ADVANCED RESEARCH JOURNAL OF BIOCHEMISTRY AND BIOTECHNOLOGY
WWW.SPRINGJOURNALS.NET

Advanced Research Journal of Biochemistry and Biotechnology, Vol. 10(7): pp 1-9, July, 202026.

Full Length Research Paper

Molecular Analyses of Genetic Variation and Phylogenetic Relationship in the Family Sapindaceae. A Review Paper

P.K. Langat*, E.W. Njenga[2], P. Jeruto[2], C.N. Nyamwamu[2], C.W. Lukhoba[3] and F.M. Musila[4].

[1]Department of Biological Sciences, University of Eldoret, Kenya
[2]Department of Biological Sciences, University of Eldoret, Kenya
[3]Department of Biological Sciences, University of Nairobi, Kenya
[4]Department of Biological and Life Sciences, Technical University of Kenya, Kenya

Abstract

The Sapindaceae, commonly known as the soapberry family, is a cosmopolitan group of approximately 1900 species across 144 genera, forming part of the economically and ecologically significant angiosperm order Sapindales. Despite prior taxonomic efforts, relationships within Sapindaceae and across Sapindales have remained poorly resolved due to complex morphological variation and incomplete infra-familial classification systems. Recent advances in molecular systematics, particularly the use of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences and Angiosperms353 target enrichment datasets, have enabled substantial progress in reconstructing evolutionary relationships within this group. ITS-based phylogenetic analyses have confirmed species-level resolution within Indian *Sapindus*, clearly distinguishing *S. emarginatus* from *S. trifoliatus*, and revealing the divergent position of *S. oligophyllus*, which clusters with *Allophylus* of tribe Thouinieae. Estimates of evolutionary divergence revealed significant variability among tribes, with the greatest divergence observed between Paullinieae and Harpullieae (0.20) and the least between Sapindeae and Lepisantheae (0.06), supporting past taxonomic hypotheses. Complementary phylogenomic analyses using Angiosperms353 markers across 123 Sapindaceae genera (86% coverage) recovered 21 clades, providing the basis for a revised classification into four subfamilies and 20 tribes, including six newly proposed tribes within Sapindoideae. Broader Sapindales-wide analyses comprising 448 samples and 85% of genera confirmed family monophyly and resolved core clades while also revealing persistent challenges in subfamily-level relationships due to paralogy, likely linked to ancient hybridisation and polyploidy events. The presence of paralogous loci, particularly in Meliaceae and Rutaceae, affirms the need for careful data curation and highlights the impact of ancient genome duplications on phylogenetic inference. This integrated molecular framework provides the most comprehensive phylogenetic resolution of Sapindaceae and Sapindales to date. It offers a robust foundation for future evolutionary, biogeographic, taxonomic, and conservation-orientated studies while emphasising the need for continued sampling and the incorporation of genomic complexity in phylogenetic reconstruction.

Key Words: Sapindaceae, phylogenomics, ITS sequences, Angiosperms353, infra-familial classification

Accepted 1/7/2025 Published 6/8/.2025

INTRODUCTION

Molecular markers, especially the internal transcribed spacer (ITS) region of nuclear ribosomal DNA, have become instrumental in resolving phylogenetic relationships in plants due to their high variability and ease of amplification (Alvarez & Wendel, 2003; Kress et al., 2005). Several studies have successfully employed

ITS sequences to explore plant systematics, offering clearer resolution where morphological data fall short (Ghada et al., 2013; Du et al., 2014; Michel et al., 2016).

Despite progress in understanding the broader systematics of Sapindaceae using morphological, palynological, phytochemical, and molecular data (Muller & Leenhouts, 1976; Buerki et al., 2009, 2010, 2011a), the phylogeny of Indian Sapindus species remains inadequately studied using molecular tools. This gap necessitates rigorous molecular-based phylogenetic studies to resolve the classification and evolutionary history of the genus.

The classification of the Sapindaceae family has historically been fraught with challenges, originating from Radlkofer's early work (Radlkofer, 1931–1934) and onwards (Meyer, 1976). Although early morphological taxonomies provided initial insights, they lacked an evolutionary basis. Molecular phylogenetics has since revealed high levels of polyphyly and paraphyly in traditional subfamilial and tribal classifications (Harrington et al., 2005; Buerki et al., 2009). For instance, genera like Arytera, Cupaniopsis, and Haplocoelum were shown to be non-monophyletic, necessitating taxonomic revisions (Buerki et al., 2010).

