SCIENCE CHINA
Life Sciences

Peking University Biological Science Education & Research (1925–2015)
• RESEARCH PAPER •

November 2015 Vol.58 No.11: 1099–1110
doi: 10.1007/s11427-015-4961-x

De novo assembly and comparative analysis of root transcriptomes from different varieties of *Panax ginseng* C. A. Meyer grown in different environments

ZHEN Gang[1,2†], ZHANG Lei[3†], DU YaNan[1,2], YU RenBo[1,2], LIU XinMin[4], CAO FangRui[4], CHANG Qi[4], DENG XingWang[1,2*], XIA Mian[3*] & HE Hang[1,2*]

[1]*State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing 100871, China;*
[2]*School of Advanced Agricultural Sciences, Peking University, Beijing 100871, China;*
[3]*Frontier Laboratories of Systems Crop Design Co., Ltd., Beijing 100085, China;*
[4]*Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100193, China*

Received May 27, 2013; accepted June 16, 2013

Panax ginseng C. A. Meyer is an important traditional herb in eastern Asia. It contains ginsenosides, which are primary bioactive compounds with medicinal properties. Although ginseng has been cultivated since at least the Ming dynasty to increase production, cultivated ginseng has lower quantities of ginsenosides and lower disease resistance than ginseng grown under natural conditions. We extracted root RNA from six varieties of fifth-year *P. ginseng* cultivars representing four different growth conditions, and performed Illumina paired-end sequencing. In total, 163,165,706 raw reads were obtained and used to generate a *de novo* transcriptome that consisted of 151,763 contigs (76,336 unigenes), of which 100,648 contigs (66.3%) were successfully annotated. Differential expression analysis revealed that most differentially expressed genes (DEGs) were upregulated (246 out of 258, 95.3%) in ginseng grown under natural conditions compared with that grown under artificial conditions. These DEGs were enriched in gene ontology (GO) terms including response to stimuli and localization. In particular, some key ginsenoside biosynthesis-related genes, including HMG-CoA synthase (*HMGS*), mevalonate kinase (*MVK*), and squalene epoxidase (*SE*), were upregulated in wild-grown ginseng. Moreover, a high proportion of disease resistance-related genes were upregulated in wild-grown ginseng. This study is the first transcriptome analysis to compare wild-grown and cultivated ginseng, and identifies genes that may produce higher ginsenoside content and better disease resistance in the wild; these genes may have the potential to improve cultivated ginseng grown in artificial environments.

***Panax ginseng*, *de novo* assembly, paired-end sequencing, comparative transcriptome analysis, ginsenoside biosynthesis, disease resistance genes**

Citation: Zhen G, Zhang L, Du YN, Yu RB, Liu XM, Cao FR, Chang Q, Deng XW, Xia M, He H. *De novo* assembly and comparative analysis of root transcriptomes from different varieties of *Panax ginseng* C. A. Meyer grown in different environments. Sci China Life Sci, 2015, 58: 1099−1110, doi: 10.1007/s11427-015-4961-x

Panax ginseng C. A. Meyer has been used in traditional medicine in eastern Asia since ancient times [1−3], and cultivation of this herb dates back at least as far as the Ming dynasty. Ginsenosides, a large group of triterpene saponins, are the main bioactive compounds that account for ginseng's medicinal properties. Modern medical science has demonstrated that various ginsenosides may be very useful for the treatment of many human diseases such as cancer, diabetes, cardiovascular disease, oxidative stress, hypoxia, and neural degeneration [4–9]. Owing to its important me-

†Contributed equally to this work
*Corresponding author (email: deng@pku.edu.cn; xm@frontier-ag.com; hehang@pku.edu.cn)

dicinal properties, *P. ginseng* has been highly valued and vigorously pursued since ancient times. However, because *P. ginseng* grows very slowly in natural conditions and wild-grown ginseng cannot meet the growing demand from the medical market. Consequently, the artificial cultivation of ginseng provides an alternative means of meeting demand for this precious herb.

Compared with that grown in the wild, cultivated ginseng grows much faster and produces a much higher yield; however, it is considered much less medicinally valuable. Because culture methods affect the concentrations of the various volatile compounds (e.g., sesquiterpenes) in *P. ginseng* [10], the differences in medicinal value probably result from the reduced ginsenoside content of cultivated ginseng compared with wild-grown ginseng. Furthermore, cultivated ginseng faces great pressure from various plant diseases and does not usually survive beyond its sixth year. Ginseng farmers typically have to harvest cultivated ginseng in its fifth year and occasionally change the plantation location. In contrast, under natural conditions, *P. ginseng* is a very long-lived herb and can grow for hundreds of years.

Kwon et al. [11] reported that a novel gene, *pNRT2* (plasma membrane localized nitrate transporter 2), was more preferentially expressed in mountain wild-grown ginseng than mountain cultivated ginseng. This gene might take part in nitrogen transport [12], and may therefore contribute to the increased survival rate of wild-grown plants exposed to adverse environmental and climatic conditions. To date, few systematic studies have been carried to identify the cause of the differences in medicinal value and disease resistance between wild and cultivated ginseng.

The advances made in next-generation sequencing technology have greatly facilitated the generation of genomic and transcriptomic data, and RNA sequencing (RNA-seq) analysis provides a powerful method for obtaining a quick measure of gene expression at the whole transcriptome level. Such analysis has been performed using *P. ginseng* and its two close relatives, *Panax quinquefolius* and *Panax notoginseng* [13–21]. These efforts have improved our understanding of the transcriptome of *P. ginseng*.

In this study, we performed deep transcriptome sequencing of ginseng grown under different conditions, and generated an improved *de novo* root transcriptome assembly. By differential expression, we then attempted to identify the differences in transcript expression among the samples. This study provides valuable transcriptome information for determining the differences in ginsenoside content and disease resistance in ginseng grown under different conditions.

1 Materials and methods

1.1 Materials

Whole root samples from all the six varieties of ginseng were collected in September 2012, at their fifth year. Among them, XKB (XunKe B type) was planted in the far north mountain area of Heilongjiang province, China, and the other five samples were planted in the mountain area of the southeast part of Jilin province, China. The XKB, JAB (JiAn B type), CBB (ChangBai B type), and FSB (FuSong B type) varieties were grown in artificial facilities. CBA (ChangBai A type) was planted in the wild with no intervention. KDC (KuanDian C type) was half-wild-grown, meaning that it was grown under artificial conditions during its early stages and then moved to wild conditions for full growth. As cultivated ginseng is very susceptible to serious plant diseases beginning in the sixth year, all root samples were collected in the fifth year. After cleaning, these samples were immediately frozen in liquid nitrogen, and transported to our laboratory in Beijing for further analysis.

1.2 RNA extraction, sequencing, reads filtering, and *de novo* assembly

We used an RNeasy Plant Mini Kit (product code: 74904; Qiagen, Shanghai, China) for total RNA extraction, and 100 mg of ginseng root tissue was used. We followed the kit manual for all extraction steps. Purified mRNA libraries derived from these root samples were sent to the Biodynamics Optical Imaging Center (BIOPIC) at Peking University for further reverse transcription and high-throughput sequencing using Illumina's pair-end sequencing platform. The KDC and JAB samples were sequenced using HiSeq 2500 with a 101-bp read length, and the other four samples were sequenced using MiSeq with a 251-bp read length. All the raw reads have been deposited in the Sequence Read Archive (SRA) under Bioproject ID: PRJNA265935.

