Physiol Mol Biol Plants (October 2021) 27(10):2215–2229
https://doi.org/10.1007/s12298-021-01076-1

REVIEW ARTICLE

De novo whole-genome assembly and discovery of genes involved in triterpenoid saponin biosynthesis of Vietnamese ginseng (*Panax vietnamensis* Ha *et* Grushv.)

Nguyen Quang Duc Tien[1,2] · Xiao Ma[5] · Le Quang Man[1] · Duong Thi Kim Chi[1] · Nguyen Xuan Huy[3] · Duong-Tan Nhut[4] · Stephane Rombauts[5] · Tran Ut[6] · Nguyen Hoang Loc[1,2]

Received: 12 August 2021 / Revised: 13 September 2021 / Accepted: 17 September 2021 / Published online: 11 October 2021

Abstract Vietnamese ginseng (*Panax vietnamensis* Ha *et* Grushv.), also known as Ngoc Linh ginseng, is a high-value herb in Vietnam. Vietnamese ginseng has been proven to be effective in enhancing the immune system, human memory, anti-stress, anti-inflammatory, anti-cancer, and prevent aging. The present study reports the first draft whole-genome of Vietnamese ginseng and the identification of potential genes involved in the triterpenoid metabolic pathway. De novo whole-genome assembly was performed successfully from a data of approximately 139 Gbps of 394,802,120 high quality reads to generate 9815 scaffolds with an N50 value of 572,722 bp from the leaf of Vietnamese ginseng. The assembled genome of Vietnamese ginseng is 3,001,967,204 bp long containing 79,374 gene models. Among them, there are 55,012 genes (69.30%) were annotated by various public molecular biology databases. The potential genes involved in triterpenoid saponin biosynthesis in Vietnamese ginseng and their metabolic pathway were also predicted.“ Three genes encoding squalene monooxygenase isozymes in Vietnamese ginseng were cloned, sequenced and characterized. Moreover, expression levels of several key genes involved in terpenoid biosynthesis in different parts of Vietnamese ginseng were also analyzed. The SSR markers were detected by various programs from both of assembly full dataset of Vietnamese ginseng genome and predicted genes. The present work provided important data of the draft whole-genome of Vietnamese ginseng for further studies to understand the role of genes involved in ginsenoside biosynthesis and their metabolic pathway at the molecular level of this rare medicinal species.

Keywords Vietnamese ginseng · Genome assembly · *Panax vietnamensis* Ha *et* Grushv. · Gene expression · Whole-genome sequencing

✉ Nguyen Hoang Loc
nhloc@hueuni.edu.vn

1 Bioactive Compound Institute, University of Sciences, Hue University, Hue 530000, Vietnam

2 Department of Biology, Bioactive Compound Institute, University of Sciences, Hue University, Hue 530000, Vietnam

3 Hue University, Hue 530000, Vietnam

4 Tay Nguyen Institute of Scientific Research, Vietnam Academy of Science and Technology, Dalat 670000, Vietnam

5 VIB-UGent Center for Plant Systems Biology, Ghent University, 9000 Ghent, Belgium

6 Ngoc Linh Ginseng and Medicinal Materials Development Center, Quang Nam, Quang Ngai 51000, Vietnam

Introduction

Vietnamese ginseng (*Panax vietnamensis* Ha *et* Grushv.), also known as Ngoc Linh ginseng, was discovered in 1973 in Mount Ngoc Linh, Central Vietnam (Duc et al. 1999). It is a famous medicinal plant in Vietnam that contains a higher saponin content than other ginseng species such as Korean ginseng (*P. ginseng*), Chinese ginseng (*P. notoginseng*), and American ginseng (*P. quinquefolius*) (Duc et al. 1999; Huong et al. 1995). In Vietnam, this species is highly valued in medicinal herbs, is very commercially valuable and is currently considered threatened (Wikipedia, retrieved 10 Nov 2020). Thirty-seven triterpene saponins including 14 new compounds were isolated from the underground parts of Vietnamese ginseng. Some of them were common to other *Panax* spp., but the general yields in

this species were very high (Yamasaki et al. 2000). This medicinal species also contains a large amount of ocotillol-type saponins whose main ingredient is majonoside-R2 (MR2) with high content of over 5% (Yamasaki et al. 2000, Zhang et al. 2015). Ocotillol-type saponins are a rare class of ginsenosides that are rarely found in natural products (Liu J et al. 2017). Ocotillol saponins from Vietnamese ginseng exhibited a strong anti-tumour activity on two-stage carcinogenesis test of mouse hepatic tumour (Konoshima et al. 1999).

Ginsenosides are high-value pharmaceutical compounds, found in *Panax* spp. with structural diversity and a variety of biological activities. Ginsenoside is pharmacological effective in memory enhancement, anti-stress, anti-inflammatory, and anti-aging (Duc et al. 1999, Nguyen B et al. 2017, Zhang et al. 2015). Recently, some studies have reported on anti-cancer drugs developed from hydrolyzed ginsenosides (Han et al. 2016; Mai et al. 2012; Wang et al. 2008). Therefore, ginsenoside is being considered a promising natural material source in the pharmaceutical and cosmetic industries. A study on the genome and the identification of genes encoding enzymes involved in the triterpenoid saponin biosynthetic pathway in ginseng species is of great interest in the field of biotechnology. Squalene epoxidase (SE), also called squalene monooxygenase, is a key enzyme for ocotillol-type saponin biosynthesis, which plays an important regulatory role in the ginsenoside metabolic pathway (Han et al. 2010). The genes encoding SE isozymes have been identified and characterized in some *Panax* species such as *P. ginseng* (Han et al. 2016), *P. notoginseng* (He et al. 2008; Liu et al. 2015; Luo et al. 2011; Niu et al. 2014), and *P. vietnamensis* var. *fuscidiscus* (Lai Chau ginseng) (Zhang et al. 2015). Recently, the transcriptome and chloroplast genome of Vietnamese ginseng has been also analyzed by using next-generation sequencing (Nguyen B et al. 2017, Vu et al. 2020). However, information on their whole-genome database and the sequence of genes involved in the triterpene saponin biosynthetic pathway has not been reported.

The present study reports for the first time the whole-genome sequencing, assembly, and annotation of Vietnamese ginseng. Genes involved in saponin biosynthesis and their metabolic pathway in Vietnamese ginseng have also been predicted. Three key genes (*PvH_SE*) encoding squalene epoxidase were identified by DNA full-length cloning and sequencing. Their genetic characteristics and phylogenetics also were described. Furthermore, we investigated the expression pattern of these genes in various parts of Vietnamese ginseng. The result will provide important information of the whole-genome of Vietnamese ginseng to more understanding at the molecular level regarding the role of genes involved in triterpene saponins biosynthetic pathway.

Materials and methods

Plant materials

Six-year-old *Panax vietnamensis* Ha *et* Grushv. (Fig. 1) were identified and provided by Mr Tran Ut (Director of Tra Linh Medicinal Station of Center for Developing Ngoc Linh Ginseng and Medicinal Herbs, Quang Nam province, Vietnam). The research sample collection was carried out in accordance with relevant guidelines and regulations of Vietnam. After cleaning, the parts such as root, stem, and leaf were put into zip bags, flash freezing with liquid nitrogen, and stored at -80 °C for further experiments.

Construction of genomic DNA library

Genomic DNA of Vietnamese ginseng was extracted from leaves by using the GeneJET Plant Genomic DNA Purification Kit (ThermoFisher, USA). The quality of genomic DNA was confirmed before preparing of DNA library with KAPA Hyper Prep Kits following the manufacturer's instructions (Roche, USA). The sequencing library was prepared by random fragmentation of the DNA sample, followed by 5' and 3' adapter ligation. Alternatively, "tagmentation" combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments were then amplified by PCR and followed by gel purification. The quality of the prepared sequencing library was evaluated by Qubit 2.0 Fluorometer using the KAPA Library Quantification Kit.

Whole-genome sequencing and assembly

The whole genome of Vietnamese ginseng was sequenced by HiSeq Illumina at a depth of approximately 40 × with a reading length of 2 × 150 paired-end and 4 lanes allocation (Admera Health, USA). The clean reads of Vietnamese ginseng were assembled by CLC Assembly Cell (Qiagen, v5.0.12), a guided-assembly method for the generation of the reads with long high-quality contigs and/or scaffolds by using the Korean ginseng genome (Jayakodi et al. 2018) as the reference. All the clean paired-end reads were mapped to the reference genome by using "clc_mapper" from the CLC Assembly Cell program with the parameters as follows "fb ss 50 250". The contigs were got by "clc_extract_consensus" module of the software. Finally, The GapCloser software (https://sourceforge.net/projects/soapdenovo2/files/GapCloser/) was used to fill gaps using the parameters including maximum read length of 100 bp, overlap param of 25, and thread number are 10.

Fig. 1 Vietnamese ginseng (*P. vietnamensis* Ha *et* Grushv.)

