility. Further, those novel isoforms were used to find the specific splicing events (i.e. skipping exon, mutually exclusive exons, alternative 5’ or 3’ splice-site, retained intron and

alternative first and last exon) using SUPPA[33]. LncRNAs were identified from the reference-guided transcriptome assembly. From the total transcripts, transcripts with ORF ≥ 100 amino acids and length ≤200 nucleotides, having homology hit to the swiss-prot protein database, Pfam domains, and other type of noncoding RNAs (tRNA, rRNA, snRNA, snoRNA) were discarded. Further, transcripts that span over 40% to repeat masked genomic region and contained partially at protein coding genes (IPGA v1.1) were discarded. Finally, the coding potential was accessed for the remaining transcripts using CPC with score ≤-1.0 and CPAT with score <0.39. A total of 19,495 lncRNA transcripts were identified using the above criteria. TF genes in the *P. ginseng* genome were identified and compared with corresponding genes in other plant genomes by using iTAK 1.6b standalone [38] with default parameters. A *P. ginseng* small RNA library generated by Mathiyalagan *et al*., [39] was used for conserved miRNA prediction using mireap v0.2 (https://sourceforge.net/projects/mireap/). The predicted miRNAs with match to miRBase (v21) (http://www.mirbase.org/) were referred to as conserved miRNAs. The target prediction was performed for conserved miRNAs using psRNATarget [41]. Functional descriptions were assigned to annotated genes using BLASTP search (*E*-value: 1E-05) to NCBI Nr, Arabidopsis, tomato and INTERPRO[40] protein databases. GO enrichment analysis was performed using Fisher's Exact Test with multiple testing correction of FDR with cutoff 0.05. The *P. ginseng* repeat library was constructed from eight reported transposable element and consensus repeats characterized with pre-identified LTR-RTs [16, 17] and RepeatModeler [80], and genome-wide repeat content of the assembled genome was calculated with RepeatMasker [81].

Identification of genes in ginsenoside biosynthetic pathway

Genes involved in ginsenoside biosynthesis were identified using KEGG annotation and BLASTP against reference enzyme genes retrieved from KEGG and MetaCyc databases (http://metacyc.org/) with *E*-value cutoff of 1E-05. The key candidate genes were identified by co-expression analysis across RNA-Seq samples from ChP with Pearson correlation coefficients (PCC). MeJA treated RNA-seq datasets in *P. ginseng* cv. CS were used from Lee *et al*., [14].

Identification of FAD and CAB genes

FAD genes in *P. ginseng* were identified using Pfam accessions PF00487, PF11960 and PF03405 from an INTERPRO scan. InterPro analysis was also used to identify FADs in other selected plant species. CAB genes in *P. ginseng* and other species were identified using Pfam domain PF00504 from Interpro annotation. Phylogenetic trees were generated by MEGA 6.0 [82].

Phylogenetic analysis

The deduced amino acid sequences were aligned using MUSCLE with default parameters and phylogenetic tree was generated by the neighbor-joining (NJ) analysis and/or maximum likelihood (ML) analysis with default parameters of MEGA 6.0[82]. Bootstrap support value was calculated by 1,000 replicates. Subgroups were divided based on previous reports of gene family in *A. thaliana* and *Oryza sativa*.

Estimation of orthologous gene copies using low-coverage WGS

The low-coverage (~10x) WGS data of from *P. ginseng* cv. YuP (tetraploid) and *P. vietnamensis* (diploid) were utilized for estimation of presence/absence for the orthologous gene copies. The paired-end reads were

quality-trimmed and pooled together as single reads. Based on OrthoMCL, the orthologs genes were selected from gene families such as FAD and CAB. The pooled reads of *P. ginseng* cv. YuP and *P. vietnamensis* were mapped to the orthologs genes (only CDS region) separately using BWA [83]. Then, the percentage of each gene coding length (bp) covered by mapping reads was determined to check whether the coding region of a gene present in the whole genome sequencing library of each species.

REFERENCES

1. Baeg IH, So SH: The world ginseng market and the ginseng (Korea). *J Ginseng Res* 2013, 37(1):1-7.

2. WE C: Ginseng: The Genus *Panax*., vol. 15. UK: Taylor & Francis e-Library; 2006.

3. Radad K, Gille G, Liu L, Rausch WD: Use of ginseng in medicine with emphasis on neurodegenerative disorders. *J Pharmacol Sci* 2006, 100(3):175-186.

4. Cho I-H: Effects of *Panax ginseng* in neurodegenerative diseases. *J Ginseng Res* 2012, 36(4):342-353.

5. Zheng SD, Wu HJ, Wu D: Roles and mechanisms of ginseng in protecting heart. *Chin J Integr Med* 2012, 18(7):548-555.

6. Xie JT, Mehendale SR, Li X, Quigg R, Wang X, Wang CZ, Wu JA, Aung HH, A Rue P, Bell GI *et al*: Anti-diabetic effect of ginsenoside Re in ob/ob mice. *Biochim Biophys Acta* 2005, 1740(3):319-325.

7. Jung HJ, Choi H, Lim HW, Shin D, Kim H, Kwon B, Lee JE, Park EH, Lim CJ: Enhancement of anti-inflammatory and antinociceptive actions of red ginseng extract by fermentation. *J Pharm Pharmacol* 2012, 64(5):756-762.

8. Wong AS, Che CM, Leung KW: Recent advances in ginseng as cancer therapeutics: a functional and mechanistic overview. *Nat Prod Rep* 2015, 32(2):256-272.

9. Shi W, Wang Y, Li J, Zhang H, Ding L: Investigation of ginsenosides in different parts and ages of *Panax ginseng*. *Food Chem* 2007, 102(3):664-668.

10. Oh JY, Kim YJ, Jang MG, Joo SC, Kwon WS, Kim SY, Jung SK, Yang DC: Investigation of ginsenosides in different tissues after elicitor treatment in *Panax ginseng*. *J Ginseng Res* 2014, 38(4):270-277.

11. Xiao D, Yue H, Xiu Y, Sun X, Wang Y, Liu S: Accumulation characteristics and correlation analysis of five ginsenosides with different cultivation ages from different regions. *J Ginseng Res* 2015, 39(4):338-344.

12. Kim YJ, Jeon JN, Jang MG, Oh JY, Kwon WS, Jung SK, Yang DC: Ginsenoside profiles and related gene expression during foliation in *Panax ginseng* Meyer. *J Ginseng Res* 2014, 38(1):66-72.

13. Jiang M, Liu J, Quan X, Quan L, Wu S: Different chilling stresses stimulated the accumulation of different types of ginsenosides in *Panax ginseng* cells. *Acta Physiol Plant* 2016, 38(8):210.

14. Lee Y, Park HS, Lee DK, Jayakodi M, Kim NH, Koo HJ, Lee SC, Kim YJ, Kwon SW, Yang TJ: Integrated transcriptomic and metabolomic analysis of five *Panax ginseng* cultivars reveals the dynamics of ginsenoside biosynthesis. *Front Plant Sci* 2017, 8:1048.

15. Hong C, Lee S, Park J, Plaha P, Park Y, Lee Y, Choi J, Kim K, Lee J, Lee J *et al*: Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences. *Mol Genet Genomics* 2004, 271(6):709-716.

16. Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park M, Choi D, Lim YP, Kwon SJ, Park BS *et al*: Major repeat components covering one-third of the ginseng (*Panax ginseng* C.A. Meyer) genome and evidence for allotetraploidy. *Plant J* 2014, 77(6):906-916.

17. Jang W, Kim NH, Lee J, Waminal NE, Lee SC, Jayakodi M, Choi HI, Park JY, Lee JE, Yang TJ: A Glimpse of *Panax ginseng* Genome Structure Revealed from Ten BAC Clone Sequences Obtained by SMRT Sequencing Platform. *Plant Breed Biotechnol* 2017, 5(1):25-35.

18. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 2008, 9(1):R7.

19. Kim D, Langmead B, Salzberg SL: HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015, 12(4):357-360.

20. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL: StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015, 33(3):290-295.

21. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010, 28(5):511-515.

22. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011, 29(7):644-652.

23. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD: Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 2003, 31(19):5654-5666.

24. Wu TD, Watanabe CK: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005, 21(9):1859-1875.

25. Iwata H, Gotoh O: Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res* 2012, 40(20):e161.

26. Gotoh O: Direct mapping and alignment of protein sequences onto genomic sequence. *Bioinformatics* 2008, 24(21):2438-2444.

27. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M: BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 2016, 32(5):767-769.

28. Lomsadze A, Burns PD, Borodovsky M: Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* 2014, 42(15):e119.

29. Stanke M, Diekhans M, Baertsch R, Haussler D: Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 2008, 24(5):637-644.

30. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G: CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 2007, 35(Web Server issue):W345-349.

31. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W: CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 2013, 41(6):e74.

32. Eddy SR: HMMER: Profile hidden Markov models for biological sequence analysis. 2001.

33. Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E: SUPPA: a super-fast pipeline for alternative splicing analysis from RNA-Seq. *bioRxiv* 2014:008763.

34. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009, 37(Database issue):D211-D215.

35. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al*: InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014, 30(9):1236-1240.

36. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21(18):3674-3676.

37. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007, 35(Web Server issue):W182-185.

38. Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, Zhang P, Banf M, Dai X, Martin GB, Giovannoni JJ *et al*: iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Mol Plant* 2016, 9(12):1667-1670.

39. Mathiyalagan R, Subramaniyam S, Natarajan S, Kim YJ, Sun MS, Kim SY, Kim YJ, Yang DC: Insilico profiling of microRNAs in Korean ginseng (*Panax ginseng* Meyer). *J Ginseng Res* 2013, 37(2):227-247.

40. Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, Schwach F, Dalmay T, Moulton V: The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 2012, 28(15):2059-2061.

41. Dai X, Zhao PX: psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res* 2011, 39(Web Server issue):W155-W159.

42. Jones-Rhoades MW, Bartel DP, Bartel B: MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol* 2006, 57:19-53.

43. Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, Givan SA, Kasschau KD, Carrington JC: Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA-and tasiRNA-directed targeting. *Plant Cell* 2007, 19(3):926-942.

44. Small ID, Peeters N: The PPR motif–a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem Sci* 2000, 25(2):45-47.

45. Rivals E, Bruyere C, Toffano-Nioche C, Lecharny A: Formation of the Arabidopsis pentatricopeptide repeat family. *Plant Physiol* 2006, 141(3):825-839.

46. Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC: Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nat Genet* 2004, 36(12):1282-1290.

47. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL: High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PloS one* 2007, 2(2):e219.

48. Xie F, Frazier TP, Zhang B: Identification, characterization and expression analysis of MicroRNAs and their targets in the potato (*Solanum tuberosum*). *Gene* 2011, 473(1):8-22.

49. An W, Gong W, He S, Pan Z, Sun J, Du X: MicroRNA and mRNA expression profiling analysis revealed the regulation of plant height in *Gossypium hirsutum*. *BMC genomics* 2015, 16(1):1.

50. Benveniste P: Biosynthesis and accumulation of sterols. *Annu Rev Plant Biol* 2004, 55:429-457.

51. Han JY, Kim HJ, Kwon YS, Choi YE: The Cyt P450 enzyme CYP716A47 catalyzes the formation of protopanaxadiol from dammarenediol-II during ginsenoside biosynthesis in *Panax ginseng*. *Plant Cell Physiol* 2011, 52(12):2062-2073.

52. Gachon CM, Langlois-Meurinne M, Saindrenan P: Plant secondary metabolism glycosyltransferases: the emerging functional analysis. *Trends Plant Sci* 2005, 10(11):542-549.

53. Khorolragchaa A, Kim YJ, Rahimi S, Sukweenadhi J, Jang MG, Yang DC: Grouping and characterization of putative glycosyltransferase genes from *Panax ginseng* Meyer. *Gene* 2014, 536(1):186-192.

54. Han JY, Hwang HS, Choi SW, Kim HJ, Choi YE: Cytochrome P450 CYP716A53v2 catalyzes the formation of protopanaxatriol from protopanaxadiol during ginsenoside biosynthesis in *Panax ginseng*. *Plant Cell Physiol* 2012, 53(9):1535-1545.

55. Tansakul P, Shibuya M, Kushiro T, Ebizuka Y: Dammarenediol-II synthase, the first dedicated enzyme for ginsenoside biosynthesis in *Panax ginseng*. *FEBS Lett* 2006, 580(22):5143-5149.

56. Han JY, In JG, Kwon YS, Choi YE: Regulation of ginsenoside and phytosterol biosynthesis by RNA interferences of squalene epoxidase gene in *Panax ginseng*. *Phytochemistry* 2010, 71(1):36-46.

57. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011, 12:323.

58. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J *et al*: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013, 14(6):671-683.

59. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26(1):139-140.

60. Wijnheijmer E, Brandenburg W, Ter Borg S: Interactions between wild and cultivated carrots (*Daucus carota* L.) in the Netherlands. *Euphytica* 1989, 40(1-2):147-154.

61. Gross KL: Predictions of fate from rosette size in four “biennial” plant species: Verbascum thapsus, Oenothera biennis, *Daucus carota*, and Tragopogon dubius. *Oecologia* 1981, 48(2):209-213.

62. Dawid C, Dunemann F, Schwab W, Nothnagel T, Hofmann T: Bioactive C17-Polyacetylenes in Carrots (*Daucus carota* L.): Current Knowledge and Future Perspectives. *J Agric Food Chem* 2015, 63(42):9211-9222.

63. Los DA, Murata N: Membrane fluidity and its roles in the perception of environmental signals. *Biochim Biophys Acta* 2004, 1666(1):142-157.

64. Martz F, Kiviniemi S, Palva TE, Sutinen M-L: Contribution of omega-3 fatty acid desaturase and 3-ketoacyl-ACP synthase II (KASII) genes in the modulation of glycerolipid fatty acid composition during cold acclimation in birch leaves. *J Exp Bot* 2006, 57(4):897-909.

65. Shewfelt R: Response of plant membranes to chilling and freezing. In: *Plant membranes*. Springer; 1992: 192-219.

66. Thomashow MF: Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. *Annu Rev Plant Physiol Plant Mol Biol* 1999, 50(1):571-599.

67. Palta J, Li P: Cell membrane properties in relation to freezing injury. *Plant cold hardiness and freezing stress* 1978, 1:93-115.

68. Román Á, Andreu V, Hernández ML, Lagunas B, Picorel R, Martínez-Rivas JM, Alfonso M: Contribution of the different omega-3 fatty acid desaturase genes to the cold response in soybean. *J Exp Bot* 2012, 63(13):4973-4982.

69. Matteucci M, D'angeli S, Errico S, Lamanna R, Perrotta G, Altamura M: Cold affects the transcription of fatty acid desaturases and oil quality in the fruit of Olea europaea L. genotypes with different cold hardiness. *J Exp Bot* 2011:62(10):3403-3420.

70. Thomashow MF: Role of cold-responsive genes in plant freezing tolerance. *Plant Physiol* 1998, 118(1):1-8.

71. Khodakovskaya M, McAvoy R, Peters J, Wu H, Li Y: Enhanced cold tolerance in transgenic tobacco expressing a chloroplast ω-3 fatty acid desaturase gene under the control of a cold-inducible promoter. *Planta* 2006, 223(5):1090-1100.

72. Cao S, Zhou XR, Wood CC, Green AG, Singh SP, Liu L, Liu Q: A large and functionally diverse family of Fad2 genes in safflower (*Carthamus tinctorius* L.). *BMC Plant Biol* 2013, 13(1):5.

73. Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury JM, Badger JH *et al*: The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* 2010, 465(7298):617-621.

74. Ganeteg U, Külheim C, Andersson J, Jansson S: Is each light-harvesting complex protein important for plant fitness? *Plant Physiol* 2004, 134(1):502-509.

75. Kovács L, Damkjær J, Kereïche S, Ilioaia C, Ruban AV, Boekema EJ, Jansson S, Horton P: Lack of the light-harvesting complex CP24 affects the structure and function of the grana membranes of higher plant chloroplasts. *Plant Cell* 2006, 18(11):3106-3120.

76. Xu YH, Liu R, Yan L, Liu ZQ, Jiang SC, Shen YY, Wang XF, Zhang DP: Light-harvesting chlorophyll a/b-binding proteins are required for stomatal response to abscisic acid in Arabidopsis. *J Exp Bot* 2012, 63(3):1095-1106.

77. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al*: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 2012, 1(1):18.

78. Marçais G, Kingsford C: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011, 27(6):764-770.

79. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J *et al*: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013, 14(6):671-683.

80. Smit A, Hubley R: RepeatModeler Open-1.0. *Repeat Masker Website* 2010.

81. Smit AF, Hubley R, Green P: RepeatMasker Open-3.0. In.; 1996.

82. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 2013, 30(12):2725-2729.

83. Li H, Durbin R: Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009, 25(14):1754-1760.

CHAPTER 2

Comparative transcriptome analysis of heat stress responsiveness between two contrasting ginseng cultivars

Abstract

Panax ginseng has been used in traditional medicine to strengthen the body and mental well-being of humans for thousands of years. Many elite ginseng cultivars have been developed, and ginseng cultivation has become well established during the last century. However, heat stress poses an important threat to the growth and sustainable production of ginseng. Efforts have been made to study the effects of high temperature on ginseng physiology, but knowledge of the molecular responses to heat stress is still limited. We sequenced the transcriptomes (RNA-Seq) of two ginseng cultivars, Chungpoong (CP) and Yunpoong (YP), which are sensitive and resistant to heat stress, respectively, after 1- and 3-week heat treatments. Differential gene expression (DEG) and gene ontology (GO) enrichment along with profiled chlorophyll contents were performed. CP is more sensitive to heat stress than YP, and exhibited a lower chlorophyll content than YP. Moreover, heat stress reduced the chlorophyll content more rapidly in CP compared to YP. A total of 329 heat-responsive genes were identified. Intriguingly, genes encoding chlorophyll ab binding (CAB) proteins, WRKY transcription factors, and fatty acid desaturase (FAD) were predominantly responsive during heat stress and appeared to inhibit photosynthesis. In addition, a genome-wide scan of photosynthetic and sugar metabolic genes revealed reduced transcript levels for *ribulose 1,5-bisphosphate carboxylase/oxygenase* (*RuBisCO*) under heat stress, especially in CP, possibly attributable to elevated levels of soluble sugars. This comprehensive genomic analysis reveals candidate loci/gene targets for breeding and functional studies related to developing high-temperature tolerant ginseng varieties.

