Sequencing of *Panax notoginseng* genome reveals genes involved in disease resistance and ginsenoside biosynthesis

Guangyi Fan[1,2,*], Yuanyuan Fu[2,3,*], Binrui Yang[1,*], Minghua Liu[4,*], He Zhang[2], Xinming Liang[2], Chengcheng Shi[2], Kailong Ma[2], Jiahao Wang[2], Weiqing Liu[2], Libin Shao[2], Chen Huang[1], Min Guo[1], Jing Cai[1], Andrew KC Wong[5], Cheuk-Wing Li[1], Dennis Zhuang[5], Ke-Ji Chen[6], Wei-Hong Cong[6], Xiao Sun[3], Wenbin Chen[2], Xun Xu[2], Stephen Kwok-Wing Tsui[4,7,]†, Xin Liu[2,]†, Simon Ming-Yuen Lee[1,]†.

[1]State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China.

[2]BGI-Qingdao, BGI-Shenzhen, Qingdao 266000, China.

[3]State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210096, China

[4]School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong, China

[5]System Design Engineering, University of Waterloo, Ontario, Canada

[6]Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing, 100091, China

[7]Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong, Hong Kong, China

*These authors contributed equally to this work.

†Correspondence authors: Simon Ming-Yuen Lee (simonlee@umac.mo), Wenbin Chen (chenwenbin@genomics.cn) and Stephen Kwok-Wing Tsui (kwtsui@cuhk.edu.hk).

Abstract

Background: *Panax notoginseng* is a traditional Chinese herb with high medicinal and economic value. There has been considerable research on the pharmacological activities of ginsenosides contained in *Panax* spp.; however, very little is known about the ginsenoside biosynthetic pathway.

Results: We reported the first *de novo* genome of 2.36 Gb of sequences from *P. notoginseng* with 35,451 protein-encoding genes. Compared to other plants, we found notable gene family contraction of disease-resistance genes in *P. notoginseng*, but notable expansion for several ATP-binding cassette (ABC) transporter subfamilies, such as the Gpdr subfamily, indicating that ABCs might be an additional mechanism for the plant to cope with biotic stress. Combining eight transcriptomes of roots and aerial parts, we identified several key genes, their transcription factor binding sites and all their family members involved in the synthesis pathway of ginsenosides in *P. notoginseng*, including dammarenediol synthase, *CYP716* and *UGT71*.

Conclusions: The complete genome analysis of *P. notoginseng*, the first in genus *Panax*, will serve as an important reference sequence for improving breeding and cultivation of this important nutraceutical and medicinal but vulnerable plant species.

Keywords: *Panax notoginseng*, Ginsenosides biosynthesis, Genome, Transcriptome, Traditional Chinese herb.

Background

Genus *Panax* in the Araliaceae, contains some medicinally and economically important ginseng species including *P. ginseng*, *P. quinquefolius* and *P. notoginseng* (Sanqi in Chinese)[1]. The United States, Canada, China and South Korea are the biggest producers of ginseng, with their total production of fresh ginseng being approximately 175.85 million pounds and constituting more than 99% of the world's total ginseng production[2]. The world ginseng market is estimated to be worth $2,084 million[2]. Unlike *P. ginseng* and *P. quinquefolius*, which are widely distributed in several countries in the northern hemisphere, growth of *P. notoginseng* is mainly restricted to mountain areas with altitudes of 1200–2000 m around 23.5°N and 104°E in Wenshan Prefecture, Yunnan Province, in China[3]. In China, *P. notoginseng* has been cultivated for about 400 years[4] and the cultivated area was estimated at around 8300 ha in late 2005, harvesting 7.03×10^6 kg of fresh roots of *P. notoginseng*[5]. The stable *P. notoginseng* population is relatively genetically homogeneous compared with the other ginseng species.

Panax notoginseng is susceptible to a wide range of pathogens and identification of the genes conferring disease resistance has been a major focus of research[6], which greatly reduces its economic benefits. The main class of disease-resistance genes (R-genes) consist of a NBS, a C-terminal LRR and a putative coiled coil domain (CC) at the N-terminus or a N-terminal domain with homology to the mammalian toll interleukin 1 receptor (TIR) domain[7]. Generally, the LRR domain is often involved in protein–protein interactions, with an important role in recognition specificity as well as ligand binding[8]. Besides, ATP-binding cassette (ABC) proteins have also been reported to play a crucial role in the regulation of resistance processes in plants[9], most of which modulate the activity of heterologous channels or have intrinsic channel activity. The reason underlying the poor resistance of *P. notoginseng* is still unclear.

Saponin constituents, also known as ginsenosides, are the major active ingredients in genus *Panax*[10, 11] and have diverse pharmacological activities, such as hepatoprotection, renoprotection, estrogen-like activities and protection against cerebro-cardiovascular ischemia and dyslipidaemia in experimental models[12-16]. Xuesaitong (in Chinese) is a prescription botanical drug, manufactured from saponins in the root of *P. notoginseng*[12], used for the prevention and treatment of cardiovascular diseases, and is among the top best-selling prescribed Chinese medicines in China[17, 18]. Compound Danshen Dripping Pill, a prescription botanical drug for cardiovascular disease[19], comprises *P. notoginseng*, *Salvia miltiorrhiza* and synthetic borneol, and has successfully completed Phase II and is undergoing Phase III clinical trials in the United States. Although different ginseng species have been one of the most important traded herbs for human health around the world, the industry is facing elevating challenges, such as increasing disease vulnerability of the cultivated plant and an as yet undetermined disease preventing continuous cultivation on the same land, leading to a trend of decreasing production[2].

Plant metabolites play very important roles in plant defense mechanisms against stress and disease, and some have important nutritional value for human consumption. For instance, capsaicinoids and caffeine present in pepper and coffee, respectively, have important health benefits**.** The recently completed whole-genome sequencing of coffee and pepper provide new clues in understanding evolution and transcriptional control of the biosynthesis pathway of these metabolites[20, 21]. Ginsenosides, the oligosaccharide glycosides of a series of dammarane- or oleanane-type triterpenoid glycosides[11], are a group of secondary metabolites of isoprenoidal natural products, specifically present in genus *Panax*. Based on the structure differentiation of the sapogenins (aglycones of saponin), dammarane-type ginsenosides can be classified into two subtypes: protopanaxadiol (PPD) and protopanaxatriol (PPT) types[11]. Up to now, over 100 structurally diversified ginsenosides have been isolated. Interestingly, the two types (PPD and PPT) of ginsenosides have been reported to exhibit opposing biological activities; for instance, pro-angiogenesis and anti-angiogenesis[22, 23]. The whole plant of *P. notoginseng* contains both PPD- and PPT-type ginsenosides but the aerial parts (e.g. leaf and flower) contain a higher abundance of PPD compared to roots[24, 25]. Herein, we characterize the genome of *P. notoginseng*, the population of which has a relatively uniform genetic background. This was aimed to determine the genetic causes of weakened resistance of *P. notoginseng* as well as to understand evolution and regulation of the triterpenoid saponin biosynthesis pathway leading to the production of specialized ginsenosides in genus *Panax*.

Data descriptions

Genomic DNA was isolated from leaf of *P. notoginseng* from Wenshan City of China and then six libraries were constructed with various insert sizes ranging from 170 bp to 10 kb (including 170bp, 500bp, 800bp, 2kb, 5kb and 10kb). In total, 377.44 Gb (~153-fold) of raw sequences were generated using the Illumina Hiseq 2000 platform and ~13 Gb (~6-fold, average read length of ~9kb) SMRT sequence data generated by Pacbio RSII sequencing system (**Supplementary Table 1**). To meet the requirement of further analysis, we obtained ~256.07 Gb data (~104.09-fold) by filtering the low quality (≤ 7) and adapter contaminated reads. Leaves and roots of *P. notoginseng* were collected from 1-, 2- and 3-year-old *P. notoginseng* and flowers collected from 2- and 3-year-old *P. notoginseng*. Total RNA was extracted from each part and then mRNA was isolated. An average of ~6.06 Gb raw data were obtained from each sample sequenced using Illumina HiSeq 2000 (**Supplementary Table 2**).

Analyses

Genome assembly and annotation

We *de novo* sequenced and assembled the *P. notoginseng* genome using ~256.07 Gb data (~104.09-fold) generated by Illumina HiSeq2000 and ~13 Gb (~6-fold, average read length of ~9kb) SMRT sequence data generated by Pacbio RSII sequencing system. The final genome assembly spanned about 2.36 Gb (~95.93% of the estimated genome size) with scaffold N50 of 72.37 kb and contig N50 of 16.42 kb (**Supplementary Table 3**, **4** and **Supplementary Fig. 1)**. To evaluate our assembly, we firstly checked the links within scaffolds based on the pair-end relationships of sequencing reads and the result revealed that pair-end information supported the links of scaffolds well (**Supplementary Fig. 2**). Secondly, we obtained the unmapped sequencing reads against our genome assembly and re-assembled them into ~359 Mb sequences which contained ~95.05% repetitive sequences (**Supplementary Table 5**). Using KAT sect tool[26], we also calculated the sequencing depth distribution of flanking sequences of gap regions, we observed that ~71% of sequences nearby junction regions have higher sequencing depth (**Supplementary Fig. 3**), indicating that majority of the gap regions were repetitive sequences. Finally, more than 95.42% of transcripts (coverage ratio >= 90%) could be unambiguously mapped into assembly sequences revealing the good integrity of genome assembly (**Supplementary Table 6**). Of the genome, 72.45% of repetitive sequences ratio was estimated by Jellyfish[27] but only 1.24 Gb (51.03%) was found to be repetitive sequences, which is more than that of carrot genome (~46%)[28], with long terminal repeats the most abundant (~95.08% of transposable elements) (**Supplementary Table 7**). We totally predicted 35,451 protein-encoding genes in the *P. notoginseng* genome by integrating the evidences of *ab initio* prediction, homologous alignment and transcripts (**Supplementary Table 8**). Assessing the *P. notoginseng* genome assembly and annotation completeness with single-copy orthologs approach (BUSCO)[29] showed that 895 (93%) complete single-copy genes containing 246 (25%) complete duplicated genes were validated, similar to that of cotton[30] and *Dendrobium officinale*[31]. In combination with the RNA-seq mapping results above, it was unambiguous that our gene set was available for further analysis.

Genome evolution of *Panax* species

Panax notoginseng is the first sequenced genome of genus *Panax*, thus we compared its genome to other related species to reveal genome evolution of *Panax* species. To analyze *P. notoginseng* evolution, we first constructed the phylogenetic tree using published closely-related species (**Supplementary Fig. 4**). We estimated the species divergence time between *P. notoginseng* and *D. carota* to be about 71.9 million years ago (**Supplementary Fig. 4**). Compared to closely related species (*D. carota*, *S. tuberosum*, *S. lycopersicum* and *C. annuum*), we found 1,144 unique gene families (containing 2,714 genes) with no orthologs in other species (**Supplementary Fig. 5**), which were possibly related to specific features of *P. notoginseng*. Thus, we performed functional enrichment of these genes to find several interesting Gene Ontology (GO) terms likely related to saponin biosynthesis, such as UDP-glucosyltransferase (UGT) activity (P = 0.003, **Supplementary Fig. 6**). To determine features and specific functions in *P. notoginseng*, we identified a total of 1,520 expanded gene families (**Supplementary Fig. 4**) and further performed Kyoto Encyclopedia of Genes and Genomes (KEGG) and GO functional enrichment analyses for these genes. We found several significant gene families related to saponin biosynthesis, such as sesquiterpenoid and triterpenoid biosynthesis (P = 0.006, **Supplementary Table 9**).