The quality of whole-genome assembled was confirmed by BUSCO before performing further analysis.

Genome annotation and prediction of the metabolic pathway for ginsenoside biosynthesis

RepeatModeler (v2.0) was used to identify the repeat families in the assembly genome of Vietnamese ginseng. RepeatMasker (v4.1) was used to discover and classify repeats based on the custom repeating libraries constructed by RepeatModeler (v2.0). Gene models were predicted by a combination of ab initio prediction and homology search. BRAKER2 was used for ab initio gene prediction using model training based on proteins of very close homology from Korean ginseng after the annotated repeats were soft masked in the assembly. For homology prediction, protein sequences from other five related species (*Arabidopsis thaliana*, *Oryza sativa*, *Solanum tuberosum*, *Brassica rapa*, and *Daucus carota*) together with Korean ginseng were used as query sequences to search the reference genome using TBLASTN with different e-values (*A. thaliana*, *O. sativa*, *S. tuberosum*, *B. rapa* and *D. carota* with e-value $\leq 1e^{-5}$, Korean ginseng with e-value $\leq 1e^{-10}$). Regions mapped by these query sequences were subjected to Exonerate. Finally, EvidenceModeler (v1.1.1) was used to integrate all of the above pieces of evidence based on different weights. All gene models were annotated by using OmicsBOX (ver 2.0.29) with the different molecular biology databases (NR, Swiss-Prot, EggNOG, KEGG) and Hayai Annotation Plants web server (http://pgdbjsnp.kazusa.or.jp/app/hayai2). Furthermore, pathway analysis was performed based on the combination of KEGG (https://www.genome.jp/kegg/) and Reactome (https://reactome.org) databases. The listing ID of KEGG Orthology (KO) was obtained and then mapped to the KEGG pathway database for visualization of the triterpenoid backbone biosynthesis pathways (Moriya et al. 2007). The deduced amino acid sequences of species including Vietnamese ginseng, *P. ginseng*, *D. carota*, *A. thaliana*, and *B. rapa* were used to compare species relationships based on orthologous clusters analysis by Orthovenn2 (Xu et al. 2019). The amino acid sequences of single-copy gene clusters were extracted and multiple sequences aligned using MUSCLE (http://www.drive5.com/muscle/) before removing the poorly aligned region by GBLOCKs (http://molevol.cmima.csic.es/castresana/Gblocks.html). The super-gene for each species was generated from high-quality blocks, which are to be used for the construction of phylogenetic tree by MEGAX.

Functional annotation and GO terms assignment

Functional annotations were performed by comparing the sequences of the predicted genes with public databases included the NCBI non-redundant protein database (NR, modified in August 2021) (http://www.ncbi.nlm.nih.gov/), Swiss-Prot database (modified in August 2021) and the Clusters of Orthologous Groups database (http://www.ncbi.nlm.nihgov/COG/) by Blastx (with an e value cut off of $1e^{-5}$) and annotated by OmicsBOX (ver 2.0.29) with default settings. Based on the GO annotation result from OmicsBOX, the top 20 gene ontology terms from three categories (biological process, molecular function, and cellular component) at level 3 was used for visualization of the functional GO classification diagram by WEGO 2.0 (http://wego.genomics.cn).

Identification of genes involved in the ginsenoside biosynthesis pathway

The predicted gene models were obtained by OmicsBOX based on the top hit after excluding those less than 100

amino acid residues in length. The predicted protein sequences were again analyzed with the Pfam database and KEGG Automatic Annotation Server (https://www.genome.jp/kegg/kaas/) with an E value cutoff of 1e − 5. Candidate Cytochrome P450 (P450s) and UDP glycosyltransferase (UGTs) genes which are responsible for the production of various types of ginsenosides in the final step of this pathway, were also identified based on Blastp (E value cutoff of 1e − 5 and percentage identity cutoff of 40%) against Pfam and KEGG databases. The deduced amino acid sequences of putative P450 and UGT genes of various species were multiply aligned by MUSCLE and constructed a phylogenetic tree by MEGAX by using the pairwise distance with 1000 bootstrap replicates and visualized by iTOL (https://itol.embl.de). The putative pathway and the related genes can be accessed in Table 3.

Cloning and characterization of squalene monooxygenase

Total RNA and genomic DNA were extracted from the root tissues of Vietnamese ginseng by Trizol reagent and GeneJET Plant Genomic DNA Purification Kit, respectively. Three genes (isoforms) encoding squalene monooxygenase (*PvH_SE*) were isolated by RT-PCR or PCR amplification (Thermo Scientific, USA) with *PvH_SE* gene-specific primers following the manufacturer's instructions. Three *PvH_SE* gene sequences were obtained from a dataset of the draft genome of Vietnamese ginseng. Their DNA flanking sequences were used to design the specific primers (Supplementary Table T1) using Primer3 (Version 4.1.0, http://primer3.ut.ee). PCR amplifications were performed in a total volume of 50 µL containing 50 ng of genomic DNA or cDNA, 10 pmol of each primer, 200 µM of dNTPs, 5 uL of 10 × Taq polymerase buffer, and 1.25 U of DreamTaq DNA Polymerase (Thermo Scientific, USA). The PCR program was as follows: 94 °C for 5 min, 35 cycles of 94 °C for 1 min, 57 °C for 55 s and 73 °C for 3 min, and a final extension at 72 °C for 10 min. The PCR products were purified by GeneJET Gel Extraction Kit (Thermo Scientific, USA), and were then cloned to pGEM-T Easy vector (Promega, USA). The successfully transformed clones were identified by using restriction digestion before conducting DNA sequencing with T7 and SP6 universal primers. Multiple sequence alignment was carried out by ClustalW with the default parameters. The phylogenetic tree was constructed by MEGA X following the neighbour-joining method (NJ) with 1000 bootstrap replicates. The signal peptide was predicted by Signal-3L 2.0 software (Shen et al. 2007) based on Neural networks (NN) and Hidden Markov models (HMM) databases of eukaryotes with the best confidence value.

RT-PCR analysis

Total RNA (1 µg) was isolated from different parts (root, leaf and stem) of Vietnamese ginseng by GeneJET Plant RNA Purification Kit (ThermoFisher, USA). RNA quality was confirmed by the OD 260/280 ratio and agarose electrophoresis. The first-strand cDNA has been synthesized from 1 µg of total RNA by RevertAid First-Strand cDNA Synthesis Kit (Thermo Scientific, USA) with oligo(dT) 18 primer. Two microliters from the first-strand cDNA synthesis were used as a template to amplify PCR in a total volume of 50 µL with specific primers for the consensus of several genes involved in the triterpenoid biosynthesis pathway. The "housekeeping" genes (*GAPDH and PMK)* were used as a reference gene (Zhang et al. 2015). A list of primers was shown in Supplementary Table T1. The RT-PCR products were analyzed along with DNA ladder marker (1 kb plus, ThermoFisher, USA) on 1% agarose gels RT-PCR products were analyzed by ImageJ software (ver. 1.53d) and Graphpad 7.

SSR marker detection and primer design

Potential simple sequence repeat (SSR) markers were detected in the genome assembly dataset and 79,374 annotated genes by MISA PERL script (Beier et al. 2017), Krait (v1.3.3) (Du et al. 2018), and SciRoKo (v3.3) (Kofler et al. 2007). The SSRs were searched with motifs ranging from mono- to hexanucleotides in size. The minimum of repeat units was set as follows: ten repeat units for mononucleotide, six for di-nucleotides, and five for tri-, tetra-, penta- and hexanucleotides. The scaffolds containing motif of di and tri-nucleotides with more than 8 repeats were employed for primer pairs design by using Primer3 (http://bioinfo.ut.ee/ primer3-0.4.0/primer3/) with default parameters.

