Botanical Journal of the Linnean Society, 2023, **XX**, 1–14
https://doi.org/10.1093/botlinnean/boad022
Advance access publication 6 July 2023
Original Article

OXFORD

Original Article

Target capture sequencing for phylogenomic and population studies in the Southeast Asian genus *Palaquium* (Sapotaceae)

Aireen Phang[1,2,3,*], Flávia Fonseca Pezzini[1], David F.R.P. Burslem[3], Gillian S. Khew[2], David J. Middleton[2], Markus Ruhsam[1], Peter Wilkie[1]

[1]Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK
[2]Singapore Botanic Gardens, National Parks Board, 1 Cluny Road, 259569 Singapore
[3]School of Biological Sciences, University of Aberdeen, Aberdeen AB24 3UU, UK

[*]Corresponding Author. Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK. E-mail: aphang@rbge.org.uk

ABSTRACT

The first phylogenomic study using a custom target capture bait panel within the Isonandreae tribe in Sapotaceae is presented. The combination of taxon-specific and universal loci from the Angiosperms353 probe set achieved high species resolution within the genus *Palaquium* and provides the first phylogenomic insights into Malesian representatives of Sapotaceae, where relationships between taxa often remain poorly understood. The results highlight that generic and some species circumscriptions require further investigation and possible revision: *Aulandra longifolia* is found to be nested in *Palaquium*, and *P. rostratum* within *P. microphyllum*. Population structure analysis produced limited resolution between and within species, but the bait set was able to recover parameters that are potentially useful in population genomic studies.

Keywords: molecular systematics; *Palaquium*; population genomics; Sapotaceae; Southeast Asia; target capture

INTRODUCTION

The genus *Palaquium* Blanco, containing ~120 species of tropical trees (POWO 2022), has a range that extends from India to Australasia and the western Pacific, with a centre of diversity in Malesia (the distinctive floristic region composed of Brunei, Indonesia, Malaysia, Papua New Guinea, the Philippines, Singapore, and Timor-Leste). Its habitat ranges from lowland to montane forest and swamp to coastal regions (Ng 1972). The most extensive revision of the genus published to date recognized 109 species in Malesia (Royen 1960), and taxonomic work has continued in subsequent regional accounts (Ng 1972, Chai and Yii 2002, Chantaranothai 2014, Jessup 2019). Notable species include the well-known and economically important *Palaquium gutta* (Hook.) Baill., the latex (known as gutta-percha) of which was used in the insulation of early deep-sea telegraph cables that greatly facilitated transatlantic communication. Although its usage has now declined with the advent of synthetic materials, gutta-percha is still used in dentistry as a material for fillings (Bansal *et al.* 2020). Several species of *Palaquium*, including the heavy hardwood of *Palaquium ridleyi* King & Gamble, have also been used in silviculture (Burkill 1935). Pennington (1991) recognized five tribes, 53 genera, and around 1100 species in his morphological synthesis of the Sapotaceae family: Chrysophylleae (19 genera), Isonandreae (which contains *Palaquium* as well as *Aulandra, Burckella, Diploknema, Isonandra, Madhuca* and *Payena*), Mimusopeae (17 genera), Omphalocarpeae (four genera), and Sideroxyleae (6 genera). Most of Sapotaceae, including representatives of *Palaquium*, were found to be diploid, with polyploidy being rare and mostly occurring in Sideroxyleae (Johnson 1991, Smedmark and Anderberg 2007).

Phylogenetic classification

The earliest family-wide phylogenetic studies based on the plastid marker *ndh*F determined that Isonandreae *sensu* Pennington was paraphyletic (Swenson and Anderberg 2005), and the authors proposed a three-subfamily classification of Sapotoideae, Sarcospermatoideae, and Chrysophylloideae. In this new classification, the subfamily Sapotoideae includes the tribes Sideroxyleae and Sapoteae, the latter of which contains 11 genera, including those previous included in Isonandreae such

Received 23 August 2022; revised 28 February 2023; accepted 15 May 2023

as *Madhuca*, *Payena*, and *Palaquium*. However, sampling was limited to one species representative of *Palaquium*, and relationships among genera in Pennington's (1991) Isonandreae have remained largely unresolved in all subsequent studies (Smedmark *et al.* 2006, Gautier *et al.* 2013, Richardson *et al.* 2014). Crown node estimation for Sapotaceae has fluctuated between studies, from 58.3 Mya (Rose *et al.* 2018) to 107 Mya (Richardson *et al.* 2014), but it is generally accepted that rapid speciation occurred in the later history of the family, with Richardson *et al.* (2014) finding that almost one-third of all species sampled across Isonandreae evolved during the Pleistocene.

The last molecular phylogenetic study based on ITS nuclear data (which contained the largest sampling of *Palaquium* species; around 40 accessions) found the genus to be non-monophyletic, with representatives of the genera *Aulandra* and *Diploknema* nested in the *Palaquium* clade (Richardson *et al.* 2014). Although there was a well-supported clade containing all *Palaquium* species, interspecific relationships were mostly unresolved. This echoes the complexities in previous morphological classifications, where Royen (1960) noted that his recognition of seven groups in the genus was 'without sharp distinction', due to overlapping characters used to distinguish species (e.g. size of leaves, leaf venation, indumentum). Most phylogenetic studies also did not support the groups proposed by Royen (1960) and have demonstrated that many morphological characters previously deemed as diagnostic are homoplasious, with characters such as staminodes lost and then re-emerging in certain clades (Swenson and Anderberg 2005, Mackinder *et al.* 2016). It is clear that intensive phylogenetic research is needed to clarify generic as well as species relationships and limits within the tribe Isonandreae *sensu* Pennington (1991), especially since the bulk of its representatives occur in Asia where many areas face severe deforestation pressures (Gautier *et al.* 2013).

Previous molecular phylogenetic studies have been based on Sanger sequencing of selected region-specific markers (e.g. ITS, ETS, *trnS-trnG*, *matK*, and *trnL-trnF*), and the intrinsic limitations of these markers (which normally have a median length of <1000 bp) have led to calls for augmentation with Next Generation Sequencing (NGS) approaches (Coissac *et al.* 2016, Wilkinson *et al.* 2017). More recently NGS methods of target capture using baits (or probes) to capture numerous orthologous, single/low-copy target loci across a genome, have exponentially raised the number of phylogenetically informative characters available and been successfully applied to clarify lineage evolution and recent radiations (Nicholls *et al.* 2015, Shah *et al.* 2021; Yardeni *et al.* 2022; Boluda *et al.* 2022). Baits used for target capture can be tailored to a group (taxon-specific kits, such as in Couvreur *et al.* 2019 or Straub *et al.* 2020), or designed to apply to a wider range of taxa (universal kits), for instance across vertebrates (Lemmon *et al.* 2012) or angiosperms (Johnson *et al.* 2019). Recent studies have highlighted the benefits and limitations of each approach, particularly in relation to their efficiency at lower taxonomic and microevolutionary scales (Kadlec *et al.* 2017, Chau *et al.* 2018, Slimp *et al.* 2021, Yardeni *et al.* 2022). Taxon-specific kits are expected to achieve higher target capture success and more phylogenetic resolution than the more conserved loci in universal kits (Woudstra *et al.* 2021). There is strong evidence that useful sequence data can be extracted from degraded herbarium and museum specimens using a target capture approach (Hart *et al.* 2016, McCormack *et al.* 2016, Johnson *et al.* 2019; O'Connell *et al.* 2022), which opens up the massive potential of collections and can mitigate limitations in collecting fresh material from a wide range of species and geographical areas.

Population genetic applications

In addition to phylogenomic studies, target sequencing loci have also been shown to be applicable to population genomic analyses and the assessment of demographic parameters such as heterozygosity and introgression (Choo *et al.* 2020; Slimp *et al.* 2021; O'Connell *et al.* 2022). Unlike RADseq approaches that require higher quality starting DNA and may result in allele-drop out (missing data) that hampers deeper phylogenetic analyses (Andrews *et al.* 2016, Leaché and Oaks 2017), target capture overcomes these limitations by being applicable to degraded DNA, with high target loci coverage allowing for the detection of rare alleles (Chung *et al.* 2016), as well as paralogues (Yardeni *et al.* 2022). Such demographic applicability can bolster assessments of intra-specific variation for taxonomic purposes and provide insight into the genetic diversity of populations.

The recent creation and application of a Sapotaceae-specific bait set for the study of Malagasy genera in the endemic tribe Tseboneae (Gautier *et al.* 2013, Christe *et al.* 2021, Boluda *et al.* 2022) provides an ideal opportunity to assess its utility for other tribes in the family from different geographical areas such as Malesia. Coupled with the ongoing *Flora of Singapore* project (Middleton *et al.* 2019), this bait set can be used to test a dense sampling of *Palaquium* species occurring in a small tropical island such as Singapore (Fig. 1) and in turn assess the validity of existing morphological species limits, particularly identifying where there may be incomplete isolation of lineages. As it is possible for recently diverged species to be non-monophyletic in gene trees, an additional aim was to increase phylogenetic signal by expanding sampling from the usual one or two samples per species to at least six accessions per species. The Sapotaceae bait set included 1019 loci in total: 531 family-specific monocopy genes, 227 Short Tandem Repeat (STR, which are short repetitive motifs with around one to six bases) markers as well as 261 of the 353 (~73%) single-copy loci from the Angiosperms353 universal target panel (Johnson *et al.* 2019) that were selected and retained according to proximity to the Sapotaceae reference transcriptomes (Christe *et al.* 2021). The combination of universal and taxon-specific loci in the same bait set afforded an opportunity to compare the efficiency of each locus type for phylogenetic informativeness, as well as test their ability to generate demographic markers that can reliably guide population-level studies.

