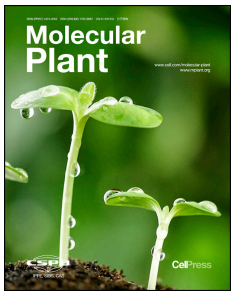
Journal Pre-proof



Evolution and functional characterization of a biosynthetic gene cluster for saponin biosynthesis in Sapindaceae

Yingxiao Mai, Huimin Hu, Wenjuan Ji, Yaxuan Xiao, Hong Zhou, Zaohai Zeng, Wenshu Lv, Xingling Su, Jiakun Zheng, Jing Xu, Yanwei Hao, Zhenhua Liu, Rui Xia

PII: S1674-2052(25)00139-X

DOI: https://doi.org/10.1016/j.molp.2025.04.013

Reference: MOLP 1902

To appear in: *MOLECULAR PLANT*

Received Date: 13 December 2024

Revised Date: 4 March 2025

Accepted Date: 29 April 2025

Please cite this article as: Mai Y., Hu H., Ji W., Xiao Y., Zhou H., Zeng Z., Lv W., Su X., Zheng J., Xu J., Hao Y., Liu Z., and Xia R. (2025). Evolution and functional characterization of a biosynthetic gene cluster for saponin biosynthesis in Sapindaceae. Mol. Plant. doi: https://doi.org/10.1016/j.molp.2025.04.013.

Evolution and functional characterization of a biosynthetic gene cluster for saponin biosynthesis in Sapindaceae

Yingxiao Mai[a,1], Huimin Hu[a,1], Wenjuan Ji[b,1], Yaxuan Xiao[a], Hong Zhou[b], Zaohai Zeng[a], Wenshu Lv[a], Xingling Su[a], Jiakun Zheng[a], Jing Xu[a], Yanwei Hao[a], Zhenhua Liu[b,*], Rui Xia[a,*]

[a] Guangdong Basic Research Center of Excellence for Precise Breeding of Future Crops, State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops (South China) at Ministry of Agriculture and Rural Affairs, College of Horticulture, South China Agricultural University, Guangzhou, Guangdong, 510642, China

[b] Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China

1 These authors contribute equally to this work

* Correspondence: Rui Xia (rxia@scau.edu.cn), Zhenhua Liu (zhenhua.liu@sjtu.edu.cn)

Short Summary:

We identified a biosynthetic gene cluster (BGC) comprising *BAS*, *CYP716A*, and *UGT87* that underlies saponin biosynthesis in Sapindaceae. Comparative genomic analyses reveal that the cluster arose through stepwise gene-translocation events, with UGT87 incorporated via a reciprocal chromosomal translocation. These findings offer key insights into the evolutionary processes that drive plant BGC assembly.

Dear Editor,

Saponins are one of the most abundant and diverse groups of plant natural products, playing critical roles in plant defense against disease and herbivores, and some saponins have gained recognition for their pharmaceutical importance. Saponins are glycosides of triterpenes, with their triterpene backbones formed through the cyclization of the 2,3-oxidosqualene by oxidosqualene cyclase (OSC). The triterpenes are oxidized mostly by cytochrome P450 (CYP450), and glycosylated by UDP-glycosyltransferases (UGTs) or cellulose synthase-like enzymes (CSLs), to form diverse saponins (Abe et al., 2004; Seki et al., 2015). Recently, it has been found that in plant, like in microbes, enzymatic genes in charge of the biosynthesis of specialized metabolites cloud be clustered within specific genomic regions, known as biosynthetic gene clusters (BGCs) (Nützmann et al., 2016). BGCs have been widely reported in plants, involving in the synthesis of various metabolites (Nützmann and Osbourn, 2014). How distant metabolic genes across chromosomes are assembled into BGCs is largely unknown.

Sapindaceae, comprising four subfamilies, 144 genera, and over 1,900 species (Buerki et al., 2009). It is also known as the soapberry family, because of its founding species soapberry (*Sapindus mukorossi*), which is well-known for its richness in saponins (Xu et al., 2022). Natural saponins in soapberry are often used as non-ionic surfactant in daily washing products like soap and detergent. Additionally, aesculin, an important medicinal saponin extracted from the dry seeds of horse chestnut (*Aesculus chinensis*), is used to treat stomach distention and chest pain(Sun et al., 2023). The fruit tissue of yellowhorn (*Xanthoceras sorbifolia*), litchi (*Litchi chinensis*) and other Sapindaceae species have been reported to contain saponins as well (Kilari and Putta, 2016; Venegas-Calerón et al., 2017) (Fig. S1A). To confirm the broad presence of saponins in Sapindaceae species, we collected the pericarp tissues for six of these species (Fig.S1B) and profiled their saponin composition using liquid chromatography-mass spectrometry (LC-MS). Indeed, multiple saponins were detected (Table S1), with the soapberry pericarp containing up to eight different types of saponins (Fig. S1C). Two saponins (#1 and #8) were identified across multiple species (Fig. S1C), indicating potential conserved saponin biosynthesis pathways within the Sapindaceae family.

