Full-length transcriptome atlas of *Panax vietnamensis* var. *fuscidiscus* reveals novel genes and alternative splicing in tissue-specific biosynthesis of ocotillol-type saponins

Wei Li[a,b], Zhe Xu[a,b], Aneela Nijabat[a,b],and Li-zhi Gao[a,b*]

[a] Engineering Research Center for Selecting and Breeding New Tropical Crop Varieties, Ministry of Education, Tropical Biodiversity and Genomics Research Center, College of Tropical Agriculture and Forestry, Hainan University, Haikou 570228, China

[b] Sanya Institute of Breeding and Multiplication, Hainan University, Sanya 572025, China

[*] Corresponding author.
E-mail address: Lgaogenomics@163.com.

Abstract

Background *Panax vietnamensis* var. *fuscidiscus* is the only known *Panax* species that accumulates exceptionally high levels of ocotillol (OCT)-type ginsenosides, particularly majonoside R2. However, the molecular mechanisms underlying OCT-type ginsenoside biosynthesis remain largely uncharacterized, due in part to the absence of comprehensive full-length transcriptome data and complete gene annotations.

Results We combined PacBio single-molecule real-time (SMRT) and Illumina RNA sequencing (RNA-Seq) technologies to construct the first tissue-specific, full-length transcriptome atlas of *P. vietnamensis* var. *fuscidiscus*. A total of 281,468 non-redundant transcripts were assembled, including 8,089 novel genes not present in existing annotations, substantially expanding the documented gene repertoire of this species. We identified 21 candidate genes putatively involved in triterpenoid saponin biosynthesis and uncovered distinct tissue-specific expression patterns across the oleanane (OA)-, protopanaxadiol (PPD)-, protopanaxatriol (PPT)-, and OCT-type biosynthetic pathways. Notably, we detected 46,917 alternative splicing (AS) events, with over 45% of expressed genes producing multiple isoforms. Several key biosynthetic genes—including dammarenediol synthase (*DDS*), cytochrome P450 716A53v2 (*CYP716A53v2*), and mevalonate diphosphate decarboxylase (*MVD*)—exhibited tissue-specific splicing patterns, such as retained introns and exon skipping, which may contribute to functional diversification of isoforms involved in

OCT-type ginsenoside biosynthesis. These results highlight the importance of post-transcriptional regulation and transcript diversity in the evolution and fine-tuning of specialized metabolic pathways.

Conclusions This study establishes a high-quality, full-length transcriptomic resource for *P. vietnamensis* var. *fuscidiscus*, revealing previously undescribed genes and extensive alternative splicing events related to ginsenoside biosynthesis. Our findings provide novel molecular insights into the mechanisms regulating OCT-type saponin accumulation and offer a foundation for future research into the functional characterization, regulatory mechanisms, and synthetic biological production of rare triterpenoid saponins.

Keywords *Panax vietnamensis*, alternative splicing, PacBio Iso-Seq, triterpenoid biosynthesis, transcriptome

Background

Medicinal plants play a crucial role in promoting human health. In developing countries, herbal medicines serve as the primary form of healthcare for approximately 75% of the population. Known as the "King of Herbs", ginseng (*Panax* spp.) has been used in traditional Chinese medicine for over 5,000 years and has gained increasing acceptance in Western medicine in recent decades. The main bioactive compounds in ginseng are ginsenosides, a group of triterpenoid saponins that are predominantly found in *Panax* species, with the highest concentrations typically located in the roots and rhizomes [1, 2]. To date, more than 300 natural ginsenosides have been identified across *Panax* species [3–5]. Based on their aglycone backbone structures, ginsenosides are categorized into two major types: dammarane-type and oleanane (OA)-type. The dammarane-type can be further divided into three subclasses: protopanaxadiol (PPD)-, protopanaxatriol (PPT)-, and ocotillol (OCT)-type ginsenosides [3].

In *Panax*, OCT-type ginsenosides represent a specialized subclass of dammarane-type ginsenosides. Oxidosqualene cyclases (*OSCs*), which initiate dammarane-type ginsenoside biosynthesis, have been shown to generate OCT-related precursors through functional diversification following gene duplication. However, the downstream biosynthetic pathways and regulatory mechanisms governing OCT-type ginsenoside accumulation remain poorly understood [6, 7].

Panax vietnamensis Ha et Grushv., a new *Panax* species, was first discovered in the central highlands of Vietnam in 1973 and is the southernmost distributed species

77 within the genus [8]. The roots and rhizomes of this species have long been secretly
78 used by the Sedang ethnic group in Vietnam as a life-saving medicinal plant, and it
79 holds significant ethnomedical value [9–11]. Notably, *P. vietnamensis* is the only
80 known *Panax* species that accumulates high levels of OCT-type ginsenosides [12].
81 Among them, the representative compound majonoside R2 accounts for over 5% of
82 the dry weight of the rhizome and approximately half of the total ginsenosides,
83 making *P. vietnamensis* an important resource for studying OCT-type ginsenosides
84 [13–15].

85 Interestingly, a new variety of *P. vietnamensis*, named *P. vietnamensis* var.
86 *fuscidiscus* K. Komatsu, S. Zhu & S.Q. Cai (commonly known as "Ye Sanqi" in
87 Chinese), was discovered in Jinping County, Southern Yunnan Province, China [16].
88 However, subsequent studies suggest that its taxonomic classification as a variety of *P.*
89 *vietnamensis* requires further investigation [17]. This variety contains significantly
90 higher levels of majonoside R2 compared to other genotypes, making it an ideal
91 model for studying the biosynthetic mechanisms of OCT-type triterpene saponins [18,
92 19]. Although recent studies have identified two UDP-glycosyltransferases (*PvfUGT1*
93 and *PvfUGT2*) responsible for the sequential glycosylation steps leading to
94 majonoside R2 formation, the upstream regulatory architecture of this pathway
95 remains largely unresolved. OCT-type saponins were first identified as minor
96 constituents from the underground parts of *P. japonicus* var. *major* [20].
97 Pharmacological studies have since demonstrated their multiple bioactivities,
98 including anti-tumor effects [21], anti-stress properties [22], and hepatoprotective

99 functions [23, 24].

100 Although substantial research has been focused on the diversity, content,
101 metabolic pathways, and pharmacological activities of ginsenosides in widely used
102 species, such as *P. ginseng*, *P. quinquefolius*, and *P. notoginseng* [3, 25–28], studies on
103 the biosynthetic pathways of OCT-type ginsenosides in *P. vietnamensis* remain limited.
104 Recent studies have applied the Illumina sequencing platform to *P. vietnamensis*,
105 including the release of draft genome assembly [19, 29], tissue-specific transcriptome
106 analyses, and mining of genes involved in ginsenoside biosynthesis [19, 30, 31].
107 However, the limitations of short-read sequencing, such as its inability to fully capture
108 complex splicing isoforms and full-length transcripts, have hindered the
109 comprehensive understanding of the molecular mechanisms underlying OCT-type
110 ginsenoside biosynthesis in *P. vietnamensis*.

111 Single-molecule real-time (SMRT) sequencing, a third-generation sequencing
112 technology, enables the direct acquisition of full-length cDNA transcripts and
113 overcomes the shortcomings of short-read sequencing in transcript assembly and
114 isoform identification. In recent years, SMRT sequencing has been successfully
115 applied in various plant research areas, including gene structure analysis, alternative
116 splicing (AS), transcription factor identification, and simple sequence repeat (SSR)
117 discovery [32–35]. In this study, we report the first integration of SMRT and
118 second-generation sequencing (SGS) technologies to generate a high-quality,
119 full-length, tissue-resolved transcriptome of *P. vietnamensis* var. *fuscidiscus*. By
120 combining PacBio Isoform Sequencing (Iso-Seq) for isoform-level transcript

121 discovery and Illumina RNA sequencing (RNA-Seq) for expression quantification and
122 annotation, we constructed a comprehensive transcriptomic atlas across the five major
123 tissues (root, rhizome, stem, leaf, and flower). This integrative approach enabled the
124 identification of 8,089 novel genes and over 46,000 alternative splicing events,
125 significantly improving the resolution of gene structure and regulatory complexity.
126 Importantly, we identified key candidate genes and splicing isoforms potentially
127 involved in the biosynthesis of OCT-type triterpenoid saponins, offering valuable
128 resources for functional genomics and metabolic engineering in *Panax* species.
129

130 # Methods

131 **Plant materials**

132 Three-year-old *P. vietnamensis* var. *fuscidiscus* plants were collected in October 2015
133 from the cultivation base of Jinping Shili Medicinal Material Development Co., Ltd.
134 in Jingping County, Yunnan Province, China (22° 44′ 24″ N, 103° 13′ 48″ E;
135 Altitude: 1750 m). The plants were grown under the forest canopy in humus-rich soil,
136 with approximately 25 - 30 cm spacing between individuals, and maintained under
137 natural temperature and light conditions typical of the local environment. During the
138 sampling period, the average temperature was 16 - 24 ℃ with a photoperiod of about
139 12 hours of daylight. Root, stem, leaf, flower, and rhizome tissues were cut into small
140 pieces, immediately frozen in liquid nitrogen, and stored at −80 ℃ until RNA
141 extraction.

142

143 **Construction of Illumina RNA-Seq library**

144 Total RNA was extracted from frozen root, stem, leaf, flower, and rhizome tissues
145 using the Plant RNA Extraction Kit (Zoonbio Biotechnology, Cat. No. RK16-50T)
146 following the manufacturer's instructions. The integrity, quality, and concentration of
147 the extracted RNA were assessed using the Agilent 2100 Bioanalyzer system (Agilent
148 Technologies, Santa Clara, CA, USA). High-quality RNA samples were used to
149 construct cDNA libraries with the Illumina mRNA-Seq Sample Preparation Kit
150 (Illumina, San Diego, CA, USA). Sequencing was performed on the Illumina HiSeq
151 2000 platform using standard protocols.

152 Raw reads were processed with Trimmomatic (v0.39) to remove adapter
153 sequences, trim low-quality bases (Phred score < 15), and cut the first 15 bases from
154 each read (HEADCROP:15). Cleaned reads were *de novo* assembled using Trinity
155 (v2.15.2), and redundant sequences were removed using CD-HIT (v4.8.1) [36].
156

157 **cDNA library preparation and PacBio SMRT sequencing**

158 Total RNA was extracted using TRIzol Reagent (Invitrogen, Beijing, China; Cat. No.
159 15596-026) according to the manufacturer's instructions. RNA integrity was
160 evaluated with an Agilent 2200 TapeStation (Agilent Technologies, USA). Full-length
161 cDNA libraries were prepared following the PacBio Isoform Sequencing (Iso-Seq)
162 protocol, using the Clontech SMARTer PCR cDNA Synthesis Kit (Cat. No. 634925)
163 and the BluePippin Size Selection System (Sage Science, USA), with minor
164 modifications. Equal amounts of high-quality total RNA (10 μg, RNA Integrity
165 Number > 8.5) were pooled from four tissues (root, stem, leaf, and flower). The
166 cDNA was size-selected into three fractions: 1-2 kb, 2-3 kb, and 3-6 kb using
167 BluePippin. Each size fraction (500 ng) was used to construct SMRTbell libraries
168 with the PacBio Template Prep Kit 1.0 (Part No. 100-259-100). A total of 13 SMRT
169 cells were sequenced on the PacBio RS II platform: five cells for both 1–2 kb and 2–3
170 kb libraries, and three cells for the 3-6 kb library. Sequencing was performed using
171 the P4-C2 chemistry (BINDINGKIT=100356300; SEQUENCINGKIT=100356200)
172 with 120-minute movies on the PacBio RS II system. An additional eight SMRT cells
173 were run: three for each of the 1-2 kb and 2-3 kb libraries, and two for the 3-6 kb

174 library. Raw reads were processed using the SMRT Portal analysis suite (PacBio) to
175 generate subreads for downstream analysis.
176

177 **Data analysis of PacBio SMRT long reads**

178 PacBio single-molecule long reads were analyzed using RS_IsoSeq (v2.3) and the
179 smrtanalysis_2.3.0.140936.p4.150482 pipeline. The script pbtranscript.py was used
180 for the characterization of full-length (FL) reads, with 5′/3′ primers and poly(A) tails
181 employed to discriminate strand-specific full-length reads.