MATERIALS AND METHODS

Plant material sampling and outgroups

Multiple accessions of individual *Palaquium* species across their natural range in Singapore were sampled in preparation for phylogenetic analysis. To check for potential conspecificity within the distribution of the genus and assess the validity of putative taxonomic relationships, accessions in Peninsular Malaysia and/or Borneo were also included for each species wherever possible. Herbarium and silica-dried material held at

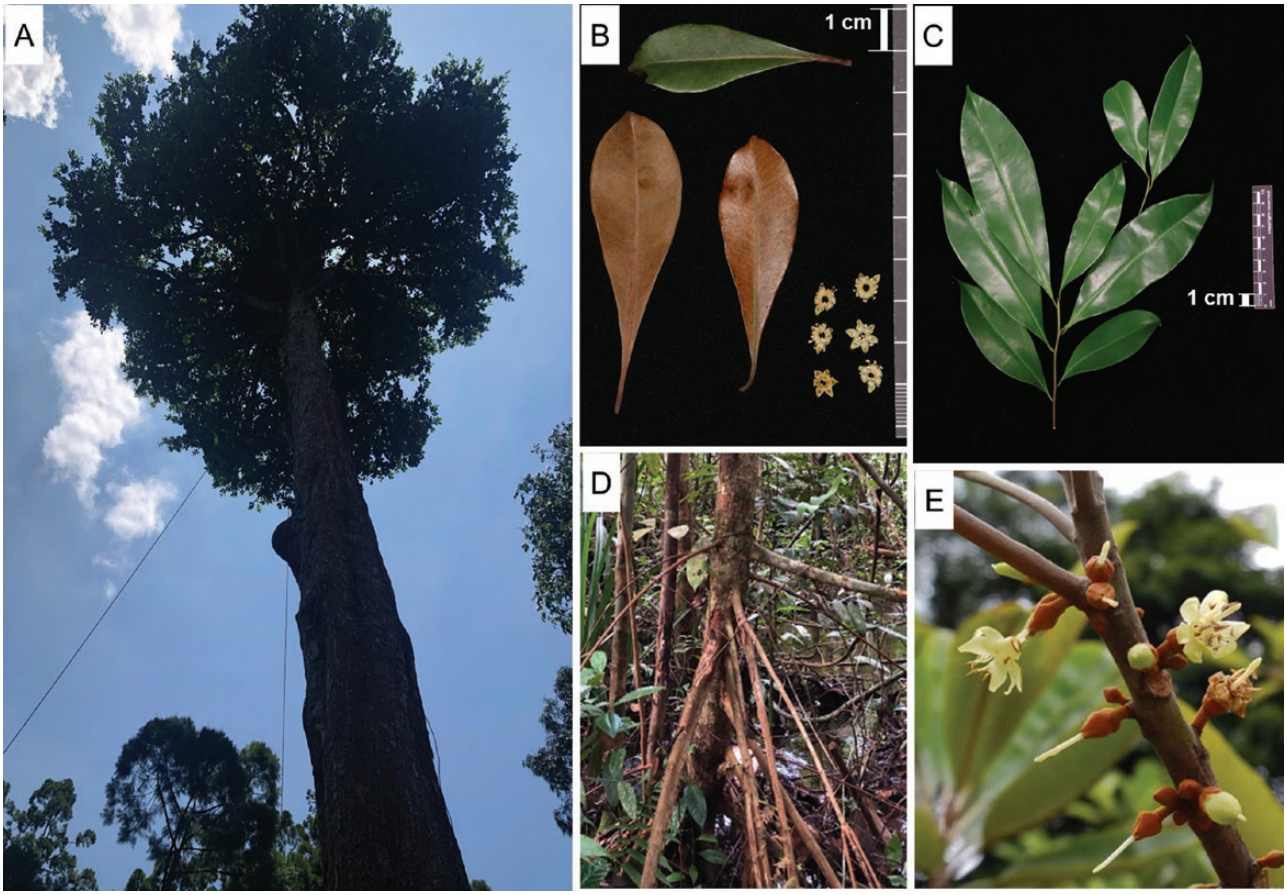

Figure 1. Diversity of representative species of *Palaquium* in Singapore. A, Heritage tree of *P. obovatum* at the Singapore Botanic Gardens, probably planted around 1897, standing at 31 m tall. B, Fallen leaves and corollas of *P. microphyllum*. C, Leaves of *P. rostratum* attached to branch. D, Stilt roots of *P. xanthochymum* in freshwater swamp forest. E, Inflorescences of *P. gutta*, displaying typical flower clusters in axils of leaf scars. Photographs: A, D, A. Phang; B, C, X.Y. Ng and E, P.K.F. Leong.

SING and E (herbaria acronyms follow Thiers, continuously updated) were assessed for suitability for DNA extraction, and only collections within the last 25 years were selected to minimize extraction failure due to sample DNA degradation. Fieldwork was conducted in Singapore's nature reserves to collect nine new vouchers for species where herbarium and silica-dried material was insufficient.

In all, 77 unique accessions (20 silica-dried and 57 herbarium samples) representing eight *Palaquium* species occurring in Singapore were sampled: *Palaquium gutta*; *P. hexandrum* (Griff.) Baill.; *P. impressinervium* Ng; *P. microphyllum* King & Gamble; *P. obovatum* (Griff.) Engl.; *P. oxleyanum* Pierre; *P. rostratum* (Miq.) Burck; and *P. xanthochymum* (de Vr.) Pierre. *Palaquium ridleyi*, the sole locally extinct species, was not sampled due to lack of available material (in Singapore only one collection, in 1892, is verifiable: the type specimen is found in K) (Supporting Information, Table S1).

Raw sequence data of nine outgroup taxa were downloaded from the NCBI Sequence Read Archive. These included eight samples generated with similar target capture methods for the study of Christe *et al.* (2021) and one for the Kew Plant and Fungal Trees of Life Project (Baker *et al.* 2022) and represented four genera in the tribe Isonandreae and one from each of the tribes Chrysophylleae, Gluemeae, Sapoteae, Sideroxyleae, and Tseboneae (Pennington 1991, Swenson and Anderberg 2005, Gautier *et al.* 2013) (Supporting Information, Table S1). Although tribal placement has not yet been fully resolved (Christe *et al.* 2021), Sideroxyleae and Chrysophylleae have been identified as outgroups that diverged early within the family (Boluda *et al.* 2022), hence taxa from these tribes were used to root phylogenies.

DNA extraction and library preparation

Genomic DNA were extracted from leaf material using a modified cetyl-tri-methylammonium bromide method, with 24:1 chloroform: isoamyl alcohol and overnight precipitation in isopropanol at −20°C (Doyle and Doyle 1987, Cullings 1992). Incubation timings before the addition of chloroform: isoamyl alcohol were increased to 12 hours for recalcitrant extracts. The samples were purified with SeraMag magnetic carboxylate-modified microparticles (Cytiva, USA) diluted to 5%, and the concentration of DNA quantified with a Nanodrop Spectrophotometer (Thermo Scientific, Singapore).

For library preparation, 100–400 ng of DNA per sample was used with 400 ng achieved for 70% of the samples. Then 80% of the extracts were sheared with a sonication device (Covaris, Woburn, MA, USA), using ultrasound cycles of 1 min for herbarium samples and 2 min for silica-dried samples, which produced an average insert size of 350 bp. Highly fragmented DNA extracts were not sonicated. After drying samples in a speed

vacuum concentrator (Eppendorf) and eluting in ddH_20, dual-indexed libraries were constructed with primers in the NEBNext Multiplex Oligos for Illumina and reagents in the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, MA, USA) using the manufacturer's protocol (Instruction Manual v.6.1) at half the recommended volumes according to the cost-reducing strategies in Hale *et al.* (2020). Size selection was only carried out for one high-quality silica-dried sample and then applying eight PCR cycles; for the rest, to preserve library complexity, fragments were bead-cleaned with SeraMag magnetic carboxylate-modified microparticles diluted to 5%, then amplified with 10 PCR cycles following adapter ligation. Libraries and fragment size were assessed with TapeStation using High Sensitivity D1000 ScreenTapes (Agilent Technologies, CA, USA) and quantification checked with a Qubit fluorometer (ThermoFisher Scientific). The final library size was on average 350 bp.

Target capture and amplification

To prepare the dual-indexed libraries for target capture, samples were pooled equimolarly into nine tubes according to type (herbarium or silica-dried) and library concentration (ranging from 1.58 to 84 ng/μL), using 100 ng of library per specimen where possible, minimally 5–10 ng for low concentration samples. Custom baits designed for the family Sapotaceae (Christe *et al.* 2021) were used. This probe design targeted 227 microsatellite and 792 protein-coding loci, with 2× tiling density, spanning around 20 000 probes of 90 bp each, of which 261 exons corresponded to monocopy genes from the Angiosperms353 pool described by Johnson *et al.* (2019), consisting of a total of 1 034 731 bp (STR and flanking regions: 152 118 bp; nuclear coding: 882 613 bp).

Hybridization of pools with specific biotinylated loci-complementary probes (myBaits Custom Target Capture Kit produced by Arbor Biosciences, Ann Arbor, MI, USA) was undertaken with the manufacturer's protocol (myBaits user manual v.5.01), except that the incubation timing was adapted to 40 hours at 65°C (Paijmans *et al.* 2016) to allow for maximum capture, with each tube including an equal volume of liquid wax (Bio-Rad Chill-Out, Hercules, CA, USA) to prevent evaporation. Streptavidin-encapsulated magnetic beads captured the hybridized fragments, and 11–17 PCR cycles using KAPA High-Fidelity DNA Polymerase (Roche, Indianopolis, IN, USA) were performed for each pool, with the lowest number of cycles for the best samples and highest for the most fragmented. PCR products were run on TapeStation to measure size distribution and concentration, and the pooled libraries were sequenced by NovogeneAIT Genomics (Singapore) on an Illumina NovaSeq instrument (San Diego, CA, USA) producing 2 × 150 bp paired-end reads.

Sequenced data processing

Raw reads were checked for initial quality with FastQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) then trimmed with Trimmomatic v.0.39 (Bolger *et al.* 2014), which removed adaptor sequences and low-quality sections (including bases from the beginning and end of reads if they fell below the Phred quality threshold, as well as reads shorter than 36 bp) with the settings LEADING: 3; TRAILING: 3; SLIDING WINDOW: 4:15; MINLEN: 36. Trimmed files were quality-checked again with FastQC.

