5-19-2017

Systematics, Biogeography, and Species Delimitation of the Malagasy Psorospermum (Hypericaceae)

Heritiana S. Ranarivelo
University of Missouri-St.Louis, hsrq98@mail.umsl.edu

Systematics, Biogeography, and Species Delimitation of the Malagasy *Psorospermum* (Hypericaceae)

Heritiana S. Ranarivelo
MS, Biology, San Francisco State University, 2010

A Dissertation Submitted to The Graduate School at the University of Missouri-St. Louis
in partial fulfillment of the requirements for the degree
Doctor of Philosophy in Biology with an emphasis in Ecology, Evolution, and Systematics

August
2017

Advisory Committee

Peter F. Stevens, Ph.D.
Chairperson

Peter C. Hoch, Ph.D.

Elizabeth A. Kellogg, PhD

Brad R. Ruhfel, PhD

ABSTRACT

Psorospermum belongs to the tribe Vismieae (Hypericaceae). Morphologically, *Psorospermum* is very similar to *Harungana*, which also belongs to Vismieae along with another genus, *Vismia*. Interestingly, *Harungana* occurs in both Madagascar and mainland Africa, as does *Psorospermum*; *Vismia* occurs in both Africa and the New World. However, the phylogeny of the tribe and the relationship between the three genera are uncertain. Using freshly collected specimens from my fieldwork as well as extant herbarium specimens, I aimed first, to generate a phylogeny of *Psorospermum*; second, to investigate its biogeography; and third, to investigate species boundaries within Malagasy *Psorospermum*.

The results of my molecular phylogenetic work, based on a combined analysis of chloroplast DNA (*ndhF; trnL-trnF*) and nuclear DNA (ITS), strongly support *Harungana* and *Psorospermum* as two genera: *Harungan*a also includes one African species of *Vismia*, *V. rubescens*, and *Psorospermum* includes the other African *Vismia* and *Psorospermum*. My results also show that Malagasy *Psorospermum* are paraphyletic, some African species being nested within the clade of Malagasy *Psorospermum* suggesting dispersal of the genus westwards back to Africa. I conducted ancestral area reconstruction studies to test this hypothesis. *Psorospermum* may have reached Madagascar by a single dispersal event from Africa during the Oligocene (ca. 34-22 Ma), followed by diversification on Madagascar after ca. 20-19 Ma. However, two recent dispersal events appear to have occurred out of Madagascar back to Africa in the late Miocene (ca. 5.5 and 5.7 Ma).

Malagasy *Psorospermum* has not had a taxonomic revision in 60 years and the total number of species is uncertain. I undertook a novel approach to investigate the species boundaries in *Psorospermum* by integrating species hypotheses delimited by both molecular and morphometric analyses. Herbarium specimens of the putative taxa in each well supported clade in the molecular phylogeny of Malagasy *Psorospermum* were subjected to morphometric analyses using gaussian mixture models. I identified 27 species of *Psorospermum* of which 11 are new. The results will be used in a taxonomic revision of Malagasy *Psorospermum* that is in progress.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

My gratitude goes to my thesis advisor, Dr. Peter F. Stevens, for his expertise, guidance, trust and support during my time at the University of Missouri-St. Louis.

My thanks and appreciation go to my thesis committee members, Dr. Peter C. Hoch, Dr. Elizabeth A. Kellogg, and Dr. Brad R. Ruhfel. They each provided invaluable help with and advice on this study.

I also thank Dr. Ivan Jiménez for his invaluable assistance during this project.

I thank the Whitney R. Harris World Ecology Center at the University of Missouri-St. Louis for providing me with funding and other support in the completion of this project.

I thank Dr. Frank Almeda, Dr. Fernanda Cabral, Dr. Rosa Gentry-Ortiz, Dr. C. Hong-Wa, Rafaela Jorge, Dr. Laura Logamarsino, Dr. Robert Marquis, Dr. Patricia Parker, John Prusky, Cindee Retkee, Dr. Oyomoare Osazuwa-Peters, Rebecca Peters, Dr. Debra Trock, Dr. Simon Uribe-Convers, and Dylis Vela for their invaluable assistance at various times throughout this project.

My thanks go to Rokiman Letsara and Jacky Andriatiana for their help during my field work.

I thank the Missouri Botanical Garden and the Department of Biology of the University of Missouri-St. Louis for their logistic support.

I would not be performing this work without the love and support of my siblings and my parents.

My special thanks to my daughter Handy and my wife Laura for their love, support, and patience.

.

Systematics and Biogeography of the African-Malagasy *Psorospermum* (Hypericaceae) with emphasis on the Malagasy species

Heritiana Ranarivelo[a,b]

[a] *Department of Biology, University of Missouri – St. Louis, One University Blvd., St. Louis, MO 63121-4000, USA*
[b] *Missouri Botanical Garden, P.O. Box 299, St. Louis, MO 63166-0299, USA*