182 Circular consensus sequencing (CCS) was executed by SMRT Link, and
183 high-quality, FL, and polished consensus transcripts were obtained using the
184 isoform-level clustering algorithm ICE (Iterative Clustering for Error Correction). FL
185 non-chimeric (FLNC) reads were identified, and high-quality (HQ) and low-quality
186 (LQ) consensus isoforms were generated after clustering. LQ isoforms were further
187 corrected with cleaned SGS reads using LoRDEC (v0.9) [37]. HQ and corrected LQ
188 isoforms were collapsed by CD-HIT (v4.8.1) to form the final PacBio isoforms.
189

190 **Merging RNA-Seq assembly and Iso-Seq isoforms**

191 The assembled transcripts from second-generation sequencing (SGS) data and
192 collapsed PacBio isoforms were combined. CD-HIT (v4.8.1) was then used to cluster
193 the transcripts and to remove redundancies. The merged transcript set was classified
194 into coding and non-coding RNAs using RNASamba (v0.2.4, classify mode) with a
195 pre-trained full-length model (full_length_weights.hdf5). Coding sequences were

196 exported as predicted proteins in FASTA format. This non-redundant, merged
197 transcript set was subsequently used as the reference for downstream gene expression
198 analysis.
199

200 **Gene functional annotation and expression analysis**

201 Gene functions were annotated using eggNOG-mapper (v2.1.12) with DIAMOND
202 mode [38], based on a locally installed eggNOG database. The annotation output was
203 processed with the emapperx.R script provided in the emcp package
204 (http://git.genek.cn:3333/zhxd2/emcp), which generated Gene Ontology (GO)
205 assignments, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway
206 annotations, and a species-specific OrgDB database for downstream enrichment
207 analyses. Expression quantification across tissues was performed using RSEM (v1.3.1)
208 with SGS reads [39].
209

210 **KEGG functional classification of expressed genes**

211 To characterize the functional profiles of different tissues, KEGG pathway
212 classification was performed on genes that were expressed (TPM > 2) in at least one
213 of the five tissues. KEGG annotation was conducted using TBtools-II (v2.376) with
214 the Plants.20211128.TBtoolsKeggBackEnd database to ensure plant-specific pathway
215 assignment [40]. KO annotations from eggNOG-mapper were mapped to KEGG
216 Level 1 and Level 2 categories through the plant-specific backend, and non-plant
217 categories were automatically excluded. The resulting KEGG functional distribution

218 was used to compare pathway composition among tissues.

219

220 **Differential expression analysis and KEGG enrichment analysis**

221 Differential expression analysis between tissues was performed using edgeR [41],
222 implemented through the Trinity utility script run_DE_analysis.pl with a fixed
223 dispersion value of 0.1. Genes with $|\log_2 FC| \geqslant 1$ and false discovery rate (FDR) $\leqslant$
224 0.05 were defined as differentially expressed genes (DEGs), while others were
225 considered not significant. As only one sample per tissue was sequenced, no
226 biological replicates were included.

227 KEGG enrichment analysis for DEGs was performed using the enricher () function
228 in the clusterProfiler package with the customized OrgDB database generated by
229 emcp, ensuring accurate mapping to plant KEGG orthology. For enrichment analysis,
230 the DEGs list was used as the input gene set, and all annotated genes were supplied as
231 the background (universe). Significance thresholds were set to pvalueCutoff = 0.05
232 and qvalueCutoff = 0.05, and enrichment significance was further adjusted using the
233 Benjamini–Hochberg method.

234

235 **Full-length transcript reconstruction and isoform identification**

236 Full-length transcript sequences were aligned to the *P. vietnamensis* var. *Fuscidiscus*
237 reference genome (https://db.cngb.org/data_resources/project/CNP0002878/) [17]
238 using minimap2 (v2.28-r1209) with spliced alignment parameters (-ax splice -uf -k14
239 --secondary=no) [42]. Resulting SAM files were converted to sorted and indexed

240 BAM format using SAMtools (v1.21) [43]. Redundant isoforms were collapsed using
241 Cupcake's collapse_isoforms_by_sam.py
242 (https://github.com/Magdoll/cDNA_Cupcake), generating high-confidence
243 non-redundant transcript loci.

244 The collapsed isoforms were processed with Cogent to cluster transcripts into
245 gene families and reconstruct representative gene models. Specifically, run_mash.py
246 computed pairwise k-mer distances (k=30), and process_kmer_to_graph.py
247 transformed these distances into similarity graphs. Connected components within the
248 graph were reconstructed using reconstruct_contig.py.

249 To evaluate isoform structure and classify transcript types, SQANTI3 [35] was
250 employed with the reference genome and GTF annotation as input. SGS reads were
251 used to support splice junction validation and saturation analysis. Low-confidence
252 isoforms were removed based on SQANTI3's filtering rules, resulting in a
253 high-confidence, structurally annotated transcriptome.

254

255 **Identification of alternative splicing events**

256 Alternative splicing (AS) events were identified using SUPPA2 [44], based on the
257 SQANTI3-corrected transcript annotations in GTF format. The generateEvents
258 function was used to detect and categorize seven major AS types: skipped exon (SE),
259 retained intron (RI), mutually exclusive exons (MX), alternative 3′ splice site (A3),
260 alternative 5′ splice site (A5), alternative first exon (AF), and alternative last exon
261 (AL).

262

263 **Identification of saponin biosynthesis-related genes**

264 Candidate genes potentially involved in saponin biosynthesis were identified through
265 homology-based searches using BLASTP (v2.16.0). A local protein database was first
266 constructed from the full set of predicted transcript-derived protein sequences using
267 the makeblastdb tool. Fifty-six known saponin biosynthetic protein sequences were
268 used as queries against this database with an E-value threshold of 1e-5.

269 Putative oxidosqualene cyclases (*OSCs*) were further identified using HMMER
270 (v3.3.2) [45]. Protein sequences were scanned against the squalene-hopene cyclase
271 N-terminal (PF13249) and C-terminal (PF13243) domain models obtained from the
272 Pfam database (v37.0) [46, 47] with the trusted cutoff parameter (-cut_tc). Only
273 sequences containing both domains were considered high-confidence *OSCs* and
274 retained for downstream analysis.

275 Putative UDP-glycosyltransferases (*UGTs*) were identified using HMMER (v3.3.2).
276 All predicted protein sequences were scanned against the PSPG-domain HMM model
277 (PF00201) downloaded from the Pfam database (v37.0) using the trusted cutoff
278 parameter (-cut_tc). Hits were further filtered by significance (E-value < 1×10^{-20}) and
279 domain length (350-600 aa) to remove truncated or non-canonical sequences. To
280 ensure glycosyltransferase specificity, candidate proteins were additionally screened
281 for the conserved PSPG-box motif ($HCGWNSX_{0-3}$) using regular-expression
282 matching. Expression data (TPM) from the five tissues were then incorporated, and
283 only genes with TPM > 2 in at least one tissue were retained as high-confidence *UGTs*.

The final *UGTs* set was subjected to multiple sequence alignment using MAFFT (-auto) and phylogenetic reconstruction using IQ-TREE (-m MFP -bb 1000) for downstream classification.Visualization of the phylogenetic tree and gene expression heatmap was performed using tvBOT (v2.6.1) [48].

Statistical analysis

All statistical analyses and visualizations were performed in R (v4.5.0) using the packages tidyverse [49], ggplot2 [50], EnhancedVolcano [51], pheatmap [52], and clusterProfiler [53].

294 # Results

295 **Full-length transcriptome sequencing and integrated assembly**

296 To construct a comprehensive transcriptome of *P. vietnamensis* var. *fuscidicus*, we
297 employed a hybrid sequencing strategy integrating third-generation PacBio
298 single-molecule real-time (SMRT) sequencing and second-generation sequencing
299 (SGS). For SMRT sequencing, high-quality total RNA was extracted from root, stem,
300 leaf, and flower tissues of three-year-old plants at the flowering stage. Equal amounts
301 of RNA from each tissue were pooled to generate three size-fractionated cDNA
302 libraries (1–2 kb, 2–3 kb, and 3–6 kb). These libraries were sequenced on the PacBio
303 RS II platform using 13 SMRT cells, producing 619,574 reads of insert (ROIs),
304 including 310,550 full-length non-chimeric (FLNC) reads identified by the presence
305 of both 5′ and 3′ primers and a poly(A) tail (**Table S1**). The length distribution of
306 FLNC reads corresponded well with the expected size ranges of the respective
307 libraries (**Fig. 1A**, **1B**), confirming effective size selection. An additional 282,547
308 non-full-length reads were also obtained. After iterative clustering and error
309 correction (ICE) and LoRDEC polishing with Illumina reads, 109,990
310 high-confidence isoforms were retained, with lengths ranging from 180 to 19,929 bp
311 (**Table S1**; **Fig. 1A**).

312 For SGS, total RNA from five tissues (root, stem, leaf, flower, and rhizome) was
313 sequenced individually on the Illumina platform, yielding 97,206,112 cleaned
314 reads. *De novo* assembly using Trinity produced 195,406 contigs with a total length of

315 approximately 197 Mb. Redundancy reduction with CD-HIT resulted in a final set of
316 191,832 unigenes (**Table S1**).

317 Merging PacBio isoforms and Illumina-assembled unigenes, followed by
318 clustering with CD-HIT, yielded a final reference transcriptome comprising 281,468
319 non-redundant transcripts with a total length of ~410 Mb, a GC content of 40.06%,
320 and an N50 of 2,471 bp (**Table S1**). The merged transcriptome showed a broader and
321 more uniform transcript length distribution than the individual assemblies, effectively
322 balancing the short contigs from Illumina and the long isoforms from PacBio (**Fig.**
323 **1C**).

324

325 **Reconstruction and annotation of transcripts**

326 Stringent filtering with SQANTI3 yielded a high-confidence reference transcriptome
327 comprising 70,835 isoforms derived from 26,258 genes. Notably, 8,089 genes were
328 newly identified and absent from the reference annotation, suggesting that Iso-Seq is
329 effective in facilitating the detection of previously unannotated genomic loci.