Further trimming and tidying of the reads to remove primers, poly A-tails, and other unwanted sequences was carried out with CutAdapt v.2.3 (Martin 2011), following the Hyb Baits pipeline (Nicholls *et al.* 2015; scripts available at https://github.com/ckidner/Targeted_enrichment). The trimmed reads were then mapped to the bait sequence (Christe *et al.* 2021; target file available at https://zenodo.org/record/4436715/#.YjIt949By3B) using Bowtie2 v.2-2.3.5 with the default value for the threshold for the alignment score (G,20,8) (Langmead and Salzberg 2012), Samtools v.1.9 (Li *et al.* 2009) to sort alignments, remove poor maps and obtain summary mapping statistics, and bcftools to obtain a consensus sequence for each locus. Sequences were aligned with the progressive method FFT-NS-2 in MAFFT v.7.407 (Katoh and Standley 2013), then AMAS v.1.0 (Borowiec 2016) was applied to obtain summary statistics for each alignment (including alignment length, missing data, and numbers of parsimony-informative sites). The output from the Nicholls *et al.* (2015) pipeline was then used to infer four of the eight phylogenies presented in this study (see the Phylogenomic Analyses section).

To assess paralogy, two methods were employed, respectively using the containerized HybPiper-RBGV and Yang-and-Smith-paralogy-resolution pipelines utilised in Schmidt-Lebuhn (2022), with scripts on https://github.com/chrisjackson-pellicle/hybpiper-nf and https://github.com/chrisjackson-pellicle/Yang-and-Smith-RBGV-scripts. First, a separate run of HybPiper (Johnson *et al.* 2016) using the HybPiper-RBGV container was undertaken using the trimmed reads as input and the baits as reference, and genes with paralog warnings for more than two individuals (Zhou *et al.* 2022) were then removed for phylogenomic inference. Second, using the paralog-flagged output files from the containerized HybPiper-RBGV run, the 'monophyletic outgroup' (MO) approach in Yang and Smith (2014) was used to infer orthologues; this approach detects gene duplication at a particular node and eliminates the side of the rooted gene tree with fewer taxa, therefore filtering for low- and single-copy genes.

Phylogenomic analyses

Both coalescent and concatenation approaches were used to infer phylogenetic relationships; the former produces a species tree from gene trees estimated at each locus, while the latter concatenates all genes before analysis. To model the multispecies coalescent (MSC), ParGenes v.1.1.2 (Morel *et al.* 2019) was used to obtain a maximum-likelihood (ML) tree for each locus (with RAxML-NG v.1.1.0: Kozlov *et al.* 2019), employing a general time reversible gamma-distribution model with automatic MRE bootstrap convergence for a maximum of 100 bootstrap replicates. Nodes with bootstrap support values of <0.1 were collapsed (this has been shown to improve species tree accuracy: Zhang *et al.* 2017) using the pxcolt function in phyx v.1.3 (Brown *et al.* 2017) before using ASTRAL-III v.5.7.8 (Zhang *et al.* 2018) to generate a species tree with posterior probability and quartet support. For concatenated supermatrices, AMAS was used to concatenate all loci alignments, and RAxML-NG used

to generate an ML species tree under the optimum general time reversible gamma-distribution model (GTR + G) for invariant site proportion and rate heterogeneity. Then 400 starting trees (200 parsimony and 200 random) were generated and bootstrap support values to estimate branch support for the best ML tree were estimated with 1000 standard replicates; these were the number of replicates required to achieve convergence with the '--bs converge' test set to '--bs-cutoff 0.02' in RAxML-NG. Trees were visualized with FigTree v.1.4 (http://tree.bio.ed.ac.uk/software/figtree) and the ape v.5.6-2 (Paradis and Schliep 2019), phytools v.1.0-3 (Revell 2012), and dendextend v.1.16.0 (Galili 2015) packages in Rstudio v.4.1.3 (R Core Team 2022) were used for tanglegram plotting. Pie charts to represent and visualize gene tree conflict were generated with Phyparts v.0.0.1 (Smith *et al.* 2015) and the PhyPartsPieCharts script (available from https://github.com/mossmatters/MJPythonNotebooks/blob/master/phypartspiecharts.py), using the species and gene trees re-rooted at the outgroup taxon *Sideroxylon gerrardianum* (Hook.f.) Aubrév.

STR loci were assessed for parsimony-informativeness, but to prevent confounding effects they were removed from the final alignments prior to the phylogenomic analyses. Eight main phylogenomic analyses were conducted (the first four with outputs from the HybPiper and Yang and Smith containerized pipelines used in Schmidt-Lebuhn 2022; the last four using the outputs from the pipeline by Nicholls *et al.*, 2015): (i) a concatenated supermatrix with paralogues (flagged by HybPiper) removed; (ii) an MSC-based species tree with paralogues (flagged by HybPiper) removed; (iii) a concatenated supermatrix with orthology inferred using the MO approach in Yang and Smith (2014); (iv) an MSC-based species tree with orthology inferred using the MO approach in Yang and Smith (2014); (v) a concatenated supermatrix assembled following Nicholls *et al.* (2015); (vi) an MSC-based species tree assembled following Nicholls *et al.* (2015); (vii) a species tree from 261 coding loci extracted from the Angiosperms353 universal pool and assembled following Nicholls *et al.* (2015); and (viii) a species tree from 531 coding loci, Sapotaceae/family-specific, identified from Christe *et al.* (2021) and assembled following Nicholls *et al.* (2015). All topologies can be found in the Supporting Information, Figures S1–S8.

Population genetic analyses

Using the aligned reads to the references (first without STR loci, then also separating the 261 loci from the Angiosperms353 universal pool) from the Nicholls *et al.* (2015) pipeline, variant detection was undertaken with GATK v.4.1.3, applying Picard v.2.19 and Samtools v.1.9 to sort, mark duplicates, and index, then HaplotypeCaller to call SNPs on single samples to generate GVCF files (default ploidy: 2), which were then consolidated with CombineGVCF and jointly genotyped in GenotypeGVCF. To minimize the influence of singletons (Linck and Battey 2019), filtering to remove multiallelic SNPs and minor alleles below a set frequency (maf 0.005) was undertaken with bcftools v.1.15 and vcftools v.0.1.16, respectively. Linkage disequilibrium (LD) was checked by applying the vcf.parser.py script from https://github.com/CoBiG2/RAD_Tools to the filtered VCF file to obtain one SNP per locus. Inference of population structure from the remaining biallelic SNPs, was done using STRUCTURE (Pritchard *et al.* 2000), processed and plotted via Structure_threader v.1.3.10 (Pina-Martins *et al.* 2017), with *K* values (i.e. number of genetic clusters) set from one to 10, replicates at 10 for each *K*, with a burn-in of 20 000 and Markov chain Monte Carlo repetitions at 100 000. The optimal value of *K* was assessed with the software Kfinder2 v.1.0.0.0 (Wang 2019), which generates three estimators of best *K* using: the $\Pr[X|K]$ method (Pritchard *et al.* 2000) and the ΔK method (Evanno *et al.* 2005) both of which consider the rate of change in log probability between successive values of *K* (ΔK), and then a parsimony index that pinpoints the ideal *K* that generates the minimum admixture.

To inspect genetic variation for each species, plink v.1.90 (Chang *et al.* 2015) was used to estimate heterozygosity and inbreeding coefficients (averaged across individuals within the species), as well as nucleotide diversity (π) and Tajima's *D* on the filtered SNP dataset (averaged across loci), with reference to Tajima (1989) for critical values to indicate deviation from mutation-shift balance. Plink was also used to convert the consolidated VCF file to eigenvector/eigenvalue ones for Principal Coordinate Analyses (PCA), and the first two dimensions were visualized with the tidyverse v.1.3.2 package (Wickham *et al.* 2019) in Rstudio v.4.1.3. PCA analysis was additionally conducted on the SNP dataset for *P. obovatum*, since the sample size (14) for that species was the largest, and a Mantel test to evaluate the association with geographical distance performed with the ade4 v.1.7-20 package (Dray and Dufour, 2007) in Rstudio v.4.1.3.

RESULTS

Target sequence capture

The final dataset consisted of 77 *Palaquium* accessions and nine outgroup taxa from related genera. Bait specificity was expectedly higher for the *Palaquium* samples that had undergone the hybridization process conducted in this study, with an average of 74.82% on-target reads (54.65%–84.23%), compared to 50.34% for outgroups. Loci were successfully retrieved at high rates regardless of the type of sampling storage: silica-dried samples yielded only a marginally higher yield of successfully mapped reads compared to herbarium samples (mean: 76.0% vs. 74.1%; median: 77.3% vs. 75.9%), although there was greater variability for the latter.

After quality checks, filtering, and trimming, the concatenated supermatrix for all loci using the Nicholls *et al.* (2015) pipeline contained 86 samples and 1019 loci, with total alignment length of 922 082 bases and 60 348 parsimony-informative sites, and a missing data average of 7.7%. Although parsimony-informative sites for microsatellite loci (including STR flanking regions) were proportionally comparable to coding loci (0.08 vs 0.06), the number of these sites was lower than for coding loci (53.73 vs 60.80 on average), and also saw more missing data (22.24% vs 3.5%). This was expected as Christe *et al.* (2021) had noted in their loci testing that STR capture was strongly dependent on phylogenetic proximity to the bait reference species (in this case *Capurodendron* Aubrév., which is in Tseboneae, a different tribe from *Palaquium*). The STR regions that were removed before the

phylogenetic inference contained a total of 144 951 bases and 12 198 parsimony-informative sites (Supporting Information, Table S2).

Putative paralogues for at least two individuals were flagged by HybPiper for 259 loci (114 from STR regions, eight from the Angiosperms353 pool, and 127 from the taxon-specific set): the supermatrix without the paralogues and prior to STR removal contained 760 loci (644 nuclear coding and 116 STR regions), with a total alignment length of 744 219 bases, 59 711 parsimony-informative sites, and 2% missing data. For the MO orthology inference, the supermatrix contained 799 loci (663 nuclear coding and 136 STR regions), with a total alignment length of 883 630 bases, 91 805 parsimony-informative sites, and 3.9% missing data. The STR regions that were removed from the HybPiper and MO orthology inferences respectively contained a total of 75 482 bases and 8599 parsimony-informative sites, and 107 248 bases and 17 256 parsimony-informative sites (Supporting Information, Tables S3, S4).