The adaptors of raw reads were first trimmed by a custom Perl script. The quality of these trimmed raw reads was screened by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and low-quality bases (Q-value<30) at both ends were cut by a custom Perl script. We then mapped the remaining reads to a eukaryote rRNA database, and removed all the reads related to eukaryotic rRNA. After that, all the clean reads were fed into Trinity software (version: r2013-02-25) [22] to carry out *de novo* transcriptome assembly with a k-mer length of 31. To avoid possible assembly errors and reduce redundancy in the *de novo* transcriptome, we first filtered those contigs shorter than 300 bp, then carried out 95% identity clustering by CD-HIT-EST [23]. Finally, we remapped all high-quality reads back to the *de novo* transcriptome and removed those contigs covered by less than an arbitrary threshold of two reads. All the mapping procedures described above used BOWTIE2 [24] with its default parameters.

1.2 Functional annotation and gene ontology (GO) analysis

We searched all transcripts in our transcriptome against

several protein databases using BLASTX (cutoff *E*-value: 1×10^{-10}); the databases included the non-redundant protein database (NR) from the National Center for Biotechnology Information (NCBI), TAIR10 (protein), and SWISS-PROT. We also compared the BLAST results against TrEMBL and an annotated *de novo* transcriptome of a similar study in *P. quinquefolius* [17] as complement. The pathway information of our assembly was assigned by KAAS (the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server) [25] (http://www.genome.jp/tools/kaas/) using the BBH method. Open reading frames (ORFs) in our transcriptome were predicted using TransDecoder (version: 2.0.1) [26], and only proteins longer than 100 amino acids were reported. For functional predictions of these ORFs, we searched them against the Pfam database using the Profile Hidden Markov Model (PHMM) by HMMScan (version: 3.1b2 Feb 2015) [27] (cutoff *E*-value: 1×10^{-5}).

We used the similarity comparison results of our assembly and TAIR10, by far the most fully annotated genome, to perform GO analysis. The most strongly hit (meaning the BLAST result with the lowest *E*-value) homology sequences (cutoff *E*-value<1×10^{-10}) in TAIR10 (protein) were extracted and submitted to the agriGO website (http://bioinfo.cau.edu.cn/agriGO/) [28] for GO analysis. The significantly enriched GO terms were identified at false discovery rates (FDRs) lower than 0.05.

1.3 Identification of ginsenosides biosynthesis genes and disease resistance genes

We used the KAAS annotation results to identify ginsenoside backbone biosynthesis genes. This backbone primarily includes the terpenoid backbone biosynthesis pathway (ko00900) and the sesquiterpenoid and triterpenoid biosynthesis pathway (ko00909, and chair-chair-chair-boat conformation). The ginsenoside biosynthesis-related genes in these two pathways were extracted for further analysis; Supportingly, we used the annotation results of the other databases to complement our work.

The Plant Resistance Gene Database (PRGDB), (http://prgdb.org) is a community-based database of plant disease resistance genes (R genes), and contains over 112 reference and 104335 putative R genes [29]. We downloaded the protein sequences of all of the curated R genes from this website and used BLASTX to search our assembled transcriptome for the protein sequences (cutoff *E*-value: 1×10^{-20}).

1.4 Co-expression analysis and differential expression analysis

Because oleanane-type ginsenosides were much less important, we analyzed only those genes that might be related to the biosynthesis of dammarane-type ginsenosides. Han et al. [30] demonstrated that the cytochrome p450 enzyme CYP716A47 transforms dammarenediol-II to protopanaxadiol during *P. ginseng* ginsenoside biosynthesis. Thus, we chose CYP716A47 and DS (dammarenediol-II synthase) for co-expression analysis. We examined the mapped raw reads of each gene and calculated the Pearson's correlation coefficient (R) to analyze all identified cytochrome p450s and glycosyltransferases for their co-expression patterns with CYP716A47 and DS.

Many genes in the *de novo* transcriptome produced by Trinity contain many splicing isoforms, and it has been argued that differential transcripts (meaning at the isoform level) expression analysis produces more false positives than differential gene expression analysis [31,32]. Thus, in all the differential expression analyses described below, we carried out the analysis at the gene level.

For differential expression analysis, we first divided the six samples into four groups. Because CBB, FSB, and JAB were all grown in nearly the same environment, they were combined to form group 41B and were treated as three replicates. We treated 41B as the control group because we wanted to determine the differences between the cultivated ginseng and that grown by different methods. We used the BOWTIE2-eXpress-DESeq pipeline to carry out differential expression analysis. To avoid possible mapping differences between reads of different lengths, all the high-quality reads derived from a sequencing platform with a read length of 251 bp were trimmed to 101 bp. We first aligned all clean pair-end reads to our assembled transcriptome using BOWTIE2 (bowtie2 –a –X 200 –x Panax_Transcriptome -1 left.fq -2 right.fq –S Results.sam). The sam files derived from BOWTIE2 were processed using eXpress software [33], and reads were assigned to different transcripts (express –o output/ Panax_Transcriptome.fa Results.sam). For downstream gene level expression analysis, reads that mapped to different isoforms of one gene were combined. Finally, the raw read count of each sample was fed into DESeq [34] to carry out differential expression analysis. A gene was defined as expressed if it was mapped by over five raw reads and as differentially expressed between two samples if the FDR was lower than 0.05.

1.5 Real-time polymerase chain reaction (PCR) analysis

After digestion with DNase I (NEB), approximately 3 μg of total RNA from each sample was converted into first-strand complementary DNA (cDNA) via the reverse-transcription reaction using oligo $(dT)_{15}$ primers and a SuperScript III Reverse Transcriptase Kit (Invitrogen). The cDNA product was diluted 10-fold using nuclease-free deionized water, and was then used as a template in real-time PCR analysis.

Specific cDNAs were amplified by SYBR Green Real-Time PCR Master Mix (TOYOBO) in a volume of 20 μL. The reaction mixture contained 10 μL SYBR Green Real-Time PCR Master Mix, 4 μmol L^{-1} each of the forward and reverse primers, and 1 μL of the template cDNA. PCR amplification was performed at an annealing temperature of 60°C using the 7500 Fast Real-Time PCR System (Applied Biosystems) according to the manufacturer's instructions. Relative transcript abundances were calculated by the comparative cycle threshold method with the glyceraldehyde 3-phosphate dehydrogenase gene (*GAPDH*) as an internal standard. The relative gene expression was calculated using the $2^{-\Delta\Delta C_t}$ method [35].

