universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Population genomics of a selected clade within the New Caledonian Diospyros radiation

verfasst von | submitted by

Florian Lukas Brück BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von | Supervisor:	Assoz. Prof. Dipl.-Ing. Dr. Ovidiu Paun

Acknowledgments

First of all, I would like to thank my thesis supervisor Ovidiu Paun for introducing me to this topic. Your expertise and insightful feedback, as well as your warm welcome to the group and support towards the end, have been invaluable. Furthermore, I would like to thank the whole plant genomics group for all the scientific and non-scientific inputs they provided me. Especially, I'd like to thank Clara, for being an exceptional office mate, who was always there to bounce ideas of off, Luiz, for our in depth chats about SFS and demographic inference (and why it wasn't working) and Katie, for inspiring me with her contagious enthusiasm for science. Additionally, I'd like to thank Teerna for dealing with all my questions when it came to naming plant populations, phylogenies and generally comparing notes whenever needed.

None of this would have been possible without the many hours invested in collecting these plants. Here a special thanks goes to Jerome Munzinger (IRD Montpellier, France) and David Bury (IRD Noumea, New Caledonia) for their support and expertise in the field. I also would like to thank Barbara Turner (BOKU Wien) and Rose Samuel (University of Vienna) for sharing their DNA extractions and silica gel fixed materials with us.

On a personal note, I would also like to thank my friends and family who have supported me during, not only this thesis, but the last years. A special thank you to Emily, for being the most supportive and understanding girlfriend anyone could wish for. In addition, I would like to thank Arne and, obviously, Emily, without whom studying and staying motivated during covid would have been a lot harder.

Funding

The research included in this MSc thesis was funded by the Austrian Science Fund (FWF) [grant DOI 10.55776/P35275].

Table of contents

Abstract

Adaptive radiations are major drivers of biodiversity in nature. Their study allows for insights into complex evolutionary processes like adaptation and speciation, and the mechanisms driving them. Due to their high biodiversity, complex geological history and distance from continental land masses the islands of New Caledonia provide ideal conditions for the study of adaptive radiations. Throughout their history the islands were colonized by plants of the genus *Diospyros* at least four times, with one colonization forming an extensive adaptive radiation of at least 24 species. To gain insights into a subclade of the recently radiated New Caledonian *Diospyros* a variety of bioinformatic tools were used on whole genome sequencing data of around 100 individuals. Besides the investigation of population structure and relatedness of individual samples and species or populations, a special interest was placed on the investigation of demographic history of populations. For this, different methods of demographic inference, based on the site frequency spectrum (SFS), were used and a range of simple to more complex models were tested. This was able to uncover widespread hybridization, as well as population substructure in some species. Demographic inference consistently showed a stable effective population size for several million years, with a strong recent decline in most species investigated. More complex models also provide support for hybridization as the origin of one population. The lack of reliable phylogenies and very diffuse evolutionary dynamics, as well as limitations when working with non-model organisms, restricted the scope of the analyses.

Zusammenfassung

Adaptive Radiationen sind ein wesentlicher Treiber der Artenvielfalt in der Natur. Ihre Studie erlaubt es, komplexe evolutionäre Prozesse, wie Adaption und Artbildung, sowie die zugrundeliegenden Mechanismen zu verstehen. Durch ihre hohe Biodiversität, geologische Geschichte und Entfernung von kontinentalen Landmassen bieten die Inseln von Neu Kaledonien eine ideale Grundlage für die Untersuchung von adaptiven Radiationen. Im Laufe ihrer Geschichte wurden die Inseln vier Mal von Pflanzen der Gattung *Diospyros* besiedelt, wobei eine der Besiedelungen eine weitgreifende adaptive Radiation mit mindestens 24 Arten zur Folge hatte. Um Einblicke in eine Untergruppe von rezent radiierten New Kaledonischen *Diospyros* zu gewinnen, wurde eine Vielfalt an verschiedenen bioinformatischen Werkzeugen auf einen Datensatz von circa 100 Genom Sequenzen angewendet. Neben der Untersuchung von Populationsstruktur und Verwandtschaftsverhältnissen von einzelnen Individuen und Arten und Populationen, wurde ein besonderes Augenmerk auf die demographische Geschichte der Populationen gelegt. Dafür wurden verschiedene Methoden der demographischen Inferenz, basierend auf Spektren der Allelfrequenz (SFS), verwendet und eine Reihe an Modellen, von simpel bis komplex, getestet. Damit war es möglich, weitgreifende Hybridisierung und auch Feinstruktur der Populationen in manchen Arten aufzudecken. Demographische Inferenz ermöglichte es, stabile Populationsgrößen über mehrere Millionen Jahre, mit einem rezenten Abfall in effektiver Populationsgröße in den meisten Arten, aufzuzeigen. Komplexere Modelle demonstrierten auch die Möglichkeit der Entstehung einer hybriden Population einer Population. Der Mangel an zuverlässigen phylogenetischen Stammbäumen und sehr diffuse evolutionäre Dynamiken, sowie Limitationen durch die Arbeit mit Nicht-Modellorganismen, schränkten den Umfang der Analysen ein.

Introduction

Adaptive radiations describe the rapid increase in species diversity through adaptive processes. They are most often characterized by a change in the environment of an ancestral population and therefore newly available resources across an array of ecological niches. This can be most pronounced when an ancestral lineage is introduced to a wholly different geographical location through long-distance dispersal (McGee et al., 2020; Seehausen, 2004; Stroud & Losos, 2016). As they are of great interest for the study of evolutionary processes and especially speciation, many animal radiations have been investigated before (Marques et al., 2019; McGee et al., 2020; Petren et al., 2005; Seehausen, 2006), but plant radiations are still under-investigated. Different studies have previously investigated the underlying genetic mechanisms by which these adaptive radiations occur. As only a limited number of individuals, and therefore a small proportion of genetic variation, will participate in colonizing the new environment, initially genetic diversity of the founder population will necessarily be low. Genetic diversity can then accumulate through multiple waves of colonization, de-novo mutation, hybridization with related lineages, chromosomal rearrangements, insertion of transposable elements or polyploidization. Much evidence points to introgression or hybridization as a way of capitalizing on this newly emerged variation, through sharing of polymorphism, or formation of homoploid hybrid populations or species. In recent radiations the low divergence time and fast speciation rates, resulting in low genetic differentiation between species would expedite hybridization and might have compounding effects (Cerca et al., 2023; García-Verdugo et al., 2013; Marques et al., 2019; Meier et al., 2017).

During their history, the islands of New Caledonia, situated in the Pacific Ocean about 1,600 km North-East off the coast of Australia, were submerged and accumulated heavy-metal rich oceanic sediments on their surface. After their emergence, which happened ca 30 mya, erosion of parts of the archipelago’s surface led to a mosaic of different soil types. The period of submergence also created a “blank slate” wiping out any previously occurring species and allowing for

new species to colonize and adapt to New Caledonia's unique and diverse conditions (Grandcolas et al., 2008). In addition the main island bears a mountain range along its length with sharp elevational gradients and differing climate conditions emerging on either side of the island. This lends itself to create an extremely diverse environment, which in turn led to the islands hosting a high level of biodiversity, much of it endemic to the archipelago (Wulff et al., 2013).

The genus *Diospyros L.* (Ebenaceae) consists of over 700 species of trees and shrubs with a wide distribution range and includes several species used for their timber or edible fruits. The islands of New Caledonia were colonized by different ancestral species of *Diospyros* at least four times during the last 25 Million years, resulting in 31 currently described species of *Diospyros* on the island. From only one of these dispersals an extensive adaptive radiation emerged, producing approximately 24 recognized species (likely more) of *Diospyros*. The radiated species are widely dispersed over the whole of the island group, being morphologically distinct and occupying many different soil types and climate conditions, in stark contrast to the non-radiated New Caledonian *Diospyros* lineages (Duangjai et al., 2009; Paun et al., 2016; Samuel et al., 2019; Turner et al., 2013).

Previous phylogenetic analyses into the species of *Diospyros* endemic to New Caledonia showed that the radiated New Caledonian *Diospyros* are most closely related to species found on other Pacific islands (Duangjai et al., 2009; Turner et al., 2013). Further studies done by Paun and colleagues (2016) used RADseq to try and resolve specifically the phylogeny of the previously described adaptive radiation. Although improving on previous phylogenies based on multiple DNA loci and genome-wide fingerprinting analyses, this however proved to be difficult with some inferred relationships not well supported. This probably stems from the limited number of generations since colonization and the severe bottleneck associated with it, leading to a large proportion of recently evolved variants found in only a few individuals. Additionally, fast diversification and admixture between the radiating species seems to have further blurred phylogenetic signals. The likely presence of cryptic species through convergent or parallel evolution complicates reliance on previously generated phylogenies and species definitions based solely on

phenotype. New Caledonian *Diospyros* plants are dioecious. Dispersal in the group is believed to happen for pollen via small flies, and for seeds via various fruit eating animals, mostly birds, such as hornbills and pigeons (Carpenter et al., 2003; Tassin et al., 2010). Both mechanisms likely promote a local gene flow, which may drive a prominent population structure.

The goals for this Master thesis have been to investigate the evolutionary history focusing on demographic events for a subgroup of the radiated New Caledonian *Diospyros* with previously unclear phylogenetic relationships and, in part, debated species delimitation. This subgroup consists of seven or eight species in the current circumscription which differ in soil and environmental preference. For this group various population genomics approaches have been used, taking advantage of newly re- sequenced whole genome (WGS) data. Further, coalescent based demographic inference has been carried out to elucidate the demographic history of the studied populations.

Methods

Sample identification

All analyses were carried out using whole genome resequencing data of New Caledonian *Diospyros* sampled over a period of over ten years by members of the group and collaborators (Figure 1). Sampling locations and metadata, as soil composition and ecology were recorded along with herbarium samples for later cross reference. Due to evolving research objectives, some locations were repeatedly sampled, hence there is a certain risk that some accessions have been sampled twice. Location names were retained from previously published studies of the same individuals (Paun et al., 2016) and proximate localities for easier comparisons. In the field individual samples were assigned to a species by morphologic and ecologic features. However, precise species identification has proven to be difficult for some members of this radiation in the past. For some localities this also traced back to the lack of fruits at sampling, which otherwise offer characteristic traits especially in the shape of their persistent calix. As all New Caledonian *Diospyros* are dioecious, a significant portion of individuals do not carry fruits at any time. To more robustly assign individuals to putative species, preliminary phylogenetic results had to be generated for samples not previously analyzed.

Chosen sequences were adapter and quality trimmed using trim-galore v0.6.10 (Krueger, 2015) using default parameters in paired end mode, with the exception of stringency for adapter removal at the end of the read (--stringency 2; default 1). Subsequently, trimmed reads were then aligned to a reference genome for *Diospyros sandwicensis* v1.0 (Paun, Khastgir and Emelianova, unpublished), a Hawaiian species closely related to members of the radiation. The short read alignment tool bwa mem v0.7.17-r1188 (Li, 2013) was used for mapping reads to the selected reference. Reads not mapping to the reference (-F "mapping_quality >= 1") were discarded and the sam alignment file was transformed into bam format with sambamba v0.6.6 (Tarasov et al., 2015). Lastly, samtools v1.13 (Li et al., 2009) was used for sorting and indexing bam files. Samtools view was then used to extract

reads mapped to the five longest scaffolds of the reference (4.1% of total sequence length of all 581 scaffolds) based on a bed file of these.

Variant calling was performed on the generated bam files using bcftools v1.9 (Danecek et al., 2021) in parallel (Tange, 2018) on each scaffold to speed up computation. For this, bcftools mpileup was run to generate genotype likelihoods only on one scaffold with the options --ignore-RG (for samples sequenced multiple times in different libraries) and -a FORMAT/AD,FORMAT/DP to add fields for per sample allelic depth and coverage. Output from bcftools mpileup was then piped to the variant caller bcftools call using the multi-allelic variant caller and only retaining variant sites (option: -mv). As the calling was made across multiple species, samples were assumed not to be in Hardy Weinberg equilibrium (HWE, option: -G -) and calls were output in compressed binary variant call format (BCF, option: -Ob). This resulted in one BCF file per scaffold, which were later concatenated using bcftools concat. For quality control statistics were generated using bcftools stats.

VCFtools v0.1.16 (Danecek et al., 2011) was then used to filter variant calls based on minor allele frequency (option: --maf 0.0077 , 2/2n with n being the number of samples), missingness, quality (options: --max-missing 0.85 --minQ 30), individual minimum depth and both mean minimum and maximum depth (options: --min-meanDP 8 --max-meanDP 40 --minDP 5). Additionally, indels were removed and only biallelic single nucleotide polymorphisms (SNPs) were retained (options: --remove-indels --min-alleles 2 --max-aleles 2). Again statistics for the filtered set of variants were generated with bcftools stats. To speed up computation in subsequent steps variants were subsampled to 10 % of filtered variants with vcfrandomsample (*Vcflib*, 2010/2023) and the subsampled VCF file was transformed into a multiple alignment fasta file using the python 3 script vcf2phylip.py (Ortiz, 2017/2023) resolving ambiguous (i.e. heterozygous) sites by choosing possible bases randomly. This multiple alignment fasta file was used as input for FastTree v2.1.11 (Price et al., 2010) to generate a maximum-likelihood phylogenetic tree.

Samples for further analyses could then be selected not only based on their initial taxonomic assignment, but also based on their clustering in the generated

phylogeny. Accensions investigated before and those with existing phylogenetic information could then be cross referenced to identify those belonging to the clade of interest. Finally, herbarium vouchers have been cross checked by Jerome Munzinger (University of Montpellier, France) for cases where initial identification in the field conflicted genetic assignment.

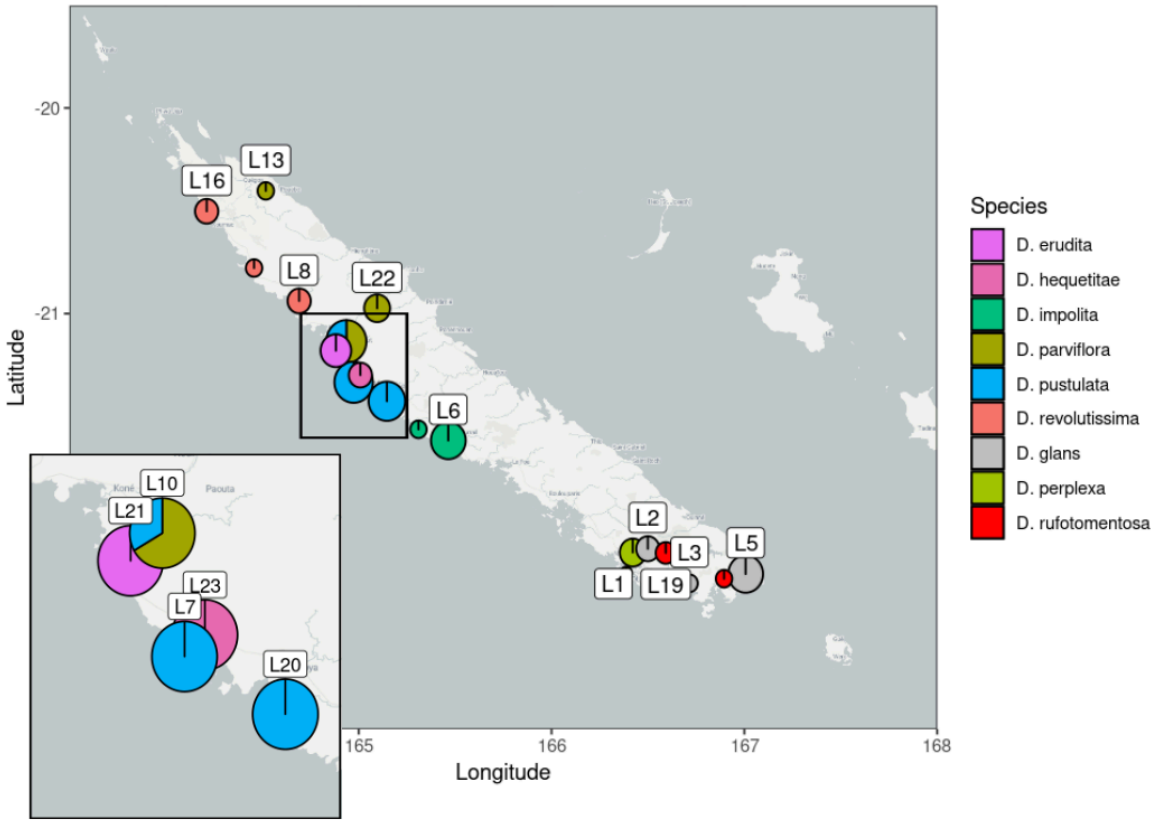


Figure 1: Map of New Caledonia with sampling locations of individuals included in analyses (n total = 102). Pie charts show composition of used species at sampling sites as initially identified, indicated by color. Size of pie charts in the total map is proportional with the number of individual samples from each site (except for the zoomed in portion). All individuals sampled in the South of the main island were later identified to be either D. rufotomentosa *or* D. perplexa *and therefore excluded. The zoomed in portion shows sampling close together, with a comparatively larger number of species within the clade growing in proximity. These locations also showed most difficulties with correct identification of species.*

Population genomics

The selected individuals belong to a subclade of the radiation of *Diospyros* in New Caledonia, with previously not fully resolved phylogenetic relations. Accessions identified to be *D. impolita*, *D. erudita* (including *D. hequetiae*), *D. revolutissima*, *D. parviflora*, and *D. pustulata* were chosen. Additionally, samples spBT196 (genetically identified as *D. parviflora*) and yahouensisH3637 (genetically identified as *D. pustulata*) were included in the group of interest. In preliminary analyses, some accessions of *D. glans*, *D. rufotomentosa*, *D. perplexa* and *D. pancheri* have also been analysed.