Resistance genes

Most cloned R-genes in the plant encode nucleotide-binding site and leucine-rich-repeat (NBS-LRR) domains. In our *P. notoginseng* genome assembly, 129 NBS-LRR-encoding genes were detected, which was notably fewer than *Arabidopsis thaliana* (183), *D. carota* (170), *S. tuberosum* (441), *S. lycopersicum* (282) and *C. annuum* (778), and similar to that of *C. sativus* (89)[32] (**Supplementary Table 10**). Considering R genes were known to be small and in repetitive areas and our assembly with much gaps in the repetitive regions, the total number of R gene of *P. notoginseng* would be likely to underestimated. However, the number of genes in the TIR_NBA_LRR subfamily was still relatively expanded according to the phylogenetic analysis using *P. notoginseng*, tomato and carrot, which possibly be related to the resistance character of *P. notoginseng* (**Fig. 1a** and **1b**).

ABC transporter gene family

We identified a total of 153 ATP-binding cassette (*ABC*) genes in *P. notoginseng*, which were classified into nine subfamilies (**Supplementary Table 11** and **Supplementary Fig. 7**). Generally, compared with *A. thaliana* (12), *Daucus carota* (5), *Solanum lycopersicum* (10), *S. tuberosum* (13) and *Capsicum annuum* (10), there were notably fewer members of subfamily A in *P. notoginseng* (3); whereas, there were notably more of subfamily B in *P. notoginseng* (28, 28, 33, 38, 34 and 47 for subfamily B, respectively) (**Supplementary Table 11**). In detail, two genes were notably contracted for subfamily A (**Supplementary Fig. 8**) **–** a subfamily reportedly involved in cellular lipid transport[33].

For subfamily B, we found a specific expanded clade in *P. notoginseng* containing 11 genes, three duplication expansions and one contraction (**Supplementary Fig. 7** and **Supplementary Fig. 9**). In particular, duplication expansions were Pno037415.1 and Pno028859.1 (AT2G36910.1 homologs); Pno009402.1 and Pno037963.1 (AT3G28860.1 homologs); and Pno034123.1 and Pno018232.1 (AT5G39040.1 homologs). One contraction was Pno003332.1 (AT4G28620.1, AT4G28630.1 and AT5G58270.1 homolog). Interestingly, the expanded genes have been reported as involved in aluminum resistance and auxin transport in stems and roots, respectively[34, 35]. The contracted gene was found to be involved in iron export[36]. In the Gpdr subfamily, we also found a notable expansion (*P. notoginseng*: 10 vs. *A. thaliana*: 1, AT1G15520.1) (**Fig. 1c**). The homolog of AT1G15520.1 is known to be related to pleiotropic drug resistance and abscisic acid (ABA) uptake transport[37, 38].

Key genes involved in ginsenoside biosynthesis of *P. notoginseng*

Dammarenediol synthase

Ginsenosides are committed to be synthesized from dammarenediol-II after hydroxylation via cytochrome P450 (*CYP450*)[39] and then glycosylation by glycosyltransferase (*GT*)[40]. Dammarenediol-II is synthesized from farnesyl-PP, which is synthesized via the terpenoid backbone biosynthesis pathway. There are three enzymes involved from farnesyl-PP to dammarenediol-II: farnesyl-diphosphate farnesyltransferase (*FDFT1*, K00801), squalene monooxygenase (*SQLE*, K00511) and dammarenediol synthase (*DDS*, K15817) (**Fig. 2a**). In total, we identified three *SQLE*, two *FDFT1* and one *DDS* in *P. notoginseng*. Combining with the results of real-time PCR, *DDS* (Pno034035.1) was found with extensive higher expression in the root than aerial parts and in 3-year old plant both root and aerial parts were found with higher expression of *DDS* (Pno034035.1). We also found that expressions of one *FDFT1* (Pno029444.1) and one *SQLE* (Pno022472.1) were higher in roots compared with flowers and leaves, and one *SQLE* (Pno009162) gene showed the opposite pattern according to RNA-seq, while the real-time PCR results suggested an increasing expression year after year of all these three genes (**Fig. 2b**). Phylogenetic analysis showed that *DDS* of *P. notoginseng* and *P. ginseng* had a very high (98.96%) amino acid sequence identity, differing in only eight amino acids, and with fewer genes than other species (**Supplementary Fig. 10**). We further found there were three amino-acid residue insertions (L194, A195 and E196) in the *DDS* sequences of *P. notoginseng* and *P. ginseng*, which were located in the cyclase-N domain (**Fig. 2c**). We used the SWISS-MODEL[41] to conduct structural modeling of *DDS* protein using the human oxidosqualene cyclase (*OSC*) in complex with Ro 48-80771[42](SMTL: 1w6j.1) as the template. When the locations of the amino acids were mapped to the protein 3D structure with amino-acid identity of 40%, two residue insertions (L194 and A195) were located between two DNA helices, which are the presumed substrate-binding pocket and the catalytic residues (**Fig. 2d**). These three amino acids of *DDS* sequences of *P. notoginseng* probably played an important role in the Dammarenediol-II biosynthesis.

CYP450s

*CYP450*s and *GT*s have been demonstrated to be involved in hydroxylation or glycosylation of aglycones for triterpene saponin biosynthesis. We identified a total of 268 *CYP450* genes in *P. notoginseng* using 243 *CYP450* homolog genes of *A. thaliana*, which is consistent with the report that there are around 300 *CYP450* genes in genomes of flowering plants[43]. To compare with other species, we also identified *CYP450* genes of four other genomes—*D. carota* (325), *S. tuberosum* (465), *S. lycopersicum* (252) and *C. annuum* (256)—using the same parameters as for *P. notoginseng* (**Supplementary Table 12** and **Supplementary Fig. 11**). We further classified the subfamilies of *CYP450* genes for each species based on the category of *CYP450* in *A. thaliana* and nine *CYP716* genes were identified in *P. notoginseng* (**Fig. 3a**). *CYP716A47* was reported to produce PPD[44] and *CYP716A53v2* was reported to be involved in the synthesis of PPT in *P. ginseng* (**Fig. 3b**). Moreover, considering the role of different types of notoginsenoside in different parts/tissues of *P. notoginseng*— that is, PPT was relatively higher in roots while PPD was higher in aerial parts (leaves and flowers) (**Fig. 3c**)— we sequenced the transcriptomes of leaves, roots and flowers collected from different ages of *P. notoginseng*. We found Pno012347.1 (*CYP716A47* homolog), Pno002960.1 (*CYP716A53v2* homolog) and Pno011760.1 were always more highly expressed in roots than in leaves and flowers for all ages; while Pno021283.1 (*CYP716A47* homolog) was the opposite (**Fig. 3d**). Thus, Pno012347.1 and Pno021283.1 were the candidate genes involved in PPD biosynthesis and Pno002960.1 was the candidate gene involved in PPT biosynthesis in *P. notoginseng*. The expression levels of three *CYP716* genes (Pno012347.1, Pno002960.1 and Pno011760.1) validated using real-time PCR were consistent with RNA-seq. Meanwhile, we identified another 22 *CYP450* genes (including *CYP78*, *CYP71*, *CYP72* and *CYP86*), which had the same expression pattern as three *CYP716A53v2*-homolog genes (**Fig. 3d**), indicated they may be involved into PPT biosynthesis.

UGTs

Three UGTs (*UGT71A27*, *UGT74AE2* and *UGT94Q2*) that participate in the biosynthesis of PPD-type ginsenosides and three UGTs (*UGTPg1*, *UGTPg100* and *UGTPg101*) participating in the formation of PPT-type saponins have previously been identified in *P. ginseng*. We identified a total of 160 UGT genes in the *P. notoginseng* genome based on 116 homologous UGT genes of *A. thaliana*[45]. We classified these 160 UGTs of *P. notoginseng* into 12 subfamilies containing 19 *UGT71A* homologs and 14 UTG74 homologs (**Supplementary Fig. 12**). *UGTPg1*, *UGTPg100*, *UGTPg101*, *UGTPg102* and *UGTPg103* of *P. ginseng* were *UGT71A27* homologs; however, *UGTPg102* and *UGTPg103* were found to have no detectable activity on PPT due to the lack of several key amino acids in their proteins (**Fig. 4a** and **b**). We identified 38 differentially expressed UGT genes between roots and aerial parts (flowers and leaves), including 23 expressed more highly in roots (HR) and 15 in aerial tissues (HA) (**Fig. 4c** and **d**). Interestingly, there were three *UGT71A27* homolog (Pno020280.1, Pno027722.1 and Pno026280.1) genes in the HR, and HA included four UGT74 genes (Pno031515.1, Pno033394.1, Pno002810.1 and Pno000722.1) and one *UGT71A27* homolog (Pno013844.1) gene (**Fig. 4c** and **d**).

Furthermore, when comparing the *UGT71A27* homolog genes of *P. notoginseng* with *UTGPg1*, *UGTPg100*, *UGTPg101*, *UGTPg102* and *UGTPg103*, we checked the known key amino-acid residues determining the function of UGTs. We found that the key amino-acid residues *UGTPg100*-A142T, *UGTPg100*-L186S, *UGTPg100*-G338R and *UGTPg1*-H82C were conserved in both Pno026280.1 and Pno27722.1. However, *UGTPg1*-H144F is a specific mutation F in the Pno026280.1 (**Supplementary Fig. 13**). All these candidate genes probably play important roles in the biosynthesis of ginsenoside.

Identification of transcription factor binding sites

We have predicted the transcription factor binding site (TFBS) in the upstream regions of ten genes encoding enzymes involved in the ginsenoside biosynthesis, including HMG-CoA synthase (HMGS), HMG-CoA reductase (HMGR), mevalonate kinase (MK), phosphomevalonate kinase (PMK), mevalonate diphosphate decarboxylase (MDD), isopentenylpyrophosphate isomerase (IPI), geranylgeranyl diphosphate synthase (GGR), farnesyl diphosphate synthase (FPS), squalene epoxidase (SE) and dammarenediol-II synthase (DS) (**Supplementary Fig. 14**). DNA-binding with one finger 1 (Dof1) and prolamin box binding factor (PBF) binding sites were detected in all these ten genes which used the same binding site motif (5'-AA[AG]G-3') (**Supplementary Table 13**). Therefore, the expression levels of Dof1 and PBF with these ten genes in different developmental stages of roots and leaves from *P. notoginseng* were compared (**Supplementary Table 14**). Most of genes' expression levels were higher in the 3 roots than that in the 3 leaves, expecting PMK and Dof1 (**Supplementary Fig. 15**). Dof1 belong to the Dof family, which is plant-specific transcription factor [46]. Yanagisawa's studies suggested that Dof1 in maize is related with the light-regulated plant-specific gene expression [47, 48]. PBF encodes Dof zinc finger DNA binding proteins, which could enhance the DNA binding of the bZIP transcription factor Opaque-2 to O2 binding site elements [49]. In three stages of roots, the trend of PBF expression level was consistent with that of DS, HMGR, SE, HMGS, MK, MDD, FPS and GGR (**Supplementary Table 15**). Whereas, in different developmental stage of leaves, there was almost no significant change in genes expression levels in these genes, excluding Dof1 and DS with a rising trend (**Supplementary Table 16**). We inferred that the PBF and Dof1 were probably associated with the transcriptional controls of these genes, involved in triterpene saponins biosynthesis pathway, in different parts, the root and the leaf of *P. notoginseng*, respectively.