330 Analysis of isoform counts showed that 52.0% of genes produced a single
331 isoform, while 48.0% generated two or more isoforms (ranging up to 16), indicating a
332 high prevalence of alternative splicing (AS) and a complex transcriptional landscape
333 in *P. vietnamensis* (**Fig. 2A**). SQANTI3 identified 120,846 splice junctions from the
334 Iso-Seq data. Among these, 67.4% (81,448) were previously annotated as canonical
335 (GT–AG) junctions, 32.5% (39,274) were novel canonical junctions, and only 0.1%

336 (124) were novel non-canonical junctions (**Fig. 2B**). Isoform classification revealed
337 that 46.43% were novel transcripts, including 9.39% from known genes (Novel In
338 Catalog, NIC) and 37.04% from unannotated loci (Novel Not in Catalog, NNC),
339 underscoring the substantial novelty captured by Iso-Seq. Known isoforms accounted
340 for 36.78% of the dataset and fell into the full splice match (FSM) and incomplete
341 splice match (ISM) categories, while the remaining transcripts were classified as
342 non-canonical types, including antisense, fusion, and intergenic transcripts (**Fig. 2C**).

343 Further analysis showed that most isoforms were 1–3 kb, with FSM and ISM
344 generally shorter. NIC and NNC isoforms showed a higher proportion of long
345 transcripts, as indicated by the extended upper tails in transcript length distribution
346 (Fig. 2E). Although their exon number distributions largely overlapped with those of
347 FSM and ISM (Fig. 2F), NIC and NNC still included numerous multi-exon isoforms,
348 suggesting a tendency toward increased structural complexity. This highlights
349 long-read sequencing's ability to capture structurally complex isoforms beyond the
350 predominant size range (Fig. 2D–F). Transcript boundary analysis revealed that FSM
351 isoforms had 3′ ends that closely matched annotated transcription termination sites
352 (TTS), supporting a high level of accuracy in termination detection. In comparison,
353 their 5′ ends often extended upstream of annotated transcription start sites (TSS),
354 suggesting that Iso-Seq may provide more complete transcript boundary information
355 (**Fig. 2G**, **2H**).

356 Together, these results highlight the value of long-read Iso-Seq in capturing
357 full-length transcripts, discovering novel isoforms, and elucidating the complex
358 transcriptomic architecture of *P. vietnamensis*.

359

360 **Functional characterization and tissue-specific transcript profiles**

361 A total of 131,784 transcripts (80.55%) were annotated in EggNOG or COG/KOG
362 databases, 84,826 (51.85%) were assigned Gene Ontology (GO) terms, and 52,683
363 (32.20%) were mapped to KEGG pathways (**Table S2**). GO classification indicated
364 comparable functional distributions across tissues, with the rhizome and flower
365 containing 34,621 and 38,262 annotated transcripts, respectively (**Fig. 3A**). Roots and
366 leaves exhibited similar GO profiles, suggesting a conserved functional repertoire
367 (**Fig. S1**).

368 KEGG analysis showed that genetic information processing was the predominant
369 functional category across all tissues. In contrast, the relative contribution of
370 metabolism-related pathways varied among tissues, accounting for 24.27% in flower,
371 17.91% in leaf, 12.45% in root, 9.53% in stem, and 6.06% in rhizome (**Fig. 3B; Fig.**
372 **S1**). Among metabolism-related subclasses, energy metabolism pathways were
373 detected in flower (690 counts, 4.13%), leaf (640 counts, 4.34%), and stem (605
374 counts, 4.22%), but were not identified in root or rhizome. Together, these results
375 suggest that while major functional classes are shared across tissues, the relative
376 representation of metabolism-related pathways differs in a tissue-associated manner.

377

378 **Alternative splicing and implication in triterpene biosynthesis**

379 A total of 56,019 expressed genes were identified, among which 45.68% produced
380 two or more isoforms, indicating extensive alternative splicing (**Fig. 4A**). We detected
381 46,917 alternative splicing (AS) events, with retained intron (RI) being the most
382 common type (44.51%), followed by alternative 3′ splice sites (A3, 19.61%),
383 alternative 5′ splice sites (A5, 13.77%), and skipped exon (SE, 9.96%) (**Fig. 4B**, **4C**).
384 Notably, key genes involved in triterpene biosynthesis exhibited considerable isoform
385 diversity, suggesting that alternative splicing may contribute to the regulation of
386 enzymatic activity. For example, *CYP716A53v2* displayed retained intron (RI) events
387 (**Fig. 4D**), which could produce truncated or non-functional proteins and thereby
388 modulate metabolic flux. Meanwhile, *dammarenediol-II synthase* (*DDS*) and
389 *mevalonate diphosphate decarboxylase* (*MVD*) showed exon-skipping (ES) patterns
390 (**Fig. S2A, B**), potentially generating isoforms with altered catalytic domains, which
391 may be associated with tissue-dependent differences in triterpene biosynthesis. These
392 observations suggest that alternative splicing may play a regulatory role in fine-tuning
393 triterpene production across different tissues.

394 Further validation using RNA-Seq data across different tissues revealed
395 substantial differential AS events ($|\Delta PSI| > 0.3$) when comparing flower, leaf, root,
396 and stem to rhizome. Specifically, 11,470, 11,323, 11,250, and 11,172 significant AS
397 events were detected in flower, leaf, root, and stem compared to rhizome, respectively,
398 each involving over 5,500 genes (**Fig. S2C**). RI and SE remained the predominant AS

399 types across all comparisons (**Fig. S2D**), suggesting their potential regulatory roles in
400 tissue-specific metabolic processes.
401
402 **Tissue-specific expression of genes involved in triterpene biosynthesis**
403 To investigate tissue-specific gene expression, we analyzed transcriptomes of 281,468
404 unigenes and identified 14,427 differentially expressed genes (DEGs) across four
405 pairwise comparisons with rhizome. The leaf *vs.* rhizome comparison had the highest
406 number of DEGs (6,696), including 3,280 upregulated and 3,416 downregulated
407 genes (**Fig. 5A**, **Fig. S3**). In contrast, the root vs. rhizome comparison showed the
408 fewest DEGs (305), including 164 upregulated and 141 downregulated genes (**Fig.**
409 **S3C**). Venn diagram analysis revealed distinct sets of DEGs across tissues, with a
410 small core set shared across all comparisons (**Fig. 5B**).

411 KEGG enrichment analysis of DEGs from comparisons between each tissue and
412 the rhizome identified the top 20 significantly enriched KEGG pathways, suggesting
413 pronounced tissue-associated differences in metabolic pathway representation (**Fig.**
414 **5C-F**). Pathways related to plant defense and triterpenoid precursor biosynthesis,
415 including phenylpropanoid biosynthesis and steroid hormone biosynthesis, were
416 significantly enriched in flower, leaf, and stem compared with the rhizome, whereas
417 no significant enrichment was observed in the root versus rhizome comparison.
418 Because KEGG enrichment reflects pathway-level overrepresentation rather than
419 gene expression levels, we further examined the expression patterns of genes involved
420 in these two pathways. Genes related to phenylpropanoid biosynthesis displayed

421 diverse expression patterns across tissues and did not show preferentially high
422 expression in rhizome or root (**Fig. S4A**). In contrast, steroid hormone
423 biosynthesis–related genes exhibited generally higher expression levels in rhizome
424 and root, with lower expression in flower, leaf, and stem (**Fig. S4B**). As this pathway
425 produces 2,3-oxidosqualene, a key precursor for triterpenoid saponin biosynthesis,
426 these results are consistent with the rhizome and root being important tissues
427 associated with saponin biosynthesis and accumulation in *P. vietnamensis*.
428

429 **Expression patterns of triterpene saponin biosynthetic genes**

430 To further characterize triterpenoid biosynthesis, we identified 21 enzyme-encoding
431 genes involved in the pathway, covering the upstream mevalonate (MVA) pathway,
432 the branching step from 2,3-oxidosqualene, and the downstream oleanane (OA)-,
433 protopanaxatriol (PPT)-, ocotillol (OCT)-, and protopanaxadiol (PPD)- type routes
434 (**Fig. 6A**; **Table S3**).

435 Expression analysis across tissues revealed distinct patterns (**Fig. 6B–F**). Genes
436 of the MVA pathway (e.g., *AACT*, *HMGS*, *HMGR*, *MVK*, *MVD*) were broadly
437 expressed, while *FPS*, *SS*, *SE*, *DDS*, and *CYP716A52* showed highest expression in
438 flowers, suggesting relatively higher precursor biosynthetic activity in this organ (**Fig.
439 6B, 6C**). Within the OA-type branch, β-amyrin synthases (*PNY1*, *PNY2*) and
440 *CYP716A52v2* were predominantly expressed in roots, suggesting a root-associated
441 preference in OA-type saponin biosynthesis. Similarly, the UDP-glycosyltransferase
442 *PgUGT8* was more highly expressed in roots (**Fig. 6D**).

443 In contrast, PPT-type biosynthetic genes such as *CYP716A53v2*, *UGTPg100*, and
444 *UGTPg101* exhibited flower-specific expression, consistent with the DEG enrichment
445 observed in the flower vs. rhizome comparison (**Fig. 5C**). This suggests that
446 hydroxylation and glycosylation steps in PPT saponin formation occur mainly in
447 floral tissues (**Fig. 6E**). Meanwhile, PPD-type glycosyltransferases including
448 *UGT71A27*, *UGT94Q2*, *UGTPg45*, and *UGTPg29* did not exhibit strong tissue
449 specificity, indicating that PPD-type glycosylation may occur broadly in both aerial
450 and underground organs (**Fig. 6F**).

451 In addition to these classical branches, we further examined the expression
452 profiles of all candidate *oxidosqualene cyclases* (*OSCs*) as well as *UGTs* potentially
453 involved in the terminal modification steps of OCT-type saponins such as majonoside
454 R2. All five *OSC* candidates were detectable across tissues, with slightly lower
455 expression in stems but no clear organ-enriched pattern overall (**Fig. 6G**). Members of
456 the *UGT* gene family also showed dispersed and heterogeneous expression across the
457 five tissues, and no consistent co-expression pattern with the *OSC* candidates was
458 observed (**Fig. S5**). These results suggest that, although these *OSCs* and *UGTs*
459 represent plausible candidates for the final cyclization and glycosylation steps of
460 OCT-type saponin biosynthesis, their transcriptional profiles do not indicate strong
461 organ specialization. This implies that these modifications may occur in multiple
462 tissues or be influenced by post-transcriptional or enzymatic regulatory mechanisms.

463 Overall, triterpene saponin biosynthesis in *P. vietnamensis* exhibits apparent
464 tissue-associated patterns of pathway activity: OA-type saponins are primarily
465 synthesized in roots, PPT-type modifications occur mainly in flowers, and PPD-type
466 glycosylation takes place across multiple tissues. In contrast, candidate genes
467 potentially involved in OCT-type saponin biosynthesis exhibit broad and non-specific
468 expression patterns, suggesting that the terminal tailoring steps of this characteristic
469 saponin class may occur in various organs.

470

471 # Discussion

472 In this study, we integrated long-read PacBio Iso-Seq and short-read Illumina
473 RNA-Seq to construct the first high-quality, full-length transcriptome atlas of *P.*
474 *vietnamensis* var. *fuscidiscus* across five tissues. This comprehensive dataset
475 uncovered widespread alternative splicing, extensive transcript diversity, and robust
476 expression of genes involved in ginsenoside biosynthesis throughout the five major
477 tissues. In contrast to a previous Illumina-based transcriptome study that assembled
478 126,758 unigenes with an average length of 1,304 bp and no isoform information [19],
479 our full-length transcriptome identified 281,468 non-redundant transcripts at isoform
480 resolution, including 8,089 novel genes and 46,917 alternative splicing events. This
481 represents a nearly 2.2-fold increase in transcript number and demonstrates the
482 capability of long-read sequencing in revealing intricate isoform diversity, providing
483 new insights into triterpenoid biosynthesis. Furthermore, we identified 21 key
484 enzyme-encoding genes potentially involved in the biosynthesis of oleanane-type
485 (OA), protopanaxadiol-type (PPD), protopanaxatriol-type (PPT), and ocotillol-type
486 (OCT) triterpenoid saponins. These findings offer important molecular insights into
487 the exceptional accumulation of OCT ginsenosides—particularly majonoside R_2—in
488 this species [1, 3].