Accensions which weren't trimmed before were now trimmed using trim-galore v0.6.10 (Krueger, 2015) as described above. Trimmed reads for all 102 individuals were then mapped against a reference genome of *Diospyros impolita* v1.1 (Paun, Khastgir and Emelianova, unpublished) using bwa mem. This reference genome was chosen as being closer to the target group, plus being more contiguous than the reference of *D. sandwicensis*. Mapped reads were filtered, sorted and indexed as described above.

ANGSD v0.941 (Korneliussen et al., 2014) was then used to create a genotype likelihood beagle file of BAM files for all previously selected accensions. To exclude sites falling within transposable elements (TEs) a prepared gff file of identified transposable elements (Emelianova, unpublished) was used. A genome index of the *D. impolita* reference was created using the first two columns of the fasta indexed reference, obtained with samtools faidx. From the resulting file and the given gff file of TEs, a list of sites excluding TEs was then generated using bedtools complement v2.30.0 (Quinlan & Hall, 2010) with the TE gff file and the genome index as input. The resulting bed file was then modified to conform to ANGSD indexing standards by adding +1 to all positions and then used in the -sites option in the ANGSD variant calling. Because of memory constraints, ANGSD was run separately scaffold by scaffold using the -r (regions) option, which also allowed to discard scaffolds which showed abnormal insert size when mapping illumina reads back onto the reference genome (TS Khastgir, pers. comm.). Minimum mapping quality and minimum variant

quality were set to 20, whereas the minimum number of individuals for a site to be considered at ~80 % of the total number of individuals (options: -minMapQ 30 -minQ 20 -minInd 80). A genotype likelihood beagle file was created (options: -GL 2 -doMajorMinor 1 -doGlf 2) and a list of all sites conforming to the selected quality filters was extracted. A custom python script was then used to further thin the list of retained sites by selecting one site at least every 1k base pairs apart to reduce analyzed sites (computational constraints) and account for linkage disequilibrium (LD). Finally, ANGSD was then rerun on only the retained sites after TE-removal and thinning with the added parameters of -SNP-pval 2e-6 and -doMaf 1 to create a beagle genotype likelihood file of single nucleotide polymorphisms (SNPs).

As sampling of multiple individuals of one species was often performed at the same site, both at the same time and during subsequent sampling trips, the presence of closely related individuals had to be investigated. Analyses of relatedness were carried out with NGSrelate (Hanghøj et al., 2019; Korneliussen & Moltke, 2015) as described in the manual. ANGSD was rerun on the selected sites from above with the option -doGlf 3 to obtain a three column genotype likelihood file and a file of allele frequencies (-doMaf 1). The allele frequency file was then manipulated as described in the github manual for NGSrelate by extracting the column containing frequency values and removing the header. NGSrelate could then be run with the generated genotype and frequency input files. In total 25 individual samples were found to be closely related (i.e. at least 1st degree relation) with a KING kinship coefficient >0.177 (Manichaikul et al., 2010). Samples erudita1106 and eruditaBT280, as well as impolita1121 and impolitaBT102 were assumed to represent repeated samplings of the same individual respectively, so only one of each pair has been retained for further analyses.

NGSadmix (Skotte et al., 2013) was used for estimation of admixture proportions in the sampled individuals. The program is able to estimate individual proportions of admixture for K ancestral populations from beagle genotype likelihood files. For each ancestral population number K, ten iterations of the program were run, with K from two to ten. The resulting log-likelihoods for each iteration for each number of ancestral populations were then compiled into a log-probability table as specified on

the CLUMPAK (Cluster Markov Packager Across K) website https://tau.evolseq.net/clumpak/bestK.html (Kopelman et al., 2015) and the best K was determined (Supplemental 11). Admixture proportions generated by NGSadmix were then visualized using pong v.1.5 (Behr et al., 2016). This approach allowed us to identify ancestral proportions and population structure across and within species. Furthermore, recent hybrids present in the dataset, which could lead to issues with demographic inference later on, could be identified. The input beagle file was then subset by removing individuals identified with this approach. In the first instance individuals from the outgroup *D. glans* and individuals of *D. rufotomentosa, D. perplexa* and *D. pancheri* were removed. Based on the retained subset NGSadmix was then run again as described above. Another subsetting step was then performed where individuals that showed signs of recent hybridisation were excluded. In addition, individuals that were identified as close relatives of other individuals were removed as well, totalling twelve individuals, and NGSadmix was rerun a last time.

Principal component analyses of the above generated SNP files were then done using PCAngsd (Meisner & Albrechtsen, 2018). The minimum minor allele frequency was set to two alleles per 2n individuals (--minMaf 0.0098; 0.0131; 0.0185) to exclude very low frequency alleles for all three datasets. PCAngsd generates a covariance matrix of all individuals present in the datasets, which was imported into R 4.2 (R Core Team, 2022). The covariance matrix was then visualized using ggplot2 (Wickham, 2016) and the eigenvalues and eigenvectors were calculated. The PCA was then visualized using ggplot2 as a coancestry heatmap as well. Individuals were colored based on species affiliation. Suspected hybrid individuals based on results from NGSadmix were then highlighted. The PCAngsd --filter function was then used for further subsetting to include only individuals of one species or population as determined by NGSadmix. This was done to identify population substructure problematic for later demographic inference. In addition to this, PCAngsd was utilized for admixture analysis (--admix option) for some subsets with automatic detection of best K by the program.

Ancestral state reconstruction

As the investigated clade also encompasses the species of *Diospyros* used to create the reference genome, *D. impolita*, special attention had to be paid to the generation of unfolded site frequency spectra while using reads aligned to this reference. For this it was attempted to reconstruct an ancestral sequence from individual representatives of species within the New Caledonian radiation as well as one outgroup (*D. sandwicensis*). Individuals were chosen to enable the reconstruction of the ancestral sequence at the last common ancestor node of the investigated clade, based on the tree topology published in Paun et al. (2016). The method was previously published in Luqman et al. (2023) and was followed with minor modifications.

In the first step reads of eight individuals aligned to the *D. impolita* v1.1 draft reference were chosen for variant calling. The eight individuals included one representative each of *D. sandwicensis*, as an outgroup to the radiation, *D. viellardii*, *D. flavorcarpa*, *D. umbrosa*, *D. cherrieri* and *D. trisulca*, as species earlier diverged in the radiation, *D. tridentata* as a representative of a sister clade, and lastly *D. impolita* as a representative of the clade of interest. Variants were called on each accession separately using freebayes v1.3.6 (Garrison & Marth, 2012) while reporting monomorphic sites and all haplotypes and outputting in genomic variant-call format (gvcf).

The resulting variant calls for each species were then subset utilizing bcftools v1.9 (Danecek et al., 2021) filtering options into one file each containing variant sites with a depth over three and a quality score over 20, variant sites which did not pass this filtering step and sites with a depth lower than three. The latter subsets were then combined using bcftools concat to create a masking file. All files were indexed using both bcftools index and gatk4 v4.4.0.0 IndexFeatureFile (McKenna et al., 2010) and a sequence dictionary was created for the impolita reference with the use of picard's CreateSequenceDictionary (available from http://broadinstitute.github.io/picard).

Following this, variants in the first subset were applied to the original reference with the help of gatk4's FastaAlternateReferenceMaker, while masking sites in the respective masking files, i.e. sites that did not pass bcftools filtering. This resulted in one fasta sequence per individual used. All eight fasta sequences were then combined by linearizing them and concatenating into one large fasta with one sequence representing each individual. Additionally the reference sequence of *D. impolita* was added to this file in the same way. As all sequences were based on the original coordinates of one reference sequence a multiple-sequence alignment is not necessary.

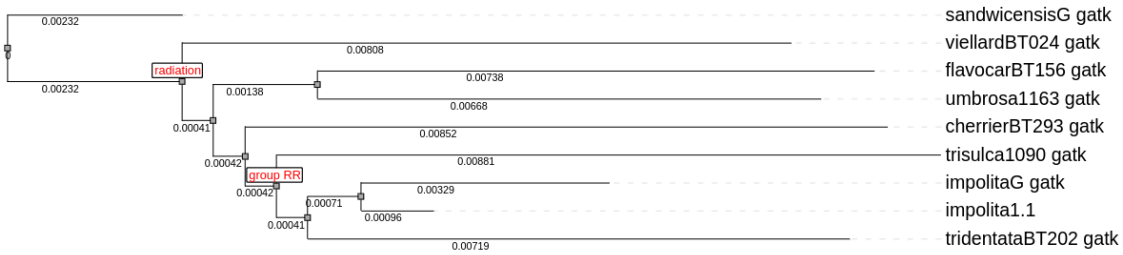


Figure 2: Phylogenetic tree with branch lengths used in the process of inferring ancestral states. Topology after Paun et al. (2016). Branch lengths generated with PHAST phylofit. Nodes of interest are highlighted, with node "radiation" representing the most basal node of the New Caledonian Diospyros radiation and "group RR" representing the last common ancestor of species investigated in this thesis.

A phylogenetic model was then computed using PHAST phyloFit v1.6 (Hubisz et al., 2011) with the tree topology for the selected individuals in newick format and the fasta file containing all individuals described above. The resulting model (Figure 2) was then modified by inserting labels for the ancestral nodes of interest and nucleotide probabilities at these nodes were computed by PHAST prequel v1.6 while keeping sequence lengths consistent (--keep-gaps). This resulted in a table of nucleotide probabilities for each position in the sequence at the specified ancestral node. The table of nucleotide probabilities was then split into scaffolds following the reference fasta index of the original sequence utilizing a custom python script. This step was necessary to minimize the memory requirements for the analysis. Another python script (*available from* https://github.com/harmsm/phylo_tools/blob/master/

structural-anc/anc/phy2fasta.py) was then used to translate nucleotide probabilities to the respective nucleotide and therefore create one fasta sequence file per scaffold, which were then concatenated to obtain the full reconstructed fasta sequence for the specified ancestral node.

Generation of site frequency spectra (SFS)

Heterozygosity for each individual accession of the clade of interest, identified earlier, was calculated using ANGSD v0.940 and realSFS (Mas-Sandoval et al., 2022). For this angsd -doSaf was run on each sample separately only on sites as defined above and a SFS for each sample was obtained using realSFS. As each SFS only contained data for one individual the second column (singletons) could be interpreted as the count of heterozygous sites. Importing in R v4.2 heterozygosity was then calculated as the number of heterozygous sites divided by the number of total sites.

Individuals for the generation of SFS for each population were chosen based on previous results. Excluded were individual samples which showed signs of recent hybridization or close relationship to other samples, based on results from NGSadmix, PCangsd and NGSrelate. Furthermore, whenever clear population substructure was visible, the individuals were split into separate populations. This resulted in eight populations, with *D. parviflora* being split into two (L10 and, respectively, L22), *D. pustulata* being split into three (L7, L10 plus L20, and cf. *pustulata*), and *D. erudita* s.s. and *D. hequetiae* being treated as one. Populations defined this way consisted of between five and nine individual samples.

Site frequency spectra were then generated for each of these populations using ANGSD with the -doSaf 1 and -GL 2 (gatk method of variant calling) options while filtering sites for quality, mapping quality and excessive mismatches (-minMapQ 30 -minQ 20 -baq 1 -C 50). Only sites contained in the above specified sites file were used to account for linkage. The SFS were polarized using the generated ancestral states at the node *D. trisulca* - *D. impolita* (“group_RR”; see Figure 2). Generation of SFS was done via realSFS by first concatenating files for all scaffolds and then

estimating SFS from the concatenated file. For populations with eight or more individual samples, folded SFS were estimated using the same method, while specifying the reference sequence as ancestral and then folding the SFS with realSFS. Joint SFS were then calculated using realSFS and previously generated single population files for population combinations of interest and for both folded and unfolded spectra.

For the calculation of Fst with ANGSD, populations were randomly subsampled to five individuals each and *D. hequetiae* was included as an independent population of four individuals, separated from *D. erudita*. Folded joint SFS were then estimated for all pairwise combinations of populations and Fst was calculated with realSFS fst index and fst print across all analyzed sites. Weighted Fst results (Bhatia et al., 2013) were then compiled into a matrix and visualized using R and ggplot2.

Demographic inference

Demographic inference was carried out on described populations (Table 1) using both folded and unfolded SFS generated with ANGSD and realSFS. For single population models fastsimcoal 2 (v27) as well as stairwayplot v2 were utilized. Both methods are based on coalescent theory. Stairwayplot 2 (Liu & Fu, 2020) uses a non-parametric or model-flexible approach for estimation of recent demographic changes of one population. Fastsimcoal 2 (Excoffier et al., 2021) is a tool that allows for the simulation of demographic events under a specified model for one or more populations, allowing estimating parameters under a composite likelihood maximization method. For demographic inference generation times for all species were estimated at 7 years and mutation rate per base per generation was set to 7e-9 as in previous studies (Paun et al., 2016). These represent general estimations based on generation times of some *Diospyros* tree species and known plant mutation rates of model organisms like *A. thaliana* and rice.

Table 1: Populations used for demographic inference, after all closely related, hybrid and outgroup individuals were removed. Species column indicates both the species and the originally assigned species.

Population name	Species	n Individuals	Location	Individual IDs
D. impolita	*D. impolita*	9	L6 (La Roche Percée)	1120;1121;1122; 1123; BT101;BT103;BT104; BT105;impolitaG
D. revolutissima	*D. revolutissima*	8	L8 (Voh); L16 (Paagoumène); Between Kaala-Gomen and Ouaco	1070;1073; BT118;BT119;BT120; BT218;BT219;BT221
D. erudita	*D. erudita s.s.* *D. hequetiae*	9	L21 (Plateau de Tia); L23 (Pindai, Baie de Nepoui)	1106;1107; BT284;BT285;BT287; 1025;1026;1029;1031
D. parvilflora L22	*D. parviflora*	5	L22 (Plateau de Tango)	BT288;BT289; BT290;BT291; BT292
D. parviflora L10	*D. parviflora*	6	L10 (Conservatoire botanique de Tiéa)	1099;1100; 1103;BT148; BT150;BT151
D. pustulata L7	*D. pustulata*	5	L7 (Pindai, Baie de Nepoui)	1115;BT111; BT112;BT113; BT114
D. pustulata L10/L20	*D. pustulata*	7	L20 (Metzdorff property, Moindah); L10 (Conservàtoire botanique de Tiéa)	1134;BT137;BT139; BT140;BT267; BT268;BT271
D. cf. pustulata	*D. pustulata*	5	L20 (Metzdorff property, Moindah, Poya South); L7 (Pindai, Baie de Nepoui)	1139;H3637; BT260;BT273; BT275

Stairwayplot v2 was run for each population, including folded SFS were applicable, following the tools manual. The provided blueprint file was modified with the respective SFS and total number of sites, break points (nrand) were set to ¼ * (2n-2), ½ *(2n-2), ¾ * (2n-2) and 1 * (2n-2) as recommended and all other options were left at default. Both best output under n random breaks estimated by the tool and other n random breaks were considered. When comparing results based on folded and unfolded input spectra the lower number of random breakpoints, chosen as best fit, was considered, to account for possible overfitting when using the unfolded spectrum.

Furthermore, fastsimcoal 2 was used to estimate effective population time for the last ~9 million years (1.3 M generations), which is the approximate age of the radiation

and dispersal into New Caledonia (Paun et al., 2016). The model was defined over three discrete time steps with variable start times and a set end time of 1.3 M generations. Effective population size was determined for all resultant timepoints, current population size, size at time point one, at time point two and ancestral size at 1.3M generations back. The program was run one hundred times to try to find the global optimum, while running 600,000 simulations and one hundred optimization cycles in each run. The best run was then determined using a custom script which chooses the run with the lowest likelihood difference under the estimated parameters.

Fastsimcoal 2 was also utilized for its parameter estimation (Supplemental 5) under more complex models encompassing two or more populations. Two species pairs (i.e., *D. revolutissima - D. erudita*, and *D. revolutissima - D. impolita*) were chosen based on the number of available sequences and previous results concerning their relationships. Multiple population divergence models were compared, simulating isolation, constant migration, ancestral migration and secondary contact. For each model sixty independent instances of fastsimcoal were run with 500,000 simulations and sixty optimization cycles. This was done with both folded and unfolded joint SFS. For all runs of fastsimcoal with multidimensional SFS cells of less than five observed counts were not taken into account (option -C 5). The best fitting run per model was determined as per the lowest likelihood difference between the observed and expected SFS, i.e. the best fit. Following this, one hundred iterations of fastsimcoal were run using the estimated best parameters of the initial run as input to generate likelihood distributions. Finally, the model with the lowest likelihood differences was determined as the best fitting model, while it was assumed that models with overlapping likelihood distributions indicated equally good fit for the observed data. Additionally, the Akaike information criterion (Akaike, 1998; Cavanaugh & Neath, 2019) was calculated in R based on the previously established maximum estimated likelihood and number of parameters for each model. This was used in addition to likelihood distributions to determine the best fitting model (Supplemental 6).

As, the model with the best fit was determined to be constant migration after split, the model was adjusted slightly to allow for more accurate parameter estimation. For

this both populations were allowed to change in effective population size at any time point before 100,000 generations in the past, based on results of single species demographic inference, which showed a stark decrease in effective population size in the more recent past. This was done to hopefully obtain more accurate parameter estimations and generate a better fit to the observed data.