The four-subfamily system proposed by Harrington et al. (2005), Dodonaeoideae, Hippocastanoideae, Sapindoideae, and Xanthoceratoideae, provided a foundation, yet many genera remained unplaced due to incomplete sampling. Notably, the tribe Melicocceae has been rendered non-monophyletic, and several morphological groupings proposed by Radlkofer have since been overturned by molecular evidence (Buerki et al., 2011a). Further compounding these issues, biogeographic analyses indicate that Sapindaceae likely originated in Eurasia during the Late Cretaceous and spread southward via several dispersal routes, with significant diversification in the Southern Hemisphere (Buerki et al., 2013). The family comprises over 140 genera, including economically vital species like Litchi chinensis, Dimocarpus longan, and Paullinia cupana. However, the delimitation of several genera remains uncertain due to insufficient molecular data, and some African genera are poorly understood or improperly placed (Acevedo-Rodríguez et al., 2011).

The genus Sapindus L., comprising approximately 13 recognised species, is distributed across tropical and subtropical regions of Asia, Australia, and North and South America (Xia & Gadek, 2007). Within India, the taxonomy of Sapindus has long been contentious, with considerable disagreement among taxonomists due to overlapping morphological characteristics. Early classification, such as Hiern (1875), proposed seven species. However, subsequent revisions, including those by Leenhouts (1969), resulted in the reclassification of certain species into the genus Lepisanthes due to evolving generic concepts. Further inconsistencies were observed in taxonomic treatments provided by Cooke (1902), Gamble (1918), and others, particularly concerning species like Sapindus trifoliatus, S. emarginatus, and S. laurifolius. Some taxonomists considered these taxa as synonyms, while others maintained their distinctiveness based on subtle phenotypic traits (Gamble, 1918; Pant, 2000). This taxonomic ambiguity affirms the need for molecular investigations to delineate infraspecific and interspecific relationships within Sapindus and other closely allied taxa of the Sapindaceae family.

Recent additions and synonymizations, such as the description of Gereaua and the merger of Tinopsis into Tina, reflect ongoing adjustments to generic boundaries as new data emerge (Buerki et al., 2010a, 2011a; Callmander et al., 2011). Within the broader context of Sapindales, the order comprises nine families—six medium to large (e.g., Sapindaceae, Meliaceae, and Anacardiaceae) and three smaller ones (e.g., Biebersteiniaceae and Kirkiaceae)—and encompasses around 479 genera and 6750 species (Joyce et al., 2023).

Economically, Sapindales species contribute over US$31 billion annually, supplying fruits, timber, pharmaceuticals, and ornamental plants (Joyce et al., 2023). Despite its significance, the internal relationships within Sapindales remain unresolved. Various studies have reconstructed Sapindaceae as sister to different clades within the order, sometimes linked with Anacardiaceae and Burseraceae (Gadek et al., 1996), and other times with Rutaceae, Simaroubaceae, and Meliaceae (Muellner-Riehl et al., 2016). These discrepancies, often based on low statistical support, have hindered a robust understanding of Sapindales' evolutionary history.

Recent phylogenomic advances, particularly in high-throughput sequencing and target capture methodologies, offer promising avenues for resolving these complex relationships. The Angiosperms353 bait kit, designed for single- or low-copy nuclear genes, enables comprehensive phylogenetic reconstructions at family and ordinal levels (Johnson et al., 2019; Baker et al., 2022). This method also facilitates the detection of paralogous loci, which, while historically excluded due to analytical challenges, can yield insights into whole-genome duplications and gene-tree discordance (Smith & Hahn, 2021). Given the prevalence of polyploidy and gene duplications in angiosperms (Soltis et al., 2009), retaining and analysing these loci is critical for understanding lineage-specific evolutionary dynamics.

The integration of molecular phylogenetics through ITS markers and target capture sequencing has the potential to resolve long-standing taxonomic ambiguities in Sapindus, clarify infrageneric and intergeneric relationships in Sapindaceae, and elucidate broader evolutionary patterns within Sapindales. A comprehensive and well-supported phylogenetic framework will not only aid systematic and taxonomic research but also inform conservation and biogeographic studies, given the ecological and economic importance of many Sapindales taxa.

MATERIALS AND METHODS

Plant Material

Studies have employed a multi-scale phylogenetic approach integrating both traditional and high-throughput molecular techniques. Initially, plant materials consisted of 25 accessions representing three Sapindus species native to India: nine accessions of S. emarginatus, six of S. mukorossi, and seven of S. trifoliatus. These were collected across diverse ecological regions in India to capture intra-species genetic variability (Mahar et al., 2011a, 2011b, 2013). Thirteen accessions of closely related taxa within Sapindaceae were also included. All voucher specimens were deposited at the CSIR-National Botanical Research Institute (LWG), Lucknow. For broader phylogenetic coverage, 72 accessions were evaluated, encompassing the 22 Sapindus samples and 50 others representing 18 genera of the Sapindaceae family (APG III, 2009). Furthermore, a global phylogenomic framework was constructed using 472 accessions, including 448 samples from all nine families of Sapindales. These samples were drawn from the Royal Botanic Gardens, Kew DNA and Tissue Collection, field collections, multiple herbarium institutions, and DNA banks. The out-group comprised 24 samples from Pentapetalae, including orders such as Brassicales, Ericales, Fabales, and others.