1.6 Measurements of ginsenoside content using ultraviolet ultra-performance liquid chromatography (UPLC-UV)

The collected fresh ginseng roots were freeze-dried and ground to a fine powder. The nine ginsenosides, namely Rg1, Re, Rf, Rg2, Rb1, Rc, Rb2, Rb3, and Rd, were identified by the UPLC-UV method described previously [36] with slight modification. Briefly, 0.5 g of ginseng powder was soaked in 80% methanol (20 mL) for 3 h, followed by 10 min ultrasonication for extraction. After centrifugation at 14,800 r min^{-1} for 15 min, the extract solution was transferred to a vial and 5 μL of the prepared sample was injected into an ACQUITY UPLC system for assay. The system (Waters, Milford, MA, USA) was equipped with a binary solvent delivery pump, an auto sampler, and a photodiode array UV detector. The chromatographic separations were carried out on an ACQUITY UPLC® BEH C18 column (50 mm×2.1 mm; internal diameter, 1.7 μm) (Waters, Milford, MA, USA) maintained at 30°C. The mobile phases comprised acetonitrile and water, and were eluted in the following gradient programs at a flow rate of 0.3 mL min^{-1}. The percentages of acetonitrile were 19% (0–3 min), 19%–21% (3–4 min), 21%–26% (4–5 min), 26%–27% (5–9 min), 27%–32% (9–12 min), 32%–43% (12–15 min), 43%–60% (15–18 min), 60% (18–20 min), 60%–70% (20–21 min), 70%–80% (21–22 min), 80%–90% (22–23 min), 90% (23–24 min), 90%–19% (24–25 min), and 19% (25–30 min). The detection wavelength was 203 nm.

2 Results

2.1 Ginseng transcriptome sequencing and *de novo* assembly

Using RNA-seq, a total of 163.2 million raw reads were generated, reaching approximately 28.6 billion base pairs (Table1). After trimming of adapter and low quality sequences, and removing ribosomal RNA, 136.4 million high-quality reads (20.8 billion base pairs) remained. These high-quality reads were used for *de novo* transcriptome assembly by Trinity software. Therefore, we obtained a *de novo* transcriptome assembly comprising 76,336 unigenes or 151,763 transcripts when all isoforms were counted. The whole transcriptome size was 152,142,167 bp. The size distribution of contigs was 300–13,248, and most (79.8%) were 300–1500 bp long (Figure 1A). The average transcript length was 1002.5 bp, and the N50 was 1439 bp. These parameters were both far longer than those of most similar ginseng studies (Table 2) [16–19]. The ginsenosides content

Table 1 Summary of *Panax ginseng* C. A. Meyer RNA sequencing results

	CBA	KDC	XKB	JAB*	FSB*	CBB*
No. of raw reads	19,243,506	49,042,984	20,166,196	33,125,304	19,560,730	22,026,986
Raw read length	251	101	251	101	251	251
No. of raw bases (Gbp)	4.83	4.95	5.06	3.35	4.91	5.53
No. of reads after filtering	18,120,158	43,793,704	19,098,670	16,066,800	18,454,930	20,868,560
No. of bases after filtering (Gbp)	3.49	4.37	3.75	1.60	3.55	4.01
Percentage in assembly	74.1%	85.1%	80.2%	78.3%	85.0%	78.3%

*, These three kinds of *Panax ginseng* constitute the group 41B, which means they were cultivated at 41° latitude

Table 2 Comparison of recent *de novo* transcriptome assemblies in *Panax* species

Species	*P. ginseng* [16]	*P. quinquefolius* [17]	*P. ginseng* [18]	*P. ginseng* [19]	*P. ginseng* (Our assembly)
No. of contigs	178,145*	41,623	60,236	35,527†	168,742
Range of length	100–7,858	300–7,719	200–13,903	410–15,918	300–13,248
Average length	424.5	895.8	652.5	1,978	999.3
N50 (bp)	620	1,114	871	2,274	1,438
Total bases	75,621,996	37,284,379	39,304,529	70,295,564	168,619,830

*, This assembly consists of 86,609 contigs and 91,536 singletons, which have a length range of 100–7,858 and 100–691, respectively. This paper did not provide any indication of N50. Its N50 number was calculated based on an incomplete assembly comprising 67,786 contigs deposited in GenBank and provided by the authors. †, This research provided two transcriptomes, CP and CS. The numbers here describe CP, while the transcriptome of CS only showed the main differences in contig numbers (27,716) and total bases (54,892,571).

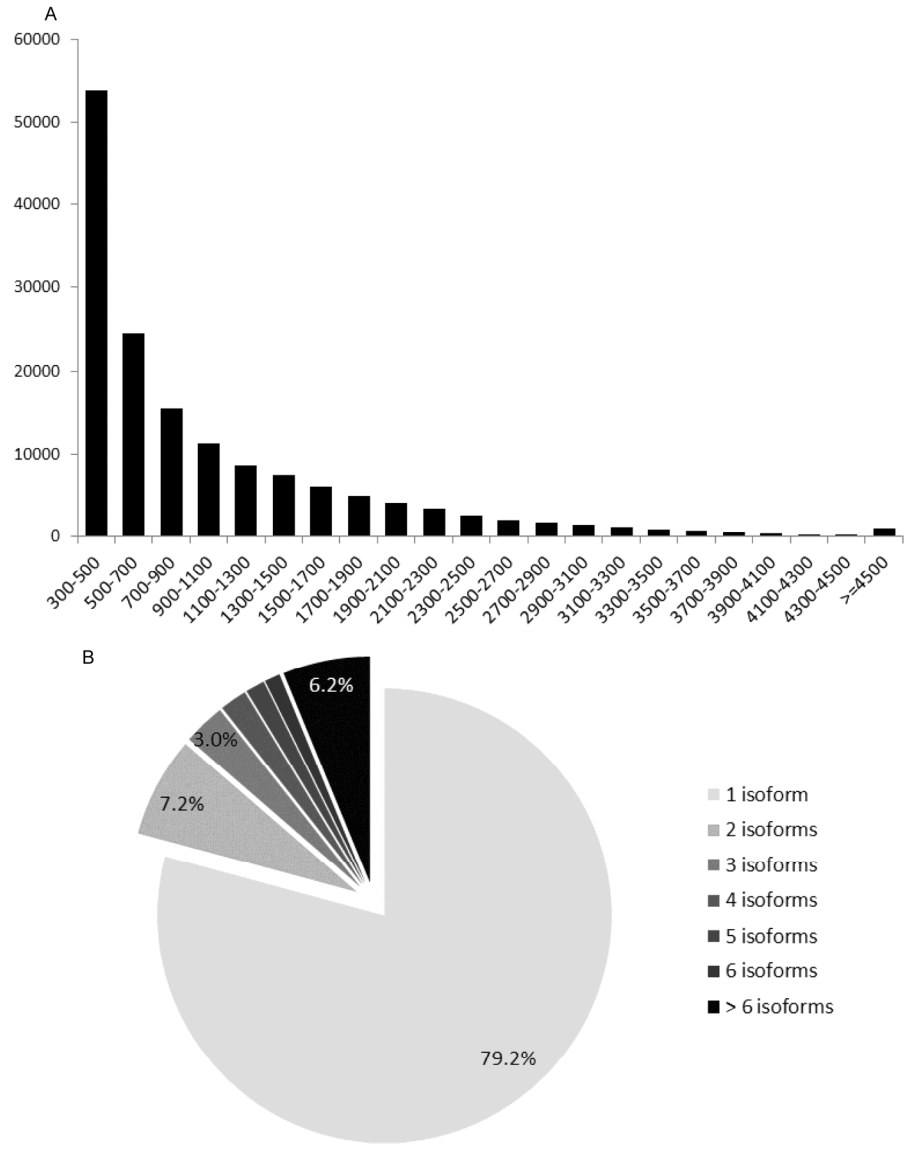

Figure 1 Contig length distribution (A) and isoform abundance (B) of our assembled transcriptome.

of our transcriptome was 39%. Approximately 19.8% of the unigenes in our assembly possessed at least two isoforms (Figure 1B), comparatively fewer than observed in a similar study using North American ginseng (*P. quinquefolius*) [17].

