Plant Cell Rep (2011) 30:2075–2083
DOI 10.1007/s00299-011-1114-y

ORIGINAL PAPER

Transcriptome profiling and *insilico* analysis of *Gynostemma pentaphyllum* using a next generation sequencer

Sathiyamoorthy Subramaniyam · Ramya Mathiyalagan · In Jun Gyo · Lee Bum-Soo · Lee Sungyoung · Yang Deok Chun

Received: 29 April 2011 / Revised: 20 June 2011 / Accepted: 21 June 2011 / Published online: 19 July 2011

Abstract Gynosaponins (Gypenosides) are major phytochemicals in *Gynostemma pentaphyllum* (Thunb.), with similarities to the ginsenosides present in *Panax ginseng*. Gynosaponins are classified as terpenoid compounds. In *G. pentaphyllum*, 25% of the total gynosaponins are similar to ginsenosides. In this study, we analyzed the transcriptional levels of the *G. pentaphyllum* genome to identify secondary metabolite genes. The complete transcriptomes for the roots and leaves were obtained using a GS-FLX pyro-sequencer. In total, we obtained 265,340 and all reads were well annotated according to biological databases. Using *insilico* analysis, 84% of sequence were well annotated and we obtained most of the secondary metabolite genes that represent mono-, di-, tri- and sesquiterpenoids. From our EST, most of the terpenoid genes were noted, among those few similar genes were studied in *P. ginseng* and these transcripts will help to characterize more triterpenoid genes in *G. pentaphyllum*. Also help to compare *P. ginseng* and *G. pentaphyllum* at transcriptome level.

Communicated by J. R. Liu.

Electronic supplementary material The online version of this article (doi:10.1007/s00299-011-1114-y) contains supplementary material, which is available to authorized users.

S. Subramaniyam · R. Mathiyalagan · I. Jun Gyo · L. Bum-Soo · Y. Deok Chun (✉)
Korean Ginseng Center and Ginseng Genetic Resource Bank, Kyung Hee University, Yongin 449-701, South Korea
e-mail: dcyang@khu.ac.kr

L. Sungyoung
Department of Computer Engineering, Kyung Hee University, Yongin 449-701, South Korea

Keywords *Gynostemma pentaphyllum* · Expressed sequence tags · Gene ontology · KEGG pathway

Introduction

Gynostemma pentaphyllum (Thunb.) Makino is a perennial creeping herb of the genus *Gynostemma* in the *Cucurbitaceae* family. Its common names include dungkulcha (Korea), Jiaogulan (China), Amachazura (Japan), five leaf ginseng, poor man's ginseng, and southern ginseng (Cui et al. 1999). *Gynostemma pentaphyllum* is naturally distributed in shaded and humid places like forests, mountain valleys, wood, scrub, and stream banks. It is a medicinal plant that reportedly has the adaptogenic nature to enhance the 'Yin' and 'Yang' properties of the human body (Razmovski-Naumovski et al. 2005). In traditional Chinese medicine, Aurvedhya and Oriental medicine in Asian countries, *G. pentaphyllum* is used for treatment of infection and inflammation, heat clearing, detoxification and relieving cough. The *G. pentaphyllum* species complex is spread throughout Asia, including India, Nepal, Bangladesh, Sri Lanka, Thailand, Myanmar, Korea, Japan, Vietnam and China. There are 21 subspecies of *G. pentaphyllum* throughout Asian countries, with the most in southwestern China (Ky et al. 2010). The plant is polyploid with variation in the number of chromosomes ($2n = 22$, 44, 66, 88) (Jiang et al. 2009). Wild type *G. pentaphyllum* was originally found in forest environments and later, due to consumption for medicinal products the increasing need for plant materials, cultivation of the plant was conducted in fields. *G. pentaphyllum* is harvested four times a year, and only the aerial parts of the plant, particularly the leaves, are used in products (Razmovski-Naumovski et al. 2005). To reduce the cost of cultivation and required

maintenance, breeders acquire advances of plant tissue culture to cultivate plant root material in the form of hairy roots and adventitious roots using suspension cultures (Chang et al. 2005). The whole plant materials are consumed in the form of herbal tea, tablet, capsule, or oral fluid or in a mixture with functional foods and other plant materials.