489

490 **Tissue-specific specialization of triterpenoid biosynthetic pathways**

491 Our transcriptomic analysis revealed that upstream mevalonate (MVA) pathway
492 genes—including *AACT*, *HMGR*, *MVK*, and *MVD*—were expressed constitutively in

493 all tissues, consistent with their essential role in supplying precursors for triterpenoid
494 biosynthesis. In contrast, downstream branch-specific genes exhibited distinct
495 tissue-specific expression patterns. The enriched expression of β-amyrin synthases
496 (*PNY1*, *PNY2*) and *CYP716A52v2* in roots suggests active biosynthesis of OA-type
497 saponins in underground tissues. Conversely, *CYP716A53v2* and glycosyltransferases
498 such as *UGTPg100* and *UGTPg101* were predominantly expressed in flowers,
499 suggesting that hydroxylation and glycosylation steps for PPT-type saponins may
500 preferentially occur in floral tissues. PPD-type glycosyltransferases (e.g., *UGT94Q2*,
501 *UGT71A27*, *UGTPg29*) were more broadly distributed across multiple tissues,
502 implying widespread PPD saponin production. These expression patterns are
503 consistent with previously reported metabolite profiles [54], further supporting the
504 tissue-specific compartmentalization of triterpenoid biosynthesis in *P. vietnamensis*.
505

506 **Alternative splicing and functional diversification**

507 The integration of Iso-Seq and RNA-Seq data revealed a high prevalence of
508 alternative splicing (AS), with more than 45% of expressed genes producing multiple
509 isoforms. Retained intron (RI) emerged as the most common type of AS event.
510 Notably, key genes including *CYP716A53v2*, *DDS*, and *MVD* exhibited tissue-specific
511 splicing patterns. These AS events may influence enzyme function, subcellular
512 localization, or stability, thereby contributing to the precise regulation of triterpene
513 biosynthesis. For instance, RI-type splicing in *DDS* could generate distinct protein
514 isoforms with altered catalytic properties, potentially influencing ocotillol scaffold

515 formation. These results highlight the role of post-transcriptional regulation in the
516 sophisticated control of specialized metabolic pathways.
517 The widespread alternative splicing observed in *P. vietnamensis* var. *fuscidiscus* (48%
518 multi-isoform genes) is consistent with long-read transcriptome studies across
519 diverse species. In closely related Panax species, SMRT sequencing of P.
520 notoginseng uncovered 20,015 AS events from 6,324 genes and 8,111 novel isoforms
521 [27], and multi-tissue Iso-Seq of P. ginseng similarly revealed extensive isoform
522 expansion [55]. Comparable patterns are reported in other dicots: 93.1%
523 multi-isoform genes in *Camellia oleifera* seed [56], 40.6–52.8% in *Phragmites*
524 *australis* [34], 59.5% in *Brassica napus* with an average of 3.81 isoforms per locus
525 [57], 34.5–36.5% in *Astragalus membranaceus* [58], and 50.83% in *Gossypium*
526 *australe* [59]. Animal long-read transcriptomes also show similar complexity—for
527 example, 3.91 isoforms per gene (60.93% multi-isoform loci) in rat hippocampus [32],
528 and 25–65% multi-isoform genes in threespine stickleback fish [33] and pig testis
529 cells [60]. Collectively, long-read transcriptomes typically reveal 30–60% of genes
530 with multiple isoforms and an average of 2–4 isoforms per locus. Therefore, the
531 isoform diversity we detected in *P. vietnamensis* falls within the range reported for
532 both *Panax* and other dicot plants, reinforcing that alternative splicing is a major
533 driver of transcriptomic and functional diversification in this lineage.
534
535 **Mechanistic implications for OCT-type ginsenoside accumulation**

536 Despite the pharmacological significance of ocotillol (OCT)-type ginsenosides,
537 particularly their anticancer and hepatoprotective properties, the biosynthetic pathway
538 responsible for formation of the OCT scaffold remains poorly understood. [17]. The
539 distinctive structure of these saponins, characterized by a tetrahydrofuran (THF) ring,
540 suggests the involvement of an atypical *oxidosqualene cyclase* (*OSC*) that differs
541 from canonical β-amyrin synthase or dammarenediol synthase. Although no *OSC*
542 specific to OCT biosynthesis was identified among the annotated transcripts, the
543 detection of alternatively spliced variants of *DDS* may point to possible
544 neofunctionalization. Further biochemical validation using heterologous expression
545 systems will be essential to identify the *OSC* that catalyzes the cyclization of
546 2,3-oxidosqualene into the OCT backbone.

547 We further investigated the biosynthesis of majonoside R2, the hallmark
548 OCT-type saponin of *P. vietnamensis*. Prior biochemical studies demonstrated that
549 *PvfUGT1* and *PvfUGT2* sequentially catalyze the two glycosylation steps converting
550 protopanaxatriol-oxide intermediates (RT4/RT5) into majonoside R2, providing key
551 enzymatic evidence for this pathway [61, 62]. To place these enzymes within the
552 broader *UGT* repertoire of *P. vietnamensis*, we incorporated *PvfUGT1* and *PvfUGT2*
553 into a phylogenetic framework and examined the expression of candidate *UGTs*
554 clustering alongside them. These *UGTs* showed broad, non-specialized expression
555 across all tissues, and the same pattern was observed for candidate *OSCs* that may
556 catalyze OCT backbone formation. The absence of tissue enrichment suggests that
557 majonoside R2 biosynthesis is not confined to a specific organ and may depend on

558 additional regulatory factors, such as substrate availability and subcellular
559 compartmentation, rather than on transcript abundance alone.

560 Together, these findings indicate that, unlike the clearly partitioned OA-, PPT-,
561 and PPD-type pathways, the OCT-type pathway is more diffusely regulated.
562 Integrating established biochemical functions of *PvfUGT1* and *PvfUGT2* with our
563 phylogenetic and expression analyses refines the current model of MR2 biosynthesis
564 and highlights targets for future functional and metabolic engineering studies.

565

566 **Evolutionary adaptations and biosynthetic innovation**

567 From an evolutionary perspective, the ability of *P. vietnamensis* to accumulate up to
568 75% ocotillol (OCT)-type saponins in its rhizomes likely arises from multiple
569 synergistic mechanisms: (1) constitutive expression of upstream MVA pathway genes
570 in underground tissues ensures a steady supply of precursors; (2) neofunctionalized
571 oxidosqualene cyclase (OSC) enzymes or alternatively spliced *DDS* isoforms may
572 enable ocotillol production; and (3) an extensive glycosyltransferase network
573 facilitates diverse modifications of the ocotillol scaffold. These adaptations
574 collectively suggest that OCT biosynthesis in *P. vietnamensis* may represent a
575 lineage-specific metabolic innovation, potentially driven by gene duplication and
576 functional divergence following whole-genome duplication (WGD), as proposed in
577 other *Panax* species.

578 While this study provides a comprehensive transcriptomic resource for *P.*
579 *vietnamensis* var. *fuscidiscus* and identifies candidate genes and alternative splicing

580 events potentially involved in OCT-type ginsenoside biosynthesis, functional
581 validation of these genes (e.g., knock-down, knock-out, overexpression) was not
582 performed. Key knowledge gaps remain, including the identity of the *OSCs*
583 responsible for forming the OCT skeleton. Nevertheless, the high-quality
584 transcriptome and tissue-specific expression profiles generated here provide a solid
585 foundation for targeted functional studies. Future work should focus on gene
586 characterization via CRISPR/Cas-mediated knockout, overexpression, or heterologous
587 expression, and on elucidating transcriptional regulators and cis-elements controlling
588 tissue-specific expression. Integrating transcriptomic, metabolomic, proteomic, and
589 single-cell RNA sequencing data will further enable a systems-level understanding of
590 OCT biosynthesis and reveal new targets for metabolic engineering.

591 In conclusion, our study provides a comprehensive transcriptomic resource for *P.*
592 *vietnamensis* and marks a substantial step toward deciphering the biosynthetic
593 pathways and regulatory mechanisms underlying OCT ginsenoside accumulation.
594 This work not only enhances our understanding of ginseng metabolic diversity, but
595 also establishes a foundation for synthetic biology approaches aimed at the
596 sustainable production of pharmacologically valuable OCT saponins.

597 # Conclusion

598 This study reports the first high-quality, full-length, tissue-specific transcriptome atlas
599 of *P. vietnamensis* var. *fuscidiscus*, constructed through integrated PacBio Iso-Seq
600 and Illumina RNA-Seq sequencing. A total of 281,468 non-redundant transcripts were
601 assembled, including 8,089 novel genes not previously annotated, substantially
602 expanding the currently available genetic repertoire of this pharmacologically
603 valuable species.

604 We identified 21 key enzyme-encoding genes involved in triterpenoid saponin
605 biosynthesis and identified tissue-associated expression patterns across the oleanane-,
606 protopanaxadiol-, protopanaxatriol-, and ocotillol-type biosynthetic pathways.
607 Notably, over 46,000 alternative splicing events were detected, including isoform
608 diversity in critical biosynthetic genes such as *DDS* and *CYP716A53v2*, underscoring
609 the importance of post-transcriptional regulation in specialized metabolite production.
610 Collectively, these findings provide new molecular insights relevant to ocotillol-type
611 ginsenoside biosynthesis and establish a valuable transcriptomic resource that may
612 support future studies aimed at identifying candidate enzymes, exploring potential
613 regulatory mechanisms, and guiding subsequent functional and metabolic engineering
614 research. This comprehensive transcriptomic resource also reinforces the value of *P.*
615 *vietnamensis* as an excellent model for studying lineage-specific metabolic
616 innovations in medicinal plants.