ARTICLE INFO

Keywords:
Psorospermum
Harungana
Phylogeny
Dispersal
Madagascar

ABSTRACT

I investigated the phylogeny and biogeography of Malagasy *Psorospermum*, a poorly-known genus in the tribe Vismieae (Hypericaceae). *Psorospermum* occurs in both mainland Africa and Madagascar. There are 26 species described in Madagascar, plus at least four additional undescribed species. The genus has not been revised since 1951; all species are endemic and its circumscription remains ambiguous. The last classification of Hypericaceae placed *Psorospermum* as synonymous with *Harungana* within the tribe Vismieae, which includes the genus *Vismia*. The latest molecular phylogenetic analysis of the whole clusioid clade confirmed the placement of *Psorospermum* within Hypericaceae and within a monophyletic clade consisting of *Harungana, Vismia* and *Psorospermum*. However, this study included only four species of *Psorospermum.* Here I present the first phylogeny of *Psorospermum* based on a combined analysis of chloroplast (*ndhF; trnL-trnF*) and nuclear DNA (*ITS*). My results strongly support *Harungana* and *Psorospermum* as distinct genera: *Harungana* includes one African species of *Vismia*, *V. rubescens*, and *Psorospermum* includes the other African *Vismia* and *Psorospermum*. Malagasy *Psorospermum* is paraphyletic, with some African species nested within the clade of Malagasy *Psorospermum* suggesting dispersal of the genus westwards back to Africa. African origins predominate within the general biogeographical relationships of Malagasy biota. With additional sampling I generated a time-calibrated phylogeny, and conducted ancestral area reconstruction to test my hypotheses of dispersal out of Madagascar, back to Africa. *Psorospermum* seems to have originated on Mainland Africa ca. 34 Mya, arrived in Madagascar ca. 21 Mya, with two recent dispersals back to Africa in the late Miocene ca. 5.5 and 5.7 Mya.

1. Introduction

Madagascar is a biodiversity hotspot (Myers et al., 2000; Bellard et al., 2014; Sloan et al., 2014). The number of Malagasy plant species is estimated at 11,200

(Callmander et al., 2011); the angiosperms alone make up 10,650 species, 84% of which are endemic (Buerki et al., 2013). Understanding the composition, origin, and evolution of such an unique flora is a major goal in systematic and biogeographic research. To be able to generalize about this, we need information about relationships of many groups of organisms. One group that needs particular attention is the genus *Psorospermum*. It grows in Madagascar and Africa, but relationships within the genus have never been investigated, and such information may be also helpful in other fields of biology such as taxonomy, phytochemistry, pharmacognosy, and conservation biology. Some species of *Psorospermum* are used in traditional medicine. For example, *P. febrifugum* Spach is used as an anti-inflammatory or febrifuge in tropical Africa, as well as a treatment for leprosy, a purgative, and a poison antidote (Epifano, 2013). Similarly, in Madagascar *P. androsaemifolium* Baker is traditionally used to treat spider and scorpion bites, as well as stomach sickness (Poumale et al. 2008, 2011; Epifano, 2013). However, *Psorospermum* has never been investigated in an evolutionary context, especially at the species level, and little is known about its morphological evolution. The study of the clusioid clade by Ruhfel et al. (2011) did include *Psorospermum*, but focused on general relationships.

The last revision of *Psorospermum* separated the 26 Malagasy *Psorospermum* from the African ones (Perrier de la Bâthie, 1951); the latter were placed into five species (Bamps, 1966). All Malagasy *Psorospermum* are endemic, but their monophyly has never been tested. Finally, little is known about the relationships between *Psorospermum s.l.* and the other genera in the tribe Vismieae, such as *Harungana* and *Vismia*. The latest classification of Hypericaceae is that by Stevens (2007), which was based on morphology and included *Psorospermum s.l.* in *Harungana*. However, the well-supported molecular phylogeny of the clusioids (Ruhfel et al., 2011) does not support this classification (Fig. 1).

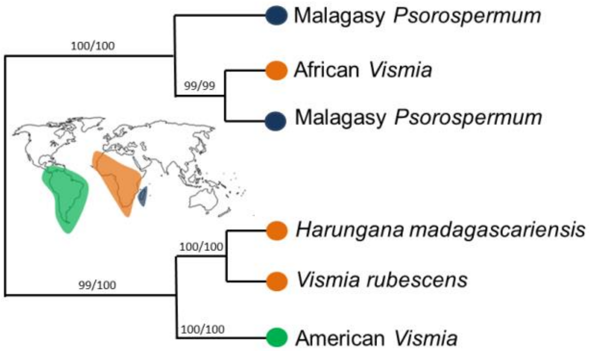


Fig. 1. Previous hypothesized relationships in Vismieae (Ruhfel et al. 2011). Values above the branches are maximum likelihood bootstrap values (left) and Bayesian posterior probabilities converted to percentages (right).

Harungana sensu Stevens (2007) is paraphyletic (Ruhfel et al., 2011, 2013). *Harungana* s.s. and the African *Vismia rubescens* together are sister to American *Vismia*, while other African *Vismia* and all *Psorospermum s.l.* form another clade (Fig. 1). Although the phylogeny in Ruhfel et al. (2011) was well supported, only four species of *Psorospermum s.l.* were included. In addition, morphological characters within this group need to be investigated to better assess its classification, to elucidate patterns of character evolution, and to identify synapomorphies for the major clades.

The biogeography of the Malagasy *Psorospermum* is also intriguing and has not been explored. In the bigger picture, African origins of Malagasy biota predominate (Yoder and Nowak, 2006; Buerki, 2013). However, although Vismieae may have originated in Africa (Ruhfel et al., 2016), *Psorospermum* and *Harungana* are found in both mainland Africa and Madagascar, while *Vismia* grows in the New World and Africa. Geological studies of Madagascar estimate that Madagascar separated from Africa ca. 180 Ma (Yoder and Novak, 2006). The current flora of Madagascar may be descendants of a mix of species some of which were living on pre-break-up Madagascar, although 180 Ma is before the origin of crown-group angiosperms according to some estimates (Stevens, 2001 onwards). Most species likely arrived by long distance dispersal (LDD) from other landmasses. Thus dated molecular studies argue that all the clades that make up the Malagasy flora are far too young to result from vicariance, making LDD the most probable explanation for their presence in Madagascar (e.g. Bacon et al., 2015; Federman et al. 2015, 2016).