Discussion

Compared with other phylogenetically-related plant species, there was notable gene family contraction of R-genes in *P. notoginseng*. The encoded resistance proteins play important roles in plant defense by controlling the host plant's ability to detect a pathogen attack and facilitate a counter attack against the pathogen[7]. Whether *P. notoginseng* has evolved other defenses and/or anti-stress mechanisms to compensate for the contraction of R-genes remains to be determined in data mining of the first draft genome of this genus.

Given that the notable contraction of R-genes of *P. notoginseng* may provide clues to its poor stress and disease resistance, we also observed expansion and contraction of different subfamilies of the ABC transporter gene family in *P. notoginseng*. The substrates of ABC proteins include a wide range of compounds such as peptides, lipids, heavy metal chelates and steroids[50]. Interestingly, one of the contracted gene candidates (Pno003332.1) in *P. notoginseng* was reported to be involved in iron export[36]. A previous experimental study showed that optimal iron concentration could significantly determine the plant biomass and stabilize arsenic content in soil to reduce arsenic contamination[51] where arsenic contamination of *P. notoginseng* is a serious agricultural problem due to arsenic pollution in the environment[51].

Phenolics, alkaloids and terpenoids are three major classes of chemicals involved in plant defenses[52]. Another possible complementary mechanism to cope with reduced pathogen resistance, which possibly resulted from R-gene contraction, is through evolving synthesis of defensive secondary plant metabolites. Pepper and tomato, two species phylogenetically close to *P. notoginseng*, produce alkaloids such as capsaicinoids and tomatine, respectively, which have been found to function as deterrents against pathogens[21, 53]. Azadiracht, a meliacane-type triterpenoid with very similar skeleton to dammarane from the neem tree (*Azadirachta indica*)[54], has insecticidal properties. Ginsenosides are the oligosaccharide glycosides of a series of dammarane- or oleanane-type triterpenoid glycosides[11]. Palazón *et al*. reported that during *in vitro* culture of ginseng hair root lines, the addition of methyl jasmonate, a signaling molecule specifically expressed by plants in response to insect and pathogenic attacks, could enhance overall ginsenoside production and conversion of PPD-type (e.g. Rb1, Rb2, Rc and Rd) to PPT-type (e.g. Re, Rf and Rg1) ginsenoside[55]. In addition, natural ginsenosides have antimicrobial and antifungal action, as shown in numerous laboratory studies[56, 57]. The exact natural roles of ginsenosides in plants, particularly controlling plant growth and defense against pathogenic infection, remain unclear and require further investigation.

DDS is the critical determinant for the initial step of producing the dammarane-type sapogenin (aglycone of saponin) template. This can be further subjected to structural diversification by *CYP450* and UGT enzymes. Along the triterpenoid saponin biosynthesis pathway, DDS of the OSC group is responsible for the cyclization of 2,3-oxidosqualene to dammarenediol-II[58]. A highly-conserved *DDS* homolog of *P. ginseng* was also identified in *P. notoginseng*. Upstream of *DDS* in the triterpenoid saponin biosynthesis pathway, three *SQLE* and two *FDFT1* genes were identified and shown to exhibit significant differential expression in different parts in *P. notoginseng*.

Han *et al*. identified that *CYP716A47* produces PPD through the hydroxylation of C-12 in dammarenediol-II[44], *CYP716A53v2* hydroxylates the C-6 of PPD to produce PPT[59], and *CYP716A52v2* is involved in oleanane-type ginsenoside biosynthesis[60] in *P. ginseng*. We were interested in determining how the transcriptional regulation of different CYP450 genes respectively correlated with synthesis of higher content of PPT- and PPD-type ginsenosides in roots and aerial parts (leaves and flowers). Combined the previous research and transcriptome analysis, we identified the key genes which produces PPD and PPT in the *P. notoginseng*.

UGTs are GTs that use uridine diphosphate (*UDP*)-activated sugar molecules as donors. Several types of UGTs, which catalyze the synthesis of specific sapogenins to ginsenosides, were found in previous studies of *P. ginseng*[61]. For example, UGTPg1 was demonstrated to region-specifically glycosylate C20-OH of PPD and PPT[62]. *UGTPg100* specifically glycosylates *PPT* to produce bioactive ginsenoside Rh1, and *UGTPg101* catalyzes *PPT* to produce F1[62]. Up to now, very few of the UGTs characterized that are able to glycosylate triterpenoid saponins have been found in *P. ginseng*[63-66] and *P. notoginseng*[67, 68]. However, in the present study, an almost complete collection of 160 UGT genes, representing many new homologs and putative UGT members, was disclosed in the *P. notoginseng* genome. In our study, we found the lack of correlation between qPCR and RNA-seq in the Pno026280, Pno002772 and Pno020280. In order to eliminate error because of improper normalization method, we used other two quantification methods. Using the raw RNA-seq reads counting and normalized by quantiles (limma) [69]and DESeq[70] showed the completely consistent tendency with FPKM results (**Supplementary Table 17**). The probable reasons are the nonspecific binding of qPCR primers to cDNA template and the bias in the RNA-seq sequencing due to sequencing library preparation, namely, it was due to nature of these two methods (RNA-seq vs. qPCR).

In conclusion, we sequenced and assembled the first *de novo* genome of *P. notoginseng*. Using a combination of data from genome and transcriptomes, we identified contraction and expansion in numerous important genome events during the evolution of this plant. The contraction of disease-resistance genes families figured out the vulnerability to disease of *P. notoginseng* on genetic level and the expansion of several *ABC* transporter subfamilies indicated the potential of *ABC*s as an additional mechanism for the plant to cope with biotic stress. Combining eight transcriptomes of roots and aerial parts, several key genes and all their family members involved in the synthesis pathway of ginsenosides were also identified including dammarenediol synthase, *CYP716* and *UGT71*. These findings provide new insight into its poor resistance and the biosynthesis of high-valued pharmacological molecules. This reference genome will serve as an important platform for improving breeding and cultivation of this vulnerable plant species to increase the medicinal and economic value of *Panax* species.

Methods

K-mer analysis

We used the total length of sequence reads divided by sequencing depth to calculate the genome size. To estimate the sequencing depth, the copy number of 17-mers present in sequence reads was counted and the distribution of sequencing depth of the assembled genome was plotted. The peak value in the curve represents the overall sequencing depth, and can be used to estimate genome size (G): $G = \text{Number}_{17\text{-mer}}/\text{Depth}_{17\text{-mer}}$.

Genome assembly

The first genome assembly version (v1.0) using *SOAPdenovo[71]* (v2.04) based on the short length reads data sequencing by Hiseq2000, was highly fragmented. It contained more than 600Mb gaps with missing sequence because of presence of high proportion of repetitive DNA sequences in the genome. The contig N50 sizes of v1.0 was 4.2kb. For the short contig N50 size, therefore, we further generated ~13Gb SMRT sequence data (~6-fold of whole genome size) with average sub-read length of ~9kb. Considering the inadequate depth of our Pacbio third generation sequencing, we used the PBJelly[72] to fill the gap and upgrade the genome assembly with long length reads. Concretely, we generated our protocol file containing full paths of the reference assembly above, the output directory and the input files. Then, we executed multiple steps (including setup, mapping, support, extraction, assembly, output) in a consecutive order. From the statistics of the result of PBJelly, we found that the gap number was lowered to 533Mb while sizes of contig N50 was increased to 10.7kb. Next, we used the libraries with the insert sizes of 500bp and 800bp as the gap-filling input data of the GapCloser (v1.0, http://soap.genomics.org.cn/). After running, the gap number was decreased to 493Mb and the length of contig N50 was increased to 16.4kb. Finally, the sizes of contig and scaffold N50 of the final version (v1.1) were

16.4kb and 70.6kb, respectively. In addition, the total length of the genome changed from 2.1Gb to 2.4Gb with 493Mb gap sequences. The detailed parameters of *SOAPdenovo* were "pregraph -s lib.cfg.1 -z 2800000000 -K 45 -R -d 1 -o SAN_63 -p 12; contig -g SAN_63 -R -p 24 ; map -s lib.cfg.1 -g SAN_63 -k 45 -p 24; scaff -g SAN_63 -F -p 24". We also adopted ABySS[73] to estimate the optimized K value for genome assembly from 35 to 75 using about 20-fold high quality sequencing reads. The result revealed the K=45 is the best K value (**Supplementary Table 18**), which is completely consistent with the best K value of SOAP*denovo*.

Transcript *de novo* assembly

Leaves and roots of *P. notoginseng* were collected from 1-, 2- and 3-year-old *P. notoginseng* and flowers collected from 2- and 3-year-old *P. notoginseng*. All of these 8 samples were randomly collected from the fileld of commercial planting base in Yanshan county of Wenshan Zhuang and Miao minority autonomous prefecture in Yunnan province. After cleansing, the leaves, roots and flowers were collected separately, cut into small pieces, immediately frozen in liquid nitrogen, and stored at -80 °C until further processing. Subsequently, mRNA isolation, cDNA library construction and sequencing were performed orderly. Briefly, total RNA was extracted from each tissue using TRIzol reagent and digested with DNase I according to the manufacturer's protocol. Next, Oligo magnetic beads were used to isolate mrna from the total rna. By mixing with fragmentation buffer, the mrna was broken into short fragments. The cDNA was synthesized using the mRNA fragments as templates. The short fragments were purified and resolved with EB buffer for end repair and single nucleotide A (adenine) addition, and then connected with adapters. Suitable fragments were selected for PCR amplification as templates. During the quality control steps, an Agilent 2100 Bioanalyzer and ABI StepOnePlus Real-Time PCR System were used for quantification and qualification of the sample library. Each cDNA library was sequenced in a single lane of the Illumina HiSeqTM 2000 system using paired end protocols. Before performing the assembly, firstly, raw sequencing data that generated by Illumina HiSeq2000, was subjected to quality control (QC) check including the analysis of base composition and quality. After QC, raw reads which contained the sequence of adapter (adapters are added to cDNAs during library construction and part of them may be sequenced), more than 10% unknown bases or 5% low quality bases were filtered into clean reads. Finally, Trinity[74] was used to perform the assembly of clean reads (detail information could be retrieved in Trinity website http://trinityrnaseq.sourceforge.net/). Based on the Trinity original assembly result, TGICL[75] and Phrap[76] were used to acquire sequences that cannot be extended on either end, the sequences are final assembled transcripts. We mapped all transcripts that were *de novo* assembled from nine transcriptome short reads by Trinity into the genome assembly using Blat[77], and then calculated their coverage and mapping rate. The sequencing depth distribution was examined by aligning all short reads against assembly using SOAP2[78] and then the depth of each base was calculated.