Keywords: *Panax ginseng*, Chunpoong, Yunpoong, heat stress, RNA-Seq

Introduction

Plants experience various abiotic and biotic stresses that adversely affect their growth and productivity worldwide [1-4]. High temperature and heat stress are a particular concern, as temperatures are projected to rise by 6.9°C by the end of this century [5] due to global warming [6-8], and temperature increase of 3–4°C is expected to reduce crop productivity by 15–35% [9]. In addition, heat stress compounds with drought stress to reduce agricultural production even more than heat or drought stress alone [10].

The major effects of heat include increased membrane fluidity, protein denaturation, and stimulated production of reactive oxygen species (ROS), which cause permanent damage to plant growth and development [4], notably during their reproductive stage [4]. These effects, and the plant's response to them, can be reflected by alterations to the transcriptome, proteome, and metabolome, and ultimately, severe cellular injury [8]. Heat stress upregulates about 5% of the plant transcriptome, including genes encoding molecular chaperones and genes involved in signaling, translation, and metabolism to mediate heat stress responses in plants [4]. Thus, it is essential to identify QTLs/genes involved in heat stress responses to develop crops with enhanced heat tolerance and ensure global food security.

Korean ginseng (*Panax ginseng* C.A. Meyer, hereafter referred to as *P. ginseng* or simply ginseng) is one of the most famous traditional medicinal herbs. Ginseng, which has been used for more than 2,000 years in Asia, produces triterpene saponins as principle bioactive secondary metabolites with various pharmacological and physiological benefits to humans [11-13]. *P. ginseng* is a shade-loving perennial plant that is susceptible to photoinhibition at light intensities exceeding 500 μmol $m^{-2} s^{-1}$. In addition, leaves of this plant are burned after exposure to 30°C for more than 5 days and its growth will completely stop. Due to this sensitivity to heat, *P. ginseng* plants can be severely damaged in summer and require careful

management. Moreover, the increasing temperatures associated with global warming represent an important threat to *P. ginseng* growth and production. A few physiological and morphological studies [14-16] have been conducted to understand the heat response in *P. ginseng*, but a comprehensive study that combines genome, transcriptome and physiological data to understand the heat response in *P. ginseng* is still needed.

In Korea, the *P. ginseng* cultivar (cv.) Yunpoong (YP) is known to have slightly higher heat tolerance (light saturation point of 400 μmol m^{-2} s^{-1}) than the heat-sensitive cv. Chunpoong (CP) (light saturation point of 200 μmol m^{-2} s^{-1}) [14-16]. In this study, we analyzed and compared transcriptomes of CP and YP under different heat treatments to understand ginseng heat-response mechanisms and identify major genes and biological processes affected by heat stress. This study reports the first transcriptome profile in response to heat stress and provides novel candidate target genes for alleviating adverse effects of heat stress in *P. ginseng*.

Materials and methods

Plant materials, growth conditions, and heat treatments

Dormant roots with healthy rhizomes of 1-year-old cv. CP and YP plants were obtained from the Korea Ginseng Corporation (KGC, Deajeon, Korea). After storage for 1 month at 4°C to break dormancy, the roots were planted in soil and grown for 4 weeks to generate plants with fully-expanded leaves under normal growth conditions (24°C, relative humidity 60%, and continuous light of 40 μmol m^{-2} s^{-1}). Samples of these plants were harvested as controls before heat treatment. For heat treatment, the plants were treated with 30(±1)°C for 1 week and 3 weeks (relative humidity and light conditions were the same as the normal growth conditions). After heat treatment, sample leaves were excised, immediately frozen using LN_2, and stored at -70°C before total RNA isolation. Three independent biological replicates were prepared, and each replicate included leaves from three or more plants.

Measurement of chlorophyll content

Chlorophyll content was measured in leaves of more than 20 plants using a chlorophyll meter (SPAD-502, Minolta, Japan) before and during heat treatment, as described by Lee et al [16].

Total RNA isolation and RNA-Seq analysis

Total RNA was isolated from leaves using the RNeasy Plant kit (QIAGEN, Germany) and/or Hybrid-R kit (GeneAll, Korea), according to the manufacturer's instructions. After examination of its quality and quantity using a Bioanalyzer (Agilent Technologies, USA), about 2 μg total RNA was used for construction of RNA-Seq libraries. RNA-Seq libraries with an insert size of 300 bp were constructed independently using an Illumina TruSeq RNA Sample Preparation Kit according to the manufacturer's instructions. Pooled libraries were sequenced using the Illumina HiSeq 2000 platform with paired-end (PE) reads of 101 bp at Macrogen Co. (Seoul, Korea) or the Illumina NextSeq 500 platform with PE read length of 150 bp at LabGenomics Co. (Seongnam, Korea). For transcriptome analysis, reads containing bacterial contaminants were removed by mapping against the available bacterial genomes using BWA [17] followed by removal of PCR duplicates and ribosomal RNA reads using FastUniq [18] and SortMeRNA [19], respectively. Finally, stringent quality control and removal of adaptor contamination was done using the NGS QC Toolkit (v2.3.3) [20].

Differential gene expression analysis

The current version of the ginseng gene set (IPGA_v1.1) was retrieved from the ginseng genome database (ginsengdb.snu.ac.kr) [21]. Trimmed, high-quality RNA-Seq reads were mapped to the ginseng gene set to calculate FPKM (Fragments Per Kilobase per Million) using RSEM [22]. The sequencing library differences (due to the use of a different Illumina platform) between replicate RNA-Seq samples were normalized using Trimmed Mean of M values (TMM) [23]. The bioconductor package edgeR [24] was used to identify differentially-expressed transcripts between heat stress samples of CP and YP. Genes exhibiting over 2-fold

changes with a significant false discovery rate (FDR) of 0.001 were considered as differentially expressed. GO enrichment analysis was performed on the differentially-expressed genes using Fisher's Exact Test with a multiple testing correction FDR limit of 0.05.

Gene family annotation

Protein domains and motifs of genes in *P. ginseng* were identified using InterProScan [25] (v5.13). Genes involved in sugar metabolism and photosynthesis were identified using the KAAS server [26]. For comparative sugar metabolic gene analysis, gene sets from carrot (http://www.ncbi.nlm.nih.gov/Traces/wgs/wgsviewer.cgi?val=LNRQ01&search=LNRQ01000000&display=contigs), tomato (ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG2.4_release/), arabidopsis (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/), grape (http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/12X/annotation/), and pepper (ftp://ftp.solgenomics.net/genomes/Capsicum_annuum/C.annuum_cvCM334/annotation/) were used.

Results and Discussion

***P. ginseng* cultivars CP and YP show different responses to heat**

Table 2-1. Summary of *P. ginseng* heat-treated transcriptome data

Samples	Raw		After filtering		SRA ID
	No. of reads	Length (bp)	No. of reads	Length (bp)	

cv. Chunpoong					
Control, rep# 1 [a)]	34,200,406	3,454,241,006	28,548,966	2,883,445,566	SRR6117049
Control, rep# 2 [b)]	12,010,556	1,671,583,519	8,401,588	1,169,602,247	SRR6117050
Control, rep# 3 [b)]	14,219,204	1,971,113,105	10,112,344	1,403,615,982	SRR6117051
1-week heat treatment, rep# 1 [b)]	12,797,522	1,776,036,672	9,176,372	1,403,615,982	SRR6117052
1-week heat treatment, rep# 2 [b)]	14,303,250	1,985,768,203	10,266,344	1,274,623,891	SRR6117053
1-week heat treatment, rep# 3 [b)]	14,795,776	2,057,976,292	10,419,368	1,427,291,425	SRR6117057
3-week heat treatment, rep#1 [a)]	40,201,970	4,060,398,970	35,188,388	1,633,028,161	SRR6117058
3-week heat treatment, rep# 2 [b)]	14,670,068	2,031,823,791	11,931,866	1,650,210,144	SRR6117059
3-week heat treatment, rep# 3 [b)]	18,163,348	2,502,843,099	14,568,344	2,004,943,384	SRR6117061
cv. Yunpoong					
Control, rep# 1 [b)]	12,468,942	1,721,266,026	11,123,218	1,528,796,322	SRR6109634
Control, rep# 2 [b)]	15,189,718	2,106,203,208	13,386,670	1,846,968,856	SRR6109657
Control, rep# 3 [b)]	14,917,750	2,062,271,071	13,121,488	1,804,691,129	SRR6109670
1-week heat treatment, rep# 1 [b)]	12,277,348	1,699,892,166	10,716,748	1,476,028,232	SRR6109675
1-week heat treatment, rep# 2 [b)]	14,461,344	2,001,009,942	12,730,788	1,752,633,304	SRR6109678
1-week heat treatment, rep# 3 [b)]	12,226,114	1,690,713,369	10,564,980	1,452,352,808	SRR6111127
3-week heat treatment, rep# 1 [b)]	13,934,774	1,925,525,852	12,403,000	1,706,350,704	SRR6111131
3-week heat treatment, rep# 2 [b)]	12,189,310	1,691,363,635	10,905,514	1,507,226,834	SRR6111139
3-week heat treatment, rep# 3 [b)]	11,997,888	1,652,914,364	10,636,142	1,458,514,772	SRR6111142
Total	295,025,288	38,062,944,290	244,202,128	29,383,939,743	

[a)] Sequenced using HiSeq2000, [b)]Sequenced using NextSeq500, rep: replicates

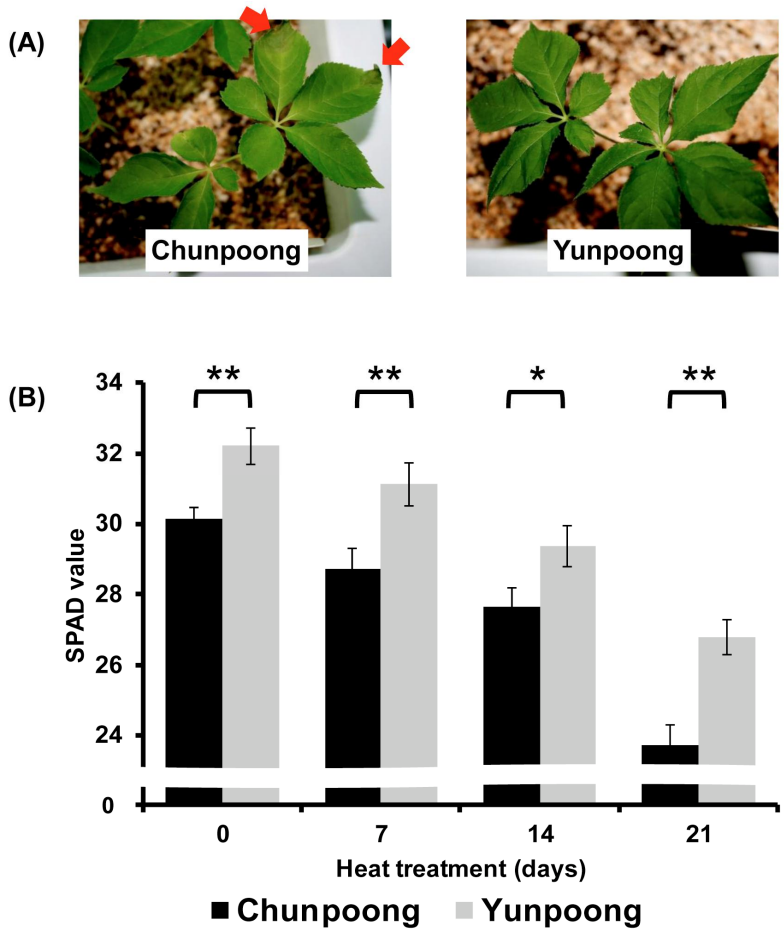

Fig. 2-1. Leaf burning and chlorophyll content of *P. ginseng* plants under heat stress. (A). Different leaf burning occurrence between CP and YP. After heat treatment for 3 weeks, leaves of CP showed damaged regions (leaf burning, red-arrows) while those of YP did not show any visibly damaged regions. (B) Changes of chlorophyll content (SPAD values) during heat treatment. Error bars indicate mean±standard deviation (SD) obtained from more than 20 plants. Asterisks indicate values with significant differences between two cultivars, based on Student's *t*-test (*P<0.05, **P< 0.01).

During the period of heat treatment, *P. ginseng* cultivars CP and YP showed different responses in both morphology and physiology. After heat treatment for 3

weeks, leaves of CP started to show damaged regions indicated by leaf burning whereas YP did not show any visibly damaged regions (**Fig. 2-1A**). The damage in leaf tissues was confirmed by measuring chlorophyll content (SPAD values) (**Fig. 2-1B**). Although the chlorophyll content was different between the two cultivars even under normal growth conditions, the difference increased proportionally to the treatment time. The chlorophyll content was 7% higher in YP than in CP under normal growth conditions. However, the content was 8% and 13% higher in YP after treatment for 1 week and 3 weeks, respectively. This difference indicated that chlorophyll content decreased more slowly in YP compared to CP under heat stress, which is in agreement with the results of a previous physiological study of the two cultivars [16].

Identification of heat-responsive genes in *P. ginseng*

In total, 59,352 genes were retrieved from the ginseng genome database and used as a reference for differentially-expressed gene (DEG) analysis. Transcriptome sequencing of three independent biological samples of both CP and YP yielded a total of 295,025,288 raw RNA-seq reads (**Table 2-1**). All raw sequencing reads were deposited into the sequencing read archive (SRA) of NCBI (accession numbers are in **Table 2-1**). Prior to DEG analysis, the raw RNA-Seq reads were processed and yielded 244 million high-quality reads. DEG analysis was performed between heat-treated samples from CP and YP independently. This analysis resulted in 883 and 448 genes showing differential expression upon 1- and 3-week heat stress in CP and YP, respectively (**Fig. 2-2A**). Further, a set of 329 genes was found in both CP and YP with a similar response of either up- or down-regulation (**Fig. 2-3**) after heat stress. We considered these genes as heat-responsive genes in ginseng irrespective of genotypes. Additionally, we grouped those DE genes based on their protein domains (**Fig. 2-2B**) and found that genes belonging to the chlorophyll ab binding protein (CAB) family primarily respond to heat stress. Photosynthesis is highly vulnerable to high temperature and is inhibited long before other symptoms or cell functions are impaired [1, 27].

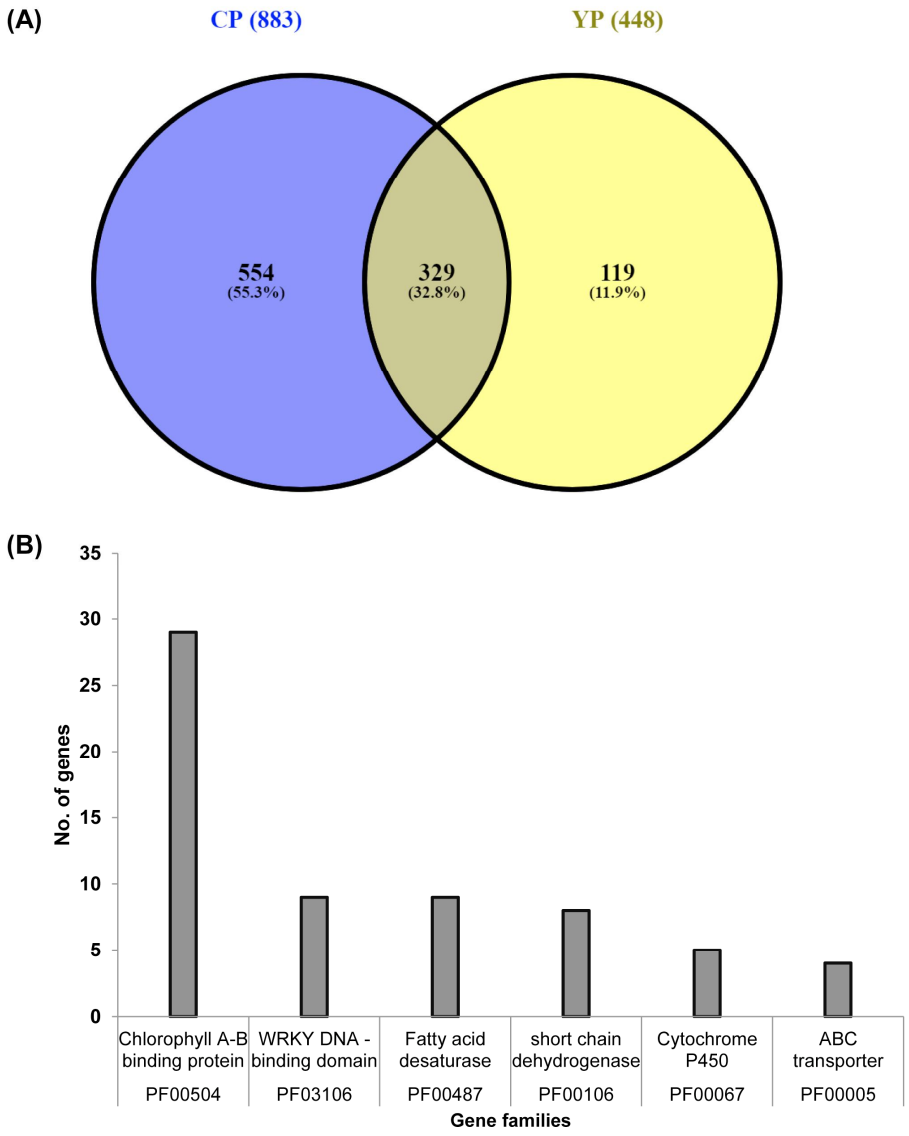

Fig. 2-2. **Comparative transcriptome of CP and YP cultivars.** (A). Venn diagram depicts number of differential expressed (DE) genes commonly found in both CP and YP cultivars. (B). Classification of DE genes based on their Pfam domain annotated by InterPro Scan.