Phylogenomic analysis

The loci recovered in the target capture process produced well-resolved phylogenetic trees, and all analyses were nearly topologically congruent for species delimitation, except for the placement of the outgroup taxon *Diploknema butyraceae* (Roxb.) H.J.Lam, which is usually in clades containing other taxa from Pennington's (1991) tribe Isonandreae, either emerging near *Madhuca insignis* H.J.Lam in most results, or closer to *Isonandra compta* Dubard in the phylogenomic analyses (ii) and (vi), though not always with strong support. For all the analyses save (vii), which was generated only from 261 of the Angiosperms353 loci, internal nodes are fully resolved in the ML trees and mostly so in the species trees (BS = 100; local PP > 0.98) except for the accession of *Palaquium hexandrum* and *P. sp.* from Borneo in the coalescent-based inferences and outgroup clades containing the other Isonandreae taxa including *Isonandra compta*, *Madhuca insignis*, and *Diploknema butyracea*. Maximum support (BS = 100; local PP = 1) was obtained in almost all trees [except analysis (vii) from the Angiosperms353 loci] for every major clade of known *Palaquium* species in Singapore. The datasets addressing paralogues recovered near-identical topologies to the analyses using the outputs from the pipeline by Nicholls *et al.* (2015), and contained the same well-supported nodes delimiting all main *Palaquium* species (Fig. 2; Supporting Information, Figs S1–S6).

The genus *Aulandra* is clearly nested within *Palaquium*. Several *Palaquium* species are not monophyletic as currently circumscribed: although four taxa (*P. impressinervium*, *P. obovatum*, *P. oxleyanum*, and *P. xanthochymum*) emerge in strongly supported distinct clades, *P. rostratum* is nested in the *P. microphyllum* clade. The high density of within-species sampling results in a short-branched tree near the tips with poorer support at the outermost nodes, which is expected given fewer phylogenetically sorted mutations that allow probabilistic modelling. Still, a few within-clade nodes presented reasonably high ML bootstrap or local posterior probability support, and geographical structure of collections (i.e. separate clades for Borneo, Peninsular Malaysia, and Singapore collections) can normally be distinguished. Three sterile specimens were re-assessed and re-determined due to phylogenetic placement: two Peninsular Malaysian specimens from *P. hexandrum* to *P. obovatum*, and one Bornean specimen from *P. gutta* to *P.* sp.

The ASTRAL tree generated from 261 loci out of the Angiosperms353 set is strongly supported for the deepest nodes of the *Palaquium* clades, including the nesting of *Aulandra*. Node support however progressively weakens towards the tips, with *P. hexandrum* from Borneo emerging outside the clade with the rest of the *P. hexandrum* accessions, and geographical sorting less defined between Peninsular Malaysia and Singapore accessions for *P. xanthochymum*, *P. rostratum*, and *P. microphyllum*. There is also greater uncertainty for the placement of sister taxa to *Isonandra compta* in the branch containing the other outgroups (local PP = 0.86). The species tree based on the 531 taxon-specific coding loci in the bait set has greater support for relationships among outgroup taxa (i.e. higher local PP = 0.92 regarding taxa sister to *Isonandra compta*), as well as clearer geographical structure equivalent to the coalescent-based species tree generated with all loci. Although the analysis of gene tree discordance showed widespread minor conflicts, there was less discordance at the deeper nodes for the species tree from the Sapotaceae-specific loci, with the node containing all *Palaquium* species having more concordant and informative gene trees (185 out of 346) than the ones from the universal probe set (51 out of 210) (Fig. 3).

Population genetic results

The LD-adjusted dataset from all loci (without STRs) with 790 SNPs produced the same STRUCTURE results and genetic clusters as the non-LD-adjusted SNPs (Supporting Information, Fig. S10), hence the latter dataset was maintained for downstream analyses. A total of 118 519 biallelic SNPs were identified, with *P. obovatum* containing the highest and *P. gutta* the lowest number of SNPs (Table 1). The proportion of heterozygosity ranged from 19% in *P. microphyllum* to 33% in *P. gutta*. Pairwise nucleotide diversity (π) ranged from 0.00038 (*P. xanthocymum*) to 0.0013 (*P. obovatum*). Tajima's *D* was negative for all species, which might indicate recent directional selection (Slimp *et al.* 2021) although there was no significant deviation from mutation-drift balance (assessed against critical values at the 95% level in Tajima, 1989). Inbreeding coefficients (*F*) averaged across each species were negative for all species except for *P. impressinervium*, *P. oxleyanum*, and *P. microphyllum*; although the *F* values for *P. hexandrum* and *P. obovatum* were not significantly different from zero ($P > 0.05$).

An analysis of population structure using the total SNP dataset did not show geographical structure within species, and in some taxa did not resolve species boundaries. The expectation of eight genetic clusters according to the number of included species was not supported by the estimators for best *K*. The most likely number of distinct genetic groups was $K = 4$ as estimated by Pritchard and the Parsimony index. Although the Evanno method identified the best *K* as 8, this seemed unlikely because of random noise introduced by the additional genetic groups in the plot (Fig. 4). The four clusters corresponded to major nodes in the earlier phylogenetic results prior to lineage separation, with only *P. impressinervium* and *P. obovatum*, respectively, emerging as distinctive gene pools. For the other two clusters,

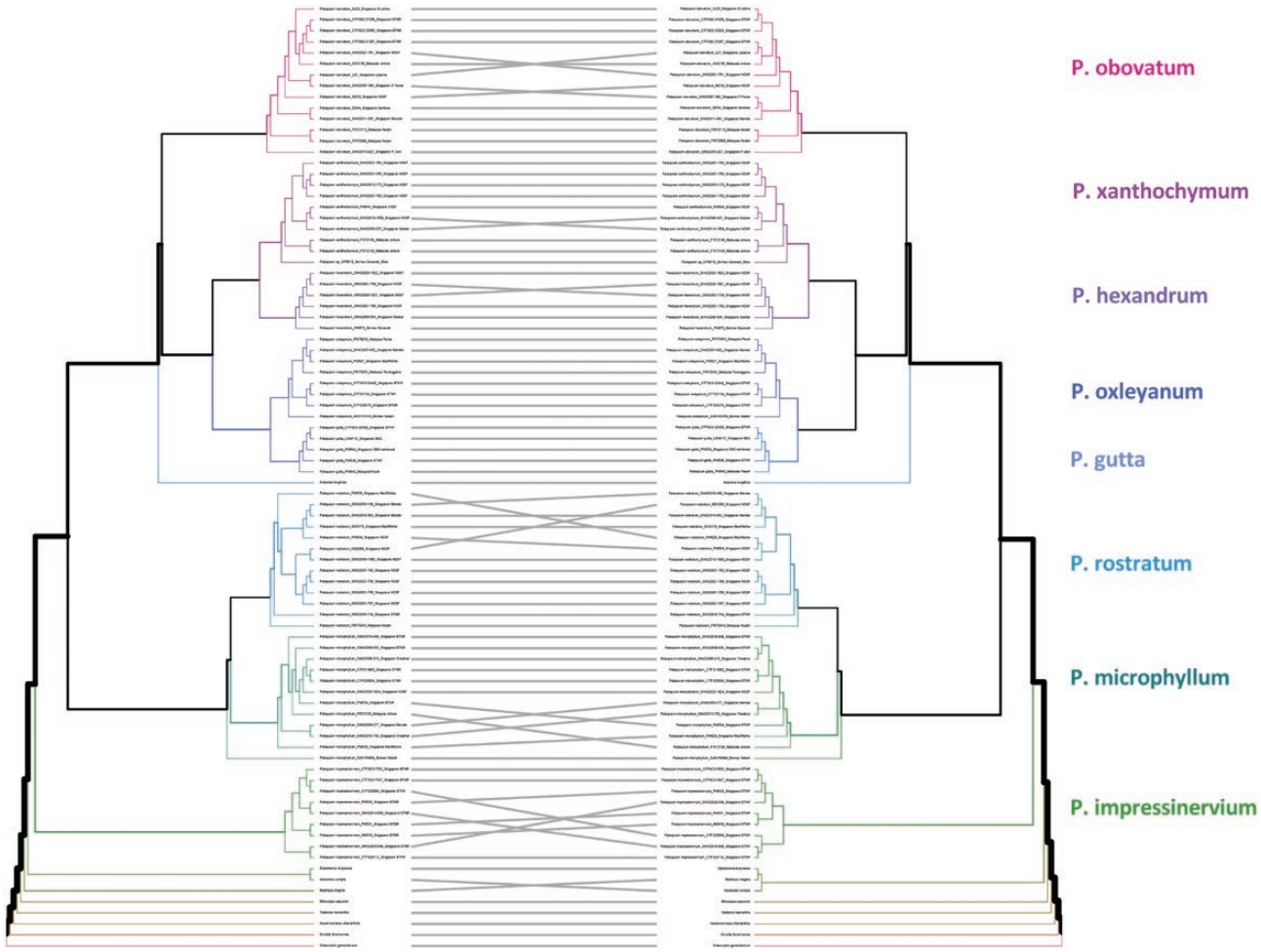

Figure 2. Tanglegram comparing coalescent-based species trees generated from different pipelines: the one using all loci from Nicholls *et al.* (2015) (left) and the paralogue-aware MO inference from Yang and Smith (2014) (right).

the first consisted of both *P. microphyllum* and *P. rostratum*, and the second comprising four species as currently defined: *P. gutta*, *P. oxleyanum*, *P. hexandrum*, and *P. xanthochymum*. *P. hexandrum* showed the most significant account of admixture at different levels of *K* up to 10. (Fig. 4). The 261 loci from the Angiosperms353 universal pool produced 15 926 biallelic SNPs and showed even less population structure than the full dataset, hence further comparative species-level analyses was not undertaken.

The clusters seen in the PCA scatter plot for the consolidated SNP dataset (Fig. 5) appeared closer to $K = 3$ than the inferred population structure ($K = 4$): *P. impressinervium* being the most clearly separated, and overlaps seen for *P. microphyllum* and *P. rostratum*, with *P. obovatum* emerging close to the grouping of *P. gutta*, *P. oxleyanum*, *P. hexandrum*, and *P. xanthochymum* (Fig. 5). A PCA on the *P. obovatum* dataset showed some degree of spatial distribution, though not necessarily varying according to proximity, as the accession from Pulau Ubin, an island ~2 km off the coast of mainland Singapore, showed greater similarity in the first two principal component axes to accessions from Kedah, a state of Malaysia >600 km away (Supporting Information, Fig. S9). The Mantel test found a negative but insignificant ($P = 0.81$) correlation between genetic variability and geographical distance.

