MASTERARBEIT | MASTER‘S THESIS

Titel | Title

Whole genome re-sequencing unravels complex speciation processes within the RRc clade of the radiated *Diospyros* in New Caledonia

verfasst von | submitted by

Johannes Zechmeister 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 066 832
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Botanik
Betreut von | Supervisor:	Assoz. Prof. Dipl.-Ing. Dr Ovidiu Paun

Acknowledgements

First of all, I want to thank my supervisor, Prof. Ovidiu Paun, for his time, efforts, and continuous availability throughout the duration of this research. His guidance and general enthusiasm for evolutionary biology have been invaluable.

Further thanks go to my colleagues, Katie Emelianova, Teerna Khastgir, Florian Brueck, Luiz Cauz, Clara Groot Crego, Amin Ghane, and Fabian Oberauer. Their companionship and dedication to study similar topics has been a source of motivation, and various kinds of conversation added a lot of enjoyment to the process.

My appreciation extends to those who contributed to the fieldwork: Ovidiu Paun, Katie Emelianova, Teerna Khastgir, Jerome Munzinger (IRD Montpellier, France), and David Bruy (IRD Noumea, New Caledonia). Their efforts and support in the field were essential. I also thank Barbara Turner (Boku) and Rose Samuel (UniVie) for providing their DNA extractions and silica gel-fixed material.

Finally, I gratefully acknowledge the Austrian Science Fund (FWF) [grant DOI 10.55776/P35275] for providing the funding necessary for the data collection in this MSc thesis.

Abstract

The purpose of this study is to unravel the dynamics of speciation within the radiated RRc clade of *Diospyros* in New Caledonia. By analyzing population structure and by inferring demographic scenarios, we aimed to shed light on the demographic history and genetic diversity to better understand its evolutionary history and to contribute to comprehensive conservation strategies. In particular, we test whether (1) speciation with gene flow was more prominent than speciation without gene flow, and if (2) the population *D. minimifolia* L17 originated by hybridization.

New Caledonia has a wide array of ecologically distinct habitats due to its geological history, contributing to its high level of endemism and biodiversity. Around 30 species of *Diospyros* are currently present in New Caledonia, over 20 of these originated through a rapid adaptive radiation in the last seven million years. The analysis of genome data from six to ten individuals per population, six populations in total, was carried out by the software tools ANGSD and fastsimcoal2.

The results of the principal component analysis, admixture analysis, Fst analysis, and demographic inference indicate complex scenarios of diversification within the RRc clade. *D. tridentata* is shown to be the most genetically distinct from the other populations, while the southern populations *D. calciphila*, *D. inexplorata* and *D.* sp. Pic N'ga, have relatively close genetic relationships. *D. minimifolia* L17 is genetically distinct from the other populations, but further research is needed to determine if it underwent speciation by hybridization. The results further indicate that gene flow played a role during and/or after the speciation processes of all analyzed groups.

Kurzfassung

Die vorliegende Arbeit zielt auf die Analyse der Artbildungsprozesse innerhalb der RRc-Klade von *Diospyros* auf Neukaledonien ab. Die sich ergebenden Einsichten sollen es uns ermöglichen, evolutionäre Dynamiken innerhalb der untersuchten Gruppe besser nachvollziehen zu können, um auf den Erhalt biologischer Diversität bezogene Schutzmaßnahmen gegebenenfalls anpassen zu können. Konkret wird mittels Analyse genetischer Daten versucht zu eruieren, ob (1) Artbildung während dem Auftreten von genetischem Austausch (engl.: gene flow) stattgefunden hat und ob (2) *D. minimifolia* L17 durch Hybridisation zweier verschiedener Arten entstand.

Die geologische Geschichte von Neukaledonien spiegelt sich in einer großen Vielfalt an Habitaten und einer hohen Anzahl an endemischen Arten sowie einer hohen biologischen Vielfalt im Allgemeinen wider. Etwa 30 Arten der Gattung *Diospyros* sind derzeit auf dem Archipel vorzufinden, über 20 dieser Arten sind Teil der adaptiven Radiation und entstanden in den letzten sieben Millionen Jahren. Mittels der bioinformatischen Werkzeuge ANGSD und fastsimcoal2 konnten genetische Daten von sechs bis zehn Individuen pro Population, insgesamt sechs Populationen, analysiert und demographische Szenarien geprüft werden.

Die Ergebnisse der Hauptkomponentenanalyse (engl.: Principal Component Analysis), Admixture-Analyse, Fst-Analyse sowie der Untersuchung der demographischen Geschichte deuten auf eine komplexe Diversifizierung innerhalb der RRc-Klade hin. *D. tridentata* ist genetisch deutlich verschieden von den anderen untersuchten Populationen. In geringerem Ausmaß gilt dies auch für die Population *D. minimifolia* L17. Im Gegensatz dazu sind die Populationen *D. calciphila*, *D. inexplorata* und *D.* sp. Pic N'ga genetisch relativ ähnlich. Die Untersuchung der demographischen Geschichte ermöglicht keine klare Aussage bezüglich eines Artbildungsprozesses durch Hybridisation von *D. minimifolia* L17. Es konnte jedoch gezeigt werden, dass genetischer Austausch während und/oder nach Artbildungsprozessen aller Populationen wahrscheinlicher ist als Artbildung ohne diesen.

Contents

List of Tables

List of Figures

1. Introduction

Adaptive radiation is an evolutionary phenomenon, in which lineages rapidly diversify into a multitude of clades, each adapted to different ecological niches. Prominent examples are Darwin's finches on the Galápagos islands (Abzhanov, 2010), and the cichlid fish in the African Great Lakes (Seehausen, 2006) and the silversword alliance in Hawaii (Carr and Kyhos, 1986). Adaptive radiations commonly begin after long distance dispersal and the arrival at geographic sites with a wide array of available niches. The high rate of speciation complicates genomic analyses due to incomplete lineage sorting (ILS), which results in the persistence of shared polymorphisms through speciation events (Suh et al., 2015). Additionally, interspecific hybridization and introgression, which are common inside radiations due to incomplete reproductive isolation between species, further complicate reconstructing evolutionary histories within radiations (Seehausen, 2004).

The islands of New Caledonia encompass a mosaic of different soil types because of their unique geological history. It is estimated that from 75 to 60 million years ago, today's landmass was submerged under the lithosphere mantle (Isnard et al., 2016; Pillon et al., 2020). As a result, overthrusts of ultramafic rocks cover around one third of the main island today (Pintaud et al., 2001). The various kinds of ultramafic soils feature various nutrient deficiency, cation imbalances and have a very low water holding capacity (Isnard et al., 2016).

The climate of New Caledonia has been relatively stable since the late Cretaceous, for example compared to New Zealand and Australia, where glaciation or desertification events occurred. During the Pleistocene (around 2 million to 10 000 years ago) Earth's climate oscillated between warm/interglacial and cold/glacial periods and water was fixed during the cold periods in polar regions as ice. These oscillations have changed the sea level by up to 200 meters, alternating the size and ecology of the New Caledonian's islands drastically. Although the climate was generally drier during ice ages, there are indications that rainforests persisted continuously on New Caledonia since 40 thousand years ago and possibly longer (Pintaud et al., 2001). The climate conditions on the main island vary due to a mountain range spanning from North to South and predominant wind direction from East to West. This creates dry conditions in the north-west, and wet conditions in the south-east and at higher elevations.

Today, the main types of primary vegetation across the archipelago are saline vegetation, marsh and swamp vegetation, wet evergreen forests, dry/sclerophyllous forest and maquis. Secondary vegetation types, which appear in relation to the intensity of human activities, are savannas and different degradation stages of forests (Morat, 1993). The diversity of topological, edaphic and climatic conditions as well as the long term isolation of the islands promote speciation and contribute to the remarkable plant diversity and high endemism. Around 2,270 vascular plant species inhabit New Caledonia and 74% of these

are endemic (Wulff et al., 2013; Isnard et al., 2016). Much of the biodiversity is threatened by human activities directly, for example via land use and mining. Around 30% of the world's nickel resources are located in New Caledonia. Furthermore, Neobiota and climate change threaten biodiversity as well. For example the New Caledonian tropical dry forests cover at present only an estimated 2% of their original area (Gillespie and Jaffré, 2003). Many of the species of sclerophyllous forests are included in the IUCN Red List. In addition, New Caledonia is one of the regions with the highest predicted land loss as a result of climate-change driven increase of sea level (Bellard and Courchamp, 2013).

The genus *Diospyros* is part of the family Ebenaceae and comprises over 700 species. The species are pan-tropically distributed trees and shrubs. Multiple colonization of New Caledonia by different lineages within the genus have occurred over the last few million years following long-distance-dispersal events. Phylogenetic surveys designate these into four clades: three with one to five species that remain restricted with regard to ecological preference and geographic area of distribution, and a fourth with over 30 species (Duangjai et al., 2009; Turner et al., 2013) that occupy all ecological niches on the archipelago, except for mangroves. The latter, speciose group is the result of a rapid adaptive radiation which started around nine million years ago and accelerated roughly exponentially over the last six million years. This acceleration is remarkable and thought to have happened because of initial adaptations to climatic conditions, followed by subsequent adaptations to a wide array of edaphic conditions (Paun et al., 2016).

The recentness of this proliferation is accompanied by genetic and morphological similarities causing significant taxonomic challenges in species delimitation. This master's thesis focuses on a monophyletic sub-group of the radiating clade, referred to as the RRc group in Paun et al. (2016). It entails six lineages: *D. calciphila* F. White, *D. inexplorata* F. White, *D. minimifolia* L10+20 F. White, *D. minimifolia* L17 F. White, *D.* sp. Pic N'ga, and *D. tridentata* F. White. This grouping is thought to be reasonable based on prior analyses (Paun et al., 2016) and our analyses.

The grouping mentioned above either represents recognized species, distinct populations of such, or not yet described species. Samples of *D. minimifolia* were split into two different populations based on the location of sampling and because they did not behave as a monophyletic group in prior phylogenetic surveys. *D. inexplorata* was formerly recognized as a species, but recently taxonomically revised and included in *D. calciphila* due to the genetic similarity and little to no ecological and morphological differences (Schatz and Lowry, 2018). *D.* sp. Pic N'ga is currently taxonomically investigated, and may represent a new undescribed species. It is an ultramafic specialist and closely related to *D. calciphila*.

This complex sampling provides an ideal setting for exploring hidden diversity and for testing modes of speciation. Characteristics of gene flow, such as time of occurrence, direction and abundance, can help to better estimate its role during speciation in this group. Introgressive hybridization might have been very important in conquering the island's niches, either by providing new combinations of ecologically important traits or by enhancing the evolutionary potential for future selective pressures. However, gene flow is also known to have a multifaceted role, as it can be disruptive by causing the loss of locally

adapted alleles if selection is not strong enough (Tigano and Friesen, 2016). Furthermore, it is assumed that gene flow has the greatest impact on variation when populations have begun genetic and phenotypic divergence, but are not yet fully reproductively isolated (Grant and Grant, 2019).

Prior analyses showed that gene flow is prevalent throughout the *Diospyros* New Caledonian radiation, implying that geographical distribution, besides phylogenetic closeness, is an important factor in this regard (Paun et al., 2016). Having an improved understanding of evolutionary factors during rapid adaptive radiations like gene flow is of critical importance for guiding conservation efforts reasonably. The research questions of this MSc thesis target the genetic structure and the demographic history of the RRc group. Specifically, two hypotheses have been formulated. Hypothesis 1: Speciation with migration has more likely happened than speciation without migration. Hypothesis 2: *D. minimifolia* L17 originated by speciation through hybridization.

In this study, genomic data of six to ten individuals per population was analyzed, first via the program ANGSD (Korneliussen et al., 2014). ANGSD is a software that can calculate various summary statistics, perform association mapping and population genetic analyses. The program is especially equipped to handle low coverage data, because of its probabilistic approach to SNP calling. Following the ANGSD analysis, fastsimcoal2 (Excoffier et al., 2013) is used to infer demographic parameters with a flexible and robust simulation-based framework under various evolutionary models. By utilizing a composite likelihood approach, it is possible to choose the best fitting model for our situation. We will test different scenarios for splits within the clade RRc: strict isolation (SI; i.e., no gene flow after speciation); isolation with migration (IM; i.e., speciation happens in the presence of continuous geneflow); ancestral migration (AM; i.e., gene flow happens in the early phases of divergence, but stops at a particular time); and secondary contact (SC; i.e., no gene flow initially after speciation, but later and until the present). By utilizing these methods, we aim to better understand the phylogenetic relationships inside and the demographic history of the RRc group in *Diospyros*.

2. Material and Methods

2.1. Dataset and data processing

The sampling took place in New Caledonia on different occasions and was carried out by collaborators and colleagues. In total 59 samples were gathered from single sampling localities of *D. calciphila*, *D. inexplorata*, *D.* sp. Pic N'ga, *D. minimifolia* L17, and *D. tridentata* and from two populations of *D. minimifolia* L10+20. The sampling localities are shown in Figure 2.1 and in Table 2.1.

DNA was extracted from leaf tissue, libraries were constructed in house and sent to Next Generation Sequencing Facility at Vienna BioCenter Core Facilities (VBCF), Austria. DNA libraries were sequenced on Illumina NovaSeq with paired-end 150bp reads. Some of the samples used in this survey were sequenced twice due to insufficient output in the first trial, and subsequently merged. Computationally demanding tasks were performed using the Life Science Compute Cluster (LiSC) of the Faculty of Life Sciences, University of Vienna. The raw sequence reads, using Trim Galore (v. 0.6.10; available from https://github.com/FelixKrueger/TrimGalore; (Krueger, 2015). This step ensured that all adapter sequences were removed, that the reads exceeded a specific quality threshold, and that all of them were at least 20 base pairs (bp) long. Thereafter, the reads were aligned to the reference genome *Diospyros impolita* v. 1.1 (Paun, Khastgir and Emelianova, unpublished). The mapping took place with BWA (v. 0.7.17; (Li and Durbin, 2009), followed by filtering out alignments with a mapping quality less than 1 with Sambamba (v. 0.6.6; (Tarasov et al., 2015)) and sorting by the reference coordinates with Samtools (v. 1.13; (Li et al., 2009)). The pipeline for the trimming and mapping procedure is shown in the code section of the supplementary. The resulting BAM files could then be used as input data for the program ANGSD (v. 940; (Korneliussen et al., 2014)).

Table 2.1.: Summary of the localities sampled, the number of accessions analyzed, the geographic coordinates, and soil types on sites.

Population	No. samples	Locality	Latitude	Longitude	Soil
D. calciphila	10	Ile des Pins, Baie de Kanumera	-22.66541797	167.4415992	Calcareous
D. inexplorata	6	Presqu'Ãőle de KuÃľbini	-22.239472	167.019833	Calcareous
D. minimifolia L17	7	Paita, Baie de Gadji	-22.16956782	166.3791002	Sedimentary
D. minimifolia L10	7	Conservatoire botanique de TiÃľa	-21.1359569	164.9377142	Sedimentary
D. minimifolia L20	4	Metzdorff property, Moindah, Poya South	-21.42411067	165.1466668	Type E
D. sp. Pic N'ga	10	Ile des Pins, Pic N'Ga	-22.654655	167.45634	Ultramafic
D. tridentata	9	Koumac	-20.4721537	164.3212192	Calcareous

2.2. Calling genotype likelihoods and filtering sites

Genotype likelihoods were inferred for the principal component analysis and admixture analysis in ANGSD (v. 0.940) by creating a BEAGLE file (- doGlf 2). The estimation was based on the GATK model (- GL 2). Loci, present in less than 80% of the individuals, were excluded by using - minInd. The minor allele frequency was calculated (- doMaf 1 - doMajorMinor 1). To filter out rare alleles, the value for the minimum minor allele frequency (- minMaf) was set to 2/(number of samples). The minimum mapping quality (- minMapQ) was set to 30 and the minimum for the q-score (- minQ) to 20. The p-value for SNPs was set to 1e-6.

