Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genome-guided prediction of UDP-glycosyltransferase candidates for triterpenoid saponin glycosylation in Psammosilene tunicoides

Zhenduo Fang
Aulin College, Northeast Forestry University
Renbo Zhang
Aulin College, Northeast Forestry University
Jing Yin
yinjing20135@163.com
College of Life Sciences, Northeast Forestry University

Research Article

Keywords: Psammosilene tunicoides, UDP-glycosyltransferase, triterpenoid saponins, RNA-seq reanalysis, candidate prioritization

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11596503/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Genome-guided prediction of UDP-glycosyltransferase candidates for triterpenoid saponin glycosylation in *Psammosilene tunicoides*

Zhenduo Fang[1], Renbo Zhang[1], Jing Yin[2,*]

1 Aulin College, Northeast Forestry University, 26 Hexing Road, Harbin 150040, China

2 College of Life Sciences, Northeast Forestry University, 26 Hexing Road, Harbin 150040, China

* Correspondence: Jing Yin; yinjing20135@163.com

Abstract

Background

Psammosilene tunicoides produces diverse triterpenoid saponins, including quillaic acid- and gypsogenin-derived compounds. UDP-glycosyltransferases (UGTs) are potential catalysts of their glycosylation, but the relevant enzymes remain incompletely characterized. Evaluating numerous UGT candidates against different triterpenoid acceptors and UDP-sugar donors is experimentally demanding. A newly available chromosome-level genome enables historical transcriptomic evidence to be connected to defined loci. We aimed to establish an evidence-based priority set for functional investigation of triterpenoid saponin glycosylation.

Results

We integrated genome-guided reanalysis of a published salicylic-acid elicitation time course with sequence-confidence assessment, Pareto prioritization and replicate-subset analysis. Among 123 retained PtUGT/UGT-like loci, standard annotation conversion covered 74; reconstructing counting-compatible exon features restored coverage for 49 additional loci. Among 62 Tier 1 sequence-confidence genes meeting strong differential-expression criteria, ten occupied Pareto Front 1. Four—*PtunChr12G250320*, *PtunChr8G185640*, *PtunChr11G239890* and *PtunChr8G179000*—retained Front 1 membership across all 27 overlapping balanced subsets. *PtunChr8G179000* belonged to the recovered loci, demonstrating the influence of counting coverage on candidate prioritization. The full-data consensus-rank-2 gene *PtunChr12G252960* retained membership in 18/27 subsets, separating high rank from persistence under changes in replicate composition.

In a separate 21-sample root dataset, *PtunChr12G250320* was higher in wild roots across all three comparisons, whereas *PtunChr11G239890* was higher in cultivated roots. *PtunChr8G185640* showed no significant individual wild-root contrast, and *PtunChr8G179000* showed a mixed pattern. Historical qRT-PCR records mapped to *PtunChr12G250320* with high confidence and *PtunChr8G179000* with medium locus-level confidence.

Conclusions

Annotation representation affected candidate coverage, while full-data rank and subset persistence supplied distinct prioritization information. The four stable-core candidates also differed in external root expression. Together with characterized UGT reference reactions and historical mappings, these findings establish a four-gene initial panel within a ten-gene expanded set for investigating triterpenoid saponin glycosylation through acceptor- and UDP-sugar-donor screening.

Keywords: *Psammosilene tunicoides*; UDP-glycosyltransferase; triterpenoid saponins; RNA-seq reanalysis; candidate prioritization

Background

Elucidating triterpenoid saponin glycosylation in *Psammosilene tunicoides* requires connecting its chemical diversity to candidate enzymes, a task now supported by a chromosome-level reference genome [1]. This medicinal member of the Caryophyllaceae accumulates oleanane-type saponins in its roots, including tunicosaponins and psammosaponins [2,3]. Quillaic acid and gypsogenin are major reported aglycone types, and the characterized saponins commonly carry sugar chains at C-3 and C-28 [3]. Identifying the enzymes that initiate and extend these glycans would help explain how the plant produces structurally diverse saponins.

Family-1 UDP-dependent glycosyltransferases (UGTs) are a major enzyme class responsible for transferring sugars to specialized metabolites and existing glycosides [4-6]. Biochemical studies in Medicago, Saponaria, soybean, Barbarea, licorice, Panax and Centella demonstrate distinct acceptor ranges, donor preferences and glycosylation steps among these enzymes [7-15]. Structural studies relate this diversity to the substrate-

binding environment [16,17], while the donor-specificity variants of licorice UGT73P12 illustrate how related sequences can support different reactions [13]. Saponin glycosylation can also involve enzymes outside family 1, including a cellulose-synthase-derived 3-O-glucuronosyltransferase [18]. The present study addresses the family-1/UGT-like repertoire as a tractable source of candidates for biochemical screening.

A central difficulty is deciding which candidates to assay first. Testing each UGT entails cloning or synthesis, heterologous expression, screening combinations of acceptors and UDP-sugar donors, and identifying the products. Achnine et al. used transcript and metabolite profiles to select Medicago glycosyltransferases for recombinant characterization [7]. In *Astragalus membranaceus*, screening reduced 517 glycosyltransferase-related unigenes to 24 cloned candidates before four enzymes were characterized [19]. A substrate-multiplexed platform profiled 85 Arabidopsis family-1 UGTs against 453 potential acceptors [20]. These studies show why a defined candidate set and biologically grounded assay panels are useful intermediates between sequence identification and enzyme characterization.

Previous *P. tunicoides* research has established both chemical and experimental starting points. Oleanane-type triterpenoid biosynthesis has been investigated in yeast [21], and historical studies reported UGT expression assays and biochemical records [22,23]. These records are valuable for interpreting new candidates, provided that historical names, cloned sequences and primer sets can be reconciled with modern genomic loci. The remaining gap is a family-wide, locus-resolved basis for deciding which additional UGTs merit initial experiments and how they relate to the previously studied enzymes.

Published SA elicitation experiments offer relevant transcriptional evidence for this purpose. Su et al. reported increased saponin accumulation and changes in quillaic acid and gypsogenin measured after acid hydrolysis of SA-treated hairy roots [24]. The accompanying RNA-seq series comprises a pretreatment baseline and 8- and 24-h samples, with three biological replicates per time point. The new reference permits these reads to be reassessed at defined loci. Annotation choice can affect RNA-seq quantification [25]; here, the distributed GFF contains gene, mRNA and CDS records but lacks explicit exon records. This raises a concrete coverage question before candidates can be compared. Other *P. tunicoides* transcriptomes examine

root rot [26] and wild-versus-cultivated roots [27], offering distinct contexts for interpreting the expression of selected loci.

Genome-wide UGT inventories and expression screens in other plants provide precedents for family-level candidate discovery [28-31], and integrated transcriptome-metabolome studies connect expression patterns to saponin biology [27,32]. Lineage-specific expansion and biochemical diversity complicate the choice of individual UGTs [33-35]. For *P. tunicoides*, candidate selection therefore requires a complete assessable repertoire and a way to distinguish a high full-data rank from persistence when the available replicate composition changes. Characterized UGT reactions and historical locus mappings can then inform experimental choices alongside the expression priorities.

Our objective was to predict UGT candidates potentially involved in triterpenoid saponin glycosylation in *P. tunicoides* and establish evidence-based priorities for their functional investigation. We used the chromosome-level reference and published SA time course to determine how annotation representation affected counting coverage and candidate inclusion. Within a predefined sequence-confidence and differential-expression universe, we distinguished full-data Pareto ranking from Front 1 persistence across balanced replicate subsets. We then examined the candidates using characterized UGT references, a separate root transcriptome and historical experimental records. Together, these analyses were designed to define an initial screening panel and an expanded candidate set with clear evidence for their selection.

Materials and methods

Genome reference and counting-compatible annotation representation

The chromosome-level *P. tunicoides* assembly GCA_054740895.1 was used as the mapping and gene-model reference [1]. The assembly comprises 14 pseudochromosomes and 30,924 annotated protein-coding genes. Genome FASTA, GFF, predicted protein and coding-sequence files were obtained from the genome-associated resources, and the correspondence between the author chromosome identifiers and the accession-style FASTA identifiers was verified before downstream analysis. Supporting datasets, plots and historical-evidence records are supplied in Additional files 1–8.

The source GFF contained 30,924 gene features, 30,924 mRNA features and 149,089 CDS features but no explicit exon features. A standard gffread -T conversion reconstructed 142,937 exon records covering 23,227 genes. To create a uniform exon-based representation for featureCounts, the same source GFF was converted with gffread 0.12.9 using -T --force-exons [36]. This counting-compatible Gene Transfer Format (GTF) file contained 161,504 exon records and provided exon features for all 30,924 genes and all 123 retained PtUGT/UGT-like loci, with no duplicate exon intervals or within-gene exon overlaps. Splice-site extraction yielded 130,335 unique genomic junction coordinates, 10,832 more than the standard-conversion catalog. The force-exons GTF and its splice-site catalog were used for the primary SA realignment and counting. The representation supplies counting features for the existing gene models. Annotation-representation statistics are supplied in Supplementary Table S1 (Additional file 1).

Historical SA RNA-seq preprocessing, alignment and quantification

The historical SA dataset comprised nine paired-end libraries representing a pretreatment 0-h baseline and 8 and 24 h after SA addition, with three biological replicates per time point [24]. The RNA-seq series did not include time-matched untreated or mock libraries at 8 or 24 h. Clean FASTQ inputs were generated with fastp 1.3.6 using paired-end adapter detection [37]. Reads were aligned to the chromosome-level reference with HISAT2 2.2.3 using --dta and the known-splice-site file derived from the counting-compatible representation [38]. Alignments were coordinate-sorted and indexed with samtools 1.24 [39].

Gene-level quantification used featureCounts from Subread 2.1.1 [40] with paired-end fragment counting (-p --countReadPairs), unstranded assignment (-s 0), exon features (-t exon) and gene_id aggregation (-g gene_id). All nine libraries were quantified against the same force-exons GTF. To isolate the effect of annotation representation, the same nine original alignment files were also quantified using the standard-conversion GTF and the corrected force-exons GTF with identical featureCounts settings. Assigned-fragment fractions were summarized as arithmetic means of the nine sample-specific percentages. A uniform-CDS count matrix generated from the same alignments was used for the representation sensitivity analysis.

Differential-expression analysis and apeglm effect-size estimation

Differential expression was analyzed in R 4.6.1 with DESeq2 1.52.0 [41]. Genes with total raw counts <10 across the nine SA libraries were excluded before the primary model. A design of ~ condition was fitted for the three time points. Pairwise contrasts were calculated for 8 h versus 0 h, 24 h versus 0 h and 24 h versus 8 h. Genes with Benjamini-Hochberg adjusted $P < 0.05$ and absolute $\log_2$ fold change ≥ 1 in at least one contrast were classified as strong-DE. A likelihood-ratio test (LRT) comparing the full ~ condition model with an intercept-only model (~1) quantified global temporal evidence.

