



Genome-wide bioinformatic mining and evidence-based validation of PAL and CHS candidate genes in kasuri methi (*Trigonella corniculata* L.)

Mohit R. Sharma, Sanjay Kumar, Vasundhara Sharma, Y.K. Sharma, and M.K. Mahatma*

Reviewed by:

Dr. Rishee K. Kalaria
Aspee Shakilam Biotechnology
Institute, Surat Navsari Agricultural
University, Gujarat
Email: risheelkal@nau.in

*Correspondence

M.K. Mahatma
Email: maheshmahatma@gmail.com

Received: 03 September 2026

Revised: 15 September 2026

Accepted: 23 September 2026

Citation:

Sharma, M.R., Kumar, S., Sharma, V.,
Sharma, Y.K., & Mahatma, M.K.,
(2026) Genome-wide bioinformatic
mining and evidence-based validation
of PAL and CHS candidate genes in
kasuri methi (*Trigonella corniculata* L.).
Int J Seed Spice. 16 (1): pp:20-31.

DOI:

<https://doi.org/10.56093/IJSS.v16i1.3>

Affiliation:

ICAR-National Research Centre on
Seed Spices, Tabiji, Ajmer, Rajasthan,
India

Abstract

Chromosome-scale genomic resources are now available for Kasuri methi, however, reliable pathway annotation remains difficult because specialized-metabolism gene families are often paralogous and predicted gene models may be incomplete. We re-evaluated 117 pathway-oriented candidates from the 64,801 predicted protein-coding genes of *Trigonella corniculata* using an evidence-integrated workflow that combined curated Swiss-Prot homology, KO/EC concordance, protein-CDS-ORF consistency, Pfam profile-HMM searches, conserved-feature checks, family phylogeny, genomic context and targeted reconstruction. Seven phenylalanine ammonia-lyase (PAL) candidates contained PF00221 and retained the tested MIO-forming ASG tripeptide, while two chalcone synthase (CHS) candidates contained PF00195 and PF02797 across approximately 96% of their proteins and retained all four tested conserved functional residues. Two hydroxymethylglutaryl-CoA reductase (HMGR) candidates contained PF00368, and a 21-sequence, 377-aa conserved-domain phylogeny supported HMGR-family affinity; however, their 293-320-aa protein models were only about half the median length of full-length plant HMGR references and retained only partial or divergent diagnostic sequence features. LOX, oxidosqualene cyclase (OSC) and terpene synthase (TPS) candidates remained strongly truncated and were not rescued by homolog-guided reconstruction, while the tested NNMT and IMT/D-pinitol-associated models lacked sufficient full-length support for confident assignment. Overall, the analysis resolves the initial screen into nine strongly supported PAL/CHS family-level candidates, two partial HMGR-family models, and a set of unresolved or unsupported annotations that require further gene-model refinement rather than functional transfer from short local similarity alone.

Keywords: *Trigonella corniculata*; Kasuri methi; secondary metabolism; PAL; CHS; HMGR; gene-model validation; Pfam; phylogeny; genome mining

Introduction

Kasuri methi or pan methi (*Trigonella corniculata*) is a leafy spice valued for its distinctive aroma and is closely related to common fenugreek (*Trigonella foenum-graecum*). A chromosome-scale assembly of genotype JKM-5 now provides a 798-Mb genomic framework containing 64,801 predicted protein-coding genes, with more than 98% of the assembly anchored to eight pseudomolecules (Gaikwad et al., 2025). This resource provides a valuable foundation for pathway-focused genome mining; however, it also underscores a fundamental challenge in non-model plant genomics: an automated functional annotation alone is insufficient to establish that a predicted locus represents a complete gene model encoding a structurally intact and

biochemically functional enzyme. Recent annotation studies similarly emphasize that reliable eukaryotic gene models benefit from integrating independent evidence types and from evaluating gene-model consistency rather than relying on a single annotation score (Gabriel et al., 2024; Nevers et al., 2025). This distinction is especially relevant to *Trigonella* specialized metabolism. Studies in *T. foenum-graecum* have implicated flavonoids, steroidal saponins, terpenoids, alkaloids, volatile compounds and cyclitols in the biochemical diversity of the genus. Transcriptomic analyses show that orthology, conserved residues, expression patterns, and phylogenetic placement can help distinguish plausible pathway genes from superficial similarity (Naika et al., 2022). More recent multi-omics work has documented extensive flavonoid-associated transcription during seed

development (Zhao et al., 2024), dynamic regulation of terpenoid biosynthesis (Javan et al., 2024), tissue-specific relationships among flavonoids, terpenoids and alkaloids (Dong et al., 2025), and comparative associations between D-pinitol accumulation and candidate biosynthetic genes in *T. foenum-graecum* and *T. corniculata* (Yadav et al., 2026). Recent reviews further highlight the value of genome mining and multi-omics integration for resolving specialized-metabolite pathways in non-model and medicinal plants (Wang et al., 2024). These studies provide strong pathway-level biological context; however, they do not, by themselves, establish whether individual *T. corniculata* gene models are structurally complete, accurately annotated, and functionally assignable.

The present study addressed a deliberately narrower but more stringent question: which *T. corniculata* gene models can be considered credible candidates for secondary metabolism when supported by multiple independent lines of evidence? Rather than maximizing the number of annotated pathway genes, we aimed to distinguish robust family-level assignments from partial, fragmentary, or weakly supported models. Unusually short proteins were therefore treated as candidate hypotheses requiring further validation, rather than being assumed to represent fragmented full-length genes. To achieve this, we integrated curated homology, annotation concordance, gene-model structure, Pfam domain evidence, conserved sequence features, phylogenetic relationships, and targeted sequence reconstruction. This

evidence-integrated approach provides not only a prioritized set of candidate genes but also an explicit assessment of uncertainty associated with their underlying gene models.