Gene expression analysis

All the QC and filter processes of raw sequencing transcriptome data were the same as that of *de novo* assembly. All the clean reads are mapped to reference gene sequences by using *SOAP*2[78] with no more than 5 mismatches allowed in the alignment. We also performed the QC of alignment results because mRNAs were firstly broken into short segments by chemical methods during the library construction and then sequenced. If the randomness was poor, read preference for specific gene region could influence the calculation of gene expression. The gene expression level of each gene was calculated by using RPKM method[79] (reads per kilobase transcriptome per million mapped reads) based on the unique alignment results. In the previous test, comparative analysis between RPKM and FPKM showed that the correlation between RPKM and QPCR was better than the correlation between FPKM and QPCR. Referring to the previous study[80], we have developed a stringent algorithm to identify differentially expressed genes between two samples. The probability of a gene expressed equally between two samples was calculated based on the Poisson distribution. P-values corresponding to differential gene expression test were generated using Benjamini, Yekutieli.2001 FDR method[81]. Moreover, because of some debates in the normalization method by gene size like

RPKM and then two new quantification methods were conducted to validate the existing results. Concretely, based on the tophat mapping results, R package 'Rsubread' was used to do the raw reads counting then we used 'limma' with the normalize parameter "quantile". Additionally, DESeq also was conducted. Finally, aiming at the differential expressed genes mentioned in the article based on RNA-seq (RPKM), we found the completely same tendency.

Identification of repetitive sequences

Transposable elements (TEs) in the *P. notoginseng* genome were annotated with structure-based analyses and homology-based comparisons. In detail, RepeatModeler[23] was used to *de novo* find these TEs based on features of structures. RepeatMasker and RepeatProteinMask were applied using RepBase[82] for TE identification at the DNA and protein level, respectively. Overlapped TEs belonging to the same repeat class were checked and redundant sequences were removed. TRF software[83] was applied to annotate tandem repeats in the genome.

Gene prediction and annotation

We predicted protein-encoding genes by homolog, *de novo* and RNA-seq evidence; and results of the former two methods were integrated by the GLEAN program[84] then filtered by threshold level of 20% percent overlap with at least 3 homolog species support. Subsequently, the inner auto pipeline was used to integrated RNA-seq data within and then checked manually. Firstly, protein sequences from five closely related species—*A. thaliana*, *S. lycopersicum*, *S. tuberosum*, *Capsicum annuum* and *Daucus carota*—were applied for homolog prediction by respectively mapping them to the genome assembly using TblastN software with E-value -1e-5. Aiming at these target areas, we used Genewise[85] to cluster and filter pseudogenes. Then Fgenesh[86], AUGUSTUS[87] and GlimmerHMM[88] were used for *de novo* prediction with parameters trained on A. thaliana. We merged three de novo predictions into a unigene set. *De novo* gene models that were supported by one more *de novo* methods were retained. For overlapping gene models, the longest one was selected and finally, we got de novo-based gene models. Using *de novo* gene set (30,660-41,495) and five homolog-based results as gene models (24,358 to 36,116) integration was done using the GLEAN program. Finally, we got the GLEAN gene set (referred to as G-set, 31,678) with parameter “value 0.01, fixed 0”. Combined with the RNA transcriptome sequencing data, we used Tophat2[89] to map and Cufflinks[90] to assemble reads into transcripts and finally integrated the transcripts into the gene set using inner pipeline. The process of inner pipeline was: (1) The assembly of transcriptome data. RNA-seq reads were mapped to *P. notoginseng* genome using tophat, and then cufflinks pipeline was used to conduct the assembly of transcripts. (2) Prediction of the ORF in transcripts (CDS length >=300bp, score >-15). (3) Comparison of transcripts and Glean results to count the overlap region (identity >=95%, total/len >=90%). (4) Integration of the existing results. Aiming at the results, we checked manually by discarding genes without any functional annotation information, any homolog support or cufflinks results support. The final gene set (35,451) was obtained. Then, with the known protein databases (SwissProt[91], Trembl, KEGG[92], InterPro[93] and GO[94]), we annotated these functional proteins in the gene set by aligning with template sequences in the databases using BLASTP (1e-5) and InterProScan.

Identification of R-genes

Most R-genes in plants encode NBS-LRR proteins. According to the conservative structural characteristics of domains, we used HMMER (V3, http://hmmer.janelia.org/software) to screen the domains in the Pfam NBS (NB-ARC) family. Then we compared all NBS-encoding genes with TIR HMM (PF01582) and LRR 1 HMM (PF00560) data sets using HMMER (V3). For the CC domains, we used the MARCOIL program[95] with a threshold probability of 0.9 and double-checked using paircoil2[96] with a P-score cut-off of 0.025.

Gene cluster analysis

We used OrthoMCL[97] to identify the gene family and so obtained the single-copy gene families and multi-gene families that were conserved among species. Then we constructed a phylogenetic tree based on the single-copy orthologous gene families using PhyML[98]. The different molecular clock (divergence rate) might be explained by the generation-time hypothesis. Here, we used the MCMCTREE program[99] (a

BRMC approach from the PAML package) to estimate the species divergence time. 'Correlated molecular clock' and 'JC69' model in the MCMCTREE program were used in our calculation.

Analysis of key gene families

The reference genes of interest in *A. thaliana* (such as ABC, CYP450 and UGT) were found in the TAIR10 functional descriptions file. Then, the target genes of *P. notoginseng* were identified and classified using BlastP with cut-off value of 1e-5 and constructing gene trees using PhyML (-d aa -b 100). The tree representation was constructed using MEGA software.

Real-time PCR analysis

Isolated RNA from different *P. notoginseng* tissues were reverse-transcribed to single-strand cDNA using the Super Script™ III First-Strand Synthesis System (Invitrogen™, USA). Quantitative reactions were performed on the Real-Time PCR Detection System (VIIA 7^{TM} Real-Time PCR System, Applied Biosystems, USA) using FastStart Universal Probe Master and Universal ProbeLibrary Probes (Roche, Switzerland). All primers used in this study are listed in **Supplementary Table 19**. The relative gene expression was calculated with the $2^{-\Delta\Delta CT}$ method. For each sample, the mRNA levels of the target genes were normalized to that of 18S rRNA.

Availability of supporting data and materials

The genome assembly and sequencing data have been deposited into NCBI Sequence Read Archive (SRA) under project number PRJNA299863 and into *Giga*DB.

Declarations

Acknowledgements

This work was supported by the Science and Technology Development Fund (FDCT) of Macao SAR (Ref. No. 134/2014/A3), Research Committee of University of Macau (MYRG139(Y1-L4)-ICMS12-LMY and MYRG2015-00214-ICMS-QRCM), the

Overseas and Hong Kong, Macau Young Scholars Collaborative Research Fund by the Natural National Science Foundation of China (Grant no. 81328025), Technology Innovation Program Support by Shenzhen Municipal Government (NO.JSGG20130918102805062) and Basic Research Program Support by Shenzhen Municipal Government (NO.JCYJ20140729163617 29 and NO. JCYJ20150831201123287).

Additional files

Additional file 1: Supplementary information includes supporting figures and supporting tables.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

S. M.Y.L, X.X and X.L designed the project. G.F, B.R.Y, Y. F, H.Z, X.L, C.S, K.M, J.H, R.G, L.S, S.D, Q.X, W.L, M.L, A.K.W, D.Z and M.C analyzed the data. W.C, X.L, G.F, Y. F, J.C and B.R.Y wrote the manuscript. B.R.Y, M.G and C. H prepared the samples and conducted the experiments.

Authors' Affiliations

(1) State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China.

(2) BGI-Shenzhen, Shenzhen 518083, China.

(3) State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210096, China

(4) Faculty of Science and Technology, Department of Civil and Environmental Engineering, University of Macau, Macao, China

Figure legends

Fig. 1. The gene families related to resistance of *P. notoginseng*. (a) The evolution of R-genes with NBS domains among *S. lycopersicum* (*Sl*), *D. carota* (*Dc*) and *P. notoginseng* (*Pn*). The numbers in circles represent the gene numbers of a species and the numbers in clades represent the gene number of expanded gene families (E), contracted gene families (C) and extinct gene families (D). NBS represents a gene that only contains the NBS domain, and NBS_LRR, TIR_NBS_LRR and CC_NBS_LRR follow a similar representation. **(b)** Phylogenetic tree of R genes of *P. notoginseng, D. carota* and *Oryza sativa* (outgroup). Blue label is *P. notoginseng*, orange is *D. carota* and purple is *Oryza sativa*. **(c)** The gene trees of ABC Gpdr subfamily of *P. notoginseng* and other five plants. The red represents genes of *P. notoginseng*, the green is pepper, the orange is potato, dark blue is tomato, the cyan is carrot and the blue is Arabidopsis.

.

Fig. 2. The preliminary steps of ginsenoside biosynthesis and several key gene families. (a) The dammarenediol-II synthesis pathway from glycolysis involves terpenoid backbone biosynthesis and finally catalysis by DDS. **(b)** Comparison of gene expressions of three key enzyme gene families between the roots and aerial parts (flowers and leaves) and the qPCR results. R1 represents one-year-old roots of *P. notoginseng*, R2: two-year-old roots, R3: three-year-old roots from plant B, L1: one-year-old leaves, L2, two-year-old leaves, L3: three-year-old leaves, F2: two-year-old flowers, F3: three-year-old flowers. **(c)** Protein sequence multiple comparisons of DDS among *P. notoginseng*, *P. ginseng* and other representative plants. Stars represent amino acids conserved in all protein sequences, red circles represent amino acids that are the same in *P. notoginseng* and *P. ginseng*, but missing in all other plants. **(d)** The protein 3D structure of DDS of *P. notoginseng* constructed by SWISS-MODEL. The arrows indicate two amino acids specific to *P. notoginseng* and

P. ginseng.

Fig. 3. The CYP450 gene families of *P. notoginseng*. (a) Phylogenetic analysis of CYP716 subfamily among *P. notoginseng*, *P. ginseng* and other plants. Red circles represent *P. notoginseng*, purple diamonds represent *P. ginseng*, green circles represent *S. tuberosum*, yellow triangles represent *D. carota*, purple squares represent *S. lycopersicum* and green diamonds represent *A. thaliana*. The lengths of clades in the gene tree show the protein similarity between two clades. **(b)** Synthesis pathway of protopanaxadiol (PPD) and protopanaxatriol (PPT) of *P. notoginseng*. **(c)**The relative amount of PPD was more than that of PPT in the aerial parts of *P. notoginseng*, but the opposite in roots. **(d)** The differentially expressed CYP450 genes of *P. notoginseng* between the aerial parts and roots and the qPCR results. The pink and green genes in the heatmap represent CYP716 subfamily genes that were involved in the synthesis pathway in *P. ginseng*.