Because low-count reference groups can yield high-variance maximum-likelihood estimates (MLEs), effect sizes used for prioritization were regularized with apeglm 1.34.0 [42]. Shrinkage was applied uniformly through coefficient-based DESeq2 lfcShrink calls. One fit used 0 h as the reference to obtain the shrunken 8 h versus 0 h coefficient; a second releveling used 8 h as the reference to obtain the shrunken 24 h versus 8 h coefficient. Contrast-based shrinkage, gene-specific censoring and pseudocount adjustments were not used. The 24 h versus 0 h contrast was retained for biological description but excluded from ranking because the corresponding unshrunken log-scale contrast is the sum of the two adjacent-time-point contrasts. This identity concerns the unshrunken contrasts from a common model; the separately regularized coefficients need not be exactly additive.

UGT family screening and confidence stratification

The predicted proteome was screened with HMMER 3.4 [43] using the PF00201 family-1 UGT Pfam model [44] and the model gathering threshold. Candidate proteins were independently evaluated for a strict plant secondary product glycosyltransferase (PSPG)-like sequence signature and for similarity to a curated set of 122 Arabidopsis family-1 UGT references using DIAMOND blastp [45]. The five best hits per query were reported at search-stage E-value $\leq 10^{-5}$; the highest-bitscore hit was retained, and final reference support required E-value $\leq 10^{-10}$ and query coverage $\geq 50\%$.

The PF00201 screen yielded 124 proteins, of which 123 met the final reference-support criteria. Tier 1 comprised retained proteins with a strict PSPG-like signature and an operational length window of 350-600 amino acids. Tier 2 comprised non-Tier-1, typical-length retained proteins with reference query coverage

≥70% and sequence identity ≥25%. All remaining retained proteins were assigned to Tier 3. These sequence-confidence assignments were defined before expression-based eligibility was assessed. Dependence on the 350-600-aa interval was evaluated across 36 alternative length windows (Supplementary Table S2B; Additional file 1). Each window recalculated sequence eligibility and recorded retention of the original Front 1 and subset-stable core members without refitting the expression or ranking models.

Pareto prioritization and replicate-stability analysis

Tier 1 genes meeting the strong-DE criterion formed the fixed formal prioritization universe. Three continuous dimensions were oriented so that larger values indicated stronger priority evidence: (i) $-\log_{10}$ of the LRT adjusted P value, (ii) the apeglm-shrunken $\log_2$ fold change for 8 h versus 0 h, and (iii) the apeglm-shrunken $\log_2$ fold change for 24 h versus 8 h. The absolute fold-change threshold made strong-DE eligibility bidirectional, whereas the signed ranking axes favored larger positive adjacent-time-point changes. Downregulated or transient profiles remained eligible and could reach the frontier through trade-offs among dimensions. The dimensions were not standardized or manually reweighted.

Pareto nondominated sorting assigned genes to successive fronts. A gene was dominated when another gene was at least as large on all three dimensions and strictly larger on at least one; Pareto Front 1 therefore represented the full-data nondominated candidate set without imposing a preselected shortlist size. For a complete ordering, each dimension was ranked in descending order with average ranks for ties, and the three component ranks were combined by their arithmetic mean. Median component rank, rank product and gene identifier were used sequentially as tie-breakers.

Stability to replicate composition was assessed across all 27 balanced 2+2+2 subsets obtained by removing one replicate from each time point. Each six-sample subset was refitted genome-wide with DESeq2, including size-factor, dispersion, LRT and pairwise-coefficient estimation, followed by apeglm shrinkage and re-ranking. Comparisons used the fixed full-data Tier 1 strong-DE universe. Front 1 membership frequency and scalar-rank frequency were recorded separately; the latter included the number of subsets in which a gene ranked first. Genes retained in Front 1 in all 27 subsets were designated the subset-stable core.

Because the subsets share libraries and eligibility was fixed, these frequencies quantify conditional stability within the available data.

Sensitivity analyses

Two principal ranking sensitivities were used. First, the same apeglm/Pareto procedure was applied to the complete retained strong-DE universe, irrespective of confidence tier, to determine whether lower-confidence genes entered Front 1. Second, the SA libraries were counted using uniform CDS features from the same primary alignments, and the full DESeq2, LRT, apeglm and Pareto workflow was repeated. A separate post hoc axis-deletion sensitivity retained the same 62-gene formal ranking universe and the original Pareto dominance, equal-weight rank aggregation and tie-breaking rules, but removed the early 8 h versus 0 h effect axis. Two-axis rankings were calculated separately using MLE and apeglm estimates of the remaining 24 h versus 8 h effect, together with LRT strength. This comparison used frozen estimates without refitting the models and was distinct from the primary three-axis apeglm analysis. Complete rankings and the verification script are provided in the repository under 09_representation_sensitivity/axis_deletion and 10_scripts/VERIFY_AXIS_DELETION.py.

Phylogenetic and PSPG context

The expanded phylogeny contained 123 retained *P. tunicoides* UGT/UGT-like proteins, 122 curated Arabidopsis UGTs and 11 experimentally characterized triterpenoid-related family-1 UGT references [7-11,13]. Amino-acid sequences were aligned with MAFFT using the L-INS-i strategy and trimmed with trimAl -automated1 [46,47]. Maximum-likelihood inference used IQ-TREE 3.1.3 [48], ModelFinder model selection [49], 1,000 ultrafast bootstrap replicates [50] and 1,000 SH-like approximate likelihood-ratio test (SH-aLRT) replicates. The fixed 256-tip tree was used for all subsequent phylogenetic summaries and was not re-estimated after candidate ranking.

For the main focused display, the fixed tree was pruned to the ten primary Front 1 candidates, three nearest tree neighbors per candidate and all 11 characterized reference UGTs, with duplicate neighbors merged, yielding 50 displayed tips. Suppressed unary branches were collapsed by summing branch lengths so that

pairwise patristic distances were preserved. Support labels were retained only for displayed nodes carrying the corresponding label in the fixed tree; labels from suppressed nodes were not inherited, and no support value was recomputed. For each Front 1 candidate, the minimum patristic distance to an experimentally characterized reference was extracted from the same fixed tree. PSPG-region alignments were used only as sequence-classification context and were not interpreted as evidence of acceptor, donor or regioselective specificity. The 50-tip circular display is shown in Fig. 4A; its numerical supports are provided in Supplementary Fig. S1, page 6. The complete fixed tree is provided in Supplementary Fig. S1, pages 1–5, and the PSPG-region display in Supplementary Fig. S3 (Additional file 4).

Duplication and pathway-homolog context

Whole-proteome similarity searches and MCScanX [45,51] were used to classify retained PtUGTs as whole-genome duplication (WGD)/segmental, tandem, proximal or dispersed duplicates. Duplication categories were interpreted descriptively for the primary Front 1 set and were not used for ranking.

For pathway context, the 30,924 predicted *P. tunicoides* proteins were searched against 16,421 reviewed Arabidopsis Swiss-Prot proteins with DIAMOND 2.2.6 using --sensitive, search-stage E-value $\leq 10^{-5}$ and up to five targets per query. Hits assigned to triterpenoid-related labels were retained as sequence-similarity homolog candidates when E-value $\leq 10^{-20}$ and query coverage $\geq 50\%$, keeping the highest-bitscore hit for each gene-by-pathway combination. Main physical-context analyses focused on squalene synthase (SQS), squalene epoxidase (SQE), oxidosqualene cyclase (OSC)/beta-amyrin-like, CYP716A and CYP72A candidates. These labels indicate similarity to Arabidopsis references and are not orthology or functional assignments.

For each retained PtUGT, the nearest same-chromosome pathway-homolog candidate was identified by midpoint-to-midpoint distance. The observed mean nearest distance for the primary Front 1 set was compared with a null distribution generated by 100,000 random samples of the same size from the 123 retained PtUGTs (seed 20260920). A one-sided permutation P value tested whether Front 1 genes were unusually close to pathway-homolog candidates.

External wild-versus-cultivated root transcriptome

BioProject PRJNA937778 [27] supplied a separate root transcriptome comprising wild plants (WP, n=3) and cultivated plants (CP) from three age groups (1CP, 2CP and 3CP; n=6 each), for 21 samples. All paired-end libraries were realigned with HISAT2 2.2.3 using the counting-compatible splice-site set and quantified together against the force-exons GTF with featureCounts 2.1.1 using paired-end reverse-strand counting (-s 2).

DESeq2 1.52.0 was used for 1CP versus WP, 2CP versus WP and 3CP versus WP contrasts and a global group LRT. Three leave-one-WP-out refits assessed dependence on individual wild samples. This analysis was conducted after the SA-derived ranking and supplied cross-context expression summaries without changing the candidate universe, Pareto fronts or consensus ranks.

Mapping previous experimental evidence to modern loci

Previous *P. tunicoides* UGT qRT-PCR primers and cloned UGT sequences were mapped to the 2026 genome by exact primer/sequence matching and uniqueness checks [22,23]. Single-primer matches in the CDS set and genome were evaluated separately from paired loci satisfying orientation and amplicon-length criteria. A mapping was classified as high confidence when the available primer pair or previous sequence uniquely supported a modern locus; cases with incomplete transcript-boundary correspondence were retained with an explicit lower-confidence qualifier. Previous expression evidence and biochemical assays were recorded separately so that qRT-PCR concordance was not conflated with enzyme-function validation. The complete 14-locus historical-evidence record, including unmatched entries and non-Front-1 loci, is provided in Supplementary Table S9 (Additional file 8).

The available historical UGT1 protein sequence supported assignment to *PtunChr11G235070*, a Tier 2 locus outside the primary frontier [23].

Statistical reporting

Multiple-testing correction used the Benjamini-Hochberg procedure unless otherwise specified. Front 1 membership and scalar ranks were reported as distinct outputs.

AI-assisted tools

During the preparation of this manuscript, the authors used AI-assisted tools for statistical analysis, code development, figure preparation, and language polishing. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

Results

Genome-resolved PtUGT repertoire and restored counting coverage

The chromosome-level genome provided a common reference for family identification and RNA-seq quantification. It contains 30,924 predicted protein-coding genes across 14 pseudochromosomes [1], but its source GFF has no explicit exon records. Standard gffread -T conversion reconstructed 142,937 exons covering 23,227 genes. Within the 123 retained PtUGT/UGT-like loci, this representation covered 74 and left 49 outside exon-based counting. The force-exons representation generated 161,504 exon records, covered all 30,924 genes and all 123 retained PtUGTs, and added 10,832 unique splice-junction coordinates. The recovered counting features allowed these 49 existing loci to enter expression assessment.

The nine historical SA libraries were realigned using the counting-compatible splice-site catalog. HISAT2 overall alignment rates ranged from 92.54% to 95.60%. Using identical alignments, the corrected exon representation increased the mean assigned-fragment fraction by approximately 3.93 percentage points and reduced NoFeatures fragment assignments by approximately 1.23 million per sample. Realignment using the same corrected GTF changed the mean assigned-fragment fraction by approximately −0.007 percentage points, giving an overall improvement of approximately 3.92 percentage points from the original to the final alignment and counting workflow. Genome-wide strong-DE counts were 5,799 for 8 h versus 0 h, 5,264 for 24 h versus 0 h and 4,573 for 24 h versus 8 h (adjusted $P < 0.05$ and $|\log_2 FC| \geq 1$). Alignment quality, sample structure and strong-DE counts are summarized in Fig. 1B–D.