The OSC including *CAS* (Cycloartenol synthase), *BAS* (β-amyrin synthase), and other functional subclasses in charge of the first committed step in triterpenoid biosynthesis (Abe et

al., 2004; Xue et al., 2012). The members in the putative *BAS* clade showed generally higher expression level than other *OSCs*, including the *BASs* previously reported in horse chestnut (Sun et al., 2023) and *Arabidopsis*(Shibuya et al., 2009) (Fig. S2). It was observed that these putative *BASs* with high expression in Sapindaceae species were all located in collinear genomic blocks, implying a potentially conserved biological role of them (Fig. 1A and S2). More intriguingly, genes encoding *CYP450s* for oxidation and *UGTs* for glycosylation, are co-located within the collinear blocks of *BASs*, forming a BGC in Sapindaceae (Fig. 1A). Each BGC in these Sapindaceae species consists of a single *BAS*, 1-7 copies of *CYP450s*, and 3-9 copies of *UGTs* arranged within a 120-820 kb region, and some of the BGCs also contain 1-2 *CSLs* (Fig. 1A). The function of *BASs* was verified by co-expressing them with truncated *HMGR* (3-hydroxy-3-methylglutaryl-coenzyme A reductase), the rate-controlling enzyme of the mevalonate pathway, using agrobacterium-mediated transient expression system in the leaves of tabacco (*Nicotiana benthamiana*). The production of β-amyrin was detected by Gas Chromatography Mass Spectrometry (GC-MS) for all the *BASs* tested (*XsBAS* from yellowhorn, *SmBAS* from soapberry, and *LcBAS* from litchi), except the *ChBAS* from balloon-vine (Fig. 1B; Fig. S3). This suggests that these Sapindaceae *BASs* are mostly functionally active. And phylogenetic analysis revealed that these *CYP450s* and *UGTs* in the BGCs belong to CYP716A and UGT87A categories, respectively (Fig. S4). CYP716A was reported to mediate the oxidation of β-amyrin to the triterpenoid acids, such as oleanolic acid (OA) (Miettinen et al., 2017). However, research on the function of UGT87A in secondary metabolite biosynthesis remains scarce, with its glycosylation activity reported only in *Eucalyptus* (Hansen et al., 2023). In order to dissect the role of the collinear BGCs in saponin biosynthesis, clustered genes in two characteristic Sapindaceae species, litchi and soapberry, were selected for further analyses.

In BGC of litchi, both the *LcBAS* and *LcCYP716A-3* exhibit a comparable expression pattern in pericarp, with remarkable expression during early stages of pericarp development in litchi (Fig. S5). Similarly, in BGC of soapberry, *SmBAS* shows a similar early expression pattern with *SmCYP716A-2/-3/-4* in the pericarp (Fig. 1C). To investigate the function of these *BAS* and *CYP716A* genes, we infiltrated *tHMGR* with both *LcBAS* and *LcCYP716A-3* (L2), as well as *SmBAS* and *SmCYP716A-2/-3/-4* (S2) into tobacco leaves for metabolic profiling. The chromatographic peak of L2 and S2 detected by LC-MS is consistent with that of the OA (Fig.

1D; Retention time (RT):15.15 minutes), validating the function of *LcCYP716A*-3 and three *SmCYP716As* in mediating the oxidization of β-amyrin synthesized by BAS into OA. The notably higher expression levels of *SmBAS* and *SmCYP716As*, compared to their counterparts in litchi (Table S5 and S6), would enormously facilitate the biosynthesis of OA, likely contributing to the high abundance of oleanane-type saponins in soapberry.