Results

Illumina sequencing and de novo assembly

A DNA library was constructed from genomic DNA of Vietnam ginseng leaves and sequenced using Illumina 150 paired-end sequencing technology (Admera Health, USA). After removal of adaptor sequences, ambiguous reads and low-quality read (Q20 < 20), a total of about 394 million clean reads were obtained after trimming bad quality read. An overview of the sequencing and assembly statistics is shown in Table 1. The high-quality reads obtained in this study are being deposited in the NCBI SRA database (SRA accession number: PRJNA665343). All the clean reads (about 394 million) were de novo assembled using the CLC

Table 1 Genome assembly and annotation statistics

Genome assembled	
Number of scaffold	9815
Scaffolds(> = 0 bp)	9837
Scaffolds(> = 100 bp)	9769
Largest scaffolds	3,664,226
Total length of scaffolds(bp)	3,001,961,343
Total length of scaffolds(> = 0 bp)	3,001,967,204
Total length of scaffolds(> = 1000 bp)	3,001,928,631
N50 of scaffolds(bp)	572,722
GC content(%)	35.11
Genome annotation	
Number of gene	79,364
Mean gene length,bp	3818
Mean exon length (bp)	230
Mean intron length (bp)	811
Mean gene locus size (first to last exon)	3820.9
Mean number of distinct exons per gene	4.5
Mean transcript size (UTR,CDS)	1022
Number of distinct exons	353,719
Number of multi-exon gene	57,781(72.8%)
Number of single –exon genes	21,583(27.2%)
Transposable elements annotatuion	(%)
Overall TE content	62.63
Class 1 LTR Gypsy/copia	31.15/5/91
Class 1 LINE	0.26
Class 2 DNA transposon	4.15
Simple repeats	11.32
Unclassified	9.49
Completeness assessment result	
Total number of core genes queried	1614
Number of core genes detected	
Complete	1491(92.4%)
Complete + Partial	1577(97.7%)
Number of missing core genes	37(2.3%)
Scores in BUSCO format*	C:92.4%[S:22.7%,D:69.7%],F:5.3%,M:2.3.%

assembly cell program (QIAGEN). The present study has already succeeded to make a guided assembly by using the *Panax ginseng* (Korea ginseng) as a reference via mapping Vietnam ginseng paired-end data. This resulted in scaffolds comparable with the Korean ginseng assembly in terms of length and number of sequences (Supplementary Figures S1-2). Due to the sequence divergence between the Vietnamese and Korean ginseng varieties and the stringent parameters for the mapping, our guided assembly of the Vietnamese ginseng is so much gaped. Therefore, gap-closing software with the Vietnamese paired-end data was employed for the generation of 9,815 scaffolds containing 9,769 long scaffolds ($\geq$ 1000 bp) with a total length of 3,001,961,343 bp. The largest size of the scaffold is 3,664,226 bp, and N50 length of 572,722 bp. Based on JCVI utility libraries v1.0.5 (Tang et al. 2015) on genome assembly, annotation, and comparative genomics (*Panax ginseng*) the result showed that the mean exon size is 229.3 bp with 4.5 exons per gene (Table 1). The completeness assessment results of genome assembly by BUSCO showed that we successfully generated a draft genome assembly of Vietnamese ginseng with high quality.

Genome annotation

Transposable elements (TEs) of the Vietnamese ginseng genome were annotated for approximately 62.63% of the *A. cruentus* genome, which contains 37.41% of long terminal repeat (LTR) elements (Table 1). After removing the transposon element sequences, the Vietnamese ginseng

genome was analyzed by EvidenceModeler (EVM) integrated with BRAKER2 (ab initio prediction) and homology prediction. The study predicted 79,374 protein-encoding gene models after masking the repeating elements, which is higher than that of Korean ginseng (59,352) and Chinese ginseng (64,742) but it is slightly lower than that of Lai Chau ginseng (84,004 unigenes) (Zhang et al. 2015) and Vietnamese ginseng (89,271 unigenes) (Vu et al. 2020). Statistics on genome annotation of Vietnamese ginseng in this study can be found in Table 1. BUSCO analysis of these predicted protein sequences based on the embryophyte database showed that 92.4% of the BUSCO genes were found to be complete, which could be increased to 97.7% with the partial BUSCO genes added. 22.7% of the BUSCO genes were single copies and 69.7% of the BUSCO genes were identified as duplicates. Besides, 5.3% of the BUSCO genes were fragmented and 2.3% were missing (Table 1). Based on the BUSCO analysis result, 97.7% of the proteins were annotated, suggesting the accuracy of the genome assembly. To further check the quality of the genome annotation, the lengths of the corresponding coding sequences (CDS) between Korean ginseng and Vietnamese ginseng were compared using reciprocal best hits (Supplementary Figure S1). The distributions of mRNA length, CDS length, exon length, and intron length in Korean ginseng and Vietnamese ginseng genome assembly were compared and presents in Supplementary Figure S2. A total of 79,374 predicted genes of Vietnamensis ginseng annotated by public databases which are showed different percentages of annotation (Table 2). The OmicsBox (ver 2.0.29) successfully annotated the functions of 55,012 of the 79,374 gen models (69.30% of total annotation) (Fig. 2). The top annotated gene ontology terms at the level 3 annotation with three categories (biological process, molecular function, and cellular component) are visualized by WEGO 2.0, which are shown in Fig. 3. The orthologous group analysis of multiple genomes was performed using Orthovenn2 software (Xu et al. 2019) based on predicted protein sequences. The result indicated that 79,374 protein sequences of Vietnamese ginseng were clustered into 27,849 orthologous groups and 2,107 singleton gene clusters. Meanwhile, a total of 59,352 protein sequences of Korean ginseng were clustered into

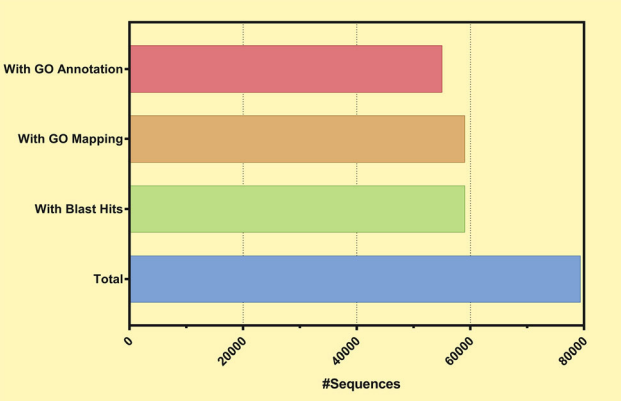


Fig. 2 Blast2Go functional annotation summary

24,024 orthologous groups (Fig. 4b). Among 22,326 common orthologous groups predicted between two ginseng species, Vietnamese ginseng has 1,314 specific groups and Korean ginseng has 5,523 specific groups (Supplementary Figure S3). An orthologous group analysis between Vietnamese ginseng and four other plant species, including Korean ginseng, *A. thaliana*, *B. rapa* and *D. carota*, found 9,504 clusters that shared all five species. Among them, Vietnamese ginseng, Korean ginseng and *D. carota* had the highest number of common clusters (2,574). The pairwise heatmap of overlapping cluster numbers between each pair of various genomes was shown in Supplementary Figure S4. The result showed that these species formed 3,8117 clusters, contains 3,7210 orthologous clusters (at least contains two species) and 907 single-copy gene clusters. (Fig. 4a). The amino acid sequences of these single-copy gene clusters were used to construct the phylogenetic tree by using the maximum-likelihood method (Fig. 4c). The result showed that the super-gene of single-copy gene clusters of Vietnamese ginseng and Korean ginseng belong to the same group of the phylogenetic tree. The super-gene of these two species shows 90.4% similarity by pairwise sequence alignment method. Besides, among all the compared genome species, *D. carota* showed closest relative to Vietnamese and Korean ginseng than *A. thaliana*, and *B. rapa*, which showed similar results to a recent study (Xu J et al. 2017).

Table 2 Genome annotation with various databases

Tools	Databases	Number of genes	Percentage (%)
OmicsBox	Nr	54,706	68.9%
	Swiss-Prot	50,412	63.5%
	EggNOG	36,926	46.52%
Hayai Annotation Plants	KusakiDB	49,897	62.87%
Kyoto Encyclopedia of Genes and Genomes	KEGG	29,212	36.8%
Total		55,012	69.30%

