Phylogenomics and taxonomy of the genus *Donella*, with emphasis on the Malagasy species.

Examined by	Prof. Dr. Hanno Schaefer Plant Biodiversity Research Technical University of Munich, Germany
Supervised by	Dr. Yamama Naciri Dr. Laurent Gautier Dr. Carlos G. Boluda Unité de Systématique et Médiation Conservatoire et Jardin botaniques, Geneva Université de Genève, Switzerland
Submitted by	Tina Kiedaisch, matriculation No. 03683526
Submitted on	30.06.2022

Declaration

I hereby affirm that the here presented master's thesis has been written without any use of not mentioned resources or help from other persons. Any part taken, analo-gously or literally, from publications or other sources has been labelled as such. This thesis has not been submitted, in identical or a similar form, to any other examination board.

30.06.2022, Livingstone, Zambia

List of Abbreviations

DNA	Desoxyribonucleic acid
dsDNA	Double stranded DNA
G	Herbarium Geneva
IUCN	International Union for Conservation of Nature
MDS	Multi-dimensional scaling
ML	Maximum-likelihood
MSCC	Multi-species coalescent cluster
NGS	Next Generation Sequencing
P	Herbarium Paris
PCA	Principal component analysis
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
STACEY	Species Tree and Classification Estimation, Yarely

Table of Contents

List of Tables

List of Figures

Abstract

Like many Malagasy angiosperm lineages, the Sapotaceae genus *Donella* shows high rates of endemism. Among the 11 currently recognized species in Madagascar, ten are endemic. Only one of the Malagasy species (*D. lanceolata*) is considered to have a wider distribution, ranging from India to Queensland and the Solomon Islands. Six further *Donella* species are found only in tropical continental Africa but not in Madagascar.

Despite a recent morphological revision, several questions about the systematics of the Malagasy *Donella* species remain open which affects threat assessments and conservation planning. In this study, we aim to resolve some of these open questions with a molecular approach. First, we aimed to unravel the relationships of the morphologically similar species *D. delphinensis*, *D. analalavensis*, and *D. fenerivensis,* occurring along a precipitation gradient. Second, we addressed a putative species complex around *D. perrieri*, a very widespread and morphologically highly variable species.

About 750 herbarium specimens were reviewed in P and G and 99 of them have been selected for genomic analysis including up to 100 years old specimens. We combined Illumina sequencing with a target enrichment method to capture almost 800 nuclear low-copy genes previously selected for Sapotaceae.

To test species delimitation, approaches such as phylogenies, gene tree clustering, genetic network, PCA, heterozygosity level and STACEY were used. With the exception of *D. ambrensis*, all remaining sixteen species of *Donella* are represented in the molecular analysis. Overall, we obtained congruent results with all analysis which were more or less consistent with morphological characters and geographical distribution patterns. All analyses displayed a clear phylogenetic delimitation between Malagasy and continental African *Donella* species with the exception of the Malagasy *D. guereliana*. According to genetic and morphological distance, we suggest that *D. guereliana* does not belong to the genus *Donella*. Concerning *D. lanceolata,* we propose a split into two species due to genetic and geographic disjunction: *D. lancoelata* from the Indo-Pacific and *D. malagassica* (*stat. & comb. nov*.) from Madagascar.

Regarding the first question, the two similar species *D. delphinensis* and *D. fenerivensis* show each two genetically well delimited clades which are however morphologically very similar. This would correspond to cases of morphological convergence. In addition, we found strong

evidence of hybridization between them. Nevertheless, no close relation to *D. analalavensis* could be found. On the other hand, there seems to be a close genetic and morphological relationship between the latter and *D. humbertii*. Second, our analyses show that the putative species complex of *D. perrieri* comprises three genetically different species but with similar morphologies. Furthermore, we suggest hybridization from *D. perrieri* with *D. masoalensis*, *D. delphinensis* and *D. fenerivensis* due to samples with high heterozygosity levels and intermediary states in all analyses. Finally, we suggest the description of four new species based on our results.

1 Introduction

Madagascar

Madagascar is an island country of exceptional interest as it features very varied ecological environments and is geographically isolated from the African mainland since the Jurassic period around 185 million years ago (Goodman & Benstead 2005). It covers an area of about 590.000 km^2, which is around the size of France, and is located in the Indian Ocean on the East African coast.

Madagascar presents both a great variety of climate and geological substrates which results in an exceptional flora (Gautier *et al.,* in press). According to its phytogeography, Madagascar can be divided in mainly six Domains (Figure 1) (Humbert, 1955). Since vegetation with human transformation cover most of the island (Lowry *et al.,* 1997), the domains should be treated associated to them. The Eastern and the Sambirano Domain are characterized by moist evergreen lowland forests and very humid climate. This vegetation form is dominated by evergreen trees, many epiphytes, and lianas but mostly lacking undergrowth. It ranges from sea level up to 800-1000 m. Due to its excellent conditions for agriculture the moist lowland forests have been largely cleared. Meanwhile, the central highlands traverse the island with its over 2000 m high mountains and a temperate climate. They have been extensively deforested and are now mainly dominated by secondary grasslands. Above 2000m the montane ericoid thicket dominates which is exempt from clearing. In the Southern Domain, xerophilous thickets persist in a subarid climate. They built a closed shrub formation consisting of spiny aphyllous plants, or plant accumulating water in their organs. This formation is considered to maintain a high level of diversity and endemism. Whereas the main threads are timber extraction and charcoal production, the destroyed landscape is often invaded by introduced plants. The seasonal Western Domain features dry deciduous forests under a seasonal climate. Characterizing is the multi-layer closed formation composed of deciduous tree, comprising liana but rarely epiphytes or ferns. It has also been widely deforested and persists here and there, especially on karstic limestone massifs that are naturally sheltered from fire hazard. It is considered as the most threatened forest ecosystem in Madagascar, mainly by slash-and-burn. At sites with higher water availability, moist semi-

deciduous forests are growing. Likewise, this vegetation from is highly threatened by clearing for agriculture (Goodman *et al*., 2018).

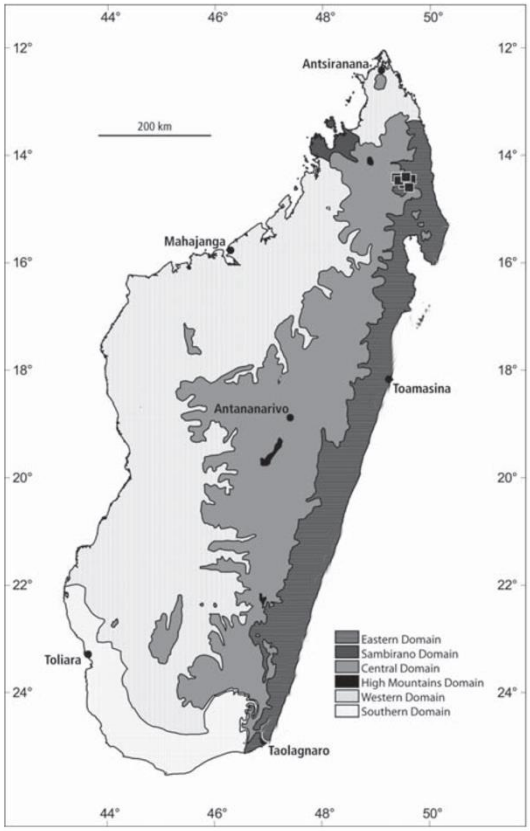

Figure 1. Phytogeographical Map showing the six domains of Madagascar; Humbert (1955) modified by Callmander and Phillipson (2011).

This wide topographic variation, the spectrum of different soil substrate conditions and its long isolation time partly explains why Madagascar is among the sixteen hotspots of biodiversity in the tropics, ranking highest in terms of vertebrate and plant endemism (Myers *et al.*, 2000; Callmander 2011; Gautier & Goodman 2003). However, the actual species diversity is still underestimated (Phillips *et al.,* 2003; Buerki *et al.,* 2013), as demonstrated for the genus *Capurodendron,* where a recent study revealed twice as many species as listed in the Flora of Madagascar (1974). *Capurodendron* is now considered the most species-rich endemic plant genus in Madagascar (Boluda *et al.*, 2022). Overall, there are 11,866 native vascular plant species in 253 families listed in the Catalogue of the Plants of Madagascar (2020).

In the past two centuries great efforts have been made to investigate the botanical diversity, which can be accessed in the 'Catalogue of the Plants of Madagascar' http://www.tropicos.org/Project/Madagascar (Madagascar Catalogue, 2020). In 2011 this

catalogue comprised a total of around 11,000 species of which 82 % are endemic to Madagascar (Callmander *et al.* 2011). In addition, large herbarium collections were established in the herbaria of Paris (France), Missouri (USA) and Geneva (Switzerland) along with those of Antananarivo (Madagascar).

Since humans arrived in Madagascar their impact on the native flora has continuously increased resulting in only 10 % of original Malagasy forest remaining (Harper *et al.*, 2007). About 81 % of the island`s surface is now covered with anthropogenic vegetation (Goodman *et al*., 2018). In addition to cropland expansion and rural populations growth, anthropogenic fires out of control are threatening especially the dry Western and Central Domain. Due to the inefficient farming, the fields are mostly abandoned after one or two years of harvest. Then, secondary thickets will grow, which could become secondary forests (Rasoanaivo et al., 2015). Logging started during colonial times and continues to put pressure on valuable timber species even in protected areas. Madagascar is ranked one of the highest priority areas for biodiversity conservation in the world and it is also one of the top recipients of biodiversity-related funding (Mittermeier *et al.* 1998; Myers *et al.,* 2000; Waeber *et al*., 2016). To slow down biodiversity loss, one of the key tools is establishing protected areas based on inventories and robust taxonomical knowledge. Nowadays Madagascar has an extended network of 122 protected areas which cover more than 10% of terrestrial landscapes and seascapes (Goodman *et al*., 2018, Coldrey and Turpie, 2021).

Sapotaceae and the genus *Donella*

Like most plant groups in Madagascar, the woody plant family Sapotaceae depends on the network of protected areas. These mostly slow-growing hard timber trees are indeed under pressure of illegal logging, one-third (36 species which are assessed to date) of the species being listed in the IUCN Red List categories (Gautier *et al.,* in press). The other species are not evaluated due to data deficiency caused among others by poor species delimitation. Therefore, conservation assessments cannot be conducted for 35% of the described species of Malagasy Sapotaceae (Boluda *et al.*, 2022). Since a clear delimitation of the species is the basis for this, taxonomic revisions are urgently needed. Nevertheless, it has remained difficult due to scarce sampling or the lack of fertile herbarium collections. Flowers and fruits provide essential characteristics for species identification but are often not available or inaccessible high in the canopy. Hence, the species keys of Aubréville (1974) were based mainly on

vegetative characteristics, which do not always allow to distinguish closely related species. However, using molecular data, species delimitation has improved, new species have been discovered in several studies (Gautier *et al.*, 2013; Boluda *et al.*, 2021, 2022), and for a growing number of species, sound conservation assessments can be performed.

In the early sixties there were two simultaneous generic monographs (Aubréville (1964) and Baehni (1965)) of Sapotaceae with a very conflicting number of genera (122, 63, respectively). Later Pennington (1991) grouped the Sapotaceae family in five tribes comprising a total of only 53 genera. Due to the analyses of molecular data this circumscription did not turn out to be monophyletic which draws us back to the monograph of Aubréville (1964). Latest revisions left Sapotaceae with c. 1300 species in 65-70 genera (the number of genera consequently increases) grouped in the three subfamilies Sarcospermatoideae, Chrysophylloideae, and Sapotoideae (Swenson *et al.,* 2020). The whole family displays a very high proportion of species endemic to Madagascar, including the strictly endemic tribe Tseboneae with three genera *Tsebona*, *Bemangidia*, and *Capurodendron* (Figure 2). Just few species in the genera *Donella, Gambeya, Manilkara, Mimusops,* and *Sideroxylon* are also found outside Malagasy region (including Mascarenes and Comoros).

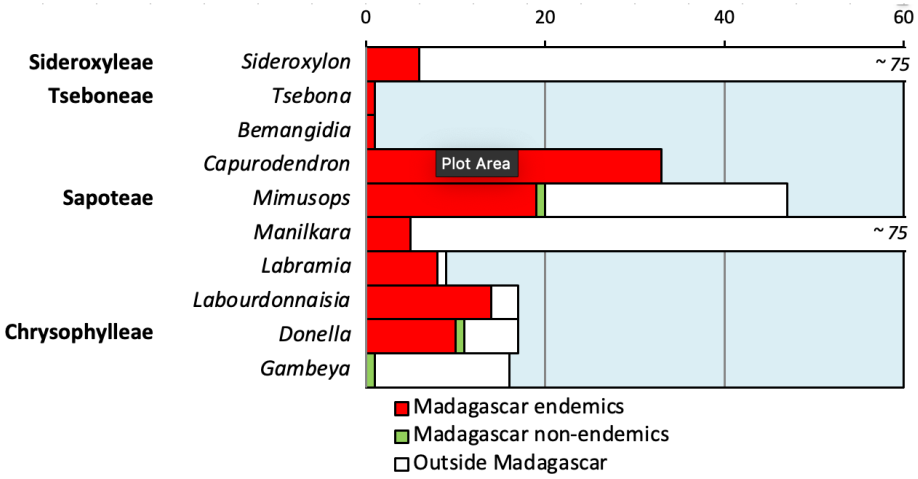

Figure 2. Endemicity graph of the Sapotaceae family showing to what extent species of each genus are endemic to Madagascar (red), non-endemic in Madagascar (green), or are found only outside Madagascar (white) (Gautier et al. in press.).

The present study focuses on the genus *Donella* (Pierre ex Baill. (1892)) with ten out of the 17 currently accepted species endemic to Madagascar. Six further species outside of Madagascar

are endemic to tropical Africa whereas only *Donella lanceolata* occurs from eastern Madagascar to Queensland and the Solomon Islands.

Donella was first recognized by Pierre but published by Baillon (1891). It was later lectotypified on the species *Donella roxburghii* (G. Don) Pierre ex Lecomte (basionym *Chrysophyllum roxburghii* G. Don 1838; syn. *Donella lanceolata* (Blume) Aubrév. (basionym *Nycterisition lanceolatum* Blume 1826). Later it was put within the broad circumscription of *Chrysophyllum* (Pennington, 1991; Schatz and Gautier, 1996), which was however shown not to be monophyletic (Bartish *et al.*, 2005; Swenson & Anderberg, 2005; Triono *et al.*, 2007, Swenson *et al.*, 2008; Bartish *et al.*, 2011). The latest revision proposes a reinstatement of the genus *Donella*, with inclusion of the small genus *Austrogambeya* Aubrév. & Pellegr. (Mackinder *et al.,* 2016). Currently, *Donella* belongs to the tribe Chrysophylleae which also comprises the genus *Gambeya*. Both genera show tiny 5-merous flowers with short corolla lobes lacking appendages and usually 5-seeded fruits (Gautier *et al.,* in press.). For distinguishing the two genera, Aubréville (1961) was using leaf venation patterns. According to this, *Donella* species could be recognized by brochidodromous venation, often with numerous secondary and parallel intersecondary veins, whereas eucamptodromous venation with prominent secondary veins, but missing intersecondary veins, are typical for *Gambeya* species. As described above, the classification of the broadly circumscribed *Chrysophyllum sensu* Pennington (1991) is based on vegetative characteristics (mainly leaf venation) while characteristics of fertile material are neglected (Aubréville, 1974; Schatz and Gautier, 1996). Likewise, the latest key of the Malagasy *Donella* species is based on leaf characteristics (Mackinder *et al*., 2016).

Among Malagasy *Donella* species, one is listed as Critically Endangered (CR) (*D. ranirisonii*), two as Endangered (EN) (*D. fenerivensis, D. guereliana*) and two as Vulnerable (VU) (*D. ambrensis, D. delphinensis*) following the IUCN criteria (Mackinder *et al.,* 2016, https://www.iucnredlist.org accessed on 23.05.2022).

Hypothesis

The first research question of this study deals with the morphologically close species *Donella delphinensis* and *D. analalavensis* and whether they are truly distinct. They are only distinguishable through leaf pubescence and blade apex shape. *Donella delphinensis* is found in humid littoral forests in the Eastern Domain (SAVA to Anosy), whereas *D. analalavensis*

occurs in dry deciduous forests in the Western Domain (SAVA to Boeny). Their distribution area overlaps in the North. Perhaps both represent only one species occurring along an environmental gradient. We further aim to investigate how they relate to *D. fenerivensis* and *D. aff. fenerivensis*. Like *D. delphinensis,* specimens attributed to the former two morphologies occur in the lowland moist evergreen forests in the Eastern Domain (Analanjirofo, Atsinanana). Morphologically, *D. fenerivensis* differs from the other two species by its obovate leaf shape, with the widest diameter in the upper third of the lamina. The available material of the morphospecies (a morphologically delimited group, described or not, which may or may not be determined to be a valid species, Boluda *et al.,* 2022) *D. aff. fenerivensis* suggests a relationship with *D. fenerivensis*, but with minor morphological variations.

The second research question relates to the *D. perrieri* complex and nearby species like *D. humbertii*, *D. capuronii* and *D. masoalensis*. *D. perrieri* is the most common and morphologically variable species, which is found in all types of moist evergreen forest, from sea level up to about 2000 m elevation along the Eastern and Sambirano Domain. In contrast, the morphologically similar species *D. humbertii* occurs in dry deciduous forests from Boeny to Melaky in the Western Domain. We hypothesize that it could be merely a seasonally dry forest form of *D. perrieri*. Likewise, *D. capuronii* is known only from a restricted number of specimens found in the parts of lowland moist evergreen forests with highest annual rainfall in the Eastern Domain (SAVA and Analanjirofo) and so could merely be a perhumid form of *D. perrieri*.

Donella masoalensis occurs in the northern parts of the Central Domain (Region SAVA) with one strange samplesite in the Eastern Domain (Fianarantsoa). It is characterized by a very strong venation and thick leaves. However there are morphological intermediates with *D. perrieri*. This issue will be also adressed with the aim of disentangeling species delimitation of these morphological similar species. All sampled specimens are pictured in the maps in Figure 3.

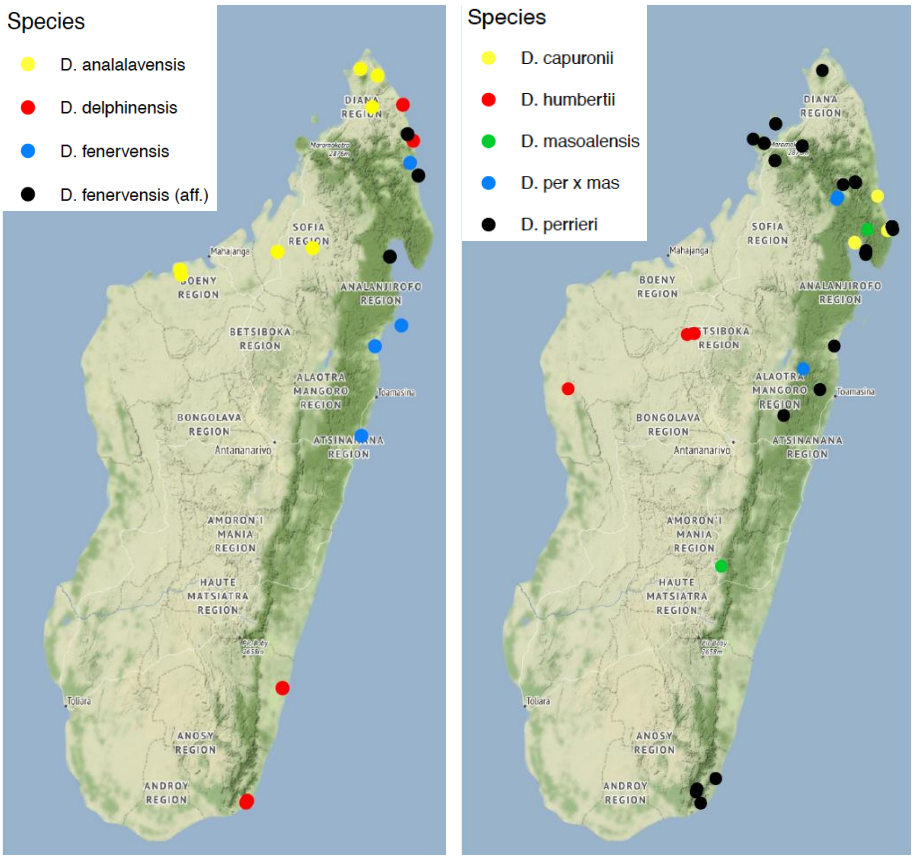

Figure 3. Maps of Madagascar, picturing the location of the sampled herbaria specimens dealing with the first (left) and the second research question (right) of this study.

Next- Generation- Sequencing

The use of herbarium material for phylogenetic studies can overcome the limitation of field collections as they contain genetic material that has been accumulated over centuries. Since this DNA material is mostly degraded over time, short-read Next-Generation Sequencing (NGS) is a suitable method to obtain useful results. However, it remains difficult to perform DNA extraction and library preparation successfully. Nevertheless, the development of NGS combined with techniques for target enrichment of selected genomic regions have revolutionized molecular biology in the last two decades (Heather and Chain, 2016). Since it is unnecessary for phylogenomic studies to assemble a full genome, sequencing effort can focus on informative low-copy genes (Jones and Good, 2016). That can be solved using specific RNA baits able to hybridize the complementary DNA regions of interest and captured genes can be amplified via PCR. This results in a higher coverage of the preselected regions, which renders this method suitable for degraded DNA from old herbarium material (Brewer

et al., 2019). This approach has been applied for Sapotaceae by Christe *et al.* (2021) who developed baits suitable for all tribes and therefore allows studies at a broad spectrum of taxonomic resolutions, from tribes to population-level studies.

2 Materials and Methods

2.1 Assembling species into morphospecies

Prior to the sampling for the molecular study, all specimens from P and G were assigned to morphospecies (a morphologically delimited group, described or not, which may or may not be determined to be a valid species, Boluda *et al.,* 2022). This was mainly based on vegetative characters, as some specimens were missing fruits or flowers. Important characters on which the classification was based are the secondary leaf venation (easy distinguishable from third venation or not), the leaf shape (eliptic, ovate, obovate, or lanceolate), the presence of pubescence and the leaf tip (elongated, pointed, blunt, rounded)

2.2 Sampling

Sampling was done from herbarium material in the Muséum National d'Histoire Naturelle de Paris (P) and the Conservatoire et Jardin Botaniques de la Ville de Genève (G). Leaf material of all 17 currently accepted *Donella* species (www.tropicos.org) was sampled with a main focus on the eleven Malagasy species. Overall, this study represents a set of 99 specimens (40 from P; 59 from G).

Table 1. Sampled and reviewed specimens of all accepted Donella species in P and G

Species name	Origin	Specimens reviewed in P	Specimens sampled in P	Specimens reviewed in G	Specimens sampled in G	Total sampled specimens
Donella ambrensis Aubrév.	Madagascar	7	2	2	2	4
Donella analalavensis Aubrév.	Madagascar	23	4	6	3	7
Donella bangweolensis (R.E.Fries) Mackinder	Africa	8	0	?	3	3
***Donella capuronii* (G.E.Schatz & L.Gaut.)** L.Gaut. & Mackinder	Madagascar	3	2	2	2	4
Donella delphinensis Aubrév.	Madagascar	17	1	11	5	6
Donella fenerivensis Aubrév.	Madagascar	23	4	7	3	7
Donella guereliana (Aubrév.) Mackinder	Madagascar	11	1	2	2	3

Donella humbertii Capuron ex Mackinder & L.Gaut.	Madagascar	5	3	0	0	3
Donella lanceolata (Blume) Aubrév.	Madagascar, Indo-Pacific	103	11	11	0	11
Donella masoalensis Aubrév.	Madagascar	16	1	11	2	3
Donella ogoouensis **(A.Chev.) Aubrév. & Pellegr.**	Africa	35	0	?	2	2
Donella perrieri Lecomte	Madagascar	112	10	44	26	36
Donella pruniformis **(Pierre ex Engl.)** Aubrév. & Pellegr.	Africa	83	0	?	2	2
Donella ranirisonii **L.Gaut. & Mackinder**	Madagascar	1	0	1	1	1
Donella ubangiensis (De Wild.) Aubrév.	Africa	0	0	?	2	2
Donella viridifolia (J.M.Wood & Franks) Aubrév. & Pellegr.	Africa	3	0	?	1	1
Donella welwitschii **(Engl.) Pierre ex Engl.**	Africa	116	0	?	2	2

To address the main research questions, the sampling prioritizes on *D. analalavensis, D. delphinensis, D. fenerivensis* and the *D. perrieri* complex including *D. capuronii D. humbertii* and *D. masoalensis*.