Both *LcUGT87A-3/-4/-6/-7* and *SmUGT87A-3.1/-3.2/-5* displayed a relatively low-level expression in the pericarp tissue, although not as stage-specific as the *BAS* and *CYP716As* in the BGCs (Fig. 1C and S5). In contrast, the two litchi *CSLs* were expressed at extremely low levels in pericarp (Fig. S5 and Table S6). Therefore, we chose to further validate the function of these *UGT87As* by expressing them in *E. coli* and performing *in vitro* enzyme assay with OA and UDP-glucose as donor substrates (Fig. 1E and S8A). A new product peak (RT at 12.24-12.26 minutes) was observed in reactions mediated by all UGT87As (Fig. 1E and S8). The peak showed similar MS/MS spectrum, especially m/z at 617.405 and 663.415 corresponding to those of pericarp tissues (Fig. S6C and S8B). The RT of the glycosylated product of UGT87As is distinct to that of the standard oleanolic acid 28-O-glucopyranosyl ester (12.54 minutes; Fig. 1E and S8), suggesting that the glycosylation reaction may occur at the C3 position of OA. Therefore, the *BASs* and *CYP716As* in the BGCs play a conserved role in the biosynthesis of OA in litchi and soapberry (Fig. 1F); *UGT87As* in the BGC may further direct the C3 glycosylation of OA (Fig. 1F). Although *UGT87As* are expressed at relatively low levels compared to the high expression of *BAS* and *CYP716A*s, all the genes are biologically functional, forming a novel and active saponin BGC in Sapindaceae.

Most of reported saponin BGCs have independently originated, indicating limited evolutionary relationships among them. The BGC of oat saponin (avenacin) were reformed in the subtelomeric region of the genome, which lacks homologous sequences in other grasses (Li et al., 2021). Similarly, the saponin BGC of astragalosides does not have homologous genes in closely related species of the Fabaceae family (Xu et al., 2024). The conservation of the saponin BGC in Sapindaceae thus provided a rare suitable case to explore its evolutionary trajectory in modern plants. Collinear analysis of the genes in the BGC revealed gymnosperms and basal angiosperms have only a single *CAS*-containing syntenic block (Fig. 1G). In contrast, monocots and eudicots possess multiple *CAS*-containing blocks (Fig. 1G and S9), likely

resulting from ancient whole genome duplication (WGD) events. In eudicots, three corresponding syntenic blocks were formed (Fig. 1G), aligning with the ancient gamma (γ) triplication event specific to core eudicots. The first collinear block underwent the loss of *CAS*, while the second one retained the *CAS* intact across species (Fig. 1G). Notably, the third homologous block identifies not only *CASs* but also putative *BASs* (Fig. 1G). In the chromosomal 9 of the *Vitis* genome, there are tandem repeats composed of *CAS* and *BAS*, suggesting the potential neofunctionalization of *CAS* into *BAS* (Xue et al., 2012). Within the clade a of eudicots, like certain lineages in Malvids, *BAS* gradually replaced *CAS* as the oxidosqualene-cyclizing gene in the syntenic region of the BGC (Fig. 1G). Subsequently, *CYP716A(s)* were recruited into the physical proximity of *BAS* in Malvids (clade b; Fig. 1G), except for the species in Brassicaceae (clade f; Fig. 1G). Afterward, *UGT87As* were integrated into the vicinity of *BAS* and *CYP716A* in species of the Sapindales (clade c; Fig. 1G). And in some Sapindaceae species (clade d) *CSLs* were further introduced into the BGC (Fig. 1G). In short, the *CYP716As* and *UGT87As* were sequentially recruited into the vicinity of BASs to collectively form saponin BGCs in Sapindales.

Next, we ask how these genes of tailoring enzymes are recruited to the neighboring region of *BAS*. It has been reported that transposable elements (TE) contribute to the BGC formation in eudicots(Boutanaev and Osbourn, 2018). Then we checked the TE content in the franking region of these three gene families, *BAS*, *CYP716A* and *UGT87A*, and found that the TE content of *CYP716A* is generally higher than that of *BAS, UGT87A* and other genes (Fig. S10A). Within the CYP450 family, the neighboring region of *CYP716A* also bear relatively higher TE abundance than many other clans or subclans of *CYP450* genes (Fig. S10B). Thus, transposon activity has likely facilitated the translocation of *CYP716A* to the proximity of *BAS* in Malvids.

In contrast, the TE content in the neighboring region of *UGT87A* is significantly lower than that of *OSC* and *CYP716A* (Fig. S10A). However, we found that *UGT87As* are preserved within conserved syntenic blocks with well-maintained collinearity, not only in the genome of Sapindales species bearing the BGC but also in many other eudicot species (Fig. S11). Interestingly, the syntenic blocks of *UGT87As* and *BAS(-CYP716As)* are on different chromosomes until the emergence of Sapindales, where these two blocks were joined together on a single chromosome in close proximity (Fig. 1H and S11). Further karyotype evolution

146 analysis uncovered that reciprocal chromosomal translocation occurred in the common
147 ancestor of Sapindales probably enabled the joining of these two syntenic blocks (Fig. 1H and
148 S11).