To test whether hybridization played a significant role in the emergence of new species another complex model was tested. Interpreting results from NGSadmix and admixture analysis with PCAngsd, the assumption was made that populations at L13 and L22 of *D. parviflora* share a large proportion of ancestry with both *D. revolutissima* and the *D. parviflora* population from L10 (Figures 4 and 12) and arose through hybridization of ancestral populations of the latter. For this scenario multiple three population models were tested, with the goal of determining which speciation scenario allows for the best fit to the observed SFS. These were constructed based on previously established phylogenies from plastid and RAD sequencing (Paun et al., 2016; Samuel et al., 2019; Turner et al., 2013), as well as preliminary results from whole genome phylogenetic analyses (Khastgir unpublished). Scenarios tested included single origin models, where *D. parviflora* splits from *D. revolutissima* and then experiences a secondary split, two origin models, where *D. parviflora* L13/L22 split from *D. revolutissima*, and a hybridization model with an ancestral split of *D. revolutissima* and *D. parviflora* L10 and a subsequent hybridization giving rise to *D. parviflora* L13/L22. Under all models parameters of effective population size, migration and divergence/hybridization times were estimated (Supplemental 7) based on pairwise unfolded jSFS for pairs of populations. Fastsimcoal was again run with 500,000 simulations and 60 optimization cycles in sixty independent instances, with the best fit for parameter estimations based on likelihood difference (Figure 3). The best model was then chosen as described above for two population models (Supplemental 8).

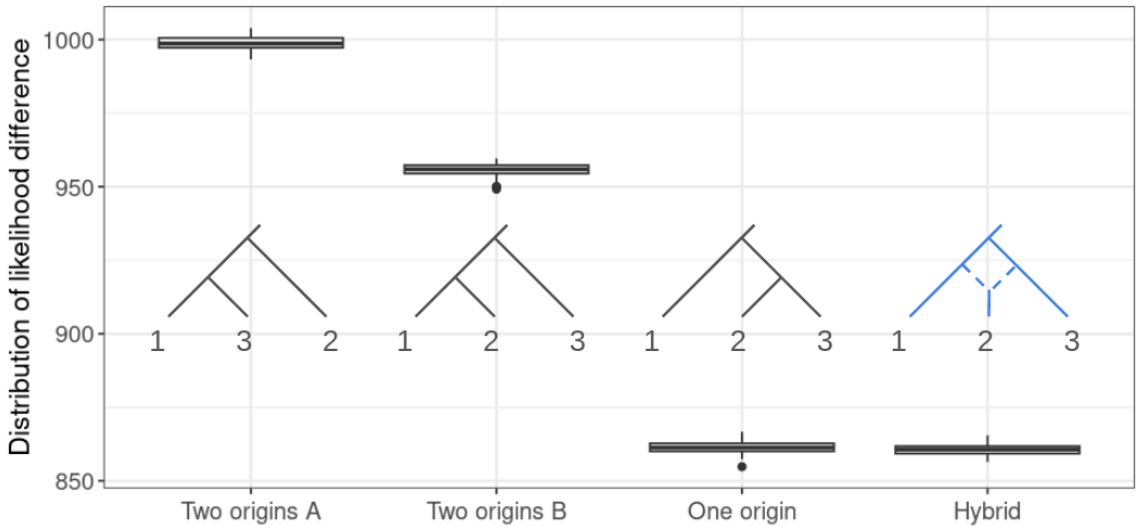


Figure 3: Plotted likelihood difference of the tested three population models. Lower values indicate a better fit. Both hybrid and single origin models performed better than two origin models. Trees show representations of the respective models tested (migration modeled but not shown). Population 1 corresponds to D. revolutissima, population 2 to D. parviflora L22 and population 3 to D. parviflora L10.

For all runs of fastsimcoal 2 estimated parameters and frequency spectra were visualized utilizing scripts provided on the fastsimcoal homepage (http://cmpg.unibe.ch/software/fastsimcoal28/). SFStools was used for plotting both observed and estimated, one- and two-dimensional, SFS for visual comparison of fit (Supplemental 9). Estimated parameters were either visualized with a modified Parfileviewer Rscript or directly in R.

Results

Read counts for the used samples ranged between 52M and 293M read pairs with accessions under 50M reads being excluded. This resulted in a set of 102 samples included as members of the clade by either taxonomic or preliminary phylogenetic assessment. The average coverage was ~17x with a minimum of 6x coverage for a *D. glans* individual (BT018) and a maximum of 38x also for an accession of the same species (1153).

Population structure

Results from both NGSadmix and PCA, as well as independently generated phylogenies, show that the used methods generally are able to recapitulate previously identified species. NGSadmix results (Figure 4A) for all individuals, including the outgroup and samples later assigned as *D. rufotomentosa* or *D. perplexa*, showed expected ancestry profiles for both clusterings K=3 and K=6, determined with Evanno's delta K (Evanno et al., 2005). For both three and six hypothetical ancestral populations *D. impolita* exhibits no admixture of other ancestral populations, with the exception of two individuals (JM4142 and VH3809). The outgroup *D. glans,* as well as *D. rufotomentosa* and *D. perplexa* share a common ancestry under K=3. All other species share varying degrees of ancestry with significant admixture of the outgroup ancestry in *D. parviflora* and *D. revolutissima*, while some individuals of *D. pustulata* share ancestry with *D. impolita*. When K=6 is considered there is a clear split between *D. glans* on one side and *D. rufotomentosa* and *D. perplexa* on the other. Furthermore, *D. parviflora* as well as *D. pustulata* both exhibit ancestry proportions mostly unique to them respectively, but with significant admixture proportions depending on the locality.

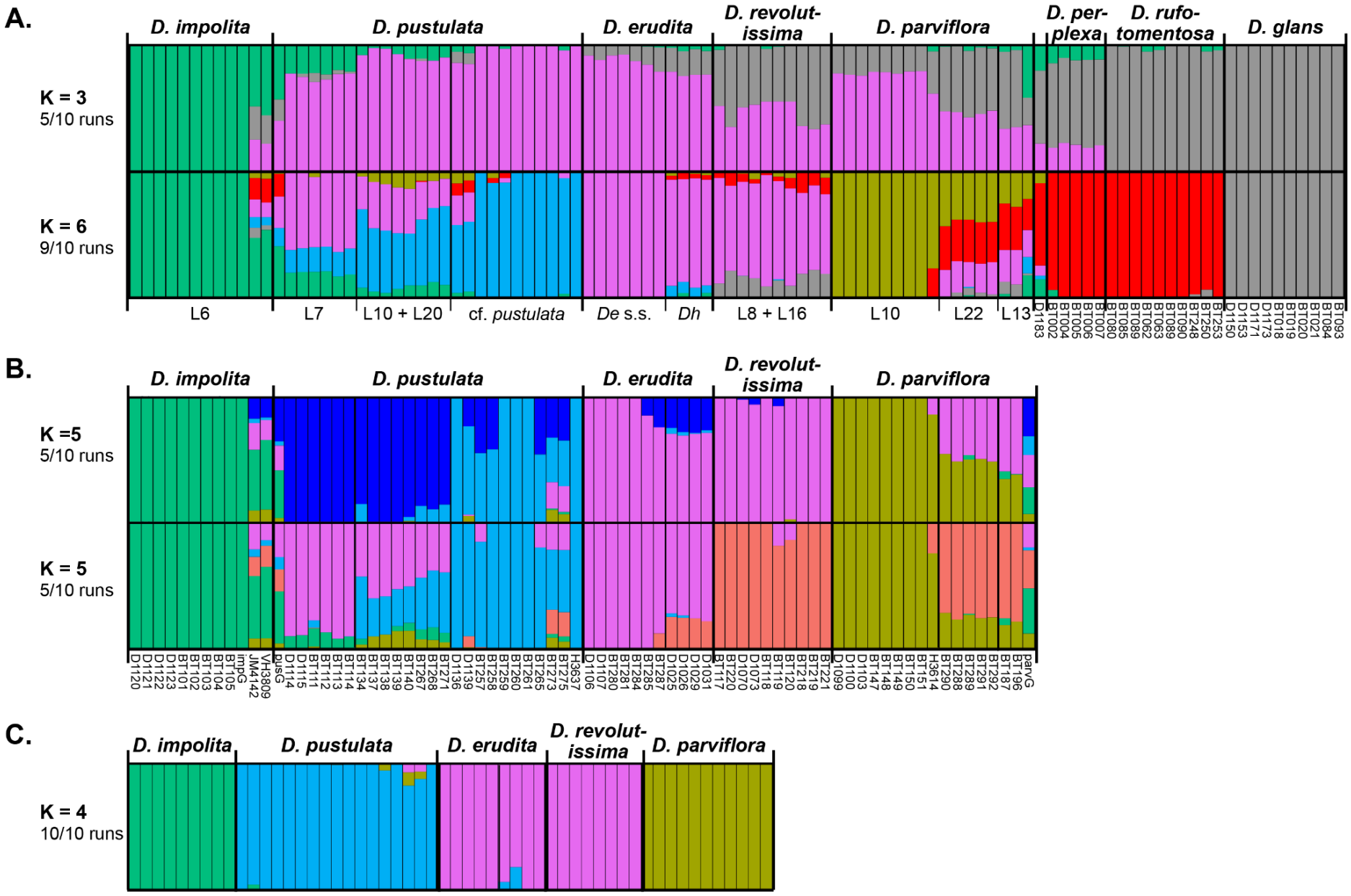

A.
D. impolita
D. pustulata
D. erudita
D. revolut-issima
D. parviflora
D. per-plexa
D. rufo-tomentosa
D. glans
K = 3
5/10 runs
K = 6
9/10 runs
L6
L7
L10 + L20
cf. pustulata
De s.s.
Dh
L8 + L16
L10
L22
L13
B.
D. impolita
D. pustulata
D. erudita
D. revolut-issima
D. parviflora
K =5
5/10 runs
K = 5
5/10 runs
C.
D. impolita
D. pustulata
D. erudita
D. revolut-issima
D. parviflora
K = 4
10/10 runs

Figure 4: Admixture plots for different sets of samples generated with NGSadmix and visualized with pong (Behr et al., 2016). Shown are majority modes for best K (Supplemental 11) determined using the Evanno method (Evanno et al., 2005). The admixture plots are given for A) all individuals initially included in the sampling (n=102); B) without outgroup accessions, and C) demographic inference individuals only. Labels above the plots show species, whereas labels below plots indicate either localities or accession numbers. Outgroup accession D1183 is a representative of D. pancheri. De s.s. stands for D. erudita s.s., Dh indicates D. hequetiae.

Principal component analysis (PCA; Figures 5 and 6) of SNPs obtained from the same group of individuals showed clustering by species, with D. rufotomentosa and *D. perplexa* forming distinct groups from *D. parviflora*. There is no separate clustering for *D. erudita* and *D. hequetiae* in the first two principal components, but a slight separation in PC three and four. When overlaying admixture proportions onto the generated PCA (Figure 5) described admixture patterns can be observed also in relative distance and clusterings, where distinct clusters towards the margins of the plot show very little admixture, while highly admixed individuals tend to take central positions in the PCA.

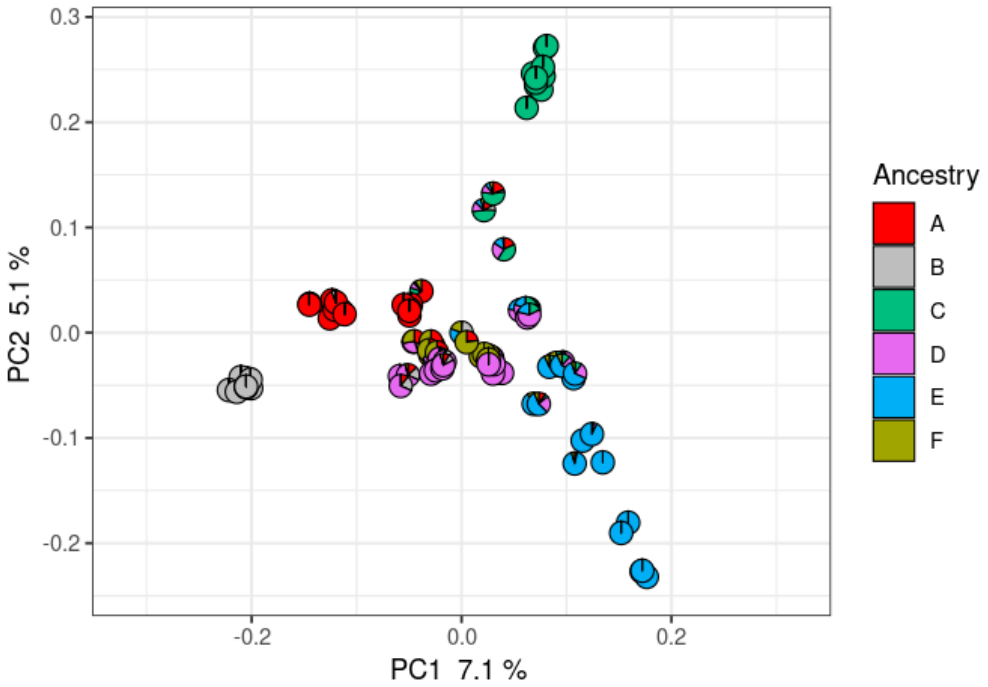


Figure 5: Plot for PC1 and PC2 of a principal component analysis of all 102 individuals investigated. Individual ancestry proportions estimated from NGSadmix under K=6 are overlaid onto the plot. Clusterings in PC1 and PC2 align closely with admixture results.

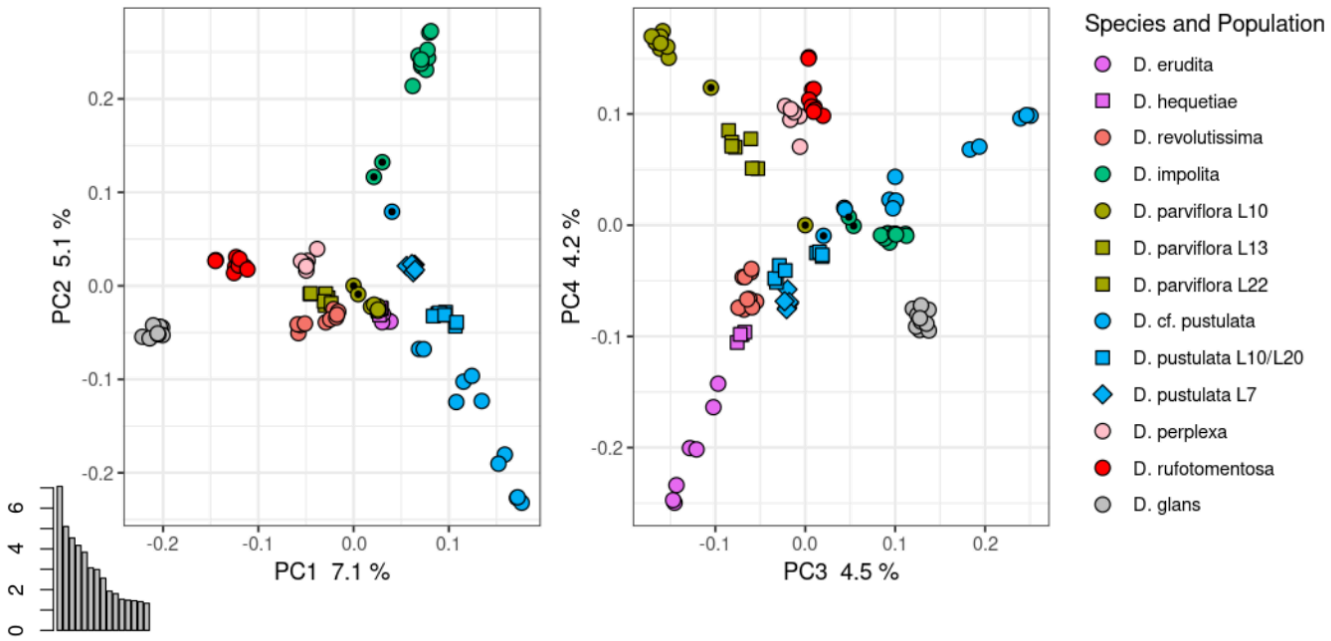


Figure 6: Scatterplots for PCA of all individuals, including outgroup (PC1 and PC2 left, PC3 and PC4 right) and scree plot of explained variance for the first 15 principal components on the bottom left (Y-axis indicates the % variance explained, with PC vectors on the X axis). Colors represent species identified from previous genetic analyses and shapes indicate distinct populations where applicable. Accessions marked by a black dot have been sampled as single individuals from a particular location and in addition could not be clearly assigned to a larger population based on genetic identification. Individuals generally cluster well according to species and population in both plots.

After outgroup samples were removed the NGSadmix results showed more consistent ancestry proportions across individuals of a species (Figure 4B). Two possible modes for best K=5 represent five out of ten runs separately. Again, *D. impolita* exhibited very clearly defined ancestry. Furthermore, *D. erudita* and *D. hequetiae* shared most of their ancestry in both modes, although *D. hequetiae* clearly shared some genes with *D. pustulata* in one alternative scenario, whereas with *D. revolutissima* in the other. Both species *D. parviflora* and *D. pustulata* are comprised of individuals with varying ancestry, where *D. parviflora* splits roughly into two groups with one being defined by admixture from ancestral *D. revolutissima* or *D. erudita* depending on the mode. In *D. pustulata* the two modes show either two distinct ancestries mostly unique to the species or one group with significant admixture of *D. erudita* ancestry. Additionally, some individuals indicated high admixture proportions. PCA for this grouping showed separation of *D. impolita* and

D. pustulata along PC one with all remaining species separating from these only along PC two, clustering together, but separating clearly in PC three and four (Figure 7).