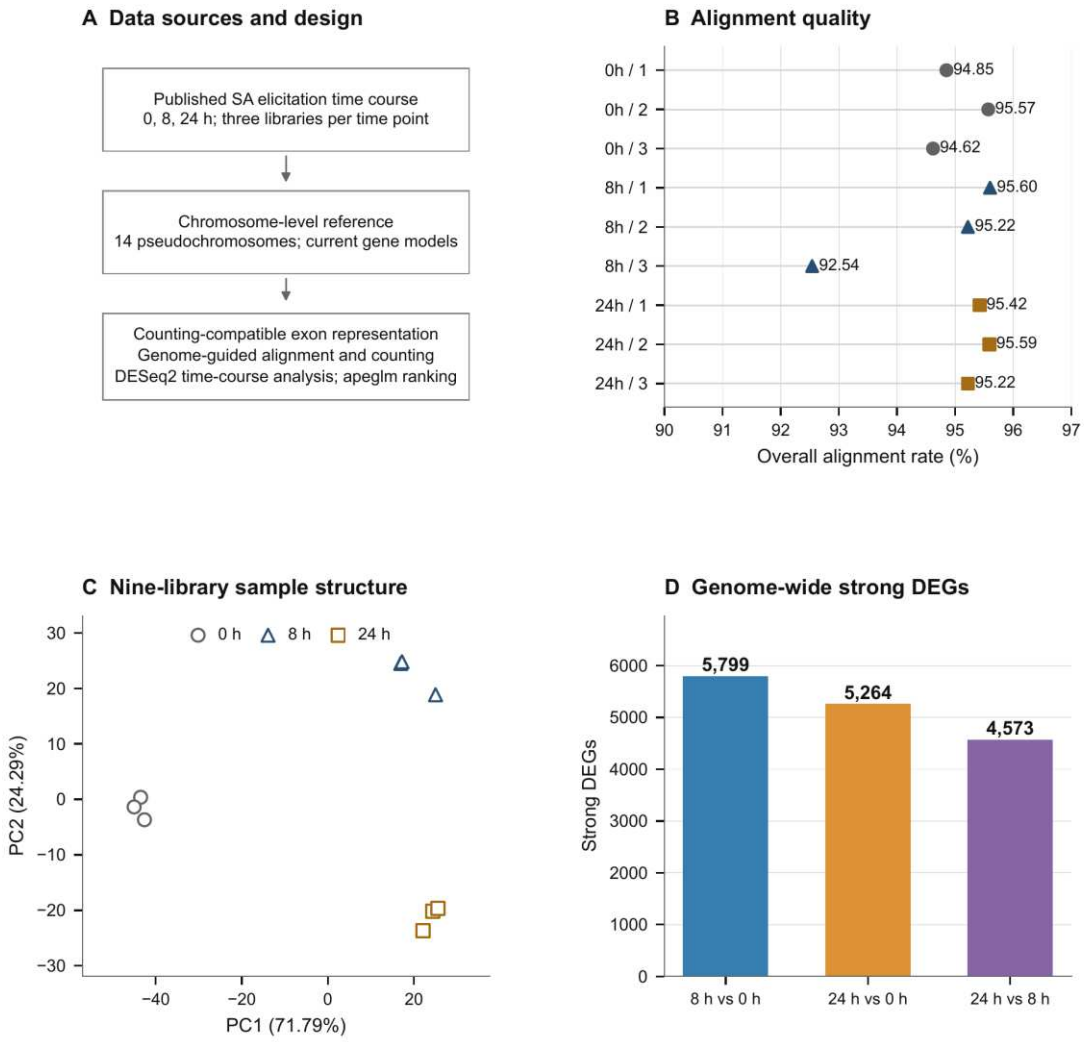


Figure 1. Data sources, analysis design and quality control. (A) Published SA elicitation libraries and the chromosome-level reference used for genome-guided reanalysis with a counting-compatible exon representation. There were three libraries at each of 0, 8 and 24 h [24]. (B) HISAT2 overall alignment rates for the nine libraries (92.54–95.60%). (C) Sample structure from the primary count-matrix principal component analysis (PCA). (D) Genome-wide strong-DE counts: 5,799 for 8 h versus 0 h, 5,264 for 24 h versus 0 h and 4,573 for 24 h versus 8 h (adjusted P <0.05 and absolute unshrunken $\log_2$ fold change ≥1). The three bar colors identify contrasts, not individual time points.

PF00201 screening identified 124 candidate proteins, one of which failed final curated-reference support, leaving 123 retained PtUGT/UGT-like proteins. Eighty-eight carried strict PSPG-like signatures. The confidence framework assigned 84 proteins to Tier 1, 31 to Tier 2 and 8 to Tier 3. Four strict-PSPG proteins remained outside Tier 1 because their lengths fell outside the operational 350-600-aa interval. Confidence-tier counts are shown in Fig. 2A and protein-length distributions in Supplementary Fig. S5A (Additional file 6).

The 123 retained loci were distributed across all 14 pseudochromosomes. Figure 2B locates each gene and distinguishes the ten Front 1 candidates and four subset-stable core genes. Supplementary Fig. S5B (Additional file 6) places the family alongside sequence-similarity candidates for major triterpenoid pathway classes.

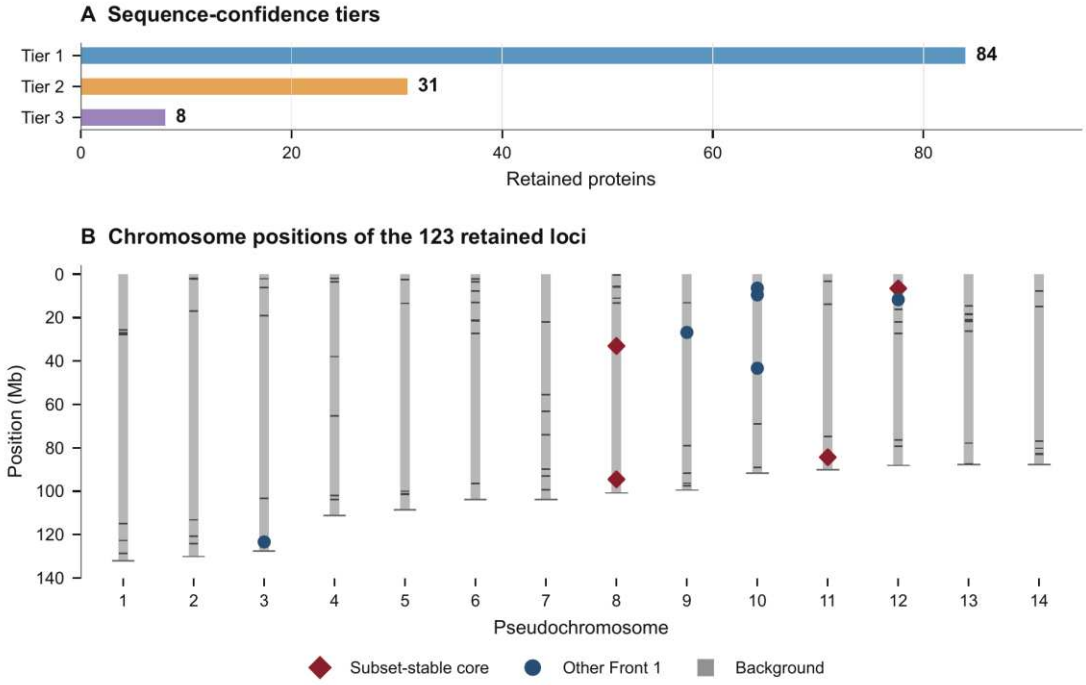


Figure 2. Confidence tiers and chromosome positions of retained PtUGT loci. (A) Sequence-confidence categories of the 123 retained proteins: 84 Tier 1, 31 Tier 2 and 8 Tier 3. Blue, orange and purple identify the tiers and match Supplementary Fig. S5A. (B) Genomic midpoint positions in Mb on the 14 pseudochromosomes; chromosome extents come from the reference FASTA index. Deep-red diamonds mark the four subset-stable core loci, deep-blue circles the six other Front 1 loci, and gray marks background loci. Protein lengths, pathway-homolog context, local clusters and physical-proximity sensitivity are shown separately in Supplementary Fig. S5A–E.

Temporal expression defines the formal candidate universe

Under the force-exons count matrix, 111 of the 123 retained PtUGTs passed the row-sum expression filter, comprising 76 Tier 1, 27 Tier 2 and 8 Tier 3 genes. Eighty-three were strongly differentially expressed in at least one time-point contrast: 62 Tier 1, 16 Tier 2 and 5 Tier 3. The 62 Tier 1 strong-DE genes formed the fixed formal ranking universe. The filtering hierarchy is shown in Fig. 3A, and complete expression profiles are provided in Supplementary Fig. S6 (Additional file 7).

Multi-objective priorities separate full-data rank from subset persistence

Multi-objective ranking identified ten primary Front 1 genes: *PtunChr12G250320*, *PtunChr12G252960*, *PtunChr8G185640*, *PtunChr10G208700*, *PtunChr10G207680*, *PtunChr11G239890*, *PtunChr8G179000*, *PtunChr3G078100*, *PtunChr10G212180* and *PtunChr9G191360* (Table 1; Fig. 3). These genes represented nondominated trade-offs among global temporal evidence and the two regularized adjacent-time-point effects. Equal-weight rank aggregation placed *PtunChr12G250320* first, followed by *PtunChr12G252960* and *PtunChr8G185640*. Profiles with weak or negative change in one interval could remain on Front 1 when other dimensions were stronger. *PtunChr8G179000*, which lacked countable exons under standard conversion, entered this frontier after its counting coverage was restored. Candidate trajectories are shown in Fig. 3B.

Across the 27 balanced subsets, Front 1 contained 6-11 genes (median 8), and 16 genes reached it at least once. Four retained membership in all 27 subsets: *PtunChr12G250320*, *PtunChr8G185640*, *PtunChr11G239890* and *PtunChr8G179000*. These genes defined the subset-stable core. Within the separate scalar-rank output, *PtunChr12G250320* ranked first in 20/27 subsets, with a rank range of 1-3. *PtunChr12G252960*, the full-data rank-2 gene, retained Front 1 membership in 18/27 subsets. Figure 3C reports membership frequencies for the ten primary candidates and six additional genes that reached Front 1 in at least one subset.

The core candidates also differed in their temporal effects. *PtunChr12G250320* and *PtunChr8G185640* had larger positive early than late effects, with apeglm $\log_2$ fold changes of 9.90/1.41 and 5.79/0.41 for 8 h versus 0 h/24 h versus 8 h, respectively. *PtunChr11G239890* had a negative early estimate (−0.68) followed by a positive late estimate (4.05), whereas *PtunChr8G179000* had little estimated early change (0.05) and a larger late effect (4.09). The continuous estimates and uncertainties are supplied in Supplementary Table S4 (Additional file 1).