Today, Madagascar is separated from Africa by only ca. 500 km. The shortest distance is from the coast of Mozambique, about 430 km away (Stankiewicz, 2006), and this distance has not changed much in the past 90 Ma as Madagascar has gradually moved north to its current position starting around 140 Ma (Rabinowitz et al. 1983; Coffin & Rabinowitz 1988; Torsvik et al. 1998; Reeves & de Wit 2000; Stankiewicz 2006). Therefore additional processes like oceanic currents may favor the dispersal of species from Africa to Madagascar (Ali and Huber, 2010). Paleooceanographic simulations predict that eastward surface currents in the early Cenozoic about 65 Ma would have facilitated the dispersal of mammals from Africa to Madagascar, confirming the traditional “sweepstake” theory (Simpson, 1940; Ali and Huber, 2010). However, after the mid-Miocene the probability of successful west-to-east transoceanic dispersal for obligate rafters decreases significantly, because the present-day westward flow, which started ca. 20 Ma, makes it nearly impossible for species to reach the island in this way (Samonds et al., 2012; Federman et al., 2015). Indeed, from ca. 20 Ma onward, some groups of plants from Madagascar could more easily have dispersed in the opposite direction: east-to-west from the island to Africa.

For example, *Exacum oldenlandioides* (S. Moore) Klack. (Gentianaceae) may have dispersed from Madagascar to mainland Africa ca. 4.7 Ma (Yuan et al., 2005). In Apocynaceae, ten species are shared between Africa and Madagascar. Dispersal events of the genus *Cynanchum* (Apocynaceae) from Madagascar to Africa have been reported, yet some species have been dispersed from Africa to Madagascar (Meve and Liede, 2002). It has been inferred that species of Celastraceae, e.g. *Brexia madagascariensis* (Lam.) Ker Gawl., *Elaeodendron* and *Pleurostylia*, that grow in both Africa and Madagascar also dispersed out of Madagascar between ca. 20-10 Ma (Bacon et al., 2015). The recent study of Ruhfel et al. (2016), suggests dispersals of Vismieae from Africa to Madagascar during the Cenozoic, but no dispersal out of Madagascar was detected.

The goals and hypotheses of this study are as follows: first, I aim to provide the first comprehensive phylogeny for Malagasy *Psorospermum*. I test whether Malagasy *Psorospermum* form a clade, if *Psorospermum s.l.* forms a clade with most other African *Vismia*, and if this African clade is sister to the clade ((*Harungana madagascariensis* + *Vismia rubescens*) + American *Vismia*). Second, I aim to assess morphological variation within the group and identify synapomorphies for the major clades. The molecular phylogeny is being integrated with morphometric analyses for the species delimitation of the Malagasy *Psorospermum* (Ranarivelo et al. in preparation). Third, with a dated phylogeny, I aim to determine the divergence times of the Malagasy and African lineages, to see if the dispersal of *Psorospermum* to Madagascar consisted of a single or multiple events, and if there were any dispersals out of Madagascar back to mainland Africa.

2. Materials and methods

2.1. Taxon sampling

The combined dataset consists of 53 taxa: 45 in the ingroup and 8 in the outgroup. Of the 45 ingroup taxa, 29 are Malagasy, 7 are American, and 9 are African. Altogether 122 samples were included, with multiple samples for each taxon, most representing Malagasy *Psorospermum*. One hundred and two of the 122 samples are new. Voucher specimens, herbarium, GeneBank numbers of sequences are listed in Appendix A.

Malagasy samples for DNA extraction were collected during fieldwork in different areas of Madagascar and include species growing in three of the five main bioclimatic regions of the island and to which *Psorospermum* is restricted (Supplementary Fig. S1). Samples were collected in diverse habitats, including coastal littoral forest vegetation on sandy soil, dry areas with karstic soil, high elevation vegetation in the highlands, and mid- and low elevation rainforests of the eastern areas of Madagascar. According to Perrier, one species, *P. cerasifolium,* occurs in the sub-arid area, but *Psorospermum* was not found when I visited the specific locality he mentioned (Supplementary Fig. S1). *Psorospermum cerasifolium* was also recorded by Perrier from in the dry area, and was collected there.

The African species of *Psorospermum* and three of the six species of African *Vismia* are included in this analysis (*V. rubescens, V. guineensis*, and *V. sp*.). The seven species of American *Vismia* were represented by sequences from GenBank, and seven species of *Hypericum* and *Eliea articulata* were used as outgroups. Voucher details and GenBank accession numbers for the sequences are provided in Appendix A.