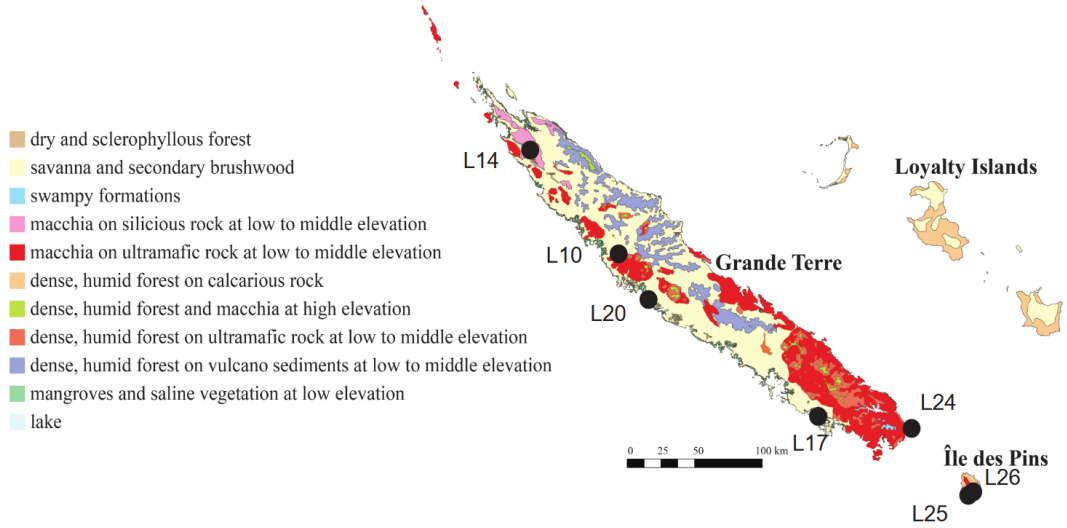

Figure 2.1.: The geographical locations of the sampling sites for the populations of the RRc group across New Caledonia. The map is modified from the vegetation types map of New Caledonia by Jaffré et al. (2012). Each point on the map represents a distinct sampling site where specimens were collected. The localities sampled include: *D. inexplorata* (L24), *D. calciphila* (L25), *D.* sp. Pic Nga (L26), *D. minimifolia* L17, *D. minimifolia* L10+20 and *D. tridentata* (L14). Numbers of sampling localities follow Paun et al. (2016). This spatial distribution map provides context for the genetic diversity and differentiation observed in the study, highlighting the specific locations of the natural populations sampled.

To exclude sites, which potentially cause technical bias, a text file was created, which could then later be used using the -sites command in ANGSD. For this task, a preliminary beagle file had to be created, in order to have information of all SNP positions, and converted to a text file encompassing all SNP positions possible. Regions of transposable elements (TE) , which were annotated in the reference genome (Emelianova unpublished), and specific scaffolds, which were problematic during the assembly of the reference genome

(TS Khastgir, personal communication), were summarized via bedtools ((Quinlan and Hall, 2010). These sites were then subtracted from the file created first, which encompassed all SNP positions. To account for linkage disequilibrium (LD), loci with less than 5,000 bp in between them were excluded. This is important to decrease the amount of false positives in further analysis. Two different text files were created in this way. For one of them, filtering was applied differently for the preliminary beagle file, in order to include monomorphic sites. This was done for the analyses with fastsimcoal2. A detailed description of the filtering procedure is included in the Figure 2.2.

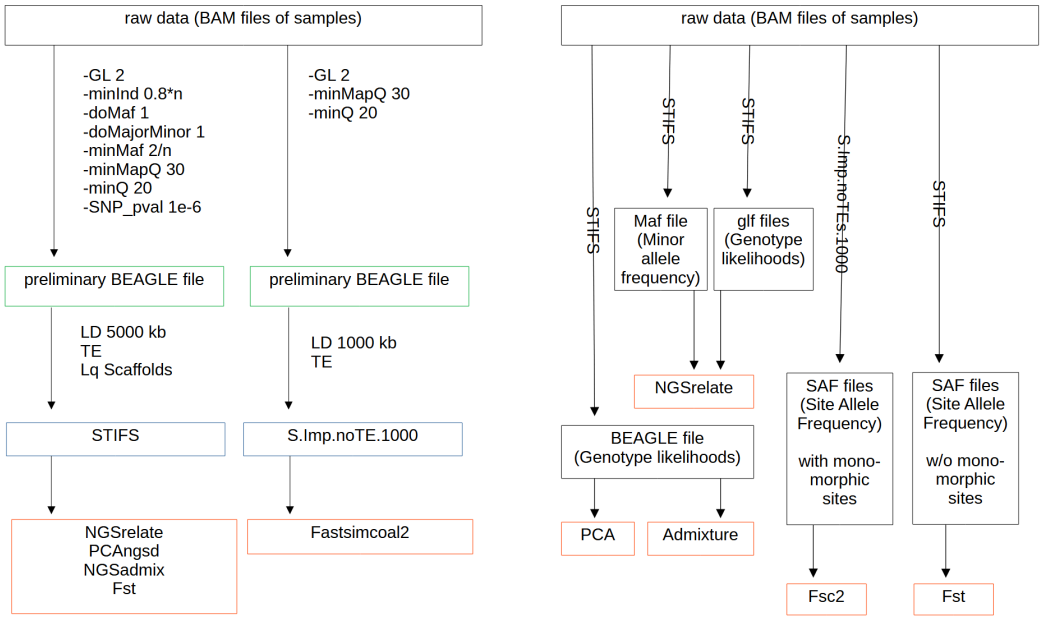

Figure 2.2.: Workflow diagrams for the population genetic analyses, with a focus on the filtering steps applied. LD, linkage disequilibrium; TE, transposable elements; Lq, low quality scaffolds.

2.3. Excluding problematic samples

The following step was to search and exclude any replicated samples of the same genotype, siblings and samples showing considerable introgression from species outside of the present sampling. Because the sampling took place on multiple occasions at the same sites, there was a risk that the same individuals were collected multiple times. This could affect the results negatively because of an overrepresentation of a single genotype. NgsRelate

was used to inspect relatedness (Hanghøj et al., 2019). The analysis was restricted to the 20 longest scaffolds of the reference for faster computation time. Various summary statistics have been calculated using this method. The selection was finally based on the relationship measurement rab (also rxy) (Hedrick and Lacy, 2015). This metric for pairwise relatedness represents the proportion of alleles shared between two individuals and is equipped for comparing inbred individuals. If the relationship measurement of two samples had a value above 0.4, one of the samples was excluded from the analysis. The decision, which of the sample to choose, was based on the number of reads available. To identify individuals with considerable introgression of other lineages, which are not part of the survey, NGSadmix (Skotte et al., 2013) and PCAngsd (Meisner and Albrechtsen, 2018) were utilized.

2.4. Population structure analysis

To characterize the genetic patterns within and between localities and species population structure was inspected. The principal component analysis is a statistical method, used to show overall patterns of variation in a data set. PCAngsd calculates a covariance matrix via principal component analysis on the basis of the BEAGLE file. The value of filtering for minor alleles (- minMaf) was set to 0.05. Additional filters including an LD pruning have been applied prior to the analyses as explained above. The resulting covariance matrix was visualized as PCA plots using Numpy (Harris et al., 2020), Matplotlib (Hunter, 2007), Seaborn (Waskom, 2021) and Scikit-learn (Pedregosa et al., 2011).

For investigating admixture between samples, NGSadmix was utilized. NGSadmix calculates the affiliation of genomic segments within a given sample to an assumed ancestral population. The number of clusters (K-values) ranged from one to nine. This range exceeds the number of putative distinctive groups. For each K, we ran the program ten times. This was done to obtain an impression on the stability of the respective splits in the data. Before running NGSadmix, the value of -minMaf was set to 0.0625, in order to filter more stringently for rare sites. The results were visualized using the software pong (Behr et al., 2016). Additionally, the Evanno method was employed to identify the K value, which captures the most structure. The method was described by Evanno et al. (2005), and is based on the calculation of the rate of change in the log probability between successive clusters (delta K). The test was executed via the Cluster Markov Packager Across K, available at https://clumpak.tau.ac.il/bestK.html (last accessed November 2023).

To have another perspective on the genetic diversity in the clade, Fst-values were calculated with the ANGSD software. The program for Fst estimates genetic differentiation based on genetic structure. For the calculation of weighted estimates, which are shown in figure 3.4, the number of samples per population is additionally considered. Fst is a commonly used parameter to estimate population divergence. For this analysis, joint site frequency spectra (JSFS) had to be calculated. In contrast to the joint site frequency spectra for the fastsimcoal2 analyses, only polymorphic sites were surveyed and the filtering steps were applied in a similar way as for the genotype likelihood file for PCAngsd

or NGSadmix (Figure 2.2). Weighted Fst values are visualized, as these are less influenced by different numbers of individuals across different populations.

2.5. Demographic inference

In order to have an impression on the demographic history of the populations fastsimcoal2 (v. 2.7) was utilized. Fastsimcoal2 (fsc2) uses a coalescent approach to calculate genetic diversity under specific circumstances stated as input for the program. Two different kinds of analysis were conducted. The first one was designed to enhance our understanding of population size changes of each population during the last 500,000 years, the second one was designed to test specific modes of speciation, investigating in particular if speciation in this group happened in the presence of gene flow. The latter approach shall provide supplemental insights into the dynamics of early speciation in our group.

To estimate the growth rate, the population sizes in the past and the present were computed. The ancestral population was set to be 70 000 generations into the past, which is approximately 500,000 years ago, considering that the generation time is estimated to be around seven years (Verdú, 2002). The input files (.est and .tpl) are shown in the code section of the supplementary. The value deltaL describes the difference between the observed and estimated maximum likelihood. The smaller this value is, the more similar the estimated SFS is to the observed SFS. If this value is relatively high, this means that the chosen model is a relatively bad fit to the observed genetic signal.

Furthermore, different models of speciation were tested on the following population pairs: *D. calciphila* - *D. inexplorata*, *D. calciphila* - *D.* sp. Pic N'ga, *D. minimifolia* L17 - *D.* sp. Pic N'ga, *D. minimifolia* L10+20 - *D. tridentata*, and *D. minimifolia* L17 - *D. minimifolia* L10+20. The models tested are: Strict Isolation (SI), Isolation with Migration (IM), Ancient Migration (AM) and Secondary Contact (SC). The exact implementations are shown in the code section in the supplementary.

As a preparation for the program fastsimcoal2 unfolded site frequency spectra and unfolded joint site frequency spectra were calculated. This was done using the software ANGSD, with its implementations -doSaf and realSFS (Mas-Sandoval et al., 2022). The reference genome *Diospyros impolita* v. 1.1 (Paun, Khastgir and Emelianova, unpublished) was used as the ancestral state, as this species is outside the target group of this thesis. A separate file for the -sites command was created. Again, the information of TE positions in the reference file was utilized and these regions were excluded as above. Additionally, the effects of linkage disequilibrium were mitigated by using only sites with at least 1,000 bp between them. The quality filters for the SAF files were set as for the creation of the BEAGLE files (-minMapQ 30 and -minQ 20).

The site frequency spectra and joint site frequency spectra were formatted to meet the specific requirements of fsc2. The number of fixed derived alleles across all individuals (last value of matrix) was set to zero and added to the value to the first value of matrix, since these values are describing the same state(recommendation by the authors, see https://groups.google.com/g/fastsimcoal/c/3-a5sz8-q94/m/TWx-dafRMwAJ (last accessed May. 2024)). Models for individual populations and population pairs were

developed as .tpl and .est files. The value for the mutation rate was set to the same value expected for Oryza sativa (Ammiraju et al., 2008), 7e-9 following previous studies (Paun et al., 2016). The recombination rate was set to zero. The commands were the same for all models and 100 runs were executed for each of them. The number of calculations (-n) was set to 500,000, and the number of loops (-L) was set to 80. We further used the commands "-d" to compute the derived SFS, "-M" to estimate parameters between iterations using the maximum likelihood of the derived SFS and "-0" to ignore monomorphic sites from the maximum likelihood estimation between iterations. This latter step was done to have more consistent results at the expense of having more robust absolute values for the parameters. Furthermore, we used the command "-C 5" to ignore entries in joint frequency spectra with values below 5 from maximum likelihood estimation for all two population models.

After running fastsimcoal2, the best run of each model was selected based on the MaxLhood value. Of these best runs, the AIC values were calculated using a script by Joana Meier, available at
https://github.com/speciationgenomics/scripts/blob/master/calculateAIC.sh (last accessed May, 2024) and compared. Of the best fitting model of each population pair, a summary table of the parameters of all runs was created, and ordered based on the likelihood value. The averages, maximums and minimum values of the ten best runs were calculated.

3. Results

The number of pairs of reads ranged between 59,960,279 and 193,445,380 per accession. The average coverage per accession was 28.9. Out of the 59 samples, three samples of the same individual (minimifolia1021a, minimifolia1021b and minimifolia1021c), and two potential sibling pairs (tridentataBT206 - tridentata1078 and spPicNgaBT318 - spPicNga1191) could be identified using NgsRelate and were excluded from further analyses (Table A.1. Two of the *D. minimifolia* L10+20 samples (minimifoliaBT269 and minimifoliaBT270) showed clear introgression from D. pustulata in the sympatric locality L20, while another one (minimifoliaBT267) was identified as part of the mentioned D. pustulata population (Figure A.1). Further, one sample of *D. calciphila* (calciphilaYP124) and *D. tridentata* (tridentataJM6697) appeared to have significant introgression from populations not analyzed in this study, as indicated by the principal component analysis and the admixture analysis (Figure A.2). However as the latter two accessions had much lower coverage than the rest of accessions, further investigations would be required to certify their exact genetic makeup.

3.1. Population structure

The total number of sites analyzed was 1 775 871 948. After filtering with the above mentioned quality thresholds, as well as for transposable elements, problematic scaffolds and linkage disequilibrium, 113 287 sites were retained. If filtering would have been applied for quality parameters only, 204 206 833 sites would have been retained.

In the PCA results (Figure 3.1) PC1 and PC2 together explain around 60% of the variation in the dataset. The clustering of accessions suggests distinct groupings and separations among the groups, reflecting the genetic variation captured by the principal components. The southern populations (*D.* sp. Pic N'ga, *D. calciphila* and *D. inexplorata*) and *D. minimifolia* L17 are clustering into a relatively small area of the PCA.

In regards to PC1, it seems that *D. tridentata* is relatively distinct from the other populations and it clusters in a relatively wide area. The accessions of *D. minimifolia* L10+20 show a sub-clustering into different subgroups. *Diospyros minimifolia* L17 is closest to the southern populations in regards to PC1, followed by its proximity to *D. minimifolia* L10+20. The close relation to *D. minimifolia* L10+20 seems to be remarkable in respect to PC2, but is strongly contrasted by PC3.