The current NCBI EST database contains 17,773 *P. ginseng* contigs with an average length of 720 bp. Similarity comparison of these NCBI database ESTs with our assembled transcriptome revealed that 96.7% had homologous sequences in our assembly (cutoff *E*-value: 1×10^{-20}), and the average identity was 97.3% (Table 3). Similarity comparison with another published *P. ginseng* transcriptome assembly and with one *P. quinquefolius* transcriptome also showed that most of the transcripts assembled in these studies (all over 95%) were well represented in our assembly. Comparison with another recently published *de novo* transcriptome of *P. ginseng* [20] also showed that about 71.3% of contigs in that assembly were represented in our transcriptome (average identity: 97.6%). Therefore, we generated an improved *de novo* transcriptome compared with former studies in the genus *Panax*, and this assembly should be valuable for future ginseng studies.

2.2 Functional annotation and GO analysis

To facilitate further downstream analysis, we performed functional annotation of all the transcripts in our assembly. Among those databases, NR annotated 86,854 contigs, SWISS-PROT annotated 59,977 contigs, and TAIR10 annotated 71,818 contigs; 16,945 transcripts were assigned to pathways by KAAS. A total of 100,648 (66.3%) transcripts in our assembly were successfully annotated. For NR anno-

Table 3 Homology searches of our assembly against other *Panax* transcriptomes (*E*-value<1×10^{-20})

	No. of sequences	Average length	Organ	Homology sequences
P. ginseng (NCBI EST database)	17,114	720.0 bp	Leaf, root, whole plant	16,553 (96.7%)
P. ginseng [16]	67,786*	558.5 bp	Leaf, root, stem, flower	64,922 (95.8%)
P. ginseng [19]	55,949	1250.6 bp	Root	39,899 (71.3%)
P. quinquefolius (NCBI EST database)	5,018	495.6 bp	Flower, root	4,788 (95.4%)
P. quinquefolius [17]	41,623	895.8 bp	Root	40,686 (97.7%)

*, This was an incomplete transcriptome assembly deposited in GenBank provided by the authors, so the average size of this assembly was different from the corresponding number in Table 2, which was provided by the authors.

tation results, the most enriched hit species was *Vitis vinifera* (10,627 hits in 86,854), *Theobroma cacao* (4,630 hits in 86,854) was the second most enriched, and *Solanum lycopersicum* (4,343 hits in 86,854) was the third most enriched. For the 16,945 contigs annotated by KAAS, 11,198 were assigned to 252 pathways, 3,729 were related to metabolism, and 3,150 were related to genetic informatics processing. Among those pathways, ribosome was the most enriched (path: ko03010, with 868 members), plant hormone signal transduction was the second most enriched (path: ko04075, with 313 members), and spliceosome (path: ko03040, with 291 members) was the third most enriched.

Among all the 151,763 transcripts in our assembly, 77,201 (50.9%) possessed ORFs. When searched against the Pfam database, 4,493 kinds of domains (186,077 domains in total), representing 50,211 contigs, were found in those ORFs. Among those domains, the 35 amino acid pentatricopeptide repeats (PPR repeat, PF01535.15, PF13812.1, PF13041.1, and PF12854.2) were the most represented domains (Supporting file 1). PPR proteins constitute a large group (over 400 members) in land plants and they usually influence the expression of transcripts from organelles, such as mitochondria and chloroplasts [37]. Leucine-rich repeats (PF00560.28, PF12799.2, PF13855.1, and PF13504.1), which are usually residue motifs of 20–29 bp and function as structural frameworks for protein–protein interactions [38], represented the second most enriched motif. The tryptophan-aspartic acid (WD) domain, G-beta repeats (PF00400.27), comprised the third most enriched motif. The WD-repeat proteins usually contain a conserved core that has a WD end, and they have many important biological functions, such as signal transduction, transcription regulation, and apoptosis [39].

The GO analysis results contain three parts: biological process (GO: 0008150), cellular component (GO: 0005575), and molecular function (GO: 0003674), and all the most strongly enriched elements are shown in Figure 2. Within biological process, cellular process (GO: 0009987), and metabolic process (GO: 0008152) were the most significantly enriched elements (FDR<1×10^{-45}); within molecular function, the most significantly enriched element was catalytic activity (GO: 0003824) (FDR<1×10^{-110}); and within cellular component, intracellular part (GO: 0044424) and organelle (GO: 0043226) groups were the most significantly enriched elements (FDR<1×10^{-120}).

2.3 Differentially expressed genes in ginseng grown under different conditions

We observed that ginseng from all four groups (CBA, KDC, XKB, and 41B) had nearly the same number of expressed genes (from 34,053 in CBA to 38,847 in KDC) (Figure 3). However, each group possessed a large proportion of genes (from 7.35% in XKB to 12.86% in KDC) that were exclusively expressed, and only 22,754 genes were expressed in all four groups.

We performed differential expression analysis between group CBA, KDC, or XKB and group 41B. Two hundred and fifty-eight DEGs were detected in CBA and 41B, 19 DEGs in KDC and 41B, and 13 DEGs in XKB and 41B (Figure 4A). The wild-planted CBA had the highest number of DEGs. Furthermore, compared with 41B, most DEGs (246 out of 258, 95.3%) were upregulated in CBA, 12 out of 19 (63.2%) were upregulated in KDC, and 12 out of 13 (92.3%) were upregulated in XKB (Figure 4B). To better characterize the functions of the DEGs, GO analyses were performed to identify functional enrichments in these comparisons (Figure 4C and D). Between CBA and 41B, the most strongly enriched GO term was response to stimulus (GO: 0050896 FDR *P*-value=0.00139), and the other enriched terms were related to localization (GO: 0051179, *P*-value= 0.0274; and GO: 0051234, FDR *P*-value=0.023). Among those DEGs in KDC and 41B, the only enriched GO term was response to stimulus (GO: 0050896, FDR *P*-value=0.0242). Moreover, there was no enriched GO term in the DEGs in XKB and 41B, owing to a lack of DEGs.

2.4 Ginsenoside biosynthesis-related genes were differentially expressed in wild-grown and cultivated ginseng

The ginsenoside content of these samples is shown in Supporting file 2 (the ginsenoside content of KDC and XKB was lacking owing to shortage of samples). The ginsenoside content of the wild-grown sample (CBA) was much higher than that of the other three cultivated samples (CBB, FSB, and JAB), and this difference could be attributed to the high proportion of protopanaxadiol in wild-grown ginseng.