Worldwide, the consumption of medicinal plants and functional foods made of medicinal plant extracts is increasing. Epidemiological data demonstrate that a plant-based diet can reduce the risk of chronic disease, and numerous clinical studies have been conducted to look at the effects of *G. pentaphyllum* crude extracts and isolated compounds on the treatment of hepatitis, hypertension, chronic bronchitis, gastritis, cancer, Parkinson's disease, asthma, leukemia, apoptosis, and diabetes (Razmovski-Naumovski et al. 2005; Huyen et al. 2010; Wang et al. 2010; Choi et al. 2010; Hsu et al. 2010; Liou et al. 2010). In holistic medical practices, plant extracts and fresh or dry materials from medicinal plants are commonly used to absorb the useful phyto-chemicals to balance the fluxes in body's metabolism. Also, the World Health Organization (WHO) estimates that 80% of the world's population relies on traditional remedies for their primary health care (Kirby and Keasling 2009). Medicinal plants consist of number of secondary metabolites such as saponin, non-saponin, phytosterols, polysaccahrides, phenol, polyacetylenes, lignans, aminoacids, alkaloids and minerals (Xiang et al. 2010). In *G. pentaphyllum*, gypenosides (Gynosaponin) are a major medicinal component with 100 dammarane-type glycosides. Similarly, *Panax ginseng*, which is a well known medicinal plant in the family *Araliaceae*, contains ginsenosides (Razmovski-Naumovski et al. 2005; Hu et al. 1996; Yin et al. 2004a, b, 2006). Phytochemical investigations identified dammarane-type glycosides, similar to protopanaxadiol-type ginsenosides (i.e., ginsenosides Rb1, Rb3, Rc, Rd, F2, Rg3) in ***G.*** *pentaphyllum*. In addition, natural norisoprenoids, which are derived from carotenoids, were isolated from ***G.*** *pentaphyllum* (Zhang et al. 2010). Even though there are additional isoforms of gypenosides and ginsenosides, very few of the minor components such as ginsenosides Rg3 and F2, and gypenosides XVII and LXXV have greater efficacy in treating diseases. As a result, scientists are trying to use microbial assays to convert gypenosides to ginsenosides, and ginsenosides to gypenosides (An et al. 2010; Cheng et al. 2007).

Genes that are involved in the triterpenoid pathway have not been well-studied. In comparison to true ginseng products, *G. pentaphyllum* is an inexpensive substitute because of the wide availability of the plant resource. Even though both plants contain similar components, they are in different families and require different cultivation times. For *P. ginseng*, it takes 4–6 years to harvest ginsenosides (Choi 2008), a longer and more expensive cultivation period than that of *G. pentaphyllum*, which takes only 3–4 months to harvest gypenosides. Scientists also have difficulties producing new cultivars that produce high yields of ginsenosides, and the genes involved in the saponin cascades remain unknown. Until now, there is no reports on genetic or molecular studies of *G. pentaphyllum*. Therefore, we performed transcriptional profiling of *G. pentaphyllum* to identify the secondary metabolite genes.

In plant science, the typical approach is studying the functional genome through sequencing of the transcriptome, rather than analyzing the whole genome. Next generation sequencing (NGS) technologies enable scientists to analyze the complete transcriptome at minimal cost. Compared to the conventional Sanger's sequencing method (Brautigam and Gowik 2010), NGS produces data for the entire transcriptome. In non-model plants, transcriptome studies are helpful for gene discovery, transcript quantification, marker discovery, and small RNA discovery (Morozova et al. 2009; Brautigam and Gowik 2010). To the best of our knowledge, this is the first study to sequence the transcriptome of *G. pentaphyllum* root and leaf.

Materials and methods

Plant materials

In this study, we used 3-months-old hydroponically cultured *G. pentaphyllum* plants harvested from Gyeryong-si, Chungcheongnam-do, Korea. Hydroponic conditions included a photoperiod of 16 h light and 8 h dark at 24°C for 3 months. The transcriptome of the leaves and roots were separately sequenced.

RNA extraction and pyrosequencing

Total mRNA was extracted using the PolyATract® (Promega) kit and then used for cDNA synthesis using the Stratascript5.0 multi-temperature reverse transcriptase (Stratagene) kit according to the manufacturer's instructions. Total cDNA were subjected to sequencing on a GS FLX sequencer (454/Roche). One segment of a sequencing plate was used for a yield of ~100,000 reads. The reads were obtained as standard flowgram format (SFF) files, which is a suitable data format for Genome Sequence assembler (GSassembler).

Generation of the unigene dataset

The unigene dataset contains a set of non-redundant sequences composed of singlets and contigs. Sequence files were produced with the SFF file, and then SFF files were processed using the GSassembler software tool kit