CAB genes are key components of photosynthesis, transferring light energy to the reaction centers of photosystem I (PS I) and photosystem II (PS II), where it is converted into chemical energy (i.e. NADPH and ATP). Various studies have revealed that photosystem II (PSII) is extremely sensitive to heat, which can greatly reduce PSII activity or even stop it [28]. Our results indicate that heat might also influence the photosynthetic process in ginseng by inhibiting the light-harvesting system.

WRKY transcription factors (TFs) are one of the largest families of transcriptional regulators and play a crucial role in the defense response against various abiotic stresses [29]. WRKY39 and CaWRKY40 are the best examples of TFs that are triggered by heat in Arabidopsis [30] and pepper, respectively [31]. Notably, WRKY TF genes responded strongly to heat in ginseng (**Fig. 2-2B**), suggesting their potential involvement during heat stress.

The accumulation of unsaturated fatty acids can aggravate heat damage [32], and transgenic tobacco [33] and rice [34] with silenced *FATTY ACID DESATURASE* (*FAD*) genes and reduced levels of trienoic fatty acid exhibit high temperature tolerance with increased photochemical efficiency of PSII. Similarly, silencing of the *LeFAD7* gene in transgenic tomato conferred tolerance to high temperature (45°C) [35]. Given these trends, it is possible that the extreme sensitivity of PSII in ginseng may be due to the effects of high temperature on chloroplast membranes. The amount of polyunsaturated fatty acid is associated with cold tolerance in plants [36, 37], and a ginseng genome study has implicated an expanded *FAD* gene family in ginseng's cold adaptation [38]. The major polyunsaturated fatty acids in plant membrane lipids, such as trienoic fatty acids including hexadecatrienoic acid (16:3) and linolenic acid (18:3) are important for ensuring the maintenance of chloroplasts during plant growth under low temperatures [39]. Using KEGG analysis, we found that most of the ginseng *FAD* genes were associated with trienoic fatty acids, indicating their role in cold acclimation. We also observed that nine genes belonging to the *FAD* family showed very strong expression after 3-week heat stress. Thus, we hypothesize that silencing

the *FAD* genes associated with trienoic fatty acid accumulation in ginseng might increase their tolerance to heat stress.

Table 2-2. Top enriched GO biological terms for common DE genes in CP and YP in response to heat stress

GO-ID	Term	Category	FDR
GO:0009768	photosynthesis, light harvesting in photosystem I	P	2.82E-36
GO:0010196	nonphotochemical quenching	P	4.93E-33
GO:0010114	response to red light	P	3.99E-31
GO:0010155	regulation of proton transport	P	5.29E-28
GO:0019344	cysteine biosynthetic process	P	1.76E-26
GO:0010218	response to far red light	P	9.68E-26
GO:0009637	response to blue light	P	4.12E-24
GO:0006364	rRNA processing	P	7.58E-24
GO:0009769	photosynthesis, light harvesting in photosystem II	P	5.60E-20
GO:0009744	response to sucrose stimulus	P	1.18E-17
GO:0006636	unsaturated fatty acid biosynthetic process	P	1.18E-15
GO:0018298	protein-chromophore linkage	P	1.19E-12
GO:0030003	cellular cation homeostasis	P	2.44E-11
GO:0009750	response to fructose stimulus	P	4.88E-11

We performed GO enrichment analysis with the DE genes to investigate their biological significance. Consistent with the above results, we found that genes involved in photosynthesis, including light harvesting, light response, cysteine biosynthesis, and sugar metabolism were primarily affected in both cultivars of ginseng (**Table 2-2**). In the *P. ginseng* genome, *CAB*, *FAD* and *WRKY* genes are significantly expanded, and such adaptive expansion might enable this plant to

acclimate to cold and low light with high photosynthetic quantum efficiency [38]. However, based on our results, these expanded genes might also cause antagonistic effects that lead to photoinhibition [16] and increased sensitivity to heat or light, or to a combination of both stresses. Clearly, more functional studies with these gene targets are needed to test these hypotheses in ginseng.

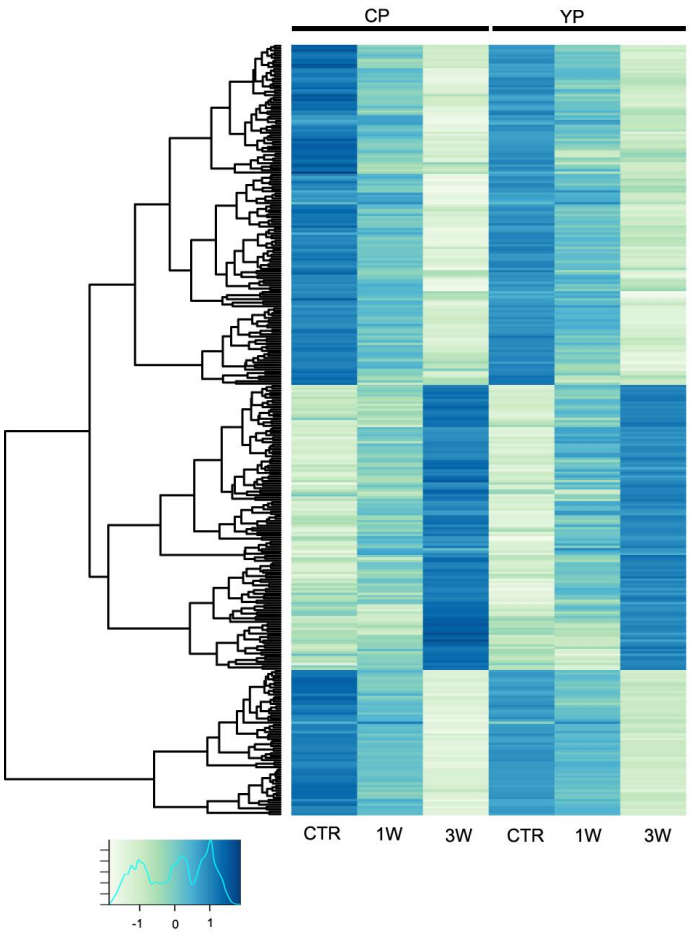

Fig.2-3. **Expression profiles of differentially expressed genes against heat stress.** A total of 329 genes showed differential expression between CP and YP cultivars. Heatmap shows the hierarchical clustering of average FPKM values obtained from individual normalized FPKM values of three replicates. CTR, 1W and 3W represent control, 1-week, and 3-week heat treatment, respectively.

Comparative expression of genes involved in photosynthesis

Since photosynthesis was markedly influenced by heat stress in ginseng, we compared the expression of genes involved in photosynthesis between CP and YP. A total of 103 photosynthetic genes were identified using ginseng functional gene annotation information. Overall, these genes showed slightly stronger expression in CP than in YP (**Fig. 2-4A**). Notably, we observed that the expression of electron carrier genes such as those encoding plastocyanin and ferredoxin $NADP^+$ reductase were significantly altered after 1-week and 3-week heat stress in CP. In agreement with Lee et al. [16], this result indicated that high temperature and light might influence ginseng by exceeding its capacity for electron transfer, subsequently shutting down the electron transfer chain in a process of photoinhibition. Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO) is a key enzyme in CO_2 fixation and determines the photosynthetic rate in response to light intensity. We identified eleven *RuBisCO* genes and observed notably high expression of six genes encoding the large subunit of RuBisCO, which is identical with chloroplast rbcL suggesting transfer of chloroplast encoded rbcL in nuclear genome of ginseng, in YP compared to CP under normal as well as heat-stress conditions (**Fig. 2-4B**). This result corresponds with the high net photosynthetic rate of YP over CP shown by Lee et al. [14] and suggests that photosynthetic rate is co-regulated with RuBisCO expression in ginseng.

Analysis of sugar metabolic genes

Sugars can serve as physiological signals that repress or activate plant genes involved in essential biological processes, including photosynthesis [40]. Further, elevated sugar levels cause reduced photosynthesis activity and stunted growth in plants [40]. Intriguingly, elevated levels of soluble sugars have been identified in *P. ginseng* [41] with significant oscillation. Therefore, we performed a genome-wide scan for key genes encoding enzymes involved in sugar metabolism in ginseng including sucrose phosphate synthase (SPS, EC 2.4.1.14), fructose-1,6-bisphosphatase (F16BPase, EC 3.1.3.11), sucrose-phosphate phosphatase (SPP EC

3.1.3.24), sucrose synthase (SuSy, EC 2.4.1.13), neutral invertase (NI, EC 3.2.1.26), fructokinase (FK, EC 2.7.1.4), hexokinase (HK, EC 2.7.1.1), and UDP-glucose pyrophosphorylase (UGPase, EC 2.7.7.9). When we compared these genes in *P. ginseng* with those of other plant species, we found increased gene numbers in ginseng compared to model plants and annuals endemic to sunny locales (**Table 2-3**). We also investigated the expression of those key genes and observed slightly higher expression in CP compared to YP, indicating that sugar metabolic processes are more active in heat-susceptible CP.

Table 2-3. Number of sugar metabolic genes in ginseng and other plant species

Species	***SPS***	***F16BPase***	***SPP***	***SUSY***	***NI***	***FK***	***HK***	***UGPase***
Ginseng	12	15	4	27	25	13	25	15
Carrot	8	6	3	9	16	7	4	7
Tomato	6	7	2	8	19	8	6	6
Arabidopsis	8	4	4	6	12	11	6	7
Grape	7	5	2	7	12	4	7	5
Pepper	4	7	2	6	11	8	7	4

We identified six sucrose transporter genes in ginseng. Intriguingly, two and four of the sucrose transporter genes were upregulated upon 1- and 3-week heat stress, respectively, in CP (**Fig. 2-4C**) when compared to YP. In other plant species, elevated sugar concentrations correlated negatively with *RuBisCO* transcript levels [42], which coincides with our results: CP showed strong expression of genes involved in sugar metabolism but showed reduced transcript levels of *RuBisCO*. Therefore, elevated sugar accumulation could also be one of the components leading CP to be more susceptible than YP to heat and light stresses.

Conclusion

Ginseng is an obligate shade species whose photosynthetic capability is significantly reduced by non-optimum light intensity and temperature. In this study, we compared the transcriptome between heat-injury tolerant YP and heat-susceptible CP ginseng cultivars. Overall, CP and YP transcriptomes showed slight variation in terms of their gene expression patterns under non-stress conditions. However, our results suggest ginseng responds to heat stress with an expanded number of *CAB*, *FAD* and *WRKY* genes. These genes may play a major role in shade and cold adaptation. Conversely, the expansion of these gene families appears to have a significant negative impact on heat and light tolerance, which might lead to an electron over-flow in the photosynthetic chain and thereby cause photoinhibition, especially in the heat-susceptible line CP. This process might also be co-regulated with elevated sugars that reduce the demand for ATP and NADPH by inhibiting the transcript level of *RuBisCO* and decreasing membrane desaturation by FAD. Altogether, this study provides novel and fundamental insight into ginseng heat responses that will be important for generating heat-tolerant ginseng varieties.

.

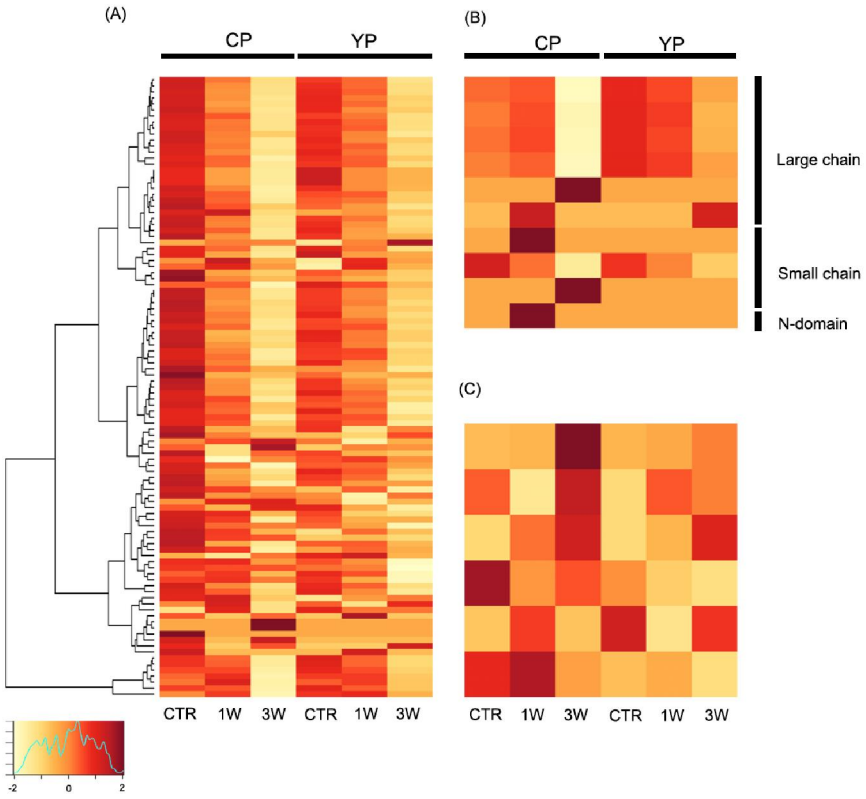

Fig. 2-4. Gene expression comparison between CP and YP. (A). Heatmap representing the comparative expression profiling of 103 photosynthetic genes between CP and YP cultivars. The list of genes and their expression can be found in Supplementary Table S3. (B). Expression of ten out of eleven *RuBisCO* genes identified in cultivars of CP and YP. Detailed information can be found in Supplementary Table S4. (C). Heatmap showing expression patterns in CP and YP of the six sucrose transporter genes identified in ginseng. Corresponding expression values can be obtained in Supplementary Table S5. Each heatmap includes hierarchical clustering of average FPKM values obtained from individual normalized FPKM values of three replicates. CTR, 1W and 3W represent control, 1-week, and 3-week heat treatment, respectively.

REFERENCES

1. Ahuja I, de Vos RC, Bones AM, Hall RD: Plant molecular stress responses face climate change. *Trends Plant Sci* 2010, 15(12):664-674.

2. Krasensky J, Jonak C: Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *J Exp Bot* 2012, 63(4):1593-1608.

3. Boyer JS: Plant productivity and environment. Science 1982, 218(4571):443-448.

4. Mittler R: Abiotic stress, the field environment and stress combination. *Trends Plant Sci* 2006, 11(1):15-9.

5. IPoCC: Climate Change 2014–Impacts, Adaptation and Vulnerability: Regional Aspects: Cambridge University Press; 2014.

6. Kotak S, Larkindale J, Lee U, von Koskull-Doring P, Vierling E, Scharf KD: Complexity of the heat stress response in plants. *Curr Opin Plant Biol* 2007, 10(3):310-316.

7. Qu AL, Ding YF, Jiang Q, Zhu C: Molecular mechanisms of the plant heat stress response. *Biochem Biophys Res Commun* 2013, 432(2):203-207.

8. Bita CE, Gerats T: Plant tolerance to high temperature in a changing environment: scientific fundamentals and production of heat stress-tolerant crops. *Front Plant Sci* 2013, 4:273.

9. Ortiz R, Braun HJ, Crossa J, Crouch JH, Davenport G, Dixon J, Dreisigacker S, Duveiller E, He Z, Huerta J, *et al*: Wheat genetic resources enhancement by the International Maize and Wheat Improvement Center (CIMMYT). *Genet Resour Crop Evol* 2008, 55(7):1095-1140.

10. Lipiec J, Doussan C, Nosalewicz A, Kondracka K: Effect of drought and heat stresses on plant growth and yield: a review. *Int Agrophys* 2013, 27(4):463-477.

11. Kim YJ, Zhang D, Yang DC: Biosynthesis and biotechnological production of ginsenosides. *Biotechnol Adv* 2015, 33(6):717-735.

12. Court WE: Ginseng: The Genus Panax. Taylor & Francis e-Library; 2006.

13. Baeg IH, So SH: The world ginseng market and the ginseng (Korea). *J Ginseng Res* 2013, 37(1):1-7.

14. Lee JS, Lee DY, Lee JH, Ahn IO, In JG: Photosynthetic characteristics of resistance and susceptible lines to high temperature injury in *Panax ginseng* Meyer. *J Ginseng Res* 2012, 36(4):461-468.

15. Lee JS, Lee KH, Lee SS, Kim ES, Ahn IO, In JG: Morphological characteristics of ginseng leaves in high-temperature injury resistant and susceptible lines of *Panax ginseng* Meyer. *J Ginseng Res* 2011, 35(4):449-456.

16. Lee JS, Lee JH, Ahn IO: Characteristics of resistant lines to high-temperature injury in ginseng (*Panax ginseng* CA Meyer). *J Ginseng Res* 2010, 34(4):274-281.

17. Li H, Durbin R: Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009, 25(14):1754-1760.

18. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S: FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* 2012, 7(12):e52249.