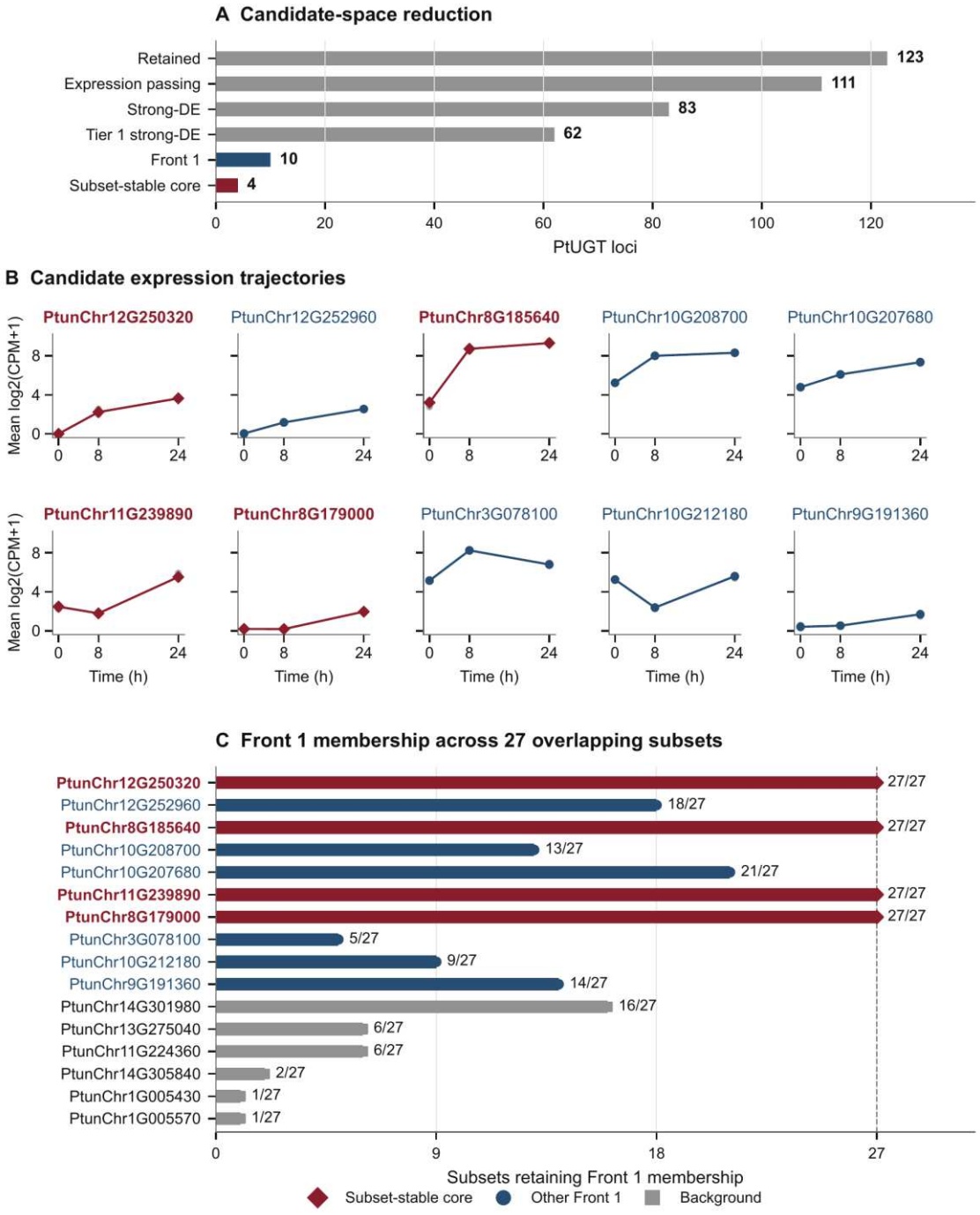


Figure 3. Candidate reduction, expression trajectories and internal ranking stability. (A) Candidate hierarchy: 123 retained loci, 111 expression-filter-passing loci, 83 strong-DE loci, 62 Tier 1 strong-DE loci, ten full-data Pareto Front 1 candidates and four subset-stable core loci. (B) Expression trajectories of the ten Front 1 candidates in full-data consensus-rank order. Lines connect the original means of $\log_2$(CPM+1) across three biological replicates at each time point, where CPM denotes counts per million. Grey points show individual replicate values. Deep red denotes subset-stable core genes and deep blue other Front 1 genes. (C) Front 1 membership counts across all 27 overlapping balanced 2+2+2 replicate subsets, conditional on the fixed 62-gene universe. Six background genes reaching Front 1 at least once are also shown. Four genes retained Front 1 membership in 27/27 subsets; *PtunChr12G252960* did so in 18/27. Complete nine-library row-standardized expression profiles for all 111 loci are retained in Supplementary Fig. S6.

Effect-size regularization left the leading candidates unchanged. Relative to maximum-likelihood estimates, *PtunChr11G224360* moved from Front 1 to Front 2 while retaining consensus rank 16. Across the 62-gene universe, the MLE and apeglm orderings remained nearly identical (Spearman rank correlation = 0.9947), indicating that shrinkage altered a boundary classification rather than restructuring the full ordering. The all-strong-DE sensitivity returned exactly the same ten Front 1 genes and admitted no Tier 2 or Tier 3 gene. Uniform-CDS counting produced a 63-gene ranking universe, compared with 62 under exon counting, and again yielded ten Front 1 genes with *PtunChr12G250320* at rank 1. Nine of the ten primary Front 1 genes overlapped between representations; *PtunChr3G078100* was exon-only and *PtunChr13G275040* was CDS-only (Supplementary Fig. S4; Additional file 5).

PtunChr12G250320 had zero raw counts in all three 0-h libraries. Its early post-treatment effect was positive under both MLE and apeglm estimation, but the absence of baseline reads limited the precision of its magnitude. Removing the early-effect axis moved *PtunChr12G250320* from Front 1 and consensus rank 1 to Front 2 and rank 3 under both MLE- and apeglm-based two-axis rankings. Under the primary three-axis apeglm analysis, it ranked first in the full dataset and retained Front 1 membership in all 27 overlapping balanced subsets. Its stability to replicate composition therefore differed from its sensitivity to deleting a ranking axis.

Characterized UGT references and genomic context of the candidates

The fixed 256-protein tree placed the ten Front 1 genes in several phylogenetic neighborhoods. Their nearest characterized references were UGT74M1 from *Saponaria vaccaria* for five candidates, UGT73C12 from *Barbarea vulgaris* for three, and UGT91H4 from *Glycine max* for two. Figure 4A displays the 50-tip tree pruned from this fixed topology; Fig. 4B reports nearest-reference distances. Numerical supports are retained in Supplementary Fig. S1 (Additional file 2), and reference reactions and evidence qualifications are listed in Supplementary Table S5 (Additional file 1).

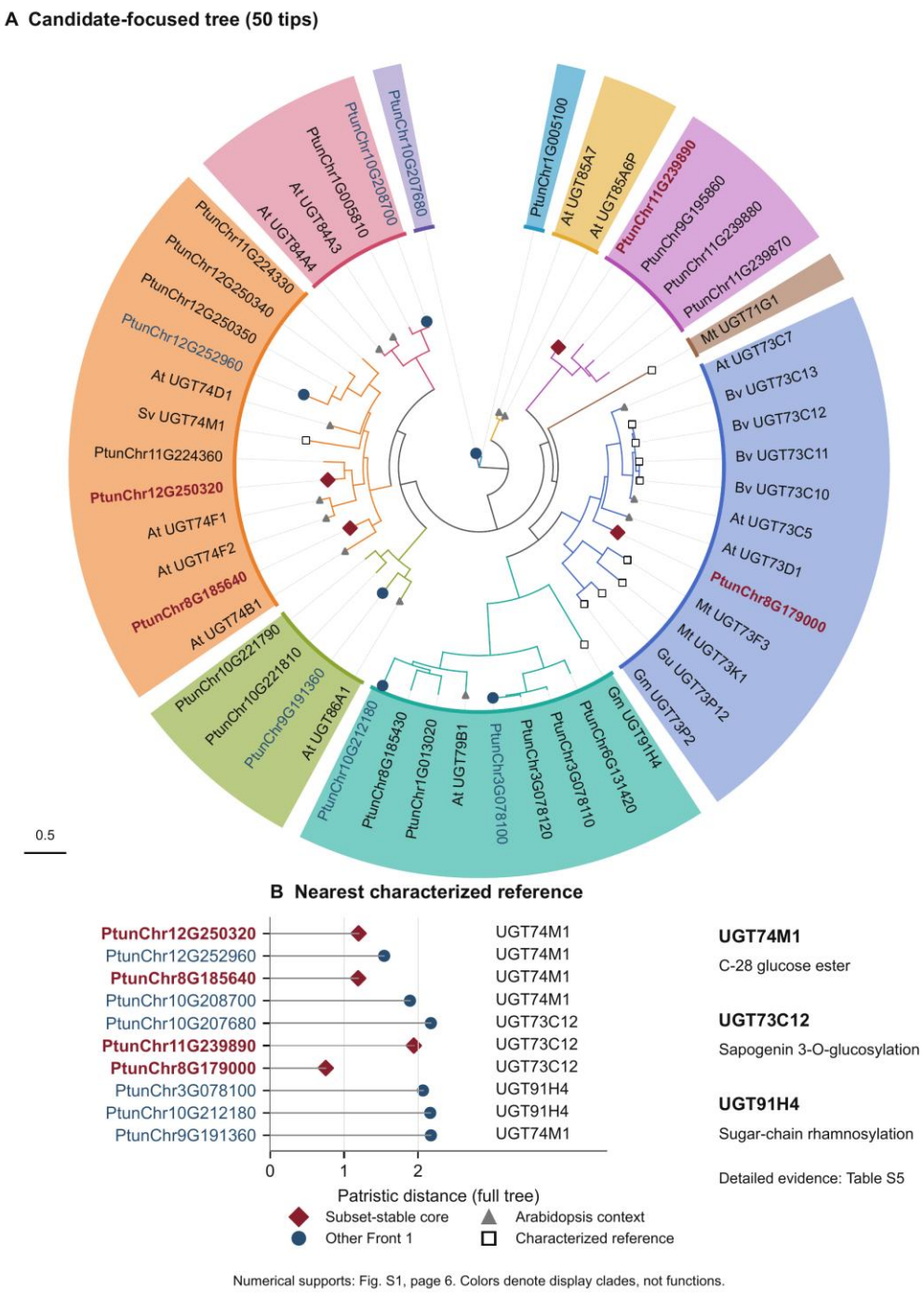


Figure 4. Phylogenetic context and nearest characterized references. (A) Circular 50-tip focused tree pruned from the fixed 256-tip maximum-likelihood tree without re-estimation or rerooting. It comprises ten Front 1 candidates, three nearest neighbors per candidate (duplicates merged), and all 11 characterized references. Unary-branch lengths were summed to preserve patristic distances. Colored sectors distinguish display clades, not UGT subfamilies or functions. Numerical SH-aLRT/ultrafast bootstrap (UFBoot) supports for displayed nodes are retained in Supplementary Fig. S1, page 6; suppressed-node supports were not transferred. (B) Nearest characterized references and patristic distances in the full tree. Reaction labels describe reference enzymes, whose biochemical evidence is listed in Supplementary Table S5. At, *Arabidopsis thaliana*; Sv, *Saponaria vaccaria*; Bv, *Barbarea vulgaris*; Mt, *Medicago truncatula*; Gm, *Glycine max*; Gu, *Glycyrrhiza uralensis*. Branch-length scale bars indicate substitutions per site.