149 Specifically, in *Vitis*, *Fragaria*, and *Theobroma*, *BAS(-CYP716As)* and *UGT87A*s are
150 located on two separate chromosomes, homologous to ACEK4 (ancient core eudicot karyotype
151 4) and ACEK6, respectively (Fig. 1H and S12). However, in Sapindales, these two
152 chromosomal blocks were brought together on a single chromosome, while retailing segmental
153 homology to ACEK4 and ACEK6 (Fig. 1H and S13). Conceivably, an ancient reciprocal
154 chromosomal translocation event resulted in a segmental exchange between the ACEK4-
155 homologous chromosome (containing the *BAS-CYP716A* block) and the ACEK6-homologous
156 (containing the *UGT87A* block), leading to the fusion of these two syntenic blocks into a single
157 *BAS-CYP716A-UGT87A* block (Fig. 1H and Fig. S13). Therefore, the other tailoring enzyme,
158 UGT87A, was brought into close proximity with BAS and CYP716A through a reciprocal
159 chromosomal translocation event.

160 Taken together, our study reports the discovery of a saponin BGC conserved in
161 Sapindaceae species and offers a compelling demonstration of the evolutionary process of its
162 formation. Initially, the single-copy *BAS* in the saponin BGC was derived from the WGD-γ event
163 and tandem duplication of *CAS* (Fig. 1I). Subsequently, the activity of transposon elements
164 likely facilitated the translocation of *CYP716A* near *BAS*, a proximity that was maintained in the
165 ancestral species of Malvids (clade b; Fig. 1G and 1I). In the common ancestor of the
166 Sapindales order (clade c; Fig. 1G), reciprocal chromosomal translocation led to the fusion of
167 *UGT87s* with the proximity of *BAS* and *CYP716A*, ultimately forming the saponin BGC (Fig. 1I).
168 Within some Sapindaceae species, *CSLs* were later integrated into the BGC. Therefore, the
169 saponin BGC in Sapindaceae was formed through the sequential recruitment of *CYP450* and
170 *UGT* to the proximity of the skeleton-forming *BAS* to achieve coordinated expression and co-
171 inheritance, ensuring the efficient biosynthesis of saponins.

172 The discovery of the conserved saponin BGC reveals the underlying genetic basis of the
173 richness of saponins in the fruits of Sapindeace species, particularly the soapberry. The
174 involvement of chromosomal reciprocal translocation in the formation of BGC offers novel
175 insight into the evolution of BGC in plants. Future investigations utilizing functional validation,

comparative genomics, and synthetic biology could further exploit the potential of these BGC, advancing biotechnological engineering of saponin biosynthesis or molecular breeding of sapindaceous plants.

Acknowledgments

This work is supported by the Key Area Research and Development Program of Guangdong Province (2022B0202070003). This work is also supported by the National Science Foundation of China (#32072547 and #32372665).

Author Contributions:

R.X. and Z.H.L. coordinated the research. Y.X.M. conducted the comparative genomic analysis, evolutionary analysis, and other bioinformatics analysis. H.M.H., W.J.J, and Y.X.X. carried out a validation experiment for the metabolic experiments. All the authors participated the data interpretation, and edited the manuscript, and approved the final manuscript.

Reference

Abe, I., Sakano, Y., Sodeyama, M., Tanaka, H., Noguchi, H., Shibuya, M., and Ebizuka, Y. (2004). Mechanism and Stereochemistry of Enzymatic Cyclization of 24,30-Bisnor-2,3-oxidosqualene by Recombinant β-Amyrin Synthase. *J. Am. Chem. Soc.* **126**:6880–6881.

Boutanaev, A. M., and Osbourn, A. E. (2018). Multigenome analysis implicates miniature inverted-repeat transposable elements (MITEs) in metabolic diversification in eudicots. *Proceedings of the National Academy of Sciences* **115**:E6650–E6658.

Buerki, S., Forest, F., Acevedo-Rodríguez, P., Callmander, M. W., Nylander, J. A. A., Harrington, M., Sanmartín, I., Küpfer, P., and Alvarez, N. (2009). Plastid and nuclear DNA markers reveal intricate relationships at subfamilial and tribal levels in the soapberry family (Sapindaceae). *Molecular Phylogenetics and Evolution* **51**:238–258.

Li, Y., Leveau, A., Zhao, Q., Feng, Q., Lu, H., Miao, J., Xue, Z., Martin, A. C., Wegel, E., Wang, J., et al. (2021). Subtelomeric assembly of a multi-gene pathway for antimicrobial defense compounds in cereals. *Nat Commun* **12**:2563.