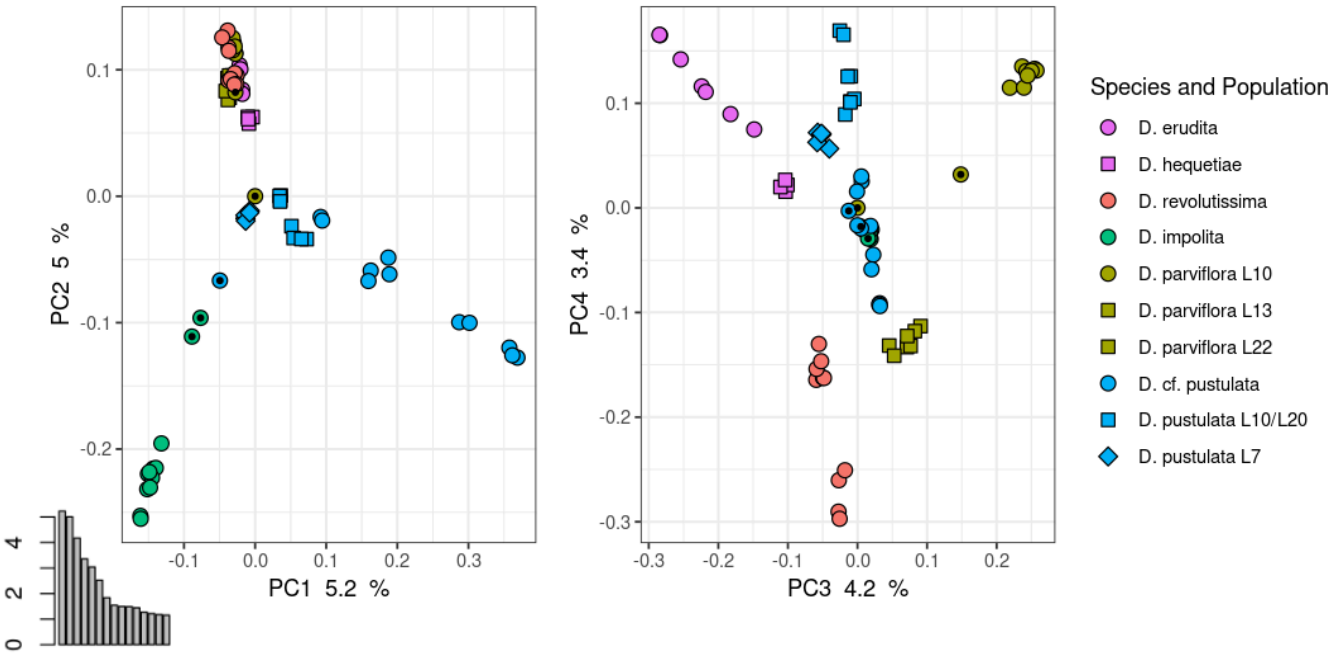


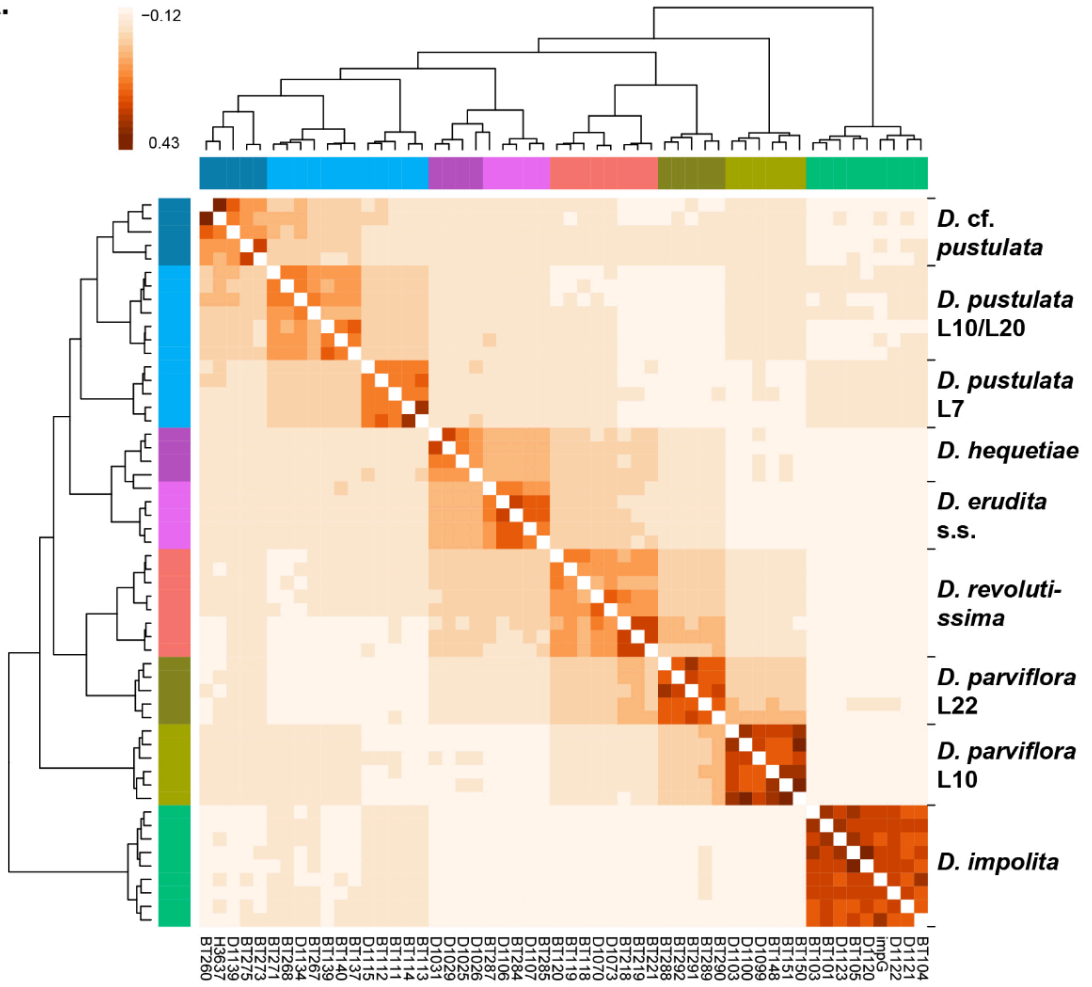
Figure 7: Scatterplots for PCA after removal of outgroups (PC1 and PC2 left, PC3 and PC4 right) and scree plot of explained variance for the first 15 principal components on the bottom left. Species and populations are indicated by colors and shapes respectively. Accessions not clearly belonging to one population are again marked by black dots. Clustering of individuals by species is only visible for all when considering the first four principal components.

When NGSadmix was run only on samples included in demographic inference, i.e. with any putatively recent hybrids and closely related individuals removed, results showed very clear ancestry proportions for all species (Figure 4C). However, all of *D. erudita*, D. *hequetiae* and D. *revolutissima* showed no difference in ancestry and best K was determined to be four, while representing six currently recognized species (Supplemental 11). In this representation, samples also showed almost no admixture. Similarly, principal component analysis showed clearly defined clusterings by species and population in the first two principal components (results not shown). A clustering heatmap of covariance (Figure 8A) for the samples shows the same patterns with the exception of some individuals of *D. erudita* and *D. hequetiae*.

Individuals cluster based on their species or population belonging and covariance between individuals of the same group is significantly higher than covariance between individuals of other species.

Fst (Figure 8B), estimated from the SFS of subsetted populations, supports findings from NGSadmix and PCA analyses. Species which show close clustering in PCA and share ancestry in admixture plots exhibit a lower weighted Fst value, meaning a lower level of genetic differentiation. The observed pattern closely mirrors the pattern visible in the covariance matrix, showing low covariance between individuals of species with high Fst and reverse. *D. impolita* shows the highest Fst-values across most species or populations, indicating it to be most differentiated from other species of the clade. Conversely, *D. revolutissima* exhibits the lowest level of differentiation from other species of the clade, with very low Fst especially when compared to *D. erudita* and *D. hequetiae*. This is counter to previously published phylogenies (Samuel et al., 2019) where the species was proposed to be sister to the rest of the clade. The lowest estimated Fst value (0.14) is present between *D. erudita* and *D. hequetiae*, recapitulating previous results based on PCA and NGSadmix, which showed these to be clustering very closely. Levels of genetic differentiation in this species pair is lower than even calculated values between populations of individuals currently included within *D. pustulata* or *D. parviflora* respectively.

A.


0.12
0.43
D. cf. pustulata
D. pustulata L10/L20
D. pustulata L7
D. hequetiae
D. erudita s.s.
D. revoluti-ssima
D. parviflora L22
D. parviflora L10
D. impolita

B.

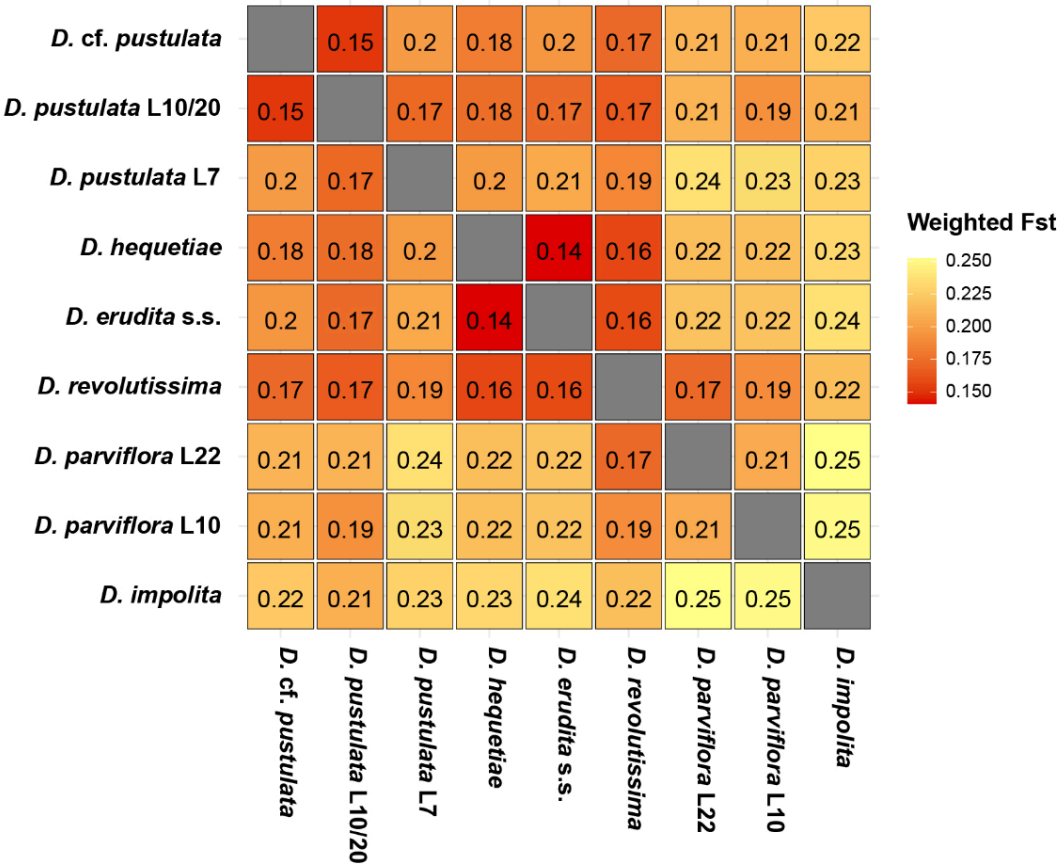

D. cf. pustulata 0.15 0.2 0.18 0.2 0.17 0.21 0.21 0.22
D. pustulata L10/20 0.15 0.17 0.18 0.17 0.17 0.21 0.19 0.21
D. pustulata L7 0.2 0.17 0.2 0.21 0.19 0.24 0.23 0.23
D. hequetiae 0.18 0.18 0.2 0.14 0.16 0.22 0.22 0.23
D. erudita s.s. 0.2 0.17 0.21 0.14 0.16 0.22 0.22 0.24
D. revolutissima 0.17 0.17 0.19 0.16 0.16 0.17 0.19 0.22
D. parviflora L22 0.21 0.21 0.24 0.22 0.22 0.17 0.21 0.25
D. parviflora L10 0.21 0.19 0.23 0.22 0.22 0.19 0.21 0.25
D. impolita 0.22 0.21 0.23 0.23 0.24 0.22 0.25 0.25
Weighted Fst
0.250
0.225
0.200
0.175
0.150

Figure 8: (A) Clustering heatmap of covariances from principal component analysis of only individuals used for demographic inference. Side color bars indicate species or population level assignment. Results for individuals of D. parviflora or D. pustulata indicate clear population structure within the species. (B) Heatmap of weighted Fst calculated from folded SFS. The plot order follows the heatmap A. The patterns of covariance between individuals of different populations given in A are reflected also in B.

Hybrid individuals and population substructure

Investigating tentative species more closely it is possible to identify population substructure within one species. Both *D. pustulata* and *D. parviflora* split into multiple groups. In the case of *D. parviflora* the split occurs along sampling locations, associated with different ecologies and soil types. A relatively major split occurs between *D. parviflora* from localities L22 and L13, with L13 representing the type locality for the species, and *D. parviflora* L10 (Figure 9). Individuals falling into *D. parviflora* were sampled from Mandjelia (L13) or Plateau de Tango (L22), both at significantly higher elevations than L10, occurring in mountain forests on ultramafic or schistous soil. The population *D. parviflora* L10 is comprised of individuals sampled at the Conservàtoire botanique de Tiéa (L10), at lower elevation in a dry forest next to a river, growing on black clays. Interestingly, based on admixture analyses, *D. parviflora* L13 and L22 showed mixed ancestry compared to *D. parviflora* L10 in most modes. Analysis of genetic differentiation between *D. parviflora* L10 and L22, using Fst (Figure 8B) estimated from the joint SFS, showed relatively high differentiation (weighted Fst = 0.21) compared to other pairings within the clade.

Similarly, there are multiple distinct populations of *D. pustulata* in the selected sampling (Figure 9). Here two groups split based on their sampling locations, with a third group occurring sympatrically with the members of other groups, but clustering in a separate group in PCA. *D. pustulata* sampled at Pindai (L7) forms one population and samples taken at Conservàtoire botanique de Tiéa (L10) as well as some samples from the Moindah (L20) fall roughly into one population. A third population consists of samples taken from both L20 and L7 and includes many individuals previously misidentified by non-molecular approaches. These results led

to a reassessment of herbarium samples for those specimens by experienced taxonomists, very familiar with New Caledonian *Diospyros*. Indeed, the respective herbarium specimens showed atypical morphology for *D. pustulata* and exhibited larger leaves, with a phenotype described as intermediate between *D. pustulata*, *D. erudita* and *D. yahouensis* and hereafter named *D.* cf. *pustulata*. The other two populations were given an identifier based on their sampling locations.

While all *D. pustulata* populations were sampled in dry forest habitats they differed in their soil. Samples taken from Pindai (L7) split into respective populations along soil types, with *D. pustulata* L7 growing on calcareous and *D.* cf. *pustulata* on sedimentary soil. Other individuals of *D.* cf. *pustulata* from location L20 were also found on sedimentary soil. Only the population *D. pustulata* L10/20 occurred on two different soil types (ultramafic and sedimentary). Looking at genetic differentiation based on calculated Fst values, *D. pustulata* from L10/20 is significantly closer to *D.* cf. *pustulata* than *D. pustulata* from L7 which shows more genetic differentiation from population L10/20. Geographically however, location L7 is situated between both other regions with sampled individuals of *D. pustulata* (L10 and L20) and genetic differentiation therefore does not strictly follow geography as observed in *D. parviflora*.

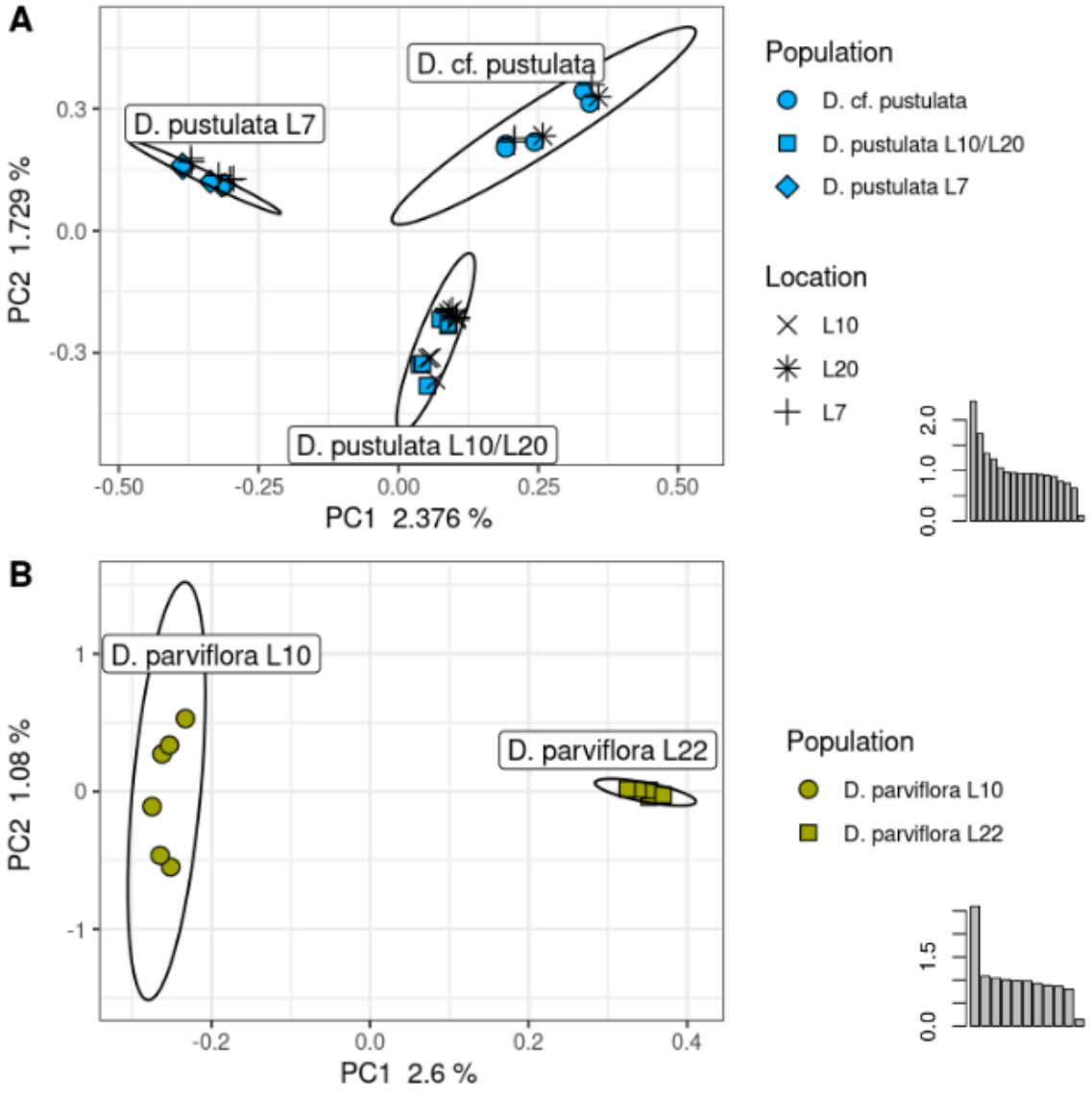


Figure 9: Scatterplots for PCA of D. pustulata (A) and D. parviflora (B) separately. D. pustulata individuals chosen for demographic inference show distinct clusterings, but no clear relationship of clusters and sampling location. In D. parviflora clusters form with respect to sampling location.