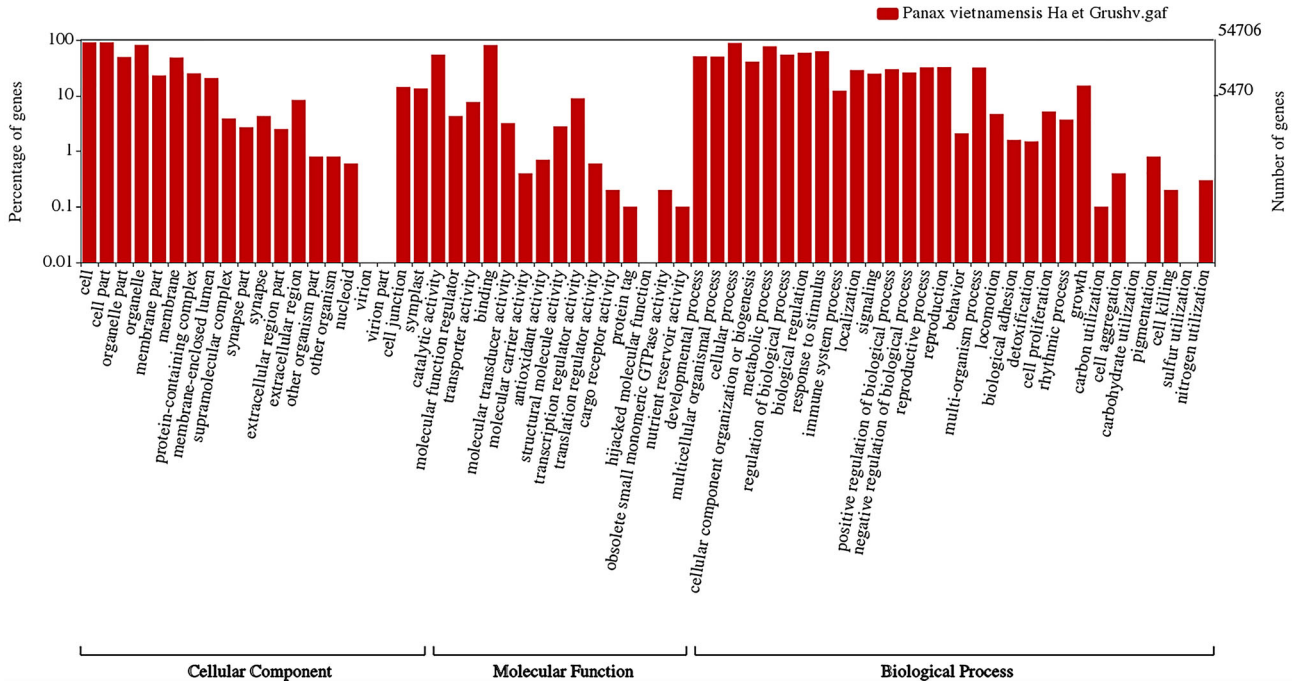


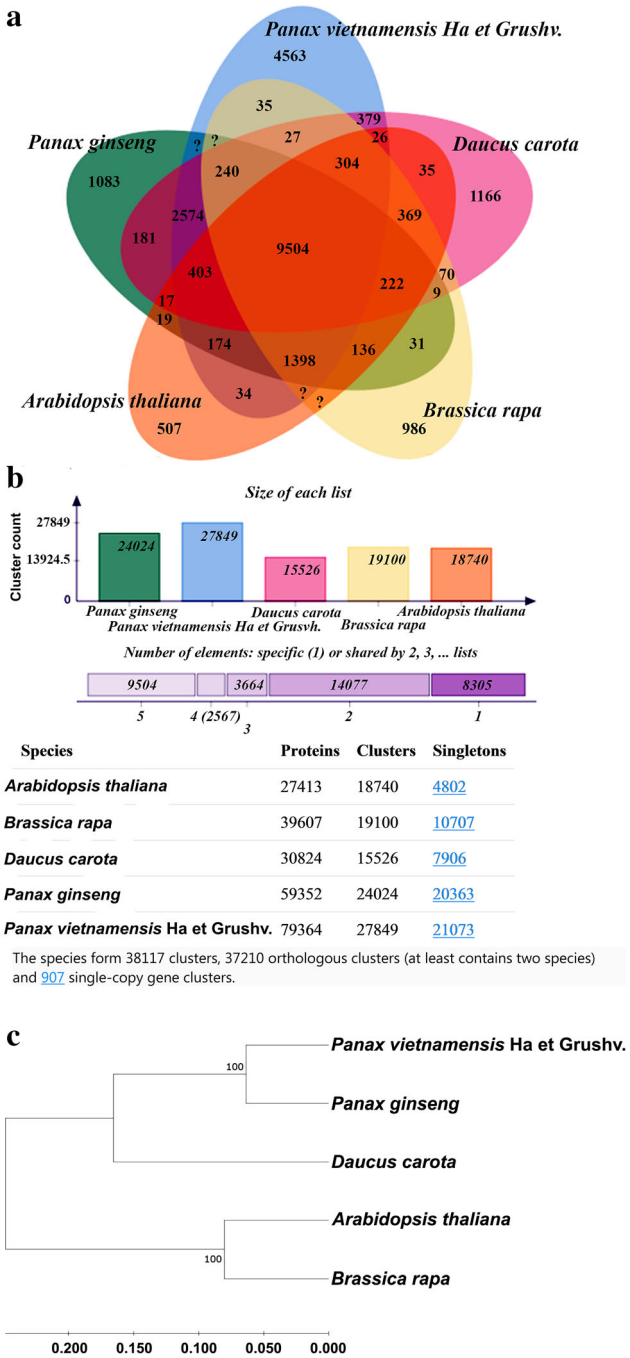
Fig. 3 Function gene ontology term of Vietnamese ginseng (Y axis in log(10) scale)

Gene functional classification

The functional classification was also performed by COG and KEGG and the results are shown in Figs. 5 and 6. All predicted gene models were searched in the COG and KEGG databases for functional prediction and classification. In total, 28,716 gene models were annotated and COG classified into 15 groups (Fig. 5). The cluster for carbohydrate transport and metabolism (G) was the largest group (3,267 genes) in COG categories of metabolism, followed by secondary metabolites biosynthesis, transport and catabolism (Q, 2,174 genes), amino acid transport and metabolism (E, 1,903 genes), energy production and conversion (C, 1,775 genes), lipid transport and metabolism (I, 1,669 genes), inorganic ion transport and metabolism (P, 1,488 genes), coenzyme transport and metabolism (H, 904 genes) and 510 genes of nucleotide transport and metabolism (F). In the cellular processes and signaling categories, the largest group (5,139 genes) is signal transduction mechanisms (T), followed by groups with a decreasing number of genes including posttranslational modification, protein turnover, chaperones (O, 5098 genes), intracellular trafficking, secretion, and vesicular transport (U, 2167 genes), cytoskeleton (Z, 851 genes), cell wall/membrane/envelope biogenesis (M, 650 genes), cell cycle control, cell division, chromosome partitioning (D, 628 genes), and defense mechanisms (V, 493 genes). A total of 29,212 genes (36.8%) were annotated and classified by KEGG databases (Fig. 6). The result showed that Protein families: genetic information process linked to the largest number of genes (5,992), followed by genetic information processing (5,151 genes), environmental information processing (1,908 genes), and 1,869 genes belong to protein families: signaling and cellular process. KEGG metabolic pathways presented in Vietnamese ginseng genome dataset include carbohydrate metabolism (2,550 genes), protein families: metabolism (1,807 genes), amino acid metabolism (898 genes), metabolism of cofactors and vitamins (576 genes), lipid metabolism (1162 genes), Unclassified: metabolism (1,070 genes), energy metabolism (905 genes), Metabolism of terpenoids and polyketides (519 genes), glycan biosynthesis and metabolism (418 genes), nucleotide metabolism (369 genes) and 120 genes of metabolism of other amino acids (Fig. 6). These annotations may provide valuable information for further research on specific metabolic pathways and functions of genes in Vietnamese ginseng.

Identification of genes involved in triterpene saponin biosynthesis and their metabolic pathway prediction

The terpenoid backbone biosynthesis pathway of Vietnamese ginseng was visualized by the KAAS system showed that ginsenoside is biosynthesized from two subpathways, namely mevalonate subpathway (MVA) and 2-c-methyl-D-erythritol-4-phosphate (MEP) (Fig. 7). In the MVA subpathway, we annotated four acetyl-CoA acetyltransferases (AACT, EC:2.3.1.9), six Hydroxymethylglutaryl-CoA synthase (HMGS, EC:2.3.3.10), eight



Species	Proteins	Clusters	Singletons
Arabidopsis thaliana	27413	18740	4802
Brassica rapa	39607	19100	10707
Daucus carota	30824	15526	7906
Panax ginseng	59352	24024	20363
Panax vietnamensis Ha et Grushv.	79364	27849	21073

The species form 38117 clusters, 37210 orthologous clusters (at least contains two species) and 907 single-copy gene clusters.

Fig. 4 Distribution of shared orthologous clusters among the five species (*a*). The numbers of shared and proteins, singletons in each genome are shown (**b**). Phylogenetic analysis of 907 single-copy gene clusters from Vietnamese ginseng and other species (**c**)