The patterns observed in the PCA are reflected also in the NGSadmix results. The respective figures depict K=1 to K=9 (Figure 3.2) and two times K=4 and K=5 (Figure 3.3). For K values above five which generated more than two outcomes, only the major pattern is shown, since the minor solutions are less stable results. To test how many

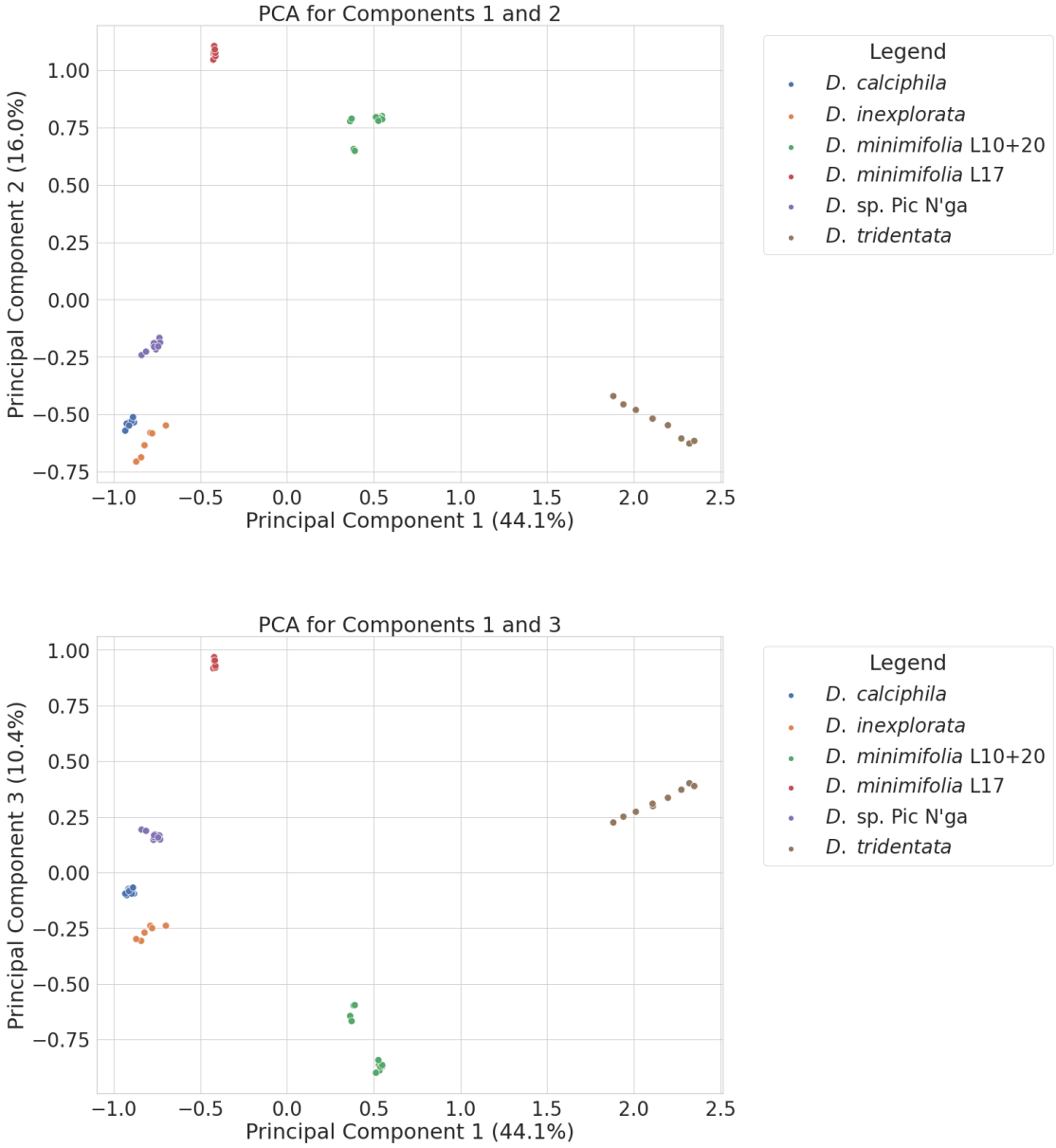

Figure 3.1.: Principal component analysis (PCA) of the accessions analysed. Each point represents an individual sample. The populations are color-coded: blue: *D. calciphila*, orange: *D. inexplorata*, purple: *D.* sp. Pic N'ga, red: *D. minimifolia* L17, green: *D. minimifolia* L10+20 and brown: *D. tridentata*. The amount of variation explained by each axis is shown in brackets.

clusters capture the most structure in our dataset, the Evanno method was employed. The optimal K by Evanno was calculated to be K=2 (Figure A.4).

The admixture plots from K=1 to K=4 show only one way to split for all ten calculations. The first split (K=2) occurs between the populations *D. tridentata* and *D.* sp. Pic N'ga, *D. calciphila* and *D. inexplorata*, and shows admixture in some samples of *D. minimifolia* L10+20 and all of *D. minimifolia* L17. Interestingly, *D. minimifolia* L17 is part of the southern group for the most part of its genome, and *D. minimifolia* L10+20 mostly grouped with *D. tridentata*. The second split (K=3), splits *D. tridentata* from the rest with small proportions of the genomes in some *D. minimifolia* L10+20 samples. Most of the *D. minimifolia* L10+20 samples and all of *D. minimifolia* L17 samples are in the same group. *D.* sp. Pic N'ga shows minor admixture with this group, but is part of the southern populations for the vast majority of its genome. K=3, in contrast to K=2, highlights a unity of *D. minimifolia* populations, which is not found using any K value other than that, except for the minor patterns observed in two runs for K=4.

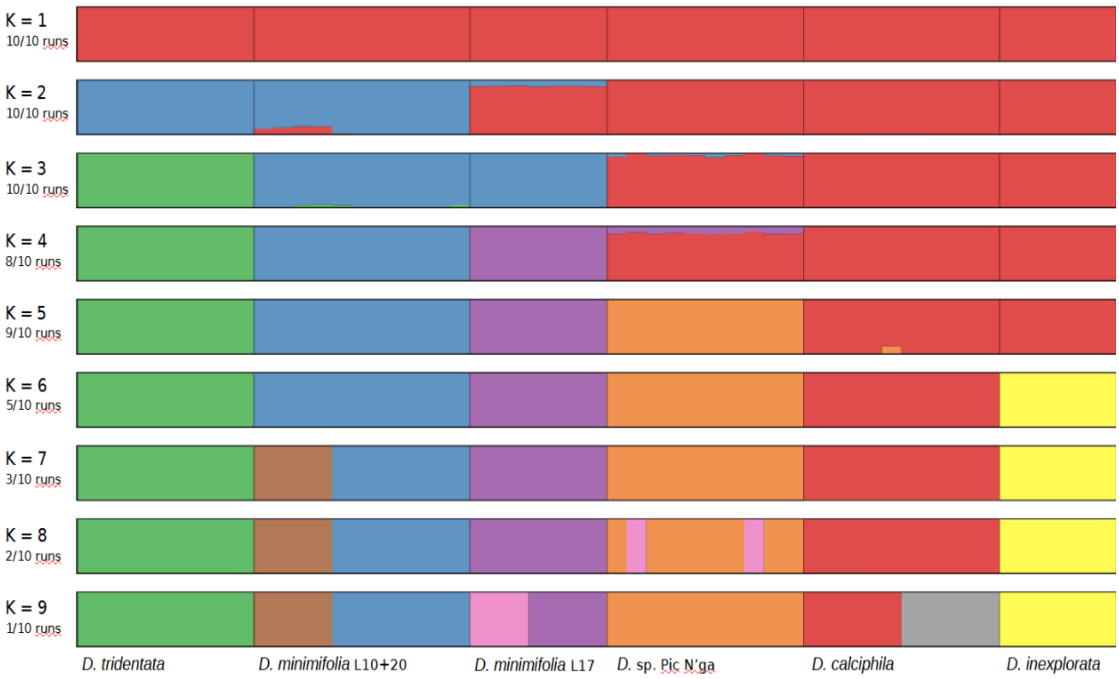

Figure 3.2.: Admixture analysis of populations in the RRc group calculated using NGSadmix and visualized with pong. Each vertical bar represents an individual sample, and the different colors within each bar denote the proportion of the individual's genome assigned to each genetic cluster (K). The analysis was performed for K values ranging from 1 to 9. This plot provides insights into the genetic diversity and admixture patterns among the analyzed *Diospyros* samples.

In the visualization of K=4, two different results are present (Figure 3.3). In one of them, *D. minimifolia* L17 forms a separate group, with part of the genome of *D.* sp. Pic N'ga. *D. tridentata* and separately *D. minimifolia* L10+20 form independent groups,

whereas *D. calciphila* and *D. inexplorata* cluster together; all these three groups show no admixture. In the other visualization of K=4, which resulted in two of the ten runs, *D.* sp. Pic N'ga forms a separate group, with a small amount of one *D. minimifolia* L17 genome and a small amount of one *D. calciphila* genome. *D. minimifolia* L10+20 clusters together with *D. minimifolia* L17, and shows little admixture with *D. tridentata* in a few samples. The results of K=5 are again shown in two different admixture plots (Figure 3.3). In this case, one of the ten runs created a different split, showing almost all populations/species as well separated, including a split within *D. minimifolia* according to the two localities L10+20, but not *D. calciphila* + *D. inexplorata*. Admixture is found in all *D.* sp. Pic N'ga accessions, which share a minor proportion of their genome with *D. minimifolia* L17. The more often occurring result for K=5 shows again well separated populations except *D. calciphila* and *D. inexplorata*. Only one of the *D. calciphila* samples shares a small part of its alleles with *D.* sp. Pic N'ga. The genomic data of the populations of *D. inexplorata* and *D. calciphila* are only split at K=6 and beyond.

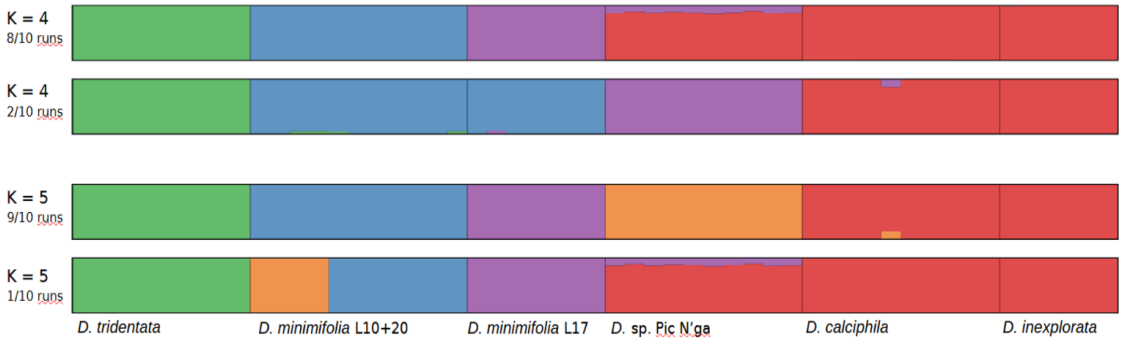

Figure 3.3.: Admixture analysis of populations in the RRc group calculated using NGSadmix and visualized with pong. Each vertical bar represents an individual sample, and the different colors within each bar denote the proportion of the individual's genome assigned to each genetic cluster (K). The analysis was performed for K values ranging from 1 to 9. This plot provides insights into the genetic diversity and admixture patterns among the analyzed *Diospyros* samples.

The weighted Fst estimates are inline with the results from the ones above in most regards. *D. inexplorata* and *D. calciphila* are shown to be very similar, as well as *D. calciphila* and *D.* sp. Pic N'ga. *D. minimifolia* L17 to *D.* sp. Pic N'ga and *D. minimifolia* L17 to *D. minimifolia* L10+20 have similar values. Interestingly, *D. minimifolia* L10+20 shows slightly lower numbers to the southern populations as to *D. minimifolia* L17. Contrary to expectation, the difference between *D. minimifolia* L17 and *D.* sp. Pic N'ga is estimated to be higher as between *D. minimifolia* L10+20 and *D.* sp. Pic N'ga. *D. tridentata* has the highest value with *D. minimifolia* L17, followed by *D. inexplorata*.

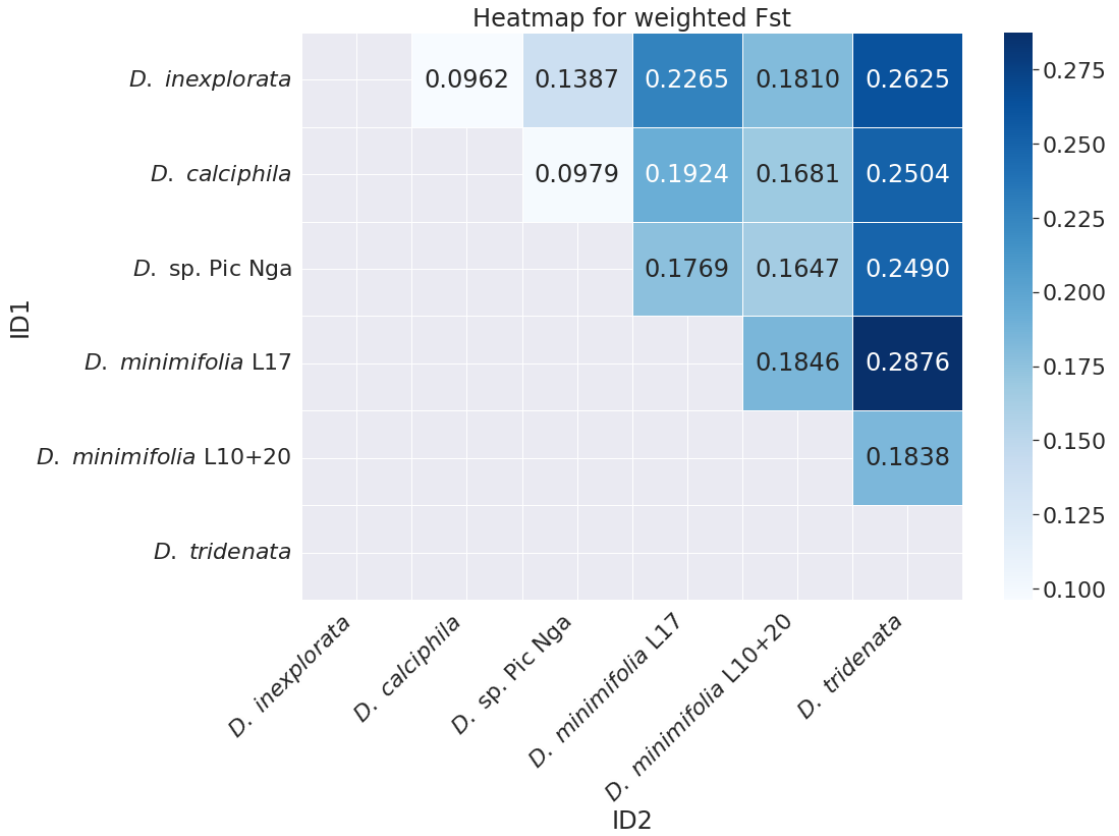

Figure 3.4.: Heatmap of pairwise weighted Fst values among all populations. Higher values represent higher genetic differentiation, shown with darker shade. The program for Fst estimates genetic differentiation based on genetic structure. For the calculation of shown weighted estimates, the number of samples per population is additionally considered.

3.2. Demographic inference - population size change

First, site frequency spectra (SFS) were used to estimate a constant rate of change of effective population size for each population via single population models simulations in fastsimcoal2. The inferred parameters are shown in Table 3.1. NANC is the parameter for the ancestral population size 70 000 generations ago. NCUR is the parameter for the current population size. R1 describes the rate of change backwards in time. It is calculated by dividing the logarithm of NANC/NCUR by 10 000. More details about the model setup are shown in the supplementary section.

The best fitting SFS were created by the program under the specific model circumstances for *D. inexplorata*, which shows a value for deltaL of around 23.5. The population for which this simple growth model is a relatively poor fit is *D. minimifolia* L10+20, followed by *D.* sp Pic N'ga. *D. inexplorata* is estimated to have experienced a decrease in effective

population size. A similar rate of decrease is also shown in *D. minimifolia* L17 and, to a lesser degree in *D. calciphila*. *D. tridentata* shows a remarkably small population size and a positive growth rate forward in time.

Table 3.1.: Results of population size and growth rate estimations using fastsimcoal2. The analysis was based on the following parameters: NANC: Ancestral population size (estimated to be 70,000 generations into the past), NCUR: Current population size, R1: Growth rate. deltaL describes the log-likelihood difference for the parameter estimates. For each population, the upper row shows the average (AVG), the middle row the smallest (MIN), and the lowest row the highest values (MAX), of the ten best runs, respectively.

Population	NCUR	NANC	R1	deltaL
D. calciphila				
AVG	85247.2	325009.5	0.0001337964	80.10
MIN	82000	304225	0.000124082	80.30
MAX	92004	363677	0.000147933	79.64
D. inexplorata				
AVG	35763.9	209365.7	0.0001766515	23.56
MIN	35150	198817	0.000173086	23.60
MAX	36420	224425	0.000183474	23.48
D. sp Pic N'ga				
AVG	101702.7	80520.1	-0.00002281931	167.42
MIN	71754	62097	-0.0000279988	177.76
MAX	127518	98694	-0.0000144544	162.77
D. minimifolia L17				
AVG	58845.2	298557.9	0.0001621814	61.96
MIN	56668	268543	0.000151378	62.15
MAX	61728	336482	0.000172678	61.66
D. minimifolia L10+20				
AVG	170813.7	111778.6	-0.00004133671	188.17
MIN	95913	65408	-0.0000496248	199.42
MAX	237416	144541	-0.0000315403	173.50
D. tridentata				
AVG	2252.8	529.9	-0.0001637534	69.20
MIN	1603	175	-0.000221751	69.58
MAX	3235	1148	-0.00010361	68.74

3.3. Demographic inference - speciation modes

In order to choose the best fitting model of speciation, AIC values were calculated. This parameter enables the comparison of likelihood values of different models, thus it incorporates the amount of parameters used for the respective models. Table 3.2 shows the best run of each model for each population pair. The rightest column shows a ranking for the respective population pair based on the AIC score.