Materials and Methods

Genome resources and candidate screen

The chromosome-scale *T. corniculata* annotation, protein FASTA, coding-sequence FASTA and GFF files were used as the common reference (Gaikwad et al., 2025). Gene models were linked through Km identifiers. A pathway-oriented screen produced 117 candidates representing phenylpropanoid, isoprenoid, aroma, saponin, trigonelline and D-pinitol-related functions. Downstream similarity integration yielded BLAST-linked evidence records for 116 unique candidates; the original 117-candidate screen count was retained and the accounting difference is documented in Supplementary Table S1. The working hypothesis was that bona fide, structurally complete pathway gene models would show concordant evidence across functional annotation, high-coverage homology, expected family domains or conserved features, and phylogenetic placement, whereas incomplete or erroneous models would be characterized by disproportionate protein shortening, limited reference coverage, disrupted feature architecture and/or failure of genomic reconstruction. The evidence hierarchy used to test this hypothesis is summarized in **Figure 1**.

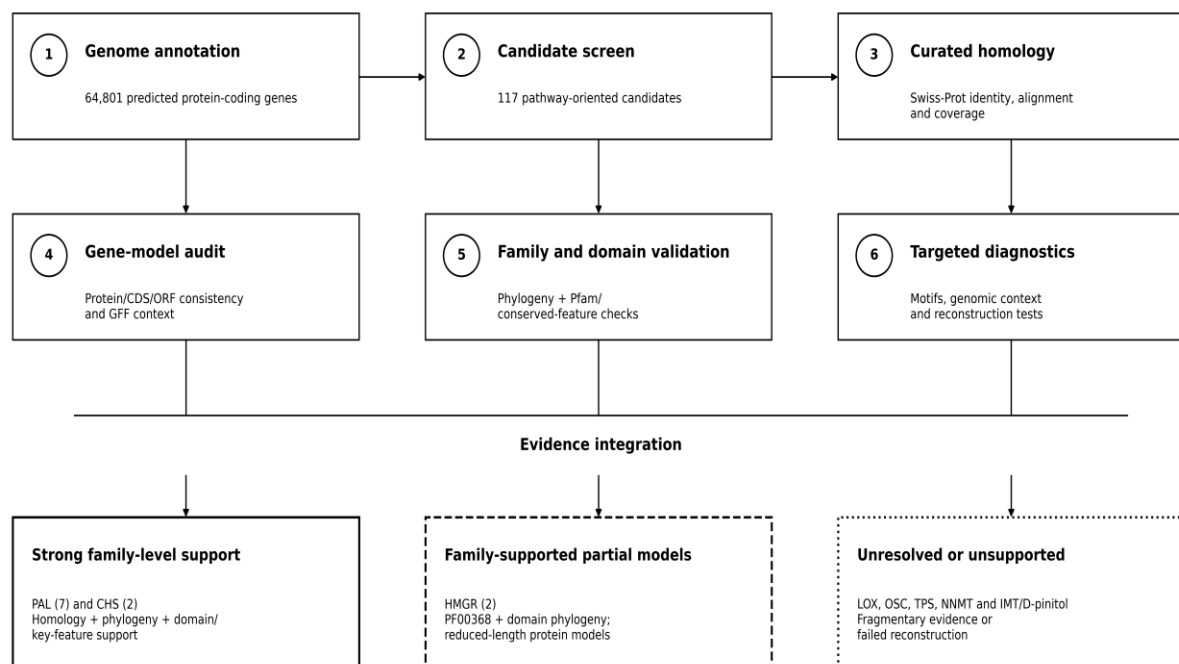


Fig 1. Evidence-integration workflow for prioritization and validation of secondary-metabolism candidate genes in *T. corniculata*. Starting from 64,801 predicted protein-coding genes, 117 pathway-oriented candidates were evaluated using curated homology, gene-model consistency, family-level phylogeny, Pfam-domain and conserved-feature assessment, and targeted motif, genomic-context, and reconstruction diagnostics. Integration of these evidence layers separated candidates into strong family-level support (PAL and CHS), family-supported partial models (HMGR), and unresolved or unsupported candidates (LOX, OSC, TPS, NNMT and IMT/D-pinitol-associated models). Short local sequence similarity alone was not considered sufficient evidence for functional assignment.

Curated homology and annotation concordance

Candidate proteins were compared with curated Swiss-Prot homologs using BLAST-based protein similarity searches (Camacho et al., 2009). Percent identity was interpreted together with alignment length, query coverage and reference coverage so that short local matches were not equated with full-length functional similarity. Exact or near-exact identities restricted to short peptides were retained as observations but were not treated as sufficient functional evidence. For prioritized candidates, enzyme names, EC numbers, KEGG Orthology (KO) identifiers and pathway annotations were cross-checked against the original genome annotation. PAL, CHS and HMGR candidates corresponded to K10775/EC 4.3.1.24, K00660/EC 2.3.1.74 and K00021/EC 1.1.1.34, respectively.

Gene-model audit and reconstruction tests

Protein lengths were compared with CDS lengths and CDS-derived translations, and alternative open reading frames were examined when a model was unusually short. GFF coordinates and local gene neighborhoods were then inspected to determine whether adjacent loci could plausibly represent fragments of a single biological gene. For LOX, OSC and TPS, a neighboring-model reconstruction was considered only when the target itself showed substantive homology, genomic order was compatible with reference order, mapped segments were complementary rather than redundant, strand orientation was biologically compatible, and the combined CDS could support a plausible open reading frame. Reconstructions that failed any of these criteria were

rejected rather than promoted as putative full-length enzymes.

Domain and conserved-feature validation

The 11 prioritized PAL, CHS and HMGR proteins were scanned against the Pfam-A profile database using HMMER hmmscan with Pfam gathering thresholds (Eddy, 2011; Mistry et al., 2021). Expected family profiles were PF00221 for PAL, PF00195 and PF02797 for CHS, and PF00368 for HMGR. Conserved sequence features were then evaluated in family-specific alignments against reviewed reference proteins. For PAL, the MIO-forming ASG tripeptide was assessed relative to parsley PAL P24481. For CHS, Cys164, Phe215, His303 and Asn336 were examined as conserved functional residues relative to *Medicago sativa* CHS P30074. For HMGR, the diagnostic sequence blocks EMPVGYVQIP, TTEGCLVA, DAMGMNM and GTVGGGT were screened as additional evidence. Domain and conserved-feature retention were interpreted as family-level support and not as proof of biochemical activity or complete protein architecture.