2.2. Gene selection

The chloroplast DNA markers *trnL-trnF*, *psbA-trnH, trnS-trnG, ndhF*, the nuclear DNA internal transcribed spacer of the ribosomal RNA genes (ITS) and the low copy gene Embryo-defective 2765 (*EMB2765*) were initially tested for variation. *ndhF, trnL-trnF* and ITS were chosen for the study since they are the most informative as summarized in Supplementary Table S1. Primers used to amplify *psbA-trnH, trnS-trnG,* and *EMB2765* are listed in Table S2. GenBank accession numbers for the sequences are provided in Appendix A. The use of ITS appeared to be successful in previous studies of Hypericaceae (Meseguer et al. 2013; Nürk et al. 2012), however, care was taken in using ITS because it is the subject of several critiques in the literature, focusing on high polymorphism, pseudogenes, and concerted evolution (Baldwin et al,. 1995; Soltis et al., 1998; Sang, 2002; Razafimandimbison et al., 2004; Alvarez and Wendel, 2013; Zimmer and Wen, 2013). A major problem with ITS is that it is a part of the rDNA cistron which has many copies. When amplifying the ITS region, amplicons can come from all of these genomic copies. Thus cloning was conducted as a precaution against a misleading interpretation of the results (see laboratory protocols). Additionally, rapid diversification, especially at the species level, is known to have occurred in the Malagasy flora (e.g. Buerki, 2013; Hong-Wa, 2013). In such cases, concerted evolution of ITS may cause problems for phylogenetic reconstruction as discussed by Sanderson and Doyle (1992) and Baldwin et al. (1995) and the rate of speciation might exceed the rate of concerted evolution; species might end up with ITS alleles that do not accurately reflect the species tree.

2.3. Laboratory protocols

DNA from leaf tissue was extracted using the cetyl trimethylammonium bromide (CTAB) method modified from Doyle and Doyle (1990). Unlike in *Hypericu*m, the DNEasy mini Kit (Qiagen, Valencia, CA) did not work with my *Psorospermum* samples; at best it yielded a very low DNA concentration that could not be amplified. One of the main causes of this problem could be the high concentration of secondary compounds in most of the species of *Psorospermum*, a problem encountered in the extraction of other clusioids such as *Garcinia, Kielmeyera* and *Caraipa* (B. Ruhfel personal communication). Extracted DNA was amplified using the polymerase chain reaction (PCR) with the following reaction mixture: 5 µl of 5X Qiagen Taq Buffer, 2.5 µl of 0.4mM MgCl2, 1 µl of 0.2mM dNTPs, 1 µl of 0.4 µM forward primer, 1 µl of 0.4 µM reverse primer, 0.125 µl of 0.4 µM Qiagen Taq DNA polymerase, plus ddH_2O to achieve a final volume of 25 µl. Primers are those used in previous studies of Hypericaceae (Nürk, 2010; Ruhfel, et al. 2011; Meseguer et al., 2013) (Table 1). Amplification of *ndh*F is difficult, and for samples that could not be amplified with primers 972F and 2110R, newly designed primers 209F and 1159R were used (Table 1).

The following thermocycler protocol was used for *ndh*F PCR: 5 min at 94° C, followed by 30 cycles of 1:30 min at 94° C, 2 min at 43° C and 4 min at 72° C, and a final elongation of 10 min at 72° C. For *trn*L-*trn*F, the following thermocycler protocol was used: 5 min at 95° C, followed by 35 cycles of 30 sec at 94° C, 30 sec at 55° C and 1 min at 72° C, and a final elongation of 5 min at 72° C. For ITS: 3 min at 95° C, followed by 35 cycles of 30 sec at 95° C, 45 sec at 53° C, and 1 min at 68° C, and a final elongation of 10 min at 70° C.

PCR products were visualized on 1.5% agarose gels; DNA bands were cut from the gel and cleaned with QIAquick PCR purification Kit (Qiagen, Valencia, CA) following the manufacturer's instructions. For ITS cloning, the PCR products were purified via gel extraction. Purified products were cloned using pGEM-T Easy Vector, and transformed into JM109 High-Efficiency Competent Cells (Promega, Madison, Wisconsin, USA). Transformed cells were plated and selected via blue-white screen on LB agar with X-Gal, isopropyl-beta-thio-galactoside (IPTG), and ampicillin. At least 3 positive clones of each PCR product were selected. Universal primers T7 and M13R were used for sequencing. Sanger sequencing was performed with an ABI 3100 Capillary Electrophoresis Genetic Analyzer with ABI BigDye Terminator v3.1 Cycle Sequencing chemistry (Applied Biosystems Inc., Foster City, CA) at the University of Missouri - St. Louis or sent to Functional BioSciences Inc. (http://functionalbio.com).

Table 1
List of the primers used in the molecular phylogenetic of *Psorospermum*

Gene region	Primer Name	Sequences	Original papers
ndhF	972 F	5'-GTCTCAATTGGGTTATATG-3'	Olmstead and Sweere, 1994
ndhF	2110R	5'-CCCCCTACTATATTTGATACCTTCTC-3'	Olmstead and Sweere, 1994
ndhF	209F	5’-GGAACCCTTTCCCTTTGTGG-3’	New
ndhF	1159R	5’-CATCAATTACTCGTCGATCCCA-3’	New
ITS	ITS-A	5'-GGAAGGAGAAGTCGTAACAAGG-3'	Blattner, 1999
ITS	ITS-B	5'-CTTTTCCTCCGCTTATTGATATG-3'	Blattner, 1999
trnL-trnF	c	5'-CGAAATCGGTAGACGCTACG-3'	Taberlet *et al.,* 1991
trnL-trnF	f	5'-ATTTGAACTGGTGACACGAG	Taberlet *et al.,* 1991

2.4. Data processing and phylogenetic analyses

DNA sequences were imported in Geneious version 6.1.8 (http://www.geneious.com) where unassembled chromatograms were checked and edited for quality and contigs were assembled using the same program and a BLAST search was used to verify their authenticity using the same program and aligned using MUSCLE (Edgar, 2004) implemented in the software. Aligned sequences from Geneious were imported into MEGA version 6 (Tamura et al., 2013) for final manual editing.