The ten Front 1 genes comprised four WGD/segmental, four dispersed and two tandem duplicates. *PtunChr11G239890* belonged to a local tandem group with *PtunChr11G239880*. Duplication composition and gene-level classifications are provided in Supplementary Table S6 (Additional file 1).

The observed mean nearest same-chromosome distance from Front 1 genes to sequence-similarity candidates for SQS, SQE, OSC/beta-amyrin-like, CYP716A or CYP72A was 18.47 Mbp (median 7.95 Mbp). The random mean was 12.97 Mbp across 100,000 size-matched samples from the 123 retained PtUGTs. The one-sided test for unusually small distances gave P=0.8463, providing no evidence of physical-proximity enrichment. Individual distances and the permutation comparison are shown in Supplementary Fig. S5D and S5E (Additional file 6).

Root expression and historical records distinguish the core candidates

Reprocessing the separate 21-sample PRJNA937778 dataset generated a unified 30,924-gene count matrix and provided counts for all ten primary Front 1 genes. Five candidates showed mixed significance or group-dependent patterns, three had no significant pairwise wild–cultivated contrast, one was consistently wild-high across the three cultivated groups, and one was consistently cultivated-high. The four subset-stable core genes spanned all four categories. These summaries were obtained after candidate selection and did not change the SA-derived priorities. External sample quality control is shown in Supplementary Fig. S2 (Additional file 3), and candidate contrasts in Fig. 5A.

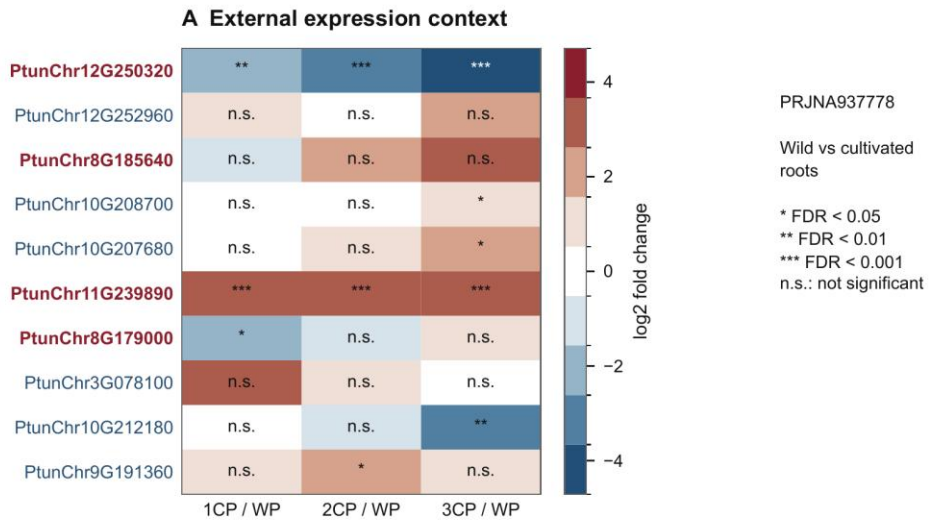
PtunChr12G250320 was consistently wild-high. Mean normalized expression was 4,685.8 in WP and 765.4, 678.2 and 179.0 in 1CP, 2CP and 3CP, respectively. Cultivated-versus-WP $\log_2$ fold changes were −2.61, −2.79 and −4.71, with adjusted P values of 1.50×10^{-3}, 6.20×10^{-4} and 4.47×10^{-10}. *PtunChr11G239890* showed the opposite group association: its cultivated-versus-WP $\log_2$ fold changes were 2.78, 3.12 and 3.01, with adjusted P values of 7.28×10^{-4}, 1.17×10^{-4} and 1.27×10^{-4}. For both genes, all leave-one-WP-out refits retained the direction of each contrast.

PtunChr8G185640 had a significant global group LRT (adjusted P=0.0143), but its individual 1CP, 2CP and 3CP versus WP adjusted P values were 0.718, 0.193 and 0.0613, respectively. *PtunChr8G179000* was wild-high in 1CP versus WP ($\log_2$ fold change −2.54; adjusted P=0.0280); its 2CP and 3CP contrasts were

nonsignificant (adjusted P=0.617 and 0.300). Thus, shared membership in the subset-stable core accompanied distinct patterns in the separate root comparison (Table 1; Supplementary Table S8A, Additional file 1).

Historical qRT-PCR records mapped to two core loci. The UGT74-2 primer pair from Li [22] mapped uniquely to *PtunChr12G250320*, with both primers on the same current transcript, giving high-confidence correspondence in the wild-versus-cultivated root context. The UGT73-2 forward primer matched one CDS, assigned to *PtunChr8G179000*. The forward primer had two matches in the full genome, but only one paired genomic locus met the orientation and amplicon-length criteria. At this locus, the reverse primer lay beyond the current annotated 3-prime boundary. This was retained as medium-confidence locus-level mapping. Figure 5B summarizes these two expression mappings alongside three historical biochemical records for other loci; all 14 checked records and their evidence qualifications are retained in Supplementary Table S9.

Historical biochemical records mapped to loci outside the primary Front 1. The available UGT1 protein sequence from Meng [23] supported assignment to *PtunChr11G235070*, a Tier 2 reference-supported noncanonical locus. The study reported yeast expression with UDP-glucose and quillaic acid or gypsogenin, followed by HPLC/LC-MS product analysis. Li reported cloning, prokaryotic expression and TLC-based reaction changes for ptUGT74E2 and ptUGT74B1, mapped here to *PtunChr3G078310* and *PtunChr5G117610*, respectively [22]. These TLC observations provide preliminary reaction evidence, without identifying the products or their glycosylation positions. These records concern different loci from the ten current frontier candidates (Fig. 5B; Supplementary Table S9; Additional file 8).



B Historical mappings and experimental records

Modern locus	Previous name	qRT-PCR mapping	Biochemical assay
PtunChr12G250320	UGT74-2	High confidence	-
PtunChr8G179000	UGT73-2	Locus-level (medium)	-
PtunChr3G078310	ptUGT74E2	Not applicable	Expression; TLC
PtunChr11G235070	UGT1	Not applicable	Yeast; HPLC/LC-MS
PtunChr5G117610	ptUGT74B1	Not applicable	Expression; TLC

Figure 5. External expression context and historical experimental evidence. (A) $\log_2$ fold changes for cultivated-root groups versus wild plants (WP) in PRJNA937778. WP has three libraries; 1CP, 2CP and 3CP each have six. Blue and red indicate lower and higher expression in cultivated groups relative to WP, respectively. Significance marks denote adjusted P values: * <0.05, ** <0.01 and *** <0.001; n.s., not significant. Subset-stable core candidates are highlighted in bold red; other Front 1 candidates are blue. These contrasts were not used to select or re-rank the SA-derived candidates. (B) Historical records mapped to five modern loci. qRT-PCR mapping confidence is high for UGT74-2 and medium at locus level for UGT73-2. Biochemical records concern three non-Front-1 loci: UGT1 has reported HPLC/LC-MS evidence; ptUGT74E2 and ptUGT74B1 have preliminary TLC evidence. A dash indicates that no corresponding historical biochemical assay was identified. All 14 locus records and evidence qualifications are provided in Supplementary Table S9 (Additional file 8).

Table 1

Primary Pareto Front 1 candidates and complementary evidence summaries.

Candidate	Consensus rank	Front 1 in 27 subsets	Subset-stable core	External root pattern	Previous evidence
PtunChr12G250320	1	27/27	Yes	wild-high in all three contrasts	High-confidence previous qRT-PCR mapping
PtunChr12G252960	2	18/27	No	No significant pairwise contrasts	No direct previous qRT-PCR match established
PtunChr8G185640	3	27/27	Yes	No significant pairwise contrasts	No direct previous qRT-PCR match established
PtunChr10G208700	5	13/27	No	mixed	No direct previous qRT-PCR match established
PtunChr10G207680	6	21/27	No	mixed	No direct previous qRT-PCR match established
PtunChr11G239890	7	27/27	Yes	cultivated-high in all three contrasts	No direct previous qRT-PCR match established
PtunChr8G179000	12	27/27	Yes	mixed	Medium-confidence previous qRT-PCR mapping
PtunChr3G078100	13	5/27	No	No significant pairwise contrasts	No direct previous qRT-PCR match established
PtunChr10G212180	15	9/27	No	mixed	No direct previous qRT-PCR match established
PtunChr9G191360	18	14/27	No	mixed	No direct previous qRT-PCR match established

External root patterns summarize the PRJNA937778 wild-versus-cultivated contrasts. “No significant pairwise contrasts” refers to the three individual cultivated-versus-wild comparisons; *PtunChr8G185640* nevertheless had a significant global group LRT (adjusted P=0.0143). Historical qRT-PCR mappings report expression evidence; biochemical records are recorded separately in Supplementary Table S9 (Additional file 8). Consensus ranks are calculated across all 62 eligible genes, so ranks within this ten-member frontier are not consecutive. The three ranking dimensions, including LRT strength and the two apeglm-shrunken $\log_2$ fold changes, are reported for each gene in Supplementary Table S4 (Additional file 1).

Discussion

Priorities for investigating triterpenoid saponin glycosylation

The biochemical diversity of *P. tunicoides* saponins creates a candidate-selection problem: multiple UGTs could act on different aglycones, glycosylated intermediates and UDP-sugar donors. We identified four subset-stable PtUGT candidates within a ten-gene Pareto priority set for biochemical investigation of triterpenoid saponin glycosylation. These priorities guide functional investigation without assigning specific catalytic reactions to individual candidates. Recovering counting coverage made the retained family available for expression assessment; Pareto analysis selected ten genes with complementary temporal evidence; and replicate-subset analysis identified four persistent frontier members. The separate root dataset then distinguished their expression backgrounds. Mapping historical names and assay records to modern loci enables those leads to be reused alongside the new expression priorities.

The SA series has a biological rationale for this task because the original study examined elicitation of saponin accumulation and measured quillaic acid and gypsogenin after acid hydrolysis [24]. Reanalysis connects that published transcriptional resource to modern loci. Its expression evidence is useful for choosing enzymes to investigate in the saponin context; the chemical reactions to test are informed separately by *P. tunicoides* phytochemistry and characterized UGTs.

Counting coverage changes which candidates can be evaluated

Annotation representation affected candidate selection before ranking began. The standard conversion supplied countable exons for 74 of 123 retained PtUGTs, whereas the force-exons representation covered the whole retained family. Genome-wide, coverage increased from 23,227 to 30,924 gene models. Annotation-dependent quantification has been documented in other systems [25]; here, the consequence can be traced to a selected locus. Once its counting features were restored, *PtunChr8G179000* entered Front 1 and retained membership across all 27 balanced subsets. A candidate screen restricted to the standard-conversion feature set would have omitted it.