19. Kopylova E, Noé L, Touzet H: SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012, 28(24):3211-3217.

20. Patel RK, Jain M: NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one* 2012, 7(2):e30619.

21. Jayakodi M, Choi BS, Lee SC, Kim NH, Park JY, Jang W, et al.Ginseng Genome Database: An open-access platform for genomics of *Panax ginseng*. Database. 2017;(under review).

22. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011, 12:323.

23. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castle D, Estelle J, *et al*: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013, 14(6):671-83.

24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26(1):139-140.

25. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWillian H, Maslen J, Mitchell A, Nuka G, *et al*: InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014, 30(9):1236-1240.

26. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007, 35(web Server issue):182-185.

27. Mathur S, Agrawal D, Jajoo A: Photosynthesis: response to high temperature stress. *J Photochem Photobiol B* 2014, 137:116-26.

28. Morales D, Rodriguez P, Dell'Amico J, Nicolas E, Torrecillas A, Sanchez-Blanco M: High-temperature preconditioning and thermal shock imposition affects water relations, gas exchange and root hydraulic conductivity in tomato. *Biologia Plantarum* 2003, 47(2):203.

29. Bakshi M, Oelmuller R: WRKY transcription factors: Jack of many trades in plants. *Plant Signal Behav* 2014, 9(2):e27700.

30. Li S, Zhou X, Chen L, Huang W, Yu D: Functional characterization of Arabidopsis thaliana WRKY39 in heat stress. *Mol Cells* 2010, 29(5):475-83.

31. Dang FF, Wang YN, Yu L, Eulgem T, Lai Y, Liu ZQ, Wang X, Qiu AL, Zhang TX, Lin J, *et al*: CaWRKY40, a WRKY protein of pepper, plays an important role in the regulation of tolerance to heat stress and resistance to Ralstonia solanacearum infection. *Plant Cell Environ* 2013, 36(4):757-774.

32. Hazel JR, Williams EE: The role of alterations in membrane lipid composition in enabling physiological adaptation of organisms to their physical environment. *Prog Lipid Res* 1990, 29(3):167-227.

33. Murakami Y, Tsuyama M, Kobayashi Y, Kodama H, Iba K: Trienoic fatty acids and plant tolerance of high temperature. *Science* 2000, 287(5452):476-479.

34. Sohn S, Back K: Transgenic rice tolerant to high temperature with elevated contents of dienoic fatty acids. *Biologia plantarum* 2007, 51(2):340-342.

35. Wang HS, Yu C, Tang XF, Wang LY, Dong XC, Meng QW: Antisense-mediated depletion of tomato endoplasmic reticulum omega-3 fatty acid desaturase enhances thermal tolerance. *J Integr Plant Biol* 2010, 52(6):568-577.

36. Upchurch RG: Fatty acid unsaturation, mobilization, and regulation in the response of plants to stress. *Biotechnol Lett* 2008, 30(6):967-977.

37. Murata N, Ishizaki-Nishizawa O, Higashi, S, Hayashi H, Tasaka Y: Genetically engineered alteration in the chilling sensitivity of plants. *Nature* 1992, 356:710–713.

38. Kim NH, Jayakodi M, Lee SC, Choi BS, Jang W, Lee J, et al.Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. P Biotec J 2017; (under review).

39. Jean-Marc R, Steven FF, John B: Trienoic Fatty Acids Are Required to Maintain Chloroplast Function at Low Temperatures. *Plant Physiol* 2000, 124(4):1697–1705.

40. Jang JC, Leon P, Zhou L, Sheen J: Hexokinase as a sugar sensor in higher plants. *Plant Cell* 1997, 9(1):5-19.

41. Miskell JH, Parmenter G, Eaton-Rye JJ: Photoperiodic changes of soluble sugar levels in *Panax ginseng*. *Science Access* 2001, 3(1).

42. Tholen D, Pons TL, Voesenek LA, Poorter H: Ethylene insensitivity results in down-regulation of rubisco expression and photosynthetic capacity in tobacco. *Plant Physiol* 2007, 144(3):1305-1315.

CHAPTER 3

Genome-wide screening of transcriptomes revealed the landscape of long noncoding RNAs in ginseng (*Panax ginseng*)

Abstract

Long noncoding RNAs (lncRNAs) have been implicated with diverse biological roles including genome regulation, various developmental processes and diseases. However, a comprehensive identification and functional prediction of lncRNAs in plants is still fall behind. Through a systematic pipeline using ~104 billion sequencing RNA reads from various tissues, stages of growth and abiotic stress treatments of *Panax ginseng*, I catalogued 19,495 lncRNAs and identified more than 100 lncRNAs involved in abiotic stress responses to drought, salt, cold, heat and methyl jasmonate (MeJA) and 2,607 involved in specialized unknown function in specific tissues and growth stages of *P. ginseng*. Gene Ontology enrichments for genes that showed expression correlation and RNA-RNA interaction with a set of lncRNAs participate in important biological processes including fatty acid desaturase activity, Photosystem II assembly and electron transport in PS I, response to cold and photoinhibition. Although lncRNAs are less conserved, I have identified two new well-conserved lncRNAs in *P. ginseng* and suggested their potential roles likely in mitochondrial protection and respiration chain. I have also proposed that transposable elements (TEs) might contribute in rapid diversification and origin of functional lncRNAs conferring tissue-specific or differential

expression against abiotic signals. This study provides new insights into the functional perspectives of lncRNAs in ginseng and plants in general.

Key words: long noncoding RNA, ginsenosides, transposable elements, *Panax ginseng*, lncRNA-mRNA interaction, single nucleotide polymorphism

Introduction

Recent advancements in high-throughput sequencing technologies have enabled us to decode the dark matter of the genomes and provided new insights into the genome and transcriptome of all organisms. Most of the eukaryotic genomes are actively transcribed into diverse ranges of protein coding and noncoding RNAs (ncRNAs) [1]. An increasing amount of evidences suggested that long noncoding RNAs (lncRNAs), a class of long RNA transcripts (>200 nucleotide) with no apparent protein coding potential [2, 3], are important regulators and implicated with distinct biological functions such as transcriptional regulation, genomic imprinting, histone modification and dosage compensation [4-8]. In addition, many lncRNAs have been associated with various diseases and recognized as potential therapeutic targets [9-12]. The expression of lncRNAs is often specific to a tissue or a developmental stage [13-16] indicating their unique role in organ developments. Due to such functional significance, the discovery and functional annotation of lncRNAs have begun to receive great attention over the past decade [17].

Majorly, lncRNAs were catalogued in mammals including mouse and humans [18] using different approaches including cDNA/EST, Chip-seq, tilling array and RNA-seq [19-21]. On the contrary, studies on lncRNAs in plants are still lagged [22]. The low expression, poor sequence conservation although some highly conserved and diverse functionality impede the standard identification of lncRNAs from genome sequences through computational methods [23]. To date, although a

genome-wide identification was performed in some plant species [24], a systematic approach was applied to only a few plant species, such as Arabidopsis [25, 26], maize [27] and rice [28]. Although many lncRNAs were being identified the functional implications of lncRNA is still in its infancy. Nevertheless, lncRNAs involved in the regulation of photoperiod-sensitive male sterility in rice [29], modulation of *Flowering Locus C* (*FLC*) expression and alternative splicing in arabidopsis [30, 31] have been major discoveries in plants. In addition, roles of lncRNAs in abiotic stress responses, reproductive developments and response to pathogen invasion were also reported [25, 32, 33]. A high-throughput and reproducible RNA sequencing (RNA-seq) is a powerful tool for profiling the complete transcriptomes of any species [34]. In plants, RNA-seq has accelerated the discovery of novel genes and functions in large and complex genomes through systematic expression studies. Therefore, expression profiling under various stress conditions and comparative expression analysis across various plant organs and growth stages of a plant is a promising approach to uncover candidate lncRNAs for further characterization of functions.

P. ginseng C. A. Meyer (2n = 4x = 48 chromosomes) is an obligate-shady medicinal herb and has been considered as a model reference genome for the family Araliaceae. In ginseng (referring major *Panax* species), ginsenosides are the major pharmacological compounds biosynthesized through isoprenoid pathway. The roots of this plant have been widely used in traditional medicines over thousands of years and have therapeutic effects of anti-cancer, diabetic, neuroprotective, antiamnestic and anti-stress [35-37]. Aside from pharmacological studies, other intriguing studies such as regulation of shady nature, long life span and environmental adaptation need to be largely explored in the perspectives of coding and noncoding RNAs to increase the production and availability for human use. Considering its therapeutic and economical importance, a large-scale genomic and transcriptomic data from various tissues, developmental stages and abiotic stress conditions have been generated to annotate the genes in *P. ginseng* as part of ginseng genome project. Ideally, such datasets can also be used for the systematic identification of lncRNAs,

global expression profiling, functional and mechanistic explorations of lncRNAs in *P. ginseng*.

In this study, a comprehensive set of lncRNAs were identified and characterized using ginseng draft genome sequence and RNA-seq datasets from 39 samples used for ginseng genome annotation. A total of 19,495 lncRNAs were identified and classified as intergenic, intronic and anti-sense lncRNAs according their genomic proximity. The dynamic expression profiling has detected the functional potential lncRNAs involved in abiotic stress, secondary metabolite biosynthesis, and environmental adaptation. Furthermore, well-conserved lncRNAs and SNPs in functional potential lncRNAs were also found. Our results provide a rich source of information to investigate functions of lncRNAs in *P. ginseng*.

Results

Genome-wide characterization of lncRNAs in *P. ginseng*

As illustrated in our pipeline (**Figure 3-1**), all the RNA-seq samples were mapped to the draft genome and assembled separately. 219,315 transcripts were yielded by merging all the separate assemblies using Cuffmerg. In total, we obtained 19,495 lncRNAs including 13,176 (67%) intergenic-, 4,704 (24%) intronic- and 1,612 (8%) antisense lncRNAs (**Figure 3-2** & **Figure 3-3A**). The size of lncRNAs ranged from 201 to 15,739 bp with an average length of 849 bp. The GC content was 35.84% which is lower than that of protein coding sequences of *P. ginseng* (43.38%). We further characterized the basic genomic features of lncRNAs in *P. ginseng*. The majority of the lncRNAs was less than 1 kb in size (**Figure 3-3B**) and consists of single exon (73%) (**Figure 3-3D**). Approximately, 47% of intergenic lncRNAs is located within 5 kb of protein coding genes (**Figure 3-3C**). Finally, we have added these lncRNA annotation features to the genome browser in ginseng genome database for visual exploration of noncoding section

Table 3-1. Details of RNA-seq data used for lncRNA prediction

	Samples	Replicates	Raw data size (Gbp)
Tissues	Leaves (1yr. old) (Control for heat treatment)	3	7.0
	Leaves (5yr. old)	1	3.9
	Main body root (1yr. old)	3	7.6
	Main body root (6yr. old)	3	7.4
	Lateral root (6yr. old)	3	8.2
	Rhizome (6yr. old)	3	8.0
	Adventitious root	1	9.1
	Dormant root (6 yr. old)	1	3.8
	Flower	1	3.3
	Shoot	1	1.7
	Immature fruit	1	3.6
	Mature fruit	1	3.2
	Seed (imbibed)	1	3.7
	Seed (stratified)	1	2.0
	Seedling (30 days)	1	3.7
Stress-treated	Controls (for cold, salt, drought treatment)	2	3.7
	Cold-treated	2	4.0
	Salt-treated	2	3.7
	Drought-treated	2	3.6
	Heat-treated for 1 week	3	5.8
	Heat-treated for 3 weeks	3	8.5

(http://ginsengdb.snu.ac.kr/Browser). Tissue-specific expression is one of the unique characteristic features of lncRNAs [16]. We therefore identified the lncRNAs that are specifically expressed in specific tissues and growth year of *P. ginseng* using RNA-seq datasets. Overall, 2,538 lncRNAs showed tissue specific expression between 10 tissues of different years of growth (**Figure 3-4**). Among them, intergenic type lncRNAs (65%) showed higher specificity of expression pattern followed by intronic (32%) and antisense lncRNAs (3%). Approximately, 2% of the total tissue-specific lncRNAs were found in lateral root tissues of six-year old roots.

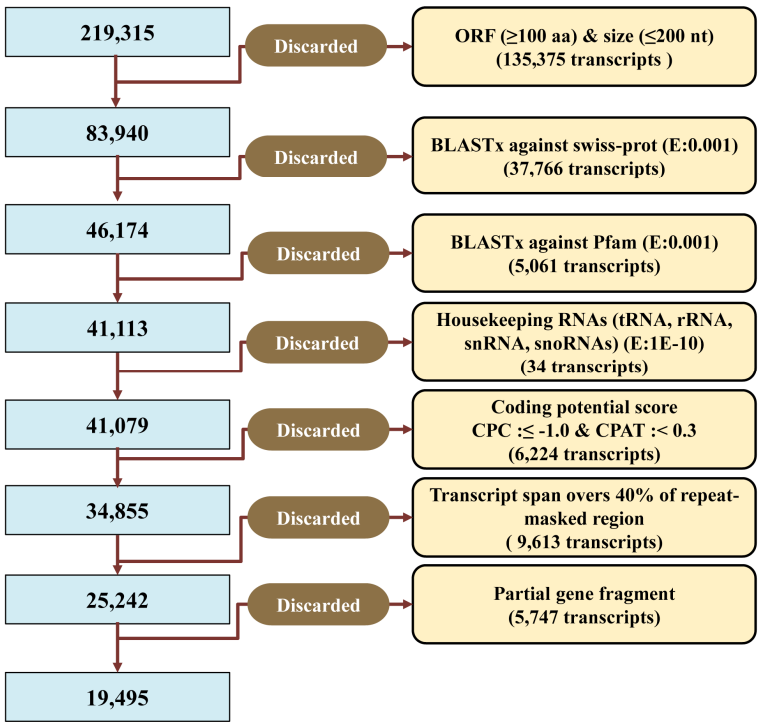

Figure 3-1. Schematic diagram of the informatics pipeline for lncRNA prediction. Step-wise removal of likely protein coding and spurious transcripts are shown for *P. ginseng*.

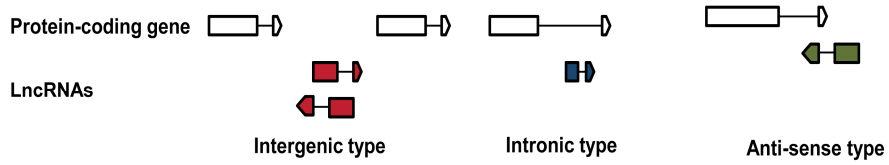

Figure 3-2. Classification of lncRNAs based on genomic locations. Intergenic lncRNAs are transcripts located from protein-coding genes at distances of more than 1~5 kb. Intronic lncRNAs are derived from the intron of a coding gene. Antisense lncRNAs are transcribed from the antisense strand and overlap with the protein-coding gene sequence.

Differential expression analysis

We used different sets of RNA-seq datasets to find out differentially expressed (DE) lncRNAs. First, a total of 41 DE lncRNAs were identified between a dataset of drought, salt and cold stress conditions. Second, 19 lncRNAs showed differential expression between one- and three-weeks heat stresses. Most of them showed increased expression pattern upon three-weeks of stress. Third, we used the methyl jasmonate (MeJA) treated adventitious root dataset from *P. ginseng* cultivar "Cheongsun (CS)" and "Sunhyang (SH)". From these datasets, 29 and 72 DE lncRNAs were identified in CS and SH respectively in response to MeJA. In total, 126 lncRNAs showed DE pattern wherein one lncRNA (TCONS_00104843) was found to be differentially regulated in all the above stress datasets (**Figure 3-5**). To confirm the results of identified lncRNA sets, we randomly selected four lncRNAs and subjected to quantitative RT-PCR to validate their expression pattern under drought, salt and cold conditions. As expected, the result was consistent with those obtained by RNA-seq data (**Figure 3-6**). Further, we grouped the DE lncRNAs on the basis of its genomic proximity classification and found that intergenic lncRNAs were majorly responded against abiotic stresses as comparted to intronic and antisense lncRNAs in *P. ginseng*.