Hydroxymethylglutaryl-CoA reductase, two Mevalonate kinase (MVK, EC: EC:2.7.1.36), four phosphomevalonate kinase (PMK, EC:2.7.4.2), and two mevalonate-5-diphosphate decarboxylases (MVD, EC:4.1.1.33). Regarding subpathway MEP pathway of Vietnamese ginseng genome, we also annotated six 1-deoxy-d-xylulose 5-phosphate synthases (DXS, EC: 2.2.1.7), two 1-deoxy-d-xylulose 5-phosphate reductases (DXR, EC: 1.1.1.267), one 2-C-methyld-erythritol 4-phosphate cytidylyltransferase (MCT, EC:2.7.7.60), four 2-C-methyld-erythritol 2,4-cyclodiphosphate synthases (MECDPS, EC:4.6.1.12), two 4-hydroxy-3-methylbut-2-enyl diphosphate synthase (ISPG, EC: 1.17.7.1), and six 4-hydroxy-3-methylbut-2-enyl diphosphate reductases (ISPH, EC:1.17.1.4). Furthermore, the potential key genes involved in the terpenoids biosynthesis pathway in Vietnamese ginseng in terpenoid backbone biosynthesis were also predicted, including the following genes: FPPS, IPPI, SS, SQE, β-AS, DDS, PPDS, PPTS, CAS, β-A28O (Table 3, Supplementary data 1). Cytochrome P450 monooxygenase (P450s) and UDP-glycosyltransferases (UGTs) are two superfamilies of important enzymes for the structural diversity of triterpenoids by their hydroxylation. There are 101 P450s and 9 UGTs were annotated from all gene models by using the Pfam and KEGG databases (Supplementary data 2 and 5). CYP450 enzyme super-family was divided into two types including of A-Type and non-A type. In the A-type of CYP450 of Vietnamese ginseng, we found CYP701, CYP703 with single genes, and CYP78 has the largest gene (8 genes) numbers than another family of the CYP71 clan. In the non-A type of P450 super-family, we found only CYP711 has a single gene, and the remaining families have more than 2 copies, in which the CYP74 family has the highest number of predicted genes. The result showed that various P450 super-family were predicted with single and multiple copies. It suggested that duplication genes and diversity functions of CYP450 in Vietnamese ginseng. Furthermore, phylogenetic analyses revealed PvH_scaffold4060.5 and PvH_scaffold0121.42 to be grouped in the CYP73A subfamily and to be most closely related to *P. ginseng* PgH2DH22.1 and PgAEY75219.1, which are involved in ubiquinone and another terpenoid-quinone biosynthesis. The putative PvH_scaffold2989.13 and PvH_scaffold0073.47 were a group to CYP82 and they have the closest relationship to *Kalopanax septemlobus* (KsALO23115.1) and *P. ginseng* (PgH2DH23.1). The remaining putative CYPs of Vietnamese ginseng are quite different from other species in *Araliaceae* family. It may be suggested that CYP450s of Vietnamese ginseng have generated expanded families due to gene duplication. The genetic connections of candidate UGTs from genus *Panax* were depicted by NJ phylogenetic tree method with a bootstrap value of 1000 (Fig. 8). The result showed that the length of the amino acid sequence of these putative UGTs ranged from 444 to 585 aa. The phylogenetic analysis indicated that gene models PvH5977.g39882 and PvH4174.g1812 had a close relationship (97.16% of similarity) to ALE15279.1 from *Panax quinquefolius*, which function is transferred sugar from UDP-sugar to the C3 position of PPD UDP-glycosyltransferase 3GT1. Gene

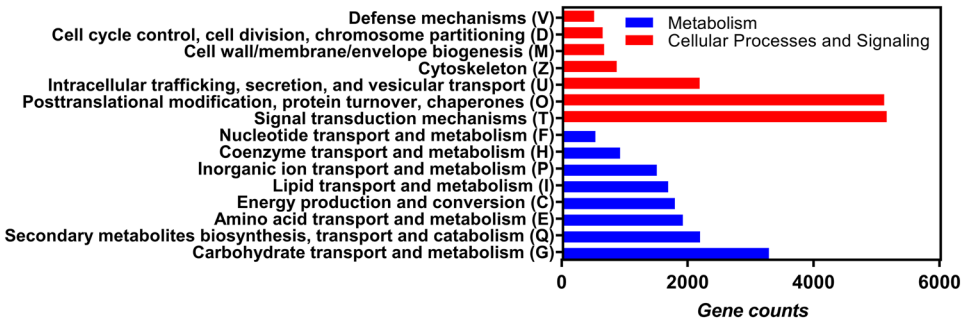


Fig. 5 COG Classification of Vietnamese ginseng genome

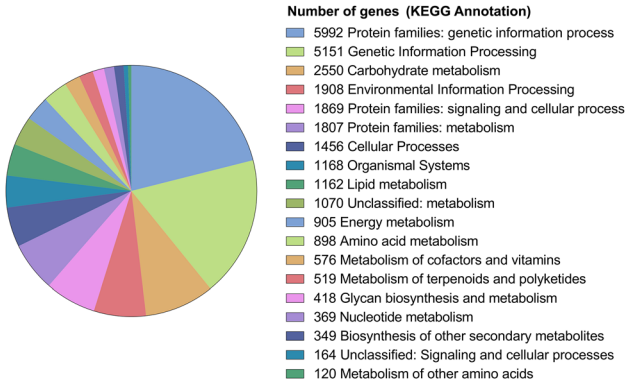


Fig. 6 KEGG Classification of Vietnamese ginseng genome. KEGG annotation shown in different colors corresponding to the number of genes in various of functional groups

model PvH6708.g20216 is highly homologous to QOJ43866.1 of *Panax notoginseng*, which belongs to the UGT29 family. The gene model PvH2902.g36227, PvH0288.g4781, PvH4174.g1821 and PvH5977.g39882 have the closest similarity with AKA44590.1 (92.78%), QEA68974.1 (96.50%), AKA44587.1 (98.27%) and A0A0K0PVL0.1 (97.47%) respectively, which belongs to UGT33, UGT76, UGT35 and UGTPg102 family of *Panax ginseng*. The result showed that Vietnamese ginseng contains different UGTs for ginsenosides diversity and they have similar function with their genes in *P. ginseng* and *P. quinquefolius and P. notoginseng.*

Combined pathway analysis by Reactome and KEGG with the blast expectation value of 1.0E-10, the result showed that a total of 10,957 sequences linked to 148 pathways (KEGG databases) and there are 5,313 genes were linked to 3,502 pathways for Reactome database analysis. (Supplementary data 4). The triterpenoid backbone biosynthesis pathway of Vietnamese ginseng based on annotated metabolic genes was visualized by the KEGG pathway databases system (Moriya et al. 2007). The metabolism pathways of terpenoids and polyketides in Vietnamese ginseng were analyzed by combination analysis of KEGG and Reactome databases. Among these pathways, the terpenoid backbone biosynthesis pathway showed the largest number of related genes (269), followed by carotenoid biosynthesis (138 genes), zeatin biosynthesis (151 genes), limonene and pinene degradation (129 genes), diterpenoid biosynthesis (101 genes), geraniol degradation (99 genes), polyketide sugar unit biosynthesis (45 genes), biosynthesis of ansamycins (23 genes), biosynthesis of siderophore group nonribosomal peptides (6 genes) and brassinosteroid biosynthesis (3 genes) (Fig. 9).

Structural characterization of *PvH_SE* genes

Ocotillol-type saponin majonoside-R2 (MR2) has been proved to be the main ginsenoside of *P. vietnamensis* (Zhang et al. 2015). The present study carried out validation of putative potential genes involved in ocotillol-type saponin biosynthesis. Three *PvH_SE* genes were isolated from genomic DNA and cDNA of roots by PCR and RT-PCR amplification. Structural analysis showed that the *PvH_SE2* gene has the longest coding sequence (1,635 bp), which contains 8 exons. While two genes *PvH_SE1* and *PvH_SE3* had shorter coding sequences, 1,581 and 1,611 bp respectively, of which the former has also 8 exons and the latter has 7 exons (Supplementary Figure S6, Supplementary data 5). The *PvH_SE* genes were deposited in GenBank with temporary accession numbers MW258698, MW258697 and MW258696. The peptide signal analysis showed that PvH_SE1 has a position to cut a signal peptide between aa 22 and 23 (VYA-LF). The signal peptide of PvH_SE2 is located at the middle of aa 28–29 (LFT-LR). Meanwhile, protein PvH_SE3 showed the most likely cleavage site between aa 39 and 40 (LLL-LN) (data not shown). Phylogenetic tree analysis by NJ method with 1000 bootstraps showed that *PvH_SE* genes from Vietnamese ginseng have a closer relationship with *SE* genes from Lai Chau ginseng than with Korean ginseng (Supplementary Figure S7).

Semi-quantitative analysis of gene expression

The expression of six important genes encoding enzymes involved in the triterpenoid biosynthesis pathway in root, leaf and stem of Vietnamese ginseng was analyzed by RT-