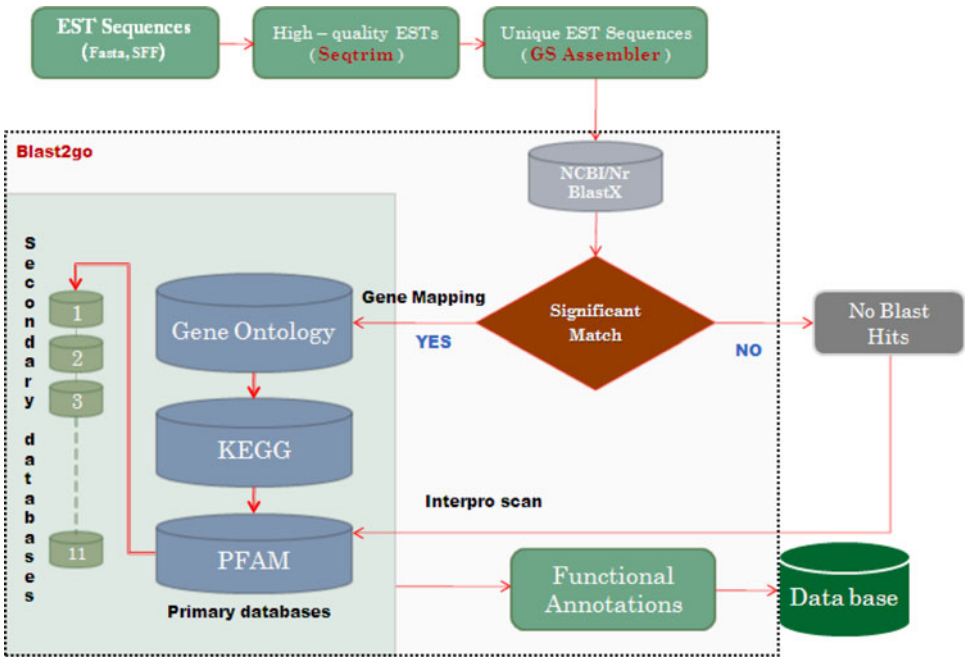

Fig. 1 Pipeline for analysis the *G. pantaphyllum* transcripts

provided by Roche. A de novo assembler project was created for the short cDNA sequence reads with default parameters. Using sffinfo tools, we obtained singleton and contig sequences. Low quality sequences were removed using seqtrim (Falgueras et al. 2010). The remaining sequences were used in the functional analysis (Fig. 1).

Gene Ontology, KEGG pathway and protein family analysis

Blast2go (B2G) is a high-throughput sequence analysis tool. Using B2G, unigenes were subjected to a BLASTX query against the national center for biotechnology information (NCBI) public non-redundant (NR) database. Based on the BLASTX results, the sequences were putatively named using the BLAST description annotator (BDA) tool embedded in B2G. More collective logic models were embedded in B2G to retrieve GOs, EC numbers, and KEGG maps. GO terms were obtained from sequence similarity and BLAST scores ($E <= 10^{-3}$) with default parameters. Those annotations were simplified into plant functional categories using the plant GOslim. Interproscan embedded with B2G was used to obtain the protein domain information for the putative sequences (Conesa and Gotz 2008) (Fig. 1).

Results and discussion

Sequencing and assembly of ESTs

G. pentaphyllum root and leaf samples were used for our study. mRNA was isolated from each sample and used for cDNA library construction. The total cDNA library was sequenced using a GS FLX sequencer (454/Roche), resulting in a total of 265,340 ESTs. All the sequences were subjected to seqtrim with the default parameter for low quality removal (Falgueras et al. 2010), and this resulted in a total of 245,376 ESTs. These ESTs had an average length of 340 bp, and represented 39.7 Mb of each library. A total of 79.5 Mb sequences were analyzed (Table 1). The 245,376 ESTs were assembled with GSassembler to produce 21,973 contigs and 26,097 singletons, for 48,070 unique sequences. The maximum numbers of ESTs per contig were 802 for leaf and 752 for root. Assembled sequences functionally annotated with blast2go. Finally, all sequences were organized into a pentaphyllum EST database (http://www.bioherbs.khu.ac.kr/pentaphyllum/).

Table 1 Sequence assemble and annotation reports

Description	No. of ESTs
Number of reads	245,376
Number of contigs	21,973
Number of reads in contigs	219,279
Number of singletons	26,097
Sequence without blast hits	11,776
Number of sequence with >70% blast similarity score	26,671
Average blast hits per sequence	18
Number of EST without GO	3,706
Number of sequence with EC	10,655
Number of enzyme codes	1,352
Sequence only with protein ids	3,787

ESTs against public non-redundant databases

Homology-based functional assignment for putative sequences was accomplished through BLASTX queries against non-redundant databases. The parameters used in the query were an *E* value of 10^{-3} or below, an HSP cut-off of 33 and a maximum of 20 blast hits per sequence (Sathiyamoorthy et al. 2010b). The BLASTX identified an average of 18 matching sequences for 36,294 (75.5%) ESTs with 11,776 (24.4%) remaining sequences having no meaningful matches (Table 1). Among the matching sequences found through the BLASTX, 86.47% were sequences from the following plants: *Vitis vinifera* (17.52%), *Arabidopsis thaliana* (13.31%), *Oryza sativa* (11.86%), *Populus trichocarpa* (11.39%), *Ricinus communis* (6.91%), *Arabidopsis lyrata* (6.53%), *Zea mays* (5.21%), *Sorghum bicolor* (3.93%), *Glycine max* (3.14%), *Physcomitrella patens* (2.39%), *Picea sitchensis* (1.58), *Medicago truncatula* (1.24%), *Nicotiana tabacum* (0.57), *Solanum lycopersicum* (0.46%), *Solanum tuberosum* (0.43%) and other plants (13.3%) . The above-mentioned plants have been relatively well-studied so there are more experimental data for them than for *G. pentaphyllum*. Annotations were obtained from the BLAST description annotator (BDA) tool from BLAST result against to NR database. These annotations of putative sequences were used to assist for further experimental analysis.