DNA Extraction and Library Preparation

Genomic DNA was extracted from fresh and silica-dried leaves using the CTAB method (Doyle and Doyle, 1987, 1990), with modifications for herbarium and degraded materials, such as adjusted precipitation times. DNA quality and quantity were assessed using the NanoDrop 1000 Spectrophotometer, Quantus Fluorometer, and gel electrophoresis. For degraded samples, multiple extractions were pooled and concentrated. Samples with fragment sizes >350 bp were sheared using a Covaris M220 ultrasonicator. Libraries were constructed using the NEBNext Ultra II DNA Library Prep Kit with dual indexing. Quality checks were performed using TapeStation and library concentrations normalized to 10 nM prior to pooling (20–24 samples/pool).

Targeted Enrichment and Sequencing

Two sequencing strategies were used. The initial ITS-based sequencing focused on amplifying the ITS region using universal primers P4 and P5 (White et al., 1990; Baldwin, 1992). PCR conditions followed Allan and Porter (2000), and products were visualised on agarose gel and purified using the QIAquick Gel Extraction Kit (Qiagen, Germany). Sequencing was conducted on an ABI 3730 automated sequencer, and electropherograms were analysed with ABI PRISM® DNA Sequencing Analysis Software v.5.0. In contrast, high-throughput phylogenomic sequencing used the Angiosperms353 probe set (Johnson et al., 2019) with hybridisations at 60–65°C for 24 hours. Enriched libraries were amplified with the KAPA HiFi ReadyMix PCR Kit and sequenced on Illumina MiSeq or HiSeq platforms by Macrogen (2 × 150 bp paired-end reads).

Gene Recovery and Sequence Processing

For phylogenomic analysis, raw reads were trimmed with Trimmomatic (Bolger et al., 2014), and exon sequences were assembled using HybPiper v1.2 (Johnson et al., 2016), mapping reads against the Sapindales subset of the mega target file (McLay et al., 2021). HybPhaser v1 (Nauheimer et al., 2021) was applied for paralogy detection and contamination filtering. Samples with >50% missing loci or loci with poor recovery were removed. Sequences with >1% SNPs were considered paralogous. Cleaned sequences were aligned using MAFFT (Katoh and Standley, 2013), with sites containing >75% missing data filtered out via Phyutility (Smith and Dunn, 2008). Loci were concatenated using AMAS (Borowiec, 2016) for supermatrix analysis.

Phylogenetic Analysis

For the ITS-based analysis, 22 Sapindus and 13 related accessions were aligned using ClustalW (Thompson et al., 1994) in MEGA-6 (Tamura et al., 2013), using Sapindus oligophyllus as a reference for ITS region boundaries (Buerki et al., 2009). Sequence variation statistics and evolutionary divergence were calculated using maximum likelihood, and phylogenetic trees were generated with 1,000 bootstrap replicates.

In the phylogenomic approach, gene trees were inferred with RAxML-NG v0.9.0 using the GTR+G model and 100 bootstrap replicates (Kozlov et al., 2019). A concatenated maximum likelihood tree was estimated with IQ-TREE (Nguyen et al., 2015), using ModelFinder Plus and 1,000 ultrafast bootstraps (Lanfear et al., 2012; Kalyaanamoorthy et al., 2017; Hoang et al., 2018). Coalescent species trees were generated using ASTRAL v5.7.8 (Zhang et al., 2018), with low-support branches collapsed using Newick Utils (Junier and Zdobnov, 2010) and outlier branches removed by TreeShrink (Mai and Mirarab, 2018). Node support was categorised as low (BS < 90, PP < 0.9), moderate (BS = 90–97, PP = 0.9–0.97), high (BS = 97–99, PP = 0.97–0.99), and maximum (BS = 100, PP = 1.0).