External homologs and phylogenetic analysis

External homologs were selected from related legumes and Arabidopsis using functional annotation and family-appropriate length filters. Close Trifolieae references, including *Medicago truncatula* and *Trifolium pratense*, were emphasized because they provide relevant comparative context for *Trigonella* (Naika et al., 2022). Protein alignments were generated with MAFFT v7.526 (Katoh and Standley, 2013), and maximum-likelihood phylogenies were inferred with IQ-TREE 3.1.3 (Wong et al., 2026) using ModelFinder, 1,000 ultrafast-bootstrap replicates and 1,000 SH-like aLRT replicates; ultrafast bootstrapping followed UFBoot2 (Hoang et al., 2018). The final PAL analysis comprised 23 proteins and 761 aligned amino-acid sites and selected Q.PLANT+R3. The CHS analysis comprised 16 proteins and 415 aligned sites and selected LG+I+R2. Because the two Kasuri HMGR models were markedly shorter than typical full-length plant HMGRs, the final HMGR analysis was restricted to the shared PF00368 region from two Kasuri candidates and 19 curated plant references. This

produced a 21-sequence, 377-aa alignment and selected Q.PLANT+G4. The HMGR tree was midpoint-rooted only for graphical presentation; the displayed root was not interpreted as an inferred evolutionary root.

Model completeness and final evidence categories

Protein length was used as a practical candidate-level diagnostic of gene-model completeness by expressing each Kasuri prediction relative to the median length of its selected curated external-homolog set. This family-normalized metric was chosen because truncated gene predictions are expected to be systematically shorter than intact homologs, while normalization to a family-specific reference set avoids imposing a single absolute length threshold across unrelated enzyme families. The median, rather than the mean, was used to reduce sensitivity to unusually long or short reference proteins. The rationale is consistent with recent gene-repertoire quality frameworks in which family-relative protein length is used to flag likely fragments; for example, OMArk classifies proteins shorter than half of their assigned gene-family median as candidate fragments (Nevers et al., 2025). In contrast, whole-proteome completeness metrics primarily evaluate the presence or absence of conserved gene repertoires and are not designed to resolve candidate-specific truncation. Transcript-dependent model-quality measures could also not be applied uniformly because matched transcript evidence was not available for every locus. Therefore, relative protein length was used only as a structural screening indicator and never as a stand-alone completeness criterion; it was interpreted jointly with curated homology, annotation concordance, family phylogeny, Pfam/conserved-feature evidence, CDS/ORF consistency and reconstruction outcomes. Final evidence categories were deliberately qualitative: strong family-level support, family-supported partial model,

unresolved/incomplete, or unsupported. These categories summarize convergence among independent computational observations and were not treated as formal probabilities of enzyme function.

Results

Evidence integration reduced the candidate set

The initial screen returned 117 pathway-oriented candidates, of which 116 generated downstream BLAST-linked evidence records. Support became markedly uneven when coverage, gene-model structure, conserved features, domains and family context were considered together. Long, high-coverage matches were concentrated in a small subset of PAL and CHS genes. By contrast, several LOX, OSC, TPS and metabolite-specific candidates showed high local identity only across short segments of the expected enzyme. Incorporating coverage and model structure as primary evidence therefore changed the biological interpretation of the screen rather than simply adding confidence scores to the original annotations.

PAL and CHS form the strongest candidate set

Seven PAL candidates and two CHS candidates showed the most internally consistent combination of curated homology and annotation concordance (Table 1). All seven PAL models carried K10775/EC 4.3.1.24, and their best curated homologs showed 78.46-96.95% identity with 91-100% query coverage. Km104323 and Km33730 were both 326-aa proteins annotated as K00660/EC 2.3.1.74 and each matched a curated *Medicago sativa* CHS at 98.47% identity across the full query. The two HMGR candidates were likewise annotated as K00021/EC 1.1.1.34, but their substantially shorter predicted proteins required separate assessment of domain retention, conserved sequence features and model completeness.

Table 1. Integrated evidence for prioritized PAL, CHS and HMGR candidates.

Family	Candidate	Protein (aa)	Curated homolog (identity / qcov %)	Domain / key-feature evidence	Phylogenetic evidence	Final interpretation
PAL	Km108112	557	M. sativa PAL (P27990); 96.95 / 100	PF00221; ASG conserved	23-seq PAL tree; Q.PLANT+R3	Strong family-level support
PAL	Km108114	557	M. sativa PAL (P27990); 96.23 / 100	PF00221; ASG conserved	23-seq PAL tree; Q.PLANT+R3	Strong family-level support
PAL	Km108115	557	M. sativa PAL (P27990); 96.77 / 100	PF00221; ASG conserved	23-seq PAL tree; Q.PLANT+R3	Strong family-level support
PAL	Km97283	557	G. max PAL1 (P27991); 92.09 / 100	PF00221; ASG conserved	23-seq PAL tree; Q.PLANT+R3	Strong family-level support
PAL	Km52307	615	C. arietinum PAL2 (Q9SMK9); 95.20 / 95	PF00221; ASG conserved	23-seq PAL tree; Q.PLANT+R3	Strong family-level support
PAL	Km31372	644	P. vulgaris PAL class 3 (P19143); 88.17 / 91	PF00221; ASG conserved	23-seq PAL tree; Q.PLANT+R3	Strong family-level support
PAL	Km113277	636	S. humilis PAL (P45732); 78.46 / 92	PF00221; ASG conserved	23-seq PAL tree; Q.PLANT+R3	Strong family-level support
CHS	Km104323	326	M. sativa CHS4 (P30075); 98.47 / 100	PF00195 + PF02797; 4/4 tested conserved functional residues	16-seq CHS tree; LG+I+R2	Strong family-level support
CHS	Km33730	326	M. sativa CHS2 (P30074); 98.47 / 100	PF00195 + PF02797; 4/4 tested conserved functional residues	16-seq CHS tree; LG+I+R2	Strong family-level support
HMGR	Km7398	320	Cannabis sativa HMGR1 (A0A7J6H013); 68.56 / 98	PF00368 (181-310); diagnostic motifs partial/divergent	21-seq PF00368 tree; Q.PLANT+G4; no robust specific ortholog	Partial HMGR-family model
HMGR	Km90795	293	Hevea brasiliensis HMGR1 (P29057); 60.71 / 97	PF00368 (163-292); diagnostic motifs partial/divergent	21-seq PF00368 tree; adjacent <i>M. truncatula</i> HMGR4 (76.1/53)	Partial HMGR-family model

qcov, query coverage. Phylogenetic statements denote family-level placement unless specific node support is stated