Since *D. analalavensis* is found in three different locations in dry deciduous forests in the north-west and extreme north, *D. delphinensis* occurs in humid littoral forests along the east coast, a minimum of two representative specimens were sampled at each locality. We further aimed to obtain a thorough sampling of the putative *D. perrieri* complex as it was collected in all types of moist evergreen forests in the north, east and south. Therefore, *D. perrieri* was divided into 11 morphospecies, while at least two representative specimens each.

Besides all Malagasy *Donella* species, six samples representing broadly the distribution of the Indo-Pacific *D. lanceolata* were included as well. Furthermore, the six African *Donella* species were sampled (1-3 specimens each) to obtain an overview of the entire genus and to examine what was their relationship with the Malagasy endemic species. See Appendix III for the list of all sampled specimens.

2.3 DNA extraction

To avoid contaminations, all plastic laboratory supplies were sterilized under UV light for at least 30 min. From each leaf fragment 10 – 15 mg were transferred in a 1.5 ml tube with two metal beads and dried for 24 hours on silica gel. Afterwards the samples were ground with metal beads at 30 hertz per second for three minutes in a TissueLyzer II (Qiagen, Hilden, Germany). DNA extraction followed a modified CTAB extraction protocol (Doyle and Doyle 1987, Appendix I). The concentration of dsDNA was measured using the DeNovix high sensitivity reaction buffer in the Qubit® Fluorimeter version 3.0 (Invitrogen, Thermo Fisher Scientific, Waltham, MA, U.S.A.). The measurement was done using 1 µl sample DNA and 199 µl of a 200:1 mix from the provided buffer (AccuClear Buffer, 1X) and dye (AccuClear Dye, 100x). The samples were subsequently stored at -20 °C until further processing.

2.4 DNA fragment analysis

The size of the extracted DNA fragments was analyzed using a bio-fragment analyzer Qsep100 (Bioptic Inc., Palm Springs, USA) following the standard protocol (https://www.labgene.ch/qsep/533-qsep100.html). Samples were prepared with 8 µl dilution buffer (Bioptic Inc., Palm Springs, USA) and 2 µl DNA template. A standard cartridge was used under the gDNA (NGS) method with a sample injection of 4 kV for 10 s and separation 8 kV for 200s. Fragment size was estimated with a C109200 size marker (15-622bp; Bioptic Inc., Palm Springs, USA) and a C109100 alignment marker (20-1K; Bioptic Inc., Palm Springs, USA). In order to use the correct ratio of magnetic beads in purification, it was important to know the size distribution of DNA fragments. This minimized the loss of DNA during the purification steps.

2.5 Illumina Sequencing

2.5.1 Library preparation

Purification

Samples were purified with Sera-MagTM Speed Beads Carboxylate-Modified Magnetic Particles (GE Healthcare, Buckinghamshire, U.K.) solubilized in a PEG 8000/ NaCl Buffer (Modified Protocol from Faircloth and Glenn, 2011).

Prior to the library preparation, samples were purified with the magnetic beads in a 0.2 ml PCR tube to remove proteins, polysaccharides, or phenolic substances. Therefore the selected bead : DNA volume ratio was 2.6 : 1 to retain fragments as small as 75 bp. To ensure that DNA longer than 75 bp could bind to the beads, the samples were mixed well and incubated for 10 min at room temperature. Then the beads were retained with a magnet and the PEG was removed. The pellet was washed two times with 180 µl of 80% ethanol. The remaining ethanol was removed by pipetting, and the beads were dried at room temperature. Finally, the pellet with the DNA was resuspended in 32 µl $H_2O_{mol.bio}$, which causes the DNA to detach from the beads. After 5 minutes of incubation at room temperature, a magnet was used to retain the beads and transfer the supernatant with the DNA into a new 0.2ml PCR tube.

DNA preparation for Indexing

Library preparation first followed a modified single tube protocol for degraded DNA (adapted from Carøe *et al.*, 2018). The preparation started from the end-repair step. In order to transform “sticky ends” into “blunt ends”, the enzyme T4 DNA Polymerase (NEB, cat#M0203S) was added to digest the 3’ overhangs and fill the 5’ overhangs. Each strand ended with an adenine which was dephosphorylated at the 3’ and phosphorylated at the 5’ by the T4 Polynucleotide Kinase (NEB, cat#M0201S). The final step in the thermocycler denatured the enzymes so they could not interfere with further processes.

In the following is the adapter ligation step, 2 µl of adapter solution (containing the hybridized IS1, IS2 and ATDC3 adapter, see Appendix II) were added to ensure the binding of the dual indexing barcodes. During the thermocycler program, the adapters were ligated to the phosphorylated 5’ DNA by the T4 DNA ligase. Finally, in the fill-in step the ligase was denaturalized, and the BstDNA polymerase completes the complementary bases to the adapters and filled the gaps at the 3’. To clean the DNA from remaining products of the

previous reactions, the samples were purified using SeraPure Magnetic Beads. Since the DNA was longer after the reactions in the above steps, a lower bead : DNA ratio was chosen (2.2 : 1). The procedure of the first washing was repeated and, at the end, the DNA was eluted in 20 µl H2Omol.bio. Then, 1 µl sample DNA was used to measure the double stranded DNA concentration with the fluorometer as described above.

Since it turned out that the protocol of Carøe *et al.* (2018) does not lead to good quality reads, we switched to the protocol following Kircher *et al*. (2012) to prepare the DNA for the Indexing. Therefore, 5 µl of the Endrepair/ A-tailing Kappa Master Mix was added 25 µl of purified DNA to transform "sticky ends" into "blunt ends". After 30 min incubation in the thermocycler at 20 °C and 30 min at 65 °C the DNA was purified with Serapure Magnetic beads at 2,8X. For the following ligation of the P5 P7 adaptors, 25 µl of the ligation mix was added and thoroughly mixed. The samples were incubated at 20 °C for 20 min and purified again with Serapure Magnetic beads at 2.8X. Finally, dNTPS were used to complementary align on the adapter site. Therefore 40 µl of the adaptor fill-in mix was added and kept for 20 min at 37 °C in the thermocycler. The purification was done at 2.8X with the Serapure Magnetic beads. In contrast to the single purification step in the end of the DNA preparation for Indexing with Carøe *et al.,* 2018 protocol, the samples undergo three purification steps in the Kircher *et al.,* 2012 protocol. Despite the potentially higher loss of target DNA during the purification steps, the sequencing results were much better with the latter. Therefore, the majority of the sample were processed with the Kircher *et al.,* 2012 protocol.

Indexing

For the indexing step the NGS P7 and P5 (5 nM) indices were annealed to the complementary sequence of the DNA fragments (See list of P7 and P5 barcode primers in Appendix II). The indices act as primers during the indexing PCR and they contain a sequence able to attach to the flow cell during the sequencing. Since each sample was provided with an individual combination of two indices, it enabled a unique identification of the samples when they are pooled in further steps.

To optimize DNA yield during the indexing PCR, around 120 ng/µl initial DNA were used and the procedure was followed according to the protocol.

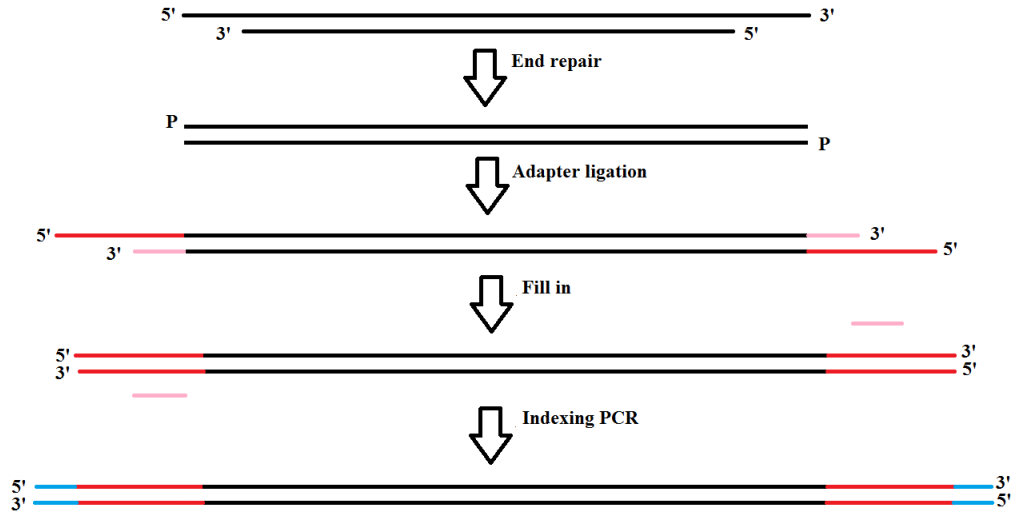

Figure 4. Illustration of the DNA during the library preparation using an adapted single tube protocol (Carøe et al., 2018). During the end repair step, the sticky DNA ends are processed into blunt ends. In the following adapter ligation step, the adaptors (red and pink) are aligned to the phosphorylated 5' ends. The gaps at the 3' are filled with complementary bases in the fill in step. Finally, the NGS P7 and P5 Indices (blue) are attached as primers during the Indexing PCR.

Quality check

Because of attached adaptors and indices, the DNA is approximately a hundred base pairs longer compared to the unprocessed DNA. To prove the increased DNA length at the end, a PCR product was used as a control. Since the genomic DNA samples contain fragments of different length (contrary to the PCR product used as control), the increase in length due to the indexing PCR cannot be visualized as a single band on a gel. The visualization was done by running a 1 % agarose gel (100 ml TAE, 1 g agarose, 5 µl GelRed (Biotium, Hayward, USA)) with the unprocessed PCR product and the PCR product after the indexing PCR. For the gel, 2 µl of a 100 bp DNA Ladder (Promega, Madison, USA) and 2 µl 6X DNA Loading Dye (Thermo Fisher Scientific, Waltham, USA) were used. After one hour at 100 V, the gel should visualize a length difference between the fragments.

Furthermore, the concentration of double-stranded DNA should increase after the indexing PCR because the indices act as primers when they are attached, which leads to DNA amplification. For this the ds DNA concentration was measured before and after the indexing PCR using a Fluorometer as described above. The measurements were compared to the calculated values (Equation 1) which simulate the DNA fragments without DNA amplification.

Equation 1. Calculation of DNA concentration of samples in ng/µl after indexing PCR, if no indices were attached to the DNA fragments

$$x = \frac{DNA\ concentration \text{ before the indexing [ng/µl]} * \text{amount of µl used for the indexing PCR}}{\text{total amount of µl in the indexing PCR}}$$

If x was equal or higher than the recorded concentration, it was interpreted as a probable library preparation failure (no DNA increase) and the protocol was therefore repeated.

Purification

DNA was cleaned from the remaining products of the previous step by using SeraPure Magnetig Beads with a bead per DNA ratio of 2.2 : 1. The procedure of the first washing was repeated and at the end the DNA was eluted in 20 µl $H_2O_{mol.bio}$. The final DNA concentration was measured using the Fluorometer as described above. The concentration of the library should be higher than 1 ng / µl otherwise the Indexing step was repeated to gain more indexed PCR products.

2.5.2 Preparing samples for Gene capture

To ensure a balanced DNA ratio between the samples, a maximum of 38 ng/µl DNA and a minimum of 19 ng/µl DNA was chosen. The volume per sample was calculated as followed:

Equation 2. Calculation of the DNA volume [µl] for pooling. The maximum quantity is 38 ng DNA per sample and for samples with less DNA concentration all DNA was used.

$$x = \frac{38\ [\text{ng}]}{Library\ concentration\ [ng/\mu l]}$$

The corresponding quantity of DNA per sample was transferred and pooled into a new 2 ml low-binding tube, so all samples could be sequenced in one lane. In total 87 samples were pooled in one tube with an approximate volume of 1.500 µl. Since a volume of 7 µl was required for the following gene capture, the volume was concentrated using the Savant SpeedVac Concentrator (Thermo Fisher Scientific, Waltham, USA).

2.5.3 Gene capture

For gene capture, the myBaits Hybridization Capture for Targeted NGS protocol (myBaits, Arbor Biosciences, Ann Arbor, Michigan, USA; v.4.01; April 2018) was used, which is an NGS

library target enrichment system. A total of three captures were performed (DON1 with 82 samples, DON2 with 36 samples and DON3 with 62 samples). In the first step, 7 µl nuclease free water was added to the tube containing the pooled libraries. Then 5 µl of the Blockers Mix following the protocol was added to the DNA and the entire liquid was transferred into a new 0.2 ml low binding tube. The mix was heated up in the thermocycler to 95 °C to denaturize the DNA (so that the baits could hybridize) and after 5 min the hybridization temperature of 62 °C was reached. Then 18.5 µl of the hybridization mix was added and the tube was incubated for 32 hours (Lane DON1); 25 hours (Lane DON2); 24,5 hours (Lane DON3) to ensure a complete hybridization of the added compounds. In this first step, the baits were able to complementarily attach to the target library molecules. Simultaneously, unwanted bonds with, e. g., adaptor molecules were blocked. The baits designed by Christe *et al.* (2021) targeted 792 low copy nuclear genes of *Sapotaceae*. In the second step, the target molecules bound to streptavidin-coated magnetic beads, which were retained using a magnet, while all non-target DNA was washed and removed. To enrich the number of targeted genes, an amplification PCR with 11 cycles was performed. This was done in duplicate to balance the bias of the PCR products (meaning the exponential increase of a random DNA fragments during the PCR which is then overrepresented in the pooled libraries) and to improve sequencing results. The concentration of dsDNA was quantified before and after the washing steps with SeraPure Magnetic Beads (2.2 : 1).

Quality control

Quality control was performed using a TapeStation System (Assay: High Sensitivity D1000 ScreenTape, Agilent Technologies Inc., Santa Clara, CA, 2019) to analyze DNA size distribution and DNA concentration of the samples. The DNA fragments should be in a range from 200 to 300 bp. Due to the fragmented herbarium DNA the size overlapped with the targed sequences and not all primer-dimers could be removed during the purification. This resulted in a visible peak ~ 140 bp. As the DNA concentration was very low in DON1, both tubes were pooled and completely dried in a vacuum using the Savant SpeedVac Concentrator. The resulting DNA was resuspended in 5 µl nuclease free H_2O and sent to sequencing. For DON2 and DON3 the PCR reaction were pooled in equal proportions to balance the bias of each PCR. The sequencing was done using an Illumina HiSeq 4000 machine (2 × 100 bp paired-end) (IGE3 Sequencing Platform, University of Geneva).

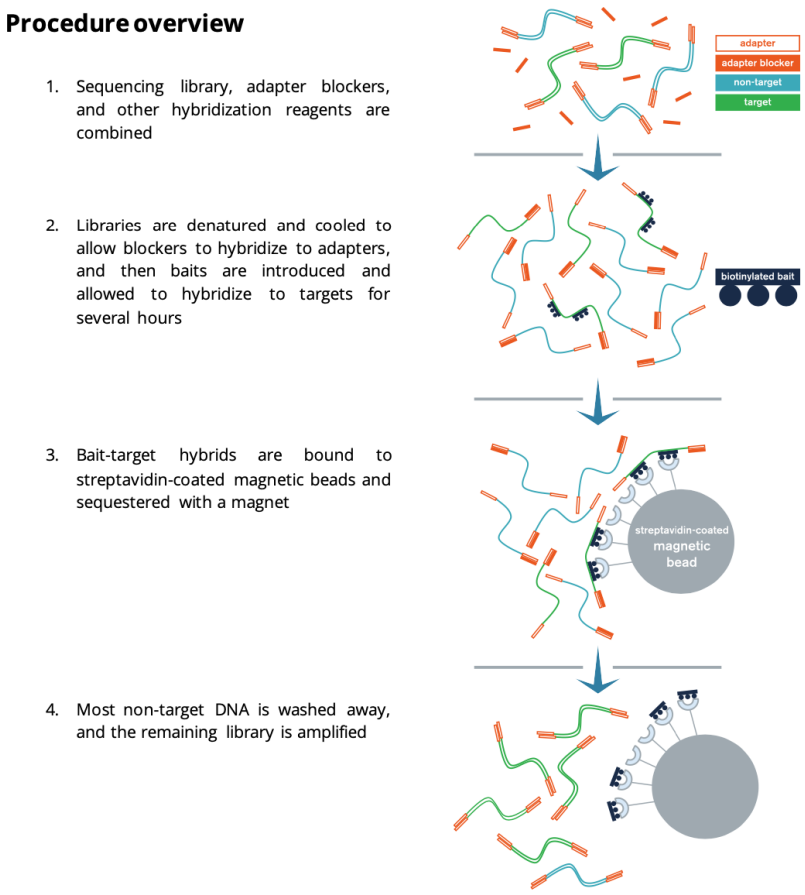

Figure 5. Schematic drawing of gene capture method which was applied to enrich the target NGS libraries. First the DNA is denatured so that the baits can bind the target DNA molecules whereas unwanted binding is blocked. This hybridization step lasts 48 hours and afterwards the bait-target hybrids bind to streptavidin-coated magnetic beads so that the non-target DNA could be washed away (myBaits Protocol – Manual v.4.01, 2018).

2.6 Phylogeny inference

The Illumina reads were demultiplexed on the iGE3 Sequencing Platform in Geneva and were processed in Baobab and Yggrasil, the high computing facilities at the University of Geneva. Firstly, a quality control check on the sequence data was done using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The thereby generated summary graphs provided a quick overview on adapter content, length distribution and duplication level. The *in silico* capture of the nuclear genes was done using two different Pipelines: Orthoskim (Pouchon *et al*., 2022) and Hybpiper version 1.3.1 (Johnsen *et al.*, 2016).

2.6.1 Orthoskim

The input files (config file, dependencies, and sample file) were prepared following https://github.com/cpouchon/ORTHOSKIM with all setting on default. One of the main advantages of Orthoskim compared to Hybpiper is, that for each library a different reference can be used for the mapping. But since there are no data available for the genus *Donella*, a nuclear reference for all samples was constructed. Therefore, the African *Donella* sample with the highest number of retrieved genes and the longest consensus sequence (*D. bangweolensis* Reekmans 7475 S103 L007) was selected as a preliminary test analysis.

Orthoskim starts with a global sequence assembly of the sequencing reads into contigs by producing one set of contigs. The reads were assembled using SPAdes (Bankevich et al., 2012) using a k-mer size of 55. Potential cpDNA, mtDNA and rDNA contaminants were removed during the cleaning step. For the following sequence capture the constructed nuclear *Donella* reference was used to assign the contigs. Thereby the minimal length of captured sequence was set on 90 with a minimal contig coverage of 3. The retrieved gene files and the gene-like files were concatenated, and multiple sequence alignments were done using MAFFT version 7 (Katoh and Standley 2013). The tool TrimAl was used for the automated removal of spurious sequences or poorly aligned regions. Afterwards each gene file was filtered, setting a threshold at 40 % missing data which discard the whole gene if the threshold was exceeded. In a second filtering step all specimen with more than 80 % missing data were removed. To infer the phylogenetic tree the maximum-likelihood (ML) tool RAxML was used on each gene (Stamatakis, 2014). The gene trees were was combined with a pseudo multispecies coalescence (MSC) method (ASTRAL-II : Mirarab & al., 2014; Mirarab & Warnow, 2015), which infers the species tree from the 692 gene trees obtained using RAxML. Since gene trees and species trees are constructed independently, ASTRAL cannot be considered a true coalescent method. The resulting species tree was visualized with FigTree v.1.4 (Rambaut, 2009) and rooted on the outgroup species.

2.6.2 Hybpiper

Before using Hybpiper, the adapter sequences were identified and removed with the tool Trimmomatic version 0.38 (Bolger *et al.*, 2014). To verify whether all adapters had been removed, FastQC was used again. Following this, Hybpiper was applied to extract the target

sequences and organize them into gene files. In contrast to Orthoskim, Hybpiper does not start with a global *de novo* sequence assembly, but first sorts the reads into genes. For doing so, the reads were mapped against the target genes which were used to capture the genes *in vitro*. Next, like in the Orthoskim pipeline, the sorted reads were assembled for each gene separately and filed into contigs using SPAdes. The coding sequences were extracted using the exonerate algorithm also applied in Orthoskim. The multiple sequence alignments as well as further downstream analysis were done the same way as in Orthoskim.

2.7 Exploring the phylogenetic tree space

A pairwise distance matrix of RAxML trees containing specimens with less than 80 % missing data was computed under the normalized Robinson-Foulds distance in the 'phangorn' R package v.2.8.1 (Schliep 2011). All gene trees were rooted on *Donella guereliana*. By using the heatmap R function directly on this distance matrix, similarity between the dendrograms could be visualized. The cells were colored according to their values in the distance matrix, which shows how all 787 gene trees resemble each other according to their topology. Furthermore, the optimal number of clusters (k) of similar gene trees was estimated with a multidimensional scaling (MDS) method. With the fviz_nbclust function from the "factoextra" R package v.1.0.7 (Kassambara & Mundt, 2020), the k value that represents the data best, was automatically chosen.

2.8 Computing the phylogenetic network

As input for Splitstree4 (Huson, 1998) a dataset was created comprising genes with less than 40% missing data and samples with less than 20 % missing data. The alignments of all 74 samples were concatenated and imported in Splitstree4. There, a Neighbor-Net network with uncorrected P-distances was computed.

2.9 Extraction of SNPs and ordination of genetic data

To retain the heterozygote sites within samples, raw data were used instead of the contigs from Hybpiper or Orthoskim. The reads were subjected to adapter and base quality using Trimmomatic version 0.38 (Bolger *et al.*, 2014). The same reference as the one used for the *in silico* capture was first indexed and then used to map the trimmed reads with BWA (Burrows-Wheeler Aligner) version 0.7.16 (Li & Durbin, 2010). The aligned bam files were

sorted, indexed and duplicates were removed using Picard tools version 2.21.1 (http://broadinstitute.github.io/picard/ (accessed on 26 April 2022) and Samtools version 1.9 (Li *et al.*, 2009). The variant calling was performed once with the Genome Analysis ToolKit (GATK) tool HaplotypeCaller version 4.1.3 in order to generate known sites for base recalibration. The GATK tool BQSR was used to detect systematic errors made by the sequencing machine and curate the base qualities. The genotype calling was done on the recalibrated BAM files by using the *MLE* subroutine of the Software ATLAS (Link *et al.,* 2017). This enables computation of the genotype likelihoods of all possible genotypes at every given SNP. The resulted VCF file was filtered for missing data (>20 %), heterozygosity and singletons using VCFtools version 0.1.16 (Danecek *et al.,* 2011). Prior to run smartpca (Zhang, 2009), .bed, .map, .fam, .bim, and .ped files were converted using plink 2.0. Afterward smartpca was executed and the resulted eigenvalues were visualized with a principal component analysis in R 4.1.2.

2.10 Bayesian evolutionary analysis

Due to the high demand of computational power of Bayesian evolutionary analyses, STACEY version 1.2.2 (Jones *et al.,* 2015; Jones, 2017) in BEAST2 (Bouckaert et al. 2014) was performed only on a small set of preselected genes processed with Hybpiper. Therefore, the alignment length, the probability parsimony informative sites and the percentage of missing data were calculated in AMAS (Borowiec, 2016). Twelve genes with an alignment length of 800 - 1250 bp, < 1.4 % missing data and a proportion of informative sites above the median (0.052) were selected and for each gene the suitable substitution model was estimated with a model test in IQ-tree version 2.2.0 (Minh *et al.,* 2020). The input for STACEY was prepared in BEAUTI2 v.2.4.5 (Bouckaert et al., 2014) putting the most variable gene in first position. The relaxed log normal clock model was chosen, and in the priors, the fossilized birth death model was selected. The growth rate was furthermore set to log normal, as well as the popPriorScale. The priors on the collapse weight was normal and the relative death rate was set on beta with alpha = 1, beta = 1. All other settings were set on default. A chain length on 200,000,000 was specified with storage every 100,000. The BEAST output was analyzed using Tracer (version 1.6) (http://beast.bio.ed.ac.uk/). Once the three separate runs in STACEY were finished, the output log files were merged in a single log file using Logcombiner (implemented in BEAST2). Out of the 7´200 of species trees from the log files, treeanotator (implemented in

BEAST2) was used to obtain a single species tree. The following species delimitation analysis was performed with the program speciesDA.jar with a burn in of 20,000 and a collapse height of 0.0001 (same as in BEAUTi). The results were visualized in R 4.1.2 by using the package “ape”. The posterior probability threshold was set on 0.01.