617

618 **Abbreviations**

619 *AACT*: Acetyl-CoA acetyltransferase

620 AF: Alternative first exon

621 AL: Alternative last exon

622 AS: Alternative splicing

623 A3: Alternative 3′ splice site

624 A5: Alternative 5′ splice site

625 CCS: Circular consensus sequence

626 CD-HIT: Cluster Database at High Identity with Tolerance

627 COG/KOG: Clusters of Orthologous Groups / Eukaryotic Orthologous Groups

628 *CYP*: Cytochrome P450

629 *DDS*: Dammarenediol synthase

630 DEG: Differentially expressed gene

631 eggNOG: Evolutionary genealogy of genes: Non-supervised Orthologous Groups

632 FDR: False discovery rate

633 FL: Full-length

634 FLNC: Full-length non-chimeric

635 *FPS*: Farnesyl diphosphate synthase

636 FSM: Full splice match

637 GO: Gene Ontology

638 *HMGR*: 3-hydroxy-3-methylglutaryl-coenzyme A reductase

639 HMGS: 3-hydroxy-3-methylglutaryl coenzyme A synthase

640 HQ: High-quality

641 ICE: Iterative Clustering for Error Correction

642 ISM: Incomplete Splice Match

643 Iso-Seq: Isoform sequencing

644 KEGG: Kyoto Encyclopedia of Genes and Genomes

645 LQ: Low-quality

646 MVA: Mevalonate pathway

647 *MVD*: Mevalonate diphosphate decarboxylase

648 *MVK*: Mevalonate kinase

649 MX: Mutually exclusive exons

650 NIC: Novel in catalog

651 NNC: Novel not in catalog

652 OCT: Ocotillol

653 *OSC*: Oxidosqualene cyclase

654 PPD: Protopanaxadiol

655 PPT: Protopanaxatriol

656 RI: Retained intron

657 RNA-Seq: RNA sequencing

658 ROIs: Reads of insert

659 SE: Skipped exon

660 *SE*: Squalene epoxidase

661 SGS: Second-generation sequencing

662 SMRT: Single-Molecule Real-Time sequencing

663 *SS*: Squalene synthase

664 SSR: Simple sequence repeat

665 TSS/TTS: Transcription start site / transcription termination site

666 *UGT*: UDP-glycosyltransferase

667

668 Authors' contributions

669 LZG conceived the study. WL performed transcriptome analyses and drafted the
670 manuscript. ZX and AN contributed to RNA extraction, sequencing, and data
671 validation. LZG revised the manuscript critically and supervised the project. All
672 authors read and approved the final manuscript.

673

674 Acknowledgements

675 Not applicable.

676

677 Funding

678 This work was supported by Natural Science Foundation of China (U1902205) and a
679 startup grant of Hainan University to Li-zhi Gao.

680

681 Data availability

682 The raw data were submitted to NCBI under the BioProject number PRJNA1312250.

683

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Clinical trial number

Not applicable.

References

1. Kim Y-J, Zhang D, Yang D-C. Biosynthesis and biotechnological production of ginsenosides. Biotechnology Advances. 2015;33:717–35. https://doi.org/10.1016/j.biotechadv.2015.03.001.
2. Xia P, Huang Y, Zhu J. The complete chloroplast genome sequences of 11 *Panax* species: Providing insights for evolution and species identification. Industrial Crops and Products. 2025;223:120160. https://doi.org/10.1016/j.indcrop.2024.120160.
3. Christensen LP. Ginsenosides: Chemistry, biosynthesis, analysis, and potential health effects. Advances in Food and Nutrition Research. 2009;:1–99. https://doi.org/10.1016/s1043-4526(08)00401-4.
4. Hou M, Wang R, Zhao S, Wang Z. Ginsenosides in *Panax* genus and their biosynthesis. ACTA PHARMACEUTICA SINICA B. 2021;11:1813–34. https://doi.org/10.1016/j.apsb.2020.12.017.
5. Ngau T, Dieu C, Hue C, Huynh D, Do B, Huynh L. Ethnopharmacology, Genetic Diversity, Phytochemistry, and Pharmacological Effects of *Panax vietnamensis* Ha et Grushv.: A Review. JOURNAL OF ENVIRONMENTAL PATHOLOGY TOXICOLOGY AND ONCOLOGY. 2025;44:73–92. https://doi.org/10.1615/JEnvironPatholToxicolOncol.2024052414.
6. Yang Z, Li X, Yang L, Peng S, Song W, Lin Y, et al. Comparative genomics reveals the diversification of triterpenoid biosynthesis and origin of ocotillol-type triterpenes in *Panax*. Plant Communications. 2023;4:100591. https://doi.org/10.1016/j.xplc.2023.100591.
7. Zhan Z, Zhang J, Huang W, Huang J. Transcriptomic strategy provides molecular insights into the growth and ginsenosides accumulation of *Panax ginseng*. Phytomedicine. 2025;143:156834. https://doi.org/10.1016/j.phymed.2025.156834.
8. Komatsu K, Zhu S, Fushimi H, Qui TK, Cai S, Kadota S. Phylogenetic analysis based on 18S rRNA gene and matK gene sequences of *Panax vietnamensis* and five related species. Planta med. 2001;67:461–5. https://doi.org/10.1055/s-2001-15821.
9. DUC N, KASAI R, OHTANI K, ITO A, NHAM N, YAMASAKI K, et al. Saponins from vietnamese ginseng, *Panax vietnamensis* HA et grushv. Collected in central vietnam 2. CHEMICAL & PHARMACEUTICAL BULLETIN. 1994;42:115–22.
10. Huong NT, Matsumoto K, Nham NT, Quang NH, Duc NM, Yamasaki K, et al. Effect of vietnamese ginseng on the phagocytosis in vitro and in vivo. Phytomedicine. 1997;4:341–6. https://doi.org/10.1016/S0944-7113(97)80044-X.
11. Konoshima T, Takasaki M, Ichiishi E, Murakami T, Tokuda H, Nishino H, et al. Cancer chemopreventive activity of majonoside-R2 from vietnamese ginseng, *Panax vietnamensis*. Cancer Letters. 1999;147:11–6. https://doi.org/10.1016/S0304-3835(99)00257-8.
12. Yunusova N, Kim JY, Lee GJ, Hong JY, Shin B, Cai S-Q, et al. Comparison of ginsenosides in radix and rhizome of wild *Panax* species using LC-ELSD and LC-Q-TOF-MS. INTERNATIONAL JOURNAL OF FOOD SCIENCE AND TECHNOLOGY. 2015;50:1607–14. https://doi.org/10.1111/ijfs.12814.
13. Nguyen MD, Kasai R, Ohtani K, Ito A, Nguyen TN, Yamasaki K, et al. Saponins from vietnamese ginseng, *Panax vietnamensis* HA et grushv. Collected in central vietnam

1. Chemical & pharmaceutical bulletin. 1994;42:115–22.
14. Nguyen TD, Villard PH, Barlatier A, Elsisi AE, Jouve E, Duc NM, et al. *Panax vietnamensis* protects mice against carbon tetrachloride-induced hepatotoxicity without any modification of CYP2E1 gene expression. Planta medica. 2000;66:714–9. https://doi.org/10.1055/s-2000-9603.
15. Yamasaki K. Bioactive saponins in vietnamese ginseng, *Panax vietnamensis*. Pharmaceutical Biology. 2000;38:16–24. https://doi.org/10.1076/phbi.38.6.16.5956.
16. ZHU S, FUSHIMI H, CAI S, CHEN H, KOMATSU K. A new variety of the genus *Panax* from southern yunnan, China and its nucleotide sequences of 18S ribosomal RNA gene and matK gene. Journal of Japanese Botany. 2003;78:86–94.
17. Yang Z, Li X, Yang L, Peng S, Song W, Lin Y, et al. Comparative genomics reveals the diversification of triterpenoid biosynthesis and origin of ocotillol-type triterpenes in *Panax*. Plant Communications. 2023;4:100591. https://doi.org/10.1016/j.xplc.2023.100591.
18. Zhu S, Zou K, Fushimi H, Cai S, Komatsu K. Comparative study on triterpene saponins of ginseng drugs. Planta medica. 2004;70:666–77. https://doi.org/10.1055/s-2004-827192.
19. Zhang G-H, Ma C-H, Zhang J-J, Chen J-W, Tang Q-Y, He M-H, et al. Transcriptome analysis of *Panax vietnamensis* var. *fuscidicus* discovers putative ocotillol-type ginsenosides biosynthesis genes and genetic markers. BMC Genomics. 2015;16:159. https://doi.org/10.1186/s12864-015-1332-8.
20. Morita T, Kasai R, Tanaka O, Zhou J, Yang T, Shoji J. Saponins of zu-tziseng, rhizomes of *Panax japonicus* C.A. MEYER var. major (BURK.) C.Y. wu et K.M. FENG, collected in yunnan, China. Chem Pharm Bull. 1982;30:4341–5. https://doi.org/10.1248/cpb.30.4341.
21. Konoshima T, Takasaki M, Tokuda H, Nishino H, Duc NM, Kasai R, et al. Anti-tumor-promoting activity of majonoside-R2 from vietnamese ginseng, *Panax vietnamensis* ha et grushv. (I). Biological & pharmaceutical bulletin. 1998;21:834–8. https://doi.org/10.1248/bpb.21.834.
22. Huong, N.T.T., Matsumoto, K., Watanabe, H. The antistress effect of majonoside-R2, a major saponin componentof vietnamese ginseng: Neuronal mechanisms of action. Methods and Findings in Experimental and Clinical Pharmacology. 1998;20:65. https://doi.org/10.1358/mf.1998.20.1.485634.
23. Tran QL, Adnyana IK, Tezuka Y, Nagaoka T, Tran QK, Kadota S. Triterpene saponins from vietnamese ginseng (*Panax vietnamensis*) and their hepatocytoprotective activity. J Nat Prod. 2001;64:456–61. https://doi.org/10.1021/np000393f.
24. Tran QL, Adnyana IK, Tezuka Y, Harimaya Y, Saiki I, Kurashige Y, et al. Hepatoprotective effect of majonoside R2, the major saponin from vietnamese ginseng (*Panax vietnamensis*). Planta medica. 2002;68:402–6. https://doi.org/10.1055/s-2002-32069.
25. Gurung B, Bhardwaj PK, Talukdar NC. Subtractive transcriptome analysis of leaf and rhizome reveals differentially expressed transcripts in *Panax sokpayensis*. FUNCTIONAL & INTEGRATIVE GENOMICS. 2016;16:619–39. https://doi.org/10.1007/s10142-016-0517-9.
26. Zhang D, Li W, Xia E, Zhang Q, Liu Y, Zhang Y, et al. The medicinal herb *Panax*

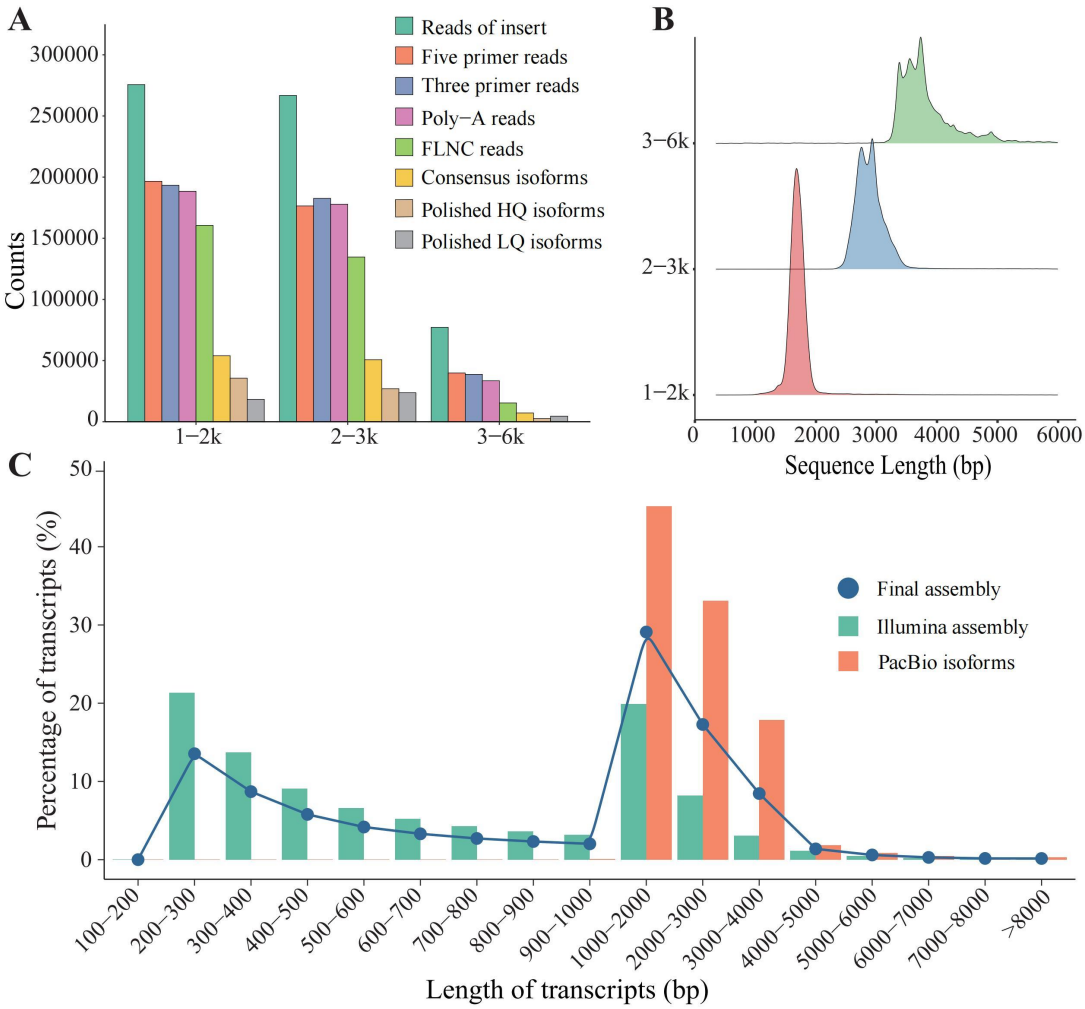
notoginseng genome provides insights into ginsenoside biosynthesis and genome evolution. Molecular Plant. 2017;10:903–7. https://doi.org/10.1016/j.molp.2017.02.011.