Domain and conserved-feature validation strengthens PAL/CHS support and qualifies HMGR

All seven PAL candidates contained PF00221 and retained the tested MIO-forming ASG tripeptide. The detected Pfam region occupied approximately 59.0-68.9% of the predicted PAL proteins; this value reflects coverage by the Pfam family profile and was not used as a measure of full-length protein completeness. Both CHS candidates contained the expected type-III polyketide-synthase profiles PF00195 and PF02797, spanning 96.32% and 96.63% of their respective proteins, and both retained the four tested conserved functional residues. The agreement among curated homology, annotation, Pfam evidence and conserved sequence features therefore provided strong independent support for PAL/CHS family membership. Both HMGR candidates contained PF00368, spanning residues 181-310 in Km7398 and 163-292 in Km90795. In contrast to PAL and CHS, however, the tested HMGR diagnostic sequence blocks were only partially conserved or divergent in the current short models. The domain evidence therefore supports HMGR-family affinity while

leaving the missing N-terminal and other architectural features unresolved. These loci were consequently retained as partial HMGR-family models rather than classified as complete enzymes.

Phylogenies support PAL/CHS membership and HMGR family affinity

The final PAL dataset contained seven *T. corniculata* proteins and 16 external homologs. The 761-aa alignment selected Q.PLANT+R3, and the maximum-likelihood tree had a log-likelihood of -7816.6160. All seven Kasuri proteins remained within the PAL comparative framework, with several close relationships to legume homologs and strong support on major internal branches (Figure 2). Together with curated homology, KO/EC concordance and the domain/conserved-feature results, the phylogeny provided an independent line of evidence supporting the seven PAL candidates at the family level.

PAL maximum-likelihood phylogeny | Q.PLANT+R3

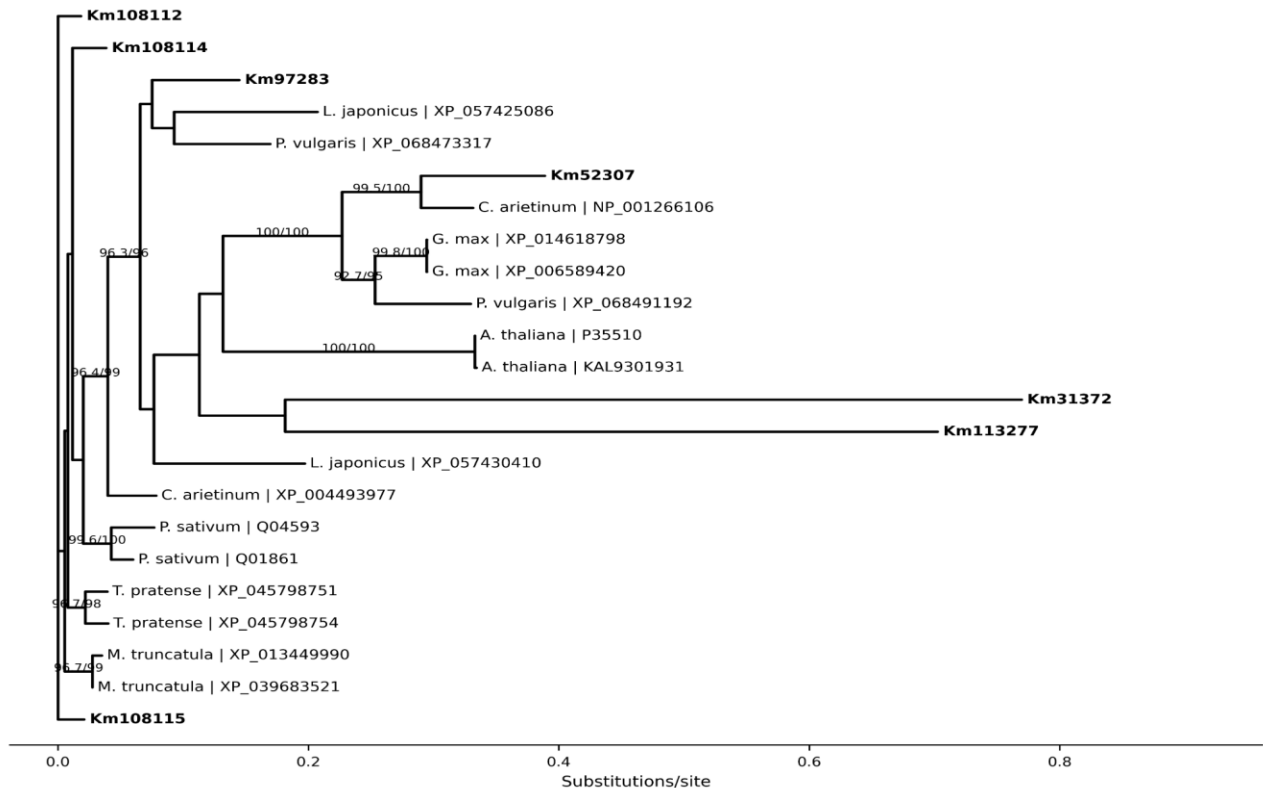


Figure 2. Maximum-likelihood phylogeny of PAL-family proteins from *T. corniculata* and representative plant species. The analysis included seven *T. corniculata* candidates and 16 external reference proteins aligned across 761 amino-acid positions. IQ-TREE selected Q.PLANT+R3 according to the Bayesian information criterion. Branch support was assessed with 1,000 SH-aLRT replicates and 1,000 ultrafast bootstrap (UFBoot) replicates. *T. corniculata* proteins are shown in bold. Branch lengths represent substitutions per site. Internal-node labels report SH-aLRT/UFBoot support (%) and are shown routinely only when SH-aLRT $\geq 80\%$ and UFBoot $\geq 95\%$ to reduce visual crowding. The rectangular representation of the unrooted topology does not imply a biological root. The CHS analysis was similarly coherent. Km104323 and Km33730 were analyzed with

14 external proteins in a 415-aa alignment; LG+I+R2 was selected and the maximum-likelihood tree had a log-likelihood of -4569.9204 (Figure 3A). For HMGR, the final conserved-domain dataset contained two Kasuri candidates and 19 curated plant references. The 377-aa PF00368 alignment included 86 parsimony-informative sites, selected Q.PLANT+G4, and yielded a log-likelihood of -3035.5445. Km90795 was adjacent to *M. truncatula* HMGR4, but support for this specific sister relationship was modest (SH-aLRT/UFBoot = 76.1/53). Km7398 also remained within the HMGR-domain comparison but lacked a strongly supported one-to-one relationship with an individual reference. Thus, phylogenetic evidence supports HMGR-family membership without establishing specific orthology or complete enzyme architecture (**Figure 3B**).