3 Results

3.1 Sequence data analysis

Before sequencing, the libraries were analyzed with a TapeStation System in order to check for DNA quantity and quality. The results show a small peak around 150 bp which is probably caused by adapter dimers (Figure 6). Due to the very fragmented DNA with an average size of 250 bp, it was not possible to remove all adapters without losing targeted DNA. A size between 140 -650 bp and a quantity of 388 pg/µl could be estimated.

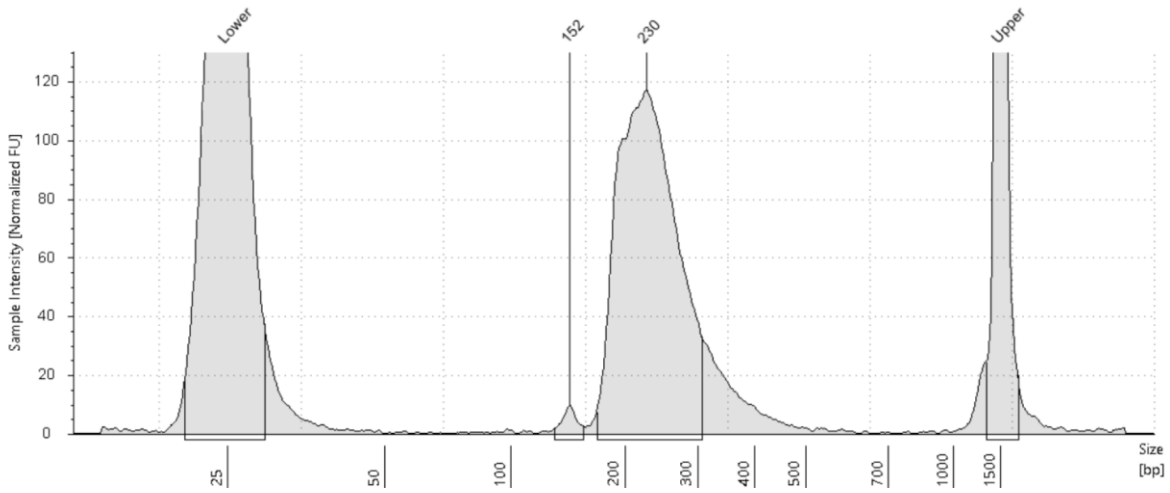

Figure 6. Quality control of constructed libraries using a TapeStation System (Assay: High Sensitivity D1000 ScreenTape). The two peaks between the lower and upper peak represent the pooled libraries and the fragment size is given on the bottom. The quantity can by estimated by the area under the curve.

The sequence read quality analysis with FastQC and MultiQC show c. 1 million unique reads per sample and 2-6 times more duplication reads. The duplication reads resulted from the PCR reaction after the gene capture and hybridization step. Furthermore, the quality score was good for all samples as well as the per base N content and the sequence length distribution (Figure 7). The adapter content remained problematic since it was not possible to remove all adapters with Trimmomatic. Nevertheless, they should not interfere with the following bioinformatic process because both pipelines are based on a k-mer assembly approach. If the depth is high enough (this was checked in Integrative Genomics Viewer, IGV; Thorvaldsdottir *et al.*, 2013) the endings of the reads should be retrieved no matter of the attached adapter.

In a first run, it was shown that only around 15 % of the reads mapped *in silico* against the baits designed in Christe et al. (2021) using the genera *Bemangidia, Capurodendron* and *Manilkara* species, so a new reference was designed. Therefore, the African *Donella* sample with the highest number of retrieved genes and the longest consensus sequence was selected (*D. bangweolensis*; collection Reekmans 7475 (G)). With this reference a mapping percentage of 20 – 80 % could be reached.

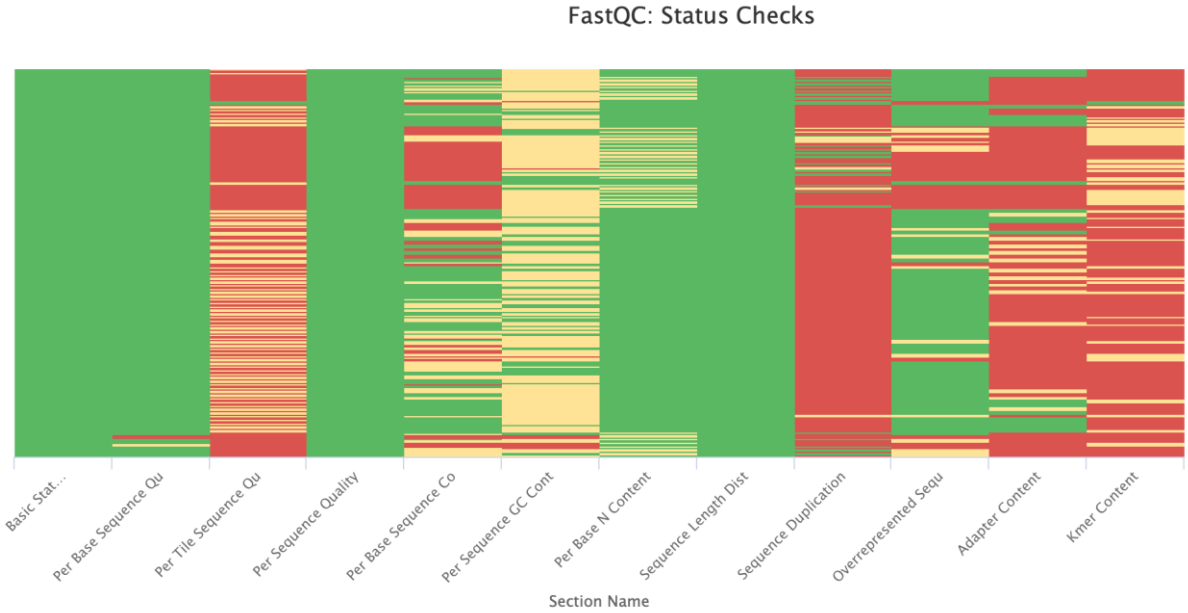

Figure 7. MultiQC report on the raw reads. Every sample (total 99) is represented by a horizontal bar. The analyzed sections are given on the bottom. Green means success, yellow means critical, red means failure.

Overall, out of 99 samples 77 passed the filtering using an upper threshold of 40 % missing data per gene and less than 80% missing data per sample. From the 24 samples processed with the Carøe *et al.*, 2018 protocol only six samples succeeded (25 % success rate). In contrast 68 samples treated with the Kircher *et al.* (2012) protocol could be retrieved in the tree (90 % success rate) (Figure 8).

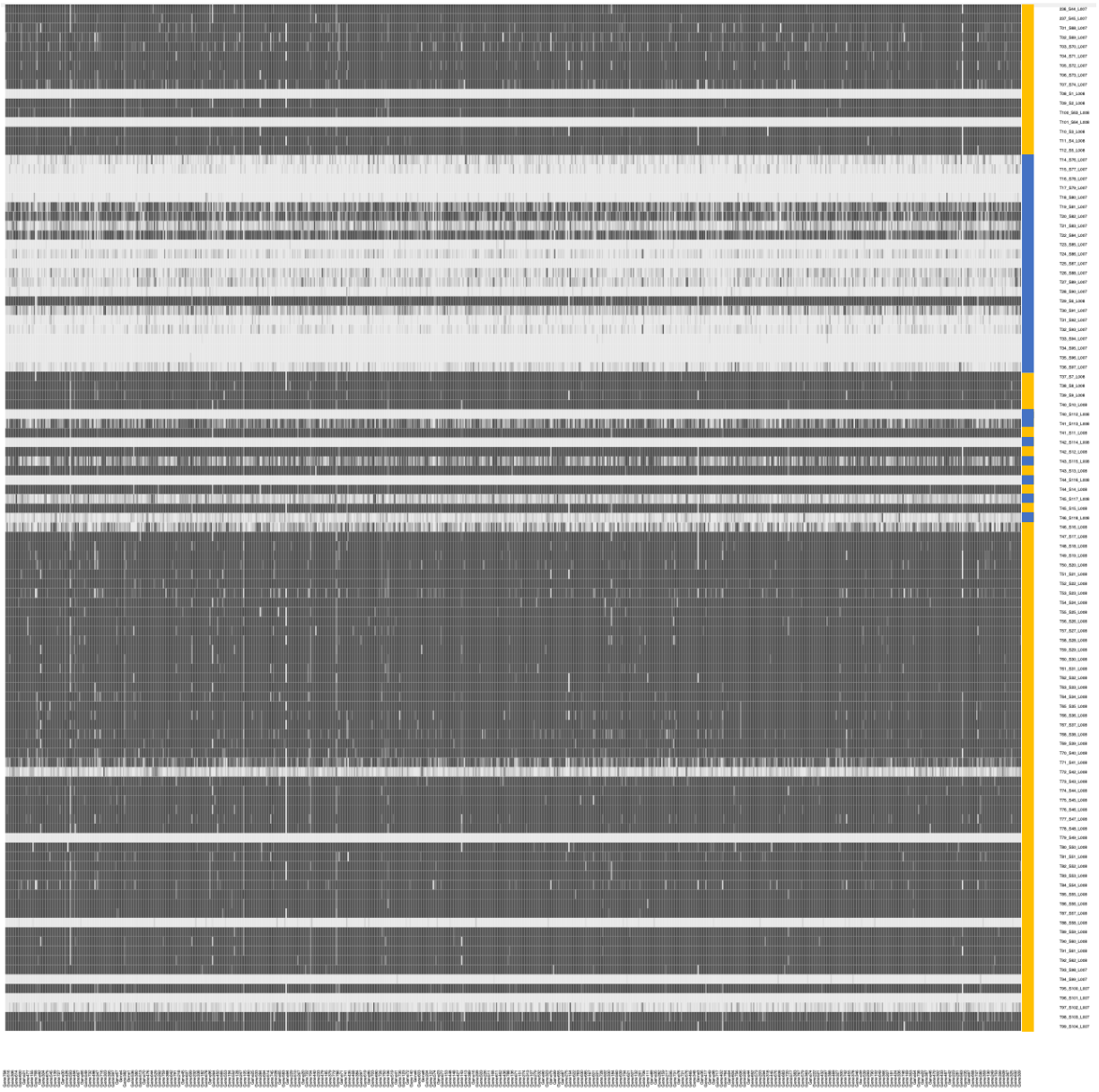

Figure 8. Heat map showing the recovery efficiency for 792 genes extracted by HybPiper. Each column is a gene, and each of the 99 rows is a sample. The shade of gray in the cell is determined by the length of sequence recovered by the pipeline, divided by the length of the reference gene (maximum of 1.0). The yellow bar on the right represents the Indexing step following the Kicherer et al. (2012) protocol, whereas the blue bar stands for the Carøe et al.*, (2018) protocol.*

With Hybpiper between 1.5 million and 17 million reads could be mapped *in silico* against the *Donella* reference. For around 90 % of the samples a maximum of 787 genes could be mapped. Only four genes (29, 241, 659, 689) could not be retrieved because they are probably not present in the genus *Donella*.

Table 2. Comparison of Hybpiper and Orthoskim using the AMAS summary.

Pipline	Hybpiper		Orthoskim	
Function	*mean*	*median*	*mean*	*median*
Nr. of taxa	75	76	73	80
Alignment length (bp)	1050	825	730	714
Missing data (%)	1,60	1,31	3,22	2,5
Nr. of variable sites	187	147	232	225
Prop. of variable sites	0,18	0,17	0,35	0,33
Pars. informative sites	59	46	125	121
Prop. of pars. informative	0,06	0,05	0,20	0,17

As shown in Table 2 the pipeline Hybpiper retrieved longer alignments with around half as much missing data than Orthoskim. However, the number and the proportion of variable sites is 24,28 % higher (respectively 94,44 %) in Orthoskim. Likewise, the parsimony informative sites were 112,83 % higher and their proportion 233,33 % higher in Orthoskim.

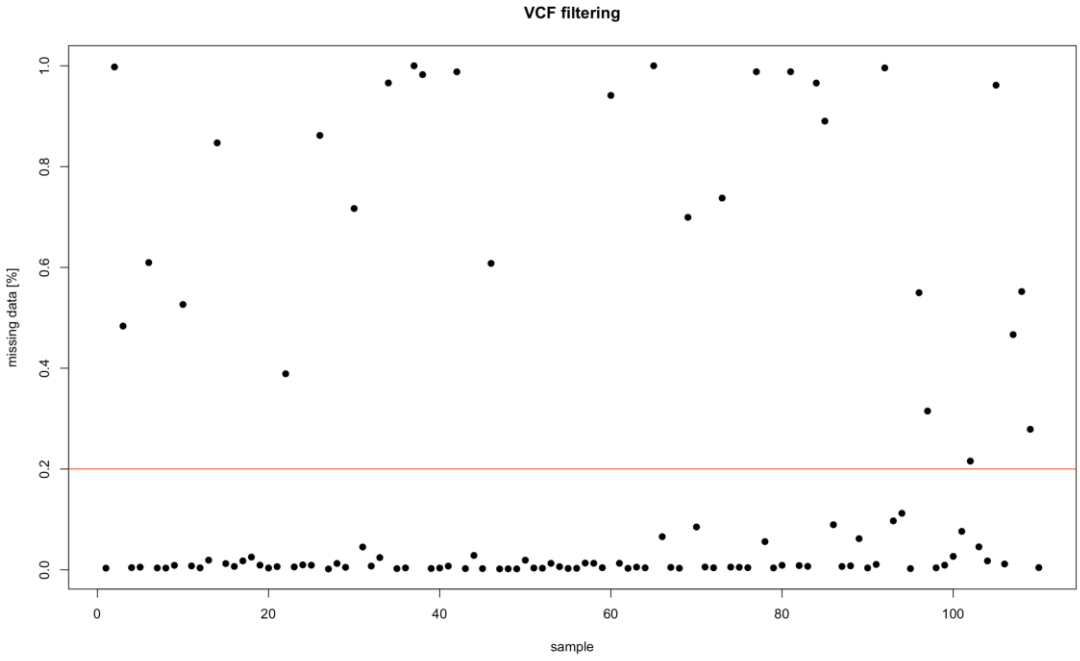

Figure 9. Overview on the missing data in the VCF file on the extracted SNPs. Each of the 110 samples is represented by a dot. The red line indicates 20 % missing data and all samples above this line were discarded.

The VCF file contains 818.623 SNPs from 110 samples (including also bad read samples processed with the Carøe et al., (2018) protocol, which were discarded from the beginning

for the *in silico* capture with Hybpiper and Orthoskim). In total, eighty-three samples show less than 20% missing data and were used for further analysis. As shown in Figure 9, 55 samples have even less than 1 % missing data. The mean depth per sample ranges from 130 to 955 which is very high.

3.2 Phylogeny inference

The phylogeny inference includes the six accepted African species and ten out of the eleven accepted Malagasy species. Only for *D. ambrensis* no sequences could be retrieved. With Orthoskim 79 *Donella* samples could be recovered in the tree, and 76 samples with Hybpiper. Among the four additional samples that Orthoskim could recover, three were lab duplicates and one is *Donella guereliana* Ratovoson 1300. On contrary, Hybpiper retrieved one more sample of *D. capuronii* (Schatz 2555).

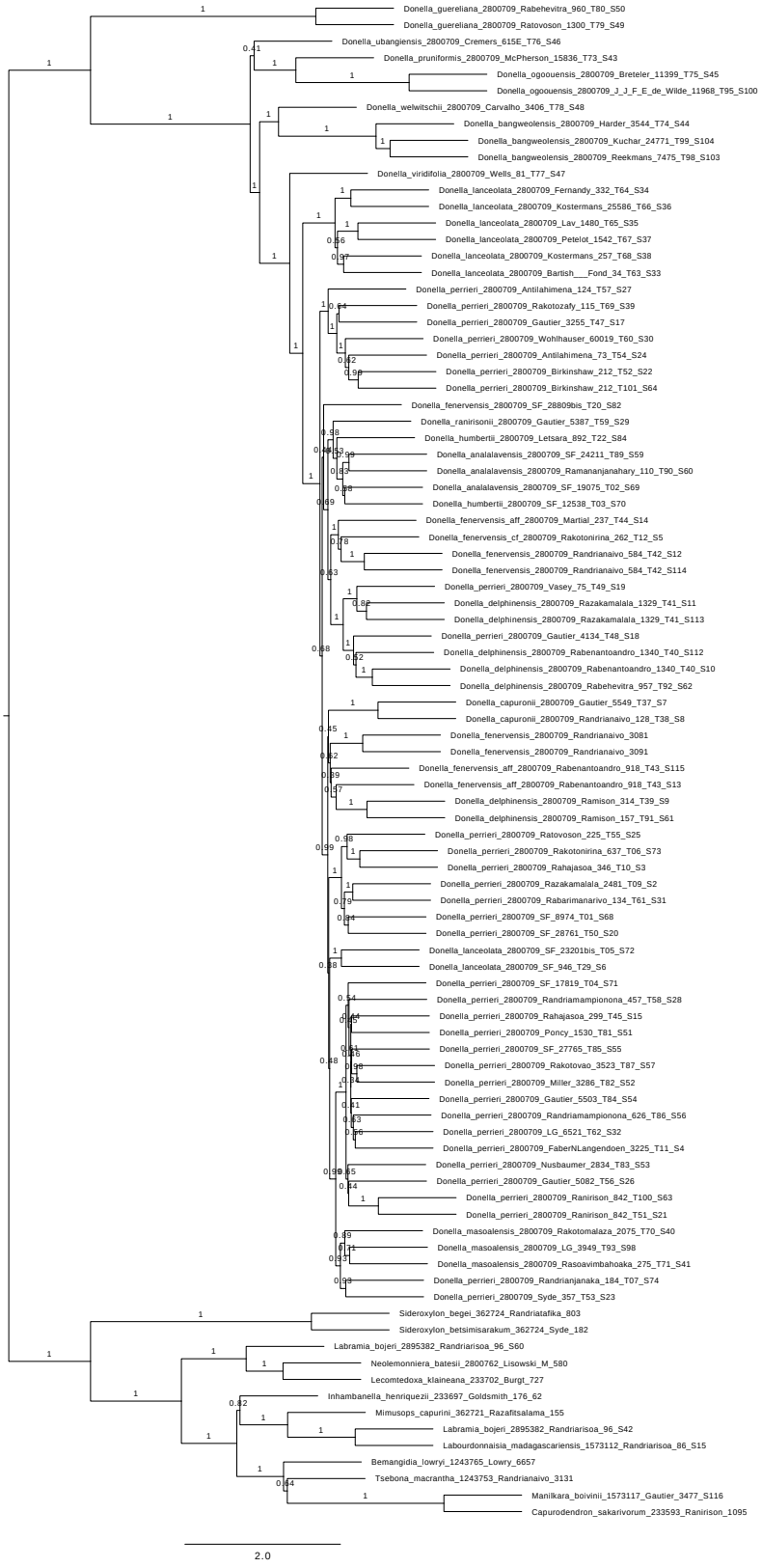

Figure 10. Pseudocoalescent phylogeny from ASTRAL inferred from 787 RAxML gene trees and rooted on species representing all other tribes of Sapotaceae. The gene sequences were retrieved with Orthoskim. All genes with more than 40 % missing data were discarded and only specimen with less than 80 % missing data over all genes are displayed in the tree. The values on the branches represent ASTRAL posterior probabilities. Species names are followed by collector and collector and collector number and end with the lab code.

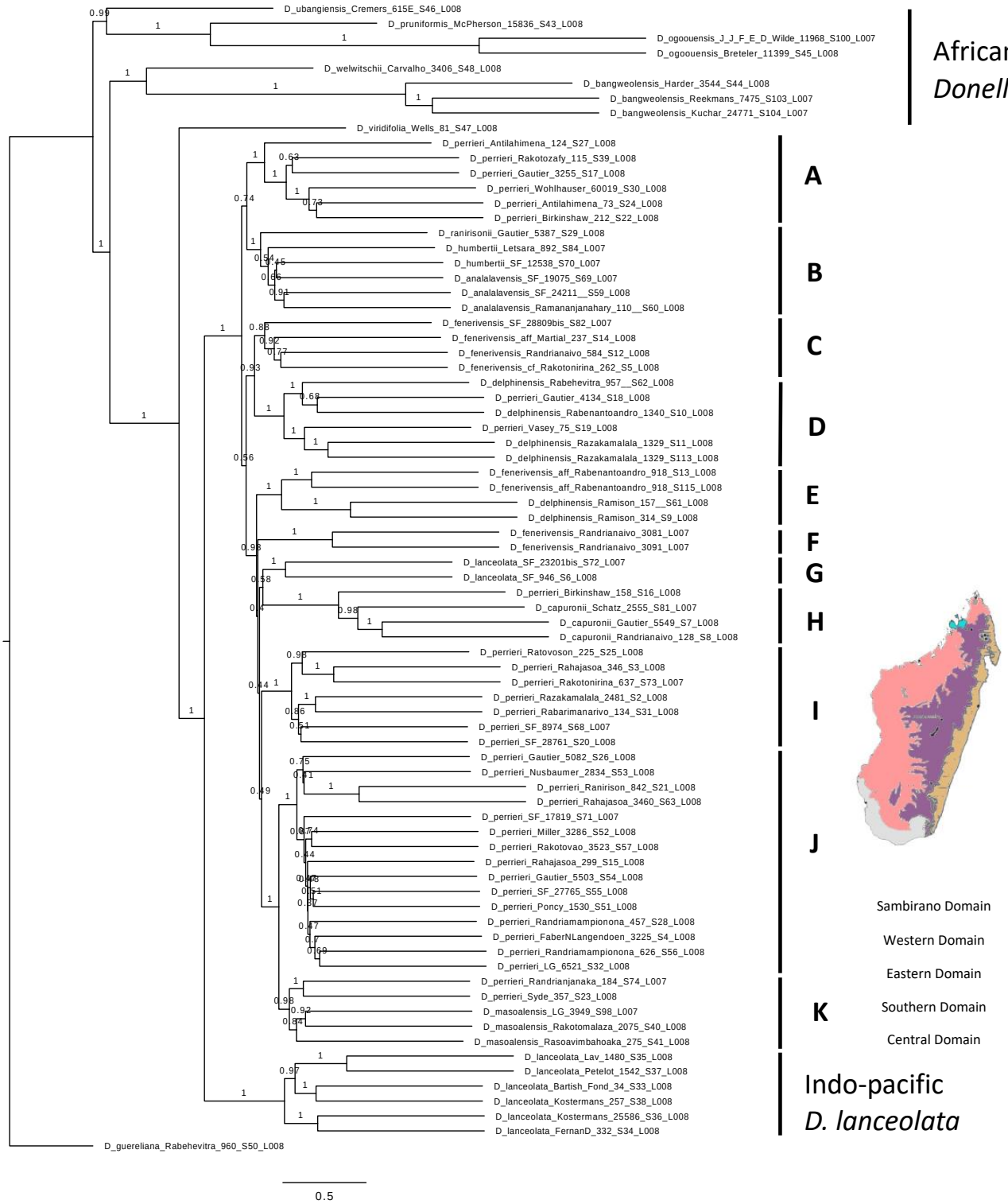

Figure 11. Pseudocoalescent phylogeny from ASTRAL inferred from 787 RAxML gene trees and rooted on D.guereliana. The gene sequences were retrieved with Hybpiper. All genes with more that 40 % missing data were discarded and only specimen with less than 80 % missing data over all genes are displayed in the tree. The values on the branches represent the ASTRAL posterior probabilities. Species names are followed by collector and collector number and end with the lab code. The Malagasy species are colored according to the sampling site and divided into clades A-K.

Both phylogenetic trees display all Malagasy species, except *D. guereliana*, in a big cluster with rather unsupported relationships among the Malagasy clades. However, the eleven smaller cluster (A-K) each have a very high support. Surprisingly, *D. guereliana* does not cluster with the other Malagasy species but appears basal to all *Donella* species, including the continental African ones (Figure 11). The mainland species split into two clades, whereas a third clade containing *D. viridifolia* appears as sister to the clade formed by the Indo-Pacific *D. lanceolata* and all remaining Malagasy species. While the branch lengths of the mainland African species show that those species have accumulated many mutations over time that allow to distinguish them, the backbone of the Malagasy species display very short and unsupported branches that blur the relationships.