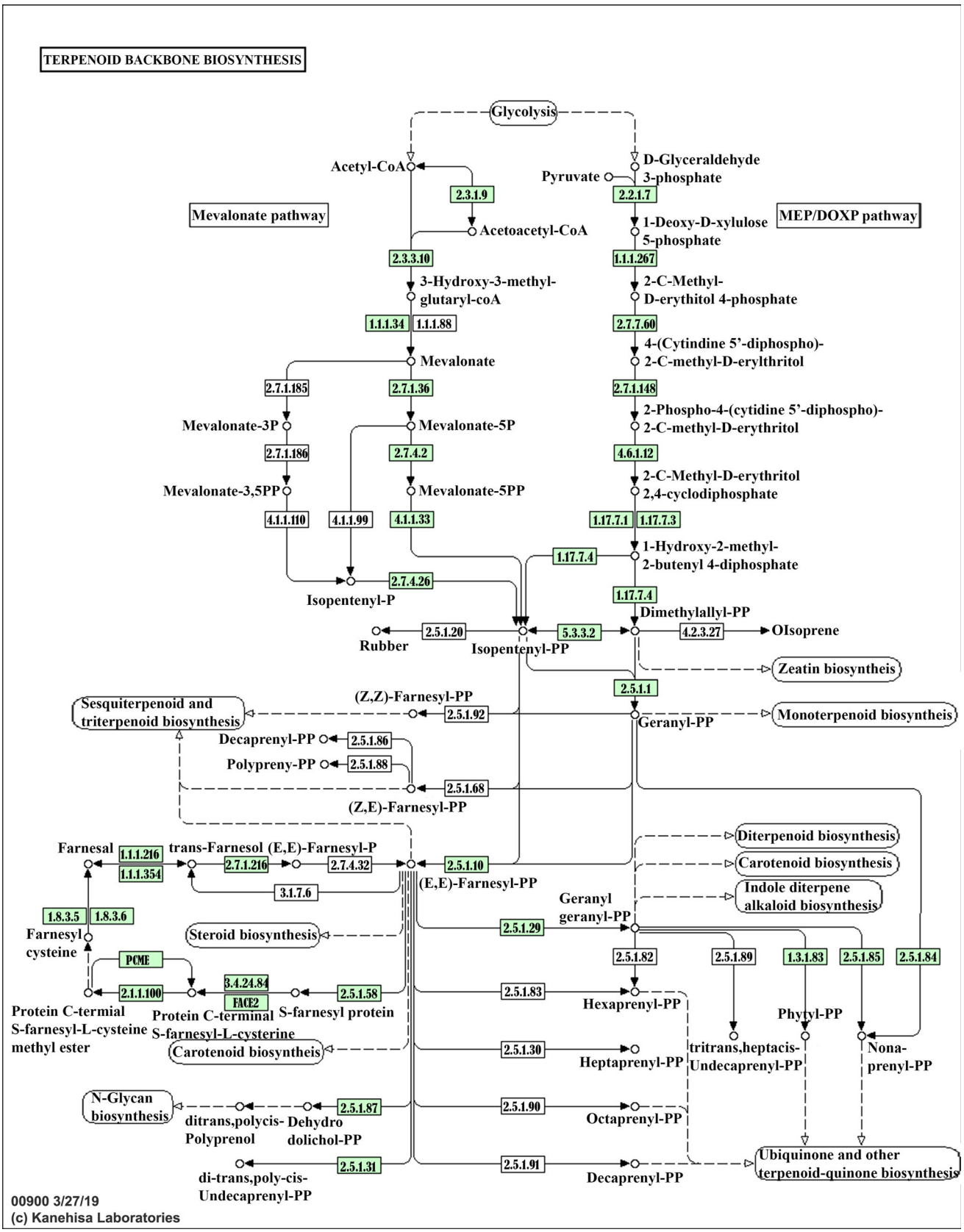


Fig. 7 Prediction of triterpenoid backbone biosynthesis pathway by KEGG analysis. The predicted genes involved in triterpenoid biosynthesis encode enzymes with EC numbers in green color

Table 3 Putative genes involved in triterpenoid saponin biosynthesis in Vietnamese ginseng

Gene name	EC number	Gene counts
IDI, Isopentenyl diphosphate isomerase	5.3.3.2	3
DXR, 1-deoxy-D-xylulose 5-phosphate reductoisomerase	2.2.1.7	2
MEP-CT, 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	2.7.1.148	1
MECDPS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4.6.1.12	4
DXS, 1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	6
ISPG, 4-hydroxy-3-methylbut-2-enyl diphosphate synthase	1.17.7.1	2
ISPH, 4-hydroxy-3-methylbut-2-enyl diphosphate reductases	1.17.1.4	6
AACT, Acetyl-CoA acetyltransferase	2.3.1.9	4
HMGS, Hydroxymethylglutaryl-CoA synthase	2.3.3.10	6
HMGR, Hydroxymethylglutaryl-CoA reductase	1.1.1.34	8
MVK, Mevalonate kinase	2.7.1.36	2
PMK, Phosphomevalonate kinase	2.7.4.2	4
MVD, Mevalonate diphosphate decarboxylase	4.1.1.33	2
FPPS, Farnesyl diphosphate synthase	2.5.1.10	2
SS, Squalene synthase	2.5.1.21	5
SQE, Squalene epoxidase	1.14.99.7	11
β-AS, β-amyrin synthase	5.4.99.39	6
OAS, Oleanic acid synthase	1.14.14.126	4
DDS, Dammarendiol synthases	4.2.1.125	3
PPDS, Protopanaxadiol synthase	1.14.14.120	2
CAS, Cycloarstenol synthase	5.4.99.8	7
PPTS, Protopanaxatriol synthase	1.14.14.121	2
Candidate UGTs	2.4.1.17	9

PCR (Fig. 10). Observations showed that most of the genes expressed in the leaves were higher than those in roots and stems, especially the genes encoding HMGS, HMGR and SS. These genes play an important role in 2,3-oxidosqualene biosynthesis in the first stage of the ginsenoside metabolic pathway. Squalene epoxidase was considered as a key enzyme involved in the formation of 2,3-oxidosqualene, an important precursor for the formation of many other valuable ginsenosides (Han et al. 2010; He et al. 2008; Zhang et al. 2015). Thus, perhaps leaves are the place to carry out biosynthesis of ginsenosides precursors in Vietnamese ginseng. The result showed that three genes encoding squalene epoxidase have various expression profiles in root, stem and leaf. Among them, gene *PvH_SE3* was expressed higher than two genes *PvH_SE1* and *PvH_SE2*. While expression of *PvH_SE1* gene is quite high in leaf and stem, *PvH_SE2* gene has higher expressions level in leaf and root. The result showed quite similarly to the report of Zhang et al. (2015) on expression analysis of *Panax vietnamensis* var. *fuscidiscus* (Zhang et al. 2015).

Identification, classification and primer design of SSR markers

SSR markers in assembly full dataset or predicted gene models of Vietnamensis ginseng were analyzed by three software including MIcroSAtellite (MISA), Krait and SciRoKo tools. There are slight differences in the number of SSRs identified from assembly data by the above analytical software (Table 4). A total average of 693,812 SSRs were detected from 9,837 sequences (3,001,967,204 bp long), which contains 30.16% (mononucleotide repeat motifs), 51.93% (dinucleotide repeat motifs), 15.74% (trinucleotide repeat motifs), 1.04% (tetranucleotide repeat motifs), 0.59% (pentanucleotide repeat motifs), 0.54% (hexanucleotide repeat motifs). An average of 4,135 SSR-containing genes was detected from 79,837 predicted genes. Among genes containing SSR, tri-nucleotide repeating motif was identified as the most common type (82.69%), followed by the motifs of mono- (6.53%), di- (6.48%), tetra- (0.12%), penta- (0.15%), and hexa-nucleotide (4.04%) (Supplementary Table T2). The SSR-containing genes were annotated by the Krait program (Du et al. 2018), showed that over 78,7% are intergenic and 21,3% are from CDS (Supplementary Figure S8). The scaffolds containing identified di- and tri-nucleotide SSR motifs with repeat numbers ≥ 8 extracted to use as a