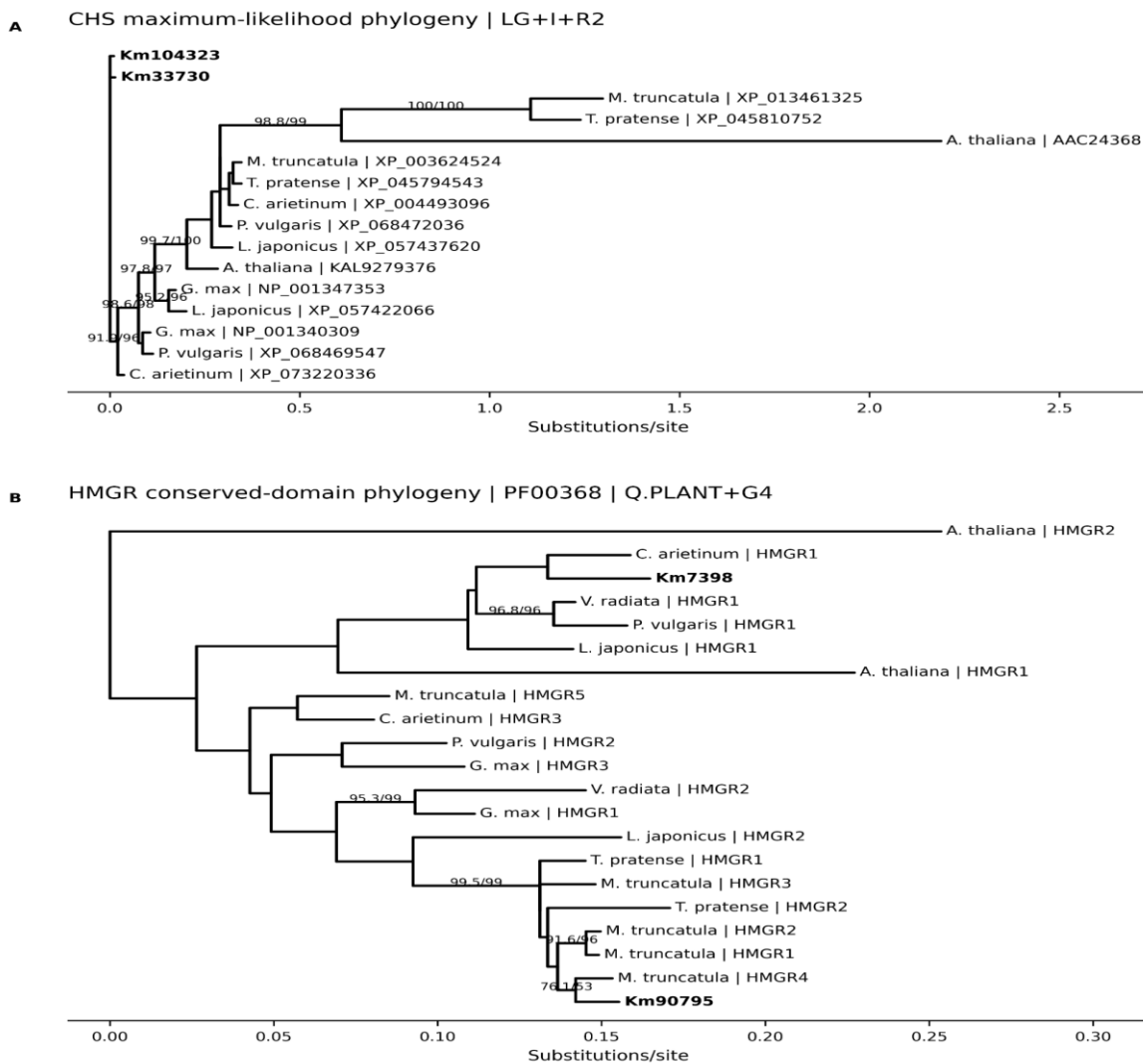


Figure 3. Phylogenetic validation of CHS and HMGR candidates in *T. corniculata*. (A) Maximum-likelihood phylogeny of two *T. corniculata* CHS candidates and 14 representative plant proteins based on a 415-aa alignment; LG+I+R2 was selected as the best-fitting model. (B) Conserved-domain phylogeny of two *T. corniculata* HMGR candidates and 19 curated plant HMGR references based on the shared PF00368 region (377 aligned amino-acid positions); Q.PLANT+G4 was selected. For both analyses, branch support was assessed with 1,000 SH-aLRT replicates and 1,000 ultrafast bootstrap (UFBoot) replicates. The HMGR tree was midpoint-rooted solely for graphical presentation and the displayed root should not be interpreted as an inferred evolutionary root. *T. corniculata* proteins are shown in bold. Routine node labels represent SH-aLRT/UFBoot support (%) and are shown when SH-aLRT $\geq 80\%$ and UFBoot $\geq 95\%$; the Km90795-*M. truncatula* HMGR4 value (76.1/53) is retained to show the limited support for

that particular relationship. Branch lengths represent substitutions per site.

Protein length and evidence synthesis distinguish plausible models from severe truncations

Protein length provided a simple but informative diagnostic of gene-model completeness (Figure 4A). PAL candidates were 76.9-89.0% of the median selected PAL-homolog length, and both CHS proteins were approximately 83.4% of the selected CHS median. The two HMGR models were only 49.8-54.4% of the full-length HMGR reference median. The discrepancy was larger for LOX, OSC and TPS, whose predicted proteins represented only about 12.7-29.0% of the relevant external-homolog medians. When this completeness signal was integrated with curated homology, phylogeny, Pfam/conserved-feature evidence and reconstruction outcomes, PAL and CHS separated clearly from partial HMGR models and from unresolved or unsupported candidate groups (Figure 4B).