Fig. 4. The UGT gene families of *P. notoginseng*. **(a)** The synthesis pathway of different PPT-type ginsenosides that were catalyzed by different UGT genes. **(b)** The synthesis pathway of different PPD-type ginsenosides that were catalyzed by different UGT genes. **(c)** Highly expressed UGT genes of *P. notoginseng* in roots than in aerial parts and the qPCR results. **(d)** Highly expressed UGT genes of *P. notoginseng* in aerial parts than in roots and the qRCR results. The purple, green and blue genes in the heatmap represent gene families that were involved in the synthesis pathway in *P. ginseng*.

References

1. Briskin DP. Medicinal plants and phytomedicines. Linking plant biochemistry and physiology to human health. Plant physiology. 2000;124(2):507-14.

2. Baeg IH, So SH. The world ginseng market and the ginseng (Korea). Journal of ginseng research. 2013;37(1):1-7. doi:10.5142/jgr.2013.37.1.

3. Guo H, Cui X, An N, Cai G. Sanchi ginseng (Panax notoginseng (Burkill) F. H. Chen) in China: distribution, cultivation and variations. Genet Resour Crop Evol. 2010;57(3):453-60. doi:10.1007/s10722-010-9531-2.

4. Lee CH, Kim JH. A review on the medicinal potentials of ginseng and ginsenosides on cardiovascular diseases. Journal of ginseng research. 2014;38(3):161-6. doi:10.1016/j.jgr.2014.03.001.

5. Guo XC. New green industry – takes Wenshan Panax notoginseng status and
its future as an example. Ecol Econ 2007;1:114-7.

6. Ou X, Jin H, Guo L, Yang Y, Cui X, Xiao Y et al. [Status and prospective on nutritional physiology and fertilization of Panax notoginseng]. Zhongguo Zhong yao za zhi = Zhongguo zhongyao zazhi = China journal of Chinese materia medica. 2011;36(19):2620-4.

7. Gururani MA, Venkatesh J, Upadhyaya CP, Nookaraju A, Pandey SK, Park SW. Plant disease resistance genes: Current status and future directions. Physiological and Molecular Plant Pathology. 2012;78:51-65. doi:http://dx.doi.org/10.1016/j.pmpp.2012.01.002.

8. Knepper C, Day B. From perception to activation: the molecular-genetic and biochemical landscape of disease resistance signaling in plants. The Arabidopsis book / American Society of Plant Biologists. 2010;8:e012. doi:10.1199/tab.0124.

9. Andolfo G, Ruocco M, Di Donato A, Frusciante L, Lorito M, Scala F et al. Genetic variability and evolutionary diversification of membrane ABC transporters in plants. BMC plant biology. 2015;15:51. doi:10.1186/s12870-014-0323-2.

10. Leung KW, Wong AS. Pharmacology of ginsenosides: a literature review. Chinese medicine. 2010;5:20. doi:10.1186/1749-8546-5-20.

11. Yang WZ, Hu Y, Wu WY, Ye M, Guo DA. Saponins in the genus Panax L. (Araliaceae): a systematic review of their chemical diversity. Phytochemistry. 2014;106:7-24. doi:10.1016/j.phytochem.2014.07.012.

12. Ng TB. Pharmacological activity of sanchi ginseng (Panax notoginseng). The Journal of pharmacy and pharmacology. 2006;58(8):1007-19. doi:10.1211/jpp.58.8.0001.

13. Son HY, Han HS, Jung HW, Park YK. Panax notoginseng Attenuates the Infarct Volume in Rat Ischemic Brain and the Inflammatory Response of Microglia. Journal of pharmacological sciences. 2009;109(3):368-79.

14. Li H, Deng CQ, Chen BY, Zhang SP, Liang Y, Luo XG. Total saponins of Panax notoginseng modulate the expression of caspases and attenuate apoptosis in rats following focal cerebral ischemia-reperfusion. Journal of ethnopharmacology. 2009;121(3):412-8. doi:10.1016/j.jep.2008.10.042.

15. Yang CY, Wang J, Zhao Y, Shen L, Jiang X, Xie ZG et al. Anti-diabetic effects of Panax notoginseng saponins and its major anti-hyperglycemic components. Journal of ethnopharmacology. 2010;130(2):231-6. doi:10.1016/j.jep.2010.04.039.

16. Xiang H, Liu Y, Zhang B, Huang J, Li Y, Yang B et al. The antidepressant effects and mechanism of action of total saponins from the caudexes and leaves of Panax notoginseng in animal models of

depression. Phytomedicine : international journal of phytotherapy and phytopharmacology. 2011;18(8-9):731-8. doi:10.1016/j.phymed.2010.11.014.

17. Yang X, Xiong X, Wang H, Yang G, Wang J. Xuesaitong soft capsule (chinese patent medicine) for the treatment of unstable angina pectoris: a meta-analysis and systematic review. Evidence-based complementary and alternative medicine : eCAM. 2013;2013:948319. doi:10.1155/2013/948319.

18. Wang L, Li Z, Zhao X, Liu W, Liu Y, Yang J et al. A network study of chinese medicine xuesaitong injection to elucidate a complex mode of action with multicompound, multitarget, and multipathway. Evidence-based complementary and alternative medicine : eCAM. 2013;2013:652373. doi:10.1155/2013/652373.

19. Luo J, Song W, Yang G, Xu H, Chen K. Compound Danshen (Salvia miltiorrhiza) dripping pill for coronary heart disease: an overview of systematic reviews. The American journal of Chinese medicine. 2015;43(1):25-43. doi:10.1142/S0192415X15500020.

20. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science. 2014;345(6201):1181-4. doi:10.1126/science.1255274.

21. Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nature genetics. 2014;46(3):270-8. doi:10.1038/ng.2877.

22. Yue PY, Mak NK, Cheng YK, Leung KW, Ng TB, Fan DT et al. Pharmacogenomics and the Yin/Yang actions of ginseng: anti-tumor, angiomodulating and steroid-like activities of ginsenosides. Chinese medicine. 2007;2:6. doi:10.1186/1749-8546-2-6.

23. Sengupta S, Toh SA, Sellers LA, Skepper JN, Koolwijk P, Leung HW et al. Modulating angiogenesis: the yin and the yang in ginseng. Circulation. 2004;110(10):1219-25. doi:10.1161/01.CIR.0000140676.88412.CF.

24. Yang WZ, Bo T, Ji S, Qiao X, Guo DA, Ye M. Rapid chemical profiling of saponins in the flower buds of Panax notoginseng by integrating MCI gel column chromatography and liquid chromatography/mass spectrometry analysis. Food Chem. 2013;139(1-4):762-9. doi:10.1016/j.foodchem.2013.01.051.

25. Wan JB, Zhang QW, Hong SJ, Li P, Li SP, Wang YT. Chemical investigation of saponins in different parts of Panax notoginseng by pressurized liquid extraction and liquid chromatography-electrospray ionization-tandem mass spectrometry. Molecules. 2012;17(5):5836-53. doi:10.3390/molecules17055836.

26. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2016. doi:10.1093/bioinformatics/btw663.

27. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764-70. doi:10.1093/bioinformatics/btr011.

28. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nature genetics. 2016;48(6):657-66. doi:10.1038/ng.3565.

29. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2. doi:10.1093/bioinformatics/btv351.

30. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J et al. Sequencing of allotetraploid cotton

(Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nature biotechnology. 2015;33(5):531-7. doi:10.1038/nbt.3207.

31. Yan L, Wang X, Liu H, Tian Y, Lian J, Yang R et al. The Genome of Dendrobium officinale Illuminates the Biology of the Important Traditional Chinese Orchid Herb. Molecular plant. 2015;8(6):922-34. doi:10.1016/j.molp.2014.12.011.

32. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W et al. The genome of the cucumber, Cucumis sativus L. Nature genetics. 2009;41(12):1275-81. doi:10.1038/ng.475.

33. Rea PA. Plant ATP-binding cassette transporters. Annual review of plant biology. 2007;58:347-75. doi:10.1146/annurev.arplant.57.032905.105406.

34. Martinoia E, Klein M, Geisler M, Bovet L, Forestier C, Kolukisaoglu U et al. Multifunctionality of plant ABC transporters--more than just detoxifiers. Planta. 2002;214(3):345-55.

35. Geisler M, Murphy AS. The ABC of auxin transport: the role of p-glycoproteins in plant development. FEBS letters. 2006;580(4):1094-102. doi:10.1016/j.febslet.2005.11.054.

36. Chen S, Sanchez-Fernandez R, Lyver ER, Dancis A, Rea PA. Functional characterization of AtATM1, AtATM2, and AtATM3, a subfamily of Arabidopsis half-molecule ATP-binding cassette transporters implicated in iron homeostasis. The Journal of biological chemistry. 2007;282(29):21561-71. doi:10.1074/jbc.M702383200.

37. Kang J, Hwang JU, Lee M, Kim YY, Assmann SM, Martinoia E et al. PDR-type ABC transporter mediates cellular uptake of the phytohormone abscisic acid. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(5):2355-60. doi:10.1073/pnas.0909222107.

38. Campbell EJ, Schenk PM, Kazan K, Penninckx IA, Anderson JP, Maclean DJ et al. Pathogen-responsive expression of a putative ATP-binding cassette transporter gene conferring resistance to the diterpenoid sclareol is regulated by multiple defense signaling pathways in Arabidopsis. Plant physiology. 2003;133(3):1272-84. doi:10.1104/pp.103.024182.

39. Shibuya M, Hoshino M, Katsube Y, Hayashi H, Kushiro T, Ebizuka Y. Identification of beta-amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay. The FEBS journal. 2006;273(5):948-59. doi:10.1111/j.1742-4658.2006.05120.x.

40. Choi DW, Jung J, Ha YI, Park HW, In DS, Chung HJ et al. Analysis of transcripts in methyl jasmonate-treated ginseng hairy roots to identify genes involved in the biosynthesis of ginsenosides and other secondary metabolites. Plant cell reports. 2005;23(8):557-66. doi:10.1007/s00299-004-0845-4.

41. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic acids research. 2003;31(13):3381-5.

42. Thoma R, Schulz-Gasch T, D'Arcy B, Benz J, Aebi J, Dehmlow H et al. Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. Nature. 2004;432(7013):118-22. doi:10.1038/nature02993.

43. Nelson D, Werck-Reichhart D. A P450-centric view of plant evolution. The Plant journal : for cell and molecular biology. 2011;66(1):194-211. doi:10.1111/j.1365-313X.2011.04529.x.

44. Han JY, Hwang HS, Choi SW, Kim HJ, Choi YE. Cytochrome P450 CYP716A53v2 catalyzes the formation of protopanaxatriol from protopanaxadiol during ginsenoside biosynthesis in Panax ginseng. Plant & cell physiology. 2012;53(9):1535-45. doi:10.1093/pcp/pcs106.

45. Yonekura-Sakakibara K, Hanada K. An evolutionary view of functional diversity in family 1 glycosyltransferases. The Plant journal : for cell and molecular biology. 2011;66(1):182-93. doi:10.1111/j.1365-313X.2011.04493.x.

46. Yanagisawa S. A novel DNA-binding domain that may form a single zinc finger motif. Nucleic Acids Res. 1995;23(17):3403-10.

47. Yanagisawa S. Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize. The Plant journal : for cell and molecular biology. 2000;21(3):281-8.