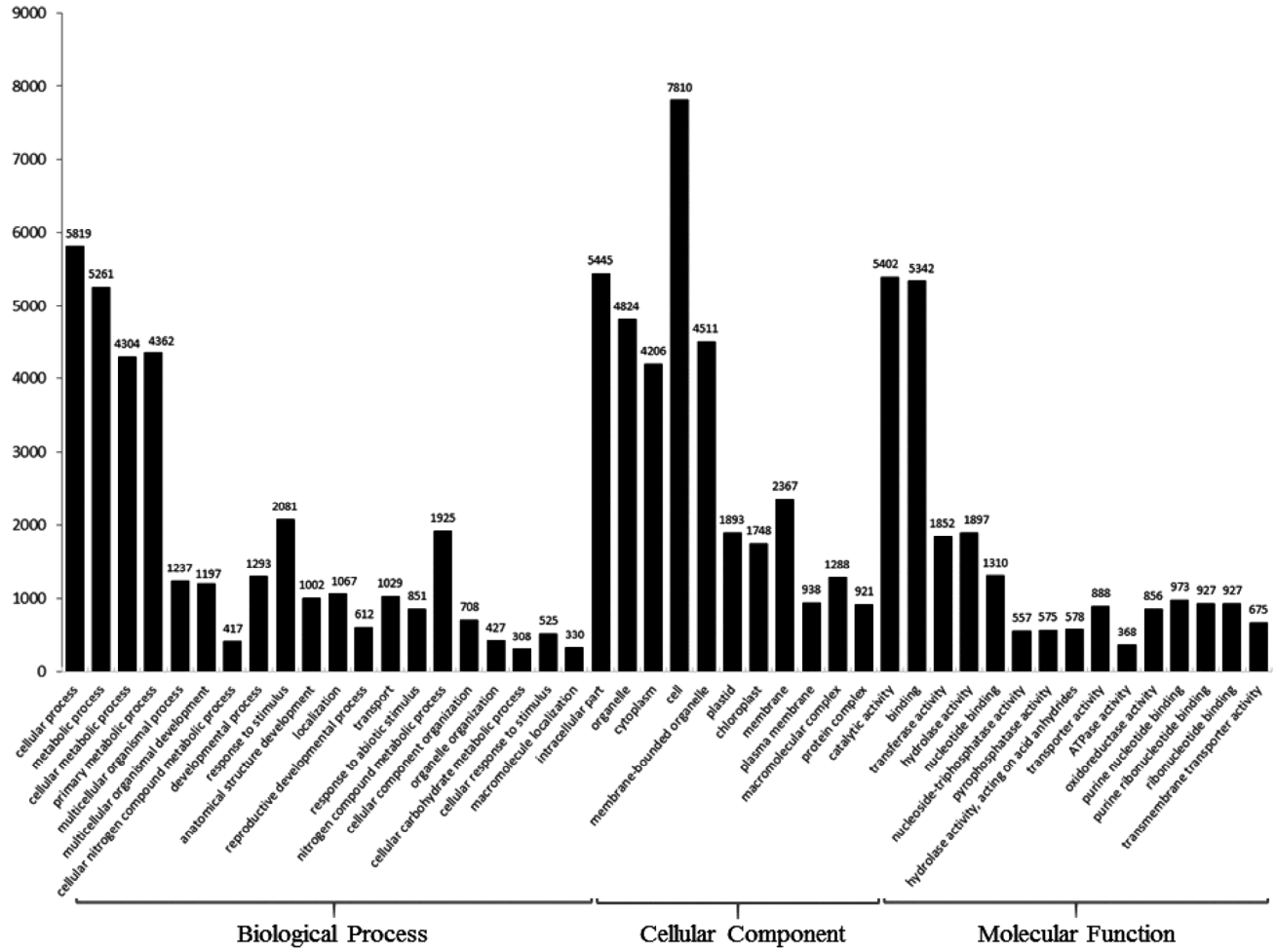

Figure 2 Functional annotation of our assembled transcriptome based on gene ontology (GO) analysis.

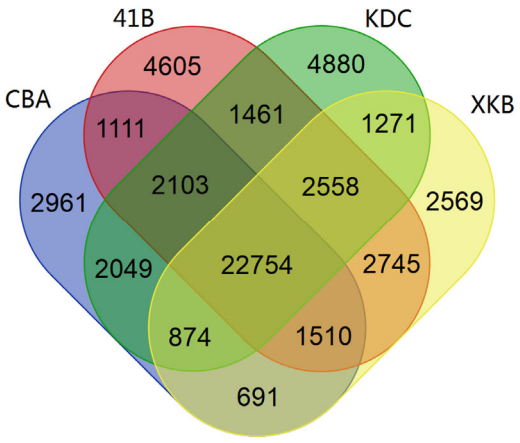

Figure 3 Venn diagram of expressed genes in the four ginseng varieties.

As ginsenosides were the most abundant of the collected biomedical components of ginseng, we attempted to identify all the ginsenoside biosynthesis-related genes in our assembly (Figure 5A). All the 13 reported ginsenoside biosynthesis-related genes were discovered when cytochrome p450s and glycosyltransferases that were possibly involved in the downstream pathway were excluded (Table 4 and Supporting file 3). All these genes contained at least two isoforms, and FPS had the most (37 isoforms).

Cytochrome p450s are very important during the downstream biosynthesis of dammarane-type ginsenosides, while certain glycosyltransferases catalyze both the downstream biosynthesis of dammarane-type ginsenosides and the biosynthesis of oleanane-type ginsenosides. In our assembly, we identified 242 genes (352 transcripts) that were annotated as cytochrome p450s and 249 genes (334 transcripts) that were annotated as glycosyltransferases. To identify potential cytochrome p450s and glycosyltransferases that might be related to the downstream ginsenoside biosynthesis, co-expression analysis was performed for these genes. The *R*-value was 0.9 for cytochrome p450 enzyme CYP716A47 and DS, indicating a close relationship between them; this result was expected because CYP716A47 was downstream

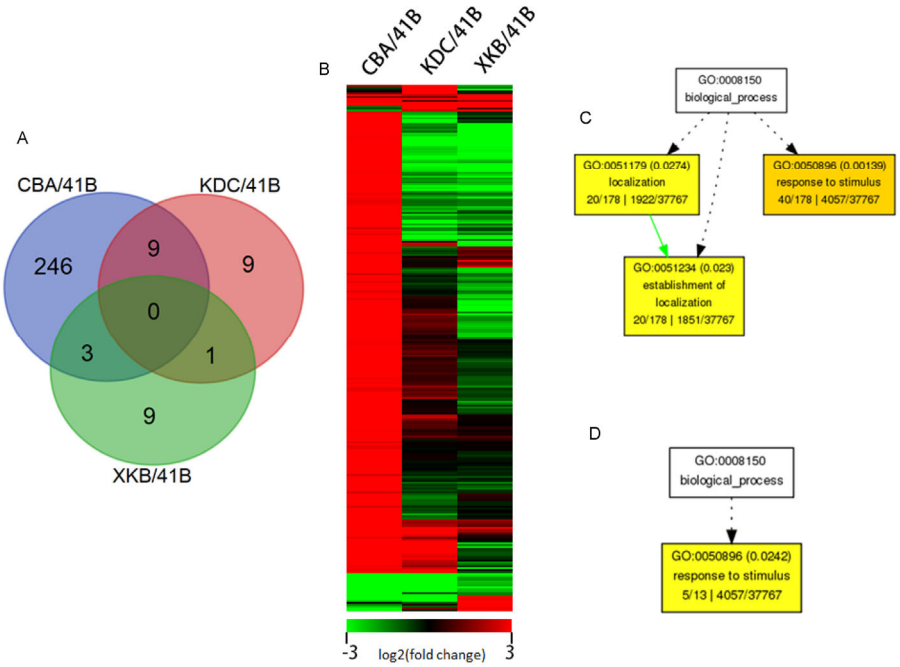

Figure 4 Differential expression pattern and gene ontology analysis of differentially expressed genes. A, Venn diagram of differentially expressed genes in the three comparison groups. B, Heatmap of all the differentially expressed genes in the three comparison groups. C, Results of gene ontology (GO) analysis of the differentially expressed genes in group CBA/41B. D, Results of GO analysis of the differentially expressed genes in group KDC/41B. For panels C and D, a green arrow means negative regulation, and a black dotted arrow means one significant node; in each box, the numbers in brackets represent the false discovery rate (FDR) of enrichment in that GO accession; the first part of the numbers in front of '|' in the bottom line represents the gene enrichment status in our sample, and the second part of the numbers represents the gene enrichment status in the background TAIR9 database.