I conducted both separate and combined phylogenetic analyses of the chloroplast genes (*ndhF* + *trnL-trnF*) and the nuclear gene ITS, using maximum likelihood (ML) and Bayesian inference (BI). Modeltest, implemented in PAUP*4.0a147 (http://people.sc.fsu.edu/~dswofford/paup_test/; Swofford, 2002), was used to assess models of nucleotide substitution rates among sites. PartitionFinder version 1.1.1 (Lanfear et al., 2012) was used to determine an appropriate data-partitioning scheme from potential partitions that were defined a priori (in this case, each locus). Substitution models for each gene were selected based on the Akaike Information Criterion (Akaike, 1973). The GTR model with rate variation among sites (GTR+G) was the best model for *ndhF* and *trnL-trnF*; the GTR model following a gamma distribution with invariant sites (GTR+G+I) was the best model for ITS. I used RAxML on CIPRES (Miller, Pfeiffer and Schwartz, 2010) to perform ML using the best model suggested by Modeltest with the rapid bootstrap option available in the program. Two independent analyses of BI were performed with MrBayes version 3.3.3 (Ronquist et al., 2012) for the combined dataset (cpDNA + ITS) under the substitution model GTR and rate variation Gamma (GTR+G). The Markov Chain Monte Carlo (MCMC) was run with 10 million generations; 0.2 heated chain temperature; 10,000 subsampling frequency; the first 20% of the trees from all runs were excluded as burn-in before making a 50% majority-rule consensus tree of the generated posterior distribution trees.

Support values are classified as follows: for BI analysis, strong support are posterior probabilities (PP) >0.95, 0.90-0.94 are moderate support, and low support is 0.70-0.89 PP. For ML analysis, strong support is bootstrap (BT) values >90%; moderate support 75-89%, and low support 50-74%. Branches with posterior probablilities lower than 0.7 PP and ML bootstrap <50% are considered unsupported.

Dataset incongruence was tested between the chloroplast and nuclear genes using the Incongruence Length Difference test (ILD) (Farris et al., 1995), implemented in the software PAUP*4.0a147 with 100 partition-homogeneity test replicates. Results indicate a non-significant *p*-value=0.06 that rejects the hypothesis of incongruence. Since the ILD test is often subject to error type I, i.e., it is biased toward incongruence of datasets (Planet, 2006), this adds more confidence to my results. The datasets were concatenated and both ML and BI analyses were carried out on the concatenated dataset. The GTR model with rate variation among sites (GTR+G) was the best model for the combined data set. The combined dataset includes a total of 2645 nucleotide bases.

2.5. Ancestral character state reconstructions

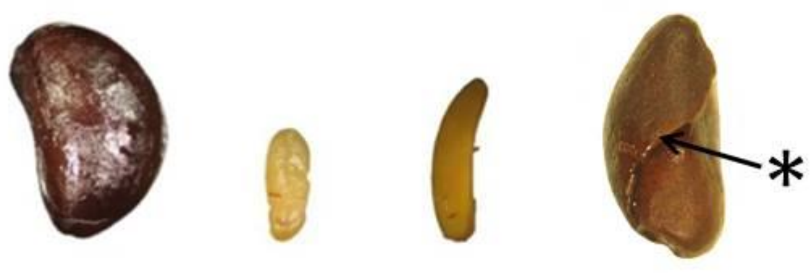
Ancestral character state reconstructions (ASR) were conducted to infer evolution of the number of anthers per fascicle, presence of anther glands, color of embryo, type of cotyledons, and size of embryo.

The stamens in Vismieae are in fascicles and the number of anthers varies within both group and species, as does the presence of glands on the anthers. *Harungana* and *Vismia rubescens* have 3 anthers; the rest of the African *Vismi*a have 6 to 25; American *Vismia* have more than 30; African *Psorospermum* have 6 to 8; and interestingly Malagasy *Psorospermum* have either 3 or 10, rarely 8, thus the number of stamens may be taxonomically important here. Glands are absent on the anthers of *Harungana* and *V. rubescens* but are always present on the anthers of the American *Vismia*. Glands are either present or absent on the anthers of the African and Malagasy *Vismia* and *Psorospermum*.

The embryo characters could also be taxonomically important in Malagasy *Psorospermum* as suggested by Bamps (1966), although he did not investigate them because he did not have enough specimens from Madagascar for his study (Bamps, 1966). The length of the embryo in Vismieae varies from 20 to 50 mm and the width varies from 2.5 to 8.5 mm. The cotyledons are yellow or green; they are rolled in African *Psorospermum* but not in American *Vismia*, African

Vismia, Harungana or Malagasy *Psorospermum* (Fig. 2). The character state codes are presented in Table 2.

Phytools (Revell, 2012) and ape (Paradis et al., 2004) packages in R (R Development Core Team, 2013) were used for the analysis; the ML tree was made ultrametric using the ape package, and the option ML was chosen so that the likelihood of changes occurring along the branches depended on their lengths. Model rates of transitions were compared and the best model chosen according to its Akaike Information Criterion (AIC) value.



1. Large embryo 2.Small embryo 3.Intermediate embryo 4.Rolled embryo

Fig. 2. Different types of embryos in Vismieae. 1. Large embryo , ca. 13 mm x 8 mm (*Psorospermum chionanthifolium*); 2. Small embryo, ca. 7 mm x 2.5 mm (*Harungana madagscariensis*); 3. Intermediate size embryo, ca. 10 mm x 3 mm (*Vismia sp*)*;* 4. Rolled embryo (*P. tenuifolium*); asterisk represents the edge of one of the cotyledons.