The specimens attributed to *D. lanceolata*, which was believed to be the only non-endemic *Donella* species in Madagascar, are found in two well separated clusters. While the Malagasy *D. lanceolata* is imbedded within the main cluster with all the other Malagasy species (clade G), the Indo-Pacific *D. lanceolata* appears as a sister clade to the Malagasy species.

The species considered the most variable and widespread in Madagascar, *D. perrieri*, occurs polyphyletic in three different well-supported clusters (A, I, J) and is also retrieved sporadically within clusters containing other species (*D. delphinensis, D. capuronii, D. masoalensis* in clades D, H and K, respectively). The specimens attributed to *D. fenerivensis* are found in two different clusters (C, F) and an additional one together with *D. delphinensis* (E). The latter also appears in another cluster together with two *D. perrieri* (D). The species *D. humbertii* and *D. analalavensis* group together with high support (B). *Donella ranirisonii* is found basal to the latter clade (B). The three *D. capuronii* samples, which comprises its type specimen (Schatz 2555), appear monophyletic in the tree (clade H) but with *D. perrieri* (Birkinshaw 158) in a basal position.

3.3 Phylogenetic tree space

The aim of this exploration was the identification of clusters of genes that share a similar evolutionary history. Therefore, the individual gene trees were compared in order to find genes that share similar topologies, by calcultating the Robinson – Fould distance. These distances are shown as a heat map within a pairwise matrix. Figure 12 A shows that most of the genes have different topologies (dark red). However, some clustering appears

sporadically: a small cluster in light red represents gene trees with similar topologies. Controversially this clustering is not reflected in the clusters from the MDS, where two clear clusters are found based on the Robinson-Fould distance.

To compare both methods, the clusters found in the MDS are displayed on the dendrogram along the heatmap. It is shown that the results are not congruent since the clusters found in the MDS do not match the clades in the dendrogram. Therefore, it is not possible to say which MDS cluster represents the more similar gene trees in the heatmap since both clusters are represented.

An ASTRAL species tree was constructed using the gene trees from the clusters found in the MDS. Cluster 1 in the MDS comprises 324 gene trees (41.54 %) while 456 gene trees (58.46 %) support a different topology. Nevertheless, the species tree from each cluster represents the same main clades as shown in Figure 11 with minor differences (Figure 19, Figure 20 in appendix V). The topology from cluster 1 differs only in the position of the very basal *D. perrieri* Antilahimena 124 and *D. fenerivensis* SF 28809bis. Both specimens appear basal to a big clade comprising F, G, H, I, J and K. Furthermore, the species *D. fenerivensis* Randrianaivo 3081 and 3091 appear as sisterclade to clade E (*D. fenerivensis* Rabenantoandro 918 and *D. delphinensis* Raminson 157 and 314) (Appendix V). The species tree from cluster 2 genes groups *D. fenerivensis* Randrianaivo 3081 and 3091 as sister to clade K and J. In general, the branches of the species trees inferred from cluster 1 and 2 are even shorter than in Figure 11 which shows fewer mutations and therefore less information to separate the clusters.

To further investigate the extent to which the gene trees of the two cluster (Figure 13 C) differ, they were analyzed according to the parsimony informative sites, the alignment length, and the missing data (Figure 13). This analysis was done using the package violinplot in R 4.1.2. The results display no significant difference in any of the variables for the two cluster (indicated by the Mann-Whitney grouping). The clusters are represented by gene trees showing a mean in parsimony informative sites of 56.21 (cluster 1) and 62.16 (cluster 2) (Figure 14 A), very few missing data (1.66, 1.51, cluster 1, cluster 2, respectively (Figure 14 B)) and a mean alignment length of 1011.34 (cluster 1) and 1107.84 (cluster 2) (Figure 14 C). In summary, the clustering in Figure 13 C is not explained by the analyzed variables in Figure 14.

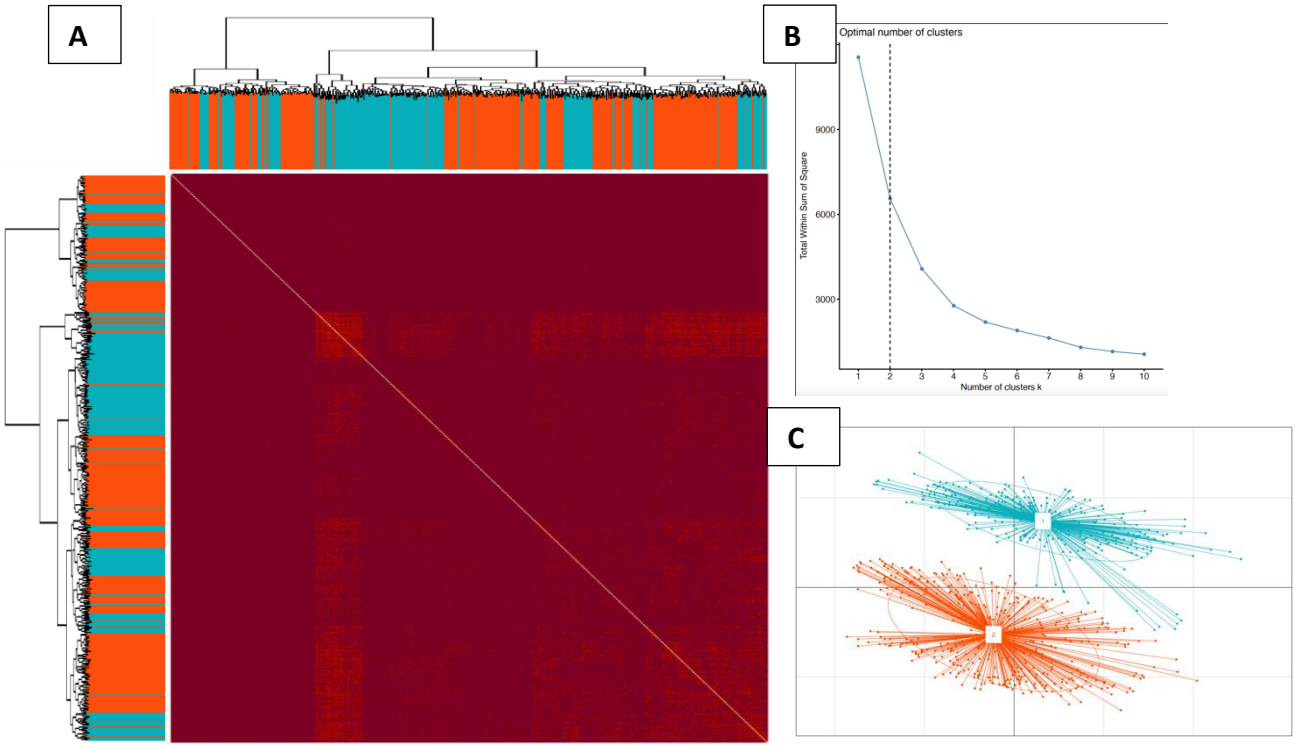

Figure 12. A: Pairwise matrix of topological distances between each pair of gene. The normalized Robinson-Foulds distance was used to compute the topological distance. Light red represents similar topology of the trees whereas dark red indicate more distance between the gene tree topologies. The dendrogram of gene trees is colored according to the clusters in C. B: optimal k-means clusters on the MDS indicated by the dotted line. C: MDS showing the variance of all gene trees topology.

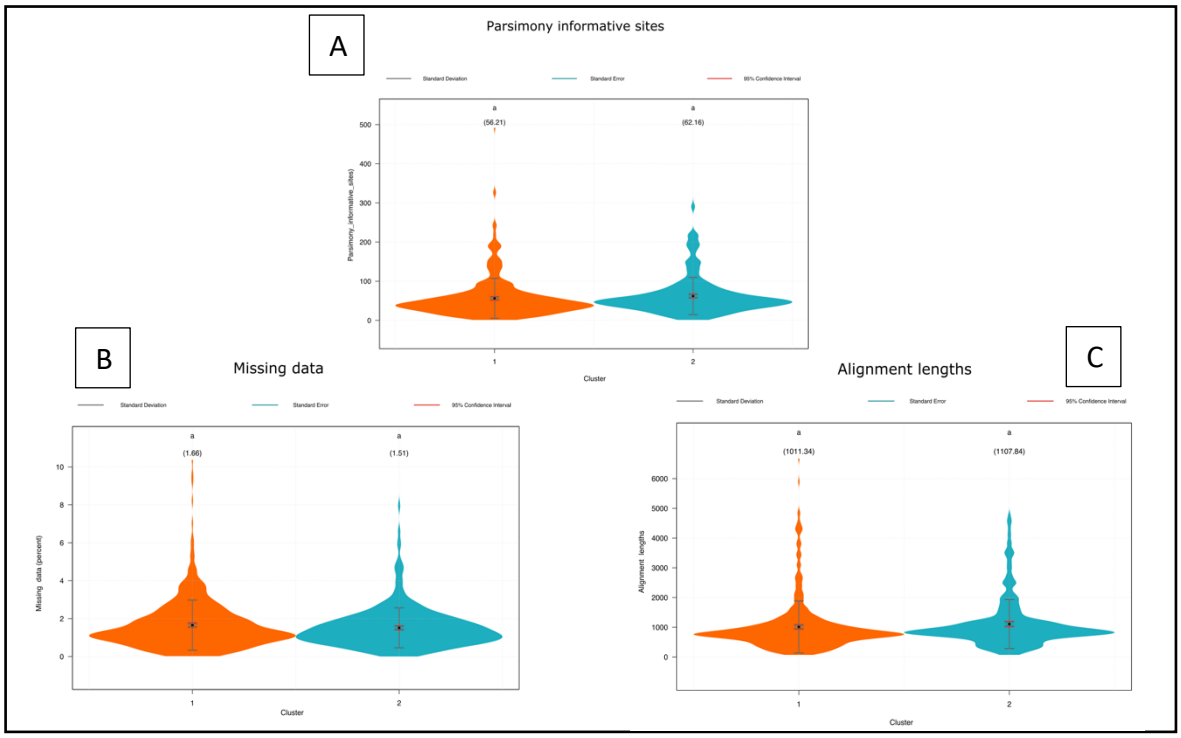

Figure 13. Violin plots comparison of the means from cluster 1 and 2 found in the k-means clustering (Figure 13) according to the parsimony informative sites (A), the missing data (B) and the alignment length (C). Mann-Whitney grouping is indicated by the letter above the respective mean.

3.4 Phylogenetic network

The computed phylogenetic network (Figure 15) displays mainly the same clusters as shown in the phylogenetic reconstruction. Striking is the very long branch of *D. guereliana*, which is four to ten times the length of the other species (Appendix V) and is shortened in (Figure 14).

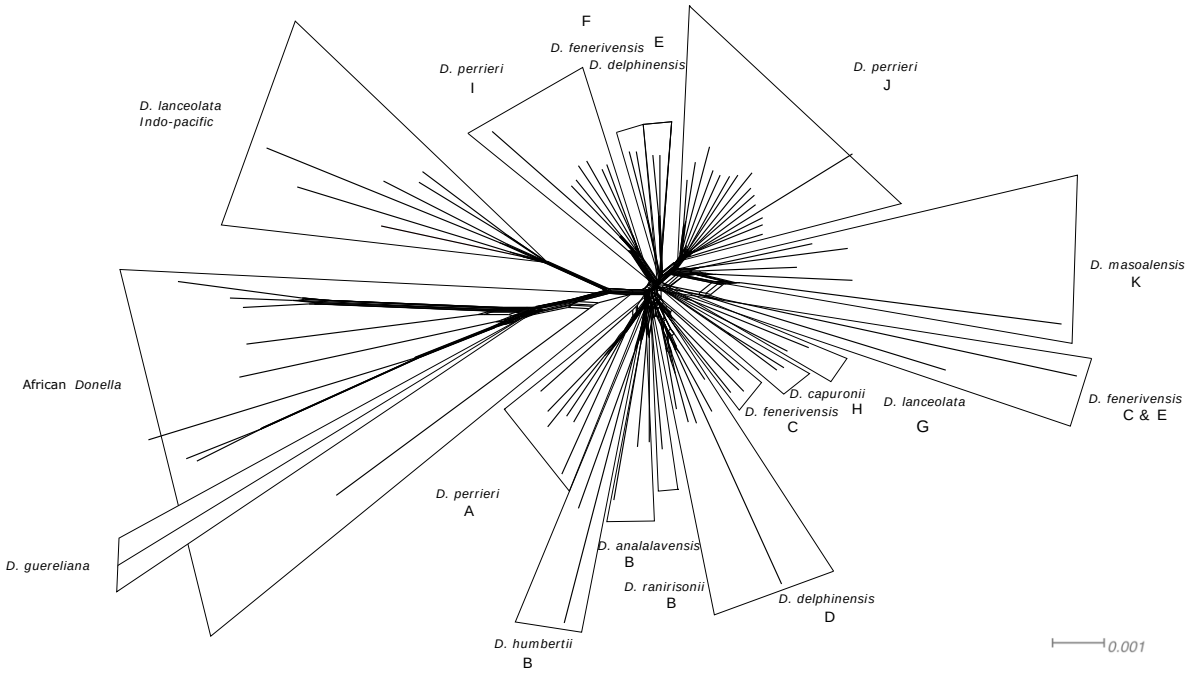

Figure 14. Phylogenetic network using the concatenated alignments of 787 genes. Sequences contained less than 20 % missing data comprising 74 samples. A Neighbor-Net with uncorrected P-distances was computed. The letters correspond to the clades in the phylogenetic tree.

The African *Donella* and the Indo-Pacific *D. lanceolata* arise from the same root which separates them from the Malagasy species. On the contrary the latter share a common root. Like in the phylogenetic tree, *D. viridifolia* appears basal to them and closest to the Malagasy species. In addition, the neighbor-net shows some gene flow between *D. viridifolia* (the lowest branch from the African *Donella* cluster) and the rest of the African species. Closest to the latter cluster is the *D. perrieri* clade A with Antilahimena 124 very basal and with some gene flow with the rest of the cluster. The *D. perrieri* clade J is found on the other side of the network being very uniform except for Randriamampionona 475 appearing on a long branch. The *D. masoalensis* clade arises close to them and from the same root and displays rather ancient and rather recent gene flow with samples assigned to *D. fenerivensis* (gray: Rabenantoandro 918 clade E and SF 28809bis) clade C. The latter two *D. fenerivensis* species

appeared in different clusters in the ASTRAL tree (Figure 11, clade C and E). While Rabenantoandro 918 appeared with *D. delphinensis* (clade E), SF 28809bis was within the light-yellow *D. fenerivensis* group (clade C). Similar to the phylogenetic tree, the light-green *D. delphinensis* northern clade D also comprises *D. perrieri* Vasey 75 and Gautier 4134, which are flanking the group. This grouping reflects the two *D. delphinensis* clades from the ASTRAL tree (Figure 11, clade D and E) but showing much gene flow within the specimens from clade D. Clade B from the ASTRAL tree can be well grouped in the three species *D. humbertii, D. analalavensis* and *D. ranirisonii*. Nevertheless, their roots are nested and separated from the nearby clades A and D.

The clades A-D (exception *D. fenerivensis* SF 28809bis) are found together on the lower side of the network. These clades are sister to clades E-K in the ASTRAL tree (Figure 11).

3.5 Ordination of genetic data

The filtering for heterozygosity was done to identify contaminations or F1 Hybrids. By using the F value (inbreeding coefficient), the probability that two alleles are identical is given. Since a low F value indicates heterozygosity, samples with a very low or even negative value were examined. This was the case for sample Birkinshaw 158 of *D. perrieri* which is basal in the ASTRAL tree to *D. capuronii* (clade H). Also, the two samples Vasey 75 and Gautier 4134 of *D. perrieri* (clade D) showed negative F values and were clustered together with the Northern *D. delphinensis* in the ASTRAL tree. In addition, very low F-values were found for Antilahimena 124 (clade A), Gautier 5503 and Syde 357 samples of *D. perrieri* (clades J and K, respectively).

The genetic principal component analysis (PCA) of all SNPs (818,623) is shown in Figure 23, Appendix VI. Since the two *D. guereliana* displayed coordinates of 0.7013 and 0.6933 (Ratovoson 1300 and Rabehevitra 960, respectively), on the first axis, they appear far away from the rest and cause that the first axis explains 20.13 % of the total variation. To see the position of the other species, a zoom on this first PCA is shown in Figure 15 (in this zoom, *D. guereliana* will appear around 350 cm out of the plot). Again, the African *D. ogouensis* and *D. pruniformis* are the most distant to the Malagasy *Donella*, which are displayed in the black cloud in the upper left corner. The other African *Donella* (*D. ubangensis*, *D. bangweolensis, D. welwitschii*) are a bit closer to the Malagasy species. *D. viridifolia*, like in the phylogenetic

reconstruction, is placed the closest to them. The Indo-Pacific *D. lanceolata* are clearly separated from the Malagasy species recovered as a sister clade.

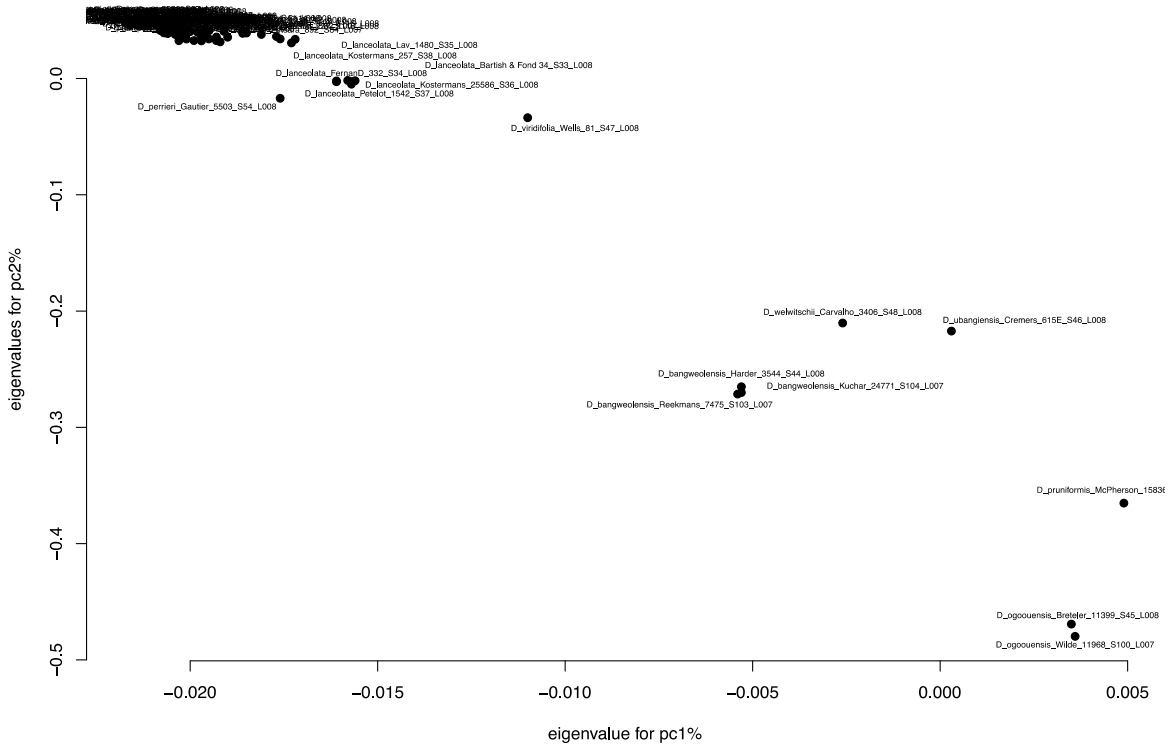

Figure 15. Principal Components Analysis (PCA) on 818.623 extracted SNPs. Both D. guereliana are not displayed as they fall far apart from all other species. Samples containing less than 20 % missing data. Species names are followed by collector name, number and lab code.

A second PCA was performed excluding *D. guereliana* as well as the African *Donella* and the Indo-Pacific *D. lanceolata.* One strange sample (*D. perrieri* Gautier 5503) which was placed within *D. perrieri* clade J in the tree, is now found far outside the other *D. perrieri* species. It showed good quality scores and no other conspicuities. As it will draw out the explained variation in the PCA from the other samples and does not lead to new findings, it was also excluded.

Figure 16 visualizes the PC1-PC2 and the PCA of PC3-PC4 is shown in Appendix VI.

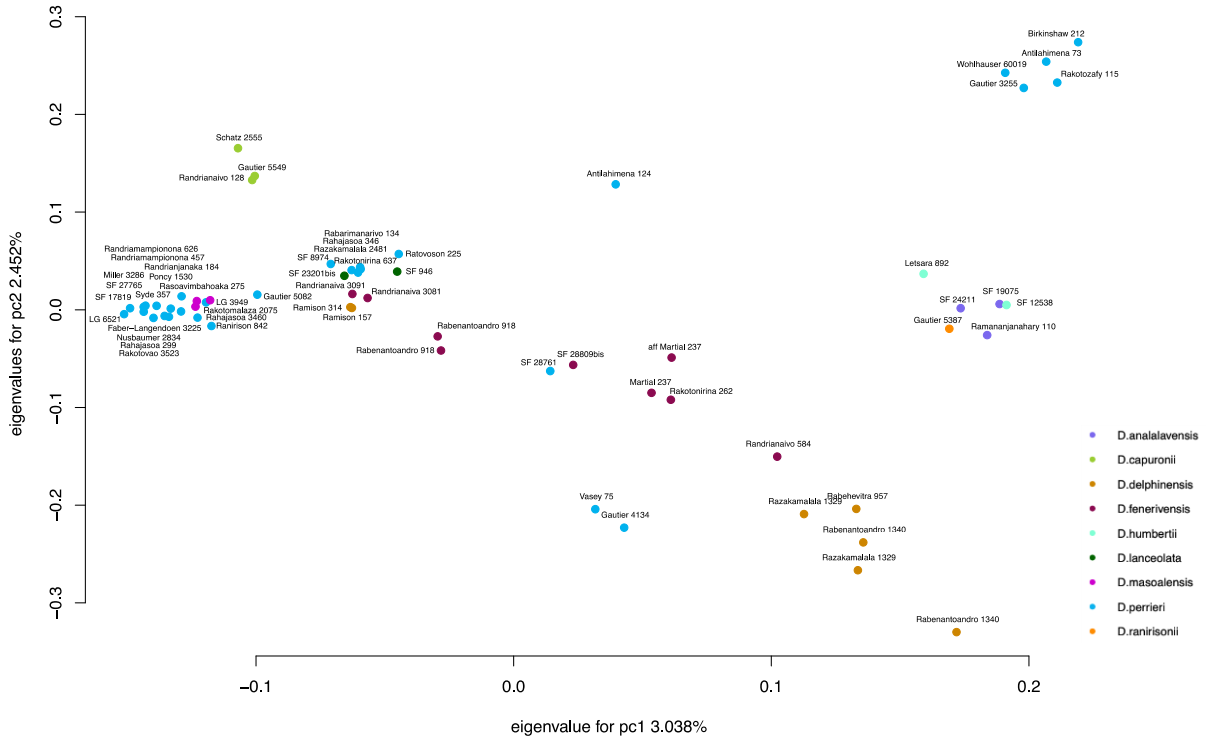

Figure 16. Principal Components Analysis (PCA) on 818.623 extracted SNPs. Samples containing less than 20 % missing data. Samples are colored by species and labeled with the collector name and number.

The first axis of the PCA in Figure 16 explains 3.0 % and the second axis 2.5 % of the variation in the data. Like in the ASTRAL tree specimens attributed to *D. analalavensis* and *D. humbertii* are grouped together (Clade B). With them D*. ranirisonii* is found. A separation for the latter has not been found in PC3 (2.4 % eigenvalue) or PC4 (2.2 %). All *D. capuronii* samples appear close to each other but missing the sample *D. perrieri* Birkinshaw 158, which was basal to them in the tree. This group was placed even further away from the rest on the PC3 and PC4. *D. perrieri* is grouped into the three same main clusters as in the phylogenetic reconstruction (A, I, J) with a few exceptions. The group on the left represents the *D. perrieri* clade J. Within this group the *D. masoalensis* clade (K) is found. The latter two groups were neither separated on PC3 nor PC4. This confirms the proximity of *D. perrieri* clade J and D*. masoalensis* since they also were sister clades in the tree. The second *D. perrieri* clade I comprises morphospecies '1' including the SF 28761 which is further apart. The third *D. perrieri* group (clade A) is clearly isolated on the right corner. Basal but within this clade is Antilahimena 124, which appears halfway to *D. perrieri* clade I and J. Interestingly the two samples Vasey 75 and

Gautier 4134 assigned to *D. perrieri*, which were clustered in the tree together with *D. delphinensis* (clade D), appear halfway to *D. delphinensis* and the *D. perrieri* clade I and clade J. Likewise, the specimen originally filed as *D. fenerivensis* (Rabenantoandro 918), which was clustered with *D. delphinensis* in the tree, is found between those species. Close to them there are two further *D. fenerivensis*, which were nevertheless clustered in a different clade (F) but as sister to the latter in the ASTRAL tree. Another clearly separated group of *D. fenerivensis* is found, which is also separated in the tree (clade C). Interestingly Randrianaivo 584, which was sampled between *D. fenerivensis* and *D. delphinensis* appears halfway to both species.