Functional analysis and phenylpropanoid pathway genes

Gene functional annotation is a more difficult task for newly sequenced non-model plants than for humans, because the plant genome has numerous genes reflecting adaptations to environmental factors. To simplify the annotation process, Gene Ontology (GO) has evolved for use in the field of functional genomics. GO has been invaluable in the annotation of putative transcripts and has become a de facto standard for annotating putative genes. GO is ideal for grouping genes into clusters based on control vocabularies and for elucidating hierarchical relationships between gene groups. Control vocabularies are grouped into three major categories, namely molecular function, biological processes, and cellular components (Sathiyamoorthy et al. 2010a, b, c; The Gene Ontology 2010). In our results, 32,588 (67.7%) of the EST sequences mapped into one, a combination of two, and three categories. We organized these groupings in a Venn diagram individually for roots (Fig. 2a) and leaves (Fig. 2b). The numbers of EST sequences in the categories of cellular compound (CC), molecular function (MF) and biological process (BP) were 23,337 (71.6%), 27,307 (83.7%) and 22,876 (70.1%), respectively. In total, 18,884 ESTs (57.9%) mapped to a combination of the two categories of CC and MF, 17,533 ESTs (53.8%) mapped to CC and BP, and 20,996 ESTs (64.4%) mapped to BP and MF. There were 16,296 ESTs (50%) mapped to all three categories. We used plant-GOslim to screen for plant-specific GO vocabularies. In our results, a large number of unique sequences were grouped under the first category of molecular function with nucleotide binding, protein binding, kinase activity, transporter activity, etc. The second category included biological processes with subcategories like transport, response to stress, protein modification process, response to abiotic stimulus, catabolic process, and cellular component organization. The third category included cellular compounds with subcategories of plastid, plasma membrane, mitochondrion, cytosol, and vacuole (Supplementary Table 1). From the GO results, ESTs which are responses to biotic and abiotic stresses, were putatively annotated to phenylpropanoid biosynthesis pathway, which, including 4-coumarate: ligase/Acyl: CoA ligase (17 ESTs), caffeic acid *O*-methyltransferase/catechol *O*-methyltransferase (10 ESTs), trans-caffeoyl-CoA

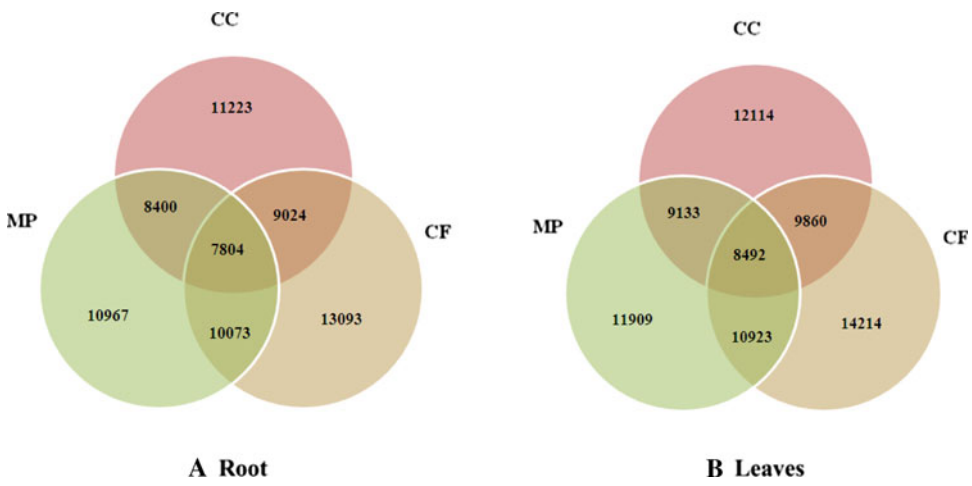

Fig. 2 Venn diagram of *G. pentaphyllum* data set showing numbers annotated to one, a combination of two and/or all three GO vocabularies (*MF* molecular function, *BP* biological process, *CC* cellular compound)

3-*O*-methyltransferase-like protein/caffeoyl-3-*O*-methyltransferase (12 ESTs), UTP-glucose glucosyltransferase (3 ESTs), aldehyde dehydrogenase (2 ESTs), anthranilate *N*-benzoyltransferase, SNG1 (sinapoylglucose 1) serine-type carboxypeptidase sinapoylglucose-malate/*O*-sinapoyltransferase (2 ESTs), cinnamoyl reductase (20 ESTs), and cinnamyl alcohol dehydrogenase/mannitol dehydrogenase (20 ESTs). Generally, the phenylpropanoid pathway genes are involved in responses to biotic and abiotic stresses. The phenylpropanoid pathway produces useful polymers like lignin, tannins, anthocyanins, flavonoids, phenylpropanoid esters, cutin and coumarins (Ferrer et al. 2008; Vogt 2010). The role of the phenylpropanoid pathway genes in responses to biotic and abiotic stress is well studied in *Arabidopsis*. Also, a few of these genes (cinnamyl alcohol dehydrogenase (Pulla et al. 2009), short-chain alcohol dehydrogenase (Kim et al. 2009), and spermidine synthase (Parvin et al. 2010)) have been studied for their involvement in biotic stress in *P. ginseng,* which has similar saponin compounds to those in *G. pentaphyllum*. In our results, there were 15,482 (32.2%) unknown ESTs that represent *G. pentaphyllum*-specific genes that might be involved in the gypenoside pathway. The partially annotated 32,588 (67.7%) ESTs that are putative/hypothetical proteins need to be functionally characterized to see if they are involved in the gypenoside pathway. Annotation data will guide further gene selection and functional experiments.