Co-expression and target interaction analysis

To find the plausible function of lncRNAs, we calculated the Pearson's correlation coefficient (PCC) between lncRNAs and protein coding genes. 58 lncRNAs showed high co-expression with 700 protein coding genes. Then, I performed Gene Ontology (GO) enrichment analysis for those co-expressed genes. Notably, the co-expressed genes were enriched for the biological process related to photosynthesis, response to red, blue and far-red light, photoinhibition, jasmonic acid signaling and response to cold. Furthermore, we identified intergenic lncRNAs that are co-expressed with key

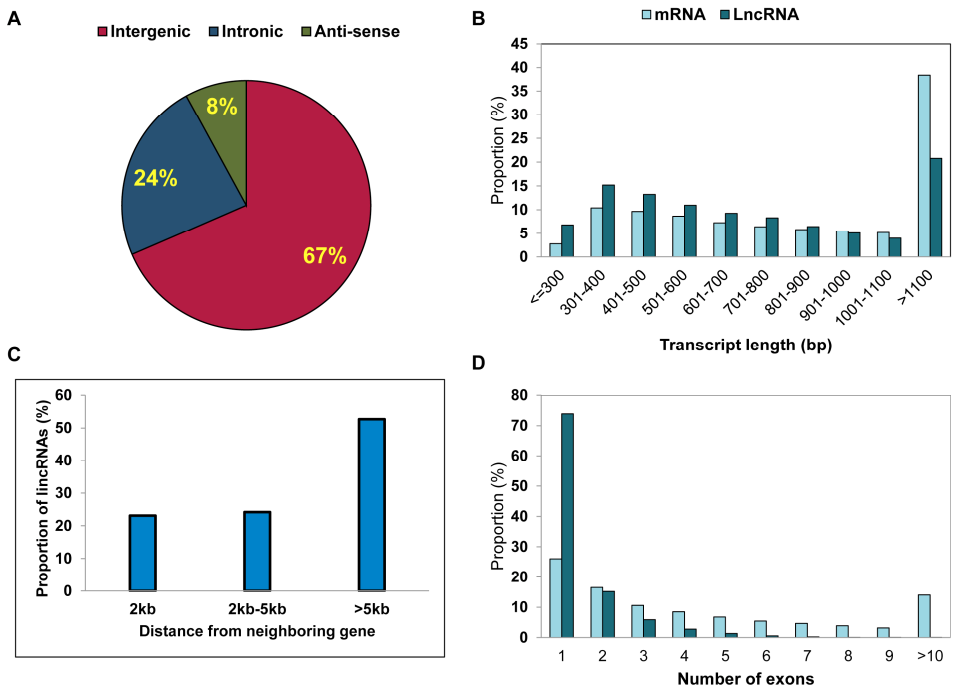

Figure 3-3. Classification and characteristics of lncRNAs. A). Percentage of lncRNAs identified in intergenic, intronic and antisense type. B). Length distribution of lncRNA and mRNAs (IPGA v1.1). C). Proportion of lncRNAs located within 2 kb, between 2 kb-5 kb and over 5 kb from nearest coding gene. D). Distribution of exon number in lncRNA and mRNAs.

candidate enzymes involved in ginsenoside biosynthesis pathway such as dammarenediol synthase (DDS: Pg_S3318.3), squalene epoxidase (SQE: Pg_S3064.5, Pg_S6308.10). In addition, we investigated RNA-RNA interaction between co-expressed protein coding genes and lncRNAs through minimum free energy joint structure of two RNA molecules by base pairing. Interestingly, except one pair (TCONS_00159594 : Pg_S5778.11), all the co-expressed gene and lncRNA pairs showed high interaction pattern suggesting those genes can be the potential targets of lncRNAs. In order to confirm this method of functional prediction, we selected a co-expressed pairs of lncRNA (TCONS_00216176) and gene (Pg_S8040.1) where the lncRNA is antisense type

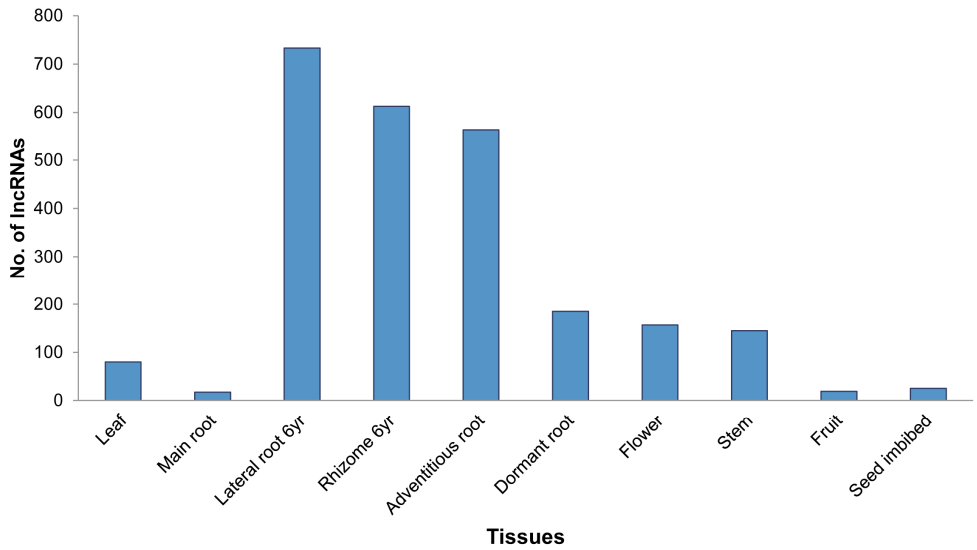

Figure 3-4. Tissue-specific expression of lncRNAs. Barplot showing number of lncRNAs in *P. ginseng* showing tissue specificity.

and showed differential expression in response to drought, salt, cold and MeJA treatments. We then subjected this pair to strand-specific quantitative RT-PCR and checked the expression pattern in 12 h, 24 h and 48 h MeJA treated adventitious roots from CS cultivar. As we identified through PCC and DE, both antisense

lncRNA and gene showed similar co-expression pattern as well as expression changes in response to MeJA conditions.

Conservation analysis

From the total lncRNAs, we found that 175, 45, 401, 197 lncRNAs were conserved in the genomes of tomato, arabidopsis, carrot and grapes respectively.

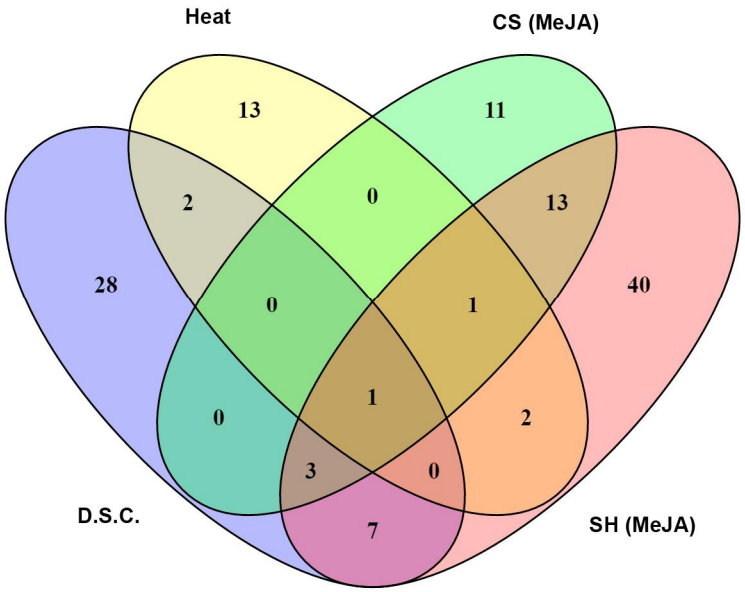

Figure 3-5. Distribution of number of differentially expressed (DE) lncRNAs. Venn diagram showing the number of intersecting DE lncRNAs in response to DSC (Drought, Salt, Cold), heat, and 12 h, 24 h and 48 h methyl jasmonate (MeJA) treated adventitious roots of Cheongsun (CS) cultivars and 12 h, 24 h MeJA treated adventitious roots in Sunhyang (SH) cultivar.

Similarly, 28 lncRNAs had homology with lncRNAs in NONCODE database. Intriguingly, we found two intergenic lncRNAs (TCONS_00046304, TCONS_00160350) that showed complete query coverage with over 80-90% sequence identity among the plant genomes (**Figure 3-7**). We also observed that these two lncRNAs were also matched with the mitochondrial genome of *P. ginseng*

as well as other plant species. These two lncRNAs have not been found in the list of differential expressed lncRNAs under abiotic stress and MeJA treatments in this study. On the other hand, we observed different level of expression pattern such as low, moderate and high expression among normal developmental tissues. Further, we investigated the expression in drought, salt and cold stress conditions through quantitative RT-PCR and found similar pattern obtained from RNA-seq data. Functional prediction of those genes by adjusted co-expression analysis revealed that they presumably associated with mito virus RNA-dependent RNA polymerase activity and electron transport activity in mitochondria respiration chain.

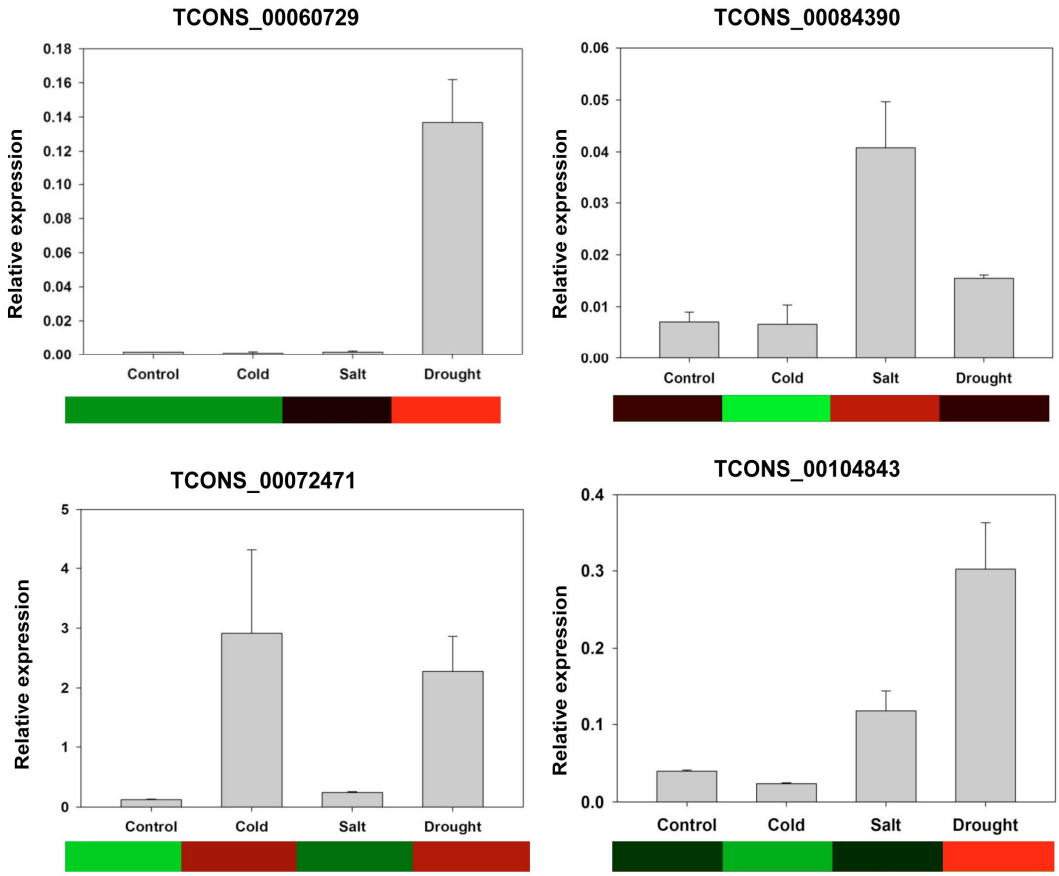

Figure 3-6. Differential expression of abiotic stress responsive lncRNAs. Expression confirmation of selected lncRNAs in drought, salt and cold conditions

by quantitative RT-PCR. A heatmap of each lncRNA was generated from normalized FPKM values to visualize the lncRNA expression pattern in corresponding stress RNA-seq data.

Single nucleotide polymorphism (SNPs) analysis between *Panax* species

We first predicted SNPs by comparison of *P. ginseng* draft genome which was made from Chunpoong (CP) cultivar with genomic reads of another cultivar Yunpoong (YP). A total of 32,970 SNPs were identified in the regions of lncRNAs in draft genome of *P. ginseng*. Among them, we observed 724 SNPs in the

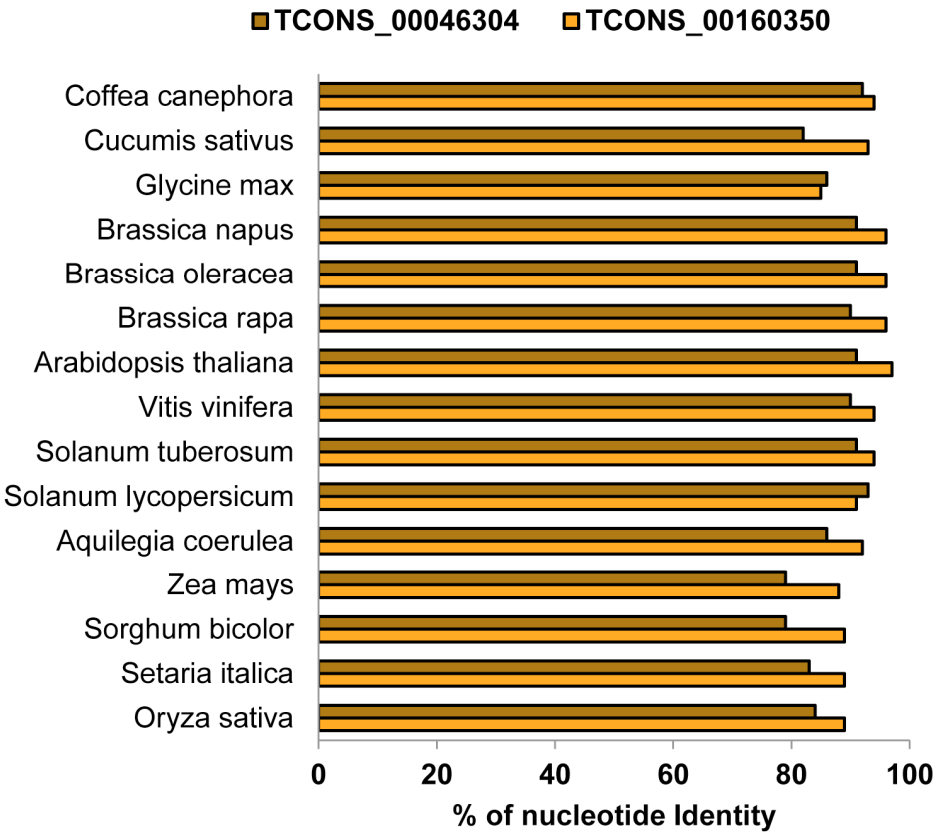

Figure 3-7. Percentage of nucleotide identity of two well-conserved lncRNA matched with other plant species.

functional potential lncRNA regions that were identified from differential and co-expression analysis. Similarly, we called SNPs between *P. ginseng* and *P. vietnamensis*. This yielded a total of 539,791 SNPs in the regions of lncRNAs wherein 5,408 SNPs in the regions of functional potential lncRNAs. In additions, we investigated SNPs in genes that are co-expressed with lncRNAs. We found that 3,792 and 48,544 SNPs in potential target genes between Yunpoong and *P. vietnamensis* respectively.

Functional repertoire of lncRNAs derived from transposons

We then investigated DE lncRNAs manually in the ginseng genome browser. Intriguingly, we observed that the flanking regions of lncRNAs were densely occupied by repetitive elements or lncRNAs were overlapped with long terminal repeat (LTR) and repeat masked regions. In turn, we selected those 126 DE lncRNAs and masked the repetitive elements by CENSOR (http://www.girinst.org/censor/index.php). 96 (76%) of total DE lncRNAs were found to be transposable element (TE) derived lncRNAs. Similarly, we selected tissue specific lncRNAs with FPKM $\geq$ 3 which we assumed that such high confident candidates enable us to make reliable inferences, for repeat masking. 100 (72%) of 137 high confident tissue-specific lncRNAs were identified to be TE-derived lncRNAs.

Discussion

Emerging evidences have demonstrated that a study on lncRNA is inevitable among eukaryotic research communities due to their diverse functionalities. *P. ginseng* is one of the well acclaimed medicinal plants in Asia and has great pharmacological values on human's health. The availability of draft genome sequence (Kim et al., 2017; Jayakodi et al., 2017)) and extensive transcriptome resources of *P. ginseng* provides a unique opportunity to perform systematic lncRNA research. Previously, small-scale efforts have been made for noncoding RNA studies in *P. ginseng* such as *de novo* micro RNA [38] (miRNA) and mRNA

like noncoding RNA [39] prediction. However, lncRNAs have a characteristic of spatial and temporal expression pattern [40] and thus, it is necessary to have transcriptome data from various tissues, growth stages and a high quality reference genome to profile lncRNAs more comprehensively. In this study, we used transcriptome data from 39 RNA-seq experiments including various tissues, stages and treatments (**Table 3-1**) and the size of high-quality RNA-seq reads (104 billion paired-reads) used for prediction is relatively higher than that used in other model plant species for lncRNA studies [25, 27, 28] to date. In transcriptome assembly, low expressed or insufficient sequencing reads for certain genes and high similar paralogs copies might produce partial and spurious transcriptome assemblies. Subsequently, many of those fragments were possibly annotated as a class of lncRNAs. Therefore, we carefully designed our lncRNA prediction pipeline that not only discarded likely protein coding transcripts from transcriptome assembly but also repetitive/transposable elements, mapping errors and partial gene fragments to excerpt decisive lncRNAs (**Figure 3-1**). Mostly, strand-specific RNA-seq data were being used to catalogue antisense type lncRNAs [26]. Nonetheless, the availability of draft genome sequence, huge RNA-seq datasets and interactive genome browser has enabled us to catalogue antisense lncRNAs in *P. ginseng*. We identified more than 15,000 reliable set of lncRNAs for accelerating functional genomic research in *P. ginseng*. The characteristic features of lncRNAs in *P. ginseng* were similar to other plant species. For instance, the number of lncRNAs identified in this study and the proportion of lincRNAs located over 5 kb (**Figure 3-3C**) apart from protein-coding genes are relatively similar to lncRNAs predicted in maize [27]. Majorly, lncRNAs in *P. ginseng* were less than 1 kb in size (**Figure 3-3B**) and consisted of a single exon (**Figure 3-3D**), which are also similar to characterization of lncRNAs in other model plant species [27, 28, 41].