27. Zhang D, Li W, Chen Z, Wei F, Liu Y, Gao L. SMRT- and illumina-based RNA-seq analyses unveil the ginsinoside biosynthesis and transcriptomic complexity in *Panax notoginseng*. Sci Rep. 2020;10:15310. https://doi.org/10.1038/s41598-020-72291-1.
28. Nguyen HT, Phan LK, Huynh K-LV, Duong T-H, Le HT, Hai-Yen NT, et al. Untargeted metabolomics approach for the differentiation between *Panax vietnamensis* var. *vietnamensis* and *Panax vietnamensis* var. *fuscidiscus*. Metabolites. 2023;13:763. https://doi.org/10.3390/metabo13060763.
29. Tien NQD. De novo whole-genome assembly and discovery of genes involved in triterpenoid saponin biosynthesis of Vietnamese ginseng (*Panax vietnamensis* Ha et Grushv.). Physiol Mol Biol Plants. 2021. https://doi.org/10.1007/s12298-021-01076-1.
30. Vu DD, Shah SNM, Pham MP, Bui VT, Nguyen MT, Nguyen TPT. De novo assembly and Transcriptome characterization of an endemic species of Vietnam, *Panax vietnamensis* Ha et Grushv., including the development of EST-SSR markers for population genetics. BMC Plant Biol. 2020;20:358. https://doi.org/10.1186/s12870-020-02571-5.
31. Luu LH, Pham HLB, Huynh HTT, Nguyen LN, Vu TT, Le HTT. Differential expression of ginsenoside biosynthesis-related genes at an early developmental stage of *Panax vietnamensis*. Plant Biotechnology Reports. 2022;16:215–28. https://doi.org/10.1007/s11816-022-00751-3.
32. Wang X, You X, Langer JD, Hou J, Rupprecht F, Vlatkovic I, et al. Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus. Nat Commun. 2019;10:5009. https://doi.org/10.1038/s41467-019-13037-0.
33. Naftaly AS, Pau S, White MA. Long-read RNA sequencing reveals widespread sex-specific alternative splicing in threespine stickleback fish. Genome Res. 2021;31:1486–97. https://doi.org/10.1101/gr.274282.120.
34. Cui J, Qiu T, Li L, Cui S. De novo full-length transcriptome analysis of two ecotypes of *Phragmites australis* (swamp reed and dune reed) provides new insights into the transcriptomic complexity of dune reed and its long-term adaptation to desert environments. BMC Genomics. 2023;24:180. https://doi.org/10.1186/s12864-023-09271-y.
35. Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, et al. SQANTI3: Curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat Methods. 2024;21:793–7. https://doi.org/10.1038/s41592-024-02229-2.
36. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2. https://doi.org/10.1093/bioinformatics/bts565.
37. Salmela L, Rivals E. LoRDEC: Accurate and efficient long read error correction. Bioinformatics. 2014;30:3506–14. https://doi.org/10.1093/bioinformatics/btu538.
38. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Molecular Biology and Evolution. 2021;38:5825–9. https://doi.org/10.1093/molbev/msab293.
39. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-seq data with

or without a reference genome. BMC Bioinformatics. 2011;12:323. https://doi.org/10.1186/1471-2105-12-323.

40. Chen C, Wu Y, Li J, Wang X, Zeng Z, Xu J, et al. TBtools-II: A “one for all, all for one” bioinformatics platform for biological big-data mining. Molecular Plant. 2023;16:1733–42. https://doi.org/10.1016/j.molp.2023.09.010.
41. Robinson MD, McCarthy DJ, Smyth GK. edgeR : A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40. https://doi.org/10.1093/bioinformatics/btp616.
42. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191.
43. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008. https://doi.org/10.1093/gigascience/giab008.
44. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, et al. SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19:40. https://doi.org/10.1186/s13059-018-1417-1.
45. Eddy SR. Profile hidden markov models. Bioinformatics. 1998;14:755–63. https://doi.org/10.1093/bioinformatics/14.9.755.
46. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, et al. The pfam protein families database: Embracing AI/ML. Nucleic Acids Research. 2025;53:D523–34. https://doi.org/10.1093/nar/gkae997.
47. Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, et al. InterPro: The protein sequence classification resource in 2025. Nucleic Acids Research. 2025;53:D444–56. https://doi.org/10.1093/nar/gkae1082.
48. Xie J, Chen Y, Cai G, Cai R, Hu Z, Wang H. Tree visualization by one table (tvBOT): A web application for visualizing, modifying and annotating phylogenetic trees. Nucleic Acids Research. 2023;51:W587–92. https://doi.org/10.1093/nar/gkad359.
49. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. JOSS. 2019;4:1686. https://doi.org/10.21105/joss.01686.
50. Wickham H. ggplot2. Cham: Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-24277-4.
51. Blighe K, Rana S, Lewis M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.
52. Kolde R. pheatmap: Pretty heatmaps. 2010;:1.0.13. https://doi.org/10.32614/CRAN.package.pheatmap.
53. Xu S, Hu E, Cai Y, Xie Z, Luo X, Zhan L, et al. Using clusterProfiler to characterize multiomics data. Nat Protoc. 2024;19:3292–320. https://doi.org/10.1038/s41596-024-01020-z.
54. Le THV, Lee GJ, Vu HKL, Kwon SW, Nguyen NK, Park JH, et al. Ginseng saponins in different parts of *Panax vietnamensis*. Chem Pharm Bull. 2015;63:950–4. https://doi.org/10.1248/cpb.c15-00369.
55. Jo I-H, Lee J, Hong C, Lee D, Bae W, Park S-G, et al. Isoform sequencing provides a more comprehensive view of the *Panax ginseng* transcriptome. Genes. 2017;8:228. https://doi.org/10.3390/genes8090228.

56. Gong W, Song Q, Ji K, Gong S, Wang L, Chen L, et al. Full-length transcriptome from *Camellia oleifera* seed provides insight into the transcript variants involved in oil biosynthesis. J Agric Food Chem. 2020;68:14670–83. https://doi.org/10.1021/acs.jafc.0c05381.
57. Yao S, Liang F, Gill RA, Huang J, Cheng X, Liu Y, et al. A global survey of the transcriptome of allopolyploid *Brassica napus* based on single-molecule long-read isoform sequencing and illumina-based RNA sequencing data. The Plant Journal. 2020;103:843–57. https://doi.org/10.1111/tpj.14754.
58. Li J. Long read reference genome-free reconstruction of a full-length transcriptome from *Astragalus* membranaceus reveals transcript variants involved in bioactive compound biosynthesis. Cell Discovery. 2017.
59. Feng S, Xu M, Liu F, Cui C. Reconstruction of the full-length transcriptome atlas using PacBio iso-seq provides insight into the alternative splicing in *Gossypium australe*. BMC Plant Biol. 2019;19:365. https://doi.org/10.1186/s12870-019-1968-7.
60. Lin W, Zhang X, Liu Z, Huo H, Chang Y, Zhao J, et al. Isoform-resolution single-cell RNA sequencing reveals the transcriptional panorama of adult baoshan pig testis cells. BMC Genomics. 2025;26:459. https://doi.org/10.1186/s12864-025-11636-4.
61. Seki H, Tamura K, Muranaka T. P450s and UGTs: Key players in the structural diversity of triterpenoid saponins. PLANT AND CELL PHYSIOLOGY. 2015;56:1463–71. https://doi.org/10.1093/pcp/pcv062.
62. Peng S, Li X, Jiang W, Wang Y, Xiang G, Li M, et al. Identification of two key UDP-glycosyltransferases responsible for the ocotillol-type ginsenoside majonside-R2 biosynthesis in *Panax vietnamensis* var. *fuscidiscus*. PLANTA. 2023;257. https://doi.org/10.1007/s00425-023-04143-x.



898

899 **Figure 1. Overview of PacBio SMRT sequencing and transcript length**
900 **distribution.** (A) Distributions of read counts and lengths from PacBio libraries with
901 size fractions of 1–2 kb, 2–3 kb, and 3–6 kb. (B) Length distribution of full-length
902 non-chimeric (FLNC) reads across different libraries. (C) Comparison of transcript
903 lengths between PacBio SMRT and Illumina assemblies.