Protein functional domain analysis

In our results, 11,776 (24.4%) ESTs did not have a significant match in the non-redundant database and 3,706 (7.7%) ESTs did not map to a GO. Therefore, we attempted functional annotations of these sequences with interproscan, which is designed for protein domain homology searches (Quevillon et al. 2005). In our interproscan results, there were 3,787 (7.8%) sequences without blast hits and GO were assigned with interpro ids. Our analysis also included secondary protein database ids like PFAM (16,429), SMART (2,916), GENE3d (9,907), PROSITE (3,530), PROFILE (5,611), PRODOM (295), SUPERFAMILY (12,067), PANTHER (25,575), PIR (186), PRINTS (2,424), TIGRFAMs (792), and SINGNALP (10,019) to EST sequences. Most of our ESTs were putatively identified as proteins and also hypothetical proteins with unknown function. These ESTs also annotated by interpro ids and a secondary protein database id. Most of ESTs with functional domain for pathogenic proteins were not mapped with GO, but were annotated with interproscan ids (Hunter et al. 2009). Pathogen proteins were classified into five groups based on protein domains; in our study all classes of protein domains were found. Those domains included serine/threonine kinases, leucine-rich repeats, leucine zippers, coiled-coils, the toll and interleukin1 receptors, and the nucleotide binding site of the WRKY family (McHale et al. 2006; Martin et al. 2003). Genes found among our ESTs included pore-forming toxin-like protein hfr-2 (IPR005830, IPR008998), ribosome inactivating protein precursor (IPR001574, IPR016138), bet-vi allergen family protein (IPR000916), gynostemmin-like protein (IPR001574, IPR016138, IPR016139, IPR017988, IPR017989), leucine-rich repeat-containing (IPR001611), major pollen allergen carb1 isoforms 1a and 1b (IPR000916), trichosanthin precursor (IPR001574, IPR016138, IPR016139, IPR017989), brassinosteroid insensitive 1-associated receptor kinase1 (IPR000719, IPR001245, IPR011009, IPR017441), wrky transcription (IPR003657, IPR018872), pathogen induced protein2-4 (IPR006016), pathogenesis-related protein1 (IPR001283, IPR014044), pathogenesis-related protein4b (IPR000726, IPR001153, IPR009009, IPR014733, IPR018226). Among these, some genes have been studied in *P. ginseng*, including the polygalacturonase-inhibiting protein assigned with a leucine-rich repeat (IPR001611), leucine-rich repeat-containing N-terminal domain, type 2 (IPR013210) (Sathiyaraj et al. 2010) and other genes like PR-10 and PR-5 (Pulla et al. 2009; Yu et al. 2009; Pulla et al. 2010).

KEGG biochemical analysis and Terpenoid backbone pathway genes

Biochemical analysis for our ESTs was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG was created with bioinformatic algorithms and is the most accessible database of major biochemical pathways (Kanehisa et al. 2004). All putative transcripts were subjected to a KEGG database query with the BLAST score to retrieve KEGG enzyme codes and pathway maps. From our collection of ESTs, 10,655 sequences (22.48%) were assigned 1382 EC numbers. From that, 799 (57.8%) enzyme numbers were assigned to 149 unique KEGG pathways and 587 (42%) were not assigned to any pathways. The KEGG pathways were in sub-categories of carbohydrate metabolism, amino acid metabolism, energy metabolism, lipid metabolism, and secondary metabolism. This study focused on the gypenoside-related genes and genes in the tri-terpenoid pathway. Terpenoids are classified into groups based on the number of carbons present in the components. The major groups are monoterpenes (C10), sesquiterpenes (C15), diterpenes (C20), and triterpenes (C30). In this results, genes for all major groups of terpenoids were present in *G. pentaphyllum* ESTs. Basically, terpenoids are derived from two isoprenoid pathways in plants, which are called the terpenoid backbone pathways. One is the cytosol MVA pathway with the end product of IPP; another is the plastidial DXP pathway with

Table 2 Secondary metabolite genes of *G. pentaphyllum* based on KEGG biochemical analysis