3.6 Species delimitation in Stacey

After combining the three runs of BEAST in Tracer, the statistics suggest a valid output as all EES values are good for all parameters (>200). Furthermore, the trace of the posterior values is neither skewed nor asymmetric which indicates a well fitted model.

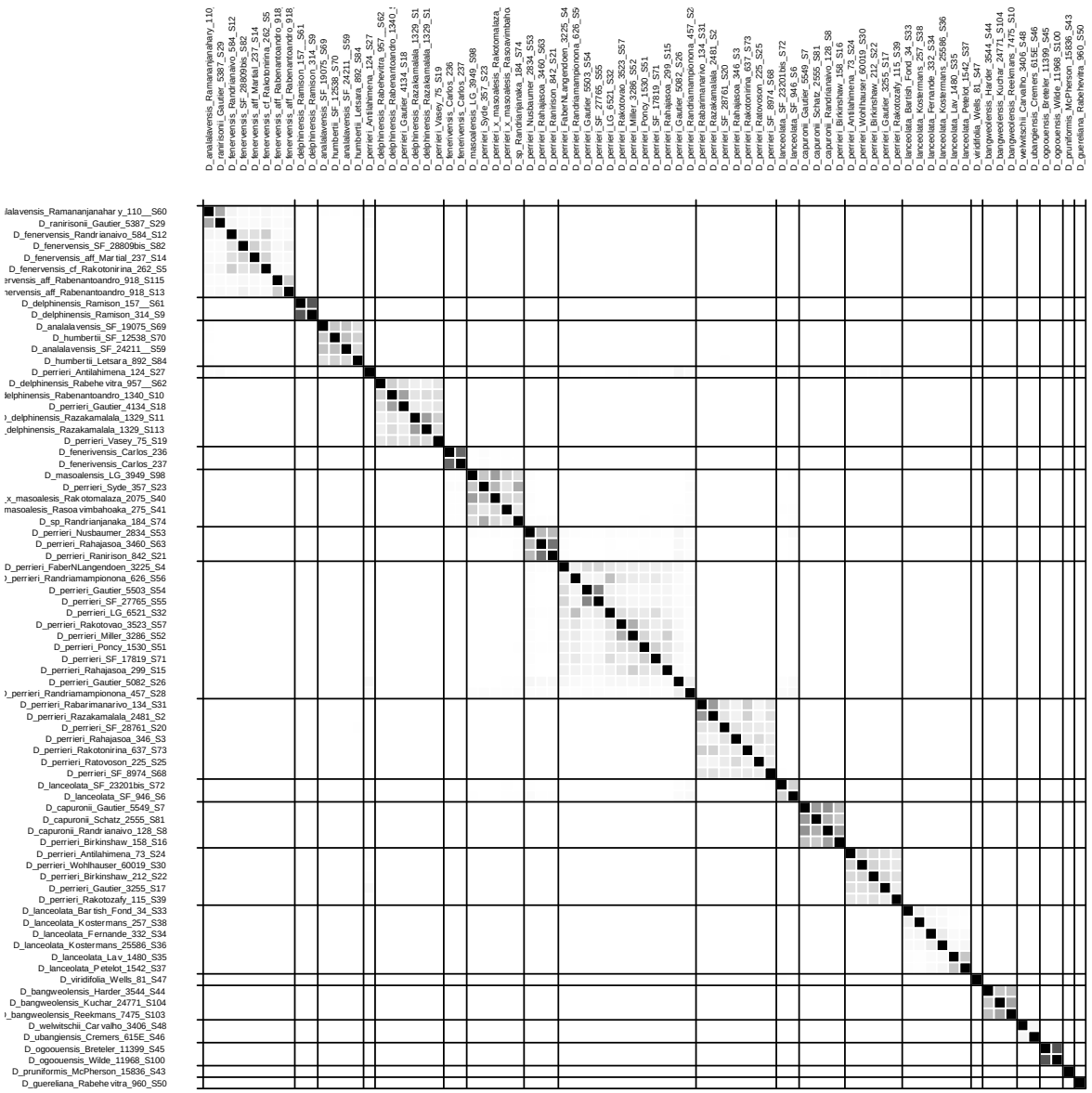

Figure 17. Similarity matrix based on a STACEY analysis performed on twelve genes with average sizes and variabilities. The grey shade of the squares indicating the posterior probability of two samples belonging to multi-species coalescent cluster (MSCC) ranging from black (PP=1) to white (PP=0). The delimitation of the MSCC was obtained automatically with at PP threshold of 0.01.

The STACEY similarity matrix based on 12 genes in Figure 17 retrieved 20 clusters (vs. 11 clades in the ASTRAL tree reconstructed from 787 genes, see Figure 11). While 14 of them group specimens originally attributed to the same species, six group specimens originally identified as different species. In comparison with the clades formed in the phylogeny (Figure 11) six clusters remain stable (D F G, H, I, K,), while one (J) split into two, another is lacking one specimen (A lacking Antalahimena124) and three clusters are formed in discrepancy with the phylogeny (B, C and E). Overall, the matrix shows very low posterior probability values for

most of the clusters. Nevertheless, the level of genetic distinctiveness indicated by the STACEY analysis is consistent with the clusters found in the phylogenetic tree. On the bottom the African *Donella* species are displayed. While samples of *D. ogoouensis* and *D. bangweolensis* are each assigned to their respective species cluster, *D. welwitschii* and *D. ubangwensis* appear in the same cluster but with low posterior probability values. Furthermore, the Indo-Pacific *D. lanceolata* are clustered together as well as the two Malagasy *D. lanceolata* which are grouped with the other Malagasy species. All *D. perrieri* clade A group together with exception of Antilahimena 124, which appears as an isolated species. In contrast, clade J is separated in one bigger and one smaller cluster consisting of Ranirison 842, Rahajasoa 3460 and Nusbaumer 2834. The latter three samples display rather good posterior probability values and were also together in a subcluster in the tree. As shown in the ASTRAL tree and the PCA, the *Donella masoalensis* with *D. perrieri* Syde 357 and Randrianjanaka 184 appear together with no pattern within the cluster. For the specimens assigned to *D. fenerivensis,* two well defined clusters are found. In contrast to the PCA and the tree, Rabenantoandro 918 appears not with *D. delphinensis* but with the big cluster of *D*. *fenerivensis*. Meanwhile the two samples *D. delphinensis* Ramison 157 and 314 form their own cluster. Like in the ASTRAL tree the two samples *D. perrieri* Vasey 75 and Gautier 4134 appear among the second cluster of specimens originally filed as *D. delphinensis*. For the first time *D. analavensis* Ramananjanahary 110 falls not within the *D. ananalavensis / D. humbertii* cluster (clade B) but group together with *D. ranirisonii* as part of the big *D. fenerivensis* cluster. *D. perrieri* Birkinshaw 158, which is basal *to D. capuronii* in the ASTRAL tree, appears within the same cluster as the latter

3.7 Morphological and geographic analysis

To understand the polyphyletic occurrence of some species, their morphology and distribution was investigated.

Donella guereliana appeared basal to all other *Donella* species in the tree but within the tribe Chrysophylleae (Figure 10). It is found in lowland deciduous forest on limestone in the Western Domain. Morphologically, *D. guereliana* displays oblanceolate leaves, which are glabrous above and below. The secondary nerves are prominent and wide and show branching before reaching the edge of the leaf. The most striking character, which

distinguishes the species from the typical *Donella,* is the 1- to 2-seeded fruit (not mentioned in Mackinder et al., 2016). Usually, *Donella* shows typically 5-seeded fruits.

Donella ogoouensis is found in Western Africa (Gabon) as well as *D. pruniformis*, which is also distributed in Central Africa and along the west coast. *D. ubangiensis* shows a distribution similar to the latter. All three species form a clade in the phylogenetic tree and are most distant to the Malagasy *Donella* species. *Donella bangweolensis* is distributed on the east side of central Africa, while *D. welwitschii* is found widely distributed on the west coast and in central Africa. The only species distributed on the African east side is *D. viridifolia*.

The phylogeny clearly shows that *D. lanceolata* is divided into two different clades (F and K): one representing the Malagasy samples and the other representing samples from the Indo-Pacific, respectively. This matches the described varieties *D. lanceolata var. malagassica* and *D. lanceolata var. lanceolata*. Nevertheless, no morphological differences in the vegetative characters could be found. *Donella lanceolata* is a morphological highly variable species throughout its large Indo-Pacific range, depending on the local climate or the soil type. Due to the clear genetic and geographic distance, no further investigation in fertile characters were made.

Since the narrow local endemic *D. ambrensis* could not be retrieved in the tree it was morphologically and geographically investigated. The only species which is distributed in the very north of Madagascar near *D. ambrensis* is *D. analalavensis*. However*, D. analalavensis* grows in inland dry deciduous seasonal forests at an elevation of 50–610 m whereas *D. ambrensis* is found in medium altitude dense evergreen forests at higher altitudes (800–1000 m). It has smaller leaves which are never pubescent and show clearly defined secondary veins, on contrary to *D. analalavensis* whose secondary veins are indistinct from, and parallel to tertiaries. Furthermore, *D. ambresis* is a large tree up to 25 m while *D. analalvensis* is a shrub or a small tree up to 12 m.

In Figure 11, the sampling site of all Malagasy species were assigned to the four Domains. Since *D. fenerivensis, D. delphinensis* and *D. perrieri* are found according to this rough categorization in the Eastern Domain, the latter was divided in a northern, a central and a southern sector following Humbert (1955).

The widespread and morphologically variable *D. perrieri* was found polyphyletic in the ASTRAL tree but forming three different purely *D. perrieri* clades (Figure 11, A, I, J). Clade A comprises samples which were mostly assigned to morphospecies '2' in the beginning of the study. They have a broad leaf base and pointed tips in common. An exception is Rakotozafy 115 which has small and narrow leaves. This demonstrates the huge variability of the vegetative characters in *D. perrieri.* All samples from clade A were collected in the Sambirano Domain, but the reverse is not true: samples originally identified as *D. perrieri* are also found outside clade A.

The *D. perrieri* clade I consists of samples which were mostly grouped in *morphospecies '1'*. They are characterized by an acute acumen and a very fine reticulated tertiary venation conspicuously raised on upper surface of driy specimens. An exception is Ratovoson 225, which is more similar to *D. capuronii* because of its very wide secondary veins and was therefore assigned to a different morphospecies in the beginning. Geographically, they are found widespread along the east coast and in the Sambirano Domain.

The third purely *D. perrieri* clade (J) consists mostly of samples assigned to the 'typical' morphospecies. There is no pattern in geographical origin as they were collected in all domains except the Central Domain. Morphologically they share a very strong midrib, and the leaves are mostly obovate with an acute leaf base, an obtuse apex sometimes with a short acumen.

For further investigations, the fertile characters of *each D. perrieri* clade were examined. All fruiting samples of clade A show rather small fruits while clade J is represented by rather big fruits. In addition, flowers of every clade were examined. All of them display five sepals and five lobes of the corolla as well as five stamens. The flower of Birkinshaw 212 from clade A show stamens attaching near the base while the corolla tube is glabrous inside. In contrast, sample SF 8974 from clade I shows stamens attaching halfway, and a rather short ovary. Furthermore, the corolla lobes are hairy on the margins but the tube is glabrous inside. Unlike the previous flowers from clade A and I, sample Nusbaumer 2834 and Faber-Langendoen 3225 from clade J show pubescence inside the corolla tube. Moreover, the two samples have the stamens attached very basal.

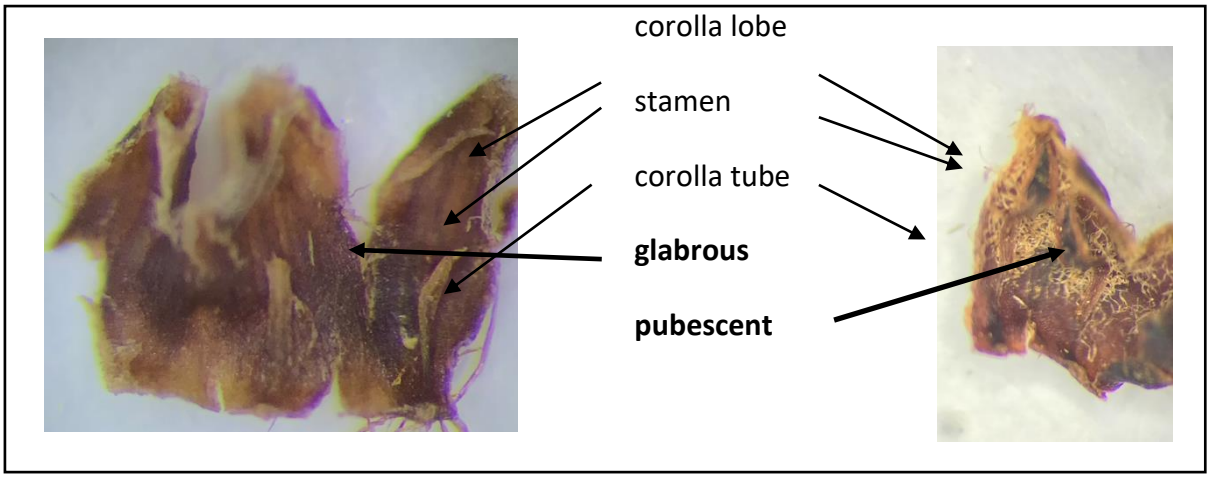

Figure 18. Zoom on the inside of the corolla in D. perrieri Nusbaumer2834 (left) and D. perrieri Birkinshaw 212 (right). Both specimens were compared in terms of the pubescence inside the corolla tube.

D. perrieri Birkinshaw 158 appears basal to *D. capuronii*. But neither the sample site nor the morphology can explain this. In contrast to *D. capuronii,* it has very narrow secondary veins which are indistinct from tertiaries, a broader leaf base and a larger leaf diameter.

The two further *D. perrieri*, that are found outside the three well supported *D. perrieri* clusters (Vasey 75 and Gautier 4134), fall together in clade D together with *D. delphinensis*. While *D. delphinensis* is characterized by secondary and tertiary veins that are hardly distinguishable from each other and clearly raised on upper leaf surface, and small obovate leaves, the two *D. perrieri* display bigger leaves that are elliptic and with reticulated tertiary veins. In addition, unlike *D. delphinensis,* the secondary and tertiary veins are easily distinguishable.

Finally, clade K comprises a group of *D. masoalensis* and two specimens originally attributed to *D. perrieri* included at the base of this clade (Syde 357 and Randrianjanaka 184). All specimens of clade K are found in the Western Domain or close by (Syde 357). Large bullate leaf blades with secondary veins that are conspicuously raised on abaxial side is the strongest character for *D. masoalensis*. Furthermore, the mature leaves keep a curly brown pubescence on the veins below. Morphologically the *D. perrieri* samples seem to have intermediate characteristics between *D. masoalensis* and *D. perrieri*. They have no bullate leaves, but stronger secondary veins than the typical D*. perrieri*. Additionally, they have a finer venation and not so much pubescence.

The specimens attributed to *D. delphinensis* are distributed in two disjunct areas along the eastern coast: one in the south of the Eastern Domain and one in the north sector of the Western Domain. Matching this distribution, two well supported clusters in the phylogenetic

tree are found (D, E). All samples of Clade D were collected in the northern part of the Western Domain or northern part of the Eastern Domain while clade E represents samples from the south of the Eastern Domain. The latter is also the type locality, and the morphology of the samples matches very well the type, with obovate leaves (as opposed to ovate in the northern specimens) that are abruptly acuminate (vs. obtuse leaf apex). However, no further major differences in vegetative morphology could be found between the two clades.

Donella analalavensis and *D. humbertii* were both collected in the Western Domain. Unfortunately, no sample of *D. analalavensis* from the extreme North (north sector of the Western Domain) yielded sufficient DNA. As shown in Table 3, the two species were analyzed for five vegetative characteristics. Both species are similar for all characteristics except for the pubescence of the leaves.

Basal to this species is *D. ranirisonii* which is characterized by very narrow leaves. The widest diameter is in the upper half of the leaf and the tip is elongated, ending with a blunt end. The venation is very fine whereas the secondary and tertiary veins are not easily distinguishable. All in all, it seems to be morphologically distant from *D. analalavensis* and *D. humbertii.* They differ especially in their leaf shape and venation pattern as well as in their geographic distribution since *D. ranirisonii* is found in the very north and the analyzed samples from the other two species are found more in the west. It should be remembered that genetic material from *D. analalavensis* samples occurring in the same locality than *D. ranirisonii* are missing in this study.

Table 3. Vegetative characteristics of D. analalavensis and D. humbertii

Species	Collector number	Herbarium	Region	Leaves glabrous --> 0 Leaves ferruginous tomentose below --> 1	Secondary and intersecondary veins so densely packed that they are not destinguishable --> 0 secondary veins destinguishable --> 1	tertiary veins parallel --> 0 tertiary veins reticulated --> 1	Leaves obovate, the widest point in the upper third of the lamina --> 0 Leaves elliptic, the widest point around the midpoint --> 1	leaf tip rounded --> 0 leaf tip pointed --> 1 leaf tip elongated --> 2
D. analalavensis Holotype	Perrier 12309	P	Boeny	**1**	**1**	**0**	**0**	**0/1**
D. analalavansis	SF 19075	P	Sofia	0	1	1	1	1
D. analalavansis	SF 24211	G	Boeny	0	1	1	1	1
D. analalavansis	Ramananjanahary 110	G	Mahajanga	0	1	1	0/1	1
D. humbertii Holotype	Perrier 8783	P	Betsiboka	**0**	**1**	**1**	**1**	**1**
D. humbertii	Letsara 892	P	Melaky	0	1	1	1	1
D. humbertii	SF 12538	P	Betsiboka	0	1	1	1	2

Specimens attributed to *D. fenerivensis* appear in three different clusters (C, E, F). These three groups are also matching with the geography but lack strong differences in vegetative morphology. However, there seems to be a gradient in leaf venation patterns. While clade F matches the type morphology having wider secondary veins that are clearly distinguishable

from tertiaries, clade C shows a finer venation with weak secondary veins. Meanwhile the sample from clade E (Rabenantoandro 918) marks and intermediate venation between the latter clades. The two samples in Clade F (Randrianaivo 3081 and 3091) were collected near the type locality Fenoarivo. In contrast, samples of clade C are found in the northern sector of the Eastern Domain. *D. fenerivensis* Randrianaivo 584 was sampled next to *D. delphinensis* Razakamalala 1329 and shows morphology (especially the venation) intermediate between these two species. *D. fenerivensis* Rabenantoandro 918 appears basal to the two D*. delphinensis* Raminson 157 and 314 in cluster E. Morphologically, the two species are easily distinguishable in their venation (*D. delphinensis* specimens with hardly distinguishable secondary and tertiary veins, whereas they are easily distinguishable in the *D. fenerivensis* specimen). In addition, the *D. fenerivensis* Rabenantoandro 918 specimen is sampled halfway from the *D. fenerivensis* type location and the *D. delphinensis* Ramison 157, 314 locations.

4 Discussion

4.1 Methods

Boluda et al. (per. comm.) showed that the single tube protocol adapted from Carøe *et al.* (2018) is very effective for old, degraded herbarium material. Nevertheless, it did not give good results in this study. Since the same problem occurred in another study conducted at the same time, we suggest that the problem is caused by the newly ordered chemicals. Our first assumption was that the ordered adapters were lacking the phosphate which could be essential for binding the DNA. But since the supplementary material of the Carøe *et al.,* (2018) protocol does not mention it as mandatory; this should not influence the binding. The question remains open however.

We found big differences in the number of variable and informative sites in the alignments retrieved with the two pipelines. Since the assembly strategy is the only main difference, we cannot explain these results. The missing data and the shorter alignment length of Orthoskim might indicate that the data retrieved with this pipeline are not as good as with Hybpiper. The higher number of informative sites retrieved with Orthoskim could then result from a lower quality of the data. Furthermore, Orthoskim recovered four more samples, which might have bad quality and therefore cause a higher number of informative sites.

Interestingly, both pipelines nevertheless led to the same topology with similar support in ASTRAL, which would mean that the errors are buffered by the huge amount of data we could obtain. We show that individual gene trees differ widely in topology and exhibit polytomies. This indicates less informative genes, which underlines the relevance of using 787 genes.

For further analysis, the paralogous sequences in each gene should be removed using the FilterParalogs.py python function from Orthoskim. Thereby polymorphic sites can be estimated in a sliding window and the respective alignment part can be discarded. This could lead to better supported branches.

In addition, the principal component analysis could be calculated on SNPs per locus to prevent overestimation of some genes. But since the PCA results are congruent with the ASTRAL tree we do not expect many changes.

To improve species delimitation, Patterson's D statistics (ABBA BABA test) and STRUCTURE can be used. The ABBA BABA test is used to test for introgression using genome-scale SNP data and therefore detect gene flow within the samples (Malinsky *et al.,* 2021). In particular, allele sharing between the putative parental species of the putative hybrids can be investigated. STRUCTURE can be used to analyze the gene pools that best fit the data (Pritchard *et al.*, 2000; Falush *et al.,* 2003). It should be used on the putative hybrids to show species isolation or introgression signals. In addition, the putative hybrids should be removed for the STACEY species delimitation analysis as it assumes random mating among species and that the coalescent of the genes is older than the speciation (Leaché *et al.* 2014, Yang and Rannala, 2010) and therefore no hybridization. Otherwise, they can obscure species boundaries of the parents (Wagner *et al*., 2020). Then this analysis could be redone using a larger number of genes to improve the posterior probability values. In another study, it was shown that 20 genes were sufficient to achieve well supported results (Boluda *et al*. 2021). It is however remarkable that nearly the same MSCC and phylogenetic clusters were recovered despite a huge difference in the number of genes used (12 and 787, respectively).

4.2 Species delimitation in a radiation

Species are a fundamental unit of biology (Mayr, 1982). However, more than 24 different named species concepts exist (Mayden, 1997), which ties the current disagreement about the theoretical concepts of the species closely to the issue of species delimitation. For instance, the stochasticity of lineage sorting within and among species, hybridization or demography impedes species delimitation as shown in Naciri and Lindner (2015). In the present study, these issues are further complicated as we aim to delimit species within a radiation process. The radiation-like evolution is indicated by the presence of many clades arising nearly at the same time in the ASTRAL tree in Figure 10 and Figure 11 (Degnan and Rosenberg, 2009). Outside this Malagasy radiation we found the continental African *Donella* well delimited. Interestingly, the Malagasy and the Indo-Pacific species splits from the mainland African *D. viridifolia* (Figure 11), which is also the geographically closest species to them. We suggest that *D. viridifolia* is the closest relative to the Malagasy and Indo-Pacific species and that some gene flow existed in the past between *D. viridifolia* and the other African *Donella*, as suggested by the neighbor network (Figure 14). The *Donella* species in central and West Africa are both genetically and geographically distant in a gradient from the Malagasy species. Since

geographic proximity is reflected in the genetic distance between mainland African and Malagasy species, we hypothesize that *Donella* has a mainland African origin and colonized Madagascar only once (already hypothesized by Bartish *et al*. 2011). This colonization event led to a fast radiation giving rise to all currently known Malagasy *Donella* species in this study. A similar pattern was observed in *Sideroxylon* and especially *Mimusops* (Boluda *et al.,* in prep). This hypothesis is supported by the network analysis in Figure 14 as it shows that there is one single root of all Malagasy species while the mainland African species belong to a longer and well separated linage (together with the Indo-Pacific *D. lanceolata* which appeared as sister to the Malagasy species in the ASTRAL trees (Figure 10, Figure 11). We suggest that the Malagasy species arose as a burst of speciation events so that it is hard to find clear differences on evolutionary histories of genes from cluster 1 and cluster 2, such as the number of informative sites (Figure 19). The two clusters shown in the MDS (Figure 13 C) were not congruent to the clustering found in the heatmap (Figure 12 A). This could result from unresolved gene trees. Besides gene properties, this might be also caused by biological reasons as suggested by Naciri and Lindner (2015). Especially hybridization and incomplete linage sorting will cause incongruence in gene tree histories. We suggest that the small light red cluster might represent genes which are highly informative. For further analyses, those should be identified and used for the STACEY species delimitation analysis. Since the species trees from cluster 1 and 2 (Figure 19, Figure 20 Appendix IV) show shorter branches and less support than the tree with all 787 genes, we conclude that more information drowns out the effects of alternative evolutionary histories on the genes.