Table 2

Coding of the character states of the anthers and embryo characters for the Ancestral state character reconstructions analysis of the Malagasy *Psorospermum.*

Characters	**Character states**	**Character coding**
Number of anthers per fascicle	One	0
	Three	1
	Four and five	2
	Seven and eight	3
	Ten or more	4
Anther glands	Absent	0
	Present	1
Color of embryo	Yellow	0
	Green	1
Type of cotyledons	Symmetrical	0
	Rolled	1
Size of embryo	Small 7 mm x 2.5 mm	0
	Intermediate ca. 10 mm x 3 mm	1
	Large ca. 13 mm x 8 mm	2

2.6. Molecular dating, divergence time and biogeographical analyses

I used the combined dataset (ITS + cpDNA) for the analysis. The molecular clock hypothesis was tested using the likelihood ratio test (LRT) in MEGA version 6. The result rejected the null hypothesis of equal evolutionary rate throughout the tree at a 5% significance level ($p=0.1^{-23}$) so the parameters were set as uncorrelated lognormal relaxed clock-model in BEAUti version 1.7.0 (Drummond et al., 2012). GTR+G was used as a model of substitution, and birth and death process was chosen as a tree prior (Gernhard, 2008) with a random starting tree. Trees generated with BEAUti 1.7.0 were run with BEAST 1.7.0 (Drummond et al., 2012) to estimate divergence times. Topological constraints were applied to include ages from previous clusioid studies. *Psorospermum* has no fossil record so as age constraints I used the crown ages of Cratoxyleae, Hypericaceae, Hypericeae, and Vismieae as secondary calibration points, following Ruhfel et al. (2016) (Table 3). Ruhfel et al. (2011) used the Eocene pollen fossil *Pachydermites diederexii* and the macrofossil *Paleoclusia chevalieri* for calibration. While *Pachydermites diederexii* was placed without ambiguity as the most recent common ancestor (MRCA) of *Symphonia* and *Pentadesma* (Meseguer et al., 2013; Ruhfel et al., 2016), the placement of *Paleoclusia* remained uncertain, *Paleoclusia* can be placed either as the MRCA of Clusieae and Symphonieae (CC), or as the MRCA of Bonnetiaceae and Clusiaceae *s.s* (BC) (see Ruhfel et al., 2013, 2016). The ages found by Ruhfel et al. differ absolutely only for those two nodes where *Paleoclusia* is placed, but most other date ranges overlap. I ran the CC analyses and BC analyses separately using CIPRES (Miller, Pfeiffer and Schwartz, 2010). The root age was set at 93.5 Ma. This is 0.3 Ma older than the maximum age of the split of *Eliea articulata* from the rest of the Hypericaceae (Vismieae + Hypericeae) in the CC analysis which is estimated at ca. 93.2 Ma (Ruhfel et al., 2016). Normal distribution was chosen for all priors so that the standard deviation covers the entire confidence interval of ages (see discussion) (Table 3).

Four independent MCMC searches of 10 million generations each were run with a sampling frequency of 1000 generations. The MCMC Effective Sample Size (EES) was checked with the software Tracer v.1.6 (Rambaut et al., 2014). ESS is large for all parameters (all values superior to 200 and less than 10,000), indicating that estimates of their posterior distribution are adequate. Samples of each run were combined with LogCombiner version 1.7.1 (Drummond and Rambaut, 2012), and the maximum clade credibility tree was generated with the software TreeAnnotator version 1.7.0 (Drummond and Rambaut, 2013). The software FigTree version 1.4.2 (Drummond and Rambaut, 2013) was used to visualize the dated tree. Additionally I used the R package ape to generate the lineage through time plot (LTT) of the phylogenetic tree under a birth-death (b-d) model (Nee 2006) with the assumption that extinction (d) and speciation (b) rates are constant through time.

Table 3
Priors used in the molecular dating analysis with Beast 1.7.0. BC= *Paleoclusia* placed at the most recent common ancestor of Bonnetiaceae and Clusiaceae *s.s.*; CC= *Paleoclusi*a placed at the most recent common ancestor of Clusieae and Symphonieae; Std dev=Standard deviation.

	BC		**CC**		
	Mean crown age (Ma)	**Std dev**	**Mean crown age (Ma)**	**Std dev**	**Prior distribution**
Cratoxyleae	43.7	15	46.7	14.9	normal
Hypericaceae	71.5	9.1	77	10	normal
Hypericeae	37.3	8	39.7	8	normal
Vismieae	40.7	7.85	44.3	9.1	normal

To infer the biogeography of *Psorospermum*, I used the biogeographical regions suggested by Buerki et al. (2011); however, Southeast Asia and Australia were excluded because none of the study taxa occur in those areas. Thus the following five biogeographical regions only were used: (A) Eurasia from western Europe to Indochina, (B) Africa, (C) Madagascar, (D) North America, and (E) Central and South America. I used the model-based statistical inference of biogeography BioGeoBEARS version 0.2.1 (Matzke, 2013) R package (R DevelopmentCore Team 2013), and compared the Akaike Information Criterion (AIC) scores calculated from the dispersal model Dispersal-Extinction Cladogenesis (DEC) and the DEC with a “jump dispersal” parameter J (DEC + J). The latter assumes the rapid formation of an independent lineage right after the dispersal event (Matzke, 2013). I did not use the model DIVALIKE, a likelihood version of the model dispersal-vicariance, because vicariance is irrelevant according to the ages of the calibration used in this study. The model BayAreaLIKE was also not used because this model was designed to accommodate analyses with numerous biogeographical regions (Matzke, 2013) and this study includes only five regions.