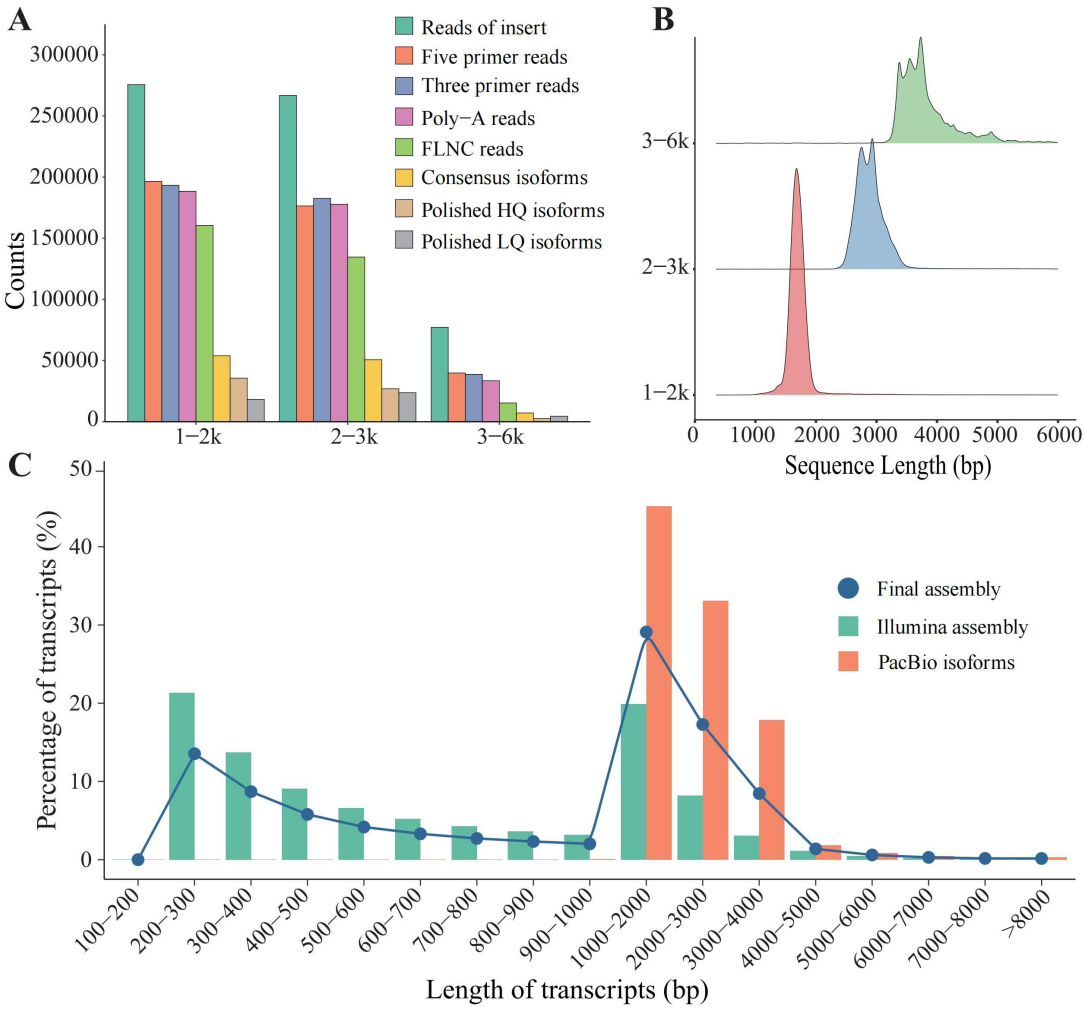


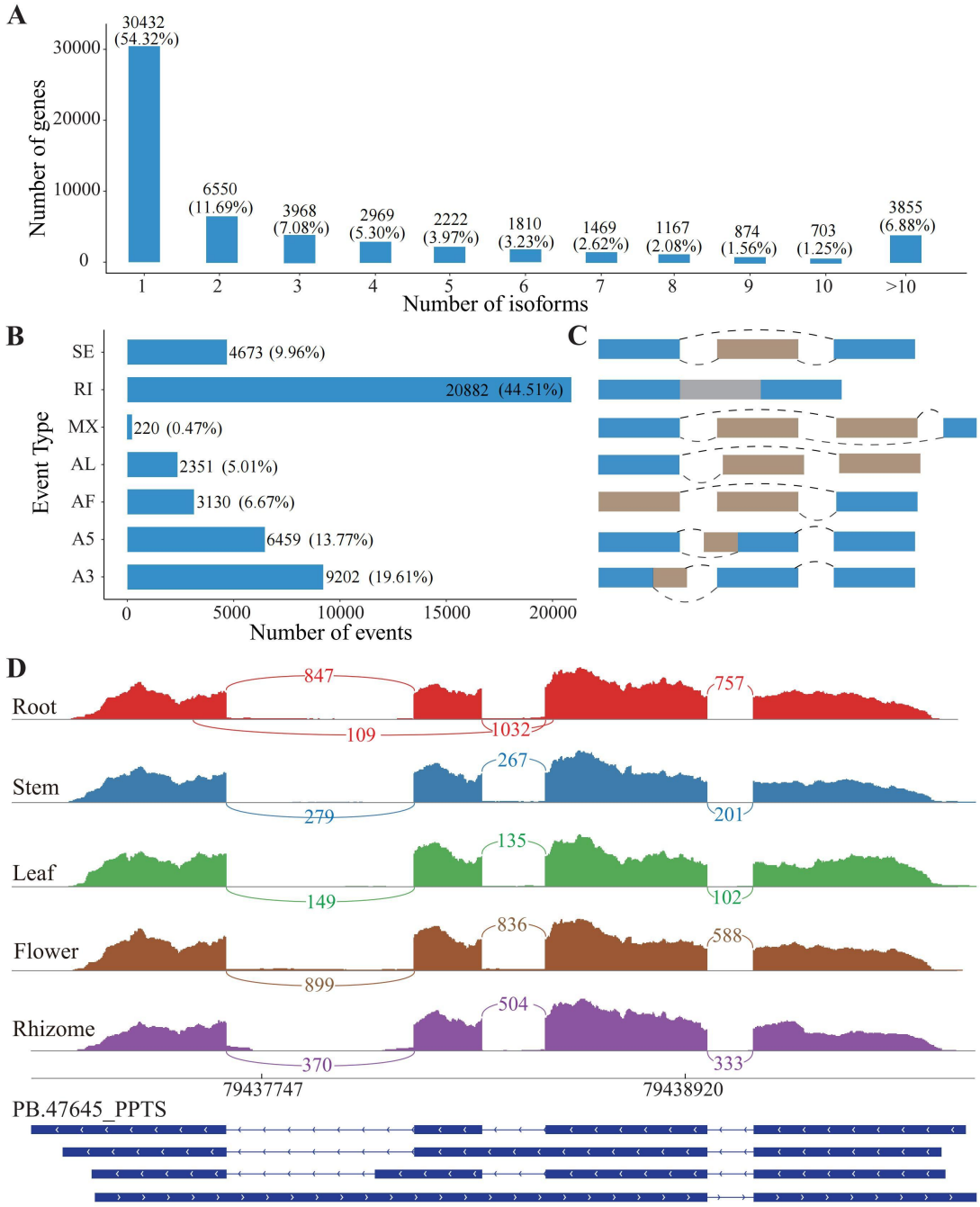
Figure 2. Structural characterization of Iso-Seq isoforms in *P. vietnamensis* var. *fuscidiscus*. (A) Gene-level splicing complexity, represented by the number of isoforms per gene. (B) Comparison of isoform counts between annotated and novel genes. (C) SQANTI3 classification of isoforms into nine structural categories: full-splice match (FSM), incomplete splice match (ISM), novel in catalog (NIC), novel not in catalog (NNC), genic genomic, antisense, fusion, intergenic, and intronic transcripts. Each category is subdivided into protein-coding (dark) and non-coding (light) isoforms. (D) Isoform length distribution across SQANTI3 categories. (E) Summary of isoform length ranges by structural class. (F) Exon count distribution

among isoforms across categories. (G) Distance between 3′ ends of FSM transcripts and annotated transcription termination sites (TTS); negative values indicate upstream positions. (H) Distance between 5′ ends of FSM transcripts and annotated transcription start sites (TSS); negative values indicate downstream positions relative to annotated TSS.

Figure3. Functional annotation and classification of *P. vietnamensis* var. *fuscidiscus* transcriptomes in rhizome and flower. (A) Gene Ontology (GO) classification of transcriptomes from rhizome and flower. (B) Kyoto Encyclopedia of

923 Genes and Genomes (KEGG) pathway classification of transcriptomes from rhizome

924 and flower.



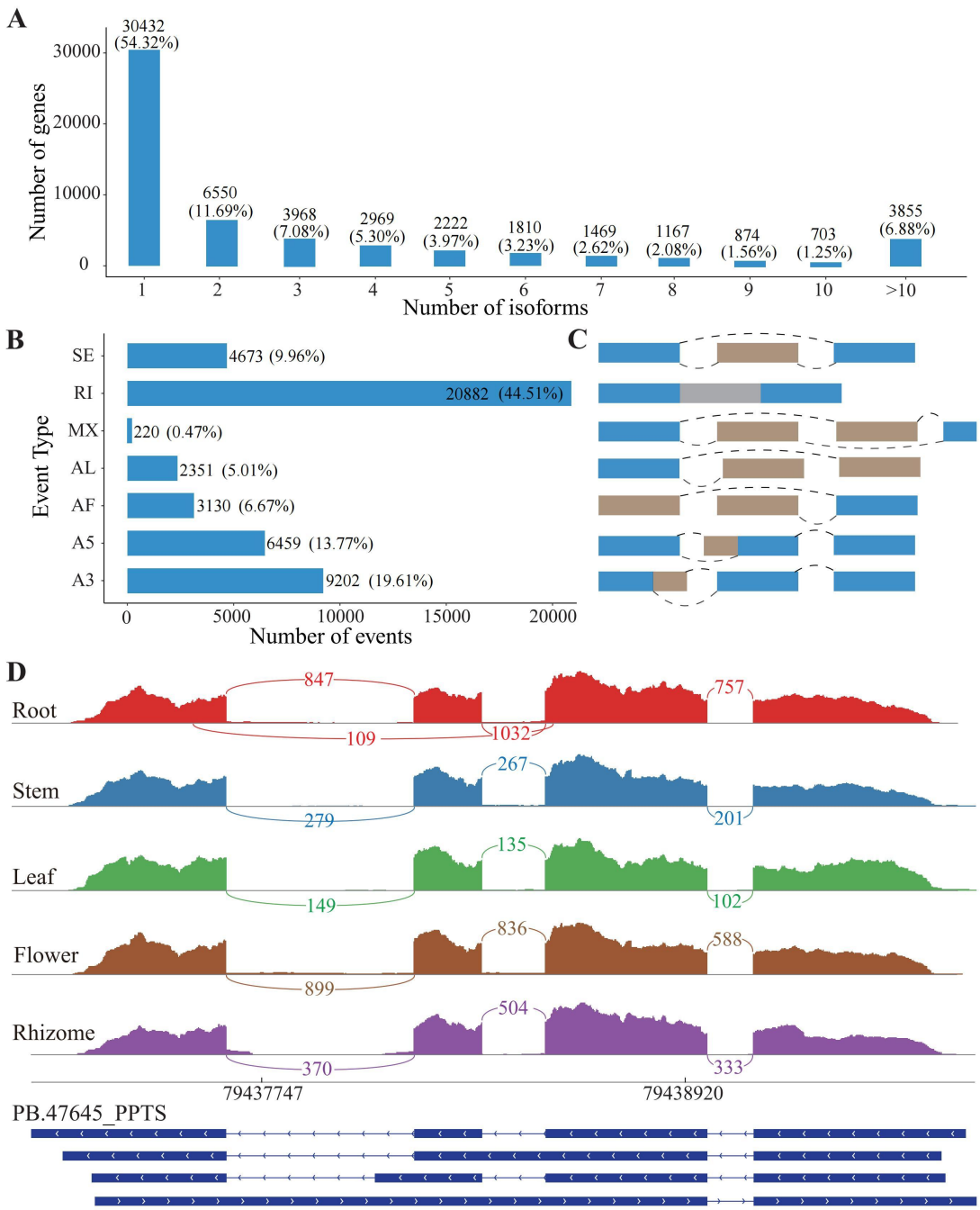
925

926 **Figure 4. Characterization of alternative splicing (AS) events in *P. vietnamensis***

927 **var. *fuscidiscus*.** (A) Proportion of genes producing one or more isoforms in the

928 Iso-Seq data; counts and percentages are indicated above bars. (B) Total number of

929 AS events identified from Iso-Seq data; counts and percentages are shown above bars.
930 (C) Schematic representation of seven major AS types: skipped exon (SE), retained
931 intron (RI), mutually exclusive exon (MX), alternative 3′ splice site (A3), alternative
932 5′ splice site (A5), alternative first exon (AF), and alternative last exon (AL). (D)
933 Sashimi plot showing transcript isoforms of the cytochrome P450 gene *CYP716A53v2*
934 (*PPTS*) based on Iso-Seq data. Peaks in red, blue, green, brown, and purple represent
935 short-read RNA-seq coverage in root, stem, leaf, flower, and rhizome, respectively,
936 and curved lines with numbers in corresponding colors represent splicing junctions
937 supported by that number of short-read RNA-seq reads. For each isoform, blocks in
938 blue represent exons and lines in-between represent introns. The presence of retained
939 intron (RI) isoforms indicates alternative splicing–mediated transcript diversity in
940 *CYP716A53v2*.