The table of the best runs for each model of speciation for each population pair (Table 3.2) enables the comparison between the models of speciation for each population pair. Secondary contact was the best option for the population pairs *D. calciphila* - *D.* sp. Pic N'ga and *D. calciphila* - *D. inexplorata*. Isolation with migration was the best fitting model for all other population pairs. Inferred migration rates were always low, and often asymmetric. The differences of the AIC values of the best model to the second best model vary between the population pairs. Sometimes, the difference is marginal as in the case of *D. minimifolia* L17 and *D. minimifolia* L10+20, whereas other times, the best fitting model is the best by much as in the case of *D.* sp. Pic N'ga and *D. calciphila*. To have a more reliable impression of the estimated values and their ranges, Table 3.3 shows the average, maximum and minimum parameter values of the ten best runs of the best model for each respective population pair. The analysis of the population pair of *D. calciphila* and *D.* sp. Pic N'ga suggests that these populations divided between a relatively wide range of generations. The probabilities of alleles to be transferred between populations varies in the population pair of *D. calciphila* and *D.* sp. Pic N'ga to a lesser degree than in the population pair of *D. calciphila* and *D. inexplorata* or other population pairs. The estimations of the population size of *D.* sp. Pic N'ga are more consistent in the population pair with *D. calciphila* than with *D. minimifolia* L17. The analysis of the population pair of *D. calciphila* and *D. inexplorata* suggests a split between a wide range of generations. In the case of *D. tridentata* - *D. minimifolia* L10+L20, the population size of *D. tridentata* is estimated to be small, which is in line with the prior results of calculating the growth rate. The population size calculated for *D. minimifolia* L10+20 spans hundredfold from minimum to maximum value in both of the population pairs calculated.

Table 3.2.: fastsimcoal2 results of two population models: Ancient Migration (AM), Isolation with Migration (IM), Secondary Contact (SC), and Strict Isolation (SI). SPNGA, CALC, INEX, MINI12, MINI3, TRID are abbreviations for *D.* sp. Pic N'ga, *D. calciphila*, *D. inexplorata*, *D. minimifolia* L10+20, *D. minimifolia* L17 and *D. tridentata*, respectively. The other parameter abbreviations: NANC: Ancestral effective population size, MIGMX: Probability of an allele being transferred from the first population to the second, MIGXM: Probability of an allele being transferred from the second population to the first, TDIV: Estimated time of population split (in generations), CHANGM: Time when migration stops in ancient migration models or starts in secondary contact models (in generations), AIC: Akaike Information Criterion value, Ranking: Model ranking based on AIC values, with 1 being the best fit.

	SPNGA	CALC	NANC	MIGMX	MIGXM	TDIV	CHANGM	AIC	Ranking
AM	333181	144348	354740	7.45863e-07	8.33650e-05	23845	13335	314221.37	3
IM	112785	46572	101247	1.64039e-05	5.21944e-05	52796		313147.05	2
SC	52046	11565	434	1.58187e-05	2.53236e-04	138567	12456	312703.71	1
SI	619013	305299	655996		37173		314273.44	4	

	CALC	INEX	NANC	MIGMX	MIGXM	TDIV	CHANGM	AIC	Ranking
AM	51085	22521	132473	5.05532e-07	5.68659e-09	3122	1860	242516.16	4
IM	78332	35448	190443	5.47206e-06	2.55640e-08	4970		242515.73	3
SC	15690	2929	705	9.10E-05	8.64E-04	45437	3607	242006.35	1
SI	76156	33926	193911		4682		242512.20	2	

	TRID	MINI12	NANC	MIGMX	MIGXM	TDIV	CHANGM	AIC	Ranking
AM	226617	1144038	668626	4.87436e-04	9.59349e-05	187367	94784	301509.21	3
IM	4224	28445	14338	1.68580e-04	1.16874e-05	8094		301060.95	1
SC	10232	80362	40602	0.0000739346	0.00000407177	25110	12301	301182.15	2
SI	29798	138953	87512		12985		302001.83	4	

	MINI3	SPNGA	NANC	MIGMX	MIGXM	TDIV	CHANGM	AIC	Ranking
AM	24772	50245	48653	1.78825e-04	5.56859e-05	10706	5410	288456.75	3
IM	240033	453633	394362	2.83790e-06	1.88300e-06	175720		287586.99	1
SC	5657	9999	8580	0.000130731	0.0000993783	3314	1614	287658.54	2
SI	23363	46163	44754		6609		288685.54	4	

	MINI12	MINI3	NANC	MIGMX	MIGXM	TDIV	CHANGM	AIC	Ranking
AM	380946	123662	225997	0.000226191	0.000714591	85554	43149	296997.63	3
IM	32190	10195	17786	0.000019376	0.0000648152	9245		296121.57	1
SC	13524	4354	7211	0.0000550425	0.000159083	3645	1808	296181.44	2
SI	600338	202797	355737		72209		297389.85	4	

Table 3.3.: Best model results for each pairwise population analysis using fastsimcoal2. The parameters estimated include NANC: Ancestral population size, MIGMX: Probability of an allele being transferred from the first population to the second, MIGXM: Probability of an allele being transferred from the second population to the first, TDIV: Estimated time of population split (in generations) and in some cases CHANGM: Time when migration stops in ancient migration models or starts in secondary contact models (in generations). The models shown are Secondary Contact (SC) or Isolation with Migration (IM). SPNGA, CALC, INEX, MINI12, MINI3, TRID are abbreviations for the populations *D.* sp. Pic N'ga, *D. calciphila*, *D. inexplorata*, *D. minimifolia* L10+20, *D. minimifolia* L17 and *D. tridentata*, respectively. For each population pair, the table presents the average (avg), maximum (max), and minimum (min) values of the best runs for each parameter.

SC								
		SPNGA	CALC	NANC	MIGMX	MIGXM	TDIV	CHANGM
	avg	30380.3	6237	1241.1	0.00003236582	0.0005260732	77027.8	6785.4
	max	52046	11565	3361	0.0000501022	0.00081733	138567	12456
	min	15309	3587	152	0.0000158187	0.000253236	39258	4128
SC								
		CALC	INEX	NANC	MIGMX	MIGXM	TDIV	CHANGM
	avg	136291.1	57497.2	303030.5	0.0000522760463	0.00009592342175	43557.9	5273.9
	max	428199	193633	1094633	0.000196809	0.000863591	313191	41543
	min	15690	2929	705	0.000000229281	0.0000000116841	1618	1
IM								
		TRID	MINI12	NANC	MIGMX	MIGXM	TDIV	
	avg	17087.4	110803.4	57534.3	0.000234409962	0.0000174627541	30521.3	
	max	78909	511719	272254	0.000675517	0.0000459797	142288	
	min	1051	7199	3663	0.00000909162	0.000000642271	1895	
IM								
		MINI3	SPNGA	NANC	MIGMX	MIGXM	TDIV	
	avg	51462.6	97834.8	84155.5	0.000091643685	0.000057013246	38475.8	
	max	240033	453633	394362	0.000298685	0.000184433	175720	
	min	2471	4791	3962	0.0000028379	0.000001883	1890	
IM								
		MINI12	MINI3	NANC	MIGMX	MIGXM	TDIV	
	avg	69126.4	20212.6	36200.9	0.000028257792	0.000088134363	19990.9	
	max	330919	93256	167858	0.0000670433	0.00021676	96910	
	min	9250	2995	5163	0.00000177596	0.00000766383	2616	

4. Discussion

The six populations of this survey are part of a recent adaptive radiation, and not all of them can be assigned to distinct species. As described in the introduction, the taxonomic situation within the RRc group is nuanced. Additionally, it is important to note that the samples are derived from specific locations and not the entire distribution range of the respective species. This has to be considered, as population structure within species potentially influences these results. This is critical for all analyses employed, such as the estimation of parameters with fastsimcoal2. This program is based on the coalescent theory and therefore assumes that population structure within analyzed groups is not present. Species of *Diospyros* are commonly dioecious (Samuel et al., 2019; Pomper et al., 2020). In the genus Ficus, dioecy is generally associated with lower genetic spatial structure compared to monoecy, however further research is needed to test the extent of genetic spatial structure for the clades in this analysis (Nazareno et al., 2013; Zhou and Chen, 2010).

For example, the species *D. tridentata* is known to have a very narrow distribution in the north of the mainland of New Caledonia in contrast to *D. calciphila*, which is relatively widely distributed, including the Loyalty islands North-East of the mainland (White and Vink, 1993). The plots of the principal component analysis shows that the populations *D. inexplorata* and *D. calciphila* are clustering separately, but in close proximity to each other. It seems reasonable to assume that other populations of *D. calciphila* from the Loyalty islands, would cluster separately in close proximity to the shown population as well. This provokes the question, how distinct the populations *D.* sp. Pic N'ga and *D. inexplorata* are from *D. calciphila* if more of its distribution range would have been covered. It can be assumed that the robustness and accuracy of demographic inference and population structure analysis benefit from adequate sampling sizes per population. Ma and Amos (2012) describe how sample size influences patterns in eigenvector-plots of principal component analysis. Moreover, they argue that relative sample sizes of populations and admixed individuals strongly affect the outcome, and thereby complicate interpretations.

New Caledonia has exceptionally heterogeneous habitats due to its geologic history. This heterogeneity of habitats appears not only spatially but also on the temporal scale, as sea level oscillations well over 100m occurred at least four times during the pleistocene, affecting the biodiversity on the island ((Isnard et al., 2016; Pillon et al., 2020; Pintaud et al., 2001; Morat, 1993; Wulff et al., 2013). Adaptive introgression, which results from hybridization and back-crossing, can lead to the transfer of few neutral alleles, but also segregating blocks of different adaptive systems. This latter possibility can result in the emergence of new adaptive traits as discussed by Caujapé-Castells et al. (2017). Gene flow can also cause the loss of unique alleles and local adaptations if selection is not strong

enough (Tigano and Friesen, 2016). This raises the question whether geneflow was an important factor in the colonization of the island by *Diospyros*, or conversely, a limiting factor.

The principal component analysis (Fig. 3.1) was used to visualize overall variation in the data set, and to have an impression of the overall genetic diversity within and between the groups of interest. PC1 does not discriminate between *D. inexplorata*, *D.* sp. Pic N'ga and *D. calciphila*, and shows that *D. minimifolia* L17 is genetically similar to this group of populations. This axis is the most important one to discuss, since it explains around 44% of the variation in the dataset. A clustering pattern for *D. minimifolia* L17 between the southern populations and *D. minimifolia* L10+20 was expected following the hypothesis that it originated by hybridization of these groups. Interestingly, *D. minimifolia* L17 is much closer to the southern populations. In regards to PC2, its closest neighbor is *D. minimifolia* L10+20, and there is much more distance to all other populations. PC3 shows the opposite, that is that *D. minimifolia* L17 and *D. minimifolia* L10+20 are spread far apart from each other. If *D. minimifolia* L17 originated by a simple scenario of recent hybrid speciation, it would be expected that its samples are found between the samples of the parent species. Based on this figure, it seems that the evolutionary history of *D. minimifolia* L17 is more complicated. This might be the case because the hybrid speciation event happened a long time ago, resulting in the accumulation of many distinctive SNPs in *D. minimifolia* L17, complicating today's genetic signal. There is substructure in *D. minimifolia* L10 + 20, possibly discriminating between the two populations, as well as individuals with recently introgressed regions from *D. pustulata* (Figure A.1). *D. tridentata* seems to account for a lot of the heterogeneity, since its samples are spread out remarkably, in regards to all shown axes. This pattern possibly results from intraspecific variation. This is especially interesting since *D. tridentata* counts as a threatened species, due to its small population size and narrow distribution (IUCN Red list, https://www.iucnredlist.org/species/173692/156963006#assessment-information). The results indicate a notable genetic similarity between southern populations, and a distinctness of *D. tridentata* and *D. minimifolia* L10+20 and to a lesser degree *D. minimifolia* L17.

To have another perspective on the genomic landscape of the populations NGSadmix was employed. The figures of NGSadmix (Figure 3.2 and Figure 3.3) shows that *D. tridentata* is the most distinct population, since it is the first population recognized as a unique group. Furthermore, it is shown that it is most similar to *D. minimifolia* L10+20. This is in line with the results of the principal component analysis and the prior analysis of Paun et al. (2016). The phylogenetic closeness of the southern populations, *D. inexplorata*, *D. calciphila* and *D.* sp. Pic N'ga, is also reflected as in the PCA results. However, *D.* sp. Pic N'ga splits from the other two populations completely, or shares parts of its genome with *D. minimifolia* at K=3 and onward. *D. minimifolia* L17 shares, similar to *D.* sp. Pic N'ga, segments with different populations until a certain value of K is surpassed.

If all populations surveyed would have originated in regular and sufficiently long time intervals through cladogenesis under strict isolation consequently, we would see one population split of the rest of the populations for each K value added. This would be the

case under the assumption that variation accumulates continuously through mutations and is not affected by selection, genetic drift, gene flow or incomplete lineage sorting (ILS). This consequent "splitting off" would not happen if cladogenesis would occur in parallel in different lineages. The latter case would result in a former splitting of populations into two groups. Each odd K value would show splits inside the genomes of the samples of populations, because clean grouping into an odd number of populations wouldn't be possible. Again, provided that cladogenesis happens after regular and sufficiently long time intervals and without effects altering variation other than continuous mutations. If we view our results under such a simplified scenario, we could argue that the first cladogenetic event occurred between the ancestor populations of *D. tridentata* + *D. minimifolia* L10+20 and *D. minimifolia* L17 + the southern populations. This split also corresponds to phylogenetic patterns described by Paun et al. (2016). The admixed genomes of *D. minimifolia* L17 and *D. minimifolia* L10+L20, and depicted admixture patterns in general, could be caused by introgression or ILS (Yan et al., 2021). ILS describes the persistence of ancient variation through speciation events, which appears as shared polymorphisms across different clades. These polymorphisms are an important biological bias in various phylogenetic analysis, especially in regards to adaptive radiations (Suh et al., 2015).

For K=3, *D. tridentata* forms a separate group with very small parts of the genomes of *D. minimifolia* L10+20. The two other groups form between the *D. minimifolia* populations and small parts of the genome of *D.* sp. Pic N'ga and the rest of the southern population genomes. Again, we see that a simplified view like the one explained above does not capture the nuances shown in the representation by NGSadmix. Similar difficulties arise for all other K values. The analysis of admixture therefore cannot support the idea that the present genetic diversity of our groups originated by cladogenesis under the proposed simplified conditions. Whether this is caused by the complexity of demographic events or biases caused for example by ILS or undetected paralogy, remains unclear (Yan et al., 2021; Scornavacca and Galtier, 2016).

Overall, the results of NGSadmix suggest that the southern populations are closely related, that *D.* sp. Pic N'ga and *D. minimifolia* L17 share similarities and that *D. tridentata* is the most distinct population. *D. minimifolia* L17 is significantly distinct from all other populations but shares similarities with *D. minimifolia* L10+20 and *D.* sp. Pic N'ga. To improve the results of NGSadmix, increased sample size could be considered, since salient variation affects the outcome of admixture analysis significantly (Lawson et al., 2018).

A similar but slightly different outcome was produced by the ANGSD implementation for calculating pairwise Fst. The least amount of differentiation can be found between the southern populations. The lowest value can be found for the pair *D. calciphila* and *D. inexplorata*, followed by the pair *D.* sp. Pic N'ga and *D. calciphila*. This suggests that the southern populations diverged the most recently, or alternatively have the most amount of gene flow with genome-wide introgression patterns. Surprisingly, the population *D. minimifolia* L10+20 shows to be less differentiated from the southern populations, than *D. minimifolia* L17. The lowest level of differentiation for *D. minimifolia* L17 can be

found in relation to *D.* sp. Pic N'ga, followed by *D. minimifolia* L10+20. All values of *D. minimifolia* L17 are relatively high compared with the other populations, only surpassed by *D. tridentata*, which is shown to be the most differentiated from the rest. The highest level of genetic differentiation is found between *D. tridentata* and *D. minimifolia* L17. These results are, similar as the other results from population structure analysis, only vaguely supporting the assumption that *D. minimifolia* L17 originated by hybrid speciation alone.