Unlike mRNA, putative functional prediction has been a major hindrance in lncRNAs studies. Nevertheless, expression profiling is a robust approach to identify functional potential lncRNAs in eukaryotic species. Mostly, tissue-specific and low-expression pattern were commonly observed for lncRNAs in different plants

including Arabidopsis, maize, rice and cucumber [25, 27, 28, 42]. In *P. ginseng*, we also identified a set of lncRNAs (**Figure 3-4**) that likely play specialized function in specific organs. In addition, many studies have used differential expression analysis to detect candidate lncRNAs associated with biotic/abiotic stress responses [24]. A total of 126 abiotic stress (**Figure 3-5**) responsive lncRNAs were identified in *P. ginseng* and ascertained the accuracy of our approach by quantitative RT-PCRs (**Figure 3-6**). This indicates that lncRNAs could also be one of the important regulators like transcription factors and miRNAs involved in environmental adaptations. Notably, one lncRNA (TCONS_00104843) was found to be differentially regulated in all the above stress conditions. This provided an opportunity where a single lncRNA manipulation might bring desirable effects to a plant against multiple stresses. In comparison, the total number of DE lncRNAs were relatively low which could be attributed to the low number sequencing reads generated for stress-treated samples or suspected that the majority of lncRNAs in ginseng might be expressed constitutively. In addition, I also expect that the lncRNAs that might induce or repress in early stage might also be missed due to a long period of stress treatments. The co-expression analysis has also been widely used to predict functions of lncRNAs. We have also performed co-expression analysis to target the functional potential lncRNAs in *P. ginseng*. In addition, we investigated the interaction between co-expressed lncRNAs and mRNAs to gain further conviction on this analysis. In *P. ginseng*, some of the most important mechanisms including cold and shade adaptation, photoinhibition are poorly understood. The genes co-expressed with lncRNAs were enriched for such biological process, indicating the significant involvements of lncRNAs in physiological and ecological responses of *P. ginseng*. Furthermore, many single lncRNA showed expression correlation and RNA level interaction with multiple genes in *P. ginseng*. Hence, like miRNAs [43], lncRNAs might also have a property of regulating multiple genes from a single locus. Notably, all the co-expressed lncRNAs were found in the list of DE lncRNAs in this study, demonstrating that either of those approaches or combining them is appropriate to find functional

potential lncRNAs in any genomes. Ginsenosides are the major targets of research in *P. ginseng* and has been found that MeJA stimulates the biosynthesis of many secondary metabolites including ginsenosides [44]. We have identified 18 lncRNAs (**Figure 3-5**) that showed differential expression pattern upon MeJA treatment in both SU and SH adventitious roots. In addition, we found lncRNAs were co-expressed with key candidate genes involved in ginsenoside biosynthesis. Taken together, lncRNAs may also be the key player in various metabolic pathways including ginsenosides biosynthesis in *P. ginseng*. Among classes of lncRNAs, long intervening/intergenic RNAs (lincRNA) were implicated with diverse functional roles in different eukaryotic species [45]. In our results, lincRNAs showed high tissue-specificity and differential expression against abiotic stresses indicating the necessity of cataloguing lincRNAs among the classes of lncRNAs for functional biological research.

Recent genome-wide studies have revealed that lncRNAs showed a low degree of evolutionary constraint in both mammals and plants [46]. In our study, we also observed very few conserved lncRNAs against tomato, Arabidopsis, carrot, and grapes genomes through systematic BLAST search. However, reports have also stated that over thousands of lncRNAs being evolutionarily conserved in mammals [47]. But, such conserved lncRNAs have not been documented in plants. In this study, we have identified two well-conserved lncRNAs that showed over 80% nucleotide identity between model plant species (**Figure 3-7**). Since these two lncRNA fragments similar with regions in mitochondria genome of *P. ginseng* and other species, we speculate that these two lncRNA fragments might have derived from mitochondrial genomes of plants during the process of evolution. The expression profiling clearly suggested that these two lncRNAs may play cellular house-keeping functions in mitochondrial compartment. Our functional prediction also indicated that they might regulate mitovirus RNA-dependent RNA polymerase activity and cellular respiration. Therefore, this study paves the way for further functional exploration of these well-conserved lncRNAs to gain more insights into regulation of energy metabolism and protection on mitochondria of *P. ginseng*.

The phenotypic variation caused by SNPs in lncRNAs has not been well studied in plants. As to accelerate such studies, we have identified SNPs in lncRNAs and compared their proportions between cultivars of *P. ginseng* and *Panax* species to extend our understandings. The functional potential of lncRNAs has also been determined by the structural features [29] and thereby SNPs in lncRNAs can lead to altered the binding potential with targets and transcriptional efficiency. We have also found SNPs in the functional potential lncRNAs in *P. ginseng* which might have functional variations between cultivars and *Panax* species. In addition, the disruption of structural motifs in lncRNAs might also affect the underlying regulatory mechanism between lncRNAs and target protein coding genes and thus can also lead to variations in phenotype. Furthermore, SNPs in lncRNA and their potential target genes identified in this study indicated that diversification of lncRNAs occurs rapidly even within *Panax* species. This has driven us to explore the reason for such a rapid diversification of lncRNAs in ginseng. The rapid proliferation of TEs creates genetic variability that allows plants to adapt in adverse environmental stresses [48]. In vertebrates, TEs were the major contributors for origin, diversification and regulation of lncRNAs [49]. Furthermore, the concept of Repeat Insertion Domains of LncRNAs (RIDLs) explained that TE-derived fragments of lncRNAs could interact with highly relevant regulatory proteins or specific genes and RNAs [50]. In this study, large proportion of functional lncRNAs including DE and high confident tissue-specific lncRNAs were found to be TE-derived lncRNAs. In addition, those lncRNAs were also flanked by highly repeat rich genomic regions. These indicate that TEs are the major contributors for the origin of regulatory lncRNAs in ginseng. Furthermore, we also attribute TEs to lncRNAs unique features including evolutionary less conservation or rapid evolution and specificity including tissue- or species- or lineage-specific lncRNAs in ginseng based on TE-derived high confident tissue-specific lncRNAs. In consequence, near future, a vast amount of species-specific lncRNAs may also be expected between *Panax* as well as plant species. Therefore, these data provided relevant resources to unravel mechanisms through which TEs embedded in

lncRNAs and become major players in regulating the expression of ginseng genomes.

Materials and method

Datasets used for lncRNA predictions

High-throughput Illumina RNA-seq data generated for *P. ginseng* cultivar "Chunpoong (CP)" were used, which were originally generated for *P. ginseng* genome annotation from the ginseng genome database (http://ginsengdb.snu.ac.kr/). This included 39 experiments from discrete tissues (leaves and roots of one-year old plants, main-body root, lateral root, and rhizome of six-year old plants, dormant root, flower, immature and mature fruits, stem, 30 days old seedling, adventitious root, stratified and imbibed seeds) and abiotic stress treatments (drought: air-drying for 24 h, salt: 100 mM NaCl for 24 h, cold: 4℃ for 24 h and heat: 30℃ for 1 week and 3 weeks) of *P. ginseng* cultivar CP. A total of 104 billion paired-raw reads with the read lengths ranging from 101 to 150 nucleotides were obtained (**Table 3-1**). Prior to read-mapping to genome, all the raw reads were subjected to the following filtration process. Firstly, the bacterial contaminant reads were removed by aligning against all published bacterial genome. Then, the duplicated and ribosomal contaminant reads were identified using FastUniq [51] and sormerna [52] respectively and discarded. Finally, the low-quality bases were trimmed using NGS-QC toolkit [53]. The draft genome sequences (v1) and gene annotation (IPGA_v1.1) were downloaded from ginseng genome database and used.

Pipeline for lncRNA identification

All the filtered RNA-seq data were mapped to the draft genome sequence of *P. ginseng* using the spliced aligner HISAT[54] (with parameters --no-mixed, --no-discordant, --max-intronlen 10000) and then assembled using stringtie [55]. An in-house Perl script was used to filter out the assembled transcripts that contain ≥200 bp in length and ORFs of ≤100 amino acids. Then, the protein coding transcripts

were eliminated by BLASTX search against Swiss-prot and Pfam protein databases with an *E-value* cutoff of 1E-03. Further, to discard non lncRNAs including tRNAs (http://gtrnadb.ucsc.edu/), rRNAs (http://www.arb-silva.de/no_cache/download/archive/current/Exports/) and other class of ncRNAs (snRNAs, snoRNAs, 7SL/SRP) (http://noncode.org/), a housekeeping noncoding RNA database was established and used to filter transcripts by alignment using Blast with E-value of 1E-10. The remaining filtered transcripts were subjected to assess the coding potential by CPC [56] (score of ≤-1.0) and CPAT [57] (score of <0.3). In addition, transcript assembly by error mapping and partial gene transcripts by low-sequencing depth have been identified and discarded. Finally, the filtered transcripts were compared with ginseng genome annotation by cuffcompare program in the cufflinks and retained the transcripts containing class codes of 'u', 'i' and 'x' which represents novel intergenic, intronic and antisense transcripts respectively.

Characterization of lncRNAs

The identified lncRNAs including intergenic, intronic and anti-sense were compared with *P. ginseng* genome annotation v1.1. To find out the conservation, all the lncRNAs sequences were aligned against the genome sequence of tomato (ITAG2.5), carrot (GenBank: LNRQ01000001-LNRQ01004826), grape (v2) and Arabidopsis (TAIR10) with a E-value cutoff of 1E-05. The lncRNAs that showed >20% query coverage were defined as conserved lncRNAs. BLAST search (E-value <1E-5) was also performed against the know lncRNAs in the NONCODE v4 database (http://noncode.org/). The intergenic lncRNAs located within 2 kb and 2 kb -5 kb of upstream or downstream of genes was identified using an in house Python script. The repetitive element contents were annotated using Repeat Masker [58].

Expression profiling

Differentially expressed lncRNAs were identified between replicated datasets of drought, salt, cold stress and between 1- and 3-week heat stress treated CP cultivars. In addition, DE analysis was also performed using RNA-seq data from methyl jasmonate (MeJA) treated adventitious roots of other *P. ginseng* cultivars [59, 60]. This included 12 h, 24 h and 48 h MeJA treatment in CS cultivar and 12 h, 24 h MeJA treatment in SH cultivar. Expression levels of all lncRNAs in each sample were determined using fragments per kilobase of transcript per million fragments (FPKM) by RSEM [61]. Further, the expression data were normalized using Trimmed Mean of M values (TMM) [62] due to the differences in sequencing platforms as well as sequencing libraries. The differentially expressed lncRNAs were identified using edgeR [63] bioconductor package with over two-folds and false discovery rate (FDR) adjusted *p*-value of 0.01. The tissue specificity (T*s*) of an lncRNA in tissue *s* was calculated as the fraction of expression (in FPKM) relative to the sum of its expression in all the stages of tissues.

Co-expression analysis

The gene expression data (39 RNA-seq experiments in CP cultivar) generated for *P. ginseng* genome study was selected for co-expression analysis. All the protein coding genes and differentially expressed lncRNAs were used for co-expression analysis using Pearson's correlation coefficient ($|r| \geq 0.9$). The Gene Ontology (GO) enrichment analysis was performed for co-expressed genes using Fisher's Exact Test with multiple testing correction of FDR with cutoff 0.05 [64]. The interaction between co-expressed lncRNAs and corresponding genes were predicted using LncTar [65].

SNP calling

SNP calling was performed with ~20X resequencing data generated from *P. ginseng* cultivar Yunpoong and a diploid ginseng species *P. vietnamensis* by Illumina Hiseq (2 x 101 bp) and Nextseq (2 x 150 bp) platform respectively. Those data sets were retrieved from Ginseng Genome Database. The raw data were

cleaned by quality control and removing low-quality bases. Illumina paired-end reads from each dataset were aligned to the reference draft genome of *P. ginseng* using BWA [66] program separately. To increase the accuracy, uniquely mapped alignments were selected for further processing. Then, the alignments were grouped as consensus alignment to remove duplicated alignments and fix the mate-pair information using Picard (http://picard.sourceforge.net). In addition, Genome Analysis Toolkit (GATK) [67] was used to investigate the misalignments caused by INDELs and to call candidate SNPs. Finally, BEDTools [68] were used to identify SNPs in lncRNA regions.

RNA extraction and quantitative RT-PCR

Total RNAs were isolated from *P. ginseng* cultivar CP plants treated with various abiotic stresses using Hybrid-R kit (GeneAll, Korea), following the manufacturer's protocol. RNA quality and quantity were examined using a spectrophotometer and formaldehyde agarose gel electrophoresis. For quantitative reverse transcription-polymerase chain reaction (qRT-PCR) 1 µg RNA was used for first-strand cDNA synthesis by reverse transcription using SuperScript II Reverse Transcriptase and random hexamers (poly T) (Invitrogen, USA). The cDNA synthesis performed using following procedure: 37 °C for 20 min and stored at -20 °C until use for RT-PCR. mRNA expression of each lncRNA was done by qRT-PCR analysis using lncRNA specific primers designed using NCBI primer BLAST tool. *Pg*Actin was used as a control for all the conditions. qRT-PCR was performed using SYBR Green Power PCR Master Mix (Applied Biosystems, Foster City, CA) in the Roche lightcycler 480 realtime PCR system (Roche, Indianapolis, IN). PCR was conducted in 20 ul total reaction mixture containing 10ng cDNA, 10 pmol of each forward and reverse primer, 250 uM dNTPs, 1 unit of Taq DNA polymerase (VIVAGEN, South Korea) and 1ul of SYBR green. The thermal cycling condition for PCR was 95°C for 5 min, 30 cycles of 95°C for 5 s, 60°C for 30 s and 72°C for 30 s, followed by 72°C for 5 min. The cycles passing threshold (Ct) were recorded

and the expression of *Pg*Actin was used as an internal control. qRT-PCR was carried out in duplicate for each sample.

Validation of antisense lncRNA by strand-specific quantitative RT-PCR

Expression levels of non-coding RNA were analyzed by quantitative real-time PCR (qRT-PCR) in CS adventitious roots collected at 0, 12, 24, 48hr following methyl jasmonate (MeJA) treatment. Total RNA was extracted from the adventitious roots were RNeasy Plant kit (Qiagen, Germany) and quantified into equal concentrations. Antisense strand-specific cDNA was synthesized with antisense-specific primer using SuperScript II Reverse Transcriptase (Invitrogen™ Life Technology, USA) according to the manufacturer's protocol. Total cDNA and sense strand-specific cDNA as negative controls were synthesized with oligo-dT and sense-specific primer, respectively. For qRT-PCR reaction, an equal volume of cDNA was added to PCR reaction and amplified with secondary nested primer by qRT-PCR using a Light cycler 480 (Roche, Mannheim, Germany). The thermal cycling conditions were as follows: 95 °C for 5 min and 40 cycles of 95 °C for 15 s, 58 °C for 10 s, and 72 °C for 10 s. Antisense strand-specific amplification of non-coding RNA was confirmed by that the negative control reactions performed with total cDNA and sense strand-specific cDNA produced no amplicon..

REFERENCES

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al*: Landscape of transcription in human cells. *Nature* 2012, 489(7414):101-108.

2. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG *et al*: The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012, 22(9):1775-1789.

3. Carninci P, Kasukawa T, Katayama S, Gough J, Frith M, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C: The transcriptional landscape of the mammalian genome. *Science* 2005, 309(5740):1559-1563.

4. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL *et al*: Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010, 464(7291):1071-1076.

5. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY: Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010, 329(5992):689-693.

6. Wang X, Arai S, Song X, Reichart D, Du K, Pascual G, Tempst P, Rosenfeld MG, Glass CK, Kurokawa R: Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* 2008, 454(7200):126-130.

7. Wilusz JE, Sunwoo H, Spector DL: Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* 2009, 23(13):1494-1504.

8. Chow JC, Yen Z, Ziesche SM, Brown CJ: Silencing of the mammalian X chromosome. *Annu Rev Genomics Hum Genet* 2005, 6:69-92.

9. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, St Laurent G 3rd, Kenny PJ, Wahlestedt C: Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med* 2008, 14(7):723-730.

10. Saayman S, Ackley A, Turner AM, Famiglietti M, Bosque A, Clemson M, Planelles V, Morris KV: An HIV-encoded antisense long noncoding RNA epigenetically regulates viral transcription. *Mol Ther* 2014, 22(6):1164-1175.

11. Chen H, Xu J, Hong J, Tang R, Zhang X, Fang JY: Long noncoding RNA profiles identify five distinct molecular subtypes of colorectal cancer with clinical relevance. *Mol Oncol* 2014, 8(8):1393-1403.

12. Crea F, Watahiki A, Quagliata L, Xue H, Pikor L, Parolia A, Wang Y, Lin D, Lam WL, Farrar WL *et al*: Identification of a long non-coding RNA as a novel biomarker and potential therapeutic target for metastatic prostate cancer. *Oncotarget* 2014, 5(3):764-774.

13. Yang X, Gao L, Guo X, Shi X, Wu H, Song F, Wang B: A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PloS one* 2014, 9(1):e87797.

14. Clark BS, Blackshaw S: Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease. *Front Genet* 2014, 5:164.

15. Paralkar VR, Mishra T, Luan J, Yao Y, Kossenkov AV, Anderson SM, Dunagin M, Pimkin M, Gore M, Sun D *et al*: Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. *Blood* 2014, 123(12):1927-1937.

16. Tsoi LC, Iyer MK, Stuart PE, Swindell WR, Gudjonsson JE, Tejasvi T, Sarkar MK, Li B, Ding J, Voorhees JJ: Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. *Genome Biol* 2015, 16(1):24.

17. Mattick JS, Rinn JL: Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol* 2015, 22(1):5-7.

18. Zhao Y, Li H, Fang S, Kang Y, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, Chen R: NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res* 2016, 44(D1):D203-D208.

19. Li L, Wang X, Stolc V, Li X, Zhang D, Su N, Tongprasit W, Li S, Cheng Z, Wang J *et al*: Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet* 2006, 38(1):124-129.

20. Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engstrom PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW *et al*: Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* 2006, 2(4):e62.

21. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008, 5(7):621-628.

22. Liu X, Hao L, Li D, Zhu L, Hu S: Long non-coding RNAs and their biological roles in plants. *Genomics Proteomics Bioinformatics* 2015, 13(3):137-147.