Nützmann, H.-W., and Osbourn, A. (2014). Gene clustering in plant specialized metabolism. *Current Opinion in Biotechnology* **26**:91–99.

Nützmann, H.-W., Huang, A., and Osbourn, A. (2016). Plant metabolic clusters – from genetics to genomics. *New Phytologist* **211**:771–789.

Seki, H., Tamura, K., and Muranaka, T. (2015). P450s and UGTs: Key Players in the Structural Diversity of Triterpenoid Saponins. *Plant and Cell Physiology* **56**:1463–1471.

Shibuya, M., Katsube, Y., Otsuka, M., Zhang, H., Tansakul, P., Xiang, T., and Ebizuka, Y. (2009). Identification of a product specific β-amyrin synthase from *Arabidopsis thaliana*. *Plant Physiology and Biochemistry* **47**:26–30.

Sun, W., Yin, Q., Wan, H., Gao, R., Xiong, C., Xie, C., Meng, X., Mi, Y., Wang, X., Wang, C., et al. (2023). Characterization of the horse chestnut genome reveals the evolution of aescin and aesculin biosynthesis. *Nat Commun* **14**:6470.

Xu, Y., Zhao, G., Ji, X., Liu, J., Zhao, T., Gao, Y., Gao, S., Hao, Y., Gao, Y., Wang, L., et al. (2022). Metabolome and transcriptome analysis reveals components regulating triterpenoid saponin biosynthesis of soapberry Advance Access published March 1, 2022, doi:10.1101/2022.02.28.482332.

Xu, B., Huang, J.-P., Peng, G., Cao, W., Liu, Z., Chen, Y., Yao, J., Wang, Y.-J., Li, J., Zhang, G., et al. (2024). Total biosynthesis of the medicinal triterpenoid saponin astragalosides. *Nat. Plants* **10**:1826–1837.

Xue, Z., Duan, L., Liu, D., Guo, J., Ge, S., Dicks, J., ÓMáille, P., Osbourn, A., and Qi, X. (2012). Divergent evolution of oxidosqualene cyclases in plants. *New Phytologist* **193**:1022–1038.

Figure legend:

Figure 1 Discovery, functional characterization, and evolution of a conserved saponin BGC in Sapindaceae.

A. The collinearity of the saponin BGC across seven Sapindaceae species.

B. Gas chromatography mass spectrometry (GC-MS) analysis of extracts of *N. benthamiana* leaves following *Agrobacterium*-mediated transient expression. Leaves were agroinfiltrated with expression constructs of different combinations, with *tHMGR* (control) infiltrated together with *BASs* from yellowhorn (*XsBAS),* soapberry (*SmBAS*), litchi (*LcBAS*), balloon-vine (*ChBAS*), respectively. The mass spectrum (Fig. S3) of the new peak is identified as β-amyrin (m/z 426.386).

C. Expression level of genes within the BGC in pericarp of soapberry. DAP, day after pollination. TPM, transcripts per million.

D. LC-MS chromatograms show the production of oleanolic acid (m/z 455.353) in infiltrated tobacco leaves which co-expressed with *tHMGR + LcBAS* (L1), *tHMGR+ LcBAS +LcCYP716A-3* (L2), *tHMGR+ SmBAS* (S1), and *tHMGR+ SmBAS +SmCYP716A-2/3/4* (S2).

E. LC-MS chromatograms show the products of the soapberry pericarp and the products generated by *SmUGT87A-3.1*, *SmUGT87A-3.2* and *SmUGT87A-5* in *E. coli*, featuring EIC corresponding to the $[M+HCOO]^-$ (m/z 663.410). EIC, Extracted Ion Chromatogram.

F. A diagram illustrating the biosynthetic pathway directed by genes within the saponin BGC.

G. A phylogenetic tree of 25 represent species from four major clades: gymnosperms, angiosperms basal group (ANA), monocots, and eudicots. And comparative syntenic analysis of BGC regions reveals the evolutionary process of the formation of the saponin BGC in the Sapindaceae. MYA, million years ago.

H. Karyotype analysis (with reference to ACEK ancestral karyotype) for the chromosomes where the BGCs are located. The 21 chromosomes of ACEK were denoted by different colors. The red-highlighted collinear blocks represent the regions where the target genes are located. And the orange, blue, and green arrows indicate the locations of *BAS*, *CYP716A*, and *UGT87A*, respectively, on the chromosomes.

I. An evolutionary model summarizing the formation process of the saponin BGC in Sapindaceae. The purple, orange, blue, green, and navy-blue arrows denote *CAS*, *BAS*, *CYP716A*, *UGT87A* and *CSLG*, respectively. And the white arrows represent the corresponding genes that are absent from the genome.