Pathway	Enzyme	EC number	ESTs
Mevalonate pathway	Acetyl-CoA C-acetyltransferase	EC:2.3.1.9	13
	Hydroxymethylglutaryl-CoA reductase (NADPH)	EC:1.1.1.34	13
	Isopentenyl-diphosphate Delta-isomerase	EC:5.3.3.2	9
	Hydroxymethylglutaryl-CoA synthase	EC:2.3.3.10	4
	Phosphomevalonate kinase	EC:2.7.4.2	4
	Mevalonate kinase	EC:2.7.1.36	3
	Diphosphomevalonate decarboxylase	EC:4.1.1.33	3
	Hydroxymethylglutaryl-CoA reductase	EC:1.1.1.88	1
MEP/DXOP pathway	1-Deoxy-D-xylulose-5-phosphate synthase	EC:2.2.1.7	19
	Farnesyltranstransferase	EC:2.5.1.29	13
	Geranyltranstransferase	EC:2.5.1.10	10
	Isopentenyl-diphosphate Delta-isomerase	EC:5.3.3.2	9
	4-Hydroxy-3-methylbut-2-enyl diphosphate reductase	EC:1.17.1.2	7
	Dimethylallyltranstransferase	EC:2.5.1.1	6
	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase	EC:2.7.1.148	3
	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	EC:4.6.1.12	3
	Trans-hexaprenyltranstransferase	EC:2.5.1.30	3
	1-Deoxy-D-xylulose-5-phosphate reductoisomerase	EC:1.1.1.267	2
	Di-trans, poly-cis-decaprenylcistransferase	EC:2.5.1.31	2
	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	EC:2.7.7.60	1
	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	EC:1.17.1.1	1
Monoterpenoid biosynthesis	Secologanin synthase	EC:1.3.3.9	6
	Myrcene synthase	EC:4.2.3.15	2
	(R)-limonene synthase	EC:4.2.3.20	1
Diterpenoid biosynthesis	Gibberellin 2beta-dioxygenase	EC:1.14.11.15	10
	Gibberellin 3beta-dioxygenase	EC:1.14.11.15	10
	Taxane 13alpha-hydroxylase	EC:1.14.13.77	6
	Taxane 10beta-hydroxylase	EC:1.14.13.76	4
	10-Deacetylbaccatin III 10-*O*-acetyltransferase	EC:2.3.1.167	3
	Ent-kaurene synthase	EC:4.2.3.19	3
	Taxadien-5alpha-ol *O*-acetyltransferase	EC:2.3.1.162	2
	Gibberellin 2beta-dioxygenase	EC:1.14.11.13	2
	Taxadien-5alpha-ol *O*-acetyltransferase	EC:2.3.1.162	2
	Taxadiene 5alpha-hydroxylase	EC:1.14.99.37	1
Sesquiterpenoid biosynthesis	(+)-delta-cadinene synthase	EC:4.2.3.13	1
Carotenoid biosynthesis	Carotene 7,8-desaturase	EC:1.14.99.30	7
	Capsanthin/capsorubin synthase	EC:5.3.99.8	6
	(+)-abscisic acid 8'-hydroxylase	EC:1.14.13.93	4
	Zeaxanthin epoxidase	EC:1.14.13.90	4
	Neoxanthin synthase	EC:5.3.99.9	3
	Phytoene synthase	EC:2.5.1.32	3
	Xanthoxin dehydrogenase	EC:1.1.1.288	3
	9-*cis*-epoxycarotenoid dioxygenase	EC:1.13.11.51	2
	Violaxanthin de-epoxidase	EC:1.10.99.3	2

Table 2 continued

Pathway	Enzyme	EC number	ESTs
Steroid biosynthesis	Cycloartenol synthase/beta amryin synthase	EC:5.4.99.8	10
	Squalene monooxygenase	EC:1.14.99.7	7
	Sterol-4alpha-carboxylate 3-dehydrogenase (decarboxylating)	EC:1.1.1.170	6
	Sterol 14-demethylase	EC:1.14.13.70	5
	Sterol 24-C-methyltransferase	EC:2.1.1.41	5
	Squalene synthase	EC:2.5.1.21	4
	Cycloeucalenol cycloisomerase	EC:5.5.1.9	3
	Delta14-sterol reductase	EC:1.3.1.70	3
	Methylsterol monooxygenase	EC:1.14.13.72	3
	Cycloeucalenol cycloisomerase	EC:5.5.1.9	3
	24-Methylenesterol C-methyltransferase	EC:2.1.1.143	2
	7-Dehydrocholesterol reductase	EC:1.3.1.21	2
	Cholestenol Delta-isomerase	EC:5.3.3.5	1

the end product of IPP and DMAPP. Both pathways are expressed in different locations. The mevalonate pathway enzymes are located in the cytosol, and the DXP enzymes are found in the plastid (Kirby and Keasling 2009; Yan et al. 2005). After the IPP and DMAPP pathways, the triterpene aglycone of ginsenoside/gypnoside, protopanaxadiol, is synthesized from 2,3-oxidosqualene. All genes involved in putative ginsenoside pathways were also present in our EST library, so we hypothesized that these genes may be responsible for gypenoside production. Few genes with a role in ginsenoside regulation, such as squalene synthase and squalene epoxidase, have been well characterized in *P. ginseng* (Han et al. 2010; Ju-Sun et al. 2010). *G. pentaphyllum* is polyploid, and recently ginseng scientists have reported three squalene synthases (Kim et al. 2010). In our data, we detected squalene synthase1-2. More isoforms were present for fatty acid genes, P450 cytochromes and glycosyl transferases. Most of the secondary metabolite genes present in our library along with EST counts were shown in Table 2.

Conclusion

G. pentaphyllum is a well known traditional medicinal plant in Asian countries and has been commercialized worldwide. Recently, scientists are searching for the novel genes that are involved in the terpenoid pathway. Our EST library contains most mono-, di-, tri-, and sesquiterpenes and steroidal pathway genes. Given the difficulties in culturing *P. ginseng*, and the similarities between gypenosides and ginsenosides, the analysis of *G. pentaphyllum* as an alternative to *P. ginseng* may be valuable to both scientists and consumers. While the two plants are in different families, they have the same chemical components. *G. pentaphyllum* is a viable alternative model plant for *P. ginseng* and is a good resource for identifying new traits and novel genes in the gypenoside and ginsenoside pathways. From our study, 75.5% of sequences with BLAST hits, that to 50% of sequence includes all three categories of Gene Ontology (GO). And those 24.4% ESTs not find any BLAST hit but 8% of ESTs annotated protein ids. Totally, 84% sequences were well annotated with biological schemas. Finally relational database were developed for easy access of the data.

Acknowledgements This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency)" (NIPA-2011-(C1090-1121-0003)).