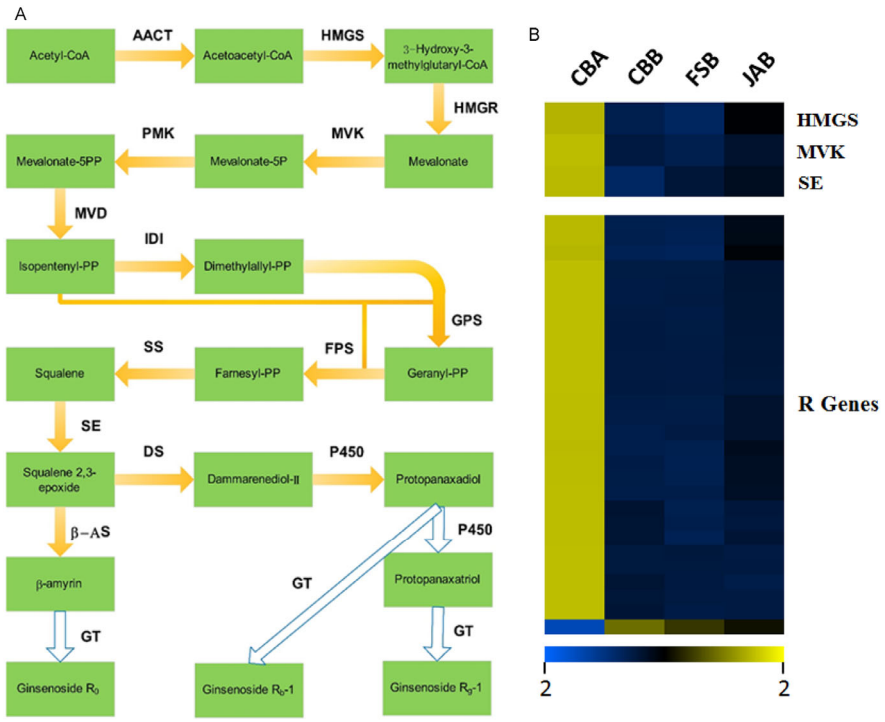

Figure 5 Putative ginsenoside biosynthesis pathway (A) and differential expression pattern of genes involved in the ginsenoside biosynthesis backbone and disease resistance (B). AACT, Acetyl-CoA acetyltransferase, HMGS, HMG-CoA synthase; HMGR, HMG-CoA reductase; MVK, mevalonate kinase; PMK, phosphomevalonate kinase; MVD, mevalonate diphosphate decarboxylase; IDI, isopentenyl diphosphate isomerase; GPS, geranylgeranyl pyrophosphate synthase; FPS, farnesyl diphosphate synthase; SS, squalene synthase; SE, squalene epoxidase; DS, dammarenediol-II synthase; β-AS, β-amyrin synthase; P450, cytochrome p450; GT, glycosyltransferase. The first three genes were involved in the ginsenoside biosynthesis backbone, and the last six genes were involved in disease resistance. We used the Z-score transformed relative expression values to draw this picture.

Table 4 Representatives of assembled genes related to ginsenoside biosynthesis

Gene ID	Blast hit*	*E*-value	Alias	KO Number	EC
comp44140_c2	sp|Q9FIK7|THIC2_ARATH	1.00×10^{-178}	AACT	K00626	2.3.1.9
comp48765_c0	sp|P54873|HMCS_ARATH	2.00×10^{-141}	HMGS	K01641	2.3.3.10
comp46397_c0	sp|O64967|HMDH2_GOSHI	0	HMGR	K00021	1.1.1.34
comp42967_c0	sp|P46086|KIME_ARATH	6.00×10^{-123}	MVK	K00869	2.7.1.36
comp44066_c0	AT1G31910	3.00×10^{-77}	PMK	K00938	2.7.4.2
comp45563_c0	sp|Q99JF5|MVD1_MOUSE	1.00×10^{-90}	MVD	K01597	4.1.1.33
comp42880_c0	sp|O48965|IDI2_CAMAC	2.00×10^{-126}	IDI	K01823	5.3.3.2
comp260010_c0	sp|Q7S565|COQ1_NEUCR	7.00×10^{-65}	GPS	K14066	2.5.1.1
comp320345_c0	sp|Q92235|FPPS_GIBFU	2.00×10^{-84}	FPS/DMAPP	K00787	2.5.1.1/10
comp50500_c0	sp|O24242|FPPS2_PARAR	1.00×10^{-92}	FPS/DMAPP		
comp48566_c0	sp|P53800|FDFT_NICBE	2.00×10^{-174}	SS	K00801	2.5.1.21
comp318783_c0	sp|P78589|FDFT_CANAX	9.00×10^{-63}	SS		
comp41754_c0	sp|O48651|ERG1_PANGI	0	SE	K00511	1.14.13.132
comp27590_c0	sp|O48651|ERG1_PANGI	0	SE		
comp47297_c1	sp|O48651|ERG1_PANGI	2.00×10^{-172}	SE		
comp46282_c0	sp|Q08IT1|DADIS_PANGI	0	DS	K15817	4.2.1.125
comp46752_c0	sp|H2DH16|C7A47_PANGI	2.00×10^{-162}	CYP716A47		
comp40861_c0	sp|O82140|BAMS1_PANGI	1.00×10^{-102}	beta-AS	K15813	5.4.99.39
comp40164_c1	sp|H2DH24|C7D47_PANGI	0	CYP82D47		
comp43002_c0	sp|H2DH21|C7A29_PANGI	2.00×10^{-21}	P450		
comp46709_c0	sp|Q9C788|C70B1_ARATH	8.00×10^{-49}	P450		
comp42593_c0	sp|B9DFU2|MAX1_ARATH	0	CYP711A1		
comp41918_c0	sp|Q8LDU5|PP298_ARATH	8.00×10^{-141}	CYP81D1		
comp48306_c1	sp|I7CT85|C7A53_PANGI	0	CYP716A53v2		
comp41434_c0	sp|P24465|C71A1_PERAE	2.00×10^{-107}	P450		
comp44789_c0	sp|H2DH17|C7A22_PANGI	8.00×10^{-176}	CYP72B1		
comp44704_c0	sp|P0C7P7|U74E1_ARATH	2.00×10^{-30}	GT		
comp43399_c0	sp|Q9FZ49|ALG9_ARATH	0	GT		
comp50219_c0	sp|Q66PF3|UFOG3_FRAAN	4.00×10^{-141}	UGT		
comp36369_c0	sp|Q6WFW1|GLT3_CROSA	6.00×10^{-91}	UGT		
comp42691_c0	sp|Q9ZQ96|U73C3_ARATH	4.00×10^{-135}	UGT		
comp43390_c0	sp|Q2V6K0|UFOG6_FRAAN	1.00×10^{-79}	GT		
comp50843_c0	sp|Q2V6K0|UFOG6_FRAAN	8.00×10^{-123}	UGT		
comp44452_c0	sp|Q9LMF0|U85A5_ARATH	5.00×10^{-54}	UGT		