Divergence Time Estimation

Molecular dating incorporated 29 fossil calibrations (Parham et al., 2012) placed at stem nodes of respective

clades. To accommodate angiosperm age uncertainties, three divergence scenarios were considered based on Ramírez-Barahona et al. (2020): CC-complete (140.33–144.29 Ma), RC-complete (143.91–147.94 Ma), and UC-complete (212.25–221.02 Ma). BEAST v2.6.6 (Bouckaert et al., 2019) was used with a fixed topology, relaxed log-normal clock, and birth-death prior. Ten independent MCMC runs of 50 million generations were sampled every 1,000 generations. Convergence was assessed in Tracer v1.7.2 (Rambaut et al., 2018), and consensus trees were constructed with logCombiner and TreeAnnotator. Sensitivity analyses employed alternative tree priors and calibration distributions.

RESULTS

Evolutionary Divergence among Sapindaceae Tribes

The molecular analyses using the nuclear ribosomal internal transcribed spacer (ITS) region offered substantial insights into the genetic diversity and evolutionary relationships within *Sapindus* and its allied taxa in the family Sapindaceae. Estimations of evolutionary divergence among Sapindaceae tribes revealed varying genetic distances, with the highest divergence (0.20) occurring between the tribes Paullinieae and Harpullieae, suggesting a deeper evolutionary split. In contrast, the lowest divergence (0.06) was observed between Sapindeae and Lepisantheae, indicating a close evolutionary relationship. This finding aligns with past taxonomic observations that led to reclassifications of species between these tribes (Leenhouts, 1969).

Species-Level Delimitation within Indian *Sapindus*

Phylogenetic reconstruction based on ITS sequences clearly separated the three Indian *Sapindus* species *S. emarginatus*, *S. mukorossi* and *S. trifoliatus* into well-supported, distinct clusters, affirming their species-level differentiation. A significant molecular similarity was noted between the Japanese species *S. delavayi* and *S. mukorossi*, supported by a bootstrap value of 99%, suggesting either recent divergence or historical gene flow between these geographically separated taxa. Particularly notable was the robust molecular distinction between *S. emarginatus* and *S. trifoliatus*, resolving long-standing taxonomic ambiguities and contradicting their previous treatment as synonyms or varietal forms in Indian flora.

Phylogenomic Resolution of Sapindaceae Genera and Tribes

Parallel to the ITS-based results, high-throughput sequencing using the Angiosperms353 target enrichment approach generated a rich phylogenomic dataset. Sequence data were successfully retrieved for 123 genera, representing 86% of the recognised genera in Sapindaceae, with an average recovery rate of 335 out of 353 targeted genes per accession. Although 19 genera were excluded due to poor DNA quality or lack of available material, prior phylogenetic data enabled the taxonomic placement of 11 of them.

Phylogenetic trees inferred using both concatenated (RAxML) and coalescence-based (ASTRAL III) methods revealed nearly identical topologies, differing only in the placement of Melicoccus bijugatus. The RAxML-based phylogeny placed Melicoccus as sister to Talisia, Tripterodendron, and Dilodendron, corroborating earlier studies (Buerki et al., 2009, 2011b). Altogether, 21 highly supported clades were resolved, including four monotypic clades, forming the foundation of a revised tribal classification of Sapindaceae.

The analyses also supported the existence of four subfamilies (Acevedo-Rodríguez et al., 2011) and recognised 20 tribes. Notably, clades such as the Litchi and Blomia groups, which had previously received limited support, were now well-resolved. Additionally, phylogenetic positions were confidently assigned to 10 previously unplaced genera.

Locus Recovery and Paralogy Across Sapindales

The extended Sapindales-wide dataset, comprising 472 samples including 24 outgroup taxa, further reinforced the robustness of phylogenetic inferences. On average, 324 loci were recovered per sample, with 74% target coverage. Data cleaning using HybPhaser removed an average of 15 loci per sample due to signs of contamination or excessive SNP divergence.

A total of 28.55% of loci were flagged as potentially paralogous, defined by a threshold of >1% SNPs. Notably, levels of paralogy varied across families, with Meliaceae showing the highest (mean 51% ± 4.11) and Kirkiaceae the lowest (mean 11% ± 3.89). These trends persisted even under stricter criteria for SNP thresholds.

Phylogenetic Reconstruction of Sapindales

The concatenated alignment used for phylogenetic inference consisted of 330 loci, totalling 194,132 base pairs with 135,613 parsimony-informative sites and 14.73% gaps or ambiguities. IQ-TREE identified 45 optimal partitions and produced a maximum-likelihood phylogeny after 173 tree searches and 1,000 bootstrap replicates, with a final log-likelihood of -10,642,396. Both concatenated and coalescent-based analyses provided full support for the monophyly of Sapindales and its constituent families. Nitrariaceae was consistently recovered as sister to all other Sapindales families, supported by bootstrap values of 96 and posterior

probability of 0.97, thus refining the deep-node relationships within the order.