DISCUSSION

Phylogenetic insights into species

This is the first phylogenomic study within the Isonandreae tribe of Sapotaceae that successfully applies a custom target capture bait panel to a densely sampled cross-section of *Palaquium* representatives. The combination of taxon-specific and universal loci within a bait set was able to achieve high resolution across a relatively wide phylogenetic distance. The loci from the universal bait set provided less population-level resolution, which was expected given the kit's general applicability to all angiosperms, but it was still able to achieve a similar level of interspecific resolution as the taxon-specific loci, showing how useful the Angiosperms353 panel can be for resolving uncertainties in species delimitation within families considered taxonomically difficult. Nonetheless, resources permitting, a taxon-specific bait set would afford much greater clarity for species across its distribution, particularly in terms of geographical structure that may guide consideration for subspecies or varietal circumscription.

Paralogue filtering did not impact the delineation of species relationships in this study, even though paralogue inclusion has been shown in other analyses to compromise species tree inferences (Altenhoff *et al.* 2019; Fernández *et al.* 2020). In this study, it is possible that regardless of the removal or

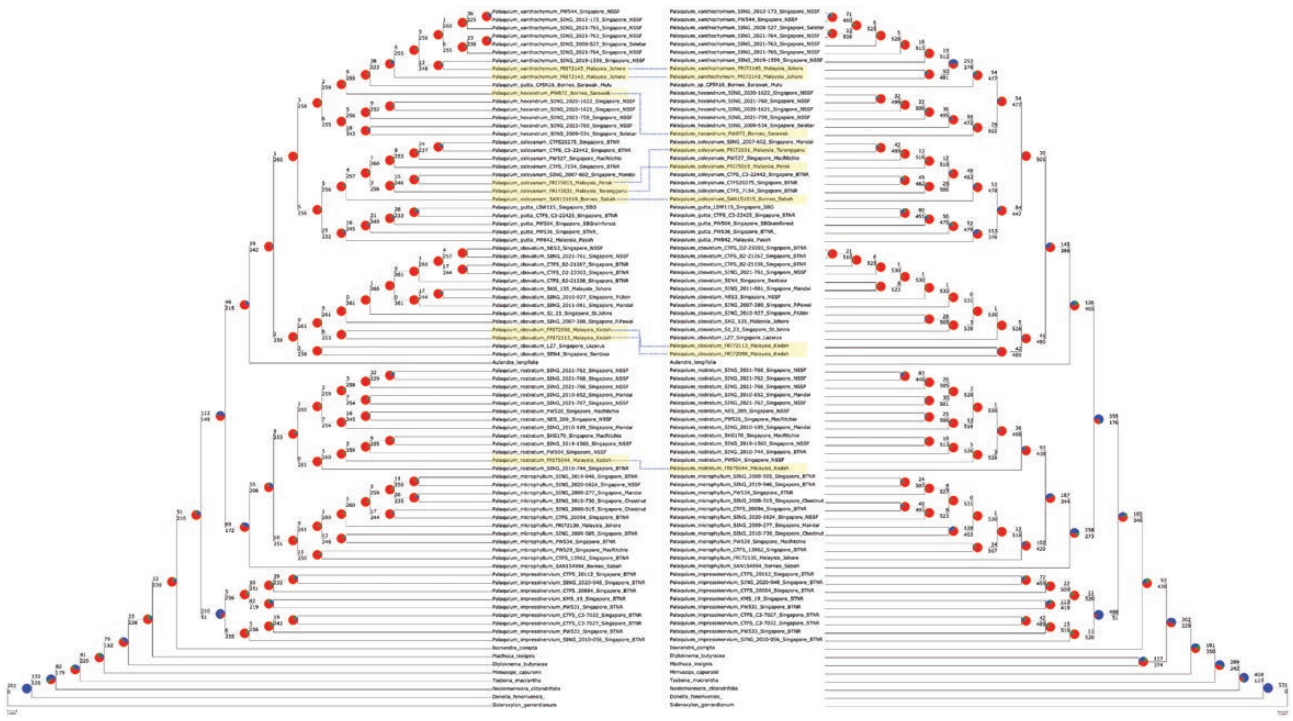

Figure 3. Coalescent-based species trees from 261 from the Angiosperm 353 set (left) and 531 genes in the family/taxon-specific probes (right); gene tree discordance is represented by pie charts at each node: proportion concordant with the shown topology (blue), the most frequent conflicting bipartition (green), other conflicting bipartitions (red), and gene trees that are uninformative (grey). The top number refers to the number of homologues concordant with the species tree at that node, and the bottom the number of homologues in conflict. Selected taxa outside Singapore are highlighted (yellow) to indicate differences in phylogenetic placement.

Table 1. Numbers of recovered SNPs and population statistics for species in this study

	Number of individuals	SNPs (unfiltered)	SNPs (filtered, no missing data)	SNPs (filtered, only biallelic markers)	Heterozygosity (average individual)	π (average across loci)	F (average individual)	Tajima's D (average across loci)
P. gutta	5	41 576	11 548	9745	0.3284	0.000404	−0.01311	−0.40621
P. hexandrum	6	52 742	23 989	21 535	0.3247	0.000881	−0.15972	−0.15972
P. impressinervium	10	54 098	31 136	27 995	0.2084	0.000721	0.553409	−0.98256
P. microphyllum	12	55 025	25 489	22 405	0.1896	0.000532	0.198521	−1.08973
P. obovatum	14	86 421	59 681	53 033	0.1986	0.001305	−0.26109	−0.86731
P. oxleyanum	8	48 328	18 818	16 549	0.2529	0.000525	0.056708	−0.70999
P. rostratum	10	53 264	22 818	20 065	0.1954	0.000483	−0.01928	−1.18776
P. xanthochymum	9	43 412	12 051	10 358	0.2915	0.000384	−0.07507	−0.18334

incorporation of the 135 loci flagged by HybPiper as paralogous, there remained sufficient phylogenetic information in the rest of the strongly recovered 644 loci. Yan *et al.* (2022) showed that gene tree-based inference in the presence of paralogues can still produce robust and accurate species tree inferences, even when addressing the presence of incomplete lineage sorting. It may also be the case that a whole-genome duplication (WGD) event, as a likely source of paralogues, has not taken place recently for the lineages under study. Sapotaceae, with numerous other families in the Ericales order, had shared an ancient WGD around the Cretaceous–Paleogene boundary, but probably has not experienced another since then (Wang *et al.* 2021; Zhang *et al.* 2022). The much higher number of parsimony-informative characters in the MO orthology inference compared to the removal of HybPiper-flagged paralogues (74 549 vs. 51 112), although having no substantial effect on phylogenetic outcomes here, also provide indication that methods such as the former (Yang and Smith 2014), which explicitly allow for gene duplication, can preserve more loci and phylogenetic signal. Other than the presence of paralogues, gene tree discordance may also be due to horizontal gene transfer or noise arising from sequence assembly (Michel *et al.* 2022), however, the similar results across pipelines provide support that node resolution appears driven by a select few genes, similar to findings in other studies (Shen *et al.* 2017; Walker *et al.* 2018).

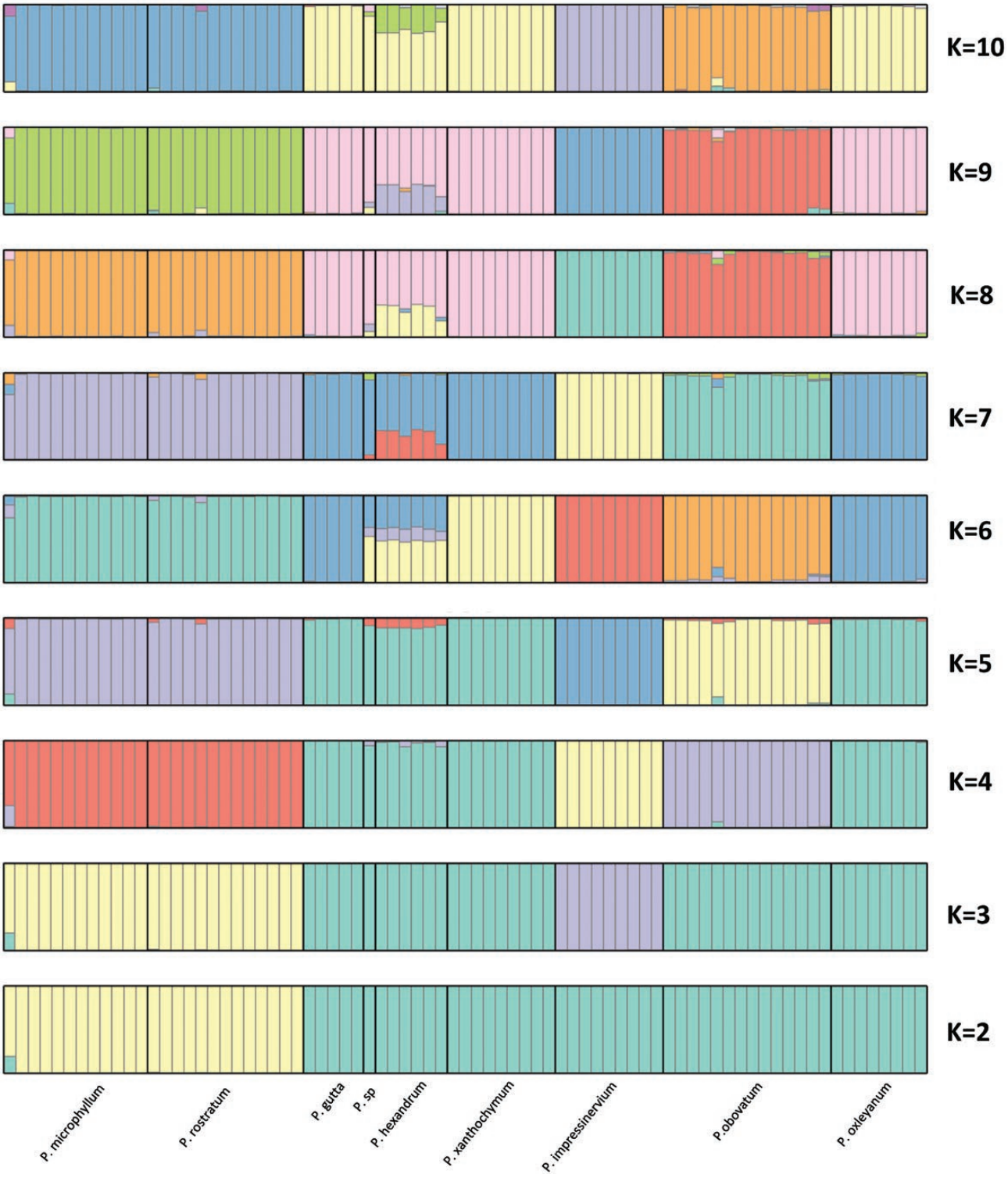

Figure 4. Population structure of selected *Palaquium* species inferred from 118 519 SNPs with the STRUCTURE software; colours represent genetically distinct clusters, and each sample is separated by vertical bars.