23. Sacco LD, Baldassarre A, Masotti A: Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis. *Int J Mol Sci* 2011, 13(1):97-114.

24. Shafiq S, Li J, Sun Q: Functions of plants long non-coding RNAs. *Biochim Biophys Acta* 2016, 1859(1):155-162.

25. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua N-H: Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell* 2012, 24(11):4333-4345.

26. Wang H, Chung PJ, Liu J, Jang IC, Kean MJ, Xu J, Chua N-H: Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome Res* 2014, 24(3):444-453.

27. Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE *et al*: Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol* 2014, 15(2):R40.

28. Zhang YC, Liao JY, Li ZY, Yu Y, Zhang JP, Li QF, Qu LH, Shu WS, Chen YQ: Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biol* 2014, 15(12):512.

29. Ding J, Lu Q, Ouyang Y, Mao H, Zhang P, Yao J, Xu C, Li X, Xiao J, Zhang Q: A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proce Natl Acad Sci* 2012, 109(7):2654-2659.

30. Bardou F, Ariel F, Simpson CG, Romero-Barrios N, Laporte P, Balzergue S, Brown JW, Crespi M: Long noncoding RNA modulates alternative splicing regulators in Arabidopsis. *Dev Cell* 2014, 30(2):166-176.

31. Swiezewski S, Liu F, Magusin A, Dean C: Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* 2009, 462(7274):799-802.

32. Kim ED, Sung S: Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends Plant Sci* 2012, 17(1):16-21.

33. Xin M, Wang Y, Yao Y, Song N, Hu Z, Qin D, Xie C, Peng H, Ni Z, Sun Q: 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* 2011, 11:61.

34. Jain M: Next-generation sequencing technologies for gene expression profiling in plants. *Brief Funct Genomics* 2012, 11(1):63-70.

35. Saito H, Yoshida Y, Takagi K: Effect of Panax Ginseng root on exhaustive exercise in mice. *Jpn J Pharmacol* 1974, 24(1):119-127.

36. Attele AS, Wu JA, Yuan CS: Ginseng pharmacology: multiple constituents and multiple actions. *Biochem Pharmacol* 1999, 58(11):1685-1693.

37. Shang W, Yang Y, Zhou L, Jiang B, Jin H, Chen M: Ginsenoside Rb1 stimulates glucose uptake through insulin-like signaling pathway in 3T3-L1 adipocytes. *J Endocrinol* 2008, 198(3):561-569.

38. Mathiyalagan R, Subramaniyam S, Natarajan S, Kim YJ, Sun MS, Kim SY, Kim Y-J, Yang DC: Insilico profiling of microRNAs in Korean ginseng (*Panax ginseng* Meyer). *J Ginseng Res* 2013, 37(2):227-247.

39. Wang M, Wu B, Chen C, Lu S: Identification of mRNA-like non-coding RNAs and validation of a mighty one named MAR in *Panax ginseng*. *J Integr Plant Biol* 2015, 57(3):256-270.

40. Fatica A, Bozzoni I: Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* 2014, 15(1):7-21.

41. Khemka N, Singh VK, Garg R, Jain M: Genome-wide analysis of long intergenic non-coding RNAs in chickpea and their potential role in flower development. *Sci Rep* 2016, 6:33297.

42. Hao Z, Fan C, Cheng T, Su Y, Wei Q, Li G: Genome-wide identification, characterization and evolutionary analysis of long intergenic noncoding RNAs in cucumber. *PloS one* 2015, 10(3):e0121800.

43. Pasquinelli AE: MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat Rev Genet* 2012, 13(4):271-282.

44. Lu M, Wong H, Teng W: Effects of elicitation on the production of saponin in cell culture of Panax ginseng. *Plant Cell Rep* 2001, 20(7):674-677.

45. Ulitsky I, Bartel DP: lincRNAs: genomics, evolution, and mechanisms. *Cell* 2013, 154(1):26-46.

46. Johnsson P, Lipovich L, Grandér D, Morris KV: Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta* 2014, 1840(3):1063-1071.

47. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP *et al*: Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009, 458(7235):223-227.

48. Sharma M, Pandey GK: Expansion and function of repeat domain proteins during stress and development in plants. *Front Plant Sci* 2015, 6:1218.

49. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C: Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* 2013, 9(4):e1003470.

50. Johnson R, Guigó R: The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* 2014, 20(7):959-976.

51. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S: FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* 2012, 7(12):e52249.

52. Kopylova E, Noé L, Touzet H: SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012, 28(24):3211-3217.

53. Patel RK, Jain M: NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one* 2012, 7(2):e30619.

54. Kim D, Langmead B, Salzberg SL: HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015, 12(4):357-360.

55. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL: StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015, 33(3):290-295.

56. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G: CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 2007, 35(Web Server issue):W345-349.

57. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W: CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 2013, 41(6):e74-e74.

58. Smit AF, Hubley R, Green P: RepeatMasker Open-3.0. In.; 1996.

59. Lee Y, Park HS, Lee DK, Jayakodi M, Kim NH, Koo HJ, Lee SC, Kim YJ, Kwon SW, Yang TJ: Integrated transcriptomic and metabolomic analysis of five *Panax ginseng* cultivars reveals the dynamics of ginsenoside biosynthesis. *Front Plant Sci* 2017.

60. Lee YS, Park HS, Lee DK, Jayakodi M, Kim NH, Lee SC, Kundu A, Lee DY, Kim YC, In JG: Comparative analysis of the transcriptomes and primary metabolite profiles of adventitious roots of five Panax ginseng cultivars. *J Ginseng Res* 2017, 41(1):60-68.

61. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011, 12:323.

62. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013, 14(6):671-683.

63. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26(1):139-140.

64. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21(18):3674-3676.

65. Li J, Ma W, Zeng P, Wang J, Geng B, Yang J, Cui Q: LncTar: a tool for predicting the RNA targets of long noncoding RNAs. *Brief Bioinform* 2015, 16(5):806-812.

66. Li H, Durbin R: Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009, 25(14):1754-1760.

67. Nekrutenko A, Taylor J: Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet* 2012, 13(9):667-672.

68. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010, 26(6):841-842.

CHAPTER 4

Ginseng Genome Database: An open-access platform for genomics of *Panax ginseng*

Abstract

The ginseng (*Panax ginseng* C.A. Meyer) is a perennial herbaceous plant that has been used in traditional oriental medicine for thousands of years. Ginsenosides, which have significant pharmacological effects on human health, are the foremost bioactive constituents in this plant. Having realized the importance of this plant to humans, an integrated omics resource becomes indispensable to facilitate genomic research, molecular breeding and pharmacological study of this herb. The first draft genome sequences of *P. ginseng* cultivar "Chunpoong" were reported recently. Here, using the draft genome, transcriptome, and functional annotation datasets of *P. ginseng*, we have constructed the Ginseng Genome Database http://ginsengdb.snu.ac.kr/, the first open-access platform to provide comprehensive genomic resources of *P. ginseng*. The current version of this database provides the most up-to-date draft genome sequence (of approximately 3,000 Mbp of scaffold sequences) along with the structural and functional annotations for 59,352 genes and digital expression of genes based on transcriptome data from different tissues, growth stages and treatments. In addition, tools for visualization and the genomic data from various analyses are provided. All data in the database were manually curated and integrated within a user-friendly query page. This database provides valuable resources for a range of research fields related to *P. ginseng* and other species belonging to the Apiales order as well as for plant research communities in general. Ginseng genome database can be accessed at http://ginsengdb.snu.ac.kr/.

Keywords: *Panax ginseng*, genome database, ginseng annotation, ginseng genome browser

Introduction

Ginseng (*Panax ginseng* C.A Meyer) is a perennial herb of the *Panax* genus in Araliaceae family and has widely been used as a traditional medicine in Eastern Asia and North America. The principle bioactive components in ginseng are ginsenosides (collectively a group of triterpene saponins), which are biosynthesized through the isoprenoid pathway [1]. Ginseng has various therapeutic effects on humans including for treatment of cancer, diabetes, cardiovascular and stress [2-6]. *P. ginseng* is known to be tetraploid (2n = 4x = 48), with an estimated genome size of approximately 3.6 Gbp [7, 8]. Its large, highly repetitive genome, which has experienced whole-genome duplication, has impeded the progress of whole-genome sequencing of *P. ginseng* [7]. In addition, the long generation time (4 years) and difficulty of maintenance in ginseng cultivation fields have limited the genetic study of *P. ginseng*. Nevertheless, with the advent of new sequencing technologies, expressed sequence tags (ESTs) and RNA-Seq data have been generated from various tissues and growth stages of *P. ginseng* [9-12], based on which a number of genes involved in ginsenoside biosynthesis pathway have been characterized [10, 11]. Recently, the complete chloroplast genome sequence of *P. ginseng* cultivars and related species were characterized [13, 14]. Furthermore, inter- and intra-species chloroplast genome diversity were also identified for authentication of ginseng cultivars and species [13-16].

At the outset of this project, a total of 17,773 ESTs from NCBI db-EST (as of January, 2017) and a database for adventitious root [9] were publicly available for ginseng. However, these data were insufficient to facilitate the functional and comparative genomics and molecular breeding of ginseng. There was no comprehensive database publicly available for ginseng despite its importance as a medicinal crop with high pharmacological value. Given the fact that ginseng shows

numerous effects on human health, a genomic and transcriptomic database is vital for ginseng research communities and other close relatives in the Apiales order. It is also anticipated that an integrated database of genetic, genomic, and metabolomic resources of ginseng would serve as a valuable resource for translational genomics. Recently, we generated extensive genomic and transcriptomic data for *P. ginseng* cultivar "Chunpoong" [17].

In this study, we built a dynamic database that integrates a draft genome sequence, transcriptome profiles, and annotation datasets of ginseng. This Ginseng Genome Database is now publicly available (http://ginsengdb.snu.ac.kr/) for the use of scientific community around the globe for exploring the vast possibilities.

This user-friendly database will serve as a hub for mining gene sequences and their digital expression data of samples from various tissues, developmental stages, and treatments. Our database interface will facilitate the easy retrieval of gene families and associated functional annotations using InterPro, KEGG, BLAST and Gene Ontology (GO) databases. To expedite metabolomics in ginseng, we have made a separate section that categorizes the genes associated with various metabolic pathways including the ginsenoside biosynthesis pathway. In addition, we have included robust tools such as BLAST and genome browser (JBrowse) [18] for survey and visualization of ginseng genomic features. This database will be updated regularly with new genome sequences and information on annotation and will provide reference genomic information for research in *P. ginseng* as well as related species.

Construction and content

Whole-genome sequencing and assembly and gene models

The genome sequence data of *P. ginseng* were generated from an elite cultivar 'Chunpoong' using Illumina HiSeq platforms. A total of 746 Gb paired-end and 365 Gb mate-paired raw data were produced and assembled, yielding the draft genome

sequence of about ~3.0 Gb in size. The repeat sequences were identified and masked using RepeatModeler [19] and RepeatMasker [20]. An automatic gene prediction was performed using evidence modeler (EVM) [21] with *ab initio* predictions (BRAKER 1 [22]), protein evidence, ESTs and RNA-Seq evidence [23]. After the removal of the transposon sequences, a total 59,352 putative protein coding genes were predicted. These genes were functionally annotated using InterPro [24], Blast2go [25], KEGG [26] and BLASTP searches with known protein databases.

Transcriptome data

The transcriptome data were generated from various tissue and abiotic stress-treated samples of ginseng using Illumina HiSeq and PacBio platforms (http://ginsengdb.snu.ac.kr/transcriptome.php). Raw RNA-Seq reads of about 120 Gb were pre-processed in four steps to obtain high quality RNA reads. Initially, the bacterial contaminant reads were removed by read mapping against the available bacterial genomes using BWA [27]. After pre-processing, the duplicated reads were filtered out using FastUniq [28]. The third step is the removal of the ribosomal RNA (rRNA) reads using SortMeRNA [29]. Finally, the low-quality reads were removed using NGS QC Toolkit [30]. The high-quality RNA-Seq reads were used for *de novo* assembly by Trinity [31] and reference-guided assembly by HISAT & stringtie [32] and then for gene prediction on the draft genome sequence. In addition, high quality PacBio sequences were used to refine the predicted gene models.

Gene families and metabolic pathways

Genes were grouped based on protein domain (Pfam) and InterPro domain. Metabolic pathways were predicted with the KAAS server [26] using the reference information on gene annotation of *Arabidopsis thaliana*, *Citrus sinensis*, *Glycine max*, *Vitis vinifera* and *Solanum lycopersicum*. This information can be accessed at http://ginsengdb.snu.ac.kr/gene_family.php and http://ginsengdb.snu.ac.kr/metabolic_pathway.php.

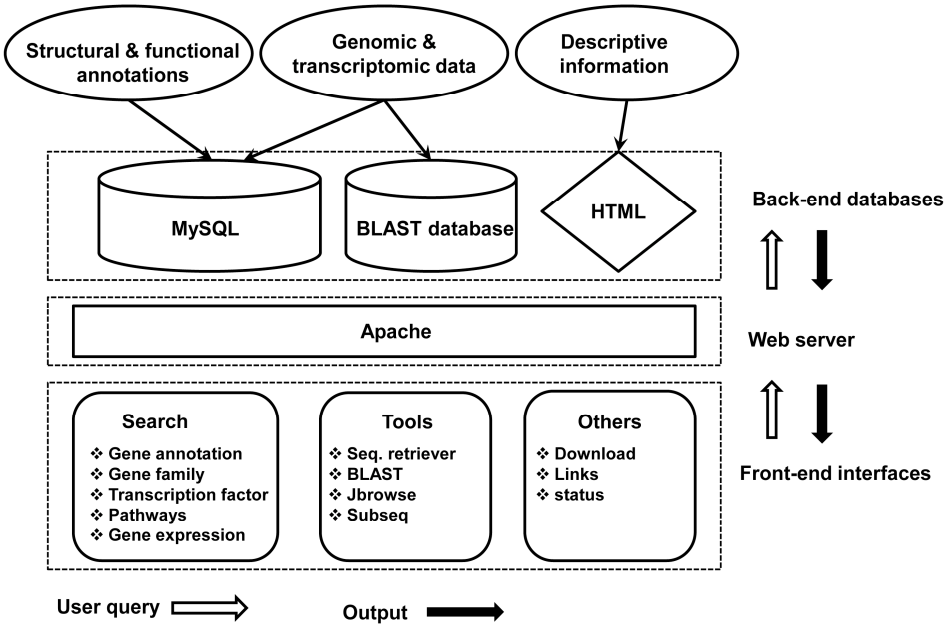

Figure 4-1. Overview of the architecture of the Ginseng Genome Database

Genome-scale metabolic network

Based on gene annotations, a compartmentalized genome-scale metabolic network was reconstructed providing the global overview of all metabolites, enzymes, reactions and pathways in ginseng. This network accounts for a total of 4,946 genes, mapped to 2,194 enzyme-catalyzed and protein-mediated transport reactions involving 2003 unique metabolites across six intracellular compartments. The global overview of ginseng genome-scale metabolic network can be accessed at http://ginsengdb.snu.ac.kr/network/index.html. This network can also be downloaded as a systems biology markup language (SBML) file.

Transcription factors

Transcription factors (TFs) are the key regulators for development and stimulus responses. TFs were identified based on the criteria of PlnTFDB [33] using iTAK (http://bioinfo.bti.cornell.edu/cgi-bin/itak/index.cgi). A total of 4,439 TF and

transcription regulator genes were identified and classified into 94 TF families (http://ginsengdb.snu.ac.kr/tf_class.php).

Genes in the ginsenoside biosynthesis pathway

Ginsenosides are biosynthesized through the mevalonate (MVA) and 2-C-methyl-D-erythritol-4-phosphate (MEP) pathways [10,34]. The number of genes that are involved in the biosynthesis of ginsenoside was identified based on KEGG as well as BLASTP annotations. UDP glycosyltransferase (UGT) genes, which are responsible for production of various types of ginsenosides in the final step of this pathway, were also identified based on InterPro ProSitePatterns (PS00375) and BLAST homology searches as well. The putative pathway and the related genes can be accessed at http://ginsengdb.snu.ac.kr/pathway.php.

Digital gene expression profiles

Digital gene expression profiles were determined using all of the RNA-Seq data. The FPKM values for all genes in each sample were calculated using RSEM [35]. Further, the expression data were normalized using Trimmed Mean of M values (TMM) to resolve the differences in the sequencing depth. The digital expression profiles can be accessed at http://ginsengdb.snu.ac.kr/gene_exp.php.

Utility and discussion

Database implementation

Ginseng Genome Database was established in the Linux (CentOS 6.6) operating system with an Apache HTTP server. PHP, HTML, JavaScript and Python scripts were used to build the user-friendly interface and design web pages. To visualize the genome, we included JBrowse version 1.11.6, which is JavaScript-based genome browser allowing visual analysis of the genome annotation [18]. We also included a BLAST server to perform homology searches with different data sets of ginseng. Moreover, we developed a Python-based tool to retrieve or download

specific scaffolds and gene sequences. An overview of the ginseng genome database architecture is shown in **figure 4-1**.