The recovery concerns counting features for existing gene models. The genome report already lists 161,504 exons [1], and the present work makes the distributed models usable in a uniform counting workflow. *PtunChr8G179000* also remained on the frontier under uniform-CDS counting, linking the coverage correction to a candidate retained under both representations. This example shows why family-level coverage checks matter when a new genome is used to reinterpret historical transcriptomes.

Full-data rank and subset persistence guide different choices

The three-axis frontier preserves trade-offs among global temporal evidence and the early and late signed expression effects. Consensus rank supplies a complete ordering, while subset membership measures persistence when one replicate per time point is removed. *PtunChr12G252960* had full-data consensus rank 2 but retained Front 1 membership in 18/27 subsets; *PtunChr8G179000* had consensus rank 12 and retained membership in 27/27. Selecting the four highest scalar ranks would therefore give a different initial panel from selecting persistent frontier members. The four-gene core makes this second criterion available for allocating the first round of assays.

PtunChr12G250320 illustrates a further distinction between changing samples and changing ranking objectives. It remained on Front 1 in all 27 subsets and ranked first in 20, yet deleting the early-effect axis moved it to Front 2 and rank 3 under both MLE and apeglm estimates. Thus, its leading position depends partly on retaining early temporal evidence, rather than solely on the estimator. The zero counts in all three baseline libraries warrant caution about the early effect's magnitude.

An experimental study in *Saponaria vaccaria* illustrates why sequence quality and expression clues should be considered together [52]. The co-induced candidate PB.41560.2 lacked the expected galactosyltransferase activity, whereas the longer coding variant PB.1747.1, which was not co-induced, was active. This supports considering complementary selection evidence alongside induction and phylogenetic proximity. The cellulose synthase-like glucuronosyltransferase characterized in the same study also places the present family-1/UGT-like screen within a broader saponin-glycosylation pathway [18,52].

The four core candidates have distinct supporting evidence

All four core genes meet Tier 1 sequence criteria and retain Front 1 membership across the balanced subsets, but their early and late expression effects and external root patterns differ. *PtunChr12G250320* combines the leading full-data rank with a strong positive early effect and a further positive late effect. Its consistently higher expression in wild roots and high-confidence mapping to the historical UGT74-2 qRT-PCR assay connect the modern locus to root expression evidence. Its nearest characterized reference is UGT74M1, making ester-forming triterpene glucosylation one reaction class to include in an initial screening panel.

PtunChr8G185640 also has a larger positive early effect than late effect and is nearest UGT74M1 in the fixed tree. In the separate root dataset, the global group LRT was significant, although none of the three individual cultivated-versus-wild contrasts passed the adjusted-P threshold. Its rationale for initial testing therefore rests chiefly on its sequence-confidence assignment and persistent SA-series priority.

PtunChr11G239890 contributes a different temporal profile: a negative early estimate followed by a strong positive late estimate. It was consistently higher in cultivated roots, in contrast to *PtunChr12G250320*. Its nearest characterized reference is UGT73C12. These observations identify a candidate with persistent prioritization under subsampling but a different root expression association, useful for examining whether contrasting expression contexts accompany different biochemical activities.

PtunChr8G179000 links restored counting coverage to a late-dominated temporal response and persistent frontier membership. It is also nearest UGT73C12, and its historical UGT73-2 qRT-PCR mapping provides medium-confidence locus correspondence. The external dataset showed higher expression in wild roots only in the 1CP comparison, with nonsignificant 2CP and 3CP contrasts. Before reusing the historical primers, its annotated 3-prime boundary should be checked.

Li et al. [27] reported greater overall triterpenoid accumulation in wild than cultivated *P. tunicoides* roots. The wild-high expression of *PtunChr12G250320* agrees with this group-level background, whereas *PtunChr11G239890* shows the opposite direction. Together with the nonsignificant individual contrasts for *PtunChr8G185640* and mixed behavior of *PtunChr8G179000*, these results show that the four subset-stable priorities capture distinct root-expression associations. These differences can guide sampling for follow-up

expression and metabolite work. Historical qRT-PCR mappings add continuity with previous research, with their mapping confidence retained; independence of those historical records from the external cohort has not been established.

Saponin chemistry and reference reactions guide assay panels

P. tunicoides phytochemistry supplies a concrete starting point for acceptor selection. Structural studies reported gypsogenin-derived saponins with C-3 glucuronic-acid-based chains and, in some compounds, C-28 chains containing fucose, rhamnose, glucose and xylose [53]. The broader quillaic acid- and gypsogenin-derived saponin repertoire [2,3] therefore poses several reaction questions: initial aglycone glycosylation, C-28 ester formation and extension of existing glycans. Quillaic acid and gypsogenin are relevant initial aglycones, supplemented by available, structurally defined glycosylated intermediates. The chemical structures identify reaction classes to investigate; the present locus-resolved priorities identify genes to enter those assays.

The nearest characterized references help specify comparison reactions. UGT74M1 uses UDP-glucose for triterpene carboxylic-acid glucosylation, with the gypsogenin C-28 glucose ester directly supported by NMR [8]. UGT73C12 glucosylates sapogenins including oleanolic acid and hederagenin using UDP-glucose; its additional products distinguish it from the more strictly 3-O-selective UGT73C10 and UGT73C11 [11,12]. The two non-core frontier genes nearest UGT91H4 extend the experimental scope: this soybean enzyme transfers rhamnose from UDP-rhamnose to soyasaponin III to form soyasaponin I, acting on an existing glycoside [10]. That reaction is a reference for sugar-chain-extension assays, rather than evidence that soyasaponin I is a *P. tunicoides* product.

These references motivate an initial UDP-glucose panel and extension to alternative donors and glycosylated acceptors as warranted by the relevant reaction classes. UGT73P12 transfers glucuronic acid from UDP-glucuronic acid to glycyrrhetinic acid 3-O-monoglucuronide, illustrating extension of an existing glycan [13]. Acceptor choice and donor choice must be tested together: a negative result with an aglycone alone leaves glycoside-dependent activity unexplored. Structural and functional studies explain why related UGTs can differ in substrate recognition and regioselectivity [16,17,34,54]. Phylogenetic position and PSPG

conservation organize these experimental hypotheses, while reaction specificity is determined from the assay products.

Acceptor modification and donor choice can be tested in stages. In coupled in vitro assays, Hudson et al. [55] found that QsFucT and SvFucT accepted unglycosylated quillaic acid, showing that prior C-3 glycosylation was not an absolute acceptor requirement. The enzymes preferentially transferred 4-keto-6-deoxy-D-glucose from its UDP donor, with subsequent reduction producing D-fucose. This supports comparing free aglycones with relevant glycosides and selecting donor panels by reaction class, rather than inferring the UDP-sugar directly from the final sugar residue. Specialized donor systems can follow initial screens when the target chemistry warrants them.

Genomic context serves a separate descriptive role. The proximity test provided no evidence that frontier genes lie unusually close to the selected pathway-homolog candidates, and the ten genes had heterogeneous duplication histories. Although tandem UGT duplication has been linked experimentally to neofunctionalization in other plants [56], the current chromosome and duplication summaries do not establish such a mechanism for the PtUGTs.

Historical enzymes provide comparisons outside the frontier

The available historical UGT1 protein sequence strongly supports correspondence to *PtunChr11G235070*, a Tier 2 locus outside the formal Tier 1 ranking universe [23]. Its reported yeast-expression and HPLC/LC-MS assays with UDP-glucose and quillaic acid or gypsogenin make it a useful biochemical comparison. Its exclusion resulted from the predefined Tier 1 sequence-eligibility criteria, rather than its reported biochemical activity. Including the historical construct alongside the core candidates would connect the new priorities to previous functional work. Because the historical ORF and modern CDS boundaries differ, the assayed sequence should be specified explicitly.

The ptUGT74E2 and ptUGT74B1 records mapped to *PtunChr3G078310* and *PtunChr5G117610* provide preliminary TLC-based reaction observations [22]. Together with the high- and medium-confidence qRT-PCR mappings to two core candidates, the fourteen-locus historical table connects historical names, primers and assay records to modern chromosome loci. Researchers can therefore retrieve earlier experimental leads

within the same genomic coordinate framework as the new candidates. The historical biochemical loci lie outside the current Front 1, making their reaction evidence complementary to the new expression priorities.

Scope and next experiments

The nine-library series contains a pretreatment baseline and two post-SA time points but no time-matched untreated or mock libraries at 8 or 24 h. Its contrasts cannot isolate SA-specific effects from other temporal changes, and metabolite measurements are not individually paired with these nine libraries. The 27 overlapping subsets quantify conditional stability within a fixed eligible universe. The signed axes favor positive adjacent-time-point effects, so the priorities reflect this particular screening objective. Individual sample correspondence between the external RNA-seq and metabolomic data was not fully resolved, so their comparison was limited to group-level context. Wild and cultivated roots provide cross-context expression evidence; their observational differences do not isolate cultivation, developmental or environmental effects or establish differences in saponin production. No new qRT-PCR or biochemical assays were performed on the ten frontier candidates, and improved enzyme-activity hit rates have not been measured.

The candidate resource supports both a general starting panel and selection by target reaction. The four core genes can be cloned or synthesized for heterologous expression and screened against the initial acceptor/donor panel, with empty-vector or enzyme-free controls and appropriate characterized reference reactions. Candidate choice within the ten-gene set can also follow the target reaction. For sugar-chain-extension studies, candidates nearest glycoside-acting UGT91H4 can be tested alongside the core panel, using the phylogenetic context to guide assays rather than assign activity. LC-MS/MS, comparison with authentic standards where available, and NMR when needed to resolve linkage position can distinguish product formation from an assigned catalytic reaction [8-10,14,20]. Liu et al. [57] reconstructed the biosynthesis of the *Quillaja saponaria* saponin QS-21 in engineered yeast, illustrating the value of characterized glycosyltransferases for assembling complex saponin pathways. Recording inactive combinations alongside active ones will make the resulting substrate and donor profiles useful for subsequent *P. tunicoides* saponin research.

Conclusions

This study establishes a genome-guided PtUGT resource for investigating triterpenoid saponin glycosylation in *P. tunicoides*. Restoring counting-compatible exon features recovered coverage for 49 existing loci, including *PtunChr8G179000*, which subsequently entered the four-candidate core. Multi-objective analysis distinguished full-data rank from persistence across replicate subsets and identified four subset-stable candidates within a ten-gene Pareto priority set. The core showed contrasting expression patterns in wild and cultivated roots. Mapping historical names, primers and assay records to modern loci preserves earlier experimental leads alongside these new priorities. Combined with reference-enzyme reactions and the plant's known saponin chemistry, these results support a four-gene initial panel and a ten-gene set from which researchers can select candidates according to their target glycosylation step.