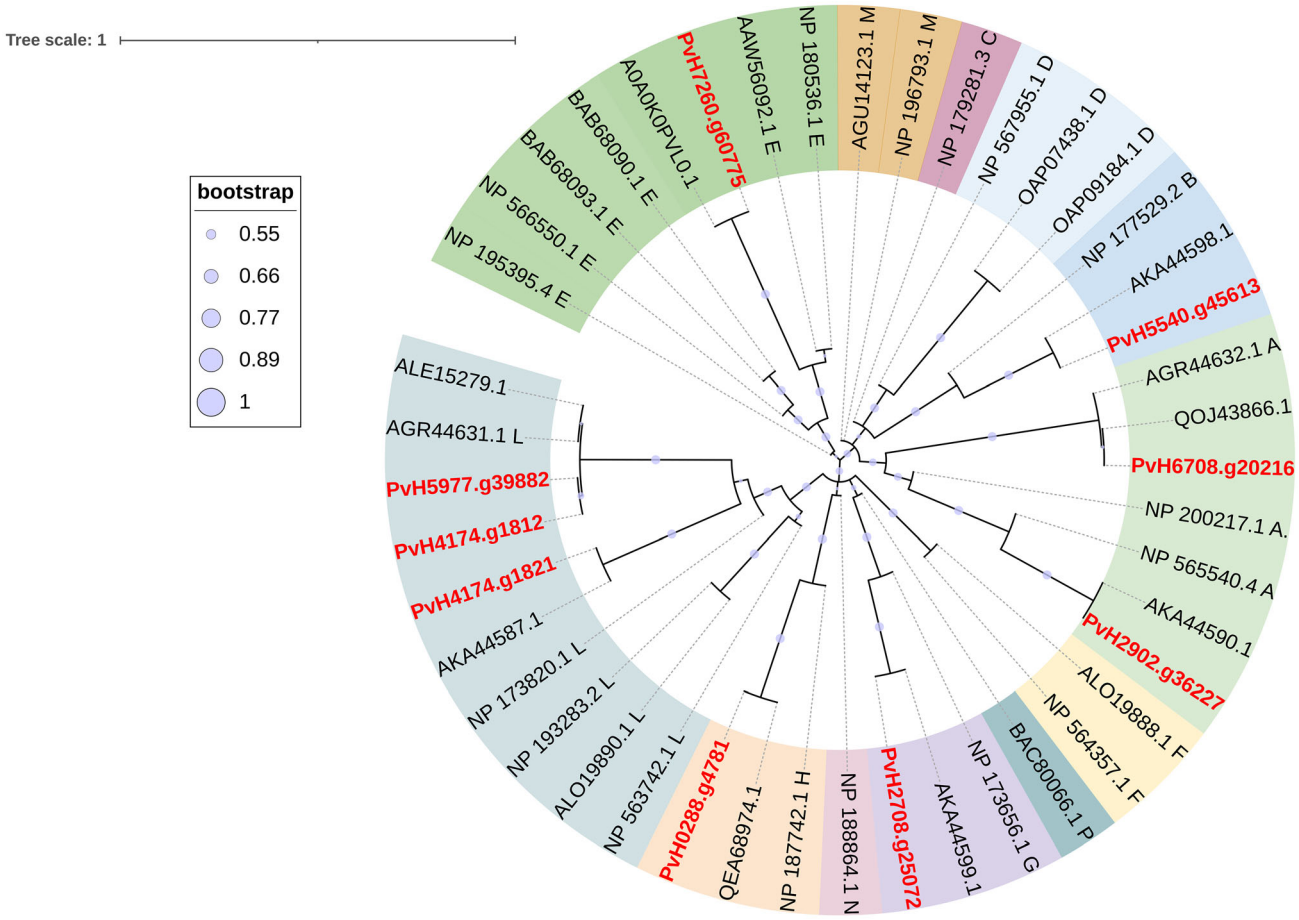


Fig. 8 Phylogenic analysis of UGTs candidates of Vietnamese ginseng and various species. UGTs of *Panax Vietnamensis* Ha et Grushv. were highlighted in red (PvH5977.g39882, PvH6708.g20216, PvH2708.g25072, PvH5540.g45613, PvH2902.g36227, PvH0288.g4781, PvH4174.g1821, PvH4174.g1812, PvH7260.g60775). Genbank accession numbers of UGTs from different species including *Panax quinquefolius* (ALE15279.1), *Panax ginseng* (AKA44599.1, AKA44598.1, AKA44590.1, QEA68974.1, AKA44587.1, A0A0K0PVL0.1, AGR44632.1, AGR44631.1), *Panax notoginseng* (QOJ43866.1), *Arabidopsis thaliana* (OAP09184.1, OAP07438.1, NP_173656.1, NP_173820.1, NP_177529.2, NP_179281.3, NP_180536.1, NP_187742.1, NP_188864.1, NP_193283.2, NP_195395.4, NP_196793.1, NP_200217.1, NP_563742.1, NP_564357.1, NP_565540.4, NP_566550.1, NP_567955.1), *Cicer arietinum* (AGU14123.1), *Camellia sinensis* (ALO19890.1, ALO19888.1), *Medicago truncatula* (AAW56092.1), and *Oryza sativa* subsp. *japonica* (BAB68090.1, BAB68093.1, BAC80066.1). The letters of A, B, C, D, E, F, G, H, L, M, N, and P are groups of UTGs

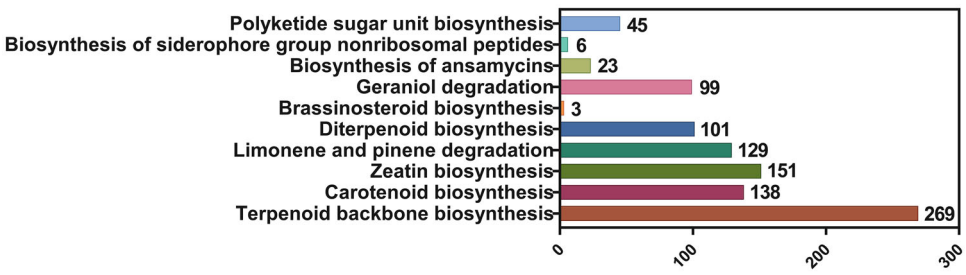


Fig. 9 Metabolism pathways of terpenoids and polyketides in Vietnamese ginseng genome assignment based on combined (KEGG and Reactome) Pathway analysis

template for the design of SSR primers using the Primer3 command-line tool (https://github.com/primer3-org/) (Supplementary data 3).

Discussion

The present study reported the first results on sequencing, assembly, and annotation of the whole genome of Vietnamese ginseng, followed by identification, expression

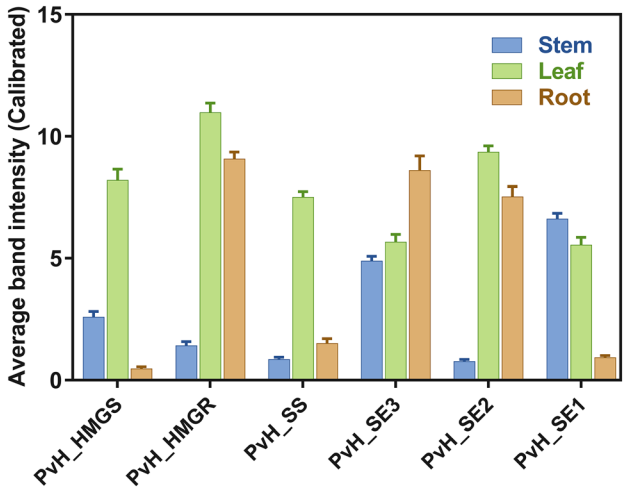


Fig. 10 Gene expression analyzed by ImageJ software based on the average intensity of the RT-PCR product bands. The expression of genes (PvH_HMGS, Hydroxymethylglutaryl-CoA synthase; PvH_HMGR, 3-hydroxy-3-methylglutaryl-CoA reductase; PvH_SS, Squalene synthase; PvH_SE3, Squalene epoxidase isoform 3; PvH_SE2, Squalene epoxidase isoform 2; PvH_SE1, Squalene epoxidase isoform 1) were analyzed based on the intensity of RT-PCR products in root (orange), leaf (green), and stem (blue) of Vietnamese ginseng

analysis of genes involved in triterpenoid biosynthesis and prediction of their metabolic pathways. The high quality of the assembled genome was generated by CLC Assembly Cell with a database of about 3.0 Gbps containing 79,316 predicted genes. Although Vietnamese ginseng genome has a fewer number of chromosomes than Korean ginseng (2n = 24 vs 2n = 48) (Kim et al. 2017), our study showed that the number of predicted gene models of Vietnamese ginseng is relatively high compared to Korean ginseng (Xu et al 2017) but slightly less compared to Lai Chau ginseng (Zhang et al. 2015). In a recent preliminary report, Vu et al. (2020) found 153,074 predicted transcripts (Vu et al. 2020) from the assembled transcriptome of Vietnamese ginseng with 89,271 unigenes. We predicted a lesser gene models number (79,374 gene models from 144,508 transcripts) than another report on the *Panax* genus. Potential proteins involved in MVA and MEP pathways for ginsenoside biosynthesis were also found with various isoforms in Vietnamese ginseng (Table 3). Identification of many isoforms of key genes involved in these secondary metabolic pathways may contribute to the regulation of triterpenoid biosynthesis in Vietnamese ginseng. To date, only one of three *SE* gene isoforms have been reported in the plant genome, including those of the ginseng genus (Han et al. 2010, He F et al. 2008). The phylogenetic tree analysis showed that the PvH_SE in Vietnamese ginseng share high similarity with their sub-species, Lai Chau ginseng, and they are not in the same group as other ginseng species. Although belonging to the same species of *Panax vietnamensis*, the amino acid sequence of PvH_SE isoenzymes of Vietnamese ginseng is significantly different from that of Lai Chau ginseng. The amino acid sequence alignment analysis result showed that these ginseng species share the identities of 99.2% for PvH_SE1, 99.3% for PvH_SE2, and 99.1% for PvH_SE3. Many reports showed that genes in a gene cluster are expressed differently in plant parts. Thus, the accumulation of secondary metabolites, the products of a cluster of genes, will also vary widely in these parts (Kim et al. 2011, Vu et al. 2020, Yamasaki et al. 2000). In this study, some of the key genes involved in the biosynthesis of ginsenoside in Vietnamese ginseng also have different levels of transcriptional expression in plant parts, which is similar to the findings in Korean ginseng and Lai Chau ginseng (Kim et al. 2011; Zhang et al. 2015). HMGR is believed to play an important role in the production of secondary metabolites in plants, specifically related to the

Table 4 Summary of novel SSR identification from Vietnamese ginseng genome assembly datasets