48. Yanagisawa S, Sheen J. Involvement of maize Dof zinc finger proteins in tissue-specific and light-regulated gene expression. Plant Cell. 1998;10(1):75-89.

49. Vicente-Carbajosa J, Moose SP, Parsons RL, Schmidt RJ. A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2. Proc Natl Acad Sci U S A. 1997;94(14):7685-90.

50. Theodoulou FL. Plant ABC transporters. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2000;1465(1–2):79-103. doi:http://dx.doi.org/10.1016/S0005-2736(00)00132-2.

51. Yan XL, Lin LY, Liao XY, Zhang WB, Wen Y. Arsenic stabilization by zero-valent iron, bauxite residue, and zeolite at a contaminated site planting Panax notoginseng. Chemosphere. 2013;93(4):661-7. doi:10.1016/j.chemosphere.2013.05.083.

52. Freeman BC, Beattie GA. An Overview of Plant Defenses against Pathogens and Herbivores. The Plant Health Instructor.; 2008.

53. Friedman M. Tomato glycoalkaloids: role in the plant and in the diet. Journal of agricultural and food chemistry. 2002;50(21):5751-80.

54. Aerts R, Mordue AJ. Feeding Deterrence and Toxicity of Neem Triterpenoids. J Chem Ecol. 1997;23(9):2117-32. doi:10.1023/b:joec.0000006433.14030.04.

55. Palazón J, Cusidó RM, Bonfill M, Mallol A, Moyano E, Morales C et al. Elicitation of different Panax ginseng transformed root phenotypes for an improved ginsenoside production. Plant Physiology and Biochemistry. 2003;41(11–12):1019-25. doi:http://dx.doi.org/10.1016/j.plaphy.2003.09.002.

56. Bernards MA YL, Nicol RW. The allelopathic potential of ginsenosides. . In Allelochemicals: Biological Control of Plant Pathogens and Diseases. Netherlands: Springer; 2006.

57. Nicol RW, Traquair JA, MA. B. Ginsenosides as host resistance factors in American ginseng (Panax quinquefolius). Canadian Journal of Botany. 2002;80:557-62. .

58. Kushiro T, Ohno Y, Shibuya M, Ebizuka Y. In vitro conversion of 2,3-oxidosqualene into dammarenediol by Panax ginseng microsomes. Biol Pharm Bull. 1997;20(3):292-4.

59. 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 physiology. 2011;52(12):2062-73. doi:10.1093/pcp/pcr150.

60. Han JY, Kim MJ, Ban YW, Hwang HS, Choi YE. The involvement of beta-amyrin 28-oxidase (CYP716A52v2) in oleanane-type ginsenoside biosynthesis in Panax ginseng. Plant & cell physiology. 2013;54(12):2034-46. doi:10.1093/pcp/pct141.

61. Jung SC, Kim W, Park SC, Jeong J, Park MK, Lim S et al. Two ginseng UDP-glycosyltransferases synthesize ginsenoside Rg3 and Rd. Plant & cell physiology. 2014;55(12):2177-88. doi:10.1093/pcp/pcu147.

62. Wei W, Wang P, Wei Y, Liu Q, Yang C, Zhao G et al. Characterization of Panax ginseng UDP-Glycosyltransferases Catalyzing Protopanaxatriol and Biosyntheses of Bioactive Ginsenosides F1 and Rh1 in Metabolically Engineered Yeasts. Molecular plant. 2015. doi:10.1016/j.molp.2015.05.010.

63. Yan X, Fan Y, Wei W, Wang P, Liu Q, Wei Y et al. Production of bioactive ginsenoside compound K in metabolically engineered yeast. Cell Res. 2014;24(6):770-3. doi:10.1038/cr.2014.28.

64. Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J et al. 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. doi:10.1186/1471-2164-11-262.

65. Chen S, Luo H, Li Y, Sun Y, Wu Q, Niu Y et al. 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in Panax ginseng. Plant cell reports. 2011;30(9):1593-601. doi:10.1007/s00299-011-1070-6.

66. Li C, Zhu Y, Guo X, Sun C, Luo H, Song J et al. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in Panax ginseng C. A. Meyer. BMC genomics. 2013;14:245. doi:10.1186/1471-2164-14-245.

67. Liu MH, Yang BR, Cheung WF, Yang KY, Zhou HF, Kwok JS et al. Transcriptome analysis of leaves, roots and flowers of Panax notoginseng identifies genes involved in ginsenoside and alkaloid biosynthesis. BMC genomics. 2015;16:265. doi:10.1186/s12864-015-1477-5.

68. Luo H, Sun C, Sun Y, Wu Q, Li Y, Song J et al. Analysis of the transcriptome of Panax notoginseng root uncovers putative triterpene saponin-biosynthetic genes and genetic markers. BMC genomics. 2011;12 Suppl 5:S5. doi:10.1186/1471-2164-12-S5-S5.

69. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47. doi:10.1093/nar/gkv007.

70. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106.

71. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1(1):18. doi:10.1186/2047-217X-1-18.

72. English AC, Richards S, Han Y, Wang M, Vee V, Qu J et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS one. 2012;7(11):e47768. doi:10.1371/journal.pone.0047768.

73. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome research. 2009;19(6):1117-23. doi:10.1101/gr.089532.108.

74. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011;29(7):644-52. doi:10.1038/nbt.1883.

75. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 2003;19(5):651-2.

76. Vogel JP, Gu YQ, Twigg P, Lazo GR, Laudencia-Chingcuanco D, Hayden DM et al. EST sequencing and phylogenetic analysis of the model grass Brachypodium distachyon. TAG Theoretical and applied genetics Theoretische und angewandte Genetik. 2006;113(2):186-95. doi:10.1007/s00122-006-0285-3.

77. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research. 2002;12(4):656-64. doi:10.1101/gr.229202. Article published online before March 2002.

78. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966-7. doi:10.1093/bioinformatics/btp336.

79. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008;5(7):621-8. doi:10.1038/nmeth.1226.

80. Audic S, Claverie JM. The significance of digital gene expression profiles. Genome research. 1997;7(10):986-95.

81. Yoav Benjamini DY. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 2001;29(4):1165-88.

82. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research. 2005;110(1-4):462-7. doi:10.1159/000084979.

83. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research. 1999;27(2):573-80.

84. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Creating a honey bee consensus gene set. Genome biology. 2007;8(1):R13. doi:10.1186/gb-2007-8-1-r13.

85. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome research. 2004;14(5):988-95. doi:10.1101/gr.1865504.

86. Salamov AA, Solovyev VV. Ab initio gene finding in Drosophila genomic DNA. Genome research. 2000;10(4):516-22.

87. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research. 2006;34(Web Server issue):W435-9. doi:10.1093/nar/gkl200.

88. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20(16):2878-9. doi:10.1093/bioinformatics/bth315.

89. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11. doi:10.1093/bioinformatics/btp120.

90. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010;28(5):511-5. doi:10.1038/nbt.1621.

91. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research. 2000;28(1):45-8.

92. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.

93. Zdobnov EM, Apweiler R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17(9):847-8.

94. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000;25(1):25-9. doi:10.1038/75556.

95. Delorenzi M, Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics. 2002;18(4):617-25.

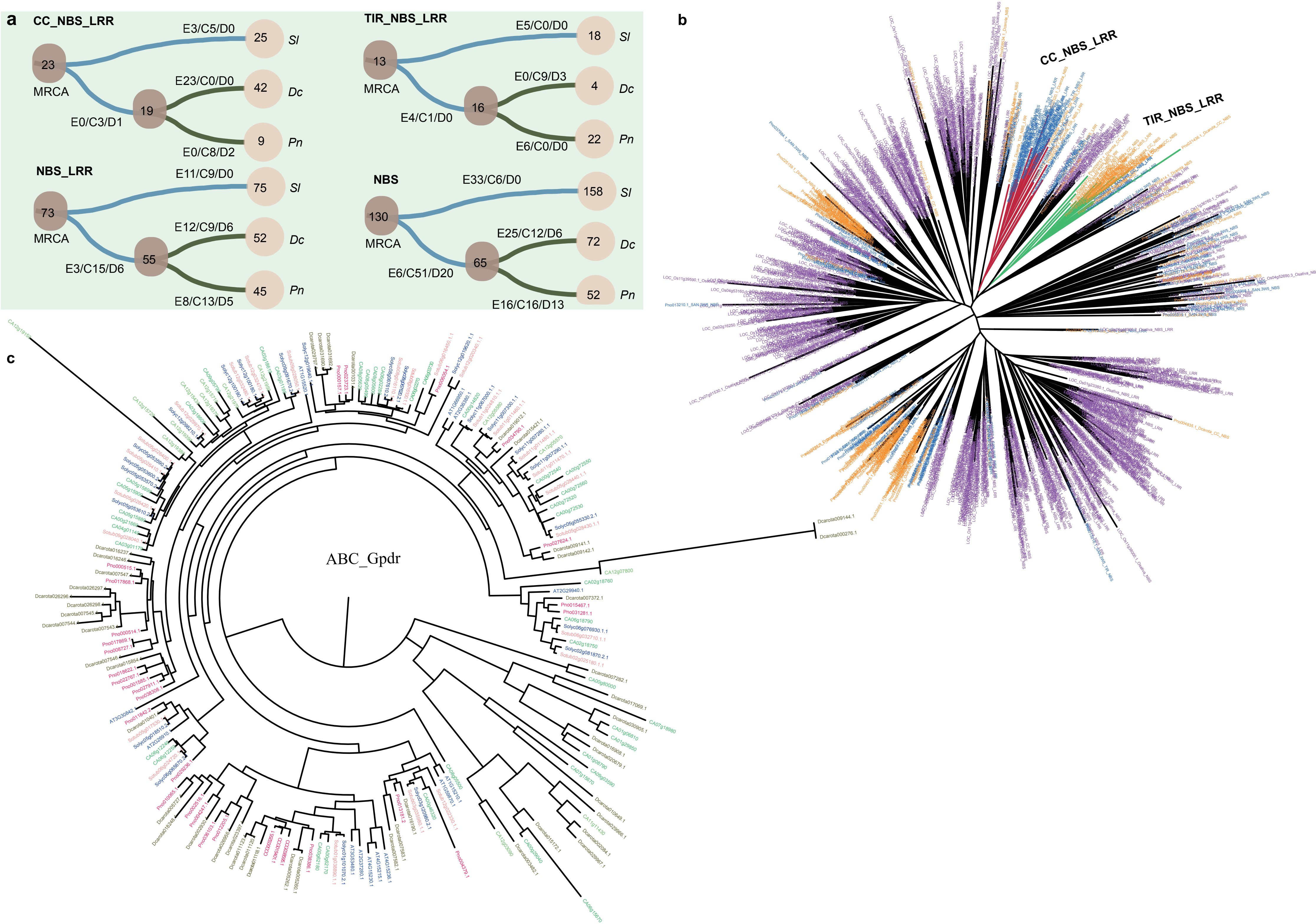
96. McDonnell AV, Jiang T, Keating AE, Berger B. Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics. 2006;22(3):356-8. doi:10.1093/bioinformatics/bti797.