We suggest that this approach might be better suited on a generic level (Randriarisoa *et al*. under review) or on species level with well-resolved gene trees. To make this approach suitable for a radiating genus like *Donella*, other ways to calculated topological distances should be explored. Since the Robinson-Foulds distance depends on the topology on the gene trees it might not be suitable for polytomic topologies. To test, if other distances are more convenient, the approximate SPR distance (subtree prun and regraft) (Oliveira Martins et al. 2008, de Oliveira Martins 2016) and the KF (branch score) distance (Kuhner and Felsenstein, 1994) can be used to calculate the pairwise matrix and the k-means clustering. The SPR distance selects and detaches (prunes) a subtree of the current best tree which is then regrafted onto another branch of the remaining tree. This procedure is repeated for all

regrafting positions that produce new topologies. On the other hand, the KF distance uses the branch lengths and sums the squared differences between the branch lengths.

Donella guereliana (Aubrév.) Mackinder

In all conducted analysis *D. guereliana* appears as an outlier and as something different from the other *Donella* species. It is basal to all *Donella* species in the phylogenetic reconstruction but within or as close relative to the tribe Chrysophylleae (Figure 10). In the phylogenetic network, *D. guereliana* is placed on a very long branch (Appendix V) and in the PCA it was found far apart from all other *Donella* species (Appendix VI). In addition, it displays morphological differences which distinguish it from all other *Donella* species, like the branching secondary nerves and the 1- to 2-seeded fruits. (*Donella* typically has 5-seede fruits.) *Donella guereliana* was placed by Aubréville (1974) into the small genus *Austrogambeya* which otherwise just comprised the central African *A. bangweolense.* In contrary to *D. guereliana, A. bangwalensis* has 4-5 seeded fruits. Genetic studies showed that *A. bangweolense* clusters together with what is nowadays accepted as the genus *Donella* (Bartish *et al*., 2010). According to this, both species of the genus *Austrogambeya* were placed in *Donella* (Mackinder et al. 2016). In this study, molecular data for *D. guereliana* were retrieved for the first time. Due the clear morphological and genetic differences we suggest placing *D. guereliana* in a different genus. In order to do this, a phylogenetic tree including *Gambeya* and representatives of the other genera of subfamily Chrysophylloideae should be constructed to see how *D. guereliana* is related to them. The genus *Gambeya*, which was also considered by Pennington (1991) as part of his large conception of genus *Chrysophyllum* would be particularly interesting in this context. *Gambeya* is a genus comprising 15 species in the humid forests of Africa. Since both genera are characterized by typically 5-seeded fruits (Gautier *et al*., in press.) we expect the 1-2 seeded *D. guereliana* basal to them. Other related genera of interest in such a study would be the African *Aningeria, Breviea and Malacantha*.

Donella lanceolata (Blume) Aubrév.

Donella lanceolata was found in two different clusters in the tree: one included in the Malagasy radiation and one as sister to the latter (Figure 10, Figure 11). This pattern is supported in the network, the PCA and the STACEY analysis ((Figure 14).

Figure 14, Figure 15, Figure 17). Despite morphological similarities of the clades, the geographic and genetic differences support the division of *D. lanceolata* in two distinct and distantly related species, the Malagasy one corresponding to what has been described as *Donella lanceolata var. malagassica* Aubrév.. Since the type of *D. lanceolata* (Blume) Aubrèv. was collected in Vietnam, the Indo-Pacific taxon will retain the name *D. lanceolata*. The variety *D. lanceolata var. malagassica* Aubrév. should be raised to species level. Its varietal name is unoccupied at species rank and could be used. Both species would then represent well supported monophyletic groups.

A dated phylogeny could resolve the question about the origin of *D. lanceolata* and *D. malagassica*. We believe that *D. lanceolata* dispersed directly from Africa which is indicate by their shared root in the network (Figure 14). This scenario is believed to be more likely by Bartish *et al.,* 2010. Since both species show no close genetic relationship, they were separated for a long time and evolved differently while developing similar morphologies. This is a case of morphological convergence (Naciri *et al.,* 2019). We can find the origin of *D. lanceolata* in the late Miocene while the African species are dated to the early Miocene (Bartish *et al*., 2010). Nevertheless, there is no dated information about *D. malagassica* available, and these questions remain therefore open.

Donella ambrensis Aubrév.

Since we retrieved no genetic data for *D. ambrensis* we will rely on morphological and geographical data. Despite its geographic distribution close to *D. analalavensis*, they are morphological different and do not share the same habitat. We therefore have no doubts that they are destinct species. Nevertheless, this should be confirmed genetically in the future.

Donella analalavensis Aubrev. **and** ***Donella humbertii*** Capuron ex Mackinder & L. Gaut.

In the ASTRAL tree as well as in the PCA, *D. analalavensis* and *D. humbertii* clustered together (clade B, Figure 11, Figure 16). *Donella humbertii* was described as a new species by Mackinder *et al*. (2016) but noted to resemble *D. perrieri*. This study refutes any close relationship between these two species. In contrast we hypothesize that *D. humbertii* is closely related to, or even conspecific with *D. analalavensis*, which would expand its distribution inland in the Western Domain. Due to its drier climate, *D. humbertii* probably got morphologically adapted to it, so that it was believed to be a new species based on morphology. It should be noted that the type of *D. humbertii* yielded no sufficiently good

sequences data to be included in the tree. However, it matches the morphology of the two samples represented in the phylogenetic reconstruction (Table 3) and has been collected geographically close to one of them. Since the three specimens sampled from *D. analalavensis* displayed similarities in vegetative character with the type of *D. humbertii* it would be important to include also the type of *D. analalavensis* in the tree. The differences in pubescence could be explained by adaption to the climate or was shown to disappear when leaves of the current season get older. But both species also differ in their leaf venation. Since we assume that venation shows not much plasticity within a species, this should be further investigated. Moreover, the fertile characteristics should be analyzed in both species. Especially for *D. analalavensis*, more specimens are needed (urgently the ones from the extreme North/ North – East, which could not yield sufficient DNA) to understand its delimitation to *D. humbertii* and *D. ranirisonii.* We further aim to investigate a strange narrow-leaved *D. analalavensis* specimen (SF 24500) which was collected in the Sahafary forest on sands and does morphologically not match with the type of *D. analalavensis*.

Finally, we suggest that the specimens assigned to the species *D. humbertii* and *D. analalavensis* are not well isolated from each other which is in accordance with the network results (Figure 14). There, the samples of both species form monophyletic groups, but the roots are intermingled and form a larger cluster.

Donella ranirisonii L. Gaut. & Mackinder

D. ranirisonii appeared basal to the *D. analalavensis* and *D. humbertii* group in the ASTRAL tress and the neighbor network and it groups with them in the PCA (Figure 11, Figure 14, Figure 16). On the contrary the species delimitation analysis clustered it with *D. analalavensis* Ramananjanahary 110 and apart from the rest of the *D. analalavensis* / *D. humbertii* group (Figure 17). This could indicate gene flow between those samples which is also why *D. ranirisonii* clustered with the whole group in the tree. We conclude that this clustering results from the lack of information in the solely twelve genes in the STACEY analysis as we found no evidence for this in all remaining analysss. *Donella ranirisonii* was described as a new species by Mackinder *et al.* (2016) who spotted some morphological resemblance with *D. delphinensis* (same venation pattern). No genetical support was found for such a resemblance since no genetic similarity between *D. ranirisonii* and *D. delphinensis* could be found in any of the analyses. *Donella ranirisonii* is only known from the type collection. Due to the

morphological differences with *D. analalavensis* and the genetic distance to *D. delphinensis*, we suggest retaining *D. ranirisonii* as a separate species pending further studies.

Donella capuronii (G. E. Schatz & L. Gautier.) Mackinder & L. Gautier

Donella capuronii is represented in the analyses by all currently accepted samples including the type. They are clustered in a clade in the ASTRAL tree and occur together in the network (Figure 11, Figure 14). Furthermore, they appear in a single MSCC in the STACEY matrix (Figure 17). *Donella perrieri* Birkinshaw 158 appeared basal to them in the tree and the network and displays negative F-values (inbreeding coefficient). Since a negative F-value can indicate a hybrid or a contamination (Furtwängler *et al.,* 2018), consideration must be given to this scenario. Besides a hybrid between *D. capuronii* and *D. perrieri* and a contamination, *D. perrieri* Birkinshaw 158 can also represent a new species. Further investigations on this scenario should be made by using STRUCTURE, and Patterson's D statistics as well as recover it in the PCA. The morphological analysis of vegetative characteristics could not reveal any conspicuities compared to the specimens assigned to *D. perrieri*.

Donella delphinensis Aubrév.

Specimens assigned to *D. delphinenesis* are found in two well supported clades (D and I, Figure 12) in the phylogenetic tree which match the geographical distribution of the samples. The two clades are not even sister but separated by accepted species. Since Ramison 314 and 157 are sampled near the type locality in the littoral forests of the extreme South of the Eastern Domain, we assume that these specimens represent the 'true' *D. delphinensis*. The northern clade builds another genetically well delimited species even if it displays weak morphological differences with the 'true' *D. delphinensis*.

Donella fenerivensis Aubrév.

We found a very similar scenario for the specimens assigned to *D. fenerivesis*. These samples are geographically distributed between the two clades of *D. delphinensis*. Likewise, we found two populations, one in the central sector of the Eastern Domain and one close to the northern species *D.* aff. *delphinensis* in the northern sector. The type of the species originates from the center where Randrianaivo 3081 and 3091 were sampled. As these specimens are also morphologically closer to the type specimen, we assume this to be the 'real' *D. fenerivensis*, while the other population is considered to be a new species. Nevertheless, just

weak vegetative morphological differences between them were found. The fertile characteristics should be analyzed in order to separate them. That species can be recognized from DNA sequences alone, as is shown in Cook *et al.,* 2010. The genetic analyses supports a division of *D. delphinensis* and *D. fenerivensis* into two well supported clades each (*D. fenerivensis* clade C and F; *D. delphinensis* clade D and E Figure 11). *Donella fenerivensis sensu lato* and *D. delphinensis sensu lato* might therefore represent cases of morphological convergence or conversely cases where ancestral morphological characters are retained and/or selected for. In both cases floral analysis is needed to confirm this hypothesis.

Donella perrieri Lecomte

Donella perrieri appeared polyphyletic in all analyses. Due to its widespread and variable morphology, this was already expected. In Aubreville (1974), two different varieties (*D. perrieri var. perrieri, D. perrieri var. pubescens*) and the species *D. sambiranensis* (which was originally described as a variety of *D. perrieri* by Lecomte (1928)) were retained. In the latest revision of *Donella* (Mackinder *et al.*, 2016), all three taxa were considered synonyms. The three varieties are rather scarcely described in the Flora of Madagascar (Aubréville, 1974) as the character states describing a variety are mostly not mentioned for the others. Aubréville mainly concentrated on fertile material which is not available for all specimens. *Donella perrieri var. perrieri* and *D. perrieri var. pubescens* are mainly distinguishable by the absence /presence of pubescence inside of the corolla tube but lack clear vegetative characteristics. *Donella perrieri var*. *sambiranensis* is described to have a rounded leaf base and a marked venation. In line with the existence of three varieties we found three well supported clusters of purely *D. perrieri* specimens in the ASTRAL tree (A, I and J; Figure 11). Nevertheless, they do not match with the described varieties. The type of *D. perrieri var. sambiranensis* matches the morphology in clade A as they show a broader leaf base than the samples of the other clades. Furthermore, they all originate from the Sambirano Domain where the type of *D. sambiranesis* was collected (Clade A; Figure 11). Due to its clear genetic distance (Figure 16), we suggest to consider this clade at species rank and to resurrect the name *D. sambiranensis*. Furthermore, we hypothesize that Antilhimena 124 is a hybrid between the latter and one of the other two *D. perrieri* clades (I and J, Figure 11). It appeared basal in the tree and the network (Figure 11, Figure 21), show a negative F-value and is found halfway between *D. perrieri* clade A and I/J in the PCA (Figure 16). *Donella sambiranensis* can be distinguished

from the other two *D. perrieri* clades by its smaller fruits and a broad and rounded leaf base. Further investigations in fertile characters should be made.

The genetic analyses show that clade J, representing mostly our preliminary morphospecies 'typical' is well supported in the ASTRAL tree, forms a clade in the network, and is found together in the PCA (Figure 11, Figure 14, Figure 16). The types of both *D. perrieri var. pubescens* and *D. perrieri var. perrieri* are matching the vegetative characteristics from clade J. Whereas *D. perrieri var. perrieri* is described as glabrous or sparsely pubescent inside the corolla tube, the *D. perrieri var. pubescens* was described as showing distinct pubescence. A restricted selection of flowering specimens from *D. perrieri* clade J (Nusbaumer 2834 and Faber-Langendoen 3225) showed pubescence inside the corolla tube, but this character appears variable and seems weak to retain a variety. As all the specimens of this clade display morphological resemblance with both types of *D. perrieri* and *D. perrieri* var. *pubescens*, we suggest assigning the name *D. perrieri* to clade J. However, further investigation in morphological characters should be made. Hence, we suggest recognizing them as the real *D. perrieri*. However, further investigation in morphological characters should be made.

The third *D. perrieri* clade (clade I) seems to be genetically closer to the putative real *D. perrieri* than to the putative *D. sambiranensis*. They are not sister clades in the ASTRAL tree and are far from each other in the axis 3 of the PCA (Figure 11, Figure 16, supplementary Figure 25). *Donella perrieri* clade I and J are distinguishable by venations pattern as the specimens show wide and strong secondary veins. Further investigations in fertile characters are badly needed. However, due to the results of the genetic analyses we would give this clade a species status. According to its venation we suggest calling it *D. nervosa*.

Since all three *D. perrieri* clades are very similar in their vegetative characters, and their distribution area is partly overlapping, we suggest that they might not be fully isolated yet.

Donella masoalensis Aubrev.

Donella masoalensis is genetically very close to the *D. perrieri* clade J but with clear morphological differences. Presumably they are separated only by a few numbers of key mutations which causes their different morphologies. Due to their genetic proximity, hybridization is likely, and we hypothesize that Syde 357 and Randrianjanaka 184 are hybrids. They display intermediary morphology between *D. masoalensis* and what has been described as *D. perrieri* and they appear basal in the ASTRAL tree and the network and are clustered

together in the species delimitation analysis (clade K, Figure 11, Figure 14, Figure 17). Alternatively, these two specimens could represent a new species. They do not appear intermixed with *D. masoalensis* but form a sister clade to them. Therefore, investigations in fertile characters and in further genetic analysis (STRUCTURE, ABBA BABA test) should be conducted.

4.3 Conservation assessments for taxonomically challenging groups

Conservation management should aim to retain biodiversity at three levels, ecosystem, species and intraspecific genetic diversity (Convention on Biological Diversity, 2007). Nevertheless, the majority of conservationists focus on species numbers as a measure of biodiversity. In this context, conservation assessments are expected to rely on a robust taxonomy where well delimited species are a prerequisite. This becomes challenging for species that are not well isolated and where species delimitation is difficult to assess. In addition to the genus *Donella*, we also see this scenario in many other Malagasy Sapotaceae genera (Boluda *et al.*, 2021, 2022; Randriarisoa *et al.,* submitted). Since genetic isolation/ speciation is most likely linked to reproductive characters like flower or fruit morphology, phenology, vegetative characters do not have great significance in this context. This makes the morphological study of *Donella*, where fruiting or flowering material is often lacking, very challenging. Therefore, DNA is a fantastic proxy, and we assume it much more trustworthy than often highly variable leaf characters.

Finally, we suggest that each taxon identified in this study is a unit, isolated or not, that should be considered to be conserved in order to maintain the genetic and morphological biodiversity.

Conclusion

Finally coming back to our hypothesis, we conclude that *D. analalavensis* and *D. delphinensis* are distinct despite of their morphological similarities. Yet we found hybridization between the latter with specimens assigned to *D. fenerivensis* while both species are polyphyletic and consist of two genetically well distinguished taxa. In addition, we conclude that *D. perrieri* is polyphyletic and consists of three distinct species, which are occasionally hybridizing with each other as well as with *D. masoalensis* and *D. delphinensis*.

Overall, some open questions stated in the recent morphological study (Mackinder et al., 2016) could be answered while new questions were raised. We found at least three new species (*D. nervosa sp. nov.* and the species found in *D. delphinensis sensu lato* and *D. fenerivensis sensu lato*), resurrected one species (*D. sambiranensis*) and raised one variety to species rank (*D. malagassica stat. & comb. nov*.). Accordingly, *Donella* would comprise 21 species (excluding *D. guereliana*) which increases the actual species richness in *Donella* to four more species comparing to Mackinder *et al.* (2016). Furthermore, we identified a putative new genus for *D. guereliana*.

Nevertheless, this genus remains a taxonomically challenging group where mismatches between genetic groups and observed phenotypes are present (Boluda *et al.,* 2021; Naciri *et al.,* 2019). Furthermore, the analysis was complicated by the rarity of specimens in certain taxa and especially by the lack of fertile specimens. This forced us to base morphological analysis on vegetative characters that are not reputed to be the best ones for taxonomy since they are more prone to convergence. We are therefore limited to draw final conclusions. Based on this issue, the available genetic data is probably more reliable than the morphology. This, however, might result in over-splitting of species due to the high resolution of the genomic data (Isaac *et al*., 2004). Where do we draw the line between evolving populations and durable species?

Acknowledgements

I would like to thank the Technical University of Munich and the University of Geneva as well as the Conservatoire et Jardin Botaniques de Genève, who enabled me to pursue an external master's thesis. I appreciate very much that this cooperation was possible. Many thanks to the University of Geneva, who supported me financially within the framework of the Swiss-European Mobility Program from September 2021 to June 2022. This research is part of a project leaded by Laurent Gautier and Yamama Naciri on the Malagasy Sapotaceae, financially supported by the Franklinia foundation (grant N° 2019-20). I thank the herbaria P and G for allowing me to study and sample their specimens. Furthermore, I am grateful to the people at the iGE3 platform for their help during the sequencing process (https://ige3.genomics.unige.ch).

Many thanks to Regine Niba, who helped me preparing the libraries after the first lane was not successful. Throughout the laboratory process, I was greatly supported by Carlos G. Boluda, whose excellent expertise helped me overcome the difficulties of the Sapotaceae samples. He furthermore supported me during the whole time including the bioinformatic process as well as in discussing the results. I further received great help within the bioinformatic process from Camille Christe and Charles Pouchon who both always had precious advice and excellent ideas. In addition, I really appreciate the support from Aina Randriarisoa who helped me whenever I had a problem and always encouraged me.

Finally, and most important I would like to thank my supervisors Hanno Schaefer from the TUM and Yamama Naciri & Laurent Gautier from the CJBG. I received great support from everybody. Hanno Schaefer supported me from afar and was always had precious advice. I greatly appreciate his expertise and experience and his look at this study from a different angle. Yamama Naciri and Laurent Gautier supervised me on site. I am very lucky to have been able to work in their highly precious team with expertise in both genetics and taxonomy. Whereas Yamama Naciri highly supported me with excellent advice in all areas of molecular genetics, Laurent Gautier advised me with his great expertise in the taxonomy of Sapotaceae. Many thanks to both of them for having me during my master’s project, for always supporting me and for inspiring me for my future path.

Literature

Aubréville, A. (1961). Notes sur les Sapotacées de l'Afrique quatoriale. *Notul. Syst. (Paris)* 16: 233–252.

Aubréville, A. (1964). Les Sapotacées; taxonomie et phytogéographie. *Adansonia, Mémoires* 1: 1-157.

Aubréville, A. (1974). *Flore de Madagascar et des Comores.* 164e famille. Sapotacées. Paris: Muséum National d'Histoire Naturelle, 164: 1-128.

Baehni, C.1965. Mémoires sur les Sapotacées. III. Inventaire des genres. *Boissiera* 11: 1–262 .

Bartish, I. V., Swenson, U., Munzinger, J., & Anderberg, A. A. (2005). Phylogenetic relationships among New Caledonian Sapotaceae (Ericales): molecular evidence for generic polyphyly and repeated dispersal. *American Journal of Botany*, 92(4), 667-673.

Bartish, I. V., Antonelli, A., Richardson, J. E., & Swenson, U. (2011). Vicariance or long-distance dispersal: historical biogeography of the pantropical subfamily Chrysophylloideae (Sapotaceae). *Journal of Biogeography*, 38(1), 177-190.

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., ... & Baker, W. J. (2019). Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in plant science,* 1102.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.

Boluda, C. G., Christe, C., Randriarisoa, A., Gautier, L., & Naciri, Y. (2021). Species Delimitation and Conservation in Taxonomically Challenging Lineages: The Case of Two Clades of *Capurodendron* (Sapotaceae) in Madagascar. *Plants*, 10(8), 1702.

Boluda, C. G., Christe, C., Naciri, Y., & Gautier, L. (2022). A 638-gene phylogeny supports the recognition of twice as many species in the Malagasy endemic genus *Capurodendron* (Sapotaceae). *Taxon* 71(2): 360-395.

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ,* 4, e1660.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., & Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4), e1003537.

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., ... & Baker, W. J. (2019). Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in plant science*, 10, 1102.

Buerki, S., Devey, D. S., Callmander, M. W., Phillipson, P. B., & Forest, F. (2013). Spatio-temporal history of the endemic genera of Madagascar. *Botanical Journal of the Linnean Society*, 171(2), 304-329.

Callmander, M. W., & Phillipson, P. B. (2011). The Genus *Vernoniopsis* Humbert (Asteraceae) in Madagascar. *Candollea*, 66(2), 409-412.

Callmander, M. W., Phillipson, P. B., Schatz, G. E., Andriambololonera, S., Rabarimanarivo, M., Rakotonirina, N., ... & Lowry, P. P. (2011). The endemic and non-endemic vascular flora of Madagascar updated. *Plant Ecology and Evolution*, 144(2), 121-125.

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S., Sinding, M. H. S., Samaniego, J. A. & Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*, 9(2), 410-419.

Christe, C., Boluda, C. G., Koubínová, D., Gautier, L., & Naciri, Y. (2021). New genetic markers for Sapotaceae phylogenomics: More than 600 nuclear genes applicable from family to population levels. *Molecular Phylogenetics and Evolution*, 160, 107123.

Coldrey, K. M., & Turpie, J. K. (2021). The future representativeness of Madagascar's protected area network in the face of climate change. *African Journal of Ecology*, 59(1), 253-263.

Cornet, A. (1974). Essai de cartographie bioclimatique à Madagascar (No. 55). Paris: Orstom.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A. & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics,* 27(15), 2156-2158.

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue.

Faircloth, B., Glenn, T., 2011. Homemade AMPure XP beads. *Ecol. and Evol. Biology*, Univ. of California – Los Angeles.

Gautier, L., & Goodman, S. M. (2003). Introduction to the flora of Madagascar. *The natural history of Madagascar,* 229-250.

Gautier, L., Naciri, Y., Anderberg, A. A., Smedmark, J. E., Randrianaivo, R., & Swenson, U. (2013). A new species, genus and tribe of Sapotaceae, endemic to Madagascar. *Taxon*, *62*(5), 972-983.

Gautier, L., and Naciri, Y. 2018. Three Critically Endangered new species of *Capurodendron* (Sapotaceae) from Madagascar. *Candollea* 73: 121-129.

Goodman, S. M., & Benstead, J. P. (2003). *Natural history of Madagascar.* University of Chicago Press.

Goodman, S. M., & Benstead, J. P. (2005). Updated estimates of biotic diversity and endemism for Madagascar. *Oryx,* 39(1), 73-77.

Goodman, S. M., Raherilalao, M. J., & Wohlhauser, S. (2018). The terrestrial protected areas of Madagascar: their history, description, and biota. Association Vahatra, Antananarivo, 701-715.

Harper, G. J., Steininger, M. K., Tucker, C. J., Juhn, D., & Hawkins, F. (2007). Fifty years of deforestation and forest fragmentation in Madagascar. *Environmental conservation*, 34(4), 325-333.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1-8

Humbert, H. (1955). Les territoires phytogéographiques de Madagascar. *Année biologique*, 31(3), 439-448.