*, These BLAST hit results were mainly derived from the BLAST results for our transcriptome against the SWISS-PROT protein database. Among those query genes, PMK did not produce a matched hit, so we used its BLAST hit result against the TAIR10 protein database instead. An entire set of all the genes related to ginsenoside biosynthesis is provided in Supporting file 3.

of DS. We then checked all identified cytochrome p450s and glycosyltransferases for their *R*-value with CYP716A47 and DS. For CYP716A47, four cytochrome p450s and seven glycosyltransferases showed a strong co-expression pattern (cutoff: absolute *R*-value>0.85), and most of them (except one cytochrome p450 and two glycosyltransferases) also showed a close relationship with DS (*R*-value>0.8). Among those four cytochrome p450s, one (comp50843_c0) was also identified by Wu et al. [17] in North American ginseng as being potentially related to dammarane-type ginsenoside biosynthesis. For DS, nine cytochrome p450s and seven glycosyltransferases showed a high expression correlation pattern, and among them, one cytochrome p450 (comp44704_c0), which catalyzes the glycosylation of certain kinds of protopanaxadiol [40] showed strong homology (*E*-value=3×10^{-23}) with *PgUGT74AE2* (GenBank: JX8985-29).

The genes involved in ginsenoside backbone biosynthesis were extracted and analyzed for their relative expression patterns. Using cutoffs including log2-fold change>2 and *P*-value<0.05, three genes, HMG-CoA synthase (*HMGS*), mevalonate kinase (*MVK*), and squalene epoxidase (SE), involved in ginsenosides biosynthesis backbone (Figure 5A) were detected as differentially expressed between CBA and 41B (Figure 5B). And all these three DEGs were up-regulated in CBA. Further realtime PCR analysis validated the upregulated expression pattern of the three genes in CBA (Supporting files 4 and 5).

2.5 Disease resistance-related genes were differentially expressed in wild-grown and cultivated ginseng

Disease resistance-related genes, or R genes, constitute a group of genes that confer resistance to pathogens by taking part in a sophisticated immune system in plants, and can be divided into several typical groups (CC-NBS-LRR (CNL), TIR-NBS-LRR (TNL), receptor-like kinase (RLK), receptor-like protein (RLP), kinase-like protein, *etc*.) based on their specific functional domains [41–43]. To identify differences in disease resistance-related gene expression between wild-grown and cultivated ginseng, we initially attempted to identify all R genes in our assembled transcriptome. As a result, 2,153 genes (5,000 transcripts) in our transcriptome were annotated as R genes (Supporting file 6), primarily falling into seven categories: RLP (611), N (442), TNL (306), NL (272), CNL (220), CN (75), and RLK (66). Among these genes, 28 were differentially expressed in CBA and 41B, while only one gene was differentially expressed in KDC and 41B, and one gene was differentially expressed in XKB and 41B (Figure 5B). The differentially expressed R genes in our samples primarily fell into five groups: TNL (6), NL (5), RLP (5), CNL (4), and N (4). Only one DEG was downregulated in CBA compared with 41B. The remaining 27 genes were upregulated in the wild-grown ginseng. Further real-time PCR analysis validated this high expression trend in CBA (Supporting files 4 and 5).

3 Discussion

P. ginseng C. A. Meyer is a very important herb, and modern science has demonstrated that the versatile medicinal properties of ginseng can be largely attributed to a group of compounds called ginsenosides. Although cultivated ginseng grows much faster than wild-grown ginseng and has played a major role in the modern ginseng market, its low medicinal value and high susceptibility to disease has caused many problems for ginseng farmers. In this research, we measured the ginsenoside content of wild-grown and cultivated *P. ginseng*, and the results showed that the wild-grown ginseng had a much higher protopanaxadiol (a type of ginsenoside) content than cultivated ginseng. Ginsenosides are the main bioactive agents in ginseng and are responsible for its medicinal properties. The difference in ginsenoside content between the wild-grown and cultivated ginseng reinforces the long-held belief that wild-grown ginseng is more medically effective than cultivated ginseng. To determine the underlying transcriptomic reasons for the differences in the ginsenoside content between the two forms, we carried out differential expression analysis to find possible DEGs involved in the ginsenoside biosynthesis backbone. Three important genes (*HMGS*, *MVK*, and *SE*) were identified as differentially expressed in CBA and 41B, and all were upregulated in CBA. This suggests a relationship between the upregulation of ginsenosides biosynthesis-related genes and the high ginsenoside content of wild-grown ginseng. It has been hypothesized in similar studies that there is an energy trade-off between primary and secondary metabolism, because wild-grown ginseng requires more energy to deal with many environmental stresses, whether biotic or abiotic. However, this study has improved our understanding of the mechanisms behind the ginsenoside content differences between wild-grown and cultivated ginseng. Further studies on the genes mentioned above may lead to improvements in cultivated ginseng when it is grown in artificial environments.

The other impending problem facing ginseng agriculture is the high susceptibility to diseases of cultivated ginseng. Our research identified a great many genes that were annotated as R genes in our assembled transcriptome, and this will be of great significance to similar future studies. Further differential gene expression analysis revealed that many of these R genes were differentially expressed in wild-grown and cultivated ginseng, and a very high proportion were upregulated in wild-grown ginseng. This result is consistent with the fact that cultivated ginseng is more susceptible to disease than wild-grown ginseng from the sixth year, and the high expression of R genes in wild-grown ginseng probably endows it with the ability to resist various kinds of pathogens from the surrounding environment. To date, although plant diseases have caused great damage to ginseng agriculture, few systematic studies have been carried out to determine the mechanism behind the high susceptibility to disease in cultivated ginseng. We now know a little more about the pathogens that cause severe diseases in cultivated ginseng, but detailed information about them is urgently needed. Future similar studies should pay particular attention to identifying host–pathogen interactions at the transcriptome level. Moreover, the specific R genes responsible for resistance to pathogens may be detected through differential expression analysis of wild-grown and cultivated ginseng in research that includes different developmental stages.

The authors declare that they have no conflict of interest.

We thank Yan Wei for advice on the analysis of disease resistance gene expression. This work was supported by the International Science and Technology Cooperation of China (2011DFA32730).