The nesting of *Aulandra longifolia* in *Palaquium* was also an outcome in a wider ITS-based phylogenetic analysis (Richardson *et al.* 2014) using the same voucher, although that result was not strongly supported. Richardson *et al.* (2014) found *Diploknema* to be polyphetic and *D. butyracea* to be nested within *Palaquium*, which is not the case here, where the taxon is clearly placed in the outgroup clades. This requires further investigation, including into the different vouchers of *D. butyracea* used in the various studies, and wider sampling of *Aulandra* and *Diploknema* species.

All phylogenetic trees raised uncertainties regarding the monophyly of *Palaquium microphyllum* and *P. rostratum*, with the latter occurring in a well-supported clade nested within the former (Fig. 2), although this might be resolved by an expanded sampling of other species within the clade. A difficulty that cuts

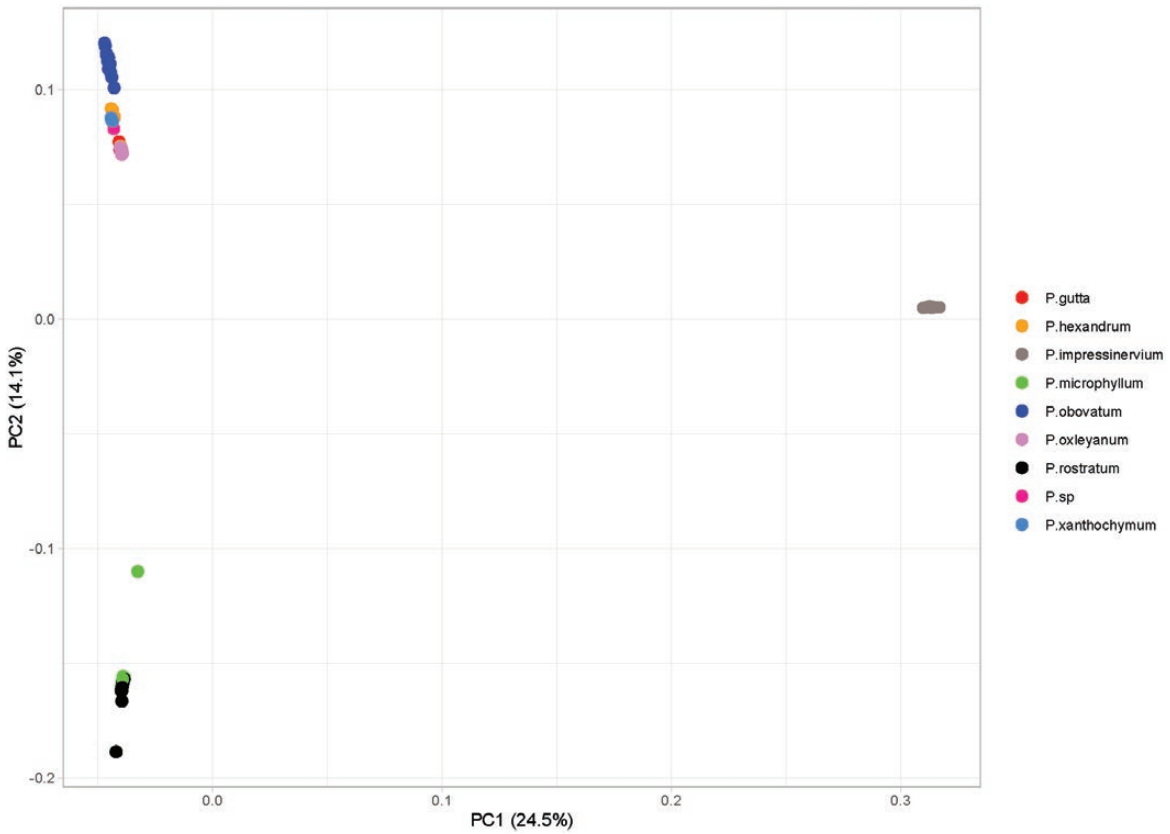

Figure 5. PCA scatter plot of SNP variability (out of 118 519 SNPs) for the 77 ingroup samples.

across the genus and family, is that while there are clear morphological distinctions (in this case *P. microphyllum* has mature leaves that are half the blade size of *P. rostratum* and floral parts substantially smaller than the latter), there are almost always overlapping characters between two species (such as inconspicuous leaf venation and glabrous surfaces). It is thus possible that those features that merit placing one species into synonymy with another, may be distinguishing factors between other species. Population genetic analyses on independent SNPs further indicate that the delineation between *P. microphyllum* and *P. rostratum* is tenuous, with both species forming a single genetic cluster. Monophyletic species concepts cannot recognize a species nested within parental taxa even if it appears to be well on its way to diverging completely, since it results in taxonomic paraphyly (Baum and Shaw 1995; De Queiroz 2007). However, a conflict between phenotypes and genotypes may lead to consideration of alternative species concepts that reconcile temporal dimensions and paraphyletic lineages (Freudenstein *et al.* 2017; Boluda *et al.* 2021).

Amid the species concept debate that continues unabated (Wells *et al.* 2022), the fundamental question is whether *P. rostratum* is sufficiently morphologically distinct, with well-characterized features separating it from *P. microphyllum*. Two possible solutions present themselves: one is the unified species concept where lineages only have to be evolving separately to be considered different species (De Queiroz 2007), or that the rank of subspecies is appropriate, conceptualized by incompletely separated lineages within a more inclusive species (De Queiroz 2020). There appear to be some genomic grounds to re-circumscribe *P. rostratum* as a subspecies of *P. microphyllum* in a historically connected and ecologically similar distribution across Malesia, but this is a decision best achieved after greater sampling outside Singapore to confirm the relationship and assess the extent of morphological and genetic dissimilarities.

It is equally possible that rainforest tree species are less likely to be resolved as monophyletic as widespread populations of significant longevity take an extremely long time to coalesce. This is due to their being predisposed to larger effective population sizes, and conceivably manifest a pattern whereby recently evolved monophyletic taxa are nested within progenitor species (Pennington and Lavin 2016). Extensive radiation in *Palaquium* from west to east of Wallace's line is well-documented (Richardson *et al.* 2014), and species including the ones in this study have large effective population sizes, which coupled with long generation times (e.g. Ridley, 1902, noted that some *P. gutta* trees have been recorded to take 21 years to produce fruit), suggest that it may take even longer than 50 million years to achieve monophyly at all loci (Baker *et al.* 2014; Naciri and Linder 2015). This may well explain the nesting pattern seen for *P. microphyllum* and *P. rostratum*, and the less-than-expected number of genetic clusters attained in the population structure analysis.

Population perspectives: potential and limitations

It may be the case that the recovered targets did not contain a sufficient diversity of SNPs to detect population structure, as it has been previously noted that slower-evolving and more conserved genes hybridize more successfully to targets (Paijmans *et al.* 2016). The Sapotaceae-specific probes, containing mostly

conserved orthologues, were designed for broad relevance across the family, and although some intron-exon boundaries were included (Christe *et al.* 2021), the intronic regions might not have contained sufficiently variable sequences for detailed assignment of populations. Using the same set of baits on taxa in the genus *Capurodendron*, STRUCTURE results in (Boluda *et al.* 2021) also did not recover gene clusters that matched species definitions; the authors had noted the possibility of insufficient signal in loci selection, due to phylogenetically distant species used in the bait design.

Nonetheless the Sapotaceae-specific probes, other than being able to resolve broad and deep phylogenetic relationships, can still be helpful to population genetic investigations. More than 100 000 SNPs were recovered in the species under study, with statistical power in the numerous loci to infer heterozygous signatures and nucleotide diversity. These parameters can be important to guide conservation assessments, for instance to pinpoint species at risk of inbreeding depression or low levels of genetic diversity that may benefit from external introductions, potentially in the form of targeted street planting of trees. Such population genomic findings in Singapore, where habitat fragmentation is significant, may help to guide conservation policy in many areas grappling with deforestation pressures amid urban encroachment (Lim *et al.* 2019) and inform perspectives in environments at risk of dynamic change and degradation.

Future work

More work is needed to define morphological structures characterizing the genus, including fertile ones. Although *Palaquium impressinervium* emerges clearly as the earliest diverging species of the genus, supporting the findings of the earlier ITS phylogenetic study (Richardson *et al.* 2014), it falls out of the clade inclusive of all other *Palaquium* species. However, the morphological traits that support such stark separation from the rest of the genus are not straightforward to assess. Only immature buds and incomplete fruit have been collected in Malaysia (Ng 1969); in Singapore, no fertile specimen has been collected or observed in the last 50 years and sterile characteristics are not sufficiently distinctive, hence a detailed study of actual fertile material would be needed for more complete modelling of ancestral trait evolution and clearer identification of morphological features separating *Palaquium* species.

Although outgroup sampling is limited, tribal limits in the coalescent-based results are congruent with the results of studies employing the same set of taxon-specific target probes to other genera (Christe *et al.* 2021; Boluda *et al.* 2022), with firm backbone support for the subfamily Sapotoideae and monophyly confirmed for Pennington's tribe Isonandreae. There still however remain multiple relationships to be clarified between Sapoteae (Christe *et al.* 2021), Isonandreae, and closely related groups such as the subtribe Gluemineae. The phylogenomic findings reported here support a similar approach for further studies containing more comprehensive sampling.