Huson, D.H. SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics* 1998, 14, 68–73.

Jones, G., Aydin, Z., & Oxelman, B. (2015). DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics,* 31(7), 991-998.

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular ecology*, *25*(1), 185-202.

Jones, G. (2017). Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of mathematical biology,* 74(1), 447-467.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J. & Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in plant sciences*, 4(7), 1600016.

Kates, H. R., Doby, J. R., Siniscalchi, C. M., LaFrance, R., Soltis, D. E., Soltis, P. S. & Folk, R. A. (2021). The effects of herbarium specimen characteristics on short-read NGS sequencing success in nearly 8000 specimens: Old, degraded samples have lower DNA yields but consistent sequencing success. *Frontiers in Plant Science,* 12, 1076.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), 772-780.

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research,* 40(1), e3-e3.

Leaché, A. D., Harris, R. B., Rannala, B., & Yang, Z. (2014). The influence of gene flow on species tree estimation: a simulation study. *Systematic biology*, 63(1), 17-30.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics, 26*(5), 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics,* 25(16), 2078-2079.

Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., & Wegmann, D. (2017). ATLAS: analysis tools for low-depth and ancient samples. *BioRxiv,* 105346.

Lowry, P. P., Schatz, G. E., & Phillipson, P. B. (1997). The classification of natural and anthropogenic vegetation in Madagascar. *Natural change and human impact in Madagascar*, 93-123.

Mackinder, B., Harris, D. J., & Gautier, L. (2016). A reinstatement, recircumscription and revision of the genus Donella (Sapotaceae). *Edinburgh Journal of Botany*, *73*(3), 297-339.

Madagascar Catalogue. (2020). Catalogue of the plants of Madagascar. Missouri Botanical Garden, St. Louis and Antananarivo.

Minh, B. Q., Trifinopoulos, J., Schrempf, D., & Schmidt, H. A. (2020). IQ-TREE version 2.0: tutorials and Manual Phylogenomic software by maximum likelihood. *University of Vienna,* 134.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics,* 30(17), i541-i548.

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12), i44-i52.

Mittermeier, R. A., Myers, N., Thomsen, J. B., Da Fonseca, G. A., & Olivieri, S. (1998). Biodiversity hotspots and major tropical wilderness areas: approaches to setting conservation priorities. *Conservation biology,* 516-520.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A., & Kent, J. (2000). Biodiversity hotspots for conservation priorities*. Nature*, 403(6772), 853-858.

Pouchon, C., Boyer, F., Roquet, C., Denoeud, F., Chave, J., Coissac, E., & Wincker, P. (2022). ORTHOSKIM: In silico sequence capture from genomic and transcriptomic libraries for phylogenomic and barcoding applications*. Molecular ecology resources*.

Pennington, T. D., & Krukoff, B. A. (1991). *The genera of Sapotaceae* (pp. 307p-307p). London: Royal Botanic Gardens, Kew.

Pierre, L. (1891). *Notes Botaniques: Sapotacées*. Paris. Published in part: 1–68.

Phillips, O. L., Martínez, R. V., Vargas, P. N., Monteagudo, A. L., Zans, M. E. C., Sánchez, W. G.& Rose, S. (2003). Efficient plot-based floristic assessment of tropical forests*. Journal of tropical ecology*, 19(6), 629-645.

Rambaut, A. (2009). FigTree, v. 1.4. 0, 2006–2012.

Rasoanaivo, N. S., Tahinarivony, J. A., Ranirison, P., Roger, E., & Gautier, L. (2015). Dynamique post-culturale de la végétation dans la presqu'île d'Ampasindava, Domaine du Sambirano, Nord-ouest de Madagascar. *Malagasy Nature*, 9, 1-14

Randriarisoa, A., Naciri, Y., and Gautier, L. 2020. A new Critically Endangered species in the Malagasy Region endemic genus *Labramia* (Sapotaceae). *Candollea* 75: 83-87.

Schatz, G. E., & Gautier, L. (1996). A New Species and Combinations in Malagasy *Chrysophyllum* L.(Sapotaceae). *Novon*, 426-428.

Schliep, K. (2011). "phangorn: phylogenetic analysis in R." *Bioinform*. 27(4): 592–593.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30(9), 1312-1313.

Swenson, U., & Anderberg, A. A. (2005). Phylogeny, character evolution, and classification of Sapotaceae (Ericales). *Cladistics,* 21(2), 101-130.

Swenson, U., Richardson, J. E., & Bartish, I. V. (2008). Multi-gene phylogeny of the pantropical subfamily Chrysophylloideae (Sapotaceae): evidence of generic polyphyly and extensive morphological homoplasy. *Cladistics*, *24*(6), 1006-1031.

Swenson, U., Lowry, P. P., Cronholm, B., & Nylinder, S. (2020). Resolving the relationships of the enigmatic Sapotaceae genera Beauvisagea and Boerlagella, and the position of Planchonella suboppositifolia. *Taxon*, 69(5), 998-1015.

THORVALDSDÓTTIR, Helga, ROBINSON, James T., et MESIROV, Jill P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics,* 2013, vol. 14, no 2, p. 178-192.

Triono, T., Brown, A. H., West, J. G., & Crisp, M. D. (2007). A phylogeny of *Pouteria* (Sapotaceae) from Malesia and Australasia. *Australian Systematic Botany*, 20(2), 107-118.
ISO 690

Waeber, P. O., Wilmé, L., Mercier, J. R., Camara, C., & Lowry, P. P. (2016). How effective have thirty years of internationally driven conservation and development efforts been in Madagascar? *PloS one,* 11(8), e0161115.

Waeber, P. O., Rafanoharana, S., Rasamuel, H. A., & Wilmé, L. (2019). Parks and reserves in Madagascar: managing biodiversity for a sustainable future. In Protected Areas, National Parks and Sustainable Future. *IntechOpen*.

Wagner, F., Ott, T., Schall, M., Lautenschlager, U., Vogt, R., & Oberprieler, C. (2020). Taming the Red Bastards: Hybridisation and species delimitation in the Rhodanthemum arundanum-group (Compositae, Anthemideae). *Molecular phylogenetics and evolution*, 144, 106702.

Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, 107(20), 9264-9269.

Zhang, Y. (2009, June). Smart pca. In Twenty-First International Joint Conference on Artificial Intelligence.

Appendix I

EXTRACTION D'ADN GÉNOMIQUE

CTAB method

Sample preparation

Prepare 10~15 mg dried leaf per 1.5 ml tube

Add 2 metal beads per tube and grained with 30 hertz per second for 3'

CTAB extraction

CTAB	**CTAB 50ml**
Tris-HCl pH8.0 100mM	1 MTris-HCl pH8.0 5ml
NaCl 1.4M	5M NaCl 14ml
EDTA 20mM	0.5M EDTA 2ml
CTAB 2%	CTAB 1g

- In each tube add 1 ml CTAB buffer plus 2 µl ß-mercaptoethanol and 1% PVP40
- Shake in the tissue lyser at 20 hertz per second for 2'
- Leave at 65°C 60' (shake at the beginning and in the end)
- Shake in the tissue lyser at 20 hertz per second for 2'

-1 Add 250µl Chloroform : IAA (24:1)

Gentry shake at 40 rpm for 20'-30' on the « multibio shaker »

Centrifuge 11,000 rpm 10'

Move supernatant to new 1.5ml tube

-2 Add 500µl Chloroform : IAA

Gentry shake at 40 rpm for 20'-30' on the « multibio shaker »

Centrifuge 11,000 rpm 10'

Move supernatant to new 1.5ml tube

Repeat this step a second time

DNA precipitation

- Add 600µl icecold isopropanol (isopropanol: > 1/3 volume of solution)
- Mix by inverting tubes
- Leave at -20ºC ~30'
- Centrifuge 11,000 rpm 10'
- Remove supernatant
- Add 500µl wash buffer
- Dry pellet (not too much)
- Dissolve in 20µl TE
- Leave at 50~65ºC ~ 5', mix well and spin down

Chemicals list

Wash buffer	**To prepare100ml of wash buffer**	
Ethanol 76%	Ethanol 96%	79ml (EtOH à 99.8%, 76ml)
Ammonium Acetate 10mM	NH4Ac 5M	200µl

TE8 100ml

20ml Tris-HCl 50 mM pH8

0.2ml EDTA 0.5mM pH 8

HCl 400 mM (8mlHCl 35% + 224ml ddH2O) to clean used metal beads

Appendix II

Single tube protocol:

This is an adaptation from Single-tube library preparation for degraded DNA (https://doi.org/10.1111/2041-210X.12871) and BEST protocol (Blunt-End Single-Tube Illumina library building for modern and ancient DNA) from Christian Carøe, GLOBE institute, Copenhagen, Denmark.

This protocol is faster and cheaper than the multi tube library construction, increasing the yield of samples with few or degraded DNA. It produces less adapter dimers, which can have the same size (around 140 bp) as the target DNA. It replaces the washing steeps to remove the enzymes by heat denaturalization.

Reagents

- T4 DNA Polymerase 3 U/µl (NEB, cat#M0203S)
- T4 Polynucleotide Kinase 10 U/µl (NEB, cat#M0201S)
- 10X T4 DNA Ligase Reaction Buffer (NEB, included with the ligase pack)
- dNTP 25 mM (or 10mM)
- PEG 4000 50%. (50 g PEG+ H2O until 100 ml)
- T4 DNA Ligase 400 U/µl (NEB, cat#M0202S)
- IS1, IS2, and ATDC3 adapter 10 µM (see annex to see how prepare it)
- Isothermal Amplification Buffer 10X (NEB, included with the Bst polymerase pack)
- Bst 2.0 Warmstart Polymerase 8 U/µl (NEB, cat#M0538S)
- Molecular grade water
- MinElute column (or SeraPure Magnetic Beads)
- P7 and P5 barcode primers (both with barcode) 10 µlM each.
- 2x KAPA HiFi Hotstart ReadyMix (Roche)

Purifying the DNA

Purification can be done using the SeraPure Magnetic Beads (with around 2.5X for degraded material) or using the MinElute columns if the DNA is highly fragmented (below 200 bp).

End-repair step:

1. Prepare the next mix (quantities for one sample):

- 0.4 µl of T4 DNA Polymerase 3 U/µl.
- 1 µl of T4 Polynucleotide Kinase 10 U/µl.
- 4 µl of 10X T4 DNA Ligase Reaction Buffer NEB.
- 0.4 µl of dNTP 25 mM (or 1 µl if is at 10 mM).
- 2.2 µl of Enhancer (0.25 g PEG-4000 (25% final concentration) + 100 µL BSA (20 mg/mL; 2 mg/mL final concentration) + 80 µL NaCl (5M stock, 400 mM final concentration) + H2O up to 1 mL.

2. Transfer 8 µl of the mix in a 0.5 ml low binding tube.
3. Add 32 µl of DNA sample (250-1500 ng of ADN in total). Add water if the 32 µl are not reached.
4. Final reaction volume 40 µl. Incubate the samples 30 min at 20°C and then 30 min at 65°C. Cool to 4ºC. (If the thermocycler is warm or takes long to heat up the lid, let the reaction tubes wait on ice before placing them in the thermocycler).

Adapter ligation step:

1. When de above reaction is finished, add to each tube 2 µL adaptor solution at 10-20 µM to each reaction (it contains the hybridized IS1, IS2 and ATDC3 adapter). Mix well.
2. Then prepare the next mix (quantities for one sample):

- 6 µl PEG 4000 50%.
- 1 µl T4 DNA Ligase Reaction Buffer 10X.
- 1 µl T4 DNA Ligase 400 U/µl.
- Mix well, it can be difficult because PEG is very viscous.

3. Add 8 µl of the Ligase prepared mix and mix well by pipetting.
4. Total reaction size of 50 µl. Incubate the samples 30 min at 20°C and 10 min at 65ºC, cool to 4ºC. (If the thermocycler is warm or takes long to heat up the lid, let the reaction tubes wait on ice before placing them in the thermocycler).

Fill-in step:

1. Prepare the next mix (quantities for one sample):

- 0.8 µl of dNTP 25 mM (or 2 µl if is at 10 mM).
- 2 µl of Isothermal Amplification Buffer 10X.
- 5.6 µl of molecular biology grade water.
- 1.6 µl of Bst 2.0 Warmstart Polymerase 8 U/µl.

2. Add 10 µl of the reaction mix to each sample.
3. Incubate 20 min at 65°C in a prewarmed thermocycler and then 20 min at 80°C, coold down to 4ºC. (If the thermocycler is warm or takes long to heat up the lid, let the reaction tubes wait on ice before placing them in the thermocycler).

Cleaning step:

Purification can be done using the SeraPure Magnetic Beads (with around 2.5X for degraded material.

Indexing step:

1. Use for each sample:
    - 4 µl of your DNA with the adaptors (more can be added decreasing the H2O).
    - 7.5 µl of ddH2O .
    - 12.5 µl of 2x KAPA HiFi Hotstart ReadyMix.
    - 1 µl of each primer premix (NGS P7 indexed and NGS P5 indexed at 5 µM each, or 0.5 µl if are at 10 µM).

    The total reaction volume is 25 µl.
2. Perform the PCR with the parameters:
    - Denaturation at 98 °C for 30 sec.
    - 8 cycles of:
        Denaturation at 98 °C for 10 sec.
        Annealing at 60 °C for 20 sec.
        Elongation at 72 °C for 20 sec.
    - Final extension at 72 °C for 20 sec.

Quantification:

Use 1 µl for Qubit® Fluorimeter DNA quantification before the cleaning. Make the calculations of the DNA you should have in the tube (using your initial amount of ng of DNA and the final volume of liquid). If the DNA concentration is higher than expected, the barcodes has been inserted and DNA replicated, if not, library construction may be wrong.

Cleaning:

Clean the samples using the SeraPure Magnetic Beads (with around 2.5X for degraded material) or using the MinElute columns if the DNA is highly fragmented (below 200 bp). Elute in 20 µl of ddH20 (will allow vacuum use for concentrating the DNA).

Quantification:

Keep the libraries at -20°C. Use 1 µl for Qubit® Fluorimeter DNA quantification, this will be the final concentration of the library, as some DNA may be lost during the cleaning step.

Adapter preparation (mix for 2000 reactions):

Sequences :
IS1_adapter-P5: 5'-A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T-3'
IS2_adapter-P7: 5'-G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T-3'
ATDC3_adapter-P5+P7: 5'-G*A*T*C*GGAA*G*A*G*C[C3spacer]-3'

1. Assemble the following hybridization reactions in separate PCR tubes. A phosphate should be at the 3' end of ATDC3, ask specifically to the provider:

Reagent	Volume (µl)	Final concentration in 100 µl
Hybridization mix for adapter P5 (200 µM):		
IS1 adapter P5.F (500 µM)	40	200 µM
ATDC3 adapter P5+P7.R (500 µM)	40	200 µM
Oligo hybridization buffer (10X)	10	1X
H_2O	10	
Hybridization mix for adapter P7 (200 µM):		
IS2 adapter P7.F (500 µM)	40	200 µM
ATDC3 adapter P5+P7.R (500 µM)	40	200 µM
Oligo hybridization buffer (10X)	10	1X
H_2O	10	

2. Mix and incubate the reactions in a thermal cycler for 10 sec. at 95°C, followed by a ramp from 95°C to 12°C at a rate of 0.1°C/sec. Combine both reactions to obtain a ready-to-use adapter mix (100 µM each adapter).
3. To reach the 10µM required for the mix, make an aliquot with 10 µl of adapter solution and 90 µl of H_2O.

NGS P5 and P7 indexed primers sequences:

#IndexP5	Name	OligoP5 sequence (index in lowercase)
AACGTTG	NGS_P5_1	AATGATACGGCGACCACCGAGATCTACAC caacgttACACTCTTTCCCTACACGACGCTCTT
AAGAGAC	NGS_P5_2	AATGATACGGCGACCACCGAGATCTACAC gtctcttACACTCTTTCCCTACACGACGCTCTT
ACCTGAT	NGS_P5_3	AATGATACGGCGACCACCGAGATCTACAC atcaggtACACTCTTTCCCTACACGACGCTCTT
ACTCTCT	NGS_P5_4	AATGATACGGCGACCACCGAGATCTACAC agagagtACACTCTTTCCCTACACGACGCTCTT
AGATATT	NGS_P5_5	AATGATACGGCGACCACCGAGATCTACAC aatatctACACTCTTTCCCTACACGACGCTCTT
AGTCCAA	NGS_P5_6	AATGATACGGCGACCACCGAGATCTACAC ttggactACACTCTTTCCCTACACGACGCTCTT

CAACTGC	NGS_P5_7	AATGATACGGCGACCACCGAGATCTACAC gcagttgACACTCTTTCCCTACACGACGCTCTT
CAAGAAT	NGS_P5_8	AATGATACGGCGACCACCGAGATCTACAC attcttgACACTCTTTCCCTACACGACGCTCTT
CAGCGCG	NGS_P5_9	AATGATACGGCGACCACCGAGATCTACAC cgcgctgACACTCTTTCCCTACACGACGCTCTT
CCAATTA	NGS_P5_10	AATGATACGGCGACCACCGAGATCTACAC taattggACACTCTTTCCCTACACGACGCTCTT
CCGTAAC	NGS_P5_11	AATGATACGGCGACCACCGAGATCTACAC gttacggACACTCTTTCCCTACACGACGCTCTT
CGTAGAG	NGS_P5_12	AATGATACGGCGACCACCGAGATCTACAC ctctacgACACTCTTTCCCTACACGACGCTCTT
CTAACGT	NGS_P5_13	AATGATACGGCGACCACCGAGATCTACAC acgttagACACTCTTTCCCTACACGACGCTCTT
CTCGGAA	NGS_P5_14	AATGATACGGCGACCACCGAGATCTACAC ttccgagACACTCTTTCCCTACACGACGCTCTT
CTGGTCT	NGS_P5_15	AATGATACGGCGACCACCGAGATCTACAC agaccagACACTCTTTCCCTACACGACGCTCTT
GAGGAGC	NGS_P5_16	AATGATACGGCGACCACCGAGATCTACAC gctcctcACACTCTTTCCCTACACGACGCTCTT
GCAGATG	NGS_P5_17	AATGATACGGCGACCACCGAGATCTACAC catctgcACACTCTTTCCCTACACGACGCTCTT
GCATCGA	NGS_P5_18	AATGATACGGCGACCACCGAGATCTACAC tcgatgcACACTCTTTCCCTACACGACGCTCTT
GCGTTCG	NGS_P5_19	AATGATACGGCGACCACCGAGATCTACAC cgaacgcACACTCTTTCCCTACACGACGCTCTT
GGCAAGA	NGS_P5_20	AATGATACGGCGACCACCGAGATCTACAC tcttgccACACTCTTTCCCTACACGACGCTCTT
GGTACCT	NGS_P5_21	AATGATACGGCGACCACCGAGATCTACAC aggtaccACACTCTTTCCCTACACGACGCTCTT
GGTCTTG	NGS_P5_22	AATGATACGGCGACCACCGAGATCTACAC caagaccACACTCTTTCCCTACACGACGCTCTT
GTCTACT	NGS_P5_23	AATGATACGGCGACCACCGAGATCTACAC agtagacACACTCTTTCCCTACACGACGCTCTT
GTTAATC	NGS_P5_24	AATGATACGGCGACCACCGAGATCTACAC gattaacACACTCTTTCCCTACACGACGCTCTT
TATATGG	NGS_P5_25	AATGATACGGCGACCACCGAGATCTACAC ccatataACACTCTTTCCCTACACGACGCTCTT
TATGCTT	NGS_P5_26	AATGATACGGCGACCACCGAGATCTACAC aagcataACACTCTTTCCCTACACGACGCTCTT
TGATGCA	NGS_P5_27	AATGATACGGCGACCACCGAGATCTACAC tgcatcaACACTCTTTCCCTACACGACGCTCTT
TGGCCGC	NGS_P5_28	AATGATACGGCGACCACCGAGATCTACAC gcggccaACACTCTTTCCCTACACGACGCTCTT
TTATCTC	NGS_P5_29	AATGATACGGCGACCACCGAGATCTACAC gagataaACACTCTTTCCCTACACGACGCTCTT
TTCCGCC	NGS_P5_30	AATGATACGGCGACCACCGAGATCTACAC ggcggaaACACTCTTTCCCTACACGACGCTCTT

#Index P7	**Name**	**OligoP7 sequence (index in lowercase)**
AACGGTC	NGS_P7_1	CAAGCAGAAGACGGCATACGAGAT gaccgttGTGACTGGAGTTCAGACGTGT
AAGTATT	NGS_P7_2	CAAGCAGAAGACGGCATACGAGAT aatacttGTGACTGGAGTTCAGACGTGT
ACTATAT	NGS_P7_3	CAAGCAGAAGACGGCATACGAGAT atatagtGTGACTGGAGTTCAGACGTGT

AGATTCT	NGS_P7_4	CAAGCAGAAGACGGCATACGAGAT agaatctGTGACTGGAGTTCAGACGTGT
AGCAGAA	NGS_P7_5	CAAGCAGAAGACGGCATACGAGAT ttctgctGTGACTGGAGTTCAGACGTGT
AGTAACG	NGS_P7_6	CAAGCAGAAGACGGCATACGAGAT cgttactGTGACTGGAGTTCAGACGTGT
ATCTGCG	NGS_P7_7	CAAGCAGAAGACGGCATACGAGAT cgcagatGTGACTGGAGTTCAGACGTGT
ATGCGTA	NGS_P7_8	CAAGCAGAAGACGGCATACGAGAT tacgcatGTGACTGGAGTTCAGACGTGT
CAGGCAA	NGS_P7_9	CAAGCAGAAGACGGCATACGAGAT ttgcctgGTGACTGGAGTTCAGACGTGT
CCATACC	NGS_P7_10	CAAGCAGAAGACGGCATACGAGAT ggtatggGTGACTGGAGTTCAGACGTGT
CCGCGAG	NGS_P7_11	CAAGCAGAAGACGGCATACGAGAT ctcgcggGTGACTGGAGTTCAGACGTGT
CGAATGG	NGS_P7_12	CAAGCAGAAGACGGCATACGAGAT ccattcgGTGACTGGAGTTCAGACGTGT
CGGAATC	NGS_P7_13	CAAGCAGAAGACGGCATACGAGAT gattccgGTGACTGGAGTTCAGACGTGT
CGTCTAA	NGS_P7_14	CAAGCAGAAGACGGCATACGAGAT ttagacgGTGACTGGAGTTCAGACGTGT
CTCGCGC	NGS_P7_15	CAAGCAGAAGACGGCATACGAGAT gcgcgagGTGACTGGAGTTCAGACGTGT
GACGCCG	NGS_P7_16	CAAGCAGAAGACGGCATACGAGAT cggcgtcGTGACTGGAGTTCAGACGTGT
GAGCCTC	NGS_P7_17	CAAGCAGAAGACGGCATACGAGAT gaggctcGTGACTGGAGTTCAGACGTGT
GCTTAGT	NGS_P7_18	CAAGCAGAAGACGGCATACGAGAT actaagcGTGACTGGAGTTCAGACGTGT
GGAGATT	NGS_P7_19	CAAGCAGAAGACGGCATACGAGAT aatctccGTGACTGGAGTTCAGACGTGT
GGCTTGA	NGS_P7_20	CAAGCAGAAGACGGCATACGAGAT tcaagccGTGACTGGAGTTCAGACGTGT
GTACGGC	NGS_P7_21	CAAGCAGAAGACGGCATACGAGAT gccgtacGTGACTGGAGTTCAGACGTGT
TAATCGC	NGS_P7_22	CAAGCAGAAGACGGCATACGAGAT gcgattaGTGACTGGAGTTCAGACGTGT
TACCTAC	NGS_P7_23	CAAGCAGAAGACGGCATACGAGAT gtaggtaGTGACTGGAGTTCAGACGTGT
TATATTG	NGS_P7_24	CAAGCAGAAGACGGCATACGAGAT caatataGTGACTGGAGTTCAGACGTGT
TCCAGCC	NGS_P7_25	CAAGCAGAAGACGGCATACGAGAT ggctggaGTGACTGGAGTTCAGACGTGT
TCTCATA	NGS_P7_26	CAAGCAGAAGACGGCATACGAGAT tatgagaGTGACTGGAGTTCAGACGTGT
TCTTCCG	NGS_P7_27	CAAGCAGAAGACGGCATACGAGAT cggaagaGTGACTGGAGTTCAGACGTGT
TTCGTCT	NGS_P7_28	CAAGCAGAAGACGGCATACGAGAT agacgaaGTGACTGGAGTTCAGACGTGT
TTGCAAT	NGS_P7_29	CAAGCAGAAGACGGCATACGAGAT attgcaaGTGACTGGAGTTCAGACGTGT
TTGGCTG	NGS_P7_30	CAAGCAGAAGACGGCATACGAGAT cagccaaGTGACTGGAGTTCAGACGTGT

Appendix III

Table 4. List of all samples including the associated morphospecies, the collector, the Herbarium, the year, the sampling country, die QR code, the applied protocol and the lane.