In turn, *D. erudita* s.s. closely clusters with *D. hequetiae*, with some overlap of the two. Results from NGSadmix and PCA suggest a high degree of similarity and Fst shows the lowest genetic differentiation in all pairings (Figures 4, 6 and 8). Genetic differentiation is less than between populations within species like *D. pustulata* and *D. parviflora*. Due to the very close relation of *D. erudita* s.s. and *D. hequetiae*, these were treated as one single population in demographic analyses. Some individual

samples initially assigned as *D. erudita* were shown to be closely related or part of *D. pustulata* (*D.* cf. *pustulata*) . This relationship was corroborated by morphological examination, suggesting at least some crossbreeding of *D. erudita* and *D. pustulata* in some locations.

The most compelling evidence for relatively common hybridisation between two species can be found between *D. impolita* and *D. pustulata*. While the later species shows considerable sub-structure, *D. impolita* individuals sampled showed remarkably little variation within the species and very tight clustering in PCA. Additionally, *D. impolita* showed the lowest average heterozygosity across all samples of all species (results not shown). Exceptions to this pattern are individuals JM4142 and VH3809 which had the lowest coverage (around 8x) but showed the highest heterozygosity within the species. Furthermore, these individual samples showed mixed ancestry in NGSadmix leading to the assumption that they represent relatively recent hybrids. A PCA (Figure 10) of just *D. impolita* and *D. pustulata* individuals was also able to show them clustering between *D. pustulata* and *D. impolita*, close to *D. pustulata* individual pustG, which also showed mixed ancestry in NGSadmix. However, pustG did not display higher heterozygosity than other *D. pustulata*, but showed the lowest heterozygosity within the species. Therefore, it is likely that this individual is a late generation hybrid.

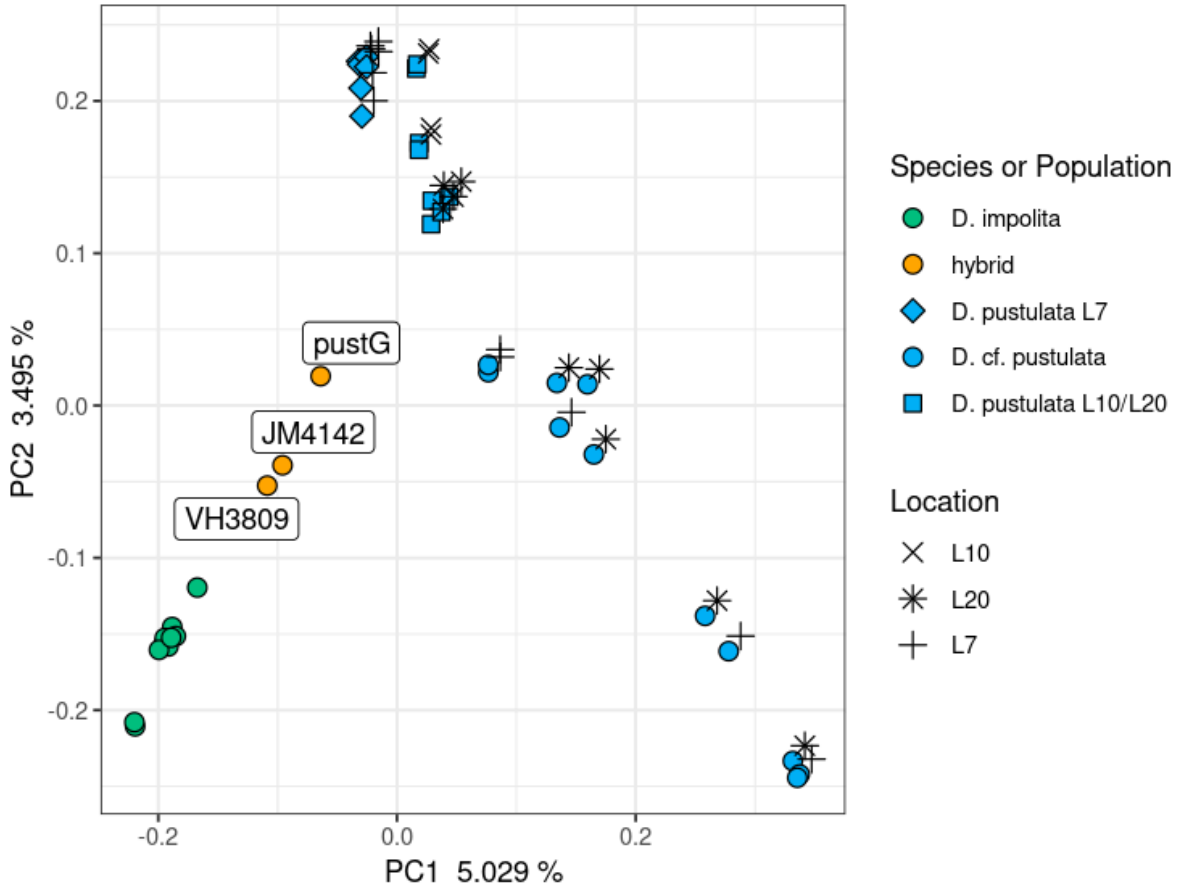


Figure 10: Principal components 1 and 2 for a PCA of only D. pustulata and D. impolita individuals. For D. pustulata sampling locations are indicated next to the respective individual. Suspected hybrid individuals identified in previous results clearly fall between clusterings of D. impolita and D. pustulata. Highlighted individuals are D. impolita JM4142 and VH3809, as well as D. pustulata pustG.

Heterozygosity within the whole clade ranged from 0.7% to 1.59% calculated as the proportion of heterozygous positions (singletons in single individual SFS) versus total sites, with the average being 1.15%. This is lower than estimates from alignment-free Kmer methods for select individuals (Ghane, unpublished), but still represents a considerable amount of variation, which is to be expected given that New Caledonian *Diospyros* is dioecious. The difference in the methods can be explained by the removal of TE-sites in the reference based approach presented here. There were no significant differences in the average heterozygosity between species, with the exception of *D. impolita* exhibiting a lower average heterozygosity (after suspected hybrids were removed).

While heterozygosity can be seen as a signal of genetic diversity, relatedness was nonetheless high. Twelve individuals were removed for showing a KING kinship coefficient of 0.177 or higher. This threshold signifies first degree relationship (i.e. siblings or parents) or duplicate samplings, whereas clonal reproduction is not known for this group. Five unique pairings were estimated to be duplicate samplings, with three of these accounting for H3637, BT261 and D1136 and their respective unique permutations (three unique pairs). It is interesting to note that even here, when sampling the same individual, initial misidentification to different species can occur in successive samplings, especially as the accessions were collected by different researchers, with different levels of taxonomic expertise. In total twelve pairings of individuals with a relatedness of at least first degree were initially identified as different species (eight pairings identified the same species). These twelve pairings were exclusively either part of *D.* cf. *pustulata* or *D. parviflora* L10 populations. For all pairings, both within and between species, kinship, expressed as the KING coefficient, showed a negative correlation with geographic distance of sampling locations (Supplemental 2).

Demographic inference

Fluctuations in effective population size

When estimating effective population size back in time (Figures 11 and 12), patterns have been found that are closely mirrored between populations and the methods used. For all but *D. revolutissima* there is a decrease in effective population size in more recent times, at around 100 thousand years (considering a generation time of 7 years) back in time in stairwayplot 2 results. Additionally, all populations show a period of very high effective population size starting at roughly 100 thousand years back in time, with some differences in timing. This period of very large and stable effective population sizes then lasts from about 100 thousand back to four to six million years ago, when population size decreases again. In some instances this decrease is part of a bottleneck of varying severity. In *D. erudita* and *D. revolutissima* recent population size decline is less severe than in other populations. These populations exhibit a more severe bottleneck at around 6 million years ago.

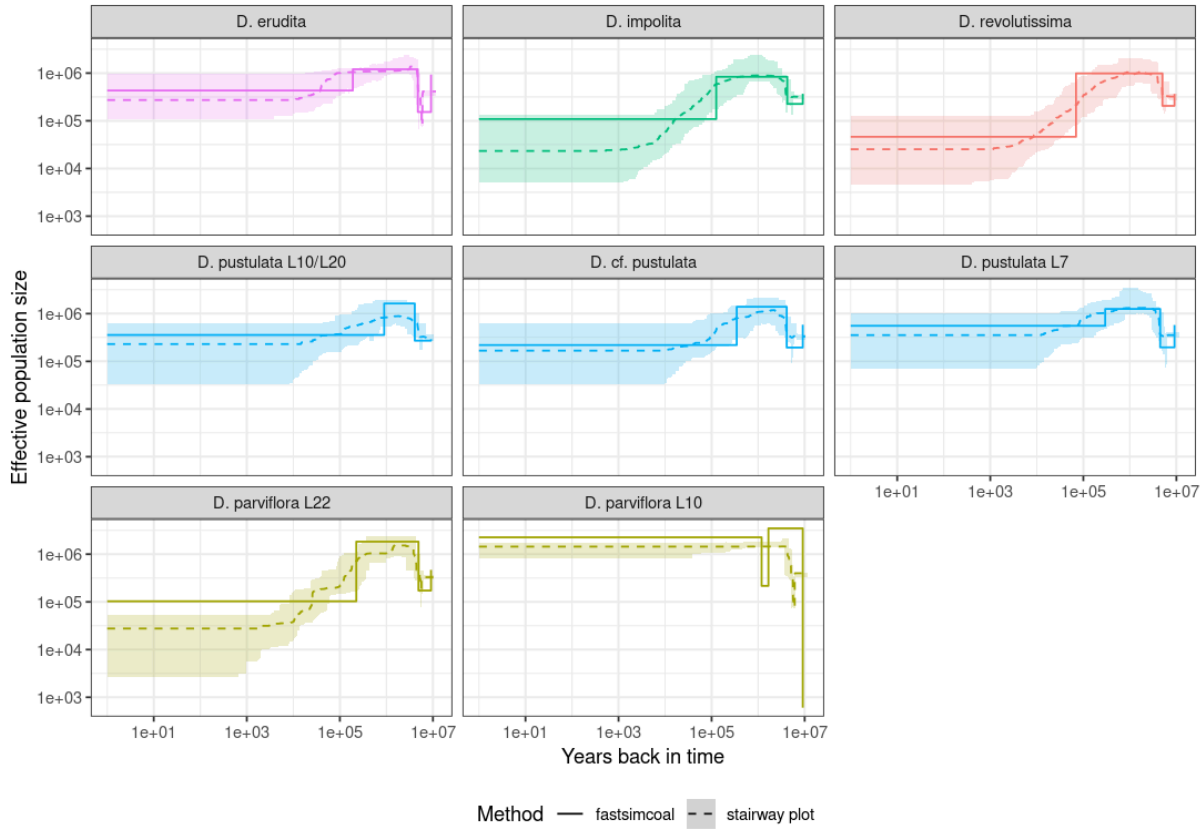


Figure 11: Effective population size back in time modeled with both fastsimcoal 2 and stairway plot 2 using the unfolded SFS. Ribbons in the stairway plot 2 estimates indicate 2.5 and 97.5 % confidence intervals as given by the program. All populations show highly congruent results for both tools.

Fastsimcoal single population models with a total of four population sizes estimated at current, ancestral and two flexible, intermediate timepoints, were able to recapitulate stairway plot results very closely. In all populations the same general patterns of population size fluctuation can be observed as estimated with stairway plot. Especially, more recent changes in demography and population size are consistent over both methods. Size estimates from fastsimcoal generally are more extreme, meaning in periods of large population sizes, Ne is overestimated compared to stairway plot and bottleneck severity is exaggerated. Fastsimcoal estimates usually are within 2.5% and 97.5% confidence intervals of stairway plot estimates.

Both methods estimate effective population sizes generally a lot higher than initially expected. As these methods estimate demographic parameters from the SFS and ancestral states were estimated with methods described above, it was tested whether the use of folded SFS would change this, by adjusting the proportion of rare to common alleles. When comparing results from unfolded and folded input SFS there was no major difference in population sizes observable (Figure 12). Results from stairway plot were very similar for all three populations tested using both SFS generation methods. Similarly, fastsimcoal results only differed slightly depending on whether the folded or unfolded SFS was used. The largest difference occurred in the timing of a bottleneck modeled in *D. revolutissima*, with the unfolded SFS resulting in the bottleneck occurring significantly earlier.

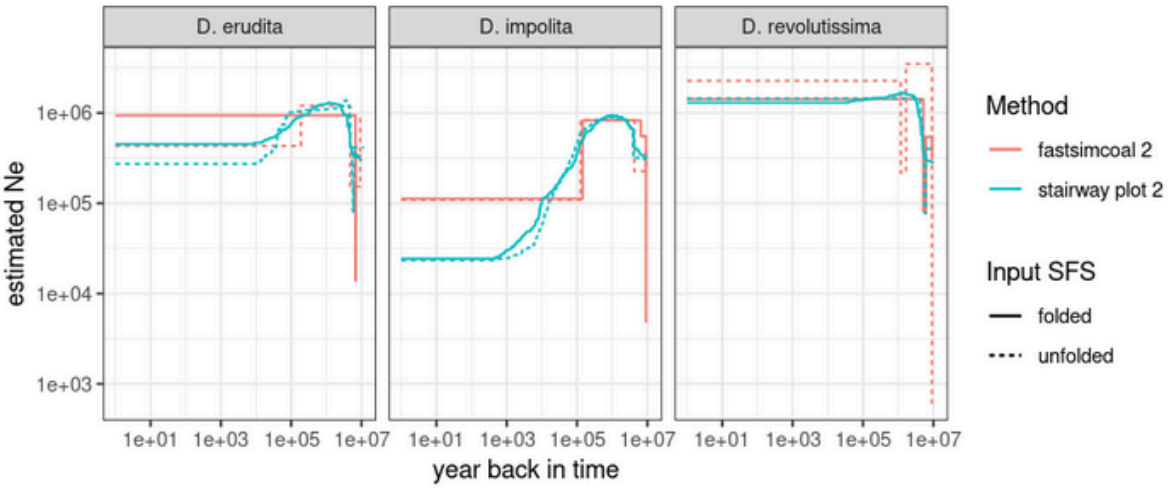


Figure 12: Single population size estimates for effective population size of three species back in time. Color indicates method used, with red for the described fastsimcoal 2 model and blue for stairway plot 2. Whether estimates were generated using a folded or unfolded SFS is indicated by the linetype.

Speciation modeling for selected pairs

Two pairs of three species were chosen to model speciation under isolation and different migration models. *D. revolutissima*, *D. erudita* and *D. impolita* were chosen, both for biological and technical reasons. Previous results showed a close relation of *D. revolutissima* and *D. erudita* with low genetic differentiation, contrary to previously

published phylogenies. Samplings for these populations were also most numerous, which allowed the usage and comparison of both folded and unfolded SFS.

Results showed the best model fit according to likelihood distributions to be a simple divergence model with constant migration (i.e., isolation with migration) for both species pairs, and for folded and unfolded input SFS (Supplemental 6). Introducing a size change event into the constant migration model improved fit significantly. Allowing for a single size change after divergence was able to account for population size fluctuations observable in single population models and stairway plots and estimated population sizes were in line with previous results. The model fit was generally better for folded input SFS, probably owing to the reduced amount of observations.

Under all models, estimated current population sizes for *D. revolutissima* were inferred to be extremely high with a size increase approximately 450 to 700 thousand years ago. *D. erudita* showed estimates for effective population size considerably lower and more stable. In *D. impolita* population size estimates were lowest, but showed a stark reduction in current effective population size.

Estimated parameters for both folded and unfolded SFS differed slightly (Supplemental 5). Time of divergence was estimated to be more recent when using the unfolded SFS as input. The relative difference between the two population pairs in time of divergence stayed similar. With both inputs, time of divergence for *D. erudita* and *D. revolutissima* was estimated to be lower, i.e. the species split more recently. Divergence times for *D. revolutissima* and *D. impolita* were estimated at ~5.5 Mya (unfolded) or ~6 Mya (folded), whereas at 4.5 Mya (unfolded) or 5.3 Mya (folded) for *D. revolutissima* and *D. erudita*, with migration rates being higher in the more recently diverged pair. These estimates are older than previously inferred ages based on RADseq and secondary calibration points in phylogenetic methods (Paun et al., 2016).

Possible hybrid population in *D. parviflora*

After results from both NGSadmix and PCAngsd suggested *D. parviflora* L22 and L13 to be composed of ancestral alleles of both *D. revolutissima* and population *D. parviflora* L10, different speciation models were tested for this group. Possible scenarios included different one and two origin configurations, as well as a hybridization model. All models also allow for migration between all extant lineages simulated back in time.