CONCLUSION

This study highlights the effectiveness of a target capture approach in phylogenetic reconstruction and indicates that including universal probes into a taxon-specific kit (Hendriks *et al.* 2021; Ufimov *et al.* 2021), can be useful in delimiting *Palaquium* species, and probably other closely related taxa in the Isonandreae. Greater taxonomic clarity is thus within reach for Southeast Asian representatives of Sapotaceae, where many gaps remain in poorly understood genera such as *Palaquium*. Our results, even on a subset of species, throw up clear issues needing further study: while morphological species limits seem clear, phylogenetic ones are often not, and greater sampling to cover all 120 or so species of *Palaquium* is likely to uncover interesting patterns and the need to revisit current species circumscriptions. Generic limits, including *Aulandra*, are also clearly needing to be investigated further and probably revised. The selected loci in the Sapotaceae bait set helped to resolve relationships in geographically widespread taxa and recover valuable parameters to study genetic diversity. However, a study of fine scale population genetic structure would require taxon-specific bait sets closer to the species under study, with a greater proportion of intronic regions in the probes to capture sufficient variability. Conservation studies require the detection of changes in effective population size to a limited number of generations (tens, at most hundreds; Maisano Delser *et al.* 2016). Therefore, methods such as RADseq may be required to supplement where a taxon-specific bait set is unable to detect sufficient divergence.

SUPPLEMENTARY DATA

Supplementary data is available at the *Botanical Journal of the Linnean Society* online.

Figure S1. ML tree from concatenated supermatrix obtained from HybPiper pipeline, paralogues, and STRs removed. Node values indicate bootstrap support percentages, and the scale bar shows per site units of substitutions.

Figure S2. Coalescent-based species tree from HybPiper pipeline with paralogues and STRs removed. Node values indicate posterior probability, and the scale bar shows coalescent units.

Figure S3. ML tree from concatenated supermatrix with orthology inferred using the MO approach in Yang and Smith (2014). Node values indicate bootstrap support percentages, and the scale bar shows per site units of substitutions.

Figure S4. Coalescent-based species tree with orthology inferred using the MO approach in Yang amd Smith (2014). Node values indicate posterior probability, and the scale bar shows coalescent units.

Figure S5. ML tree from concatenated supermatrix assembled following Nicholls *et al.* (2015). Node values indicate bootstrap support percentages, and the scale bar shows per site units of substitutions.

Figure S6. Coalescent-based species tree assembled following Nicholls *et al.* (2015). Node values indicate posterior probability, and the scale bar shows coalescent units.

Figure S7. Coalescent-based species tree from 261 coding loci extracted from the Angiosperms353 universal pool and assembled following Nicholls *et al.* (2015). Node values indicate posterior probability, and the scale bar shows coalescent units.

Figure S8. Coalescent-based species tree from 531 coding loci, Sapotaceae/family-specific, identified from Christe *et al.* (2021) and assembled following Nicholls *et al.* (2015). Node

values indicate posterior probability, and the scale bar shows coalescent units.

Figure S9. PCA scatter plot of SNP variability (out of 53 033 SNPs) from 14 samples of *P. obovatum* collected in various localities in Singapore and Peninsular Malaysia.

Figure S10. Population structure of selected *Palaquium* species inferred from 790 unlinked SNPs with the STRUCTURE software; colours represent genetically distinct clusters, and each sample is separated by vertical bars.

Table S1. Sample information.

Table S2. Summary statistics for each alignment generated from the pipeline following Nicholls *et al.* (2015) (alignment names with 'Block' indicate STR loci; Genes 1–531 are family-specific; Genes 532–792 are extracted from the Angiosperms353 universal pool).

Table S3. Summary statistics for each alignment generated using the MO approach in Yang and Smith (2014) (alignment names with 'Block' indicate STR loci; Genes 1–531 are family-specific; Genes 532–792 are extracted from the Angiosperms353 universal pool).

Table S4. Summary statistics for each alignment generated from the HybPiper pipeline with paralogues removed (alignment names with 'Block' indicate STR loci; Genes 1–531 are family-specific; Genes 532–792 are extracted from the Angiosperms353 universal pool).

ACKNOWLEDGEMENTS

We thank the management of the National Parks Board and Singapore Botanic Gardens for providing continuous support, facilities, and fieldwork approval (Permit NP/RP17-098-3) in Singapore's nature reserves. We appreciate the digital research team at the University of Aberdeen and the research/scientific computing teams at The James Hutton Institute and NIAB for providing technical support while using the computational resources on the respective HPC clusters, 'Maxwell' and the 'UK Crop Diversity Bioformatics HPC' (BBSRC grant BB/S019669/1), which have contributed to the results within this paper. We thank numerous staff at the Singapore Botanic Gardens, particularly Khoo-Woon Mui Hwang and Neo Wei Ling for strong technical support in the molecular laboratory, Matti Niissalo for generous sharing of laboratory protocols and bioinformatics tips, Choo Le Min for patient help with laboratory troubleshooting, Chong Kwek Yan and Chan Pin Jia for forest survey expertise, and Louise Neo for thoughtful field and R advice. Our sincere thanks are due to Herbarium staff at the Royal Botanic Garden Edinburgh, including Lesley Scott, for efficient assistance in sending silica-dried samples to the Singapore Botanic Gardens. We also thank Camille Christe at the Conservatoire et Jardin botaniques de la Ville de Genève for kind advice related to the taxon-specific baits' setup and purchase. Lastly, we thank the Associate Editor and anonymous reviewers for their invaluable feedback regarding the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

FUNDING

We are grateful to the Lady Yuen Peng McNeice Charitable Foundation for financial support of this study.

DATA AVAILABILITY

Raw sequence data (Table S1 in the Supporting Information) are available from National Center for Biotechnology Information under the BioProject number PRJNA912780: https://www.ncbi.nlm.nih.gov/bioproject/912780.

REFERENCES

Altenhoff AM, Glover NM, Dessimoz C. Inferring orthology and paralogy. In Anisimova M, ed, *Evolutionary Genomics: Statistical and Computational Methods*. New York: Springer. 2019:149–175.

Andrews KR, Good JM, Miller MR, *et al*. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 2016;**17**:81–92.

Baker WJ, Bailey P, Barber V, *et al*. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology* 2022;**71**:301–19.

Baker TR, Pennington RT, Magallon S, *et al*. Fast demographic traits promote high diversification rates of Amazonian trees. *Ecology Letters* 2014;**17**:527–36.

Bansal M, Bansal R, Wazir ND, *et al*. A comparative evaluation of different chemical agents and herbal products in disinfecting Gutta-Percha cones: an in vitro study. *Dental Journal of Advance Studies* 2020;**8**:36–9.

Baum DA, Shaw KL. Genealogical perspectives on the species problem. *Experimental and Molecular Approaches to Plant Biosystematics* 1995;**53**:123–4.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.

Boluda CG, Christe C, Naciri Y, *et al*. A 638-gene phylogeny supports the recognition of twice as many species in the Malagasy endemic genus *Capurodendron* (Sapotaceae). *Taxon* 2022;**71**:360–95.

Boluda CG, Christe C, Randriarisoa A, *et al*. Species delimitation and conservation in taxonomically challenging lineages: the case of two clades of *Capurodendron* (Sapotaceae) in Madagascar. *Plants* 2021;**10**:1702.

Borowiec ML. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 2016;**4**:e1660.

Brown JW, Walker JF, Smith SA. Phyx: phylogenetic tools for UNIX. *Bioinformatics* 2017;**33**:1886–8.

Burkill IH. *A Dictionary of the Economic Products of the Malay Peninsula. Volume II* **(IZ)**: London: Crown Agents for the Colonies, 1935.

Chai P, Yii P. Sapotaceae. In: Soepadmo E, Saw LG, Chung RCK (eds), *Tree Flora of Sabah and Sarawak* Kuala Lumpur: Forest Research Institute of Malaysia; 2002;**4**: 203–342.

Chang CC, Chow CC, Tellier LC, *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 2015;**4**:1–16.

Chantaranothai P. *Sapotaceae. Flora of Thailand* Bangkok: The Forest Herbarium; 2014;**11**:610–55.

Chau JH, Rahfeldt WA, Olmstead RG. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Applications of Plant Sciences* 2018;**6**:e1032.

Choo LQ, Bal TM, Choquet M, *et al*. Novel genomic resources for shelled pteropods: a draft genome and target capture probes for *Limacina bulimoides*, tested for cross-species relevance. *BMC Genomics* 2020;**21**:1–14.

Christe C, Boluda CG, Koubínová D, *et al*. New genetic markers for Sapotaceae phylogenomics: More than 600 nuclear genes applicable from family to population levels. *Molecular Phylogenetics and Evolution* 2021;**160**:107123.

Chung J, Son D-S, Jeon H-J, *et al*. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. *Scientific Reports* 2016;**6**:1–10.

Coissac E, Hollingsworth P, Lavergne S, *et al*. From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 2016;**25**:1423–8.

Couvreur TL, Helmstetter AJ, Koenen EJ, *et al*. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. *Frontiers in Plant Sciences* 2019;**9**:1941.

Cullings K. Design and testing of a plant-specific PCR primer for ecological and evolutionary studies. *Molecular Ecology* 1992;**1**:233–40.

De Queiroz K. Species concepts and species delimitation. *Systematic Biology* 2007;**56**:879–86.