It also seems likely that a bifurcating phylogenetic tree might be a drastic simplification, missing out on the nuances of diversification within the RRc clade. The displayed relationships between the populations suggest that factors like selection, introgression and/or ILS have contributed to the genetic diversity, obscuring observed genetic signals.

To test modes of speciation and to gain insights into demographic factors with single and two population models, fsc2 was utilized. Several preliminary analyses showed that focusing on polymorphic sites by using the "-0" option provides more consistent results, possibly because estimated SFS/jSFS fitted better to the observed data. However, it should be noted that the absolute values for parameters are less reliable using this approach (Excoffier et al., 2013). The first table shown of the analysis with fsc2 (Table 3.1) provides the results for the analysis of growth rate in a single and isolated population. The smallest value for deltaL was estimated for the population of *D. inexplorata*. However, its sample size is relatively low with only six samples, which is critically small for this analysis (Excoffier et al., 2013). *D. tridentata* has a low value for deltaL too and the estimated population size is remarkably low, similar as in the estimates shown in the other tables, indicating that conservation efforts targeted at this species are of great importance.

The second table (Table 3.2) of fsc2 results shows a comparison of different models for each pair of populations. The two most common modes of speciation are secondary contact and isolation with migration, with a low and asymmetric gene flow rate estimation for all comparisons. This suggests that gene flow likely took place during or after diversifications in this clade and possibly still is. The fact that some of the groups are adapted to different ecological niches than others raises the question, how much gene flow actually occurred between an adapting population to its sister clade. As explained above, gene flow can have positive effects on adaptive processes, because of adaptive introgression, but conversely also counteract local adaptations if selection is weak (Tigano and Friesen, 2016).

The third table of fsc2 (Table 3.3) results provides insights into the range of values for the parameters and shows that the parameter estimations are in many cases very variable, making an interpretation difficult. The estimations for the parameters of *D.* sp. Pic N'ga and *D. calciphila* are relatively stable. The migration rate is estimated to be much higher from *D.* sp. Pic N'ga to *D. calciphila* than vice versa. Assuming that an adaptation to limestone is ancestral to both, and that *D.* sp. Pic N'ga adapted to ultramafic soil after these populations split, we could argue that populations adapted to these harsh edaphic conditions have a limited capacity to integrate genetic segments from populations non-adapted to these conditions. On the contrary, the populations adapted to more benign conditions could be more susceptible to introgression, because alleles, which

are needed for harsh environments like ultramafic soils, might interfere less negatively than the other way around. However, it is critical to consider that asymmetrical gene flow is also influenced by physical properties. Considering that *D.* sp. Pic N'ga is situated at a higher elevation than *D. calciphila*, but in close proximity, it seems plausible that gravity plays an important role in this situation too.

D. minimifolia L17 shows high variability in all parameters in relation to *D.* sp. Pic N'ga and *D. minimifolia* L10+20. This makes it difficult to draw any conclusion in regards to the hypothesis of an origin of *D. minimifolia* L17 by hybrid speciation, without testing and comparing models composed of three populations, which encompass hybrid speciation scenarios. However, the results suggest that migration and gene flow were likely occurring during or after divergence and are likely still ongoing. It seems plausible that tested demographic scenarios are a vast simplification of a network-like phylogeny made possible by incomplete reproductive isolation. To imagine the evolutionary history of this clade as a bifurcating tree is possibly missing out on essential drivers during the diversification in this group. Introgression might be especially profound in our groups because of a correlation with a smaller number of chromosomes and a higher proportion of transposable elements in the genomes (Cannon, 2021; Samuel et al., 2019).

Our results cannot verify or falsify if a hybrid speciation event lead to the formation of *D. minimifolia* L17. This population however proved to be significantly distinct from other populations on the genetic level. It is therefore advisable to describe this population as a new species, in order to protect its genetic uniqueness. This study therewith exemplifies the importance of investigating genetic diversity in nature conservation. If morphological features cannot discriminate the population from *D. minimifolia* L10+20, a taxonomic revision could potentially lead to the description of it as a novel cryptic species (Korshunova et al., 2019).

To enhance our understanding of evolution during rapid adaptive radiations, future research could focus on the extent of porosity of genomes and test if it correlates with time of divergence (Kane et al., 2009). Furthermore, the identification of regions with decreased introgression might be important to understand how species- and biodiversity is maintained during the presence of gene flow. Moreover, a better understanding of the reproductive biology of the entire genus on the island, for example targeted at the pollination syndrome variation, could also shed light on targeted exchange of alleles between clades. However, it is important to consider that correlations with genetic markers might be difficult to test, since polygenic routes of adaptation potentially obscure genetic signals (Sendell-Price et al., 2021).

Studies targeted at evolutionary biology benefit from the general settings of oceanic islands as they typically have clear boundaries, are relatively small compared to continental systems and their age is relatively easy to determine. Moreover, biodiversity on islands is often threatened by climate change, because of physical barriers limiting migration (Linan et al., 2021). In such scenarios, reproductive isolation might be an important disadvantage, especially for long lived organisms (Caujapé-Castells et al., 2017). Conservation efforts might benefit from enabling gene flow to increase genetic diversity and ensure better evolvability. However, any endeavors need to be thought through critically beforehand, to

ensure that the potential loss of biodiversity due to genetic homogenization is not more severe. An example project, which turned out successfully heighten genetic diversity, was carried out to save a population of florida panthers (Puma concolor coryi) from negative effects of inbreeding, by enabling gene flow with Pumas from Texas (Puma concolor stanleyana) (van de Kerk et al., 2019).

Insights from conservation efforts like these exemplify the importance of gene flow for the maintenance of genetic diversity. Adaptive radiations, especially ones occurring on oceanic islands, are remarkable systems to study, enabling us to better understand how novel ecological niches become exploited and how species develop and diversify. Similar to other diversifying groups like *Quercus* sect. *Protobalanus* ((Trel.) A. Camus) our results demonstrate a complex situation with hybridizations and speciation as a continuum with diffuse limits (Ortego et al., 2018).

Rapid adaptive radiations propose a tremendous challenge to modern population genetic tools due to the multitude of effects on the genomic landscapes and the small amount of time during diversifications. Adequate methods can be found not only in genomics, but also in modeling niche filling. Applying more unified frameworks could potentially help comparing different systems undergoing rapid adaptive radiations (Soulebeau et al., 2015).

Bibliography

Arhat Abzhanov. Darwin's galapagos finches in modern biology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1543):1001–1007, 2010. doi: 10.1098/rstb.2009.0321.

Jetty S.S. Ammiraju, Fei Lu, Abhijit Sanyal, Yeisoo Yu, Xiang Song, Ning Jiang, Ana Clara Pontaroli, Teri Rambo, Jennifer Currie, Kristi Collura, Jayson Talag, Chuanzhu Fan, Jose Luis Goicoechea, Andrea Zuccolo, Jinfeng Chen, Jeffrey L. Bennetzen, Mingsheng Chen, Scott Jackson, and Rod A. Wing. Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *The Plant Cell*, 20(12):3191–3209, 2008. doi: 10.1105/tpc.108.063727.

Aaron A. Behr, Katherine Z. Liu, Gracie Liu-Fang, Priyanka Nakka, and Sohini Ramachandran. Pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, 32(18):2817–2823, 2016. doi: 10.1093/bioinformatics/btw327.

Camille Bellard and Franck Courchamp. Potential impact of sea level rise on french islands worldwide. *Nature Conservation*, 5:75–86, 2013. doi: 10.3897/natureconservation.5.5533.

Charles H. Cannon. Is speciation an unrelenting march to reproductive isolation? *Molecular Ecology*, 30(18):4349–4352, 2021. doi: 10.1111/mec.16129.

Gerald D. Carr and Donald W. Kyhos. Adaptive radiation of the hawaiian silversword alliance (compositae- madiinae). ii. cytogenetics and natural hybrids. *Evolution*, 40(5): 959–976, 1986. doi: 10.1111/j.1558-5646.1986.tb00565.x.

Juli Caujapé-Castells, Carlos García-Verdugo, Águedo Marrero-Rodríguez, José M. Fernández-Palacios, Daniel J. Crawford, and Mark E. Mort. Island ontogenies, syngameons, and the origins and evolution of genetic diversity in the canarian endemic flora. *Perspectives in Plant Ecology, Evolution and Systematics*, 27:9–22, 2017. doi: 10.1016/j.ppees.2017.03.003.

Sutee Duangjai, Rosabelle Samuel, Jérôme Munzinger, Félix Forest, Bruno Wallnöfer, Michael H. J. Barfuss, Gunter A. Fischer, and Mark W. Chase. 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, 2009. doi: 10.1016/j.ympev.2009.04.020.

Guillaume Evanno, Sebastien Regnaut, and Jerémie Goudet. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, 14(8):2611–2620, 2005. doi: 10.1111/j.1365-294X.2005.02553.x.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C. Sousa, and Matthieu Foll. Robust demographic inference from genomic and snp data. *PLOS Genetics*, 9(10):e1003905, 2013. doi: 10.1371/journal.pgen.1003905.

Thomas W. Gillespie and Tanguy Jaffré. Tropical dry forests in new caledonia. *Biodiversity and Conservation*, 12:1687–1697, 2003. doi: 10.1023/A:1023649831355.

Peter R. Grant and B. Rosemary Grant. Hybridization increases population variation during adaptive radiation. *Proceedings of the National Academy of Sciences*, 116(46): 23216–23224, 2019. doi: 10.1073/pnas.1913534116.

Kristian Hanghøj, Ida Moltke, Philip Alstrup Andersen, Andrea Manica, and Thorfinn Sand Korneliussen. Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, 8(5):giz034, 2019. doi: 10.1093/gigascience/giz034.

Charles R. Harris, Jarrod K. Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with numpy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Philip W. Hedrick and Robert C. Lacy. Measuring relatedness between inbred individuals. *Journal of Heredity*, 106(1):20–25, 2015. doi: 10.1093/jhered/esu072.

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Sandrine Isnard, Laurent L'huillier, Frédéric Rigault, and Tanguy Jaffré. How did the ultramafic soils shape the flora of the new caledonian hotspot? *Plant and Soil*, 403(1): 53–76, 2016. doi: 10.1007/s11104-016-2910-5.

Tanguy Jaffré, Frédéric Rigault, and Jérôme Munzinger. *La végétation*, pages 77–80. IRD-Congrès de la Nouvelle-Calédonie, 2012.

Nolan C. Kane, Matthew G. King, Michael S. Barker, Andrew Raduski, Sophie Karrenberg, Yoko Yatabe, Steven J. Knapp, and Loren H. Rieseberg. Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution*, 63(8):2061–2075, 2009. doi: 10.1111/j.1558-5646.2009.00703.x.

Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. Angsd: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356, 2014. doi: 10.1186/s12859-014-0356-4.

Tatiana Korshunova, Bernard Picton, Giulia Furfaro, Paolo Mariottini, Miquel Pontes, Jakov Prkić, Karin Fletcher, Klas Malmberg, Kennet Lundin, and Alexander Martynov. Multilevel fine-scale diversity challenges the 'cryptic species' concept. *Scientific Reports*, 9(1):6732, 2019. doi: 10.1038/s41598-019-42297-5.

Felix Krueger. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 2015. Available at: `https://github.com/FelixKrueger/TrimGalore` (01.06.2023).

Daniel J. Lawson, Lucy van Dorp, and Daniel Falush. A tutorial on how not to over-interpret structure and admixture bar plots. *Nature Communications*, 9(1):3258, 2018. doi: 10.1038/s41467-018-05257-7.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. doi: 10.1093/bioinformatics/btp324.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16): 2078–2079, 2009. doi: 10.1093/bioinformatics/btp352.

Alexander G. Linan, Porter P. Lowry, Allison J. Miller, George E. Schatz, Jean C. Sevathian, and Christine E. Edwards. Rad-sequencing reveals patterns of diversification and hybridization, and the accumulation of reproductive isolation in a clade of partially sympatric, tropical island trees. *Molecular Ecology*, 30(18):4520–4537, 2021. doi: 10.1111/mec.15736.

Jianzhong Ma and Christopher I. Amos. Principal components analysis of population admixture. *PLOS ONE*, 7(7):e40115, 2012. doi: 10.1371/journal.pone.0040115.

Alex Mas-Sandoval, Nathaniel S. Pope, Knud N. Nielsen, Isin Altinkaya, Matteo Fumagalli, and Thorfinn S. Korneliussen. Fast and accurate estimation of multidimensional site frequency spectra from low-coverage high-throughput sequencing data. *GigaScience*, 11, 2022. doi: 10.1093/gigascience/giac032.

Jonas Meisner and Anders Albrechtsen. Inferring population structure and admixture proportions in low-depth ngs data. *Genetics*, 210(2):719–731, 2018. doi: 10.1534/genetics.118.301336.

Philippe Morat. Our knowledge of the flora of new caledonia: Endemism and diversity in relation to vegetation types and substrates. *Biodiversity Letters*, 1(3/4):72–81, 1993. doi: 10.2307/2999750.

Alison G. Nazareno, Ana L. Alzate-Marin, and Rodrigo A. S. Pereira. Dioecy, more than monoecy, affects plant spatial genetic structure: the case study of *Ficus*. *Ecology and Evolution*, 3(10):3495–3508, 2013. doi: 10.1002/ece3.739.

Joaquín Ortego, Paul F. Gugger, and Victoria L. Sork. Genomic data reveal cryptic lineage diversification and introgression in californian golden cup oaks (section *Protobalanus*). *New Phytologist*, 218(2):804–818, 2018. doi: 10.1111/nph.14951.

Ovidiu Paun, Barbara Turner, Emiliano Trucchi, Jérôme Munzinger, Mark W. Chase, and Rosabelle Samuel. Processes driving the adaptive radiation of a tropical tree (*Diospyros*, ebenaceae) in new caledonia, a biodiversity hotspot. *Systematic Biology*, 64(2):212–227, 2016. doi: 10.1093/sysbio/syv076.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.48550/arXiv.1201.0490.

Yohan Pillon, Tanguy Jaffré, Philippe Birnbaum, David Bruy, Dominique Cluzel, Marc Ducousso, Bruno Fogliani, Thomas Ibanez, Hervé Jourdan, Louis Lagarde, Audrey Léopold, Jérôme Munzinger, Robin Pouteau, Jennifer Read, and Sandrine Isnard. Infertile landscapes on an old oceanic island: the biodiversity hotspot of new caledonia. *Biological Journal of the Linnean Society*, 133(2):317–341, 2020. doi: 10.1093/biolinnean/blaa146.

Jean-Christophe Pintaud, Tanguy Jaffré, and Henri Puig. Chorology of new caledonian palms and possible evidence of pleistocene rain forest refugia. *Comptes Rendus de l'Académie des Sciences. Série III, Sciences de la Vie*, 324:453–463, 2001. doi: 10.1016/S0764-4469(01)01312-9.

Kirk W. Pomper, Jeremiah D. Lowe, Sheri B. Crabtree, Jacob Vincent, Andrew Berry, Clifford England, and Krit Raemakers. Ploidy level in american persimmon (*Diospyros virginiana*) cultivars. *HortScience*, 55(1):4–7, 2020. doi: 10.21273/HORTSCI14274-19.

Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. doi: 10.1093/bioinformatics/btq033.

Rosabelle Samuel, Barbara Turner, Sutee Duangjai, Jérôme Munzinger, Ovidiu Paun, Michael H. J. Barfuss, and Mark W. Chase. 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, 2019. doi: 10.1093/botlinnean/boy081.

George E. Schatz and Porter P. Lowry. Novitates neocaledonicae. ix. taxonomic notes on new caledonian *Diospyros* (ebenaceae) with new synonymy and the description of two new species. *Candollea*, 73(1):91–100, 2018. doi: 10.15553/c2018v731a8.

Céline Scornavacca and Nicolas Galtier. Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66(1):112–120, 2016. doi: 10.1093/sysbio/syw082.

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

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

Ashley T. Sendell-Price, Kristen C. Ruegg, Bruce C. Robertson, and Sonya M. Clegg. An island-hopping bird reveals how founder events shape genome-wide divergence. *Molecular Ecology*, 30(11):2495–2510, 2021. doi: 10.1111/mec.15898.

Lasse Skotte, Thorfinn S. Korneliussen, and Anders Albrechtsen. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, 2013. doi: 10.1534/genetics.113.154138.