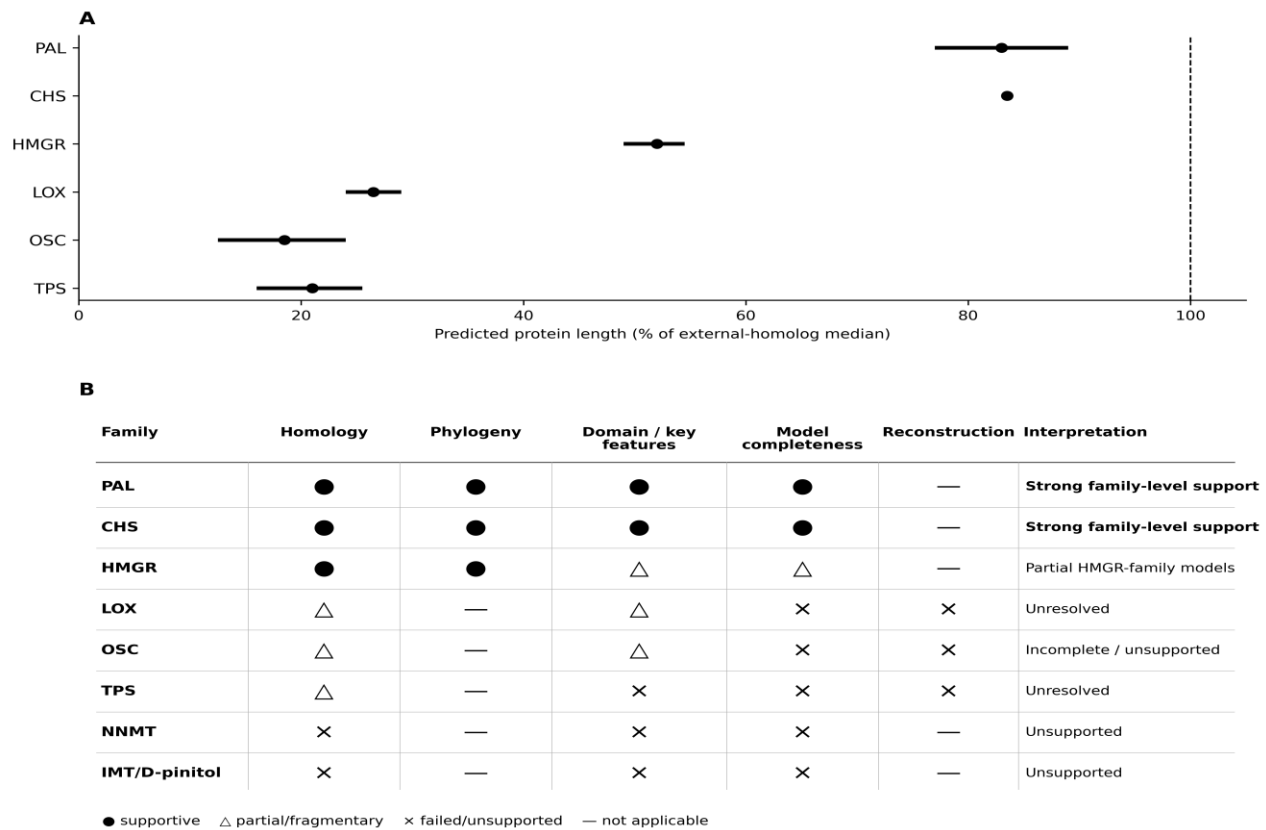


Figure 4. Protein-model completeness and integrated evidence supporting secondary-metabolism candidate assignments in *T. corniculata*. (A) Predicted protein lengths of PAL, CHS, HMGR, LOX, OSC and TPS candidate families expressed relative to the median length of curated external homologs. Points mark the midpoint of the observed family range and horizontal bars indicate the minimum-to-maximum range; the dashed vertical line denotes 100% of the external-homolog median. (B) Evidence matrix integrating curated homology, family-level phylogeny, Pfam-domain/conserved-feature support, predicted model completeness and targeted reconstruction tests. Filled circles indicate supportive evidence, open triangles indicate partial or fragmentary evidence, crosses indicate failed or unsupported evidence, and dashes indicate analyses that were not applicable. PAL and CHS candidates received the strongest family-level support, whereas Km7398 and Km90795 were retained as partial HMGR-family models.

Targeted reconstruction did not rescue LOX, OSC or TPS models

Reconstruction was tested only when neighboring short models appeared to map to different portions of the same external reference. Km27440 and Km27441 jointly covered 28.21% of an 865-aa *Phaseolus vulgaris* LOX reference, but the mapped segments involved mixed strand orientation, reversed reference order and overlap, so the proposed merge was rejected. The OSC hypothesis involving candidate models Km117295 and Km117299 together with neighboring model Km117301 covered 23.20% of a 763-aa reference but again showed reversed order and overlapping segments. The strongest TPS hypothesis, Km14477/Km14481, also failed a decisive continuity test: the combined analysis produced a longest exploratory ORF of only 28 aa and covered 15.45% of the 589-aa reference. These results show that apparent multi-fragment similarity was insufficient to justify reconstruction of any of the tested LOX, OSC or TPS models (Table 2; Supplementary Figure S1).

Table 2. Final status of unresolved or incomplete candidate groups.

Group	n	Kasuri length (aa)	Strongest evidence / failure point	Final interpretation
LOX	3	217-264	Km27440/Km27441 covered 28.21% of an 865-aa <i>P. vulgaris</i> reference, but mapped on mixed strands with reversed order and overlap.	Reject merge; unresolved
OSC / saponin	8	97-183	Candidate models Km117295/Km117299 plus neighboring Km117301 covered 23.20% of a 763-aa reference but showed reversed reference order and overlapping segments.	Incomplete / unsupported
TPS / aroma	16	95-149	No DDxxD motif; Km14477/Km14481 yielded a 28-aa exploratory ORF and 15.45% coverage of a 589-aa reference.	Reconstruction rejected; unresolved
NNMT	1	128	Km74137 lacked acceptable full-length NNMT reference support.	Unsupported as the tested NNMT model
IMT / D-pinitol	1	106	Km60703 showed 100% local identity to an IMT-like reference but only 2.83% query coverage (0.81% reference coverage); no acceptable OEPA/OEPB support.	Unsupported as the tested IMT model

Coverage values refer to the strongest homolog-informed hypothesis retained for detailed review. “Unsupported” refers to the tested gene model, not to absence of the corresponding metabolite or pathway in *Trigonella*.