De Queiroz K. An updated concept of subspecies resolves a dispute about the taxonomy of incompletely separated lineages. *Herpetological Review* 2020;**51**:459–61.
Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 1987;**19**:11–5.
Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* 2007;**22**:1–20.
Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 2005;**14**:2611–20.
Fernández R, Gabaldon T, Dessimoz C. Orthology: definitions, prediction, and impact on species phylogeny inference. In: Scornavacca C, Delsuc F, Galtier N, eds. *Phylogenetics in the Genomic Era*. Author open access book. 2020:2.4:1–2.4:13. Available at: https://hal.archives-ouvertes.fr/hal-02535414/
Freudenstein JV, Broe MB, Folk RA, *et al*. Biodiversity and the species concept—lineages are not enough. *Systematic Biology* 2017;**66**:644–56.
Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 2015;**31**:3718–20.
Gautier L, Naciri Y, Anderberg AA, *et al*. A new species, genus and tribe of Sapotaceae, endemic to Madagascar. *Taxon* 2013;**62**:972–83.
Hale H, Gardner EM, Viruel J, *et al*. Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Applications in Plant Sciences* 2020;**8**:e11337.
Hart ML, Forrest LL, Nicholls JA, *et al*. Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 2016;**65**:1081–92.
Hendriks KP, Mandáková T, Hay NM, *et al*. The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. *Applications in Plant Sciences* 2021;**9**(7):e11438.
Jessup L. A taxonomic revision of Sapotaceae for mainland Australia. *Austrobaileya* 2019;**10**:321–82.
Johnson MAT. Cytology. In: Pennington TD (ed.), *The Genera of Sapotaceae*. London: Royal Botanic Gardens Kew; 1991:15–22.
Johnson MG, Gardner EM, Liu Y, *et al*. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 2016;**4**:1600016.
Johnson MG, Pokorny L, Dodsworth S, *et al*. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 2019;**68**:594–606.
Kadlec M, Bellstedt DU, Le Maitre NC, *et al*. Targeted NGS for species level phylogenomics: 'made to measure' or 'one size fits all?'. *PeerJ* 2017;**5**:e3569.
Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 2013;**30**:772–80.
Kozlov AM, Darriba D, Flouri T, *et al*. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 2019;**35**:4453–5.
Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012;**9**:357–9.
Leaché AD, Oaks JR. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution and Systematics* 2017;**48**:69–84.
Lemmon AR, Emme SA, Lemmon EM. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 2012;**61**:727–44.
Li H, Handsaker B, Wysoker A, *et al*.; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
Lim RCJ, Leong-Škorničková J, Lindsay S, Niissalo MA, Yam TW, Middleton DJ. Plant conservation in Singapore II: practical implementation. In. Middleton, D.J., Leong-Škorničková, J. & Lindsay, S. (eds), *Flora of Singapore*. Singapore: National Parks Board. 2019; **1**:83–106.
Linck E, Battey C. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources* 2019;**19**:639–47.
Mackinder B, Harris D, Gautier L. A reinstatement, recircumscription and revision of the genus *Donella* (Sapotaceae). *Edinburgh Journal of Botany* 2016;**73**:297–339.
Maisano Delser P, Corrigan S, Hale M, *et al*. 2016. Population genomics of C. melanopterus using target gene capture data: demographic inferences and conservation perspectives. *Scientific Reports* 2016;**6**:33753.
Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* 2011;**17**:10–2.
McCormack JE, Tsai WL, Faircloth BC. Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources* 2016;**16**:1189–203.
Michel T, Tseng Y-H, Wilson H, *et al*. A hybrid capture bait set for *Begonia*. *Edinburgh Journal of Botany* 2022;**79**:1–33.
Middleton DJ, Leong-Škorničková J, Lindsay S. *Flora of Singapore*. Singapore: National Parks Board; 2019;**1**.
Morel B, Kozlov AM, Stamatakis A. ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics* 2019;**35**:1771–3.
Naciri Y, Linder HP. Species delimitation and relationships: the dance of the seven veils. *Taxon* 2015;**64**:3–16.
Ng FSP. Notes on the systematy of Malayan phanerogams V. Sapotaceae. *Gardens Bulletin Singapore* 1969;**24**:7.
Ng FSP. Sapotaceae. In: Whitmore TC. (ed.), *Tree Flora of Malaya*. Kuala Lumpur: Longman Malaysia, 1972;**1**:388–439.
Nicholls JA, Pennington RT, Koenen EJ, *et al*. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Sciences* 2015;**6**:710.
O'Connell KA, Mulder KP, Wynn A, *et al*. Genomic library preparation and hybridization capture of formalin-fixed tissues and allozyme supernatant for population genomics and considerations for combining capture-and RADseq-based single nucleotide polymorphism data sets. *Molecular Ecology Resources* 2022;**22**:487–502.
Paijmans JL, Fickel J, Courtiol A, *et al*. Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Molecular Ecology Resources* 2016;**16**:42–55.
Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019;**35**:526–8.
Pennington RT, Lavin M. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytologist* 2016;**210**:25–37.
Pennington TD. *The Genera of Sapotaceae*. London: Royal Botanic Gardens Kew, 1991.
Pina-Martins F, Silva DN, Fino J, *et al*. Structure_threader: an improved method for automation and parallelization of programs structure, fastStructure and MavericK on multicore CPU systems. *Molecular Ecology Resources* 2017;**17**:e268–74.
POWO. *Plants of the World Online*. Kew: Facilitated by the Royal Botanic Gardens. 2022. Available at: http://www.plantsoftheworldonline.org/ [Accessed 01 August 2022].
Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;**155**:945–59.
Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 2012;**3**:217–23.
R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. https://www.R-project.org/
Richardson JE, Bakar AM, Tosh J, *et al*. The influence of tectonics, sea-level changes and dispersal on migration and diversification of Isonandreae (Sapotaceae). *Botanical Journal of the Linnean Society* 2014;**174**:130–40.
Ridley H. 1902. *Annual Report on the Botanic Gardens, Singapore, for the Year 1902*: Singapore: Government Printing Office.
Rose JP, Kleist TJ, Löfstrand SD, *et al*. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient

Neotropical and East Asian connections. *Molecular Phylogenetics and Evolution* 2018;**122**:59–79.
Royen P. Revision of the Sapotaceae of the Malaysian area in a wider sense. XXIII. *Palaquium*Blanco. *Blumea* 1960;**10**:432–606.
Schmidt-Lebuhn AN. Sequence capture data support the taxonomy of Pogonolepis (Asteraceae: Gnaphalieae) and show unexpected genetic structure. *Australian Systematic Botany* 2022;**35**:317–25.
Shah T, Schneider JV, Zizka G, *et al*. Joining forces in Ochnaceae phylogenomics: A tale of two targeted sequencing probe kits. *American Journal of Botany* 2021;**108**:1201–16.
Shen X-X, Hittinger CT, Rokas A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* 2017;**1**:1–10.
Slimp M, Williams LD, Hale H, *et al*. On the potential of Angiosperms353 for population genomic studies. *Application of Plant Sciences* 2021;**9**:e11419.
Smedmark JE, Anderberg AA. Boreotropical migration explains hybridization between geographically distant lineages in the pantropical clade Sideroxyleae (Sapotaceae). *American Journal of Botany* 2007;**94**:1491–505.
Smedmark JE, Swenson U, Anderberg AA. Accounting for variation of substitution rates through time in Bayesian phylogeny reconstruction of Sapotoideae (Sapotaceae). *Molecular Phylogenetics and Evolution* 2006;**39**:706–21.
Smith SA, Moore MJ, Brown JW, *et al*. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 2015;**15**:1–15.
Straub SC, Boutte J, Fishbein M, *et al*. Enabling evolutionary studies at multiple scales in Apocynaceae through Hyb-Seq. *Applications of Plant Sciences* 2020;**8**:e11400.
Swenson U, Anderberg AA. Phylogeny, character evolution, and classification of Sapotaceae (Ericales). *Cladistics* 2005;**21**:101–30.
Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;**123**:585–95.
Thiers, B. *Index Herbariorum: A Global Directory of Public Herbaria and Associate Staff*. Facilitated by New York Botanical Garden's Virtual Herbarium, Bronx, NY, USA. continuously updated. Available at: http://sweetgum.nybg.org/science/ih/ [Accessed 01 August 2022].
Ufimov R, Zeisek V, Píšová S, *et al*. Relative performance of customized and universal probe sets in target enrichment: A case study in subtribe Malinae. *Applications in Plant Sciences* 2021;**9**:e11442.
Walker JF, Brown JW, Smith SA. Analyzing contentious relationships and outlier genes in phylogenomics. *Systematic Biology* 2018;**67**:916–24.
Wang J. A parsimony estimator of the number of populations from a STRUCTURE-like analysis. *Molecular Ecology Resources* 2019;**19**:970–81.
Wang Y, Chen F, Ma Y, *et al*. An ancient whole-genome duplication event and its contribution to flavor compounds in the tea plant (*Camellia sinensis*). *Horticulture Research* 2021;**8**:176.
Wells T, Carruthers T, Muñoz-Rodríguez P, *et al*. Species as a heuristic: reconciling theory and practice. *Systematic Biology* 2022;**71**:1233–43.
Wickham H, Averick M, Bryan J, *et al*. Welcome to the Tidyverse. *Journal of Open Source Software* 2019;**4**:1686.
Wilkinson MJ, Szabo C, Ford CS, *et al*. Replacing Sanger with Next Generation Sequencing to improve coverage and quality of reference DNA barcodes for plants. *Scientific Reports* 2017;**7**:1–11.
Woudstra Y, Viruel J, Fritzsche M, *et al*. A customised target capture sequencing tool for molecular identification of *Aloe vera* and relatives. *Scientific Reports* 2021;**11**:1–13.
Yan Z, Smith ML, Du P, *et al*. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Systematic Biology* 2022;**71**:367–81.
Yang Y, Smith SA. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* 2014;**31**:3081–92.
Yardeni G, Viruel J, Paris M, *et al*. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources* 2022;**22**:927–45.
Zhang C, Rabiee M, Sayyari E, *et al*. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 2018;**19**:15–30.
Zhang C, Sayyari E, Mirarab S. ASTRAL-III: increased scalability and impacts of contracting low support branches. In. Meidanis J, Nakhleh L. (eds), *Comparative Genomics. RECOMB-CG 2017. Lecture Notes in Computer Science*. Switzerland: Springer. 2017;**10562**:53–75.
Zhang Q, Zhao L, Folk RA, *et al*. Phylotranscriptomics of Theaceae: generic-level relationships, reticulation and whole-genome duplication. *Annals of Botany (London)* 2022;**129**:457–71.
Zhou W, Soghigian J, Xiang Q-Y. A new pipeline for removing paralogs in target enrichment data. *Systematic Biology* 2022;**71**:410–25.