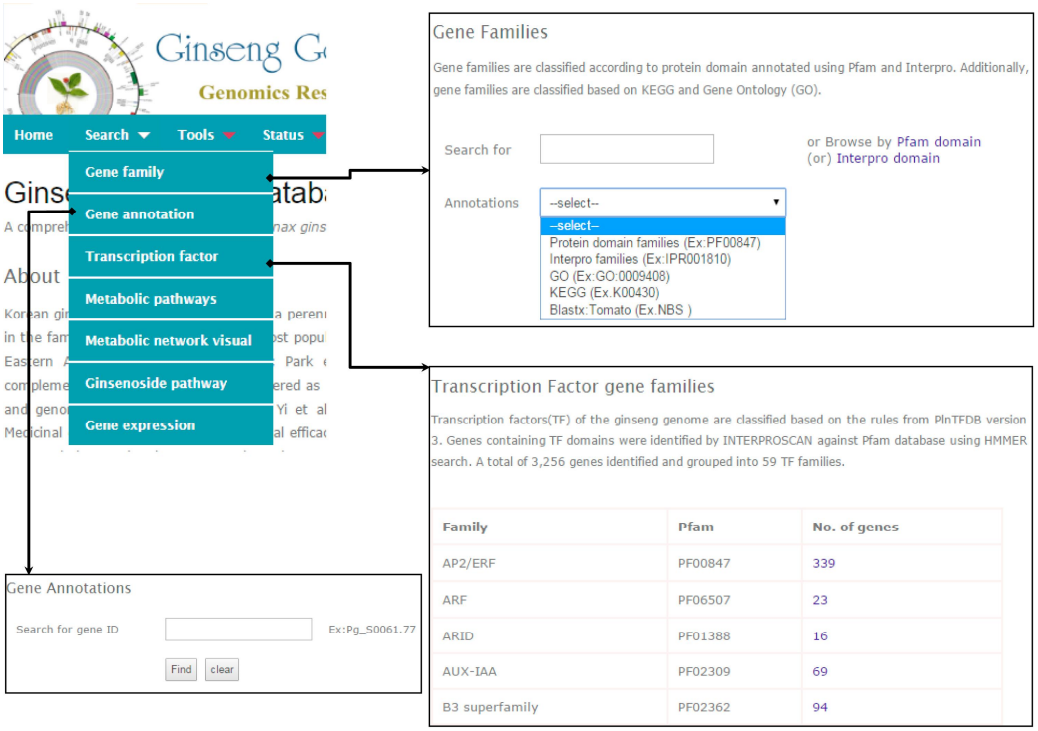

Figure 4-2. Query interface to retrieve information on gene annotations and transcription factors

Query search

Ginseng Genome Database provides two major panels, namely, a 'Search' panel and a 'Tool' panel, both of which comprise of all the information in an easy-to–use mode. Under the 'Search' panel, genes or gene families can be searched by gene ID, InterPro domain, Pfam domain, GO and KEGG orthology (KO) identifier (ID) and keywords **(Figure 4-2**). Furthermore, users can browse gene families categorized using 'InterPro' and 'Pfam' domains. The 'Gene family' option provides a sub-menu to retrieve the group of genes related to user-defined functional domains or keywords. Users can download all coding sequences (CDS) in a specific gene

family or user-selected CDSs in FASTA format from the output page. The 'Gene annotation' section provides the detailed annotations including both structural and functional annotations of the user-queried genes (**Figure 4-3**). In the output page, users can find the scaffold in which the specified genes, CDS and proteins were annotated and then can visualize those through JBrowse. Further, functional descriptions based on InterPro annotation including Pfam, Prositepatterns, and Superfamily, GO, KEGG and BLAST can also be browsed.

A list of annotated TF families is included in the 'Transcription factor' section (**Figure 4-2**). Users can explore the TF genes related to specific TF families and download the corresponding CDS. Under the 'Metabolic pathways' section, the users can simply enter a pathway name or click browse pathways to retrieve the genes involved in a particular pathway. Our database also provides links, so that users can check the 'enzyme commission (EC) number' for the corresponding genes and the complete pathway from the KEGG database. The known pathway of biosynthesis of ginsenosides and genes corresponding to each enzyme are listed under the 'Ginsenoside pathway' section. In the 'Gene expression' section, users can input a specific gene identifier and can choose to compare expressions between ginseng plant tissues or between abiotic stresses. This will return the expression data in bar-chart form using different colors.

Sequence retriever

We incorporated a 'sequence retrieving tool' using Python script. This can be utilized by entering single or batch gene (CDS and peptides) or scaffold IDs in the input query box. We customized the output options with 'view' and 'download'. Users can view single or multiple sequences in FASTA format on the web page by choosing the 'view' option. In case of many sequences, the user may select the 'download' option to download the sequences in FASTA format.

BLAST

This database also offers a homology search tool, 'BLAST', which was embedded in the database using the ncbi-wwwblast package (v2.2.26) to provide a graphic interface for the users. A BLAST-able database of the whole draft genome sequence, coding sequences (CDS), and protein sequences was made for BLAST searches. Additionally, the transcriptome data from various tissues and abiotic stress treatment of ginseng generated for the whole-genome study, the RNA-Seq assembly that were previously published and the ESTs were provided for BLAST searches. The users can perform BLAST searches by directly pasting the query sequences in the 'query text box', by choosing the appropriate search program (BLASTN, BLASTP, BLASTX, TBLASTN or TBLASTX), where BLASTP and TBLASTN are queried only against amino acid sequences. Options to filter low complexity and to set the *E*-value are available under the 'Other options' section. The result format can also be customized using the options under 'Result options'.

JBrowse

Under the 'Tools' panel, 'JBrowse' was included to visualize the genomic features of ginseng. All of the assembled scaffolds and the predicted genes were used in constructing the genome browser. The main page of 'JBrowse' contains several tracks under different sub-sections. Users can choose the 'scaffold' (only 30 scaffolds can be seen in the drop-down menu) or type the name of the scaffold with or without a location in the search box. Users can visualize various genomic features such as 'gene models', '*ab initio* gene models' generated for the gene annotation pipeline, 'assembled transcriptome structure' and 'repeats'. In addition, the alignments of RNA-Seq reads to the genome sequence generated directly from Binary Alignment/MAP (BAM) and PacBio contig alignment using GMAP [36] were also incorporated to perk up the structural annotation of the gene. Furthermore, protein sequences of non-coding genes including microRNA (miRNA) and long non-coding RNAs (lncRNA) can be seen along with their gene features. Apart from

Panax ginseng, we have incorporated the genome-guided transcriptome assembly of other *Panax* species, namely, *P. notoginseng* and *P. quinquefolius* which would aid in comparing the gene structure or find missing genes any other *Panax* species.

Figure 4-3. A detailed snapshot of the structural and functional annotations of a queried gene

Downloads

All the assembled genomic and transcriptomic sequences are available at http://ginsengdb.snu.ac.kr/data.php. Our database provides HTTP links to download the draft genome sequences (v1) and putative CDS and protein sequences (v1.1) in FASTA format. The gene and repeat structure annotations are available in Generic File Format (GFF3). The list of data files including *de novo* and reference-guided transcriptome assembly generated for whole genome study as well as the previously published transcriptome sequences generated from our research are also accessible in FASTA format. Besides, the filtered RNA-seq data used for genome analysis and the genome-scale metabolic network of ginseng can also be downloaded as a SBML file.

Conclusions

Ginseng Genome Database, the original, all-inclusive database for ginseng, is built on the most recent information of its draft genome sequence and accurate annotations. It serves as an open-access interface to retrieve genomic information from genome to gene level and to visualize all diverse components of the genome. The Ginseng Genome Database will form a valuable resource enhancing various research fields like functional/comparative genomics, metabolomics, molecular breeding, and evolutionary analysis of ginseng.

REFERENCES

1. Wang J, Gao WY, Zhang J, Zuo BM, Zhang LM, Huang LQ: Advances in study of ginsenoside biosynthesis pathway in *Panax ginseng* C. A. Meyer. *Acta Physiol Plant* 2012, 34(2):397-403.

2. Saito H, Yoshida Y, Takagi K: Effect of *Panax Ginseng* root on exhaustive exercise in mice. *Jpn J Pharmacol* 1974, 24(1):119-127.

3. Peng D, Wang H, Qu C, Xie L, Wicks SM, Xie J: Ginsenoside Re: Its chemistry, metabolism and pharmacokinetics. *Chin Med* 2012, 7:2.

4. Attele AS, Wu JA, Yuan CS: Ginseng pharmacology: multiple constituents and multiple actions. *Biochem Pharmacol* 1999, 58(11):1685-1693.

5. Shang W, Yang Y, Zhou L, Jiang B, Jin H, Chen M: Ginsenoside Rb1 stimulates glucose uptake through insulin-like signaling pathway in 3T3-L1 adipocytes. *J Endocrinol* 2008, 198(3):561-569.

6. Radad K, Gille G, Liu L, Rausch WD: Use of ginseng in medicine with emphasis on neurodegenerative disorders. *J Pharmacol Sci* 2006, 100(3):175-186.

7. Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park M, Choi D, Lim YP, Kwon SJ, Park BS *et al*: Major repeat components covering one-third of the ginseng (*Panax ginseng* C.A. Meyer) genome and evidence for allotetraploidy. *Plant J* 2014, 77(6):906-916.

8. Waminal NE, Park HM, Ryu KB, Kim JH, Yang TJ, Kim HH: Karyotype analysis of *Panax ginseng* C.A.Meyer, 1843 (Araliaceae) based on rDNA loci and DAPI band distribution. *Comp Cytogenet* 2012, 15:425-441.

9. Jayakodi M, Lee SC, Park HS, Jang W, Lee YS, Choi BS, Nah GJ, Kim DS, Natesan S, Sun C *et al*: Transcriptome profiling and comparative analysis of *Panax ginseng* adventitious roots. *J Ginseng Res* 2014, 38(4):278-288.

10. Jayakodi M, Lee SC, Lee YS, Park HS, Kim NH, Jang W, Lee HO, Joh HJ, Yang TJ: Comprehensive analysis of *Panax ginseng* root transcriptomes. *BMC Plant Biol* 2015, 15:138.

11. Lee Y, Park HS, Lee DK., Jayakodi M, Kim NH, Koo HJ, Lee SC, Kim YJ, Kwon SW, Yang TJ: Integrated transcriptomic and metabolomic analysis of five *Panax ginseng* cultivars reveals the dynamics of ginsenoside biosynthesis. *Front. Plant Sci* 2017, 8:1048.

12. Li C, Zhu Y, Guo X, Sun C, Luo H, Song J, Li Y, Wang L, Qian J, Chen S: Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng* C. A. Meyer. *BMC Genomics* 2013, 14:245.

13. Kim K, Lee SC, Lee J, Lee HO, Joh HJ, Kim NH, Park HS, Yang TJ: Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. *PloS one* 2015, 10(6):e0117159.

14. Kim K, Lee SC, Lee J, Yu Y, Yang K, Choi BS, Koh HJ, Waminal NE, Choi HI, Kim NH, *et al*: Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. *Sci Rep* 2015, 5:15655.

15. Kim K, Nguyen VB, Dong JZ, Wang Y, Park JY, Lee SC, Yang TJ: Evolution of the Araliaceae family inferred from complete chloroplast genomes and 45S nrDNAs of 10 *Panax*-related species. *Sci Rep* 2017, 7(1):4917.

16. Nguyen VB, Park HS, Lee SC, Lee J, Park JY, Yang TJ. Authentication markers for five major *Panax* species developed via comparative analysis of complete chloroplast genome sequences. *J Agric Food Chem* 2017, 65(30):6298-6306.

17. Kim NH, Jayakodi M, Lee SC, Choi BS, Jang W, Lee J, Kim HH, Waminal NE, Lakshmanan M, Binh NV, *et al*: Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. *Genome Biol* 2017, (under review).

18. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: JBrowse: a next-generation genome browser. *Genome Res* 2009, 19(9):1630-1638.

19. Smit A, Hubley R: RepeatModeler Open-1.0. *Repeat Masker Website* 2010.

20. Smit AF, Hubley R, Green P: RepeatMasker Open-3.0. In.; 1996.

21. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 2008, 9(1):R7.

22. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M: BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 2015, 32(5):767-769.

23. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD: Improving the Arabidopsis

genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 2003, 31(19):5654-5666.

24. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L: InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009, 37:D211-D215.

25. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21(18):3674-3676.

26. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007, 35:W182-185.

27. Li H, Durbin R: Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009, 25(14):1754-1760.

28. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S: FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One 2012*, 7:e52249.

29. Kopylova E, Noé L, Touzet H: SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012, 28(24):3211-3217.

30. Patel RK, Jain M: NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one* 2012, 7(2):e30619.

31. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011, 29(7):644-652.

32. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL: StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015, 33(3):290-295.

33. Jin J, Zhang H, Kong L, Gao G, Luo J: PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res* 2013: 42: D1182- D1187.

34. Zhao S, Wang L, Liu L, Liang Y, Sun Y, Wu J: Both the mevalonate and the non-mevalonate 5pathways are involved in ginsenoside biosynthesis. *Plant Cell Rep* 2014, 33(3):393-400.

35. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011, 12:323.

36. Wu TD, Watanabe CK: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005, 21(9):1859-1875.

ABSTRACT IN KOREAN

약용작물의 제왕으로 알려진 *Panax ginseng* C. A. Meyer(인삼)은 성장속도가 느리고 세대가 길며 종자 생산이 적고 복잡한 지놈 구조를 가지고 있어 연구가 어려웠다. 또한 인삼의 다양한 비생물적 스트레스에 대한 분자적 반응 기전에 대한 연구가 미비하였다. 따라서 본 논문에서는 인삼의 기능 유전체학, 범유전체학, 육종학적 연구를 수행하였고 이를 네 개의 장으로 정리하여 구성하였다. 시퀀싱 기술이 급격히 성장하면서 인삼의 게놈 프로젝트는 2011년 시작하여 2016년에 처음으로 유전체 지도의 초안이 완성되었다. 첫번째 장에서는 사배체의 *P. ginseng* 유전체에서 총 59,352개의 유전자 주석달기(annotation)를 완성하였다. 이 중 약 97퍼센트에 해당하는 유전자가 기능적 특성을 갖는 것으로 확인되었고, 3,558개의 전사인자와 851개의 전사조절인자를 94개의 유전자 그룹으로 나누어 구분하였다. 기능 및 진화학적 분석을 통해 약리학적으로 중요한 dammarane 타입의 ginsenosides가 *Panax*에서 유래하였으며, 주로 shoot조직에서 생산되어 뿌리로 이동함을 알 수 있었고 *P. ginseng*의 지방산 불포화 효소는 동결에 대한 내성을 지닌 형태로 새롭게 진화하였으며, 엽록소 a/b 결합 단백질 유전자를 보유함으로서 저조도에서도 효율적인 광합성이 가능하게 된 것을 확인하였다. 또한 다양한 시퀀싱 플랫폼을 활용한 전사체 분석과 대사체 분석 데이터를 통해 11개의 새로운 UDP-glucuronoxyltransferase (UGTs)를 동정할 수 있었다.

인삼에서 열과 빛 스트레스는 식물의 성장과 지속적인 생산에 주요한 방해 요소인 것으로 알려져 있다. 이러한 식물 생장과 관련한

고온 스트레스가 식물에게 미치는 생리학적 연구는 다양하게 이루어져 왔지만, 기저에 있는 분자 생물학적 반응은 거의 알려지지 않았다. 본 연구에서는 두 가지 인삼 품종인 청풍(CP)과 연풍(YP) 인삼을 각각 1주, 3주 동안 열처리한 뒤 각 샘플의 전사체 데이터를 이용하여 열 스트레스에 민감한 품종과 저항성이 있는 품종의 전사체 발현 패턴을 비교하였다. 차등발현유전자분석(DEG)과 유전자 기능분석, 염록소 함량 분석 결과, 열 스트레스에 민감한 CP품종은 YP품종에 비해 염록소 함량이 낮았으며, CAB 단백질, WRKY 전사인자, FAD를 번역하는 유전자가 열 스트레스에 대한 반응에 관여하는 것으로 나타났고 이를 통해 광합성이 저해되는 것을 발견하였다. 또한 광합성과 당대사 관련 유전자들이 열 스트레스에 민감한 CP 품종에서 RuBisCO의 전사를 감소시킴으로서 식물체내의 수용성 당 축적을 증가시키는 것으로 나타났다.

Long noncoding RNAs (lncRNAs)는 유전체 조절과 다양한 발달 과정 및 질병 등의 다양한 생물학적 역할에 관여하는 것으로 알려져 있다. 본 논문의 세번째 장에서는 체계적인 분석 파이프라인을 이용하여 다양한 조직과 성장 단계별, 비생물학적 스트레스 처리를 통해 얻어진 약 1,040 억개의 시퀀싱 리드를 이용하여 19,495개의 lncRNA 목록을 만들었다. 또한 가뭄, 염분, 추위, 더위, methyl jasmonate (MeJA)에 대한 비생물학적 스트레스 반응에 관련된 100개 이상의 후보 lncRNA와 이들 중 2,607개의 조직 특이적, 성장 단계별로 특이적인 lncRNA의 분석 결과를 정리하였다. 또한 전이인자가 인삼의 lncRNA의 기능에 영향을 주는 요인이 될 것으로 예상되었다.

마지막 장에서는 open-access 인삼 유전체 데이터베이스 (http://ginsengdb.snu.ac.kr/) 개발에 대한 결과를 정리하였다. 데이터 베이스의 현재 버전은 유전자의 구조적, 기능적 특성과 조직별, 성장단계별, 처리군 별 전사체 데이터 분석에 기초한 유전자의 디지털 발현 결과의 확인이 가능하며 또한 지놈 데이터의 시각화 및 분석을 위한 도구를 제공하고 있다. 모든 데이터는 사용자가 편리하게 사용할 수 있도록 수동으로 선별 및 통합하였다. 본 연구 결과는 인삼의 새로운 품종 개발을 위한 연구에 있어서 생물학 및 비생물학적 스트레스와 직사광선에 대한 내성을 지니면서 약리학적 효과는 향상시킨 새로운 인삼품종의 유전학적, 대사공학적 연구 개발에 활용 될 수 있을 것이다.