Metabolite-specific assignments remained unsupported

The metabolite-focused searches produced two informative negative results. Km74137, evaluated as an NNMT/trigonelline candidate, lacked acceptable full-

length NNMT reference support. Km60703 produced a 100% identity IMT-like local match, but the alignment covered only 2.83% of the query and 0.81% of the reference, and no acceptable OEPA or OEPB support was obtained. These findings do not challenge the occurrence

Discussion

The central contribution of this study is an evidence-based separation of gene-family identity from gene-model reliability. Starting from a broad set of pathway-oriented annotations, the analysis required sequence coverage, curated homology, annotation concordance, Pfam evidence, conserved sequence features, gene-model structure and, where informative, phylogenetic or reconstruction support to converge before a candidate was promoted. This distinction is important in specialized-metabolism gene families, where duplication, paralogy and local sequence conservation can make short alignments appear functionally convincing. The resulting candidate set is therefore intentionally smaller than the initial screen, but each retained interpretation is more traceable and biologically defensible.

PAL and CHS illustrate the value of this convergence most clearly. The seven PAL loci combine high-coverage curated homology, K10775/EC 4.3.1.24 concordance, PF00221 detection, retention of the tested MIO-forming ASG tripeptide and supportive family phylogeny. The two CHS loci show an equally coherent pattern, with full-query curated matches, K00660/EC 2.3.1.74 concordance, the expected PF00195/PF02797 domain pair and retention of the four tested conserved functional residues. These results complement transcriptomic and metabolomic studies in common fenugreek that have emphasized phenylpropanoid and flavonoid metabolism during seed development and in tissue-specific secondary-metabolite accumulation (Naika et al., 2022; Zhao et al., 2024; Dong et al., 2025). The present work adds a different level of resolution by identifying nine specific *T. corniculata* loci whose family-level assignments remain robust after explicit gene-model and sequence-quality checks. They are therefore strong priorities for expression profiling and downstream functional validation, although the present evidence does not by itself establish isoform-specific catalytic roles or metabolite flux.

HMGR provides a useful counterexample in which family membership is supported but gene-model completeness is not. HMGR is a key enzyme of the mevalonate pathway and is relevant to terpenoid biosynthesis in fenugreek

of trigonelline or D-pinitol metabolism in *Trigonella*; rather, they show that the particular short gene models tested here cannot be assigned confidently to those pathway steps on the available sequence evidence.

(Javan et al., 2024). Both Km7398 and Km90795 retain PF00368 and fall within the plant HMGR conserved-domain comparison, indicating genuine HMGR-family affinity. At the same time, their 293-320-aa predictions are roughly half the length of typical full-length plant HMGR proteins and retain only partial or divergent diagnostic sequence blocks. This combination argues against treating the current annotations as complete HMGR enzymes. The placement of Km90795 next to *M. truncatula* HMGR4 is also insufficient for a one-to-one orthology claim because the supporting node is modest (76.1/53). Thus, the HMGR result is not a failed annotation; rather, it identifies two biologically plausible loci whose current structural models require transcript-supported refinement before functional interpretation can be strengthened.

The LOX, OSC and TPS analyses show why negative reconstruction evidence is equally important. These families contained very short predictions relative to curated external proteins, and several loci initially appeared compatible with a fragmented-gene explanation. However, the strongest hypotheses failed biologically necessary checks of strand orientation, reference order, segment complementarity or ORF continuity. The Km14477/Km14481 TPS pair is particularly illustrative: neighboring models and partial reference coverage could suggest a split locus, yet the combined CDS produced only a 28-aa exploratory ORF. In gene families that often occur in local clusters or expanded paralogous regions, genomic proximity alone is therefore not sufficient evidence for reconstruction. Rejecting such models prevents annotation fragments from being propagated as apparently complete biosynthetic enzymes.

The metabolite-specific results reinforce the same distinction between pathway occurrence and gene assignment. D-pinitol accumulation is well documented in fenugreek (Lahuta et al., 2018), and comparative *Trigonella* multi-omics has linked this metabolite with candidate biosynthetic-gene expression (Yadav et al., 2026). Similarly, previous fenugreek transcriptome studies have used orthology, conserved residues and expression evidence to support candidate genes associated with trigonelline metabolism (Naika et al., 2022). The inability to validate Km60703 as an IMT/D-pinitol-associated

enzyme or Km74137 as a full-length NNMT therefore should not be interpreted as evidence that these pathways are absent from *T. corniculata*. Instead, it demonstrates that pathway presence cannot be used to justify a specific locus when sequence coverage and gene-model completeness are inadequate.

Taken together, the results position gene-model refinement as the most productive next step for the unresolved families. For Km7398 and Km90795, transcript-supported exon structure should be established before additional functional claims are made. Local synteny and orthology with *M. truncatula*, *T. pratense* and, where comparable resources are available, *T. foenum-graecum* would provide an independent test of locus identity. Tissue-resolved RNA sequencing could then determine whether the nine PAL/CHS candidates and the two partial HMGR models are expressed as complete transcripts in metabolically relevant organs. Ultimately, biochemical assays, genetic perturbation or metabolite-associated expression will be required to move from family-level assignment to catalytic function. Recent syntheses of plant specialized-metabolite genomics likewise emphasize that gene organization and genome mining are most informative when coupled to orthogonal omics and experimental evidence (Wang et al., 2024; Dixon, 2025). More broadly, the study shows that a chromosome-scale assembly is only the first layer of pathway annotation: combining positive and negative evidence is essential for preventing short or fragmented gene models from being converted into overconfident functional claims in non-model crop genomes.

Conclusions

Evidence-integrated genome mining resolved the *T. corniculata* secondary-metabolism screen into three distinct evidence classes: seven PAL and two CHS candidates with strong family-level support, two HMGR-family loci whose current protein models are partial, and a set of LOX, OSC, TPS, NNMT and IMT/D-pinitol-associated models that remain unresolved or unsupported. The principal value of the analysis is the distinction between plausible gene-family identity and complete functional annotation. By retaining strong candidates while explicitly rejecting unsupported reconstruction and short-match functional transfer, the study provides a compact set of loci for transcript, expression and biochemical validation and identifies specific annotation gaps that should be prioritized for future gene-model refinement.