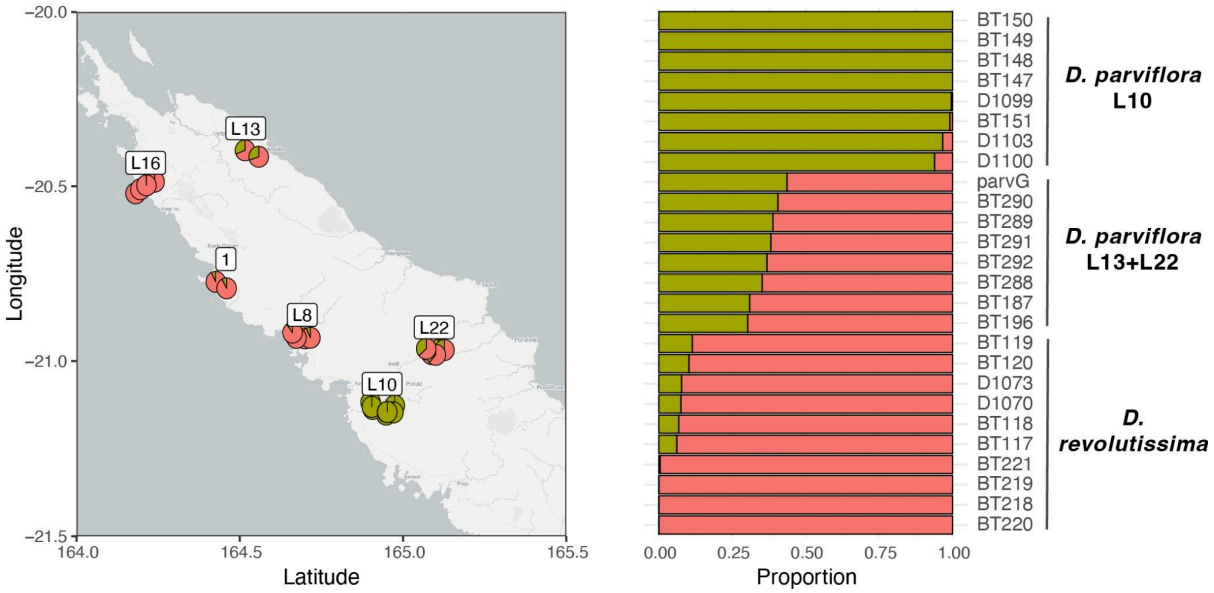


Figure 13: Map of sampling locations of the tested individuals (left). Individuals taken at locations L16, L6 and "1" are samplings of D. revolutissima, while L10 and L22 + L13 represent D. parviflora L10 and D. parviflora L10/L13 respectively. L13 also represents the type locality of D. parviflora. Ancestry proportions as calculated with PCAngsd are shown (left) as pie charts for each individual and (right) as bars on the left. There is a clear gradient of ancestry proportions visible in geography and on a population level.

Models based on two separate origins for *D. parviflora* populations generated a worse fit to the observed frequency spectra, than the other models, with the scenario describing a split of *D. parviflora* L13/22 from *D. revolutissima*, and a subsequent split of *D. parviflora* L10 from *D. revolutissima* performing worse than the models including a single origin for *D. parviflora*. This result was expected and confirmed previous results regarding tree topology. Scenarios with a single origin for all *D. parviflora* populations or a hybrid origin for *D. parviflora* L22 fit the underlying data

better. When likelihood distributions for these models were generated, there was significant overlap in both models, suggesting them to be equally capable of explaining observed patterns. As not all models possessed the same amount of parameters AIC was also used for model selection and suggested a hybrid origin of *D. parviflora* L22 as the best fitting model (Supplemental 8).

Best parameter estimates for hybrid and single origin models showed very similar timings for both the *D. revolutissima* - *D. parviflora* split and for the hybridization or divergence event giving rise to *D. parviflora* L22. The ancestral split was estimated at ~660 thousand generations (4.6 mya), with the second split at ~140 thousand (980 kya). Equally, estimates of effective population size were consistent over both models, and were in line with results from single population modeling (see Figure 14 for hybridisation model). Migration rates between both models also stayed consistent with the exception of migration between *D. revolutissima* and *D. parviflora* L10 before the divergence of *D. parviflora* populations in the single origin model, which was significantly lower than in the hybridization model (Supplemental 7). Taking into account the significant admixture event estimated at the time of hybridization this puts gene flow between *D. revolutissima* and *D. parviflora* L22 at much higher levels as gene flow between *D. revolutissima* and *D. parviflora* L10 or the ancestral *D. parviflora* populations.

Based on these results it can be concluded that *D. parviflora* consists of multiple populations which differ in their genetic makeup mostly due to different proportions of admixture/introgression from *D. revolutissima*. Further, a hybrid origin could be modeled to be at least as likely an explanation under the observed data as a single origin model.

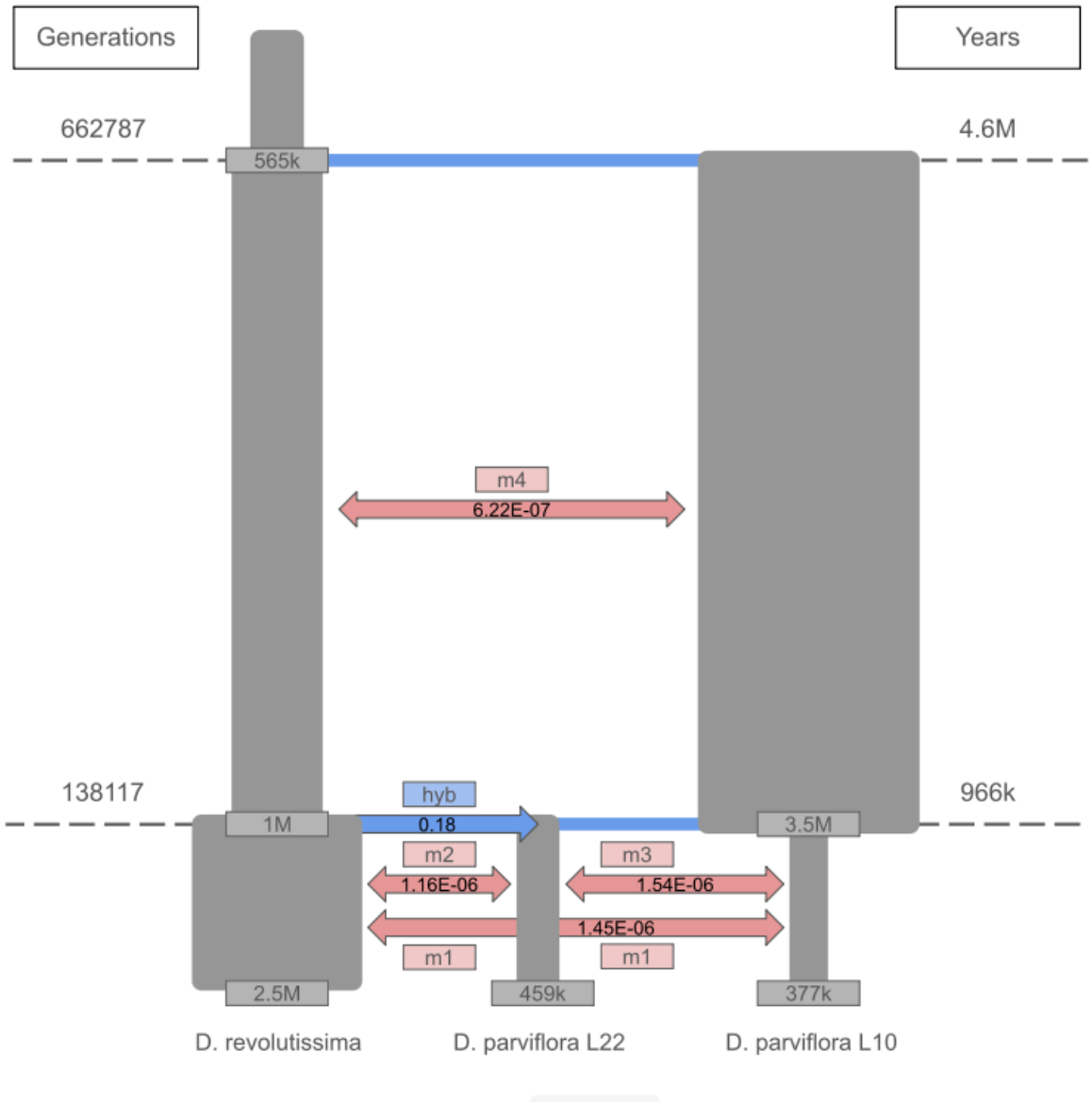


Figure 14: Visualization of the best fitting hybrid model with estimated parameters. Estimates of population size are in grey, migration in red. Single admixture or divergence events are in blue. Estimates of admixture pulse at hybridization and migration are relative to the respective population sizes at the time of the event. Timings are indicated at the right and left side in years and generations (generation time 7 years).

Discussion

The analyses conducted in this MSc thesis were able to distinguish species or genetic populations of individual samples with more confidence than taxonomic identification by morphology. The presence of very similar or cryptic species (Shin & Allmon, 2023) in the case of *D.* parviflora and *D. rufotomentosa*, and to some extent *D. perplexa* greatly complicates species identification without the use of molecular methods. The admixture analysis and PCA results confirmed that individuals sampled in the south of the island correspond to *D. rufotomentosa*, and therefore are more closely related to *D. glans*. These results further support the introduction of the complex *D. rufotomentosa* by (Schatz & Porter, 2018). Species identification is further complicated by individuals which represent probable hybrids with morphologies presenting as intermediates of two or more species as present in *D. pustulata, D. erudita* and *D. impolita*.

Results from admixture analyses and genomic PCA showed very heterogeneous genetic makeup especially for the species *D. parviflora* and *D. pustulata*. These species could be split into multiple populations with, in the case of *D. pustulata*, populations occurring sympatrically and showing only minute differences in ecological adaptation, if any. Evidence points to hybridisation with other species as one possible mechanism for the formation of these populations (Cerca et al., 2023; Meier et al., 2017; Seehausen, 2004). Genetic differentiation estimated from pairwise Fst shows less genetic differentiation between populations across certain species boundaries than between populations of these species.

Hybridisation is also a possible explanation of very high levels of heterozygosity estimated in the samples, as possible cross breeding with other species, within the radiation, could introduce a large amount of genetic variation (Edelman & Mallet, 2021; Suarez-Gonzalez et al., 2018). Importantly, as dioecious plants, *Diosypros* accessions are expected to show significantly higher heterozygosity (Balloux, 2004; Muyle et al., 2021) than most hermaphroditic plant systems. The reported heterozygosity values were also cross referenced for some samples with

GenomeScope (Ranallo-Benavidez et al., 2020) derived heterozygosity estimates (Ghane, unpublished). These estimates were even higher than estimates from this thesis, most likely owing to analyzing the entire genome including transposable element regions in GenomeScope. High estimates for heterozygosity could also, in part, stem from paralog regions collapsed in the reference or incorrect mapping due to removal of some contigs from the draft assembly. New Caledonian radiated *Diospyros* has a very high genome dynamics in its past, which led to almost a two fold increase in genome size compared to non-radiated relatives (Samuel et al., 2019), which suggests a high prevalence of diverse gene duplication in the study group.

Overestimations of heterozygosity would also influence other measures, many of which have been derived from the SFS. Fst, which can generally be explained as a proportion of diversity within a population that is not shared with others, is heavily influenced by heterozygosity and was calculated from folded SFS. Similarly, excess variants would influence the SFS directly, leading to larger effective population size estimates, when Ne is estimated from current genetic variation (Wang, 2005). As all demographic inference in this thesis is done on the basis of the computed SFS, parameters estimated would deviate widely from real values in case SFS would not be accurately estimated. With the probability of coalescence being 1/(2*Ne) and the average time of coalescence of two randomly chosen genes T=2Ne (Nordborg & Krone, 2002; Wakeley & Sargsyan, 2009), large estimates for Ne, which in turn are a result of high genetic diversity, lead to very low probabilities of coalescence. This would lead to the overestimation of divergence times and related parameters. While these factors make the concrete estimation of parameters like Ne and time of divergence very challenging, relative estimates should still be valid, under the assumption that excess variants present in the SFS are distributed equally across its cells and different species.

Estimates for effective population size with both stairway plot and fastsimcoal show very large and stable effective population sizes for at least some time in the last ~6M years. Many populations also exhibit a strong decline in recent effective population size starting at around ~100 thousand years ago. It is unclear what is causing the

recent decline in effective population size. All populations also show their highest effective population size from approximately one to five million years ago. This time frame coincides with time of speciation estimated in other studies (Paun et al., 2016; Turner et al., 2013). However, previously discussed considerations about the accuracy of these estimates have to be taken into account. Most populations also show significant growth in the period before and lower Ne from six million years ago and back. This is expressed as a bottleneck in some populations. When comparing between results generated with the unfolded versus the folded SFS for three populations, the bottleneck visible in two of the populations is less pronounced when using the folded SFS. While there is still a large reduction in Ne back in time, there is no pronounced “rebound”. This leads to the conclusion that the observed bottleneck in the results based on the unfolded SFS might be a technical artifact from the inference of the ancestral state. The slightly differing timings of changes in effective population size, especially the possible bottleneck, suggest that an estimate of 7 years generation time across all species and growth habit types, for example across shrubs and trees, is not accurate, with some slight deviations from this.

While absolute parameter estimates from multi population models are not thought to be entirely accurate, it was possible to observe relative differences in the divergence time estimates for two species pairs, in line with a priori expectations. Divergence times for *D. revolutissima* and *D. erudita* were more recent (12/20% lower) than for *D. impolita* and *D. revolutissima,* and migration rates were higher in the more recently diverged pair. This was in line with admixture, PCA and pairwise Fst estimates from this thesis, but did not reflect results from previous studies (Paun et al., 2016; Samuel et al., 2019) where *D. revolutissima* is placed as either the sister or within a sister clade to all other investigated species. A simulation of a three population model showed the validity of speciation through hybridisation for a population of *D. parviflora*. Still, a single origin model for the two populations of *D. parviflora* performed almost as well and showed very similar parameters.

Demographic inference with model based approaches like fastsimcoal always poses inherent limitations as overfitting (Lapierre et al., 2017; Tredennick et al., 2021). This is complicated in this case, as previous knowledge of the study system is sparse.

Optimizations could be possible with the use of a more accurate mutation rate and specific generation time for New Caledonian *Diospyros*. Furthermore, more accurate phylogenies would allow for the refinement or construction of more complex models. A more complete reference would ideally also be useful for removing some of the possibly excess heterozygosity, leading to more accurate parameter estimates through the reduction of superfluous variants. Due to the complexity of the system, with over twenty species in total and probable hybridization across even subclades of the radiation, very complex models would be needed. Excluding some species from models, even when species are phylogenetically far, could lead to the loss of crucial information, as in the case of *D. rufotomentosa* and D*. parviflora*, which show some signs of significant shared ancestry or hybrid speciation.

Considering this it would be useful to include other methods less reliant on previous knowledge of the radiation. Rigorous species delimitation with methods like SNAPP (Bryant et al., 2012) should be done, to allow for more clarity in some cases. Unfortunately, species delimitation can be limited within adaptive radiations (Bamberger et al., 2022; Parker et al., 2022). Species trees, but also phylogenies not following a classic tree structure generated by tools like phylonet could be beneficial. For the detection of hybridization excess allele sharing in ABBA-BABA and D-statistics or treemix could be used.

Additional sampling, including a more continuous sampling across the distribution range of each species, would also allow to address multiple issues and potentially elucidate some phenomena within the clade and larger radiation. As previously discussed, sampling is heavily complicated by population substructure and the high number of related individuals at one location, due to dispersal limitations (Carpenter et al., 2003; Heilbuth et al., 2001; Tassin et al., 2010). Ideally, species would be sampled only from one location, with at least ten individuals per population, which are not siblings or closely related. This would lessen the impact of geography and therefore population substructure in studied populations. More samples per population would also allow for the exclusive use of folded site frequency spectra, which are suspected to be less prone to technical biases compared to the inferred ancestral state. Many methods also are sensitive to unbalanced samplings, where

individuals of one group greatly outnumber other groups (Meirmans, 2019; Puechmaille, 2016). Lastly, some geographical regions were identified in this thesis which seem to exhibit a large amount of potential hybrids. The most interesting of these is location L20 (Metzdorff property, Moindah, Poya South), which could be a potential hybrid zone of two or more species and needs to be studied further.

Conclusion

Even though whole genome resequencing allowed for a huge increase of available data, compared to previous plastid and RAD-seq approaches, the complexity of the New Caledonian *Diospyros* adaptive radiation makes the generation of new insights from these dataset challenging. Nevertheless, it was possible, with the aid of bioinformatic tools and the re-examination of select individuals, to elucidate some previously unclear relationships within the radiation. Firstly, this thesis was able to confirm the existence of two different species *D. parviflora* and *D. rufotomentosa* (Schatz & Lowry, 2018), which show characteristics of cryptic species within the radiation. In the original circumscription (White, 1993; White, 1992) these had been included in the complex D. parviflora. Previous molecular studies (Paun et al., 2016; Turner et al., 2013) had already shown separating clades and the generated results support these findings, as *D. rufotomentosa* presents as a clear outgroup to the studied clade. Secondly, results suggest a closer relationship of *D. revolutissima*, which was previously thought to be sister species to the rest, to the investigated clade with especially *D. erudita* and *D. parviflora*. Species *D. parviflora* and *D. pustulata* show clear population substructure that, in the case of *D. pustulata*, does not always follow geography or ecology. *D. erudita* and *D. hequetiae* showed only very low levels of genetic differentiation in this thesis, so they could either be part of the same species, or be very recently diverged species.