Integrated Molecular Framework for Sapindaceae Phylogeny

Collectively, the results integrate traditional ITS markers with high-resolution phylogenomic data, offering a comprehensive and reliable framework for understanding evolutionary relationships within Sapindaceae and across the order Sapindales. The convergence of outcomes across different datasets and analytical approaches underscores the robustness of the inferred phylogenies and provides a strong basis for taxonomic re-evaluation and future comparative evolutionary studies within this diverse plant lineage.

DISCUSSIONS

Evolutionary Divergence Among Sapindaceae Tribes

The study revealed notable evolutionary divergence among the tribes of the family Sapindaceae, based on nuclear ribosomal DNA internal transcribed spacer (ITS) sequences and phylogenomic data. The greatest divergence (0.20) was observed between Paullinieae and Harpullieae, suggesting a deep evolutionary split, while the least divergence (0.06) was found between Sapindeae and Lepisantheae. This finding affirms previous reports of a close genetic affinity between Sapindus and Lepisanthes (Harrington et al., 2005). ITS-based phylogenetic trees generated through maximum likelihood analysis also reinforced this pattern as Lepisanthes species formed a sister clade to Sapindus with 71% bootstrap support. Additionally, several tribes, including Doratoxyleae, Harpullieae, Dodonaeeae, and Schleichereae, clustered together, suggesting shared ancestry, though some relationships received modest bootstrap support. Koelreuterieae and Paullinieae showed clear separation, highlighting substantial divergence. These observations confirm the ITS region's reliability as a marker for estimating evolutionary divergence among tribes.

Species-Level Delimitation within Indian *Sapindus*

Maximum likelihood analyses provided clear molecular resolution among the Indian species of Sapindus, strongly supporting their species-level distinctness. Sapindus emarginatus and S. trifoliatus were placed in separate clades with 99% bootstrap support, affirming their status as distinct species and negating previous assumptions of infraspecific variation. Interestingly, S. oligophyllus diverged significantly from the other Indian Sapindus species and grouped with Allophylus of tribe Thouinieae, indicating potential taxonomic misplacement.

Earlier taxonomists had debated the classification of S. oligophyllus, oscillating its placement across genera such as Aphania, Sapindopsis, Howethoa, and Sapindus (Rauschert, 1982; Xia and Gadek, 2007). The current results validate the species-level classification proposed by Prakash and Mehrotra (1990) and Pant (2000) while contradicting earlier morphological interpretations. These findings affirm the necessity of integrating molecular data into taxonomic evaluations, particularly in morphologically variable genera such as Sapindus.

Phylogenomic Resolution of Sapindaceae Genera and Tribes

Using the Angiosperms353 targeted sequencing approach, the study achieved high-resolution phylogenomic insights into Sapindaceae. With successful gene recovery from 123 genera (86% of all recognised genera), the analysis reinforced the existence of four previously described subfamilies (Acevedo-Rodríguez et al., 2011), which led to the recognition of 20 tribes.

The placement of previously ambiguous genera such as Melicoccus, Tristiropsis, and Guindilia was clarified, and genera from the economically vital Litchi group, such as Litchi, Nephelium, and Dimocarpus, were confidently grouped into one clade. New tribes were proposed based on unique morphological or genetic features in monotypic genera like Blomia and Guindilia. These results show a high congruence with earlier phylogenetic frameworks (Buerki et al., 2009, 2011a, 2011b), reinforcing the robustness of the new tribal classification.

Locus Recovery and Paralogy Across Sapindales

Comprehensive locus recovery was achieved across 472 Sapindales samples, including 24 outgroup species. On average, 324 loci per sample were recovered, and after cleaning through HybPhaser, loci affected by contamination or lab errors were removed. Approximately 28.55% of angiosperm 353 loci were identified as paralogous due to having >1% SNPs, with variation observed across families. Meliaceae exhibited the highest levels of paralogy (51% of loci with >1% SNPs), while Kirkiaceae showed the lowest (11%). This heterogeneity reflects differences in genome complexity or duplication histories among Sapindales families. Importantly, defining paralogous loci is critical for constructing accurate phylogenies, especially in diverse lineages such as Sapindaceae.

Phylogenetic Reconstruction of Sapindales

Phylogenetic analyses using both concatenated and multispecies coalescent approaches yielded highly resolved topologies. The concatenated alignment

comprised 194,132 base pairs, with 135,613 parsimony-informative sites. IQ-TREE identified 45 optimal partitions with tailored substitution models, and the final consensus tree had strong bootstrap and posterior probability support across nodes.