Anaëlle Soulebeau, Xavier Aubriot, Myriam Gaudeul, Germinal Rouhan, Sabine Hennequin, Thomas Haevermans, Jean-Yves Dubuisson, and Florian Jabbour. The hypothesis of adaptive radiation in evolutionary biology: hard facts about a hazy concept. *Organisms Diversity & Evolution*, 15:747–761, 2015. doi: 10.1007/s13127-015-0220-z.

Alexander Suh, Linnéa Smeds, and Hans Ellegren. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLOS Biology*, 13(8): e1002224, 2015. doi: 10.1371/journal.pbio.1002224.

Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015. doi: 10.1093/bioinformatics/btv098.

Anna Tigano and Vicki L. Friesen. Genomics of local adaptation with gene flow. *Molecular Ecology*, 25(10):2144–2164, 2016. doi: 10.1111/mec.13606.

Barbara Turner, Jérôme Munzinger, Sutee Duangjai, Eva M. Temsch, Reinhold Stockenhuber, Michael H. J. Barfuss, Mark W. Chase, and Rosabelle Samuel. Molecular phylogenetics of new caledonian *Diospyros* (ebenaceae) using plastid and nuclear markers. *Molecular Phylogenetics and Evolution*, 69(3):740–763, 2013. doi: 10.1016/j.ympev.2013.07.002.

Madelon van de Kerk, David P. Onorato, Jeffrey A. Hostetler, Benjamin M. Bolker, and Madan K. Oli. Dynamics, persistence, and genetic management of the endangered florida panther population. *Wildlife Monographs*, 203(1):3–35, 2019. doi: 10.1002/wmon.1041.

Miguel Verdú. Age at maturity and diversification in woody angiosperms. *Evolution*, 56 (7):1352–1361, 2002. doi: 10.1111/j.0014-3820.2002.tb01449.x.

Michael L. Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021.

Frank White and Wim Vink. *Flore de la Nouvelle-Calédonie et Dépendances*, volume 19 of *Flore et Faune Tropicales*. Muséum National d'Histoire Naturelle, Paris, 1993. ISBN 2-85654-196-8.

Adrien S. Wulff, Peter M. Hollingsworth, Antje Ahrends, Tanguy Jaffré, Jean-Marie Veillon, Laurent L'Huillier, and Bruno Fogliani. Conservation priorities in a biodiversity hotspot: Analysis of narrow endemic plant species in new caledonia. *PLOS ONE*, 8(9): e73371, 2013. doi: 10.1371/journal.pone.0073371.

Zhi Yan, Megan L. Smith, Peng Du, Matthew W. Hahn, and Luay Nakhleh. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Systematic Biology*, 71(2):367–381, 2021. doi: 10.1093/sysbio/syab056.

Hui-Ping Zhou and Jin Chen. Spatial genetic structure in an understorey dioecious fig species: the roles of seed rain, seed and pollen-mediated gene flow, and local selection. *Journal of Ecology*, 98(5):1168–1177, 2010. doi: 10.1111/j.1365-2745.2010.01683.x.

A. Appendix

A.1. Excluding problematic samples

Table A.1.: List of accession pairs identified as duplicates or with potential first degree relationships, along with their respective relatedness coefficient (rab) values. These pairs were identified using NgsRelate and only one accession from each pair has been included in further analyses.

Accession pairs	rab value
minimifolia1021b - minimifolia1021c	0.548683
minimifolia1021a - minimifolia1021c	0.544138
tridentataBT206 - tridentata1078	0.511583
minimifolia1021a - minimifolia1021b	0.499798
spPicNgaBT318 - spPicNga1191	0.430217

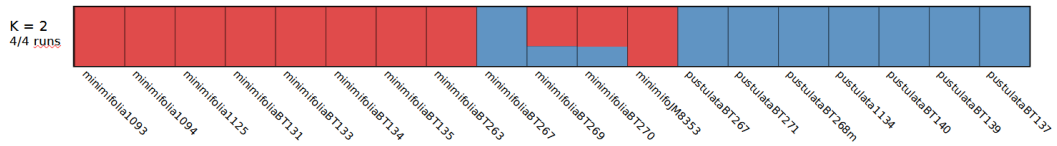

Figure A.1.: Admixture plot for identifying misclassified and introgressed samples in *D. minimifolia* L10+20. The analysis was performed with K=2 and four runs. The colors indicate the proportion of the genome assigned to each of the two genetic clusters. One sample initially classified as *D. minimifolia* L10+20 (accession: minimifoliaBT267) is shown to be actually part of *D. pustulata*.

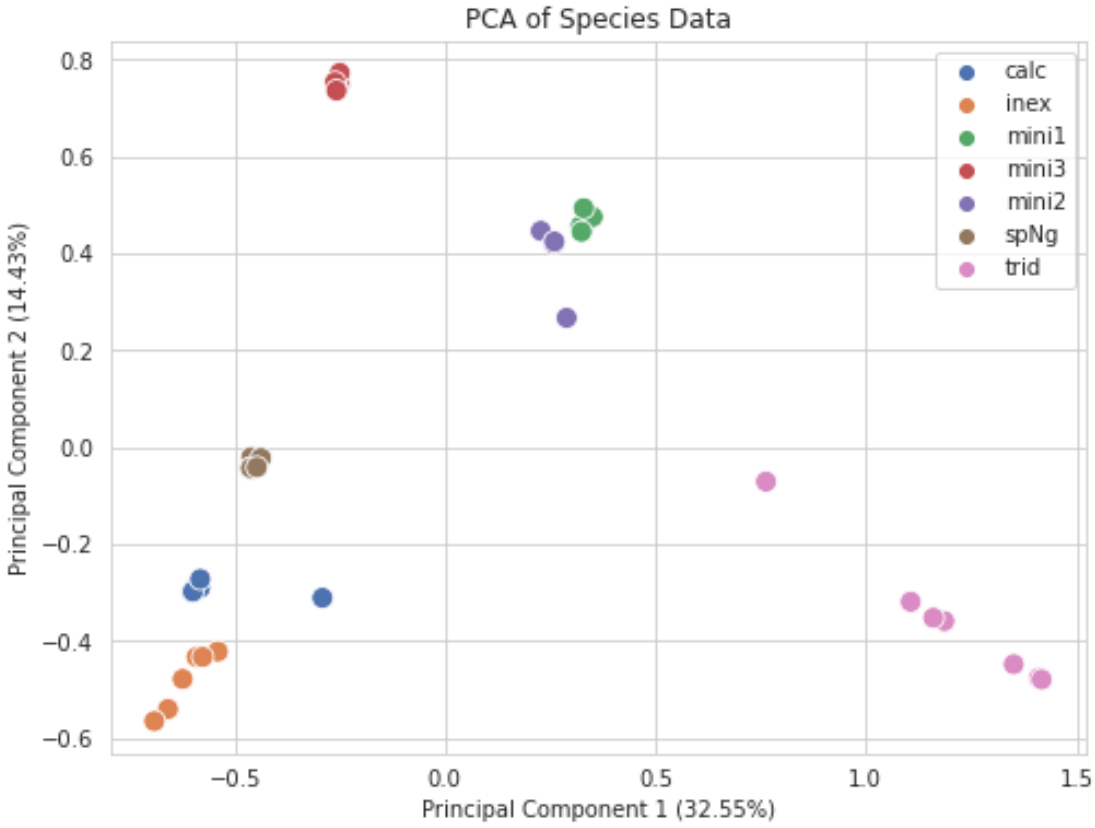

Figure A.2.: Principal Component Analysis (PCA) for excluding outliers. This PCA plot shows the genetic variation among the analyzed populations, aiming to identify and exclude misclassified samples in *D. calciphila* and *D. tridentata*. The plot is based on the first two principal components (PC), which together explain 46.9% of the total genetic variance. Each point represents an individual sample, color-coded by populations, according to legend. The PCA plot highlights the clustering of samples within each population, with distinct separation between populations. The accessions calciphilaYP124 and tridentataJM6697 were identified as outliers based on their PCA positions and have been excluded from further analysis to ensure the accuracy and reliability of the genetic study.

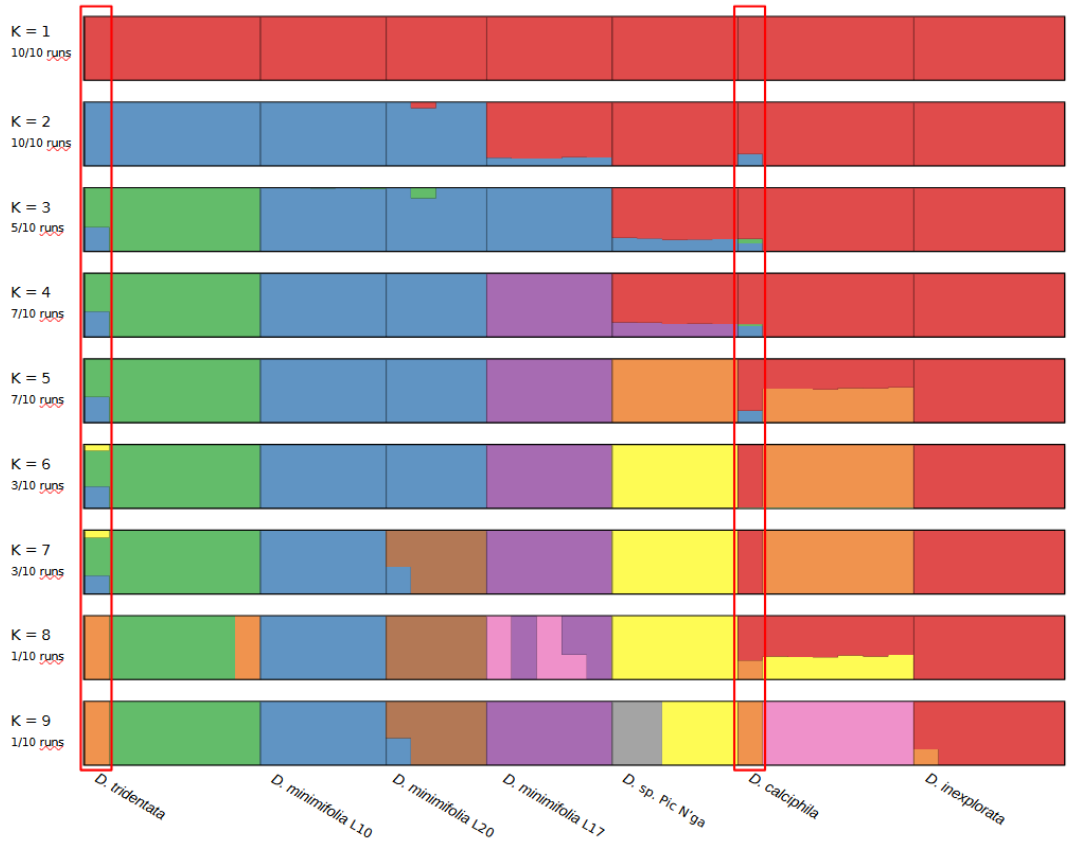

Figure A.3.: Admixture analysis for excluding outliers. This admixture plot displays the genetic structure of the populations across K values (number of genetic clusters), ranging from K=1 to K=9. The analysis was conducted using NGSadmix, with the aim of identifying and excluding genetic outliers. For each K value, the vertical bars represent individual samples, and the colors within each bar indicate the proportion of the genome assigned to each genetic cluster. One sample of *D. calciphila* (calciphilaYP124) and *D. tridentata* (tridentataJM6697), framed in red, could be identified as outliers.

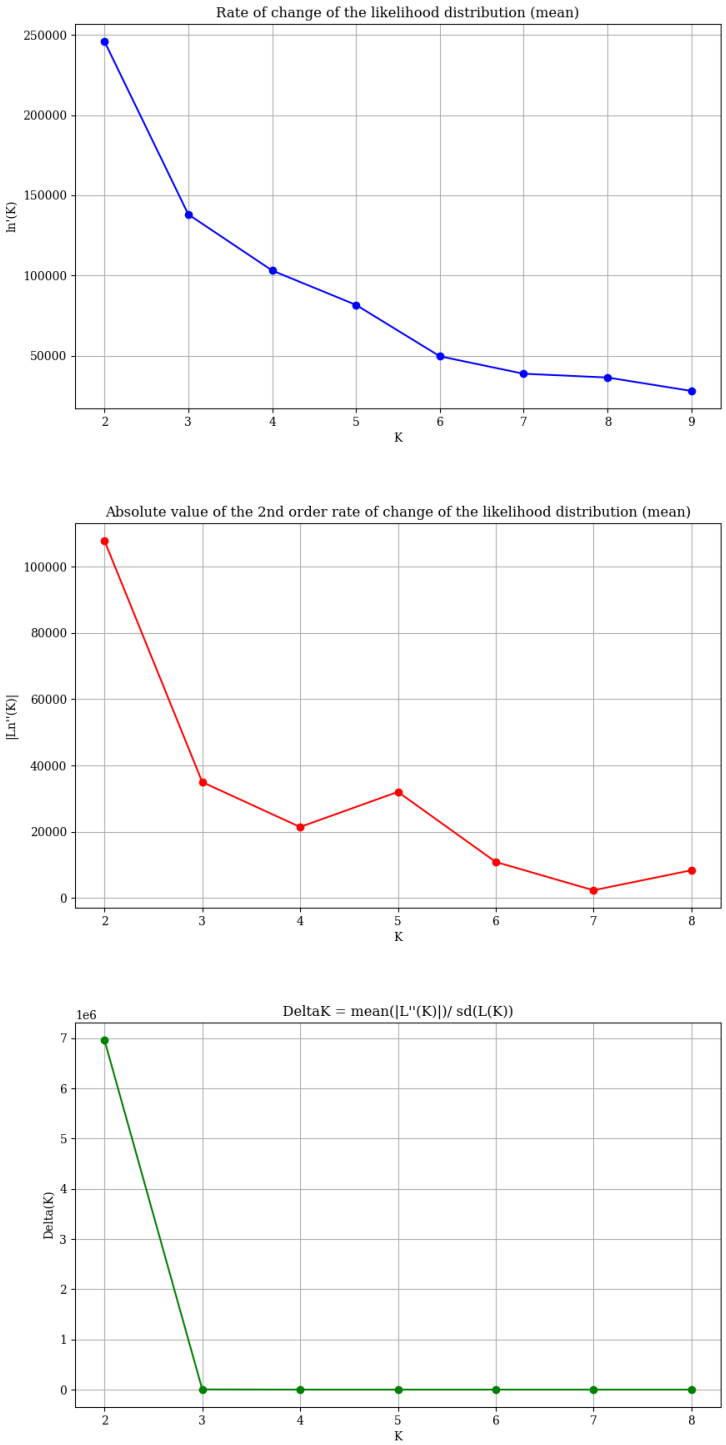

Figure A.4.: This figure displays the results of the Evanno method, which was applied on the output of NGSadmix.