1 Baeg IH, So SH. The world ginseng market and the ginseng (Korea). J Ginseng Res, 2013, 37: 1–7
2 Goldstein B. Ginseng: its history, dispersion, and folk tradition. Am J Chin Med (Gard City N Y), 1975, 3: 223–234
3 Yun TK. Panax ginseng--a non-organ-specific cancer preventive? Lancet Oncol, 2001, 2: 49–55
4 Rausch WD, Liu S, Gille G, Radad K. Neuroprotective effects of

ginsenosides. Acta Neurobiol Exp (Wars), 2006, 66: 369–375
5 Choi JS, Chun KS, Kundu J, Kundu JK. Biochemical basis of cancer chemoprevention and/or chemotherapy with ginsenosides (Review). Int J Mol Med, 2013, 32: 1227–1238
6 Lu JM, Yao Q, Chen C. Ginseng compounds: an update on their molecular mechanisms and medical applications. Curr Vasc Pharmacol, 2009, 7: 293–302
7 Karmazyn M, Moey M, Gan XT. Therapeutic potential of ginseng in the management of cardiovascular disorders. Drugs, 2011, 71: 1989–2008
8 Gu J, Li W, Xiao D, Wei S, Cui W, Chen W, Hu Y, Bi X, Kim Y, Li J, Du H, Zhang M, Chen L. Compound K, a final intestinal metabolite of ginsenosides, enhances insulin secretion in MIN6 pancreatic beta-cells by upregulation of GLUT2. Fitoterapia, 2013, 87: 84–88
9 Xu Y, Lin L, Tang L, Zheng M, Ma Y, Huang L, Meng W, Wang W. Notoginsenoside R1 attenuates hypoxia and hypercapnia-induced vasoconstriction in isolated rat pulmonary arterial rings by reducing the expression of ERK. Am J Chin Med, 2014, 42: 799–816
10 Lee KS, Kim GH, Kim HH, Chang YI, Lee GH. Volatile compounds of Panax ginseng C.A. Meyer cultured with different cultivation methods. J Food Sci, 2012, 77: C805–810
11 Kwon KR, Park WP, Kang WM, Jeon EY, Jang JH. Identification and analysis of differentially expressed genes in mountain cultivated ginseng and mountain wild ginseng. J Acupunct Meridian Stud, 2011, 4: 123–128
12 Xu G, Fan X, Miller AJ. Plant nitrogen assimilation and use efficiency. Annu Rev Plant Biol, 2012, 63: 153–182
13 Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S. De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics, 2010, 11: 262
14 Chen S, Luo H, Li Y, Sun Y, Wu Q, Niu Y, Song J, Lv A, Zhu Y, Sun C, Steinmetz A, Qian Z. 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in Panax ginseng. Plant Cell Rep, 2011, 30: 1593–1601
15 Luo H, Sun C, Sun Y, Wu Q, Li Y, Song J, Niu Y, Cheng X, Xu H, Li C, Liu J, Steinmetz A, Chen S. Analysis of the transcriptome of Panax notoginseng root uncovers putative triterpene saponin-biosynthetic genes and genetic markers. BMC Genomics, 2011, 12 Suppl 5: S5
16 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
17 Wu D, Austin RS, Zhou S, Brown D. The root transcriptome for North American ginseng assembled and profiled across seasonal development. BMC Genomics, 2013, 14: 564
18 Subramaniyam S, Mathiyalagan R, Natarajan S, Kim YJ, Jang MG, Park JH, Yang DC. Transcript expression profiling for adventitious roots of Panax ginseng Meyer. Gene, 2014, 546: 89–96
19 Jayakodi M, Lee SC, Park HS, Jang W, Lee YS, Choi BS, Nah GJ, Kim DS, Natesan S, Sun C, Yang TJ. Transcriptome profiling and comparative analysis of Panax ginseng adventitious roots. J Ginseng Res, 2014, 38: 278–288
20 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
21 Liu MH, Yang BR, Cheung WF, Yang KY, Zhou HF, Kwok JS, Liu GC, Li XF, Zhong S, Lee SM, Tsui SK. Transcriptome analysis of leaves, roots and flowers of Panax notoginseng identifies genes involved in ginsenoside and alkaloid biosynthesis. BMC Genomics, 2015, 16: 265
22 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol, 2011, 29: 644–652
23 Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics, 2010, 26: 680–682
24 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods, 2012, 9: 357–359
25 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
26 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc, 2013, 8: 1494–1512
27 Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol, 2011, 7: e1002195
28 Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res, 2010, 38: W64–70
29 Sanseverino W, Hermoso A, D'Alessandro R, Vlasova A, Andolfo G, Frusciante L, Lowy E, Roma G, Ercolano MR. PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants. Nucleic Acids Res, 2013, 41: D1167–1171
30 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: 2062–2073
31 Davidson NM, Oshlack A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol, 2014, 15: 410
32 Gonzalez E, Joly S. Impact of RNA-seq attributes on false positive rates in differential expression analysis of de novo assembled transcriptomes. BMC Res Notes, 2013, 6: 503
33 Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods, 2013, 10: 71-73
34 Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol, 2010, 11: R106
35 Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods, 2001, 25: 402–408
36 Zhang CY, Dong L, Chen SL, Xie CX, Chang DL. UPLC fingerprint for quality assessment of ginsenosides of ginseng radix et rhizoma (in chinese). Yao Xue Xue Bao, 2010, 45: 1296–1300
37 Barkan A, Small I. Pentatricopeptide repeat proteins in plants. Annu Rev Plant Biol, 2014, 65: 415–442
38 Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol, 2001, 11: 725–732
39 Li D, Roberts R. WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. Cell Mol Life Sci, 2001, 58: 2085–2097
40 Jung SC, Kim W, Park SC, Jeong J, Park MK, Lim S, Lee Y, Im WT, Lee JH, Choi G, Kim SC. Two Ginseng UDP-Glycosyltransferases Synthesize Ginsenoside Rg3 and Rd. Plant Cell Physiol, 2014, 55: 2177–2188
41 Ellis J, Dodds P, Pryor T. Structure, function and evolution of plant disease resistance genes. Curr Opin Plant Biol, 2000, 3: 278–284
42 Sanseverino W, Ercolano MR. In silico approach to predict candidate R proteins and to define their domain architecture. BMC Res Notes, 2012, 5: 678
43 Belkhadir Y, Subramaniam R, Dangl JL. Plant disease resistance protein signaling: NBS-LRR proteins and their partners. Curr Opin Plant Biol, 2004, 7: 391–399

Biographical Sketch

Deng XingWang is a university endowed professor of plant biology at Peking University. He graduated from Peking University with B.S. (1982) and M.S. (1982) degrees, and then graduated from University of California at Berkeley in 1989 with Ph.D. degree in plant biology. Before moving back to China, he was a faculty of Yale since 1992, and promoted to Daniel C. Eaton Professor of Yale University from 2003 till 2014. He worked on signaling process in plant photomorphogenesis, noncoding RNAs, heterosis and molecular design breeding in plant. He has been awarded the Kumho Science International Award by the International Society for Plant Molecular Biology (ISPMB) in 2003 and elected Member of US National Academy of Sciences in 2013.

Supporting Information

Supporting file 1 Top 100 hit Pfam accessions of all the domains in the open reading frames (ORFs) of our assembly.

Supporting file 2 Ginsenoside content of different kinds of *Panax ginseng* grown in different environments.

Supporting file 3 Sequences of all ginsenoside biosynthesis-related genes in our assembly.

Supporting file 4 Real-time polymerase chain reaction (PCR) analysis results. The first three genes are involved in the ginsenoside biosynthesis backbone, and the last six genes are involved in disease resistance.

Supporting file 5 Sequences of primers for real-time polymerase chain reaction (PCR) analysis.

Supporting file 6 Sequences of all disease resistance genes in our assembly.

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.