Notably, Nitrariaceae was recovered as a sister to all other Sapindales families (BS = 96; PP = 0.97), confirming earlier studies. The placement of Melicoccus bijugatus differed slightly between RAxML and ASTRAL trees, yet this variation did not affect broader clade resolution. Ultimately, the data confirm the monophyly of Sapindales and each constituent family.

Integrated Molecular Framework for Sapindaceae Phylogeny

Combining nuclear ITS sequences and extensive phylogenomic data from target enrichment has yielded an integrated and comprehensive framework for the phylogeny of Sapindaceae. The ITS region effectively resolved inter-specific and inter-generic relationships, while the Angiosperms353 dataset facilitated deeper tribal and subfamilial classifications.

The identification of well-supported clades, refined placement of problematic taxa, and recognition of new tribes enhance our understanding of evolutionary trajectories in Sapindaceae. These results have significant implications for taxonomy, conservation, and crop improvement, especially in economically important clades like the Litchi group. The study emphasises the need for continued phylogenomic investigations and broader taxon sampling to resolve the remaining uncertainties in this complex and diverse plant family.

CONCLUSION

The integrated molecular approach has significantly advanced the understanding of genetic variation, phylogenetic relationships, and taxonomic boundaries within the Sapindaceae family. Through the application of nuclear ribosomal DNA ITS sequences and phylogenomic data from the Angiosperms 353 loci, the study has provided robust evidence of evolutionary divergence among various tribes and affirmed the phylogenetic integrity of the four previously recognised subfamilies. Notably, the deep divergence observed between Paullinieae and Harpullieae and the close genetic affinity between Sapindeae and Lepisantheae reflect distinct evolutionary trajectories within the family and corroborate prior molecular and morphological studies.

At the species level, the clear separation of Sapindus emarginatus and S. trifoliatus with strong bootstrap support affirms their distinct taxonomic identities, resolving longstanding controversies regarding their infraspecific classifications. Furthermore, the divergent placement of S. oligophyllus, aligning more closely with Allophylus of the Thouinieae tribe, suggests the need for a taxonomic reassessment of this species. These findings highlight the critical role of molecular markers like the ITS region in clarifying complex taxonomic issues that cannot be adequately resolved by morphology alone.

Phylogenomically, the study marks a major step forward in the systematic classification of Sapindaceae. The high recovery of Angiosperms (353 loci from 123 genera) allowed for a refined and largely resolved phylogeny that supports the formal recognition of 20 tribes. The accurate placement of previously ambiguous or unplaced genera, along with the definition of new monotypic tribes, demonstrates the power of target enrichment strategies in resolving deep and shallow phylogenetic relationships. In particular, the improved resolution of the Litchi group home to economically important species has significant implications for conservation, breeding programs, and understanding evolutionary diversification in tropical tree crops.

The assessment of locus recovery and paralogy also provided critical insights into the genomic architecture of Sapindales. Variation in paralogous gene content across families underscores the need for tailored data-cleaning strategies when working with high-throughput phylogenomic data. Moreover, the well-supported phylogenetic trees constructed using both concatenated and coalescent approaches confirm the monophyly of Sapindales and its constituent families while also validating the placement of Nitrariaceae as sister to the rest of the order.

Thus, the combination of ITS-based and phylogenomic analyses presents a robust and comprehensive molecular framework for the taxonomy and evolutionary study of Sapindaceae. This framework not only resolves previously ambiguous relationships but also offers a foundation for future research into functional trait evolution, biogeography, and conservation of this ecologically and economically vital plant family.

REFERENCES

Acevedo-Rodríguez PP, van Welzen F, Adema RW, van der Ham JM. (2011). Sapindaceae. In K. Kubitzki [ed.], Flowering plants. Eudicots. The families and genera of vascular plants, vol. 10. Springer, Berlin, Germany.

Allan GJ, Porter JM. (2000). Tribal delimitation and phylogenetic relationships of Loteae and Coronilleae (Faboideae: Fabaceae) with special reference to Lotus: evidence from nuclear ribosomal ITS sequences. Am. J. Bot. 87, 871–1881.

Alvarez I, Wendel JF. (2003). Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution, 29(3), 417–434.

APG III. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society, 161(2), 105–121.

Baker WJ, Bailey P, Barber V, Barker A, Bellot S, Bishop D. (2022). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. Syst. Biol. 71, 301–319. doi: 10.1093/sysbio/syab035

Baldwin BG. (1992). Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositaogy. Mol. Phylogenet. Evol. 1(1), 3–16.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Borowiec ML. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. PeerJ 4, e1660. doi: 10.7717/peerj.1660

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Drummond AJ. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Computational Biology, 15(4), e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Buerki S, Forest F, Acevedo-Rodríguez P, Callmander MW, Nylander JAA, Harrington MG, Sanmartin I, Kupfer P, Alvarez N. (2009). Plastid and nuclear DNA markers reveal intricate relationships at subfamilial and tribal levels in the soapberry family (Sapindaceae). Mol. Phylogenet. Evol. 51, 238–258.