	Misa	Krait	SciRoKo	Average
Total number of sequences examined:	9,837	9,837	9,837	
Total size of examined sequences (bp):	3,001,967,204	3,001,967,204	3,001,967,204	
Total number of identified SSRs:	694,391	693,587	693,457	693,812
Type SSR				
Mononucleotide	209,428	209,175	209,107	209,237 (30,16%)
Dinucleotide	360,376	360,237	360,121	360,245 (51,93%)
Trinucleotide	109,369	109,150	109,149	109,223 (15,74%)
Tetranucleotide	7283	7188	7258	7,243 (1,04%)
Pentanucleotide	4149	4096	4090	4,112 (0,59%)
Hexanucleotide	3786	3741	3732	3,753 (0,54%)
Elapsed time	3 min 20 s	2 min 45 s	3 min 5 s	

accumulation of ginsenoside in the roots of Vietnamese ginseng (Liu et al. 2016; Luo et al. 2011). The expression analysis results showed that *PvH_HMGR* transcribed more in leaf tissue and roots than in plant stem. Among three *PvH_SE* genes in Vietnamese ginseng, *PvH_SE3* has the strongest level of transcriptional expression in all tested parts compared with two genes *PvH_SE1* and *PvH_SE2*. These results suggest that *PvH_SE3* may play a more important role in regulating terpenoid biosynthesis during the growth and development of Vietnamese ginseng. Up to now, the molecular understanding of ginsenoside accumulation and related genes in Vietnamese ginseng remains mysterious due to the lack of information about their genome. This study provided the genomic sequence of Vietnamese ginseng and constructed the ginsenosides biosynthesis pathway. The identification and expression analysis of key genes involved in the MVA pathway will provide valuable information, which will be an effective resource that will facilitate further studies of the metabolic biosynthesis pathway in Vietnamese ginseng. Furthermore, the SSR markers detected in the genome of Vietnamese ginseng will be useful for genetic diversity analysis, as well as actively assist in the selective breeding of this precious medicinal species.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12298-021-01076-1.

Acknowledgements This research was financially supported by the National Foundation for Science and Technology Development (NAFOSTED), Vietnam (Grant number 106.02-2018.49).

Author contributions Loc NH, Tien NQD and Nhut DT conceived, designed, and performed the overall study. Ma X and Rombauts S performed genome assembly. Ma X, Rombauts S and Tien NQD performed genome annotation, gene ontology analysis, and prediction of genes involved in triterpenoid biosynthesis in Vietnamese ginseng. Tien NQD performed gene clusters analysis, detection of SSR markers, and prediction of metabolic pathways. Tien NQD, Man LQ, Chi DTK conducted full-length cloning of genes encoding squalene epoxidase isoenzymes and the analysis of their characteristics. Tien NQD, Nhut DT, Man LQ and Huy NX analyzed gene expression. Ut T provides samples. Tien NQD and Loc NH wrote and approved the final manuscript.

Declarations

Conflict of interest All authors declared no conflict of interest.

Research involving human and animal rights This study did not involve human and animal subjects.

References

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33(16):2583–2585

Du L, Zhang C, Liu Q, Zhang X, Yue B (2018) Krait: An ultrafast tool for genome-wide survey of microsatellites and primer design. Bioinformatics 34(4):681–683

Duc NM, Kasai R, Yamasaki K, Nham NT, Tanaka O (1999) New dammarane saponins from Vietnamese ginseng. Studies in Plant Sci. 6:77–82

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

Han S, Jeong AJ, Yang H, Kang KB, Lee H, Yi EH, Ye SK (2016) Ginsenoside 20 (S)-Rh2 exerts anti-cancer activity through targeting IL-6-induced JAK2/STAT3 pathway in human colorectal cancer cells. J Ethnopharmacol 194:83–90

He F, Zhu Y, He M, Zhang Y (2008) Molecular cloning and characterization of the gene encoding squalene epoxidase in *Panax notoginseng*. DNA Seq 19(3):270–273

Huong NTT, Matsumoto K, Yamasaki K, Duc NM, Nham NT, Watanabe H (1995) Crude sponin extracted from Vietnamese ginseng and its major constituent majonoside-R2 attenuate the psychological stress-and foot-shock stress-induced antinociception in mice. Pharmacol Biochem Behav 52(2):427–432

Jayakodi M, Choi BS, Lee SC, Kim NH, Park JY, Jang W, Yang TJ (2018) Ginseng Genome Database: an open-access platform for genomics of *Panax ginseng*. BMC Plant Biol 18(1):62. https://doi.org/10.1186/s12870-018-1282-9

Kim TD, Han JY, Huh GH, Choi YE (2011) Expression and functional characterization of three squalene synthase genes associated with saponin biosynthesis in *Panax ginseng*. Plant Cell Physiol 52(1):125–137

Kim K, Dong J, Wang Y, Park JY, Lee SC, Yang TJ (2017) Evolution of the *Araliaceae* family inferred from complete chloroplast genomes and 45S nrDNAs of 10 *Panax*-related species. Sci Rep 7(1):1–9

Kofler R, Schlötterer C, Lelley T (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23(13):1683–1685

Konoshima T, Takasaki M, Ichiishi E, Murakami T, Tokuda H, Nishino H, Cancer Yamasaki K (1999) chemopreventive activity of majonoside-R2 from Vietnamese ginseng *Panax vietnamensis*. Cancer Lett. 147(1–2):11–16

Liu MH, Yang BR, Cheung WF, Yang KY, Zhou HF, Kwok JSL, Tsui SKW (2015) Transcriptome analysis of leaves roots and flowers of *Panax notoginseng* identifies genes involved in ginsenoside and alkaloid biosynthesis. BMC Genom. 16(1):265

Liu WJ, Lv HZ, He L, Song JY, Sun C, Luo HM, Chen SL (2016) Cloning and bioinformatic analysis of HMGS and HMGR genes from *Panax notoginseng*. Chinese Herbal Medic. 8(4):344–351

Liu J, Xu Y, Yang J, Wang W, Zhang J, Zhang R, Meng Q (2017) Discovery semisynthesis biological activities and metabolism of ocotillol-type saponins. J Ginseng Res 41:373–378

Luo H, Sun C, Sun Y, Wu Q, Li Y, Song J, Chen S (2011) Analysis of the transcriptome of *Panax notoginseng* root uncovers putative triterpene saponin-biosynthetic genes and genetic markers. BMC Genom 12(S5):S5

Mai TT, Moon J, Song Y, Viet PQ, Van Phuc P, Lee JM, Cho SK (2012) Ginsenoside F2 induces apoptosis accompanied by protective autophagy in breast cancer stem cells. Cancer Lett 321(2):144–153

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS. 2007. An automatic genome annotation and pathway reconstruction server. Nucleic Acids Research. 101093/nar/gkm321

Nguyen B, Kim K, Kim YC, Lee SC, Shin JE, Lee J, Yang TJ. 2017. The complete chloroplast genome sequence of *Panax vietnamensis* Ha *et* Grushv. (*Araliaceae*). Mitochondrial DNA Part A. 28 (1):85–86

Niu Y, Luo H, Sun C, Yang TJ, Dong L, Huang L, Chen S (2014) Expression profiling of the triterpene saponin biosynthesis genes FPS, SS, SE, and DS in the medicinal plant *Panax notoginseng*. Gene 533(1):295–303

Shen HB, Chou KC (2007) Signal-3L: A 3-layer approach for predicting signal peptides. Biochem Biophys Res Commun 363(2):297–303

Tang H, Krishnakumar V, Li J. Jcvi. Jcvi Utility Libraries

Vu DD, Shah SNM, Pham MP, Nguyen MT, Nguyen TPT. 2020. *De novo* assembly and transcriptome characterization of an endemic species of Vietnam *Panax vietnamensis* Ha et Grushv. including the development of EST-SSR markers for population genetics. BMC Plant Biol. 20 358. https://doi.org/10.1186/s12870-020-02571-5

Wang W, Rayburn ER, Hao M, Zhao Y, Hill DL, Zhang R, Wang H (2008) Experimental therapy of prostate cancer with novel natural product anti-cancer ginsenosides. Prostate 68(8):809–819

Xu J, Chu Y, Liao B, Xiao S, Yin Q, Bai R, Chen S (2017) *Panax ginseng* genome examination for ginsenoside biosynthesis. Gigascience 6(11):1–15

Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, Wang Y (2019) OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res 47(1):52–58

Yamasaki K (2000) Bioactive saponins in Vietnamese ginseng. Panax Vietnamensis Pharmaceut. Biol 38(sup1):16–24

Zhang GH, Ma CH, Zhang JJ, Chen JW, Tang QY, He MH, Yang SC (2015) Transcriptome analysis of *Panax vietnamensis* var *fuscidiscus* discovers putative ocotillol-type ginsenosides biosynthesis genes and genetic markers. BMC Genom. 16(1):159