A.2. Bioinformatic pipelines and scripts

```
###############_________trimming_____________############

#!/bin/bash
#
#SBATCH --job-name=TrimTrid
#SBATCH --cpus-per-task=8
#SBATCH --mem=16G
#SBATCH --partition=basic
#SBATCH --time=0-06:00:00
#SBATCH --mail-type=ALL
#SBATCH --output=%x-%j.out
#SBATCH --error=%x-%j.err

module load conda
eval "$(conda shell.bash hook)"
conda activate trim-galore-0.6.10

for file in ./Diospyros/Isamples/TridRaw/*.1.fq.gz
do
    trim_galore --paired -j 8 --stringency 2 -o ./../TridTrim/ $file ${
    file/.1.fq.gz/.2.fq.gz}
done
```

Listing A.1: Example bash script for trimming

```

###################____mapping______#################

#!/bin/bash
#
#SBATCH --job-name=MapMini
#SBATCH --cpus-per-task=32
#SBATCH --mem=40G
#SBATCH --partition=basic
#SBATCH --time=2-12:00:00
#SBATCH --mail-type=ALL
#SBATCH --output=%x-%j.out
#SBATCH --error=%x-%j.err

module load conda
eval "$(conda shell.bash hook)"
conda activate /home/user/zechmeister/.conda/envs/map-sort/

FILE="/scratch/botany/zechmeister/diospyros/Isamples/MiniTrimmed/*.trim
    .1.fq.gz"

for f in $FILE
do
    SAMPLE=$(echo ${f} | sed "s/\.trim\.1\.fq\.gz//")
    bwa mem \
    /scratch/botany/zechmeister/diospyros/ImpoRef/impolita1.1.fa -t 32 \
```

```
    ${SAMPLE}.trim.1.fq.gz \
    ${SAMPLE}.trim.2.fq.gz | \
    sambamba view -t 32 -h -f bam -F "mapping_quality >= 1" -S /dev/stdin
     |\
    samtools sort -@32 -o ${SAMPLE}.impolita_sort.bam -
    mv ${SAMPLE}.impolita_sort.bam /scratch/botany/zechmeister/diospyros/
    Isamples/MiniMapped/
done
```

Listing A.2: Example bash script for mapping

```
###################_________merging______###############

cat minimifolia1_sort.bam minimifolia2_sort.bam > minimifolia12_sort.bam
```

Listing A.3: Example bash script for merging bam files

```

###############_________ANGSD___________###############

#first, creating preliminary beagle file for filtering with -sites:

#with bamfiles
#Infer genotype likelihoods. Create beagle file.
#Adjust -minInd to the least amount of individuals, you want to see the
    analyzed SNPs to be present in. typically: 0.8 x all Ind
#Adjust -minMaf for frequency of 2 individuals like this:
#59 ind ... 1
#2 ind .... x
#x = 2/59 = 0,034

#to select SNPs which are not too close to each other (-sites). create a
    preliminary beagle file (ideally already with the filters you'd like
    to have):

#!/bin/bash
#
#SBATCH --cpus-per-task=1
#SBATCH --mem=800GB
#SBATCH --partition=himem
#SBATCH --job-name=ANGSD_RRc
#SBATCH --mail-type=ALL
#SBATCH --time=2-18:00:00
#SBATCH --output=%x-%j.out
#SBATCH --error=%x-%j.err

ulimit -v 800000000
#memory limit, so if my job uses too much and crashes, the other ones
    keep running

module load angsd

cd /scratch/botany/zechmeister/diospyros/
```

```
angsd -bam list -GL 2 -doMajorMinor 1 -doMaf 1 -SNP_pval 1e-6 -minMapQ 20
    -minQ 20 -minInd 44 -minMaf 0.02 -doGlf 2 -out RRc1
```

Listing A.4: ANGSD analysis script for preliminary beagle file

```
########___filtering Sites (LD, TE, and Scaffolds)_______###############

#first, filter out problematic regions, i.e. TE regions and Scaffolds
    with low quality Scaffolds (lq because weird stats of reference genome
    , put simply: mis-assembled)
#then filter out SNPs with less than 5000bp in between them. Therefore
    the information of all SNP positions from a preliminary beagle file is
     needed.

#create bed file of TEs
#get rid of the header, adjust spacing to bed format, get important
    columns and correct index to bed format

grep -v "^#" impolita1.1.fa.out.gff | awk 'BEGIN {OFS="\t"} {print $1, $4
    - 1, $5}' > ImpoTEs.bed

samtools faidx impolita1.1.fa
cut -f1,2 impolita1.1.fa.fai | awk '{OFS="\t"; print $1, 0, $2}' >
    impolita1.1.bed

#create bed file with only scaffolds to exclude
grep -Fwf ScaffsToExclude impolita1.1.bed > ScaffsToExclude.bed

#then merge:

awk '{print $1"\t"$2"\t"$3}' ScaffsToExclude.bed >
    ScaffsToExcludeModified.bed

cat ImpoTEs.bed ScaffsToExcludeModified.bed | sort -k1,1 -k2,2n |
    mergeBed -i stdin > RegionsToExcludeMerged.bed

#the bed file includes all regions not wanted in the analysis. for the
    sites file in angsd, the opposite/inverse is needed

#to get SNP positions (of existing beagle file)
gzip -cd *.beagle.gz | cut -f1 > *.txt
#takes a while (30m)
#then to get a nice bed file
awk -F"_" 'NR>1 {print $1"_"$2, $3-1, $3}' SNPpositions.txt >
    SNPpositions.bed

#not yet tab delimited:
awk '{$1=$1; print}' OFS="\t" SNPpositions.bed > SNPpositionsTab.bed

bedtools subtract -a SNPpositionsTab.bed -b RegionsToExcludeMerged.bed >
    SNPsToInclude.bed

##################### filtering for SNP postions#######

```

```
awk 'BEGIN {last_chr=""; last_pos=0}
{
    if ($1 != last_chr || ($1 == last_chr && $3 - last_pos >= 5000)) {
        print $1"\t"$3;
        last_chr = $1;
        last_pos = $3;
    }
}' SNPsToInclude.bed > STIfilteredUnix.txt

#then to sort

sort -V -k1,1 STIfilteredUnix.txt > STIFS.txt

#indexed with ANGSD

angsd sites index STIFS.txt
```

Listing A.5: Filtering sites

```
#########################################################
#Create beagle file (no need to redo the filters) (copy files to lisc,
    alter run.sh, etc.):

#!/bin/bash
#
#SBATCH --cpus-per-task=1
#SBATCH --mem=600GB
#SBATCH --partition=himem
#SBATCH --job-name=ANGSD_RRc3
#SBATCH --mail-type=ALL
#SBATCH --time=2-18:00:00
#SBATCH --output=%x-%j.out
#SBATCH --error=%x-%j.err

# memory limit, so if my job uses too much and crashes, the other ones
    keep running
ulimit -v 600000000

module load angsd

cd /scratch/botany/johannes/diospyros/
angsd -bam ./list -GL 2 -sites ./STIFS.txt -doMajorMinor 1 -doMaf 1 \
-minMapQ 20 -minQ 20 -minInd 44 -doGlf 2 -SNP_pval 1e-6 -out ./RRc4/RRc4

########################

#to see how many sites were analyzed:

tail ANGSD-*.err

#first beagle file (before -sites filter):
# -> Total number of sites analyzed: 1 759 754 387
# -> Number of sites retained after filtering: 204 206 833

```

```
#final beagle file:
# -> Total number of sites analyzed: 1 759 754 387
# -> Number of sites retained after filtering: 111 011
```

Listing A.6: ANGSD analysis for final beagle file

```
#############__________NgsRelate__________###############

#checking the 5 longest scaffolds:
samtools view -H calciphilaBT1001.impolita_sort.bam | grep '^@SQ' | awk '
    {print $2, $3}' | sed 's/SN://g; s/LN://g' | sort -k2 -nr
#then creating a text file for -rf command in angsd

Scaffolds_1145
Scaffolds_652
Scaffolds_1287
Scaffolds_597
Scaffolds_1238
Scaffolds_82
Scaffolds_1146
Scaffolds_62
Scaffolds_94
Scaffolds_1210
Scaffolds_703
Scaffolds_843
Scaffolds_115
Scaffolds_149
Scaffolds_76
Scaffolds_1132
Scaffolds_579
Scaffolds_16
Scaffolds_1194
Scaffolds_543

#################Indexing Bams

#!/bin/bash
#
#SBATCH --cpus-per-task=1
#SBATCH --mem=10MB
#SBATCH --partition=basic
#SBATCH --job-name=idxBams2
#SBATCH --mail-type=ALL
#SBATCH --time=0-4:00:00
#SBATCH --output=%x-%j.out
#SBATCH --error=%x-%j.err

module load samtools

cd /scratch/botany/johannes/diospyros/RRcSamples/

for file in ./*.bam; do
    if [ ! -f "$file.bai" ]; then
        samtools index "$file"
```

```
    fi
done

#####################################
#then run NGSrelate:

#!/bin/bash
#
#SBATCH --cpus-per-task=1
#SBATCH --mem=200GB
#SBATCH --partition=himem
#SBATCH --job-name=20NGSrelate
#SBATCH --mail-type=ALL
#SBATCH --time=0-18:00:00
#SBATCH --output=%x-%j.out
#SBATCH --error=%x-%j.err

module load angsd

cd /scratch/botany/johannes/diospyros/

angsd -bam ./list -GL 2 -doMajorMinor 1 -SNP_pval 1e-6 -doMaf 1 -minMaf
    0.05 -doGlf 3 -sites ./STIFS.txt -rf ./20ScaffRegions.txt -out ./
    NGSrelate2/20NGSrelate

#######################################
#then downloaded all to local, downloaded program and:

zcat 20NGSrelate.mafs.gz | cut -f5 |sed 1d > freq

/usr/local/bin/ngsRelate -g 20NGSrelate.glf.gz -n 55 -f freq -O newres

####################################
#open matrix in libre office, delete all columns not needed
#rab and theta values are higher for closely related individuals, J1 is
    the probabiliy for a clone
#KING value is especially equipped for low coverage data
#in python:

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Load your data
file_path = '/media/botanik/data4Tb/Diospyros/NGSrelate2/Table.ods'
data = pd.read_excel(file_path, engine='odf')

# Load your sample names
sample_names_file = '/media/botanik/data4Tb/Diospyros/NGSrelate2/list'
with open(sample_names_file, 'r') as file:
    sample_names = file.read().splitlines()

```

```
# Create a matrix for the heatmap
num_samples = len(sample_names)
heatmap_matrix = np.full((num_samples, num_samples), np.nan)

# Fill the matrix with the rab values
for _, row in data.iterrows():
    i, j = int(row['a']), int(row['b'])
    heatmap_matrix[i, j] = row['zygosity']
    heatmap_matrix[j, i] = row['zygosity'] # Filling symmetrically

# Plotting the heatmap
plt.figure(figsize=(20, 15))
sns.heatmap(heatmap_matrix, annot=False, cmap='viridis', xticklabels=
    sample_names, yticklabels=sample_names)
plt.title('Heatmap of zygosity Values')
plt.xticks(rotation=90)
plt.yticks(rotation=0)
plt.show()
```

Listing A.7: NGSrelate analysis

```
##########_________PCAngsd___________#################

#calculate covariance matrix. Use the same -minMaf as above
pcangsd --beagle *.beagle.gz --minMaf 0.05 -o ./PCAngsd/*

#create Pop file

#hereâĂŹs a python code for visualization:

import numpy as np
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns

RRC1_1PCA = "RRC1.1.cov"
CovM = np.loadtxt(RRC1_1PCA, delimiter=' ')

n_components = 4

pca = PCA(n_components=n_components)
projected_data = pca.fit_transform(CovM)

explained_variance_ratio = pca.explained_variance_ratio_

with open('RRCPop', 'r') as p:
    names = [line.strip() for line in p.readlines()]

names_array = np.array(names)

sns.set_style(style='whitegrid')
plt.figure(figsize=(8, 6))
sns.scatterplot(x=projected_data[:, 0], y=projected_data[:, 1], hue=
    names_array, palette='deep', s=100)
```

```
plt.xlabel(f'Principal Component 1 ({explained_variance_ratio[0]*100:.2f}%)')
plt.ylabel(f'Principal Component 2 ({explained_variance_ratio[1]*100:.2f}%)')
plt.title('PCA of Species Data')
plt.show()
```

Listing A.8: PCAngsd analysis

```
##############__________NGSadmix__________###################

#create bash script for NgsAdmix (adjust names of folders and files):

#!/bin/bash

NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 1 -o ./NGSadmix/gatk4_1_1
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 2 -o ./NGSadmix/gatk4_1_2
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 3 -o ./NGSadmix/gatk4_1_3
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 4 -o ./NGSadmix/gatk4_1_4
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 5 -o ./NGSadmix/gatk4_1_5
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 6 -o ./NGSadmix/gatk4_1_6
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 7 -o ./NGSadmix/gatk4_1_7
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 8 -o ./NGSadmix/gatk4_1_8
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 9 -o ./NGSadmix/gatk4_1_9
NGSadmix -likes gatk4.beagle.gz -K 1 -minMaf 0.0625 -seed 10 -o ./NGSadmix/gatk4_1_10

NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 1 -o ./NGSadmix/gatk4_2_1
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 2 -o ./NGSadmix/gatk4_2_2
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 3 -o ./NGSadmix/gatk4_2_3
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 4 -o ./NGSadmix/gatk4_2_4
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 5 -o ./NGSadmix/gatk4_2_5
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 6 -o ./NGSadmix/gatk4_2_6
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 7 -o ./NGSadmix/gatk4_2_7
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 8 -o ./NGSadmix/gatk4_2_8
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 9 -o ./NGSadmix/gatk4_2_9
```

```
NGSadmix -likes gatk4.beagle.gz -K 2 -minMaf 0.0625 -seed 10 -o ./
    NGSadmix/gatk4_2_10

...

NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 1 -o ./NGSadmix
    /gatk4_9_1
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 2 -o ./NGSadmix
    /gatk4_9_2
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 3 -o ./NGSadmix
    /gatk4_9_3
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 4 -o ./NGSadmix
    /gatk4_9_4
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 5 -o ./NGSadmix
    /gatk4_9_5
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 6 -o ./NGSadmix
    /gatk4_9_6
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 7 -o ./NGSadmix
    /gatk4_9_7
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 8 -o ./NGSadmix
    /gatk4_9_8
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 9 -o ./NGSadmix
    /gatk4_9_9
NGSadmix -likes gatk4.beagle.gz -K 9 -minMaf 0.0625 -seed 10 -o ./
    NGSadmix/gatk4_9_10

#run the bash file:
bash ./NGSadmix/NGSadmixBash.sh
```

Listing A.9: NGSadmix analysis

```
######___to get an best K estimate via the clumpak method____#####

for file in *.log; do echo $file; grep 'best like' $file; done
#copy to new text file and run use this python code:

import re

input_file = "BestLikes" # Replace with the name of your input file
output_file = "Klogs3" # Replace with the name of your output file

with open(input_file, "r") as infile, open(output_file, "w") as outfile:
    lines = infile.readlines()

    for i in range(0, len(lines), 2):
        log_line = lines[i]

        if i + 1 < len(lines): # Check if the index is within the list
    bounds
            like_line = lines[i + 1]
        else:
            break # Stop processing if there is no corresponding "best
    like" line

```

```
        # Extract numbers from the lines
        match1 = re.search("gatk7_(\d+)_(\d+).log", log_line)
        match2 = re.search("best like=(-[\d\.]+)\s+after", like_line)

        if match1 and match2: # Check if both patterns were found
            column1 = match1.group(1) # Only use the first number from the log file name
            column2 = match2.group(1)

            # Write the numbers to the output file
            outfile.write(f"{column1}\t{column2}\n")

#follow instructions, at clumpak.tau.ac.il/bestK.html
```

Listing A.10: Best K estimate using Clumpak (Evanno method)

```
#########plot results with pong#####

#pong:
#adjust necessary files (Filemap, Pop, Poporder)

pong -m PongFilemap -i pop -n pop_order
```

Listing A.11: Admixture visualization with Pong

```
#############_____________Fst___________###############

#creating .saf files:
angsd -bam ./SFS/listCalc -GL 2 -anc ./ImpoRef/impolita1.1.fa -minMapQ 30 -minQ 20 -minInd 7 -doSaf 1 -sites ./sites/STIFS.txt -nThreads 2 -out ./SFS/Calc/Calc

angsd -bam ../SFS/listInex -GL 2 -anc ../samples/ImpolitaRef/impolita1.1.fa -minMapQ 30 -minQ 20 -minInd 5 -doSaf 1 -sites -nThreads 2 -out ./SFS/Inex/Inex

#Fst for 2 populations

#ANGSD pipeline: http://www.popgen.dk/angsd/index.php/Fst

realSFS Calc.saf.idx Inex.saf.idx > jCaIn.sfs

realSFS fst index Calc.saf.idx Inex.saf.idx -sfs jCaIn.sfs -fstout CaIn

realSFS fst stats CaIn.fst.idx > FstValues.txt
```

Listing A.12: Fst analysis

```
######_____second -sites file (incl. monomorphic sites)____########

#starting from impolita1.1.fa.out.gff in order to have a -sites file with TE free positions