Buerki S, Lowry II PP, Alvarez N, Razafimandimbison SG, Kupfer P, Callmander MW. (2010). Phylogeny and circumscription of Sapindaceae revisited: molecular sequence data, morphology and biogeography support recognition of a new family, Xanthoceraceae. Plant. Ecol. Evol. 143, 148–159.

Buerki S, Forest F, Alvarez N, Nylander JAA, Arrigo N, Sanmartin I. (2011a). An evaluation of new parsimony-based versus parametric inference methods in biogeography: a case study using the globally distributed plant family Sapindaceae. J. Biogeogr. 38, 531–550.

Buerki S, Forest F, Salamin N, Alvarez N. (2011b). Comparative performance of supertree algorithms in large data sets using the soapberry family (Sapindaceae) as a case study. Syst. Biol. 60(1), 32–44.

Buerki S, Forest F, Stadler T, Alvarez N. (2013). The abrupt climate change at the Eocene-Oligocene boundary and the emergence of South-East Asia triggered the spread of sapindaceous lineages. Annals of Botany, 112(1), 151-160.

Callmander MW, Buerki S, Phillipson PB. (2011). Nomenclatural changes in the Malagasy endemic genus Tina Schult. (Sapindaceae). Candollea 66: 124–126.

Cooke T. (1902). The Flora of the Presidency of Bombay. Botanical Survey of India, Calcutta.

Doyle J, Doyle JL. (1987). Genomic plant DNA preparation from fresh tissue- CTAB method. Phytochem. Bull. 19, 11–15.

Doyle JJ, Doyle JL. (1990). Isolation of plant DNA from fresh tissue. Focus, 12(1), 13–15.

Du YP, He HB, Wang ZX, Li S, Wei C, Yuan XN, Cui Q, Jia GX. (2014). Molecular phylogeny and genetic variation in the genus Lilium native to China based on the internal transcribed spacer sequences of nuclear ribosomal DNA. J. Plant Res. 127, 249–263.

Gamble JS. (1918). Flora of the Presidency of Madras. vol. I Botanical Survey of India, Calcutta.

Gadek PA, Fernando ES, Quinn CJ, Hoot SB, Terrazas T, Sheahan MC, Chase MW. (1996). Sapindales: molecular delimitation and infraordinal groups. Am. J. Bot. 83, 802–811.

Ghada B, Ahmed BA, Messaoud M, Amel SH. (2013). Genetic diversity and molecular evolution of the internal transcribed spacer (ITSs) of nuclear ribosomal DNA in the Tunisian fig cultivars (Ficus carica L.; Moraceae). Biochem. Syst. Ecol. 48, 20–33.

Harrington MG, Edwards KJ, Johnson SA, Chase MW, Gadek PA. (2005). Phylogenetic inference in Sapindaceae sensu lato using plastid matK and rbcL DNA sequences. Systematic Botany, 30(2), 366–382.

Hiern WP. (1875). Sapindaceae. In: Hooker, J.D. (Ed.), The Flora of British India. vol I Reeve & Co, London.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution, 35(2), 518–522. https://doi.org/10.1093/molbev/msx281

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target

enrichment. Applications in Plant Sciences, 4(7), 1600016. https://doi.org/10.3732/apps.1600016

Johnson MG, Pokorny L, Dodsworth S, Botigue LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT, Leebens-Mack JH, Leitch IJ, Maurin O, Soltis DE, Soltis PS, Wong GKS, Baker WJ, Wickett NJ. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Systematic Biology, 68(4), 594–606. https://doi.org/10.1093/sysbio/syy086

Junier T, Zdobnov EM. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. Bioinformatics, 26(13), 1669–1670. https://doi.org/10.1093/bioinformatics/btq243

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods, 14(6), 587–589. https://doi.org/10.1038/nmeth.4285

Katoh K, Standley DM. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics, 35(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Kress WJ, Wurdack KJ, Zimmer EA, Weight LA, Janzen DH. (2005). Use of DNA barcodes in phylogenetic studies. Proceedings of the National Academy of Sciences USA, 102(23), 8369–8374.

Lanfear R, Calcott B, Ho SYW, Guindon S. (2012). PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701. doi: 10.1093/molbev/mss020

Leenhouts PW. (1969). Florae Malesianae Praecursores. A revision of Lepisanthes (Sapindaceae). Blumea 17, 33–91.