List of abbreviations

SA, salicylic acid; UGT, UDP-dependent glycosyltransferase; PSPG, plant secondary product glycosyltransferase; DE, differential expression; FDR, false discovery rate; LRT, likelihood-ratio test; CDS, coding sequence; WP, wild plants; CP, cultivated plants; WGD, whole-genome duplication; LFC, $\log_2$ fold change; MLE, maximum-likelihood estimate; CPM, counts per million; PCA, principal component analysis.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The chromosome-level genome assembly is available under accession GCA_054740895.1 [1]. The historical SA transcriptome is available from CRA002795 [24]. The external wild-versus-cultivated root transcriptome

is available under BioProject PRJNA937778 [27]. Processed count matrices, differential-expression tables, the PtUGT master table, tree files, candidate-ranking outputs, historical-evidence mappings, figure source data and analysis scripts are deposited in a private Zenodo draft (reserved DOI: https://doi.org/10.5281/zenodo.23809359). A private review link is supplied to the editor and reviewers for peer review. The repository will be made publicly accessible before manuscript acceptance. The complete checked historical-evidence record is additionally supplied as Supplementary Table S9 (Additional file 8).

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by Key Projects of the Heilongjiang Provincial Natural Science Foundation (ZD2023C004) and Heilongjiang Chunyan Innovation Team Program: Green Manufacturing of Agricultural and Forestry Bioproducts (CYCX24023), and the Undergraduate Innovation Training Program of Northeast Forestry University (202410225042).

Authors' contributions

ZF: Conceptualization, Data curation, Methodology, Formal analysis, Validation, Writing – original draft. RZ: Data curation, Visualization, Validation, Writing – review and editing. JY: Conceptualization, Supervision, Project administration, Resources, Investigation. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the researchers who generated and made publicly available the genomic and transcriptomic datasets used in this study. Their contributions provided essential resources for the present analysis.

Graphical abstract legend

Restoring countability to prioritize candidate PtUGTs. Published SA-treated hairy-root transcriptomes provide the input context [24]. A counting-compatible exon representation expands the countable retained repertoire from 74 to 123 loci. Expression-based prioritization yields ten Front 1 candidates, of which four retain Front 1 membership in all 27 overlapping balanced subsets. *PtunChr8G179000* links restored countability to the subset-stable core. The dashed "Functional priorities" connector denotes candidates for functional investigation. The lower scheme illustrates representative quillaic-acid glycosylation context: R1 and R2 are H or sugar residues, with at least one sugar residue. It assigns no exact sugar, product, reaction position or catalytic function to a candidate. The plant photograph is adapted from Li et al. [26], Fig. 1C, under the Creative Commons Attribution license (CC BY 4.0; https://creativecommons.org/licenses/by/4.0/); it was cropped and the background removed. The flask is a schematic. UGTs are family-1 glycosyltransferases.

Supplementary figure legends

Supplementary Figure S1. Complete phylogeny and focused tree with numerical supports. Page 1 shows a circular overview of the 256-tip tree, comprising 123 retained *P. tunicoides* proteins, 122 Arabidopsis references and 11 characterized UGT references. Branch-length differences are retained. Pages 2–5 show rectangular continuation views with a common branch-length scale and all 253 numerical node supports. Page 6 shows the 50-tip focused tree with all 47 supports corresponding to displayed nodes. Supports are SH-aLRT/UFBoot values. Colors distinguish clades for visualization.

Supplementary Figure S2. Quality control of the 21 external root libraries. (A) variance-stabilizing transformation (VST)-based PCA coordinates for WP (n=3), 1CP (n=6), 2CP (n=6) and 3CP (n=6). Numeric labels map to the full accessions in panel B. (B) Pearson correlations of the VST profiles, displayed using a continuous light-to-dark blue scale spanning the observed minimum (r=0.853239) to 1. Top and left annotation bars identify the groups using the colors in panel A. PCA uses the 500 most variable genes; correlations use all 26,850 VST-transformed gene profiles retained for QC.

Supplementary Figure S3. PSPG-region sequence context of candidates and selected characterized references. (A) Alignment of ten Front 1 proteins and seven of the eleven characterized reference proteins listed in Supplementary Table S5. Grey columns indicate at least 80% conservation among the displayed sequences. Outlined columns mark the HCGWNS consensus-region positions, including substitutions. (B) PSPG-region positions within the full-length candidates; protein lengths exclude terminal stop symbols. Candidate colors follow the main figures. The alignment summarizes family-level sequence conservation.

Supplementary Figure S4. Sensitivity to exon and uniform-CDS counting representations. Front 1 membership and consensus ranks are shown for the primary apeglm analysis and the uniform-CDS sensitivity analysis. Each representation has ten Front 1 members, with nine shared; *PtunChr3G078100* is EXON-only and *PtunChr13G275040* CDS-only. The eligible ranking universes contain 62 and 63 genes, respectively, so the comparison includes both feature-representation and eligibility sensitivity. Cell numbers indicate consensus rank within each representation; a dash denotes absence from that representation's Front 1.

Supplementary Figure S5. Genomic context and physical-proximity analysis. (A) Protein lengths by confidence tier, with the operational 350–600-aa interval. (B) Pathway-homolog candidates by class and pseudochromosome; marker area represents sequence identity and deep red denotes strong-DE status. (C) Local PtUGT clusters and Front 1 membership. Ten operational clusters contain 22 retained loci: eight clusters have two members and two have three. TC03 contains the Front 1 locus *PtunChr11G239890*. Bars show members per cluster. (D) Nearest same-chromosome pathway-homolog distances for the ten Front 1 loci. (E) Comparison with 100,000 random sets. Mean distances were 18.47 Mb for Front 1 and 12.97 Mb for random sets (one-sided P=0.8463). The error bar shows the standard deviation of random-set means.

Supplementary Figure S6. Expression profiles of the 111 expression-filter-passing PtUGT loci. Row z-scores show within-gene expression variation across the nine libraries. The ten Front 1 genes appear first in consensus-rank order, followed by background loci. Pages 1 and 2 show 56 and 55 loci, respectively. Candidate colors follow the main figures. Figure 3B shows the corresponding expression-level trajectories for the ten Front 1 candidates.

Supplementary Information

Additional file 1 (.xlsx). Supplementary Tables S1–S8. SA and external sample metadata and quality control; annotation-representation checks; PtUGT sequence confidence; expression and differential-expression results; candidate ranks and sensitivities; characterized reference-enzyme provenance; and genomic-context summaries.

Additional file 2 (.pdf). Supplementary Figure S1. Six-page complete fixed 256-tip phylogeny and the 50-tip focused tree with numerical supports.

Additional file 3 (.pdf). Supplementary Figure S2. Quality-control summaries for the 21 external root libraries.

Additional file 4 (.pdf). Supplementary Figure S3. PSPG-region sequence context for ten frontier candidates and seven selected characterized reference proteins; all eleven phylogenetic references are documented in Supplementary Table S5.

Additional file 5 (.pdf). Supplementary Figure S4. Sensitivity of frontier membership and ranks to exon and uniform-CDS counting representations.

Additional file 6 (.pdf). Supplementary Figure S5. Protein lengths, genomic pathway-homolog and cluster contexts, and physical-proximity analyses.

Additional file 7 (.pdf). Supplementary Figure S6. Two-page nine-library expression profiles for all 111 expression-filter-passing PtUGT loci.

Additional file 8 (.tsv). Supplementary Table S9. Complete 14-locus historical expression and biochemical-evidence record, including unmatched entries, non-frontier loci, mapping qualifications and source details.

References

1. Xie Z, Zhang Y, Yang C, Zhang T, Zhang A, Li G. High quality chromosome-level genome assembly of *Psammosilene tunicoides* (Caryophyllaceae), an endangered medicinal plant. Sci Data. 2026;13:619. doi:10.1038/s41597-026-06991-0.

659 2. Li YH, Bai XS, Yang XX, Li YX, Li HR, Wang ZL, et al. Triterpenoid saponins from *Psammosilene*
660 *tunicoides* and their antinociceptive activities. Phytochemistry. 2023;214:113795.
661 doi:10.1016/j.phytochem.2023.113795.

662 3. Wen B, Tian JM, Huang ZR, Xiao SJ, Yuan WL, Wang JX, et al. Triterpenoid saponins from the roots of
663 *Psammosilene tunicoides*. Fitoterapia. 2020;144:104596. doi:10.1016/j.fitote.2020.104596.

664 4. Bowles D, Lim EK, Poppenberger B, Vaistij FE. Glycosyltransferases of lipophilic small molecules. Annu
665 Rev Plant Biol. 2006;57:567-597. doi:10.1146/annurev.arplant.57.032905.105429.

666 5. Louveau T, Osbourn A. The Sweet Side of Plant-Specialized Metabolism. Cold Spring Harb Perspect
667 Biol. 2019;11:a034744. doi:10.1101/cshperspect.a034744.

668 6. Seki H, Tamura K, Muranaka T. P450s and UGTs: Key Players in the Structural Diversity of Triterpenoid
669 Saponins. Plant Cell Physiol. 2015;56:1463-1471. doi:10.1093/pcp/pcv062.

670 7. Achnine L, Huhman DV, Farag MA, Sumner LW, Blount JW, Dixon RA. Genomics-based selection and
671 functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*.
672 Plant J. 2005;41:875-887. doi:10.1111/j.1365-313X.2005.02344.x.

673 8. Meesapyodsuk D, Balsevich J, Reed DW, Covello PS. Saponin biosynthesis in *Saponaria vaccaria*:
674 cDNAs encoding beta-amyrin synthase and a triterpene carboxylic acid glucosyltransferase. Plant Physiol.
675 2007;143:959-969. doi:10.1104/pp.106.088484.

676 9. Naoumkina MA, Modolo LV, Huhman DV, Urbanczyk-Wochniak E, Tang Y, Sumner LW, et al.
677 Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in
678 *Medicago truncatula*. Plant Cell. 2010;22:850-866. doi:10.1105/tpc.109.073270.

679 10. Shibuya M, Nishimura K, Yasuyama N, Ebizuka Y. Identification and characterization of
680 glycosyltransferases involved in the biosynthesis of soyasaponin I in *Glycine max*. FEBS Lett.
681 2010;584:2258-2264. doi:10.1016/j.febslet.2010.03.037.

682 11. Augustin JM, Drok S, Shinoda T, Sanmiya K, Nielsen JK, Khakimov B, et al. UDP-Glycosyltransferases
683 from the UGT73C Subfamily in *Barbarea vulgaris* Catalyze Sapogenin 3-O-Glucosylation in Saponin-
684 Mediated Insect Resistance. Plant Physiol. 2012;160:1881-1895. doi:10.1104/pp.112.202747.

685 12. Erthmann PØ, Agerbirk N, Bak S. A tandem array of UDP-glycosyltransferases from the UGT73C
686 subfamily glycosylate sapogenins, forming a spectrum of mono- and bisdesmosidic saponins. Plant Mol Biol.
687 2018;97:37-55. doi:10.1007/s11103-018-0723-z.

688 13. Nomura Y, Seki H, Suzuki T, Ohyama K, Mizutani M, Kaku T, et al. Functional specialization of UDP-
689 glycosyltransferase 73P12 in licorice to produce a sweet triterpenoid saponin, glycyrrhizin. Plant J.
690 2019;99:1127-1143. doi:10.1111/tpj.14409.