References

An DS, Cui CH, Lee HG, Wang L, Kim SC, Lee ST, Jin F, Yu H, Chin YW, Lee HK, Im WT, Kim SG (2010) Identification and characterization of a novel Terrabacter ginsenosidimutans sp. nov. beta-glucosidase that transforms ginsenoside Rb1 into the rare gypenosides XVII and LXXV. Appl Environ Microbiol 76(17):5827–5836. doi:10.1128/AEM.00106-10

Brautigam A, Gowik U (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol (Stuttg) 12(6):831–841. doi:10.1111/j.1438-8677.2010.00373.x

Chang CK, Chang KS, Lin YC, Liu SY, Chen CY (2005) Hairy root cultures of *Gynostemma pentaphyllum* (Thunb.) Makino: a promising approach for the production of gypenosides as an alternative of ginseng saponins. Biotechnol Lett 27(16):1165–1169. doi:10.1007/s10529-005-8653-7

Cheng LQ, Ju RN, Myung KK, Myun HB, Deok CY (2007) Microbial conversion of Ginsenoside Rb1 to minor Ginsenoside F2 and Gypenoside XVII by *Intrasporangium* sp. GS603 isolated from soil. J Microbiol Biotechnol 17(12):1937–1943

Choi HS, Park MS, Kim SH, Hwang BY, Lee CK, Lee MK (2010) Neuroprotective effects of herbal ethanol extracts from

Gynostemma pentaphyllum in the 6-hydroxydopamine-lesioned rat model of Parkinson's disease. Molecules 15(4):2814–2824. doi:10.3390/molecules15042814

Choi Kt (2008) Botanical characteristics, pharmacological effects and medicinal components of Korean *Panax ginseng* C A Meyer. Acta Pharmacol Sinica 29(9):1109–1118

Conesa A, Gotz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008:619832. doi:10.1155/2008/619832

Cui J, Eneroth P, Bruhn JG (1999) *Gynostemma pentaphyllum*: identification of major sapogenins and differentiation from Panax species. Eur J Pharm Sci 8(3):187–191. S0928098799000135 [pii]

Falgueras J, Lara AJ, Fernandez-Pozo N, Canton FR, Perez-Trabado G, Claros MG (2010) SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics 11:38. doi:10.1186/1471-2105-11-38

Ferrer JL, Austin MB, Stewart C Jr, Noel JP (2008) Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. Plant Physiol Biochem 46(3):356–370. doi:10.1016/j.plaphy.2007.12.009

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. doi:10.1016/j.phytochem.2009.09.031

Hsu HY, Yang JS, Lu KW, Yu CS, Chou ST, Lin JJ, Chen YY, Lin ML, Chueh FS, Chen SS, Chung JG (2010) An experimental study on the antileukemia effects of Gypenosides in vitro and in vivo. Integr Cancer Ther. doi:10.1177/1534735410377198

Hu L, Chen Z, Xie Y (1996) New triterpenoid saponins from *Gynostemma pentaphyllum*. J Nat Prod 59(12):1143–1145. doi:10.1021/np960445u

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37(Database issue):D211–D215. doi:10.1093/nar/gkn785

Huyen VT, Phan DV, Thang P, Hoa NK, Ostenson CG (2010) Antidiabetic effect of *Gynostemma pentaphyllum* tea in randomly assigned type 2 diabetic patients. Horm Metab Res 42(5):353–357. doi:10.1055/s-0030-1248298

Jiang LY, Qian ZQ, Guo ZG, Wang C, Zhao GF (2009) Polyploid origins in *Gynostemma pentaphyllum* (Cucurbitaceae) inferred from multiple gene sequences. Mol Phylogenet Evol 52(1):183–191. doi:10.1016/j.ympev.2009.03.004

Ju-Sun S, Lee OK, Kim YJ, Lee JH, Kim JH, Jung DY, In JG, Lee BS, Yang DC (2010) Overexpression of PgSQS1 increases Ginsenoside production and negatively affects ginseng growth rate in *Panax ginseng*. J Ginseng Res 34(2):86–91. doi:10.5142/jrg.2010.34.2.086

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32(Database issue):D277–D280. doi:10.1093/nar/gkh063

Kim TD, Han JY, Huh GH, Choi YE (2010) Expression and functional characterization of three squalene synthase genes associated with saponin biosynthesis in *Panax ginseng*. Plant Cell Physiol. doi:10.1093/pcp/pcq179

Kim YJ, Shim JS, Lee JH, Jung DY, Sun H, In JG, Yang DC (2009) Isolation and characterization of a novel short-chain alcohol dehydrogenase gene from *Panax ginseng*. BMB Rep 42(10):673–678

Kirby J, Keasling JD (2009) Biosynthesis of plant isoprenoids: perspectives for microbial engineering. Annu Rev Plant Biol 60:335–355. doi:10.1146/annurev.arplant.043008.091955

Ky PT, Huong PT, My TK, Anh PT, Kiem PV, Minh CV, Cuong NX, Thao NP, Nhiem NX, Hyun JH, Kang HK, Kim YH (2010) Dammarane-type saponins from *Gynostemma pentaphyllum*. Phytochemistry 71(8–9):994–1001. doi:10.1016/j.phytochem.2010.03.009

Liou CJ, Huang WC, Kuo ML, Yang RC, Shen JJ (2010) Long-term oral administration of *Gynostemma pentaphyllum* extract attenuates airway inflammation and Th2 cell activities in ovalbumin-sensitized mice. Food Chem Toxicol 48(10):2592–2598. doi:10.1016/j.fct.2010.06.020