Labcode	Species name	Morphospecies	Collector N°	Herbarium	Collection year	Origin	Barcode	Protocol	Lane
T01	D. perrieri	Variation 1	SF 8974	P	1954	Madagascar	P04569175; P04569173; P04569174	old protocol	TA/TB --> DON2
T02	D. analalavensis		SF 19075	P	1958	Madagascar	P04596206	old protocol	TA/TB --> DON2
T03	D. humbertii		SF 12538	P		Madagascar	P04596207; P04596208; P04596209	old protocol	TA/TB --> DON2
T04	D. perrieri	Variation 1	SF 17819	P		Madagascar	P05193480; P04596204	old protocol	TA/TB --> DON2
T05	D. lanceolata		SF 23201bis	P	1964	Madagascar	P04596201	old protocol	TA/TB --> DON2
T06	D. lanceolata/ D. perrieri		Rakotonirina 637	P	2014	Madagascar	P01030481	old protocol	TA/TB --> DON2
T07	D. sp.		Randrianjanaka 184	P	1994	Madagascar	P04592682	old protocol	TA/TB --> DON2
T08	D. lanceolata		SF 55-B-R-230	P	1951	Madagascar	P04604601	old protocol	DON3
T09	D. perrieri	Variation 1	Razakamalala 2481	P	2005	Madagascar	P04568931	old protocol	DON3
T10	D. perrieri	Variation 1	Rahajasoa 346	P	1994	Madagascar	P04592686	old protocol	DON3
T11	D. perrieri	Typical	Faber-Langendoen 3225	P	1990	Madagascar	P04596202	old protocol	DON3
T12	D. fenervensis (cf.)		Rakotonirina 262	P	2013	Madagascar	P00870671	old protocol	DON3
T13	D. humbertii		Perrier 8783	P	1905	Madagascar	P00752279; P00752278	single tube	TA/TB --> DON2
T14	D. capuronii (cf.)		SF 27738bis	P	1967	Madagascar	P04592684	single tube	TA/TB --> DON2
T15	D. analalavensis		SF 18935	P	1958	Madagascar	P04596194; P04596195	single tube	TA/TB --> DON2
T16	D. analalavensis (cf.)		SF 24500	P	1966	Madagascar	P05193472; P04596190; P04596191; P04596193	single tube	
T17	D. ambrensis		SF 14877	P	1955	Madagascar	P04596200; P0596199	single tube	TA/TB --> DON2
T18	D. ambrensis		SF 11277	P	1954	Madagascar	P00109341; P00109342; P00109343; P00109344	single tube	TA/TB --> DON2
T19	D. capuronii		Schatz 2555	P	1989	Madagascar	P00417607	single tube	TA/TB --> DON2
T20	D. fenervensis		SF 28809bis	P	1969	Madagascar	P04609632	single tube	TA/TB --> DON2
T21	D. fenervensis		SF 8568	P	1953	Madagascar	P04609639; P04609637; P04609638; P0609634	single tube	TA/TB --> DON2
T22	D. humbertii		Letsara 892	P	2009	Madagascar	P00722324	single tube	TA/TB --> DON2
T23	D. perrieri	Typical	McPherson 14623	P	1989	Madagascar	P04604606	single tube	TA/TB --> DON2
T24	D. perrieri	Variation 0	Randrianasolo 73	P	1990	Madagascar	P04604618	single tube	TA/TB --> DON2
T25	D. delphinensis		Andriamiarinoro 274	P	2012	Madagascar	P01065533	single tube	TA/TB --> DON2
T26	D. perrieri	Variation 6	Ravelonarivo 1196	P	2000	Madagascar	P00859546; P0140673	single tube	TA/TB --> DON2
T27	D. analalavensis		SF 11345	P	1954	Madagascar	P04596189; P04596188; P04596186; P04596187	single tube	TA/TB --> DON2
T28	D. guereliana		SF 18959	P	1958	Madagascar	P04919468; P04919470; P04919469	single tube	TA/TB --> DON2
T29	D. lanceolata		SF 946	P		Madagascar	P04609620; P04609617; P04609619	old protocol	DON3
T30	D. masoalensis		SF 23215bis	P	1964	Madagascar	P04609614	single tube	TA/TB --> DON2
T31	D. fenervensis		SF 6993	P	1953	Madagascar	P04609631; P04609630	single tube	TA/TB --> DON2
T32	D. lanceolata		McPherson 14977	P	1990	Madagascar	P04609618	single tube	TA/TB --> DON2
T33	D. lanceolata		SF 488-R-56	P		Madagascar	P04604590; P04604594	single tube	TA/TB --> DON2
T34	D. humbertii		SF 12538	P		Madagascar	P04596207; P04596208; P04596209	single tube	TA/TB --> DON2
T35	D. ambrensis		Bernardi 12008	G	1967	Madagascar	G00074569	single tube	TA/TB --> DON2
T36	D. ambrensis		Leeuwenberg 14314	G	1994	Madagascar	G00074568	single tube	TA/TB --> DON2
T37	D. capuronii		Gautier 5549	G	2010	Madagascar	-	old protocol	DON3
T38	D. capuronii		Randrianaivo 128	G	1997	Madagascar	-	old protocol	DON3
T39	D. delphinensis		Ramison 314	G	2007	Madagascar	-	old protocol	DON3
T40	D. delphinensis		Rabenantoandro 1340	G	2003	Madagascar	-	both	DON1 + DON3
T41	D. delphinensis		Razakamalala 1329	G	2004	Madagascar	G00160396	both	DON1 + DON3
T42	D. fenervensis		Randrianaivo 584	G	2002	Madagascar	-	both	DON1 + DON3
T43	D. fenervensis (aff.)		Rabenantoandro 918	G	2002	Madagascar	-	both	DON1 + DON3
T44	D. fenervensis (aff.)		Martial 237	G	2013	Madagascar	-	both	DON1 + DON3
T45	D. perrieri	Variation 5	Rahajasoa 299	G	1994	Madagascar	-	both	DON1 + DON3
T46	D. perrieri	Variation 2	Birkinshaw 158	G	1992	Madagascar	G00075319	both	DON1 + DON3
T47	D. perrieri	Variation 2	Gautier 3255	G	1997	Madagascar	G00075320	old protocol	DON3
T48	D. perrieri	Variation 5	Gautier 4134	G	2001	Madagascar	G00007511	old protocol	DON3
T49	D. perrieri	Variation 5	Vasey 75	G	1998	Madagascar	-	old protocol	DON3
T50	D. perrieri	Variation 1	SF 28761	G	1969	Madagascar	-	old protocol	DON3
T51	D. perrieri	Variation 5	Ranirison 842	G	2004	Madagascar	G00019598	old protocol	DON3
T52	D. perrieri	Variation 9	Birkinshaw 212	G	1992	Madagascar	G00075316	old protocol	DON3
T53	D. perrieri	Variation 4	Syde 357	G	2017	Madagascar	-	old protocol	DON3
T54	D. perrieri	Variation 2	Antilahimena 73	G	1994	Madagascar	G00075317	old protocol	DON3
T55	D. perrieri	Variation 6	Ratovoson 225	G	2000	Madagascar	G00075315	old protocol	DON3
T56	D. perrieri	Variation 0	Gautier 5082	G	2006	Madagascar	G00170388	old protocol	DON3
T57	D. perrieri	Variation 2	Antilahimena 124	G	1994	Madagascar	G00075313	old protocol	DON3
T58	D. perrieri	Variation 0	Randriamampionona 457	G	1993	Madagascar	-	old protocol	DON3
T59	D. ranirisonii		Gautier 5387	G	2010	Madagascar	G00304192	old protocol	DON3
T60	D. perrieri	Variation 2	Wohlhauser 60019	G	1998	Madagascar	G00075321	old protocol	DON3
T61	D. perrieri	Variation 2	Rabarimanarivo 134	G	2005	Madagascar	-	old protocol	DON3
T62	D. perrieri	-	LG 6521	G		Madagascar	-	old protocol	DON3
T63	D. lanceolata		Bartish & Fond 34	P		N Australia	P0612350	old protocol	DON3
T64	D. lanceolata		Fernandy 332	P		India	P4550024	old protocol	DON3
T65	D. lanceolata		Lav 1480	P		Hainan	P4550051	old protocol	DON3
T66	D. lanceolata		Kostermans 25586	P		Sri Lanka	P4550020	old protocol	DON3
T67	D. lanceolata		Petelot 1542	P		Tonkin	P4550013	old protocol	DON3
T68	D. lanceolata		Kostermans 257	P		New Guinea	P04592290	old protocol	DON3
T69	D. perrieri	Variation 7	Rakotozafy 115	G	2013	Madagascar	-	old protocol	DON3
T70	D. perrieri x masoalesis	Variation 4	Rakotomalaza 2075	G	1999	Madagascar	G00160414	old protocol	DON3
T71	D. perrieri x masoalesis	Variation 4	Rasoavimbahoaka 275	G	1994	Madagascar	-	old protocol	DON3
T72	D. masoalensis		SF 8833	G	1982	Madagascar	G00014786	old protocol	DON3
T73	D. pruniformis		McPherson 15836	G	1992	Gabon	G00160419	old protocol	DON3
T74	D. bangweolensis		Harder 3544	G	1996	Zambia	G00160389	old protocol	DON3
T75	D. ogoouensis		Breteler 11399	G	1992	Gabon	-	old protocol	DON3
T76	D. ubangiensis		Cremers 615E	G	1967	Cote Divoire	G00160415	old protocol	DON3
T77	D. viridifolia		Wells 81	G	1961	Natal	G00160422	old protocol	DON3
T78	D. welwitschii		Carvalho 3406	G	1988	Guinea	G00160423	old protocol	DON3
T79	D. guereliana		Ratovoson 1300	G	2007	Madagascar	-	old protocol	DON3
T80	D. guereliana		Rabehevitra 960	G	2004	Madagascar	G0075312	old protocol	DON3
T81	D. perrieri	Typical	Poncy 1530	G	2001	Madagascar	G00160418	old protocol	DON3
T82	D. perrieri	Typical	Miller 3286	G	1988	Madagascar	-	old protocol	DON3
T83	D. perrieri	Typical	Nusbaumer 2834	G	2008	Madagascar	G00181663	old protocol	DON3
T84	D. perrieri	Typical	Gautier 5503	G	2010	Madagascar	-	old protocol	DON3
T85	D. perrieri	Typical	SF 27765	G	1967	Madagascar	-	old protocol	DON3
T86	D. perrieri	Typical	Randriamampionona 626	G	1993	Madagascar	-	old protocol	DON3
T87	D. perrieri	Typical	Rakotovao 3523	G	2006	Madagascar	-	old protocol	DON3
T88	D. analalavensis		Perrier 14834	G	1922	Madagascar	-	old protocol	DON3
T89	D. analalavensis		SF 24211	G	1956	Madagascar	-	old protocol	DON3
T90	D. analalavensis		Ramananjanahary 110	G	2013	Madagascar	-	old protocol	DON3
T91	D. delphinensis		Ramison 157	G	2006	Madagascar	-	old protocol	DON3
T92	D. delphinensis		Rabehevitra 957	G	2004	Madagascar	-	old protocol	DON3
T93	D. masoalensis		LG 3949	G	2001	Madagascar	-	old protocol	TA/TB --> DON2
T94	D. welwitschii		J. Miège	G	1961	Cote Divoire	-	old protocol	TA/TB --> DON2
T95	D. ogoouensis		J.J.F.E. de Wilde 11968	G	1998	Gabon	-	old protocol	TA/TB --> DON2
T96	D. ubangiensis		Germain 4527	G	1948	Congo	-	old protocol	TA/TB --> DON2
T97	D. pruniformis		Koning 5887	G	1975	Cote Divoire	-	old protocol	TA/TB --> DON2
T98	D. bangweolensis		Reekmans 7475	G	1979	Burundi	-	old protocol	TA/TB --> DON2
T99	D. bangweolensis		Kuchar 24771	G	2001	Tanzania	-	old protocol	TA/TB --> DON2
T100	see T51							old protocol	DON3
T101	see T52							old protocol	DON3

Appendix IV

Figure 19. Pseudocoalescent phylogenetic tree from ASTRAL inferred from 324 RAxML gene trees corresponding to cluster 1 from MDS in 3.4. The tree is rooted on D. guereliana. The gene sequences were retrieved with Hybpiper. All genes with more than 40 % missing data were discarded and only specimen with less than 80 % missing data over all genes are displayed in the tree. The values on the branches represent the ASTRAL posterior probabilities. Species names are followed by collector and collector number and end with the lab code.

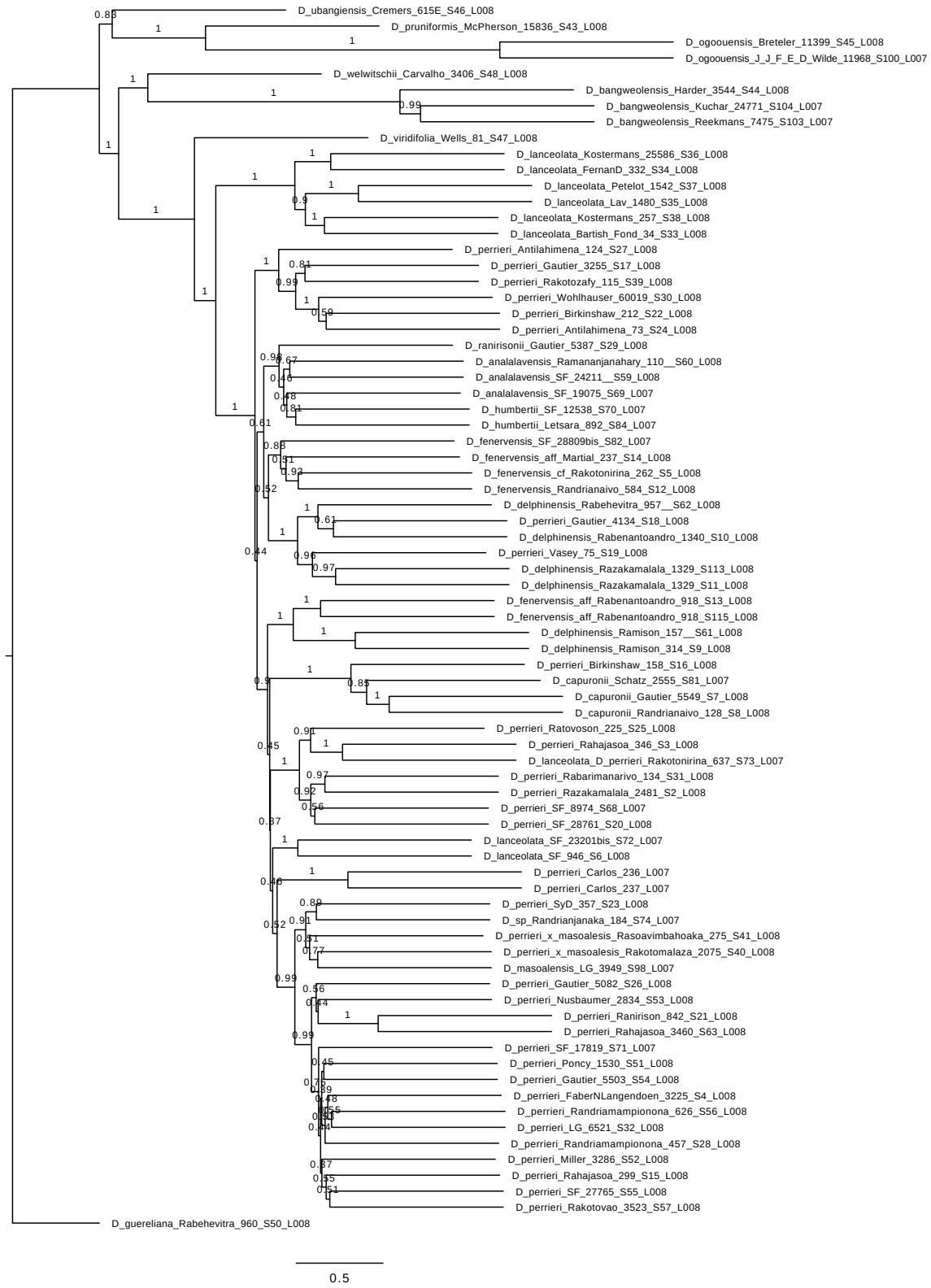

Figure 20. Pseudocoalescent phylogenetic tree from ASTRAL inferred from 456 RAxML gene trees corresponding to cluster 2 from MDS in 3.4. The tree is rooted on D. guereliana. The gene sequences were retrieved with Hybpiper. All genes with more than 40 % missing data were discarded and only specimen with less than 80 % missing data over all genes are displayed in the tree. The values on the branches represent the ASTRAL posterior probabilities. Species names are followed by collector and collector number and end with the lab code.

Appendix V

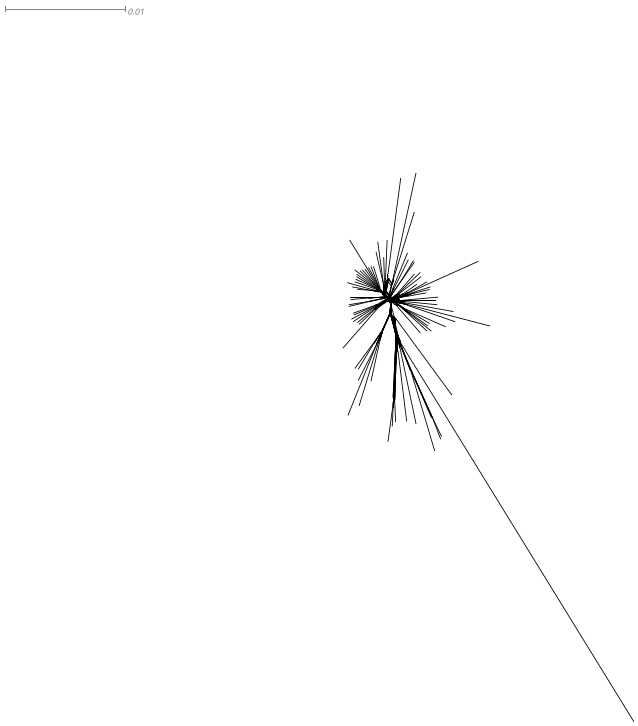

Figure 21.Phylogenetic network using the concatenated alignments of 787 genes. Sequences contained less than 20 % missing data comprising 74 samples. A Neighbor-Net with uncorrected P-distances was computed.

Appendix VI

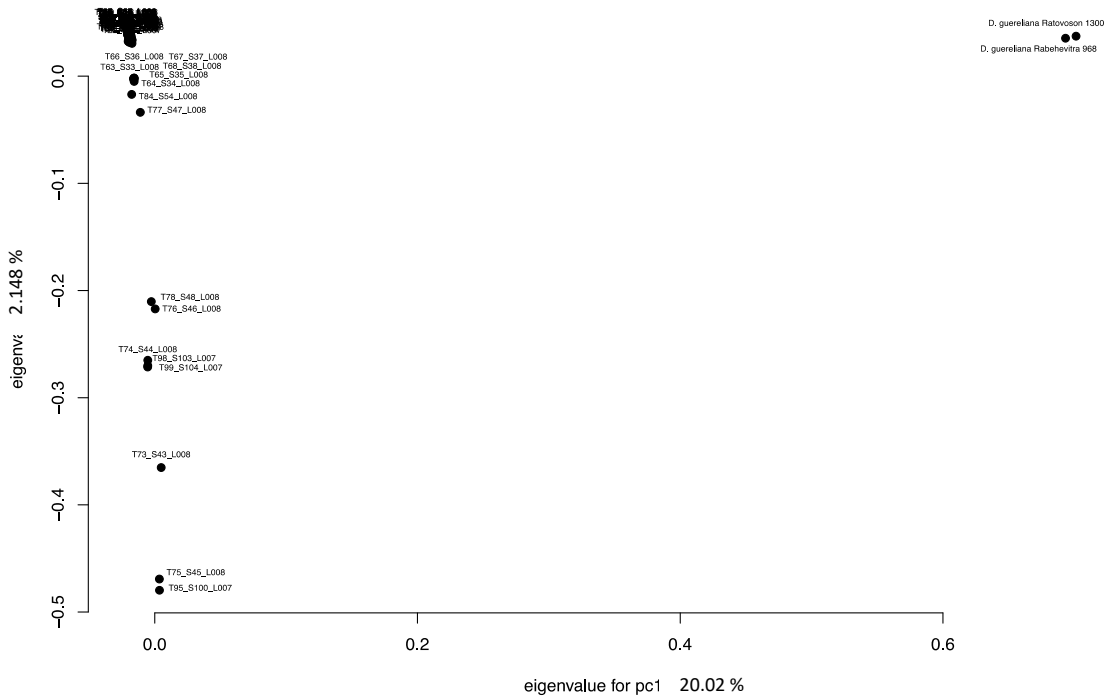

Figure 23. Principal Components Analysis (PCA) on 818.623 extracted SNPs. Samples containing less than 20 % missing data. Lab codes are given for the African Donella and the Indo-Pacific D. lanceolata. Far apart on the right corner the two D. guereliana samples are displayed.

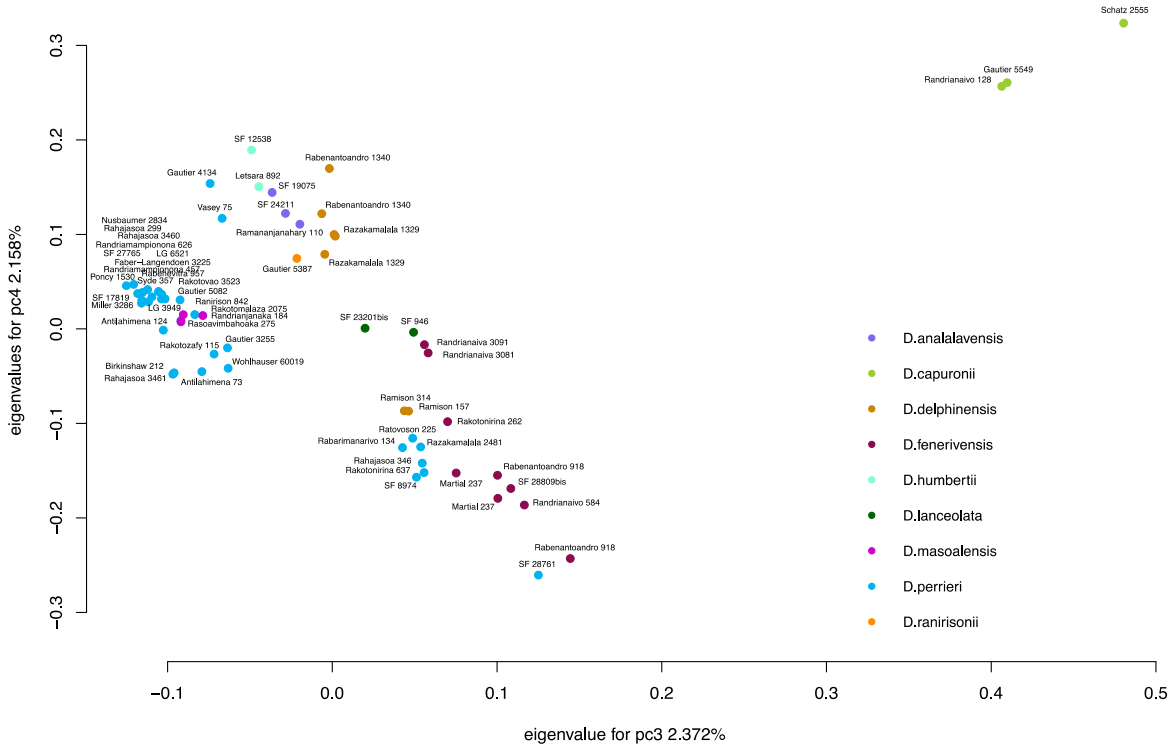

Figure 22. Principal Components Analysis (PCA) on 818.623 extracted SNPs. Samples containing less than 20 % missing data. Samples are colored by species and labeled with the collector's name and number.