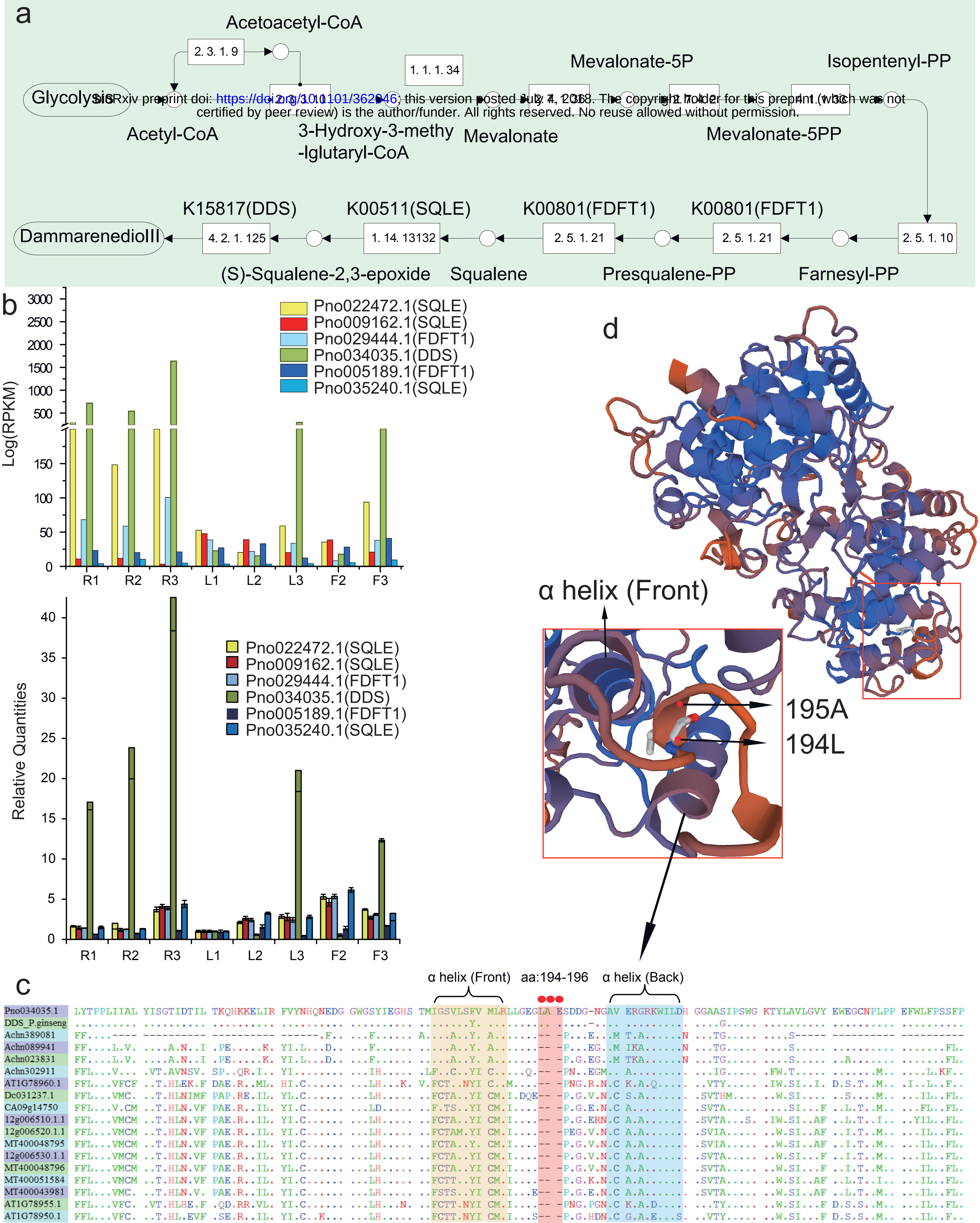
97. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research. 2003;13(9):2178-89. doi:10.1101/gr.1224503.

98. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology. 2010;59(3):307-21. doi:10.1093/sysbio/syq010.

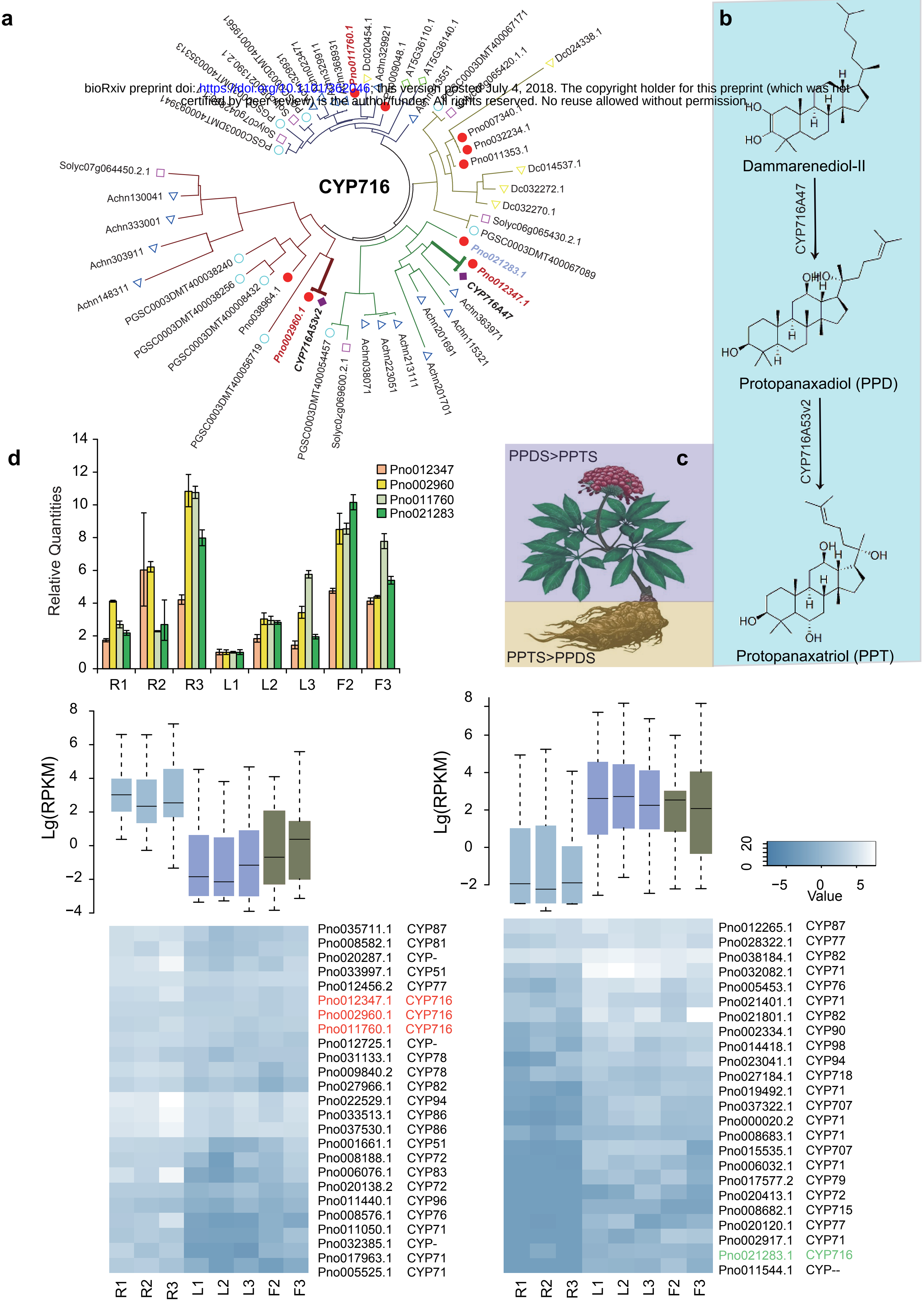
99. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution.

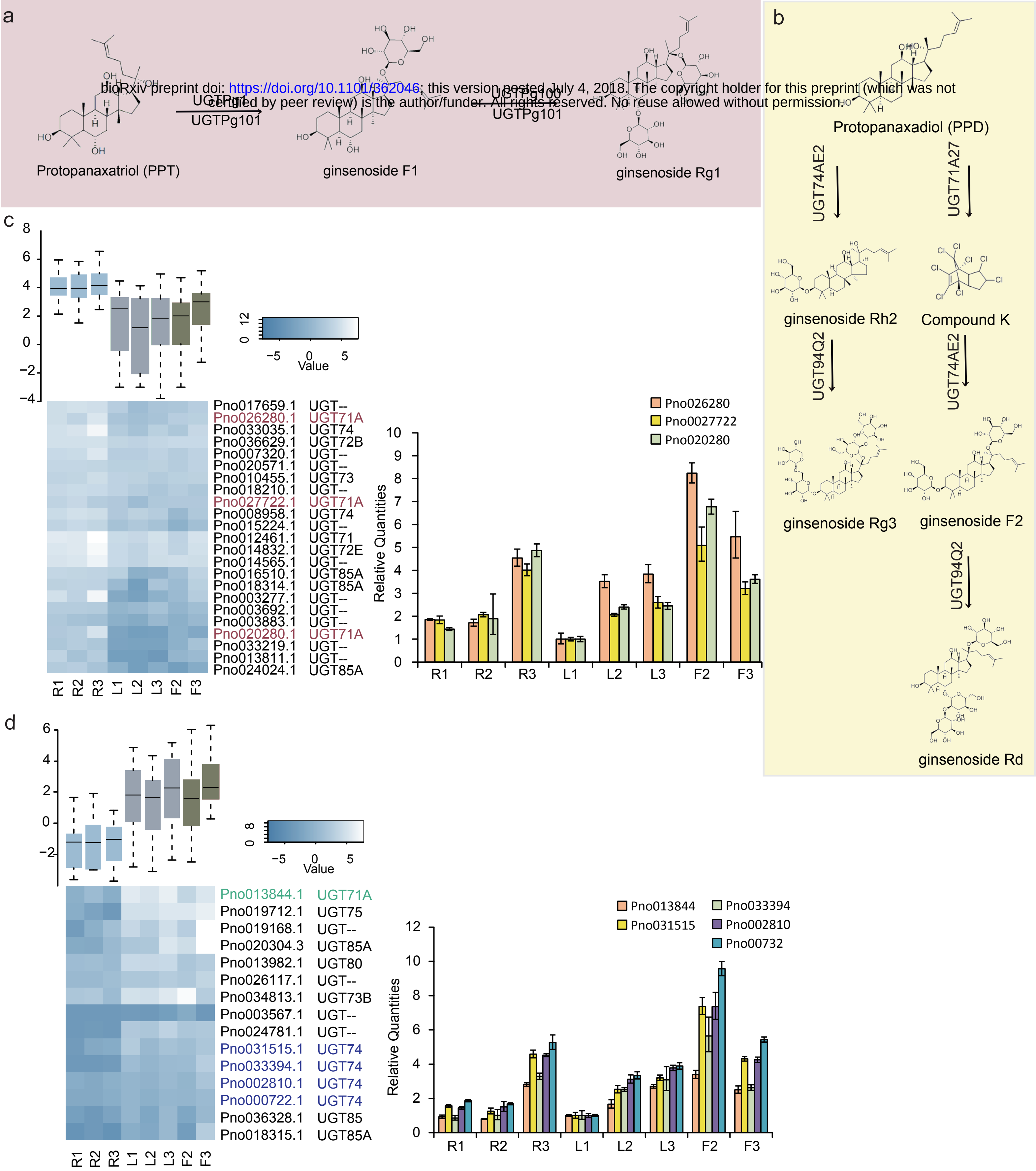
2007;24(8):1586-91. doi:10.1093/molbev/msm088.