Martin GB, Bogdanove AJ, Sessa G (2003) Understanding the functions of plant disease resistance proteins. Annu Rev Plant Biol 54:23–61. doi:10.1146/annurev.arplant.54.031902.135035

McHale L, Tan X, Koehl P, Michelmore R (2006) Plant NBS-LRR proteins: adaptable guards. Genome Biol 7(4):212

Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10:135–151. doi:10.1146/annurev-genom-082908-145957

Parvin S, Kim YJ, Pulla RK, Sathiyamoorthy S, Miah MG, Wasnik NG, Yang DC (2010) Identification and characterization of spermidine synthase gene from *Panax ginseng*. Mol Biol Rep 37(2):923–932. doi:10.1007/s11033-009-9725-x

Pulla RK, Lee OR, In JG, Kim YJ, Senthil K, Yang DC (2010) Expression and functional characterization of pathogenesis-related protein family 10 gene, PgPR10-2, from *Panax ginseng* C. A. Meyer. Physiol Mol Plant Pathol. doi:10.1016/j.pmpp.2010.05.001

Pulla RK, Shim JS, Kim YJ, Jeong DY, In JG, Lee BS, Yang DC (2009) Molecular cloning and characterization of the gene encoding cinnamyl alcohol dehydrogenase in *Panax ginseng* C.A. Meyer. Korean J Med Crop Sci 17(4):266–272

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33(Web Server issue):W116–W120. doi:10.1093/nar/gki442

Razmovski-Naumovski V, Huang T, Tran V, Li G, Duke C, Roufogalis B (2005) Chemistry and pharmacology of *Gynostemma pentaphyllum*. Phytochem Rev 4(2):197–219. doi:10.1007/s11101-005-3754-4

Sathiyamoorthy S, In JG, Gayathri S, Kim YJ, Yang D (2010a) Gene ontology study of methyl jasmonate-treated and non-treated hairy roots of *Panax ginseng* to identify genes involved in secondary metabolic pathway. Genetika 46(7):932–939

Sathiyamoorthy S, In JG, Gayathri S, Kim YJ, Yang DC (2010b) Generation and gene ontology based analysis of expressed sequence tags (EST) from a *Panax ginseng* C. A. Meyer roots. Mol Biol Rep 37(7):3465–3472. doi:10.1007/s11033-009-9938-z

Sathiyamoorthy S, In JG, Lee OR, Lee BS, Devi SR, Yang DC (2010c) In silico gene expression analysis in *Codonopsis lanceolata* root. Mol Biol Rep 38(5):3541–3549. doi:10.1007/s11033-010-0464-9

Sathiyaraj G, Srinivasan S, Subramaniyam S, Kim YJ, Kwon WS, Yang DC (2010) Polygalacturonase inhibiting protein: isolation, developmental regulation and pathogen related expression in *Panax ginseng* C.A. Meyer. Mol Biol Rep 37(7):3445–3454. doi:10.1007/s11033-009-9936-1

The Gene Ontology C (2010) The Gene Ontology in 2010: extensions and refinements. Nucl Acids Res 38(suppl_1):D331–D335. doi:10.1093/nar/gkp1018

Vogt T (2010) Phenylpropanoid biosynthesis. Mol Plant 3(1):2–20. doi:10.1093/mp/ssp106

Wang P, Niu L, Guo XD, Gao L, Li WX, Jia D, Wang XL, Ma LT, Gao GD (2010) Gypenosides protects dopaminergic neurons in primary culture against MPP(+)-induced oxidative injury. Brain Res Bull 83(5):266–271. doi:10.1016/j.brainresbull.2010.06.014

Xiang WJ, Guo CY, Ma L, Hu LH (2010) Dammarane-type glycosides and long chain sesquiterpene glycosides from *Gynostemma yixingense*. Fitoterapia 81(4):248–252. doi:10.1016/j.fitote.2009.09.009

Yan L, Hong W, He-Chun YE, Guo-Feng LI (2005) Advances in the plant isoprenoid biosynthesis pathway and its metabolic engineering. J Integr Plant Biol 47(7):769–782

Yin F, Hu L, Lou F, Pan R (2004a) Dammarane-type glycosides from *Gynostemma pentaphyllum*. J Nat Prod 67(6):942–952. doi:10.1021/np0499012

Yin F, Hu L, Pan R (2004b) Novel dammarane-type glycosides from *Gynostemma pentaphyllum*. Chem Pharm Bull (Tokyo) 52(12):1440–1444. doi:JST.JSTAGE/cpb/52.1440

Yin F, Zhang YN, Yang ZY, Hu LH (2006) Nine new dammarane saponins from *Gynostemma pentaphyllum*. Chem Biodivers 3(7):771–782. doi:10.1002/cbdv.200690079

Yu JK, Jung HL, Dae YJ, Gayathri S, Ju SS, Jun GI, Deok CY (2009) Isolation and characterization of pathogenesis-related protein 5 (PgPR5) gene from *Panax ginseng*. Plant Pathol J 25(4):400–407

Zhang Z, Zhang W, Ji YP, Zhao Y, Wang CG, Hu JF (2010) Gynostemosides A–E, megastigmane glycosides from *Gynostemma pentaphyllum*. Phytochemistry 71(5–6):693–700. doi:10.1016/j.phytochem.2009.12.017