Demographic inference was greatly limited in its ability to accurately estimate parameters due to limitations with available information, samplings and potential excess heterozygosity. Relative estimates however are believed to be accurate and could be reproduced between different methods. Most populations (exception *D. revolutissima* and *D. erudita/hequetiae*) exhibited a stark decrease in effective population size in recent times, highlighting concerns for conservation. A period of very high effective population sizes from 100 thousand years ago to four to six million years ago coincides roughly with modeled divergence times in two species pairs and previously published times of divergence. Importantly, interspecific gene flow appears prevalent across historical periods and species in this radiation, with

important potential consequences for the amount of genetic diversity present in this group and its adaptability. Additionally, speciation through hybridisation was shown to be a valid model for speciation in the tested subset of three populations.

References

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Balloux, F. (2004). HETEROZYGOTE EXCESS IN SMALL POPULATIONS AND THE HETEROZYGOTE‐EXCESS EFFECTIVE POPULATION SIZE. *Evolution*, *58*(9), 1891–1900. https://doi.org/10.1111/j.0014-3820.2004.tb00477.x

Bamberger, S., Xu, J., & Hausdorf, B. (2022). Evaluating Species Delimitation Methods in Radiations: The Land Snail Albinaria cretensis Complex on Crete. *Systematic Biology*, *71*(2), 439–460. https://doi.org/10.1093/sysbio/syab050

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, *32*(18), 2817–2823. https://doi.org/10.1093/bioinformatics/btw327

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*(9), 1514–1521. https://doi.org/10.1101/gr.154831.113

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932. https://doi.org/10.1093/molbev/mss086

Carpenter, R. J., Read, J., & Jaffré, T. (2003). Reproductive traits of tropical

rain-forest trees in New Caledonia. *Journal of Tropical Ecology*, *19*(4), 351–365. https://doi.org/10.1017/S0266467403003407

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, *11*(3), e1460. https://doi.org/10.1002/wics.1460

Cerca, J., Cotoras, D. D., Bieker, V. C., De-Kayne, R., Vargas, P., Fernández-Mazuecos, M., López-Delgado, J., White, O., Stervander, M., Geneva, A. J., Andino, J. E. G., Meier, J. I., Roeble, L., Brée, B., Patiño, J., Guayasamin, J. M., Torres, M. de L., Valdebenito, H., Castañeda, M. del R., … Martin, M. D. (2023). Evolutionary genomics of oceanic island radiations. *Trends in Ecology & Evolution*, *38*(7), 631–642. https://doi.org/10.1016/j.tree.2023.02.003

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Duangjai, S., Samuel, R., Munzinger, J., Forest, F., Wallnöfer, B., Barfuss, M. H. J., Fischer, G., & Chase, M. W. (2009). A multi-locus plastid phylogenetic analysis of the pantropical genus Diospyros (Ebenaceae), with an emphasis

on the radiation and biogeographic origins of the New Caledonian endemic species. *Molecular Phylogenetics and Evolution*, *52*(3), 602–620. https://doi.org/10.1016/j.ympev.2009.04.021

Edelman, N. B., & Mallet, J. (2021). Prevalence and Adaptive Impact of Introgression. *Annual Review of Genetics*, *55*(Volume 55, 2021), 265–283. https://doi.org/10.1146/annurev-genet-021821-020805

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*(24), 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

García-Verdugo, C., Friar, E., & Santiago, L. S. (2013). Ecological Role of Hybridization in Adaptive Radiations: A Case Study in theDubautia arborea–Dubautia ciliolata (Asteraceae) Complex. *International Journal of Plant Sciences*, *174*(5), 749–759. https://doi.org/10.1086/669929

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing* (arXiv:1207.3907). arXiv. https://doi.org/10.48550/arXiv.1207.3907

Grandcolas, P., Murienne, J., Robillard, T., Desutter-Grandcolas, L., Jourdan, H., Guilbert, E., & Deharveng, L. (2008). New Caledonia: A very old Darwinian island? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1508), 3309–3317. https://doi.org/10.1098/rstb.2008.0122

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019).

Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, *8*(5), giz034. https://doi.org/10.1093/gigascience/giz034

Heilbuth, J. C., Ilves, K. L., & Otto, S. P. (2001). THE CONSEQUENCES OF DIOECY FOR SEED DISPERSAL: MODELING THE SEED-SHADOW HANDICAP. *Evolution*, *55*(5), 880–888. https://doi.org/10.1554/0014-3820(2001)055[0880:TCODFS]2.0.CO;2

Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). PHAST and RPHAST: Phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, *12*(1), 41–51. https://doi.org/10.1093/bib/bbq072

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, *15*(5), 1179–1191. https://doi.org/10.1111/1755-0998.12387

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Korneliussen, T. S., & Moltke, I. (2015). NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, *31*(24), 4009–4011. https://doi.org/10.1093/bioinformatics/btv509

Krueger, F. (2015). Trim galore. *A Wrapper Tool around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files*, *516*, 517.

Lapierre, M., Lambert, A., & Achaz, G. (2017). Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population. *Genetics*, *206*(1), 439–449. https://doi.org/10.1534/genetics.116.192708

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), Article 16. https://doi.org/10.1093/bioinformatics/btp352

Liu, X., & Fu, Y.-X. (2020). Stairway Plot 2: Demographic history inference with folded SNP frequency spectra. *Genome Biology*, *21*(1), 280. https://doi.org/10.1186/s13059-020-02196-9

Luqman, H., Wegmann, D., Fior, S., & Widmer, A. (2023). Climate-induced range shifts drive adaptive response via spatio-temporal sieving of alleles. *Nature Communications*, *14*(1), 1080. https://doi.org/10.1038/s41467-023-36631-9

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Marques, D. A., Meier, J. I., & Seehausen, O. (2019). A Combinatorial View on Speciation and Adaptive Radiation. *Trends in Ecology & Evolution*, *34*(6), 531–544. https://doi.org/10.1016/j.tree.2019.02.008

Mas-Sandoval, A., Pope, N. S., Nielsen, K. N., Altinkaya, I., Fumagalli, M., & Korneliussen, T. S. (2022). Fast and accurate estimation of multidimensional site frequency spectra from low-coverage high-throughput sequencing data. *GigaScience*, *11*, giac032. https://doi.org/10.1093/gigascience/giac032

McGee, M. D., Borstein, S. R., Meier, J. I., Marques, D. A., Mwaiko, S., Taabu, A., Kishe, M. A., O'Meara, B., Bruggmann, R., Excoffier, L., & Seehausen, O.

(2020). The ecological and genomic basis of explosive adaptive radiation. *Nature*, *586*(7827), Article 7827. https://doi.org/10.1038/s41586-020-2652-7

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, *8*(1), 14363. https://doi.org/10.1038/ncomms14363

Meirmans, P. G. (2019). Subsampling reveals that unbalanced sampling affects Structure results in a multi-species dataset. *Heredity*, *122*(3), 276–287. https://doi.org/10.1038/s41437-018-0124-8

Meisner, J., & Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics*, *210*(2), 719–731. https://doi.org/10.1534/genetics.118.301336

Muyle, A., Martin, H., Zemp, N., Mollion, M., Gallina, S., Tavares, R., Silva, A., Bataillon, T., Widmer, A., Glémin, S., Touzet, P., & Marais, G. A. B. (2021). Dioecy Is Associated with High Genetic Diversity and Adaptation Rates in the Plant Genus Silene. *Molecular Biology and Evolution*, *38*(3), 805–818. https://doi.org/10.1093/molbev/msaa229

Nordborg, M., & Krone, S. M. (2002). Separation of Time Scales and Convergence to the Coalescent in Structured Populations. In M. Slatkin & M. Veuille (Eds.), *Modern developments in theoretical population genetics* (pp. 194–232). Oxford University PressOxford.

https://doi.org/10.1093/oso/9780198599623.003.0012

Ortiz, E. M. (2023). *Vcf2phylip* [Python]. https://github.com/edgardomortiz/vcf2phylip (Original work published 2017)

Parker, E., Dornburg, A., Struthers, C. D., Jones, C. D., & Near, T. J. (2022). Phylogenomic Species Delimitation Dramatically Reduces Species Diversity in an Antarctic Adaptive Radiation. *Systematic Biology*, *71*(1), 58–77. https://doi.org/10.1093/sysbio/syab057

Paun, O., Turner, B., Trucchi, E., Munzinger, J., Chase, M. W., & Samuel, R. (2016). Processes Driving the Adaptive Radiation of a Tropical Tree (*Diospyros* , Ebenaceae) in New Caledonia, a Biodiversity Hotspot. *Systematic Biology*, *65*(2), 212–227. https://doi.org/10.1093/sysbio/syv076

Petren, K., Grant, P. R., Grant, B. R., & Keller, L. F. (2005). Comparative landscape genetics and the adaptive radiation of Darwin's finches: The role of peripheral isolation. *Molecular Ecology*, *14*(10), 2943–2957. https://doi.org/10.1111/j.1365-294X.2005.02632.x

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Puechmaille, S. J. (2016). The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*(3), 608–627. https://doi.org/10.1111/1755-0998.12512

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432. https://doi.org/10.1038/s41467-020-14998-3

RhEA/Ancestral sequence reconstruction/PHAST2fasta.py at main · hirzi/RhEA · GitHub. (n.d.). Retrieved April 26, 2024, from https://github.com/hirzi/RhEA/blob/main/Ancestral%20sequence%20reconstruction/PHAST2fasta.py

Samuel, R., Turner, B., Duangjai, S., Munzinger, J., Paun, O., Barfuss, M. H. J., & Chase, M. W. (2019). Systematics and evolution of the Old World Ebenaceae, a review with emphasis on the large genus Diospyros and its radiation in New Caledonia. *Botanical Journal of the Linnean Society*, *189*(2), 99–114. https://doi.org/10.1093/botlinnean/boy081

Schatz, G. E., & Porter, P. L. (2018). Novitates neocaledonicae. IX. Taxonomic notes on New Caledonian Diospyros (Ebenaceae) with new synonymy and the description of two new species. *Candollea*, *73*(1), 91. https://doi.org/10.15553/c2018v731a8

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, *19*(4), 198–207. https://doi.org/10.1016/j.tree.2004.01.003

Seehausen, O. (2006). African cichlid fish: A model system in adaptive radiation research. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1597), 1987–1998. https://doi.org/10.1098/rspb.2006.3539

Shin, C. P., & Allmon, W. D. (2023). How we study cryptic species and their biological implications: A case study from marine shelled gastropods. *Ecology*

and Evolution, *13*(9), e10360. https://doi.org/10.1002/ece3.10360

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693–702. https://doi.org/10.1534/genetics.113.154138

Stroud, J. T., & Losos, J. B. (2016). Ecological Opportunity and Adaptive Radiation. *Annual Review of Ecology, Evolution, and Systematics*, *47*(1), 507–532. https://doi.org/10.1146/annurev-ecolsys-121415-032254

Suarez-Gonzalez, A., Lexer, C., & Cronk, Q. C. B. (2018). Adaptive introgression: A plant perspective. *Biology Letters*, *14*(3), 20170688. https://doi.org/10.1098/rsbl.2017.0688

Tange, O. (2018). *GNU Parallel 2018*. Ole Tange. https://doi.org/10.5281/zenodo.1146014

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics (Oxford, England)*, *31*(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Tassin, J., Boissenin, M., & Barré, N. (2010). Can Ptilinopus greyii (Columbidae) Disperse Seeds in New Caledonia's Dry Forests?1. *Pacific Science*, *64*(4), 527–532. https://doi.org/10.2984/64.4.527

Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, *102*(6), e03336. https://doi.org/10.1002/ecy.3336

Turner, B., Munzinger, J., Duangjai, S., Temsch, E. M., Stockenhuber, R., Barfuss, M. H. J., Chase, M. W., & Samuel, R. (2013). Molecular phylogenetics of New Caledonian Diospyros (Ebenaceae) using plastid and nuclear markers. *Molecular Phylogenetics and Evolution*, *69*(3), 740–763.

https://doi.org/10.1016/j.ympev.2013.07.002

Vcflib. (2023). [C++]. vcflib. https://github.com/vcflib/vcflib/blob/fdf45678d420ce311de960575a08512d45f7e80e/doc/vcfrandomsample.md (Original work published 2010)

Wakeley, J., & Sargsyan, O. (2009). Extensions of the Coalescent Effective Population Size. *Genetics*, *181*(1), 341–345. https://doi.org/10.1534/genetics.108.092460

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1395–1409. https://doi.org/10.1098/rstb.2005.1682

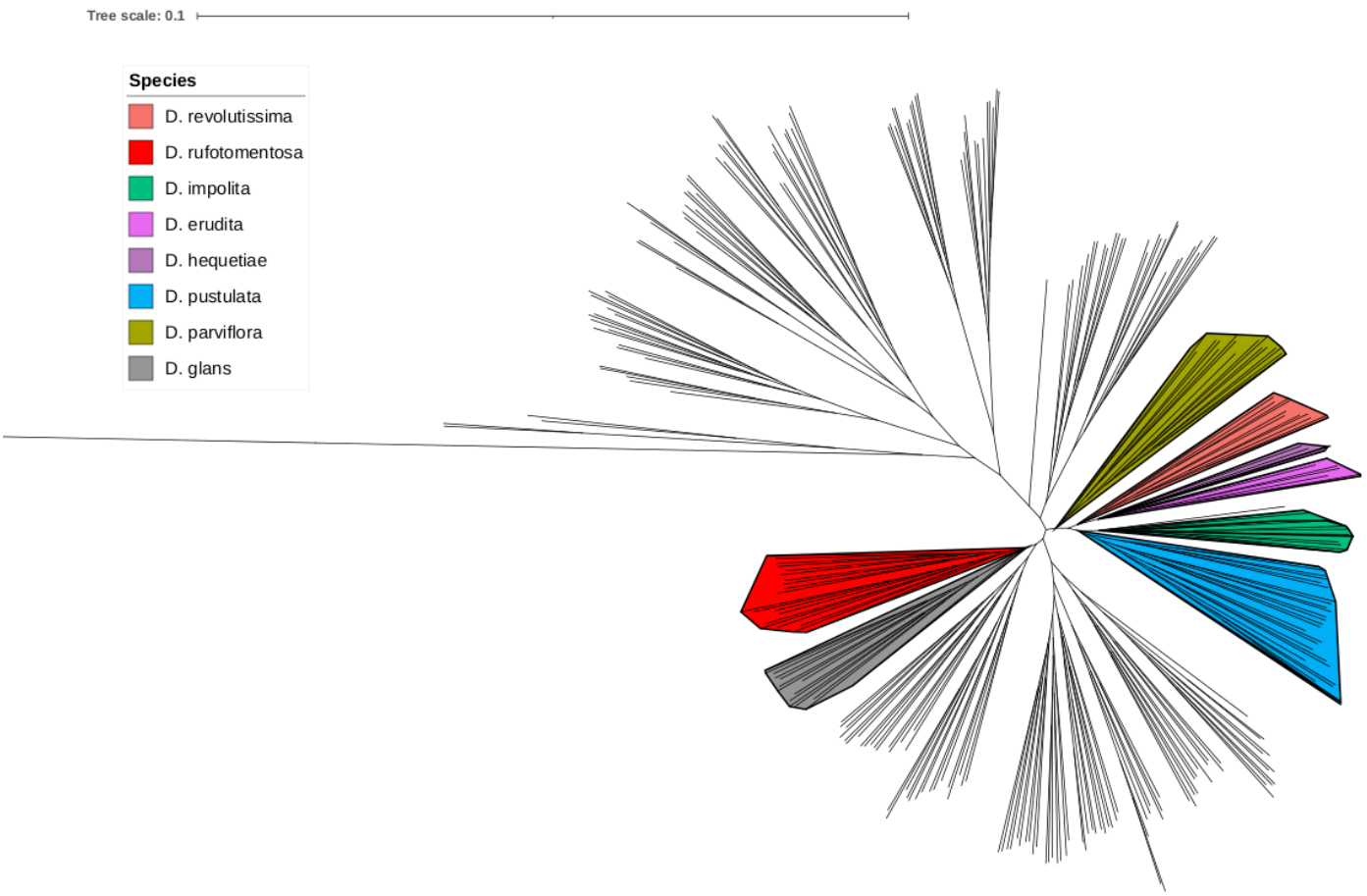
White, F. (n.d.). Ebenaceae Winteraceae. *Flore de La Nouvelle-Calédonie et Dépendances, 19*. Retrieved May 23, 2024, from https://agris.fao.org/search/en/providers/122376/records/6511a27560f8dcc51c5fffc3

WHITE, F. (1992). Twenty-two new and little-known species of Diospyros (Ebenaceae) from New Caledonia with comments on section Maba. *Twenty-Two New and Little-Known Species of Diospyros (Ebenaceae) from New Caledonia with Comments on Section Maba*, *14*(2), 179–222.