3. Results

3.1. Molecular phylogeny

Data statistics of the partitions (Cp DNA data and ITS data) are presented in Table 4. The combined aligned data matrix is 2645 base pairs long, 1021 (38.60%) of which are parsimony informative**,** and the combined data is used in subsequent analyses.

Table 4

Data statistics in the molecular phylogenetic analysis of *Psorospermum*.

	Number of sequences in the ingroup	Length (base pairs)	Variable characters	% Parsimony informative	% CG content
Combined dataset (Cp DNA + ITS)	114	2645	988	31.00% (820/2645)	37.1
CpDNA (*ndhF* + *trnL-trnF*)	114	1865	663	31.09% (580/1865)	29.9
ITS cloned	191	776	419	37.37% (290/776)	58.3

Fig. 3, displayed from TreeGraph2 (Stöver and Müller, 2010), shows a ML tree with a single taxon name representing the combined accessions of that taxon; nodes with <50% ML support were collapsed. As the BI and ML analyses yielded similar topology, PP values are displayed above the branches as well. BI and ML trees with all individuals are displayed in the Appendix (Figs. S2 and S3). Previous relationships ((*Harungana madagascariensis* + *Vismia rubescens*) + American Vismieae), named clade A, are strongly supported in both analyses (BT: 100%; PP: 1). Clade A is sister to (clade B + clade C + clade D) (Fig. 3), made up of all other Malagasy and African Vismieae (i.e. *Psorospermum* and *Vismia* from Africa). Clade B consists of (Malagasy morphospecies *P. androsaemifolium*, *P. cf. androsaemifolium*, and *P. sp 19* + African Vismieae). Clade C consists of Malagasy morphospecies (*P. cerasifolium* + *P. malifolium* + *P. sp 22* + *P. sp 2*). Five clades in clade D (D1–D5) were recovered in both ML and BI phylogenies. Clade D1 consists of ((*P. ferrovestitum*+ *P. fanerana* + P. sp16) + P. sp17); clade D2 shows *P. revolutum* as sister of *P. cf. lanceolatum*; the two African *Psorospermum* species, *P. membranaceum* and *P. staudtii*, are in clade D3 and the overall realtionships are (((*P. membranaceum*+*P. staudtii*)+*P. lamianum*) + (*P. nanum* + *P. humile*)); in clade D4 ((*P. crenatum* + *P. chionanthifolium*) + (*P. brachypodum* + *P. cf. brachypodum*)); a clade including *P. rienanense*, *P. sexlineatum* in clade D5 is sister to ((*P. atro-rufum* + *P. cf. atro-rufum*) + (*P. sp11* + *P. sp13*)).

Malagasy *Psorospermum* are paraphyletic; the clade, Clade B, is sister to the remaining Malagasy and African Vismieae (clade C + clade D). The node has low support in ML (BT: 60 %) but strong support in BI (PP: 0.97). The clade (African *P.staudtii* + African *P. membranaceum*) is nested in Clade D and is a clade sister to Malagasy morphospecies *P. lamianum* (BT: 99%, PP: 0.99). The placement of morphospecies *P. rubrifolium*, *P. sp 10, P. sp 15*, and the African *P. tenuifolium* remains ambiguous, although all are clearly members of Clade D.

3.2. Divergence times and biogeographical inference

Differences between the mean crown ages of the clades in the two analyses (BC vs CC) are very small (Table 5). However, the Highest Posterior Density intervals (HPD) of the dates at those clades are large (Table 5). When compared to ages in Ruhfel et al., the mean crown ages of Vismieae (BC and CC analyses) retrieved from this analysis are substantially older, yet the 95% HPD intervals overlap (Table 5).

Table 5
Comparison of the crown ages of the major clades in Vismieae

Clades	**Ages from the present study (Ma)**		**Ages from Ruhfel *et al.* 2016(Ma)**	
	CC	**BC**	**CC**	**BC**
Vismieae	51.05 (95%HPD=67.5-34.5)	50.79 (95%HPD=68-33.5)	44.3 (95%HPD=60.9-29.5)	40.7 (95%HPD=55.5-28.3)
(*H.m.*+*V.r.*) + Am. *V.*)	31.45 (95%HPD=67.5-34.5)	30.31 (95%HPD=68-33.5)	_	_
(M.*Ps.* + Afr. *V.*)	29.7 (95%HPD=46.5-15.6)	29.37 (95%HPD=46.2-13.5)	_	_

BC= *Paleoclusia* placed at the most common recent ancestor of Bonnetiaceae and Clusiaceae *s.s.*; CC= *Paleoclusia* placed at the most common ancestor of Clusieae and Symphonieae. H.m.= *Harungana madagascariensis*; V.r.= *Vismia rubescens*; Am. V.=American *Vismia*; M.Ps.= Malagasy *Psorospermum*; Afr.V.=African Vismieae).