a
CC_NBS_LRR
23
MRCA
E3/C5/D0
25
Sl
E0/C3/D1
19
E23/C0/D0
42
Dc
E0/C8/D2
9
Pn
TIR_NBS_LRR
13
MRCA
E5/C0/D0
18
Sl
E4/C1/D0
16
E0/C9/D3
4
Dc
E6/C0/D0
22
Pn
NBS_LRR
73
MRCA
E11/C9/D0
75
Sl
E3/C15/D6
55
E12/C9/D6
52
Dc
E8/C13/D5
45
Pn
NBS
130
MRCA
E33/C6/D0
158
Sl
E6/C51/D20
65
E25/C12/D6
72
Dc
E16/C16/D13
52
Pn
b
CC_NBS_LRR
TIR_NBS_LRR
c
ABC_Gpdr



```
Pno034035.1    LYTPPLIIAL YISGTIDTIL TKQHKKELIR FVYNHQNEDG GWGSYIEGHS TMIGSVLSFV MLRLLGEGLA ESDDG-NGAV ERGRKWILDH GGAASIPSWG KTYLAVLGVY EWEGCNPLPP EFWLFPSSFP
DDS_P.ginseng  .......... .......... .......... .......... .......... ........Y. .......... .....-.... .......... ....G..... .......... .......... ..........
Achn389081     FF....---- ---------- ---------- ------.F.. ...F...... A....A..Y. A.......-- -P...E...M T.A......N ...TG..... ....S----- ---------- ----------
Achn089941     FF....L.V. ....A.N..I .PE.....K. YI.L...D.. ...F...... .....A..Y. A.......-- -P...EG..M IKA......N ...TG..... ....S..... .......... .......FL.
Achn023831     FF....L.V. ....A.N..I .PE.....K. YI.L...D.. ...F...... .....A..Y. A.......-- -P.G.EG..M TKA......N ...TG..... ....S..... ..A....... .......FL.
Achn302911     FFL...V... .VT.AVNSV. SP..QR.I.. YI........ ...IH..... .LF..C..YI C.....Q.-- -PN..EGM.. A......... ...IG..... .FW.T..... .......M.. ....L.KF..
AT1G78960.1    FFL...VFCF ..T.HLEK.F DAE.R..ML. HI.C...... ...LH...K. V.FCT..NYI C..M....-- -PNG.R.N.C K.A.Q..... ..VTY..... .IW.SI..I. D.S.T..M.. .I..L..F..
Dc031237.1     FFL...VMC. ..T.HLNIMF PAP.RE.IL. YL.C...... ...LH..... ..FCTA..YI CM.I.DQE-- -P.G.V.N.C S.A....... .SVTHM.... ..W.SI.... ..S....M.. ...IL..FL.
CA09g14750     FFL...V.CM ..T.HLN.VF PAE.R..IL. YI.C...... ...LD..... ..F.TS..YI CM.I....-- -P.G.E.N.C A......... .SVTA..... ....SI...F ..L.T..M.. ...FL..FL.
12g006510.1.1  FFL...VMCM ..T.HLN.VF PAE.R..IL. YI.C...... ...LH..... ..FCTA..YI CM.I....-- -P.G.ERN.C A.A....... .SVTA...S. .IW.SI..AF ..L.T..I.. ..YIL..FL.
12g006520.1.1  FFL...VMCM ..T.HLN.VF PAE.R..IL. YI.C...... ...LH..... ..FCTAM.YI CM.I....-- -PEG.V.N.C A.A....... .SVIA..... ..W.SI..AF ..I.T..M.. ...IL..FL.
MT400048795    FFL...VMCM ..T.HLN.VF PAE.R..IL. YI.C...... ...LH..... ..FCTA..Y. CM.I....-- -P.G.V.N.C A.A....... .SVIA..... ..W.SI..AF ..I.T..M.. ...IL..FL.
12g006530.1.1  FFL...VMCM ..T.HLN.VF PAE.R..IL. YI.C...... ...LH..... ..FCTA..YI CM.I....-- -P.G.V.N.C A.A....... .SVTA..... ..W.SI...F ..I.T..M.. ...IL..FL.
MT400048796    FFL...VMCM ..T.HLN.VF PAE.R..IL. YL.C...... ...LH..... ..FCTA..YI CM.I....-- -P.G.V.N.C A.A....... .SVTA..... ..W.SI...F ..I.T..M.. ...IL..FL.
MT400051584    FFL...VMCM ..T.HLN.VF PAE.R..IL. YI.C...... ...LH..... ..FCTT..YI CM.I....-- -P.G.K.N.C A.A....... .SVTA..... ..W.SI..AF ..L.T..I.. ...IL..FL.
MT400043981    FFL...VMC. ..T.HLN.V. PAE.R..IL. YI.C...... ...LH..... ..FSTS..YI CM.I...E-- -P.G.E.N.C A.A....... .SVTA..... ..W.SI..AF ..L.T..M.. ...IL..FL.
AT1G78955.1    FFL...VFC. .VT.HLHE.F .QD.RR.VL. YI.C...... ...LH...N. ..FCTT.NYI CM.I....-- -PNG.PGN.C K.A.D..... ...TY..... ..W.SI...F D.S.S..M.. ...IL..FL.
AT1G78950.1    FFL...VFC. ..T.HL.EVF .SE.R..IL. YI.C..K... ...LH..... ..FCTT.NYI CM.I...S-- -P.G.HDN.C G.A.E...S. ..VTY..... ..W.SI...F D.S.S..M.. ...IL..F..
```


a
CYP716
Pno011760.1
Dc020454.1
Achn329921
Pno009048.1
AT5G36110.1
AT5G36140.1
PGSC0003DMT400067171
Solyc06g065420.1.1
Dc024338.1
Pno007340.1
Pno032234.1
Pno011353.1
Dc014537.1
Dc032272.1
Dc032270.1
Solyc06g065430.2.1
PGSC0003DMT400067089
Pno021283.1
Pno012347.1
CYP716A47
Achn363971
Achn115321
Achn201691
Achn201701
Achn213111
Achn223051
Achn038071
Solyc07g069600.2.1
PGSC0003DMT400054457
CYP716A53v2
Pno002960.1
PGSC0003DMT400056719
Pno038964.1
PGSC0003DMT400008432
PGSC0003DMT400038256
PGSC0003DMT400038240
Achn148311
Achn303911
Achn333001
Achn130041
Solyc07g064450.2.1
PGSC0003DMT400039941
PGSC0003DMT400035313
Solyc02g021390.2.1
PGSC0003DMT400019561
b
HO
HO
H
H
H
Dammarenediol-II
CYP716A47
OH
HO
H
H
H
Protopanaxadiol (PPD)
CYP716A53v2
OH
OH
HO
OH
H
H
H
Protopanaxatriol (PPT)
c
PPDS>PPTS
PPTS>PPDS
d
Relative Quantities
14
12
10
8
6
4
2
0
Pno012347
Pno002960
Pno011760
Pno021283
R1
R2
R3
L1
L2
L3
F2
F3
Lg(RPKM)
8
6
4
2
0
−2
−4
20
0
−5
0
5
Value
Pno035711.1 CYP87
Pno008582.1 CYP81
Pno020287.1 CYP-
Pno033997.1 CYP51
Pno012456.2 CYP77
Pno012347.1 CYP716
Pno002960.1 CYP716
Pno011760.1 CYP716
Pno012725.1 CYP-
Pno031133.1 CYP78
Pno009840.2 CYP78
Pno027966.1 CYP82
Pno022529.1 CYP94
Pno033513.1 CYP86
Pno037530.1 CYP86
Pno001661.1 CYP51
Pno008188.1 CYP72
Pno006076.1 CYP83
Pno020138.2 CYP72
Pno011440.1 CYP96
Pno008576.1 CYP76
Pno011050.1 CYP71
Pno032385.1 CYP-
Pno017963.1 CYP71
Pno005525.1 CYP71
Pno012265.1 CYP87
Pno028322.1 CYP77
Pno038184.1 CYP82
Pno032082.1 CYP71
Pno005453.1 CYP76
Pno021401.1 CYP71
Pno021801.1 CYP82
Pno002334.1 CYP90
Pno014418.1 CYP98
Pno023041.1 CYP94
Pno027184.1 CYP718
Pno019492.1 CYP71
Pno037322.1 CYP707
Pno000020.2 CYP71
Pno008683.1 CYP71
Pno015535.1 CYP707
Pno006032.1 CYP71
Pno017577.2 CYP79
Pno020413.1 CYP72
Pno008682.1 CYP715
Pno020120.1 CYP77
Pno002917.1 CYP71
Pno021283.1 CYP716
Pno011544.1 CYP--



a
Protopanaxatriol (PPT)
UGTPg1
UGTPg101
ginsenoside F1
UGTPg100
UGTPg101
ginsenoside Rg1
b
Protopanaxadiol (PPD)
UGT74AE2
UGT71A27
ginsenoside Rh2
Compound K
UGT94Q2
UGT74AE2
ginsenoside Rg3
ginsenoside F2
UGT94Q2
ginsenoside Rd
c
Value
Pno017659.1 UGT--
Pno026280.1 UGT71A
Pno033035.1 UGT74
Pno036629.1 UGT72B
Pno007320.1 UGT--
Pno020571.1 UGT--
Pno010455.1 UGT73
Pno018210.1 UGT--
Pno027722.1 UGT71A
Pno008958.1 UGT74
Pno015224.1 UGT--
Pno012461.1 UGT71
Pno014832.1 UGT72E
Pno014565.1 UGT--
Pno016510.1 UGT85A
Pno018314.1 UGT85A
Pno003277.1 UGT--
Pno003692.1 UGT--
Pno003883.1 UGT--
Pno020280.1 UGT71A
Pno033219.1 UGT--
Pno013811.1 UGT--
Pno024024.1 UGT85A
R1 R2 R3 L1 L2 L3 F2 F3
Pno026280
Pno0027722
Pno020280
Relative Quantities
R1 R2 R3 L1 L2 L3 F2 F3
d
Value
Pno013844.1 UGT71A
Pno019712.1 UGT75
Pno019168.1 UGT--
Pno020304.3 UGT85A
Pno013982.1 UGT80
Pno026117.1 UGT--
Pno034813.1 UGT73B
Pno003567.1 UGT--
Pno024781.1 UGT--
Pno031515.1 UGT74
Pno033394.1 UGT74
Pno002810.1 UGT74
Pno000722.1 UGT74
Pno036328.1 UGT85
Pno018315.1 UGT85A
R1 R2 R3 L1 L2 L3 F2 F3
Pno013844
Pno031515
Pno033394
Pno002810
Pno00732
Relative Quantities
R1 R2 R3 L1 L2 L3 F2 F3