```

```
sort -k1,1V -k4,4n -k5,5rn -k3,3r impolita1.1.fa.out.gff > impolita1.1.
    sorted.gff

bedtools complement -i impolita1.1.sorted.gff -g my.genome

cut -f 1,2 impolita1.1.fa.fai | sort -k1,1V > my.genome

bedtools complement -i impolita1.1.sorted.gff -g my.genome > noTEs.sites

awk '{print $1, $2+1, $3+1}' noTEs.sites > impolita1.1.angsd.noTEs.
    regions

sed 's/ \([0-9]*\) \([0-9]*\)$/:\1-\2/' impolita1.1.angsd.noTEs.regions >
     Imp.noTEs.regions

#i then created a file with sites instead on regions:

#!/bin/bash

input_file="/media/botanik/data4Tb/Diospyros/Heterozygosity/Hete4/
    impolita1.1.angsd.noTEs.regions"

output_file="Imp.noTEs.sites"

> "$output_file"

while read line; do
    read scaffold start end <<< $(echo $line)

    for ((i=start; i<=end; i++)); do
        echo "$scaffold $i" >> "$output_file"
    done
done < "$input_file"

# filtered for 1000bp between snps

awk 'BEGIN {last_chr=""; last_pos=0}
{
    if ($1 != last_chr || ($1 == last_chr && $2 - last_pos >= 1000)) {
        print $1"\t"$2;
        last_chr = $1;
        last_pos = $2;
    }
}' Imp.noTEs.sites > Imp.noTEs.1000.sites

# sorted the file and indexed it with ANGSD

sort -V -k1,1 Imp.noTEs.1000.sites > S.Imp.noTEs.1000.sites

angsd sites index S.Imp.noTEs.1000.sites
```

Listing A.13: Creating a sites file including monomorphic sites

```
########______sfs, jsfs_________###########

#i used scripts like the one below for all my .saf file for my latest
    models (Speciation, Growth).

#!/bin/bash
#
#SBATCH --cpus-per-task=2
#SBATCH --mem=200GB
#SBATCH --partition=himem
#SBATCH --job-name=CaSFS
#SBATCH --mail-type=ALL
#SBATCH --time=2-00:00:00
#SBATCH --output=%x-%j.out
#SBATCH --error=%x-%j.err

cd /scratch/botany/johannes/diospyros/

ml angsd

angsd -bam ./SFS/lists/listCalc -GL 1 -anc ./ImpoRef/impolita1.1.fa -ref
    ./ImpoRef/impolita1.1.fa \
    -doSaf 1 -minMapQ 30 -minQ 20 -baq 1 -C 50 -sites ./sites/S.Imp.noTEs
    .1000.sites \
    -nThreads 2 -out ./SFS/Speciation/Calc/Calc

#for sfs creation:

realSFS ../SFS/Calc2/Calc.saf.idx -P 8 > ../SFS/Calc/Calc.sfs

#and jsfs creation:

realSFS ../SFS/Calc/Calc.saf.idx ../SFS/SpNga/SpNga.saf.idx -P 8 > ../SFS
    /jSFS/2/jCaSp.sfs
```

Listing A.14: SFS and jSFS Creation

```
########################____to adjust SFS for fsc2

with open('../SFS/Inex3/Inex.sfs', "r") as file:
    sfs_data = file.read().strip().split()

sfs_data = [float(x) for x in sfs_data]

header_row = "1 observations" + "\t" * len(sfs_data)
column_names = ["d0_{}".format(i) for i in range(len(sfs_data))]
column_names_row = "\t".join(column_names)
data_row = "\t".join(map(str, sfs_data))

formatted_sfs_content = header_row + "\n" + column_names_row + "\n" +
    data_row + "\n"

formatted_sfs_content.splitlines()[:10]

```

```
def save_to_file(content, file_path):
    with open(file_path, "w") as file:
        file.write(content)

output_path = '/media/botanik/data4Tb/Diospyros/fsc2/InBot/InGro_DAFpop0.obs'
save_to_file(formatted_sfs_content, output_path)
```

Listing A.15: Adjusting SFS for FSC2

```
#####################____to adjust jSFS for fsc2

import numpy as np

def unflatten_sfs(sfs_file, n1, n2):
    with open(sfs_file, 'r') as file:
        sfs_data = file.read().strip().split()
        sfs_data = [float(x) for x in sfs_data]
        ncols = 2 * n1 + 1
        nrows = 2 * n2 + 1
        return np.array(sfs_data).reshape((nrows, ncols))

def round_matrix(matrix):
    """Round all values in the matrix to integers."""
    return [[int(round(val)) for val in row] for row in matrix]

def format_matrix_to_obs(matrix):
    header_row = "1\tobservations"
    column_names = ["d0_{}".format(i) for i in range(len(matrix[0]))]
    column_names_row = "\t" + "\t".join(column_names)
    data_rows = ["d1_{}\t{}".format(i, "\t".join(map(str, row))) for i, row in enumerate(matrix)]
    return header_row + "\n" + column_names_row + "\n" + "\n".join(data_rows)

def save_to_file(content, file_path):
    with open(file_path, "w") as file:
        file.write(content)

# Path to the raw jSFS file
sfs_file = 'jCaSp.sfs'
n1 = 11 # Adjust based on your population
n2 = 9 # Adjust based on your population

# Processing steps
matrix = unflatten_sfs(sfs_file, n1, n2)
matrix = round_matrix(matrix)
matrix = format_matrix_to_obs(matrix)

output_path = '/media/botanik/data4Tb/Diospyros/SFS/jSFS/jCaSp.obs'
save_to_file(matrix, output_path)
```

Listing A.16: Adjusting jSFS for FSC2

```
#####################_____.est for estimating the growth rate

// Priors and rules file
// *********************
[PARAMETERS]
//#isInt? #name #dist.#min #max
//all Ns are in number of haploid individuals
1 NCUR logunif 100 2000000 output
1 NANC logunif 100 2000000 output

[COMPLEX PARAMETERS]
0 RATIO_OOA_EA = NANC/NCUR hide
0 RTEA = log(RATIO_OOA_EA) hide
0 R1 = RTEA/10000 output
```

Listing A.17: FSC2 .est file for growth rate

```
###################_____.tpl for estimating the growth rate

//Number of population samples (demes)
1
//Population effective sizes (number of genes)
NCUR
//Sample sizes
20
//Growth rates : negative growth implies population expansion
R1
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new size, new growth
    rate, migr. matrix
1 historical event
70000 0 0 0 NANC 0 0 absoluteResize #or 2000
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of linkage blocks
1
//per Block: data type, num loci, rec. rate and mut rate + optional
    parameters
FREQ 1 0 7e-9 OUTEXP
```

Listing A.18: FSC2 .tpl file for growth rate

```
###################_____.est for AM

// Priors and rules file
// *********************

[PARAMETERS]
//#isInt? #name #dist.#min #max
//all Ns are in number of haploid individuals
1 SPNGA logunif 100 2000000 output
1 CALC logunif 100 2000000 output
```

```
1 NANC logunif 100 2000000 output
0 MIGMX logunif 1e-9 0.001 output
0 MIGXM logunif 1e-9 0.001 output
1 TDIV logunif 1500 400000 output
0 TROP logunif 0.5 1 hide

[COMPLEX PARAMETERS]

1 CHANGM = TDIV * TROP output
```

Listing A.19: FSC2 .est file for AM

```
##########################_____.tpl for AM

//Number of population samples
2
//Population effective sizes (number of genes)
SPNGA
CALC
//Haploid samples sizes
20
20
//Growth rates: negative growth implies population expansion
0
0
//Number of migration matrices : 0 implies no migration between demes
3
//Migration matrix 0
0 0
0 0
//Migration matrix 1
0 MIGMX
MIGXM 0
//Migration matrix 2
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new
    growth rate, migration matrix index
2 historical event
CHANGM 0 0 1 1 0 1 //migration starts
TDIV 0 1 1 NANC 0 2 absoluteResize //ancestral pop splits
//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 generation recombination and
    mutation rates and optional parameters
FREQ 1 0 7e-9 OUTEXP
```

Listing A.20: FSC2 .tpl file for AM

```
#####################_____.est for SC
```

```

// Priors and rules file
// *********************

[PARAMETERS]
//#isInt? #name #dist.#min #max
//all Ns are in number of haploid individuals
1 SPNGA logunif 100 2000000 output
1 CALC logunif 100 2000000 output
1 NANC logunif 100 2000000 output
0 MIGMX logunif 1e-9 0.001 output
0 MIGXM logunif 1e-9 0.001 output
1 TDIV logunif 1500 400000 output
0 TROP logunif 0.0001 0.5 hide

[COMPLEX PARAMETERS]

1 CHANGM = TDIV * TROP output
```

Listing A.21: FSC2 .est file for SC

```
#######################_____.tpl for SC

//Parameters for the coalescence simulation program : fastsimcoal.exe
2 samples to simulate :
//Population effective sizes (number of genes)
SPNGA
CALC
//Haploid samples sizes
20
20
//Growth rates: negative growth implies population expansion
0
0
//Number of migration matrices : 0 implies no migration between demes
3
//Migration matrix 0
0 MIGMX
MIGXM 0
//Migration matrix 1
0 0
0 0
//Migration matrix 2
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new
    growth rate, migration matrix index
2 historical event
CHANGM 0 0 1 1 0 1 //migration starts
TDIV 0 1 1 NANC 0 2 absoluteResize //ancestral pop splits
//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 generation recombination and
    mutation rates and optional parameters
FREQ 1 0 7e-9 OUTEXP
```

Listing A.22: FSC2 .tpl file for SC

```
#########################_____.est for IM

// Priors and rules file
// **********************

[PARAMETERS]
//#isInt? #name #dist.#min #max
//all Ns are in number of haploid individuals
1 SPNGA logunif 100 2000000 output
1 CALC logunif 100 2000000 output
1 NANC logunif 100 2000000 output
0 MIGMX logunif 1e-9 0.001 output
0 MIGXM logunif 1e-9 0.001 output
1 TDIV logunif 1500 400000 output

[COMPLEX PARAMETERS]
```

Listing A.23: FSC2 .est file for IM

```
#######################_____.tpl for IM

//Parameters for the coalescence simulation program : fastsimcoal.exe
2 samples to simulate :
//Population effective sizes (number of genes)
SPNGA
CALC
//Haploid samples sizes
20
20
//Growth rates: negative growth implies population expansion
0
0
//Number of migration matrices : 0 implies no migration between demes
2
//Migration matrix 0
0 MIGMX
MIGXM 0
//Migration matrix 1
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new
    growth rate, migration matrix index
1 historical event
TDIV 0 1 1 NANC 0 1 absoluteResize //ancestral pop splits
//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 generation recombination and
    mutation rates and optional parameters
FREQ 1 0 7e-9 OUTEXP
```

Listing A.24: FSC2 .tpl file for IM

```
##########################_____.est for SI

// Priors and rules file
// *********************

[PARAMETERS]
//#isInt? #name #dist.#min #max
//all Ns are in number of haploid individuals
1 SPNGA logunif 100 2000000 output
1 CALC logunif 100 2000000 output
1 NANC logunif 100 2000000 output
1 TDIV logunif 1500 400000 output

[COMPLEX PARAMETERS]
```

Listing A.25: FSC2 .est file for SI

```
##########################_____.tpl for SI

//Parameters for the coalescence simulation program : fastsimcoal.exe
2 samples to simulate :
//Population effective sizes (number of genes)
SPNGA
CALC
//Haploid samples sizes
20
20
//Growth rates: negative growth implies population expansion
0
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new
    growth rate, migration matrix index
1 historical event
TDIV 0 1 1 NANC 0 0 absoluteResize //ancestral pop splits
//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 generation recombination and
    mutation rates and optional parameters
FREQ 1 0 7e-9 OUTEXP
```

Listing A.26: FSC2 .tpl file for SI

```
#######______run fsc2______#######

#!/bin/bash
#
#SBATCH --cpus-per-task=12
#SBATCH --mem=200Mb
#SBATCH --partition=basic
#SBATCH --job-name=Fsc2AM
#SBATCH --time=6-00:00:00
#SBATCH --mail-type=ALL
#SBATCH --output=%x-%j.out
#SBATCH --error=%x-%j.err

ml fastsimcoal/

model="AM"

folders=(CaSp InCa Mi12Tr SpMi3)

for i in "${folders[@]}"; do
    cd "/scratch/botany/johannes/diospyros/Fsc2/Speciation/$model/$i/" ||
     exit

    for j in {1..100}
    do
        mkdir run$j
        cp $model.tpl $model.est ${model}_jointDAFpop1_0.obs run$j"/"
        cd run$j
        fsc -t $model.tpl -e $model.est -d -n 500000 -M -L 80 -C 5 -0 -c
    12 -q
        cd ..
    done
done
```

Listing A.27: Running FSC2

```
############_____________Script to get the best run

#!/bin/bash
#to be run in folder where all the folders with runs are (after doing 100
     fsc2 runs or similar)
# Bash script by Joana Meier to select the best fastsimcoa2 run of
    multiple runs under the same demographic model
# The script expects output files of different runs to be found in
    folders starting with run

m=-1000000000000000
p=$1
c=0
best="xxxx"
```

```
diff=0
diffBest=0

for i in run*;
do
    a=$(ls | grep ${p}".tpl" | sed s'/.tpl//')

    # if the file is in the run directory
    if [ -e $i/$a.bestlhoods ]
    then
        l=$(cat $i/$a.bestlhoods | awk '{print $(NF-1)}' | tail -1 | cut -f 1 -d ".")
        diff=$(cat $i/$a.bestlhoods | awk '{print $NF-$(NF-1)}' | tail -1 | cut -f 1 -d ".")
        ((c++))
    else
        # if the file is in a subdirectory
        if [ -e $i/$a/$a.bestlhoods ]
        then
            l=$(cat $i/$a/$a.bestlhoods | awk '{print $(NF-1)}' | tail -1 | cut -f 1 -d ".")
            diff=$(cat $i/$a/$a.bestlhoods | awk '{print $NF-$(NF-1)}' | tail -1 | cut -f 1 -d ".")
            ((c++))
        else
            echo "no .bestlhoods file found in "$i
        fi
    fi

    if [ $l -gt $m ]
    then
        m=$l; x=$i; best=$i;
        diffBest=$diff
    fi
done

if [ -z ${x+1} ]
then echo "Error: No run with lik>-1000000000000000"
else mkdir bestrun
    cp $x/* ./bestrun/
    cp $x/$a/* ./bestrun/
fi

echo -e "\n"$c" bestlhoods files found, "$best" with "$diffBest" Lhood diff fits best."
```

Listing A.28: Script to get the best run

```
################_________Script for AIC

#! /usr/bin/Rscript
# Joana Meier

# Usage: calculateAIC.sh modelprefix
```

```
7
8  # This script calculates AIC from fsc modeling results
9  # Run in the folder with the highest likelihood
10
11 # Read model name
12 args=commandArgs(TRUE)
13
14 # Checks if model name was given
15 if(length(args)<1){
16     stop("ERROR: No input / model name given\nUsage: fsc-calculateAIC.R
       modelname")
17 }
18
19 # Check if model.bestlhoods file exists
20 if(file.exists(paste(args[1],".bestlhoods",sep=""))){
21     bestlhoods<-read.delim(paste(args[1],".bestlhoods",sep=""))
22 }else{
23     stop(paste("ERROR: Aborted. No file ",args[1],".bestlhoods file
       exists",sep=""))
24 }
25
26 # Check if model.est file exists
27 if(file.exists(paste(args[1],".est",sep=""))){
28     est<-readLines(paste(args[1],".est",sep=""))
29 }else{
30     stop(paste("ERROR: Aborted. No file ",args[1],".est file exists in
       this directory!\nUsage: fsc-calculateAIC.R modelname",sep=""))
31 }
32
33 # Count parameters in [PARAMETERS] section
34 k_parameters <- (grep("[COMPLEX PARAMETERS]", est)) - (grep("//all Ns are
       ", est) + 1)
35
36 # Count parameters in [COMPLEX PARAMETERS] section
37 k_complex <- length(est) - grep("[COMPLEX PARAMETERS]", est)
38
39 # Total number of parameters
40 k <- k_parameters + k_complex
41
42 # Calculate AIC
43 AIC<-2*k-2*(bestlhoods$MaxEstLhood/log10(exp(1)))
44
45 # Calculate delta-likelihood
46 deltaL<-bestlhoods$MaxObsLhood-bestlhoods$MaxEstLhood
47
48 # Output model.AIC file in simulation folder
49 write.table(cbind(deltaL,AIC),paste(args[1],".AIC",sep=""),row.names = F,
       col.names = T,sep = "\t",quote = F)
```

Listing A.29: Script for AIC calculation

```
1 ################_____To get a summary table of .MaxLhood files
2
3 awk 'NR==1 {print $0; exit}' run1/Growth/*.bestlhoods > Summary.csv
```

```
4 for i in {1..60}; do awk 'NR==2 {print $0}' run${i}/Growth/*.bestlhoods
    >> Summary.csv; done
```

Listing A.30: Summary table of .MaxLhood files