Data and Code Availability

The chromosome-scale *T. corniculata* genome is available under GenBank accession JBLEBM0000000000 (assembly GCA_047302345.1; BioProject PRJNA1122319; Gaikwad et al., 2025).

Supplementary Information

Supplementary Figure S1 summarizes the decisive reconstruction failures for LOX, OSC and TPS, and Supplementary Figure S2 provides pathway context without implying measured flux or direct experimental validation. Supplementary Table S1 contains the integrated downstream output for 116 BLAST-linked candidates recovered from the 117-candidate initial screen; Supplementary Tables S2-S4 provide domain/conserved-feature results, reconstruction diagnostics, and external phylogenetic references with software/model settings, respectively.

Competing interests

The authors declare no conflicts of interest.

Acknowledgements

This research was supported by the Indian Council of Agricultural Research, Department of Agricultural Research and Education, Government of India. Authors are also thankful to the Director, ICAR-NRCSS, Ajmer for providing the necessary facilities to conduct this study.

References

- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L., (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421. DOI: <https://doi.org/10.1186/1471-2105-10-421>.
- Dixon, R.A., (2025). The ins and outs of plant specialized metabolite gene organization. *Proc. Natl. Acad. Sci. USA*, 122(17): e2504934122. DOI: <https://doi.org/10.1073/pnas.2504934122>.
- Dong, S., Han, X., Man, X., Deng, X., & Chen, Z., (2025). Tissue-specific biosynthesis and regulation of alkaloids, flavonoids, and terpenoids in fenugreek (*Trigonella foenum-graecum* L.): insights from integrated metabolomics and transcriptomics analysis. *Front Plant Sci.*, 16:1669610. DOI: <https://doi.org/10.3389/fpls.2025.1669610>
- Eddy, S.R., (2011). Accelerated profile HMM searches. *PLoS Comput Biol.*, 7(10):e1002195. DOI: <https://doi.org/10.1371/journal.pcbi.1002195>.

- Gabriel, L., Bruna, T., Hoff, K.J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M., (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.*, 34(5):769-777. DOI: <https://doi.org/10.1101/gr.278090.123>.
- Gaikwad, A.B., Yadav, S., Kumari, R., Maurya, W., Rangan, P., Singh, R., & Singh, G.P., (2025). Chromosome-scale genome assembly of *Trigonella corniculata* (L.) L. (Nagauri pan/Kasuri methi), an important spice. *Sci Data.*, 12:509. DOI: <https://doi.org/10.1038/s41597-025-04858-4>.
- Hoang, D.T., Chernomor, O., von, Haeseler, A., Minh, B.Q., & Vinh, L.S., (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.*, 35(2):518-522. DOI: <https://doi.org/10.1093/molbev/msx281>.
- Javan, S.L., Kashkooli, A.B., Shojaeiyan, A., & Majidian, S., (2024). Transcriptomic data reveals the dynamics of terpenoids biosynthetic pathway of fenugreek. *BMC Genomics*, 25:390. DOI: <https://doi.org/10.1186/s12864-024-10253-x>.
- Katoh, K., & Standley, D.M., (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.*, 30(4):772-780. DOI: <https://doi.org/10.1093/molbev/mst010>.
- Lahuta, L.B., Szablińska, J., Ciak, M., & Górecki, R.J., (2018). The occurrence and accumulation of D-pinitol in fenugreek (*Trigonella foenum graecum* L.). *Acta Physiol Plant.*, 40:155. DOI: <https://doi.org/10.1007/s11738-018-2734-4>.
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D. & Bateman, A., (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.*, 49(D1): D412-D419. DOI: <https://doi.org/10.1093/nar/gkaa913>.
- Naika, M.B.N., Sathyanarayanan, N., Sajeevan, R.S., Bhattacharyya, T., Ghosh, P., Iyer, M.S., Jarjapu, M., Joshi, A.G., Harini, K., Mohamed Shafi, K., Kalmankar, N., Karpe, S.D., Mam, B., Pasha, S.N. & Sowdhamini, R., (2022). Exploring the medicinally important secondary metabolites landscape through the lens of transcriptome data in fenugreek (*Trigonella foenum graecum* L.). *Sci Rep.*, 12:13534. DOI: <https://doi.org/10.1038/s41598-022-17779-8>.
- Nevers, Y., Vesztrocy, A.W., Rossier, V., Train, C.M., Altenhoff, A.M., Dessimoz, C. & Glover, N.M., (2025). Quality assessment of gene repertoire annotations with OMArk. *Nat Biotechnol.*, 43(1):124-133. DOI: <https://doi.org/10.1038/s41587-024-02147-w>.
- Wang, M., Zhang, S., Li, R., & Zhao, Q., (2024). Unraveling the specialized metabolic pathways in medicinal plant genomes: a review. *Front Plant Sci.*, 15:1459533. DOI: <https://doi.org/10.3389/fpls.2024.1459533>.
- Wong, T.K.F., Ly-Trong, N., Ren, H., Demotte, P., Baños, H., Roger, A.J., Susko, E., Bielow, C., De Maio, N., Goldman, N., Hahn, M.W., dos Reis, M., Vinh, L.S., Huttley, G., Lanfear, R., & Minh, B.Q., (2026). IQ-TREE 3: phylogenomic inference software using complex evolutionary models. *Mol Biol Evol.*, 43(5): msag117. DOI: <https://doi.org/10.1093/molbev/msag117>.
- Yadav, S., Maurya, W., Kumari, R., Sharma, S., Rangan, P. & Gaikwad, A.B., (2026). Identification of candidate genes governing key metabolic pathways in fenugreek (*Trigonella spp.*) through integrated transcriptomic and metabolomic analysis. *Funct Integr Genomics.*, 26:122. DOI: <https://doi.org/10.1007/s10142-026-01893-1>.
- Zhao, Q., Wu, G., Yang, P., Shi, Y., Fu, Z., Mo, H., Shi, C., & Yu, S., (2024). Metabolomic and transcriptomic analyses reveal the molecular mechanism underlying the massive accumulation of secondary metabolites in fenugreek (*Trigonella foenum-graecum* L.) seeds. *Genes (Basel)*, 15(3):343. DOI: <https://doi.org/10.3390/genes15030343>.