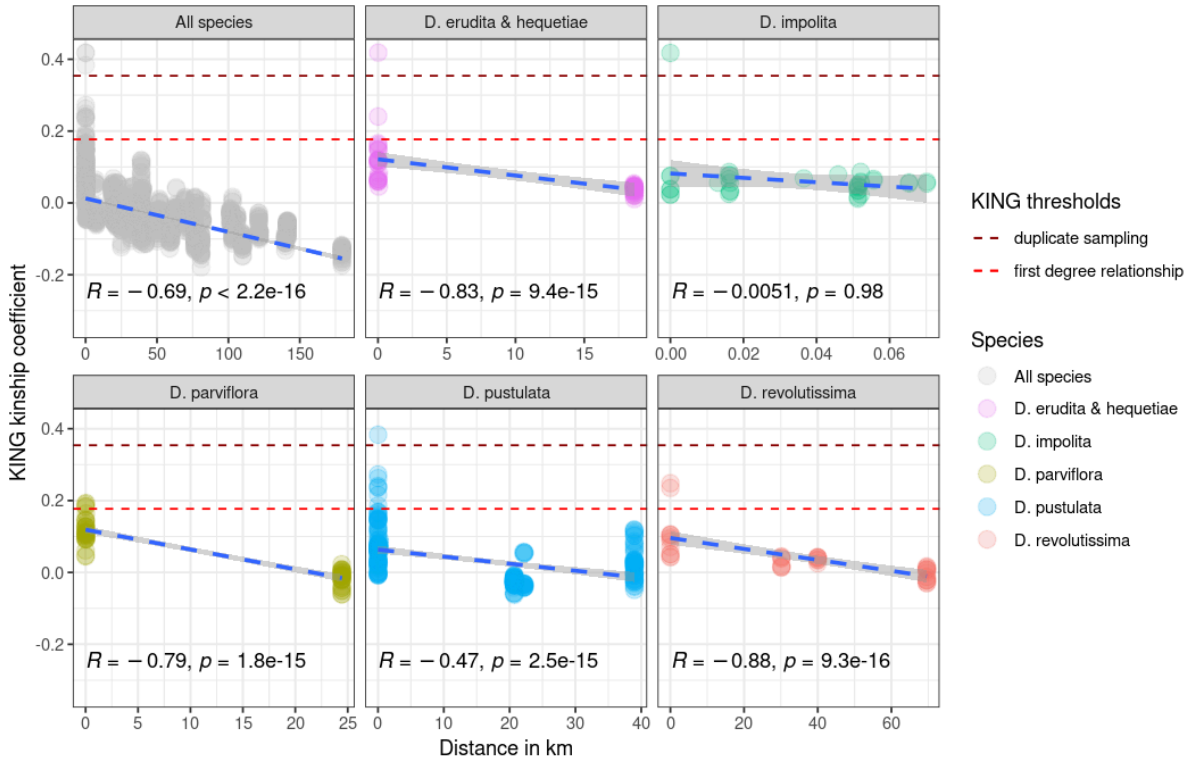
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wulff, A. S., Hollingsworth, P. M., Ahrends, A., Jaffré, T., Veillon, J.-M., L'Huillier, L., & Fogliani, B. (2013). Conservation Priorities in a Biodiversity Hotspot: Analysis of Narrow Endemic Plant Species in New Caledonia. *PLOS ONE*, *8*(9), e73371. https://doi.org/10.1371/journal.pone.0073371

Supplemental



Supplemental 1: Phylogeny generated from all individual samples of New Caledonian Diospyros (Khastgir, unpublished). Samples used in this thesis are highlighted according to species assignment after results presented in the thesis. D. rufotomentosa, with individuals originally described as D. parviflora and/or D. perplexa.



Supplemental 2: Relationship of KING coefficient and geographic distance in kilometers for all individuals and populations as used for demographic inference (suspected recent hybrids already removed). Each point represents a pairwise comparison of KING coefficient and geographical distance for one sample and another. Dashed red lines indicate thresholds for first degree relationships and duplicate samplings (clonal relationship). In all but D. impolita there was a significant relationship between geographical distance and relationship (Spearman). All observations of first degree relationship or higher occurred at the same sampling location (distance = 0 km). The range of the x-axis for the individual populations also highlights differences in sampling range for the populations, with individuals of D. impolita all occurring within ~80 m of each other and some D. revolutissima samplings more than 60 km from each other.

```
#example blueprint file
#input setting
popid: D.impolita # id of the population (no white space)
nseq: 18 # number of sequences
L: 494023 # total number of observed nucleic sites, including polymorphic and monomorphic
whether_folded: false # whethr the SFS is folded (true or false)
SFS: 5532.070694 3645.452577 2234.439142 1664.010267 1216.272206 906.818256 722.972998 560.171266 554.890707 342.052094
390.561411 365.720295 271.915855 273.119967 261.991935 283.922131 194.359234 # snp frequency spectrum: number of singleton,
number of doubleton, etc. (separated by white space)
#smallest_size_of_SFS_bin_used_for_estimation: 2 # default is 1; to ignore singletons, uncomment this line and change this
number to 2
#largest_size_of_SFS_bin_used_for_estimation: 21 # default is n-1; to ignore singletons, uncomment this line and change this
number to nseq-2
pct_training: 0.67 # percentage of sites for training
nrand: 4        8       12      16 # number of random break points for each try (separated by white space)
project_dir: moresites_impolita # project directory
stairway_plot_dir: stairway_plot_es # directory to the stairway plot files
ninput: 200 # number of input files to be created for each estimation
#random_seed: 6
#output setting
mu: 7e-9 # assumed mutation rate per site per generation
year_per_generation: 7 # assumed generation time (in years)
#plot setting
plot_title: two-epoch # title of the plot
xrange: 0,0 # Time (1k year) range; format: xmin,xmax; "0,0" for default
yrange: 0,0 # Ne (1k individual) range; format: xmin,xmax; "0,0" for default
xspacing: 4 # X axis spacing
yspacing: 4 # Y axis spacing
fontsize: 12 # Font size
```

Supplemental 3: Example blueprint file for stairway plot v2 for D. impolita.

```
//Parameters for the coalescence simulation program : simcoal.exe
3 samples to simulate :
//Population effective sizes (number of genes)
NREVO
NPARV
NPAPE
//Samples sizes and samples age
16
10
12
//Growth rates: negative growth implies population expansion
0
0
0
//Number of migration matrices : 0 implies no migration between demes
3
//Migration matrix 0
0 M01 M02
M01 0 M12
M02 M12 0
//Migration matrix 1
0 0 MN02
0 0 0
MN02 0 0
//Migration matrix 2
0 0 0
0 0 0
0 0 0
//historical event: time, source, sink, migrants, new deme size, growth rate, migr mat index
3 historical event
THYB  1 0 PHYB NHREVO 0 1 absoluteResize
THYB  1 2    1 NHPAPE 0 1 absoluteResize
TDIV  0 2    1   NANC 0 2 absoluteResize
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of contiguous linkage Block: a block is a set of contiguous loci
1
//per Block:data type, number of loci, per gen recomb and mut rates
FREQ 1 0 7e-9 OUTEXP
```

Supplemental 4: Template file for hybridisation model used in three population modeling.

Parameter estimates for revolutissima and erudita											
NPOP0 (revo)	NPOP1 (eru)	NP0C (revo)	NP1C (eru)	NANC	TCH1 (revo)	TCH2 (eru)	TDIV	MIG01	MIG10	pop	sfs
2940433	1238121	1786761	954702	504559	82559	72337	759401	1.27E-06	2.31E-06	eru x revo	fold
3137778	888109	1817624	1078982	600162	64137	3700	647776	1.18E-06	2.53E-06	eru x revo	unfold
Parameter estimates for revolutissima and impolita											
NPOP0 (revo)	NPOP1 (impo)	NP0C (revo)	NP1C (impo)	NANC	TCH1 (revo)	TCH2 (impo)	TDIV	MIG01	MIG10	pop	sfs
3725256	79180	1925720	565868	453313	69037	2256	852693	5.04E-07	2.00E-06	impo x revo	fold
5549159	342473	954131	1914659	590844	98800	78615	780317	7.60E-07	1.68E-06	impo x revo	unfold

Supplemental 5: Parameter estimates for best fitting run of two population models of divergence with migration and size change. Included are both population pairings and folded and unfolded input SFS.

sfs	**model**	**deltaL**	**AIC**	**deltaAIC**	**wi**
folded	eru_revo_ancestralM	1649.437	559991.0481	4274.865711	0
folded	eru_revo_constantM	721.59	555716.18	0.00	1.00
folded	eru_revo_currentM	744.14	555822.029	105.85	0.00
folded	eru_revo_isolation	2104.69	562081.55	6365.36	0.00
folded	*eru_revo_constantM_sizechng*	*256.09*			
sfs	**model**	**deltaL**	**AIC**	**deltaAIC**	**wi**
folded	impo_revo_ancestralM	1716.863	530453.5326	2640.647387	0
folded	impo_revo_constantM	1143.888	527812.8852	0	1
folded	impo_revo_currentM	1159.533	527886.9331	74.04788756	8.33E-17
folded	impo_revo_isolation	1906.508	531320.8801	3507.994887	0
folded	*impo_revo_constantM_sizechng*	*340.789*			
sfs	**model**	**deltaL**	**AIC**	**deltaAIC**	**wi**
unfolded	eru_revo_ancestralM	1847.1	572472.5652	4819.887602	0
unfolded	eru_revo_constantM	800.909	567652.6776	0	1
unfolded	eru_revo_currentM	811.383	567702.9122	50.23455253	1.24E-11
unfolded	eru_revo_isolation	2352.222	574792.738	7140.060377	0
unfolded	*eru_revo_constantM_sizechng*	*376.072*			
sfs	**model**	**deltaL**	**AIC**	**deltaAIC**	**wi**
unfolded	impo_revo_ancestralM	2362.1	539750.7219	2967.236847	0
unfolded	impo_revo_constantM	1718.207	536783.485	0	1
unfolded	impo_revo_currentM	1756.456	536961.6282	1.78E+02	2.07E-39
unfolded	impo_revo_isolation	2674.508	541183.4139	4399.928854	0
unfolded	*impo_revo_constantM_sizechng*	*1003.343*			

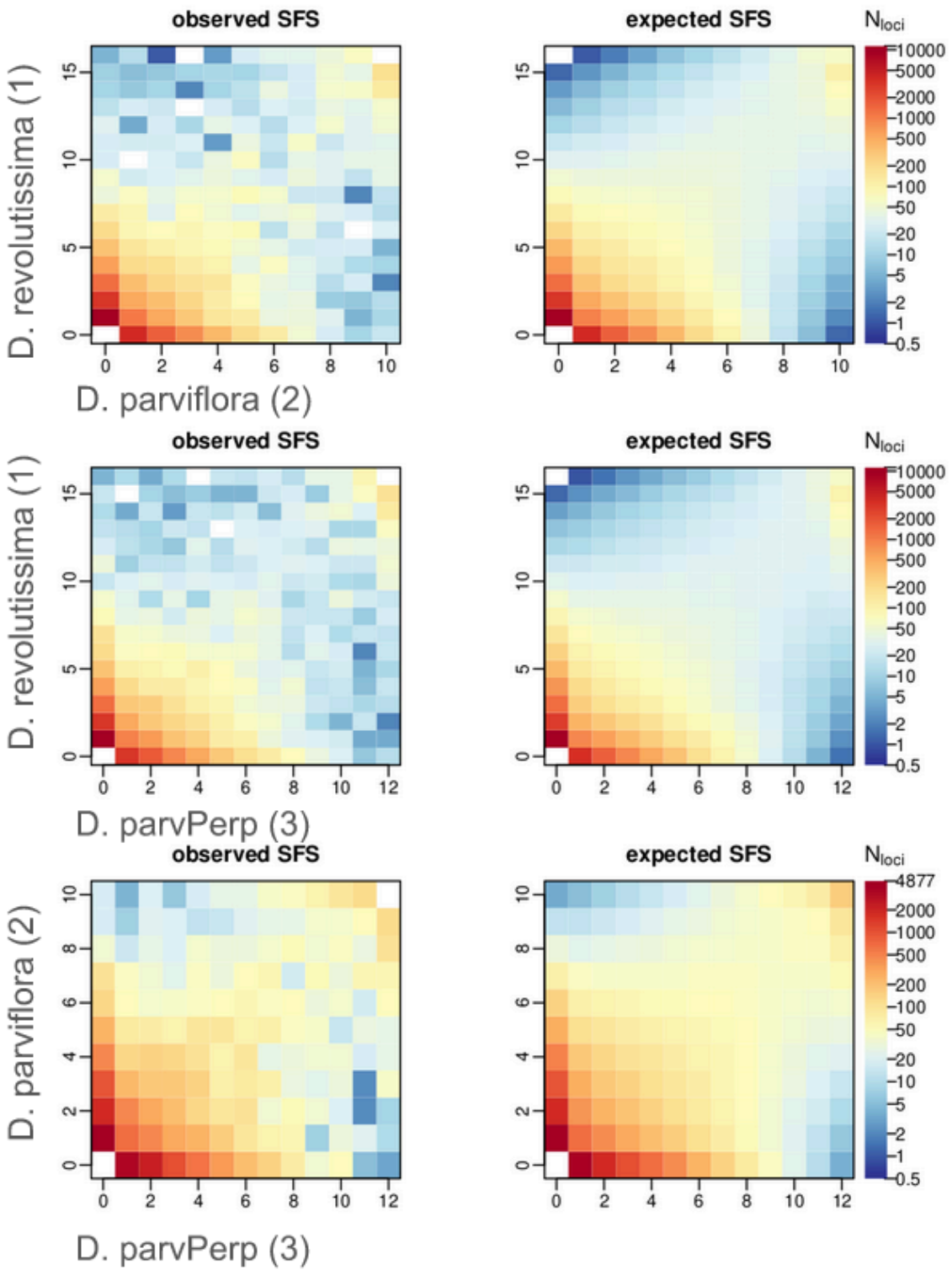
Supplemental 6: Delta likelihood and AIC for all models tested for two populations. AIC was not calculated for sizechange model, as it performed by far the best.

Parameter estimates for hybridisation model												
NREVO	NPARV	NPAPE	NHREVO	NHPAPE	NANC	THYB	TDIV	M01	M02	M12	MN02	PHYB
2490147	459521	376953	1067072	3506607	565592	138117	662787	1.16E-06	1.45E-06	1.54E-06	6.22E-07	0.1756774601
Parameter estimates for single origin model												
NREVO	NPARV	NPAPE	NDREVO	NDPAPE	NANC	TDIV1	TDIV2	M01	M02	M12	MN02	
2424053	586373	384672	1385438	4388560	523563	179351	736436	3.16E-07	1.25E-06	1.91E-06	4.31E-07	

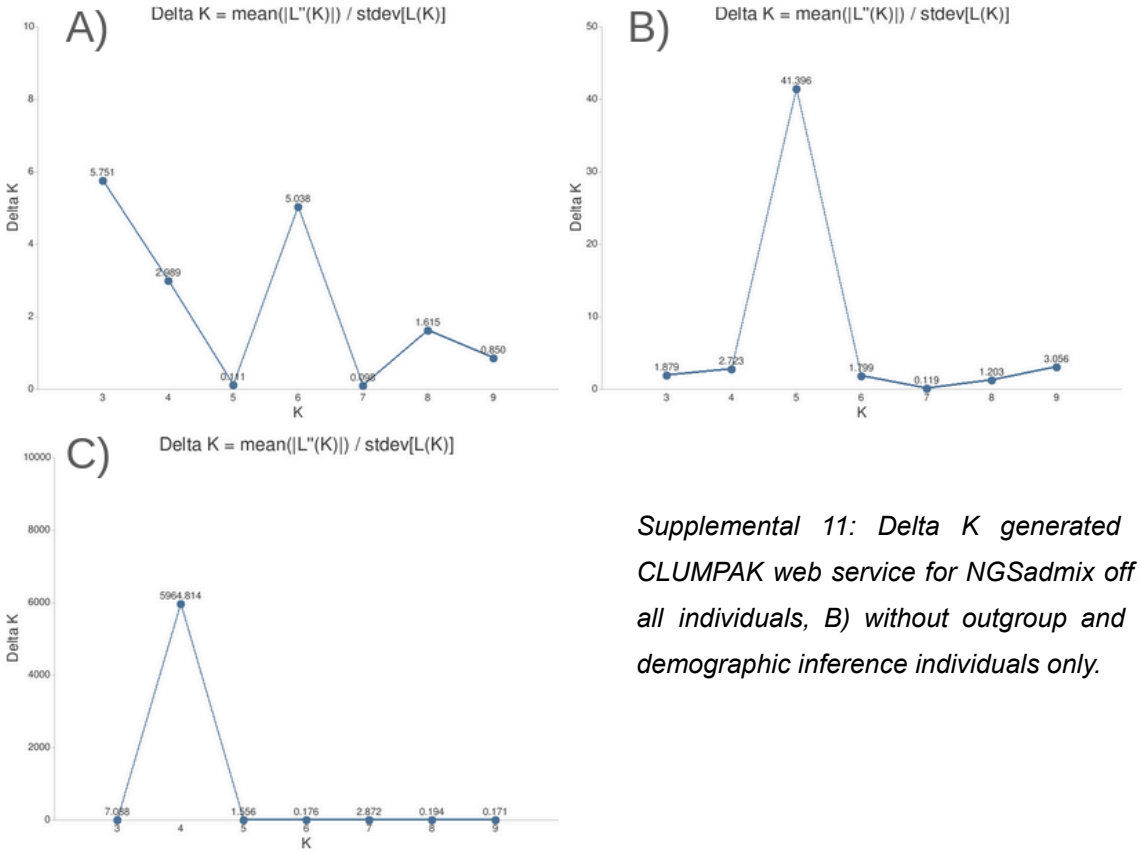
Supplemental 7: Parameter estimates for both hybridisation and single origin of D. parviflora models (best run).

model	**deltaL**	**AIC**	**deltaAIC**	**wi**
prp_2origA	990.058	1375702	621.885431	8.79E-136
prp_2origB	949.512	1375516	435.164201	3.09E-95
prp_1orig	856.469	1375087	6.685351	3.41E-02
prp_hybA	*854.583*	*1375081*	*0*	*9.66E-01*

Supplemental 8: Delta likelihood and AIC computed for three population models of D. revolutissima and D. parviflora (D. parviflora and D. parviflora L10).



Supplemental 9: Two dimensional input SFS (observed) used for demographic inference with expected SFS under best parameter estimates for hybridisation model of D. parviflora.



Supplemental 11: Delta K generated by CLUMPAK web service for NGSadmix off A) all individuals, B) without outgroup and C) demographic inference individuals only.