Mai U, Mirarab S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics, 19(5), 272. https://doi.org/10.1186/s12864-018-4620-2

Mahar KS, Rana TS, Ranade SA. (2011a). Molecular analyses of genetic variability in soap nut (Sapindus mukorossi Gaertn.). Ind. Crop. Prod. 34(1), 1111–1118.

Mahar KS, Rana TS, Ranade SA, Meena B. (2011b). Genetic variability and population structure in Sapindus emarginatus Vahl from India. Gene 485(1), 32–39.

Mahar KS, Rana TS, Ranade SA, Pande V, Palni LMS. (2013). Estimation of genetic diversity and population structure in Sapindus trifoliatus L., using DNA fingerprinting methods. Trees-Struct. Funct. 27(1), 85–96.

McLay TGB, Birch JL, Gunn BF, Ning W, Tate JA, Nauheimer L. (2021). New targets acquired: Improving locus recovery from the Angiosperms353 probe set. Appl. Plant Sci. 9, e11420. doi: 10.1002/aps3.11420

Meyer FG. (1976). A revision of the genus Koelreuteria (Sapindaceae). Journal of the Arnold Arboretum, 57(2), 129-166.

Michel CI, Meyer RS, Taveras Y, Molina J. (2016). The nuclear internal transcribed spacer (ITS2) as a practical plant DNA barcode for herbal medicines. J. Appl. Res. Med. Arom. Plant. 3, 94–100.

Muellner-Riehl AN, Weeks A, Clayton JW, Buerki S, Nauheimer L, Chiang YC. (2016). Molecular phylogenetics and molecular clock dating of sapindales based on plastid rbcL, atpB and trnL-trnF DNA sequences. Taxon 65, 1019–1036. doi: 10.12705/655.5

Muller J, Leenhouts PW. (1976). A general survey of pollen types in Sapindaceae in relation to taxonomy. In: Ferguson, I.K., Muller, J. (Eds.), The Evolutionary Significance of the Exine. Academic Press, London.

Mau, L. et al. (2021). HybPhaser: A workflow for the detection and phasing of hybrids in target capture data sets. Appl Plant Sci, 9, 43–51. doi: 10.1002/aps3.11441

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Molecular Biology and Evolution, 32(1), 268–274. https://doi.org/10.1093/molbev/msu300

Pant PC. (2000). Sapindaceae. In: Singh, N. P., Bohra, J. N., Hajra, P. K., Singh, D. K. (Eds.), Flora of India. vol. V Botanical Survey of India, Calcutta.

Parham JF, Donoghue PCJ, Bell CJ, Calway TD, Head JJ, Holroyd PA, Inoue JG, Irmis RB, Joyce WG, Ksepka DT, Patané JSL, Smith ND, Tarver JE, van Tuinen M, Yang Z, Angielczyk KD, Greenwood JM, Hipsley CA, Jacobs LL, … Benton MJ. (2012). Best practices for justifying fossil calibrations. Systematic Biology, 61(2), 346–359. https://doi.org/10.1093/sysbio/syr107

Prakash V, Mehrotra BN. (1990). Indian species of

Sapindus L. (Sapindaceae). J. Econ. Taxon. Bot. 14, 75–79.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. (2018). Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst. Biol. 67, 901–904. doi: 10.1093/sysbio/syy032

Ramírez-Barahona S, Sauquet H, Magallón S. (2020). The delayed and geographically heterogeneous diversification of flowering plant families. Nature Ecology & Evolution, 4.

Rauschert S. (1982). Nomina nova generica et combinationes novae Spermatophytorum et Pteridophytorum. Taxon 31, 554–563.

Smith SA, Dunn CW. (2008). Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics, 24(5), 715–716. https://doi.org/10.1093/bioinformatics/btm619

Smith ML, Hahn MW. (2021). New approaches for inferring phylogenies in the presence of paralogs. Trends in Genetics, 37(2), 174-187.

Soltis PS, Soltis DE. (2009). The role of hybridization in plant speciation. Annual review of plant biology, 60, 561-588.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution, 30(12), 2725–2729. https://doi.org/10.1093/molbev/mst197

Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22(22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

White TJ, Bruns T, Lee S, Taylor J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In PCR protocols: A guide to methods and applications (pp. 315–322). Academic Press.

Xia N, Gadek PA. (2007). Flora of China. 12. pp. 1–6. Available from www.efloras.org.

Zhang M, Dai S, Du B, Ji L, Hu S. (2018). Mid-Cretaceous hothouse climate and the expansion of early angiosperms. Acta Geol. Sin. 92, 2004–2025. doi: 10.1111/1755-6724.13692