691 14. Li Y, Li J, Diao M, Peng L, Huang S, Xie N. Characterization of a Group of UDP-Glycosyltransferases
692 Involved in the Biosynthesis of Triterpenoid Saponins of *Panax notoginseng*. ACS Synth Biol. 2022;11:770-
693 779. doi:10.1021/acssynbio.1c00469.

694 15. Han X, Zhao J, Chang X, Li Q, Deng Z, Yu Y. Revisiting the transcriptome data of *Centella asiatica*
695 identified an ester-forming triterpenoid: UDP-glucose 28-O-glucosyltransferase. Tetrahedron.
696 2022;129:133136. doi:10.1016/j.tet.2022.133136.

697 16. Osmani SA, Bak S, Møller BL. Substrate specificity of plant UDP-dependent glycosyltransferases
698 predicted from crystal structures and homology modeling. Phytochemistry. 2009;70:325-347.
699 doi:10.1016/j.phytochem.2008.12.009.

700 17. Kurze E, Wüst M, Liao J, McGraphery K, Hoffmann T, Song C, et al. Structure-function relationship of
701 terpenoid glycosyltransferases from plants. Nat Prod Rep. 2022;39:389-409. doi:10.1039/D1NP00038A.

702 18. Chung SY, Seki H, Fujisawa Y, Shimoda Y, Hiraga S, Nomura Y, et al. A cellulose synthase-derived
703 enzyme catalyses 3-O-glucuronosylation in saponin biosynthesis. Nat Commun. 2020;11:5664.
704 doi:10.1038/s41467-020-19399-0.

19. Duan Y, Du W, Song Z, Chen R, Xie K, Liu J, et al. Functional characterization of a cycloartenol synthase and four glycosyltransferases in the biosynthesis of cycloastragenol-type astragalosides from *Astragalus membranaceus*. Acta Pharm Sin B. 2023;13:271-283. doi:10.1016/j.apsb.2022.05.015.

20. Sirirungruang S, Blay V, Rodriguez EP, Scott YF, Vuu KM, Barnum CR, et al. A substrate-multiplexed platform for profiling enzymatic potential of plant family 1 glycosyltransferases. Nat Commun. 2025;16:6366. doi:10.1038/s41467-025-61530-6.

21. Li W, Ma X, Li G, Zhang A, Wang D, Fan F, et al. De Novo Biosynthesis of the Oleanane-Type Triterpenoids of Tunicosaponins in Yeast. ACS Synth Biol. 2021;10:1874-1881. doi:10.1021/acssynbio.1c00065.

22. Li J. Research on the key enzyme genes and functions in the biosynthetic pathway of triterpenoid saponins in *Psammosilene tunicoides* based on transcriptome and metabolome association analysis (in Chinese). MSc thesis. Yunnan University of Chinese Medicine; 2024.

23. Meng W. Functional analysis of glycosyltransferase gene and preliminary study of MYC transcription factors in *Psammosilene tunicoides* (in Chinese). MSc thesis. Yunnan University of Chinese Medicine; 2020.

24. Su L, Li S, Qiu H, Wang H, Wang C, He C, et al. Full-Length Transcriptome Analyses of Genes Involved in Triterpenoid Saponin Biosynthesis of *Psammosilene tunicoides* Hairy Root Cultures With Exogenous Salicylic Acid. Front Genet. 2021;12:657060. doi:10.3389/fgene.2021.657060.

25. Chisanga D, Liao Y, Shi W. Impact of gene annotation choice on the quantification of RNA-seq data. BMC Bioinformatics. 2022;23:107. doi:10.1186/s12859-022-04644-8.

26. Li Y, Li Q, Wang L, Qi S, Jiang Z, Zhang W, et al. Integrated analysis of transcriptomics and metabolomics and high-throughput amplicon sequencing reveals the synergistic effects of secondary metabolites and rhizosphere microbiota on root rot resistance in *Psammosilene tunicoides*. Front Microbiol. 2025;16:1554406. doi:10.3389/fmicb.2025.1554406.

27. Li J, Yang W, Li Y, Han L, Li G, Zhang A. Integration of transcriptome, metabolome and high-throughput amplicon sequencing to compare the performance of wild and cultivated *Psammosilene*

730 *tunicoides* to reveal the beneficial plant-microbe interactions for domestication. Environ Exp Bot.
731 2024;218:105587. doi:10.1016/j.envexpbot.2023.105587.

732 28. Zhou M, Fan J, Gao Y, Zheng C, Xu Y, Jia L, et al. Identification and analysis of UGT genes associated
733 with triterpenoid saponin in soapberry (*Sapindus mukorossi* Gaertn.). BMC Plant Biol. 2024;24:588.
734 doi:10.1186/s12870-024-05281-4.

735 29. Yang Y, Han F, Wang J, Gao M, Wu L, Zhao Y, et al. Genome-wide investigation of family-1 UDP
736 glycosyltransferases reveals the Fusarium resistance role of VfUGT90A2 in tung trees. Ind Crops Prod.
737 2024;222:119471. doi:10.1016/j.indcrop.2024.119471.

738 30. Kamal KA, Shah FA, Zhao Y, Chen Z, Fu S, Zhu Z, et al. Genome-wide identification of the UGT genes
739 family in *Acer rubrum* and role of ArUGT52 in anthocyanin biosynthesis under cold stress. BMC Plant Biol.
740 2025;25:288. doi:10.1186/s12870-024-06043-y.

741 31. Liu L, Yang G, Wang S, Wang C, Gu L. Genome-wide identification and expression analysis of the
742 PtrUGT gene family in *Populus trichocarpa*. BMC Plant Biol. 2025;25:1527. doi:10.1186/s12870-025-
743 07572-w.

744 32. Zhao Q, Wang J, Li Q, Zhang J, Hou R, Wang Z, et al. Integrated transcriptome and metabolome analysis
745 provide insights into the mechanism of saponin biosynthesis and its role in alleviating cadmium-induced
746 oxidative damage in *Ophiopogon japonicum*. Plant Physiol Biochem. 2024;210:108634.
747 doi:10.1016/j.plaphy.2024.108634.

748 33. Yonekura-Sakakibara K, Hanada K. An evolutionary view of functional diversity in family 1
749 glycosyltransferases. Plant J. 2011;66:182-193. doi:10.1111/j.1365-313X.2011.04493.x.

750 34. Rahimi S, Kim J, Mijakovic I, Jung KH, Choi G, Kim SC, et al. Triterpenoid-biosynthetic UDP-
751 glycosyltransferases from plants. Biotechnol Adv. 2019;37:107394. doi:10.1016/j.biotechadv.2019.04.016.

752 35. Wang R, Wang C, Du J. Plant UGTs: Functional diversity, catalytic mechanisms and rational design for
753 metabolic engineering. Biotechnol Adv. 2026;89:108872. doi:10.1016/j.biotechadv.2026.108872.

754 36. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9:304.
755 doi:10.12688/f1000research.23297.2.

756 37. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics.
757 2018;34:i884-i890. doi:10.1093/bioinformatics/bty560.

758 38. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with
759 HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907-915. doi:10.1038/s41587-019-0201-4.

760 39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
761 format and SAMtools. Bioinformatics. 2009;25:2078-2079. doi:10.1093/bioinformatics/btp352.

762 40. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence
763 reads to genomic features. Bioinformatics. 2014;30:923-930. doi:10.1093/bioinformatics/btt656.

764 41. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data
765 with DESeq2. Genome Biol. 2014;15:550. doi:10.1186/s13059-014-0550-8.

766 42. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the
767 noise and preserving large differences. Bioinformatics. 2019;35:2084-2092.
768 doi:10.1093/bioinformatics/bty895.

769 43. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7:e1002195.
770 doi:10.1371/journal.pcbi.1002195.

771 44. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The
772 protein families database in 2021. Nucleic Acids Res. 2021;49:D412-D419. doi:10.1093/nar/gkaa913.

773 45. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods.
774 2015;12:59-60. doi:10.1038/nmeth.3176.

775 46. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in
776 performance and usability. Mol Biol Evol. 2013;30:772-780. doi:10.1093/molbev/mst010.

777 47. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in
778 large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972-1973. doi:10.1093/bioinformatics/btp348.

779 48. Wong TKF, Ly-Trong N, Ren H, Demotte P, Baños H, Roger AJ, et al. IQ-TREE 3: phylogenomic
780 inference software using complex evolutionary models. Mol Biol Evol. 2026;43:msag117.
781 doi:10.1093/molbev/msag117.

782 49. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model
783 selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587-589. doi:10.1038/nmeth.4285.

784 50. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast
785 Bootstrap Approximation. Mol Biol Evol. 2018;35:518-522. doi:10.1093/molbev/msx281.

786 51. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and
787 evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49.
788 doi:10.1093/nar/gkr1293.

789 52. Chen X, Hudson GA, Mineo C, Amer B, Baidoo EEK, Crowe SA, et al. Deciphering triterpenoid saponin
790 biosynthesis by leveraging transcriptome response to methyl jasmonate elicitation in *Saponaria vaccaria*.
791 Nat Commun. 2023;14:7101. doi:10.1038/s41467-023-42877-0.

792 53. Zhong HM, Ni W, Hua Y, Chen YZ, Chen CX. A new tritepenoid saponin from *Psammosilene*
793 *tunicoides*. Acta Bot Yunnanica. 2002;24:781-786.

794 54. Zhang L, Liu X, Li Y, Feng X, Li C. Uncovering the Functional Role of UGT73F/P Subfamilies toward
795 Triterpenoids Glycosylation in Legume Medicinal Plants. ACS Synth Biol. 2026;15:1327-1339.
796 doi:10.1021/acssynbio.5c00740.

797 55. Hudson GA, Pereira JH, Winegar PH, FitzGerald DM, DeGiovanni A, Chen X, et al. Enzymology and
798 Structural Basis of Glycosyltransferases Involved in Saponin C28 Carboxylic Acid O-D-Fucosylation. JACS
799 Au. 2025;5:6011-6024. doi:10.1021/jacsau.5c00907.

800 56. Wang G, Li M, Dai X, Zhang Y, Wu Y, Yan Z, et al. Tandem Duplication-Driven Neofunctionalization
801 of UDP-Glycosyltransferases Shapes the Diversification of Triterpenoid Saponins in the Cucurbitaceae. Adv
802 Sci (Weinh). 2026;13:e76159. doi:10.1002/advs.76159.

803 57. Liu Y, Zhao X, Gan F, Chen X, Deng K, Crowe SA, et al. Complete biosynthesis of QS-21 in engineered
804 yeast. Nature. 2024;629:937-944. doi:10.1038/s41586-024-07345-9.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1SupplementaryTables.xlsx
- Additionalfile2SupplementaryFigureS1.pdf
- Additionalfile3SupplementaryFigureS2.pdf
- Additionalfile4SupplementaryFigureS3.pdf
- Additionalfile5SupplementaryFigureS4.pdf
- Additionalfile6SupplementaryFigureS5.pdf
- Additionalfile7SupplementaryFigureS6.pdf
- Additionalfile8SupplementaryTableS9.tsv