941

942 **Figure 5. Tissue-specific differential gene expression and KEGG pathway**

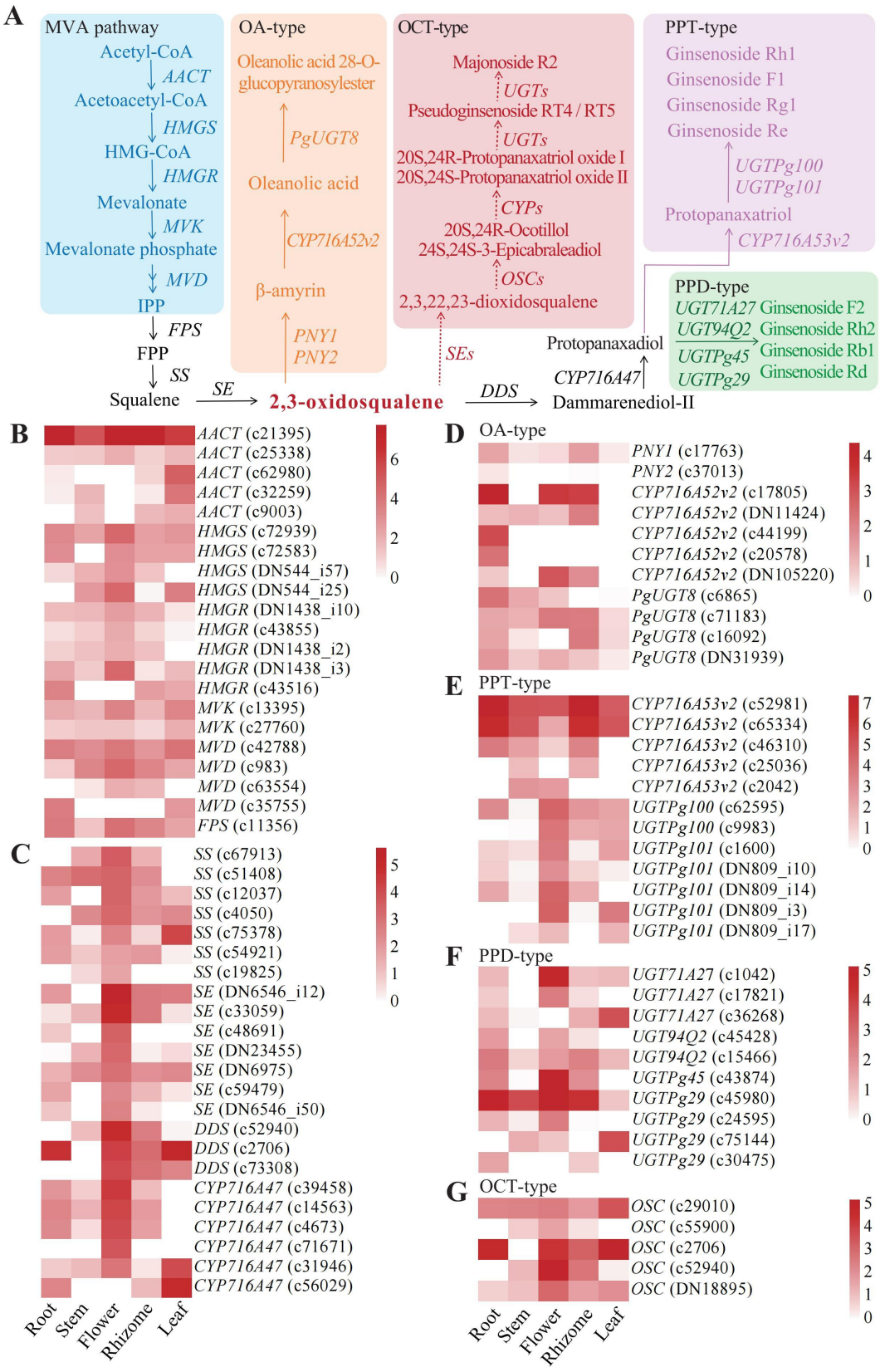
943 **enrichment in *P. vietnamensis* var. *fuscidiscus*.** (A) Number of differentially

944 expressed genes (DEGs) in pairwise comparisons between rhizome and other tissues;

945 red and blue bars represent up- and down-regulated genes, respectively. (B) Venn

946 diagram showing overlap of DEGs across four tissue comparisons (flower, leaf, root,

947 and stem vs. rhizome). (C–F) Top 20 KEGG pathways enriched in DEGs for flower
948 vs. rhizome (C), leaf vs. rhizome (D), root vs. rhizome (E), and stem vs. rhizome (F).
949 Pathways related to phenylpropanoid biosynthesis and steroid hormone biosynthesis
950 are highlighted in red.



951

952 **Figure 6. Triterpenoid saponin biosynthesis pathways and tissue-specific**

953 **expression of candidate genes in *P. vietnamensis* var. *fuscidiscus*.** (A) Proposed

954 biosynthetic pathways from the mevalonate (MVA) pathway to oleanane (OA)-type,
955 protopanaxadiol (PPD)-type, protopanaxatriol (PPT)-type, and ocotillol (OCT)-type
956 saponins. Candidate genes are shown in italics. (B–F) Heatmaps showing
957 tissue-specific expression patterns of candidate genes involved in different steps of
958 triterpenoid saponin biosynthesis, including enzymes from the upstream MVA
959 pathway and homologous genes associated with the biosynthetic pathways of OA
960 -type, PPD-type, PPT-type, and OCT-type saponins. (G) Heatmap showing the
961 expression patterns of all identified oxidosqualene cyclase (*OSC*) genes across five
962 tissues (root, stem, flower, leaf, and rhizome). Gene expression levels are shown as
963 $\log_2$(TPM + 1). The abbreviations are indicated as follows, *AACT*: acetyl-CoA
964 acetyltransferase, *HMGS*: 3-hydroxy-3-methylglutaryl coenzyme A synthase, *HMGR*:
965 3-hydroxy-3-methylglutaryl coenzyme A reductase, *MVK*: mevalonate kinase, *MVD*:
966 mevalonate diphosphate decarboxylase, *FPS*: farnesyl diphosphate synthase, *SS*:
967 squalene synthase, *SE*: squalene epoxidase, *DDS*: dammarenediol synthase, *CYP*:
968 cytochrome P450, *UGT*: UDP-glycosyltransferase.

969

970 # Supplementary Information

971 **Figure S1. Functional annotation and classification of *P. vietnamensis* var.**
972 ***fuscidiscus* transcriptomes in leaf, root, and stem.** (A, B) GO terms and KEGG
973 pathways assigned to leaf transcriptomes. (C, D) GO terms and KEGG pathways
974 assigned to root transcriptomes. (E, F) GO terms and KEGG pathways assigned to
975 stem transcriptomes.

976 **Figure S2. Comparative analysis of alternative splicing (AS) events across tissues**
977 **in *P. vietnamensis* var. *fuscidiscus*.** (A) Sashimi plot illustrating isoform structures
978 of the *dammarenediol synthase* (*DDS*) gene based on Iso-Seq data. (B) Sashimi plot
979 illustrating isoform structures of the *mevalonate-5-pyrophosphate decarboxylase*
980 (*MVD*) gene based on Iso-Seq data. In panels (A) and (B), rectangular boxes represent
981 exons, with box lengths proportional to exon size, while connecting lines indicate
982 splicing junctions. Numbers displayed on the junction arcs denote the number of
983 full-length Iso-Seq reads supporting each splicing event. Colored coverage peaks (red,
984 blue, green, brown, and purple) represent short-read RNA-seq coverage in root, stem,
985 leaf, flower, and rhizome, respectively. Curved lines with numbers in corresponding
986 colors indicate splice junctions supported by short-read RNA-seq data in each tissue.
987 For each isoform, blue blocks represent exons and the connecting lines indicate
988 introns. (C) Counts of significant AS events ($|\Delta PSI| > 0.3$) and associated genes in
989 pairwise comparisons between flower, leaf, root, or stem and rhizome. (D)
990 Distribution of seven major AS event types (SE, RI, MX, A5, A3, AF, AL) across
991 tissue comparisons. RI and SE were the most abundant types.

992 **Figure S3. MA plots of differential gene expression between rhizome and other**
993 **tissues in *P. vietnamensis* var. *fuscidiscus*.** (A–D) MA plots for flower (A), leaf (B),
994 root (C), and stem (D) versus rhizome. Log_2 fold change (logFC) is plotted against
995 average expression (logCounts). Significantly differentially expressed genes (DEGs)
996 are shown in red; non-significant genes are in black.

997 **Figure S4. Expression patterns of genes involved in phenylpropanoid and steroid**
998 **hormone biosynthesis across tissues.** (A) Heatmap showing the expression patterns
999 of genes involved in phenylpropanoid biosynthesis (KEGG pathway 00940) across
1000 five tissues (root, stem, flower, leaf, and rhizome). (B) Heatmap showing the
1001 expression patterns of genes involved in steroid hormone biosynthesis (KEGG
1002 pathway 00140) across five tissues. Gene expression levels are shown as log_2(TPM +
1003 1). Genes are displayed without hierarchical clustering to facilitate direct comparison
1004 of expression levels among tissues.

1005 **Figure S5. Expression patterns of *UDP-glycosyltransferase* (*UGT*) genes**
1006 **phylogenetically related to *PvfUGT1* and *PvfUGT2* in *P. vietnamensis* var.**
1007 ***fuscidiscus*.** Heatmap showing the expression profiles of *UGT* genes that cluster with
1008 the functionally characterized enzymes *PvfUGT1* and *PvfUGT2* in phylogenetic
1009 analysis. *PvfUGT1* and *PvfUGT2*, which have been previously demonstrated to
1010 catalyze the two sequential glycosylation steps leading to Majonoside R2 biosynthesis,
1011 are included as reference genes. Expression patterns of the phylogenetically related

1012 *UGTs* are shown across five tissues (root, stem, flower, leaf, and rhizome). Gene
1013 expression levels are presented as $\log_2$(TPM + 1).

1014 **Table S1. Summary of final reference transcripts assembled from integrated**
1015 **RNA-Seq and Iso-Seq data.**

1016 **Table S2. Functional annotation statistics of unigenes across tissues and**
1017 **transcriptomes based on EggNOG, GO, COG/KOG, and KEGG databases.**

1018 **Table S3. Candidate genes involved in triterpenoid saponin biosynthesis in *P.***
1019 ***vietnamensis* var. *fuscidiscus*, including protein annotations, gene names, and**
1020 **transcript abundance (TPM) across root, stem, flower, rhizome, and leaf.**