The time-calibrated phylogeny suggests three pairs of African-Malagasy lineages (Fig. 4; Fig. S7), one ca. 20.7 million years (Ma), a second ca. 5.84 Ma, and a third ca. 5.35 Ma (nodes B, C and D, Figs. 4 and 5). Major diversification of Malagasy *Psorospermum* may have begun ca. 20 Ma (see discussion). The DEC model fits the data best compared to the DEC + J model, with AIC scores shown in table 6. The DEC model suggests that *Psorospermum* arrived by dispersal from Africa to Madagascar ca. 29 MY (node A, Fig. 5) followed by a subsequent in situ radiation (node C Fig. 5). At node B, ca. 20-19 Ma, *Psorospermum s.l.* may have been in both Africa and Madagascar, and one African lineage is widespread (Fig. 5). The model indicates also at least two dispersal events back to Africa ca. 5.5 and 5.7 MY (Fig. 5).

Table 6
Biogeographic model fit comparison

Model	**LnL**	**d**	**e**	**j**	**AIC**	**AIC_wt**
DEC	-43.53	0.0022	$1.0e^{-12}$	0	91.06	0.017
DEC + J	-40.04	$1.0e^{-12}$	$1.0e^{-12}$	0.0059	86.08	0.20

LnL: loglikelihood; d=estimated dispersal rate; e=estimated extinction rate; j=founder event speciation rate; AIC_wt=AIC weight

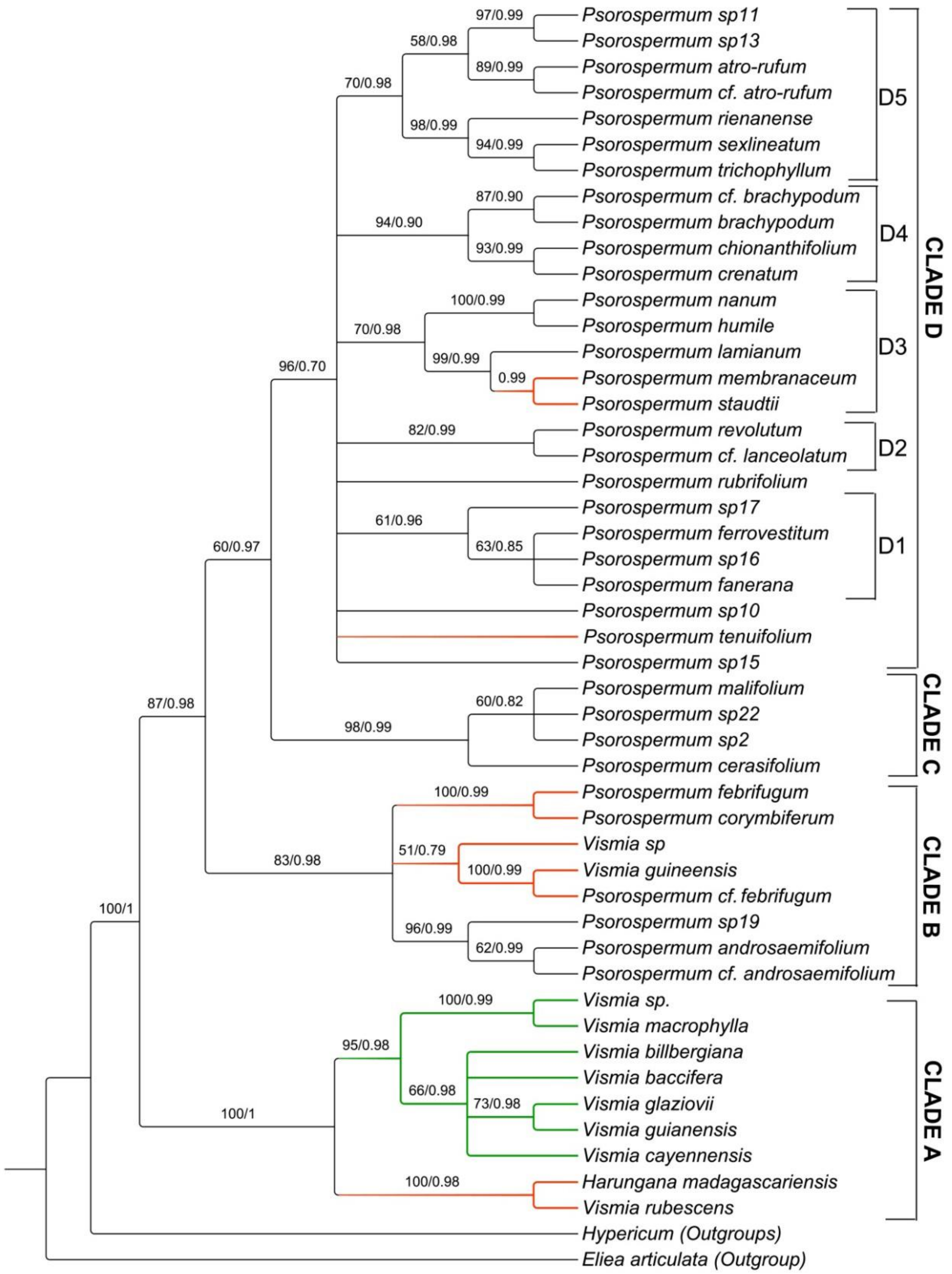


Fig. 3. Maximum Likelihood tree (ML) tree inferred from analyses using combined chloroplast DNA regions (*trnL-F, ndhF*) and nuclear ribosomal DNA (ITS). Values above branches on the left denote maximum likelihood bootstrap support (BT %) and those on the right are Bayesian posterior probabilities (PP).In the ingroup, green branches indicate American taxa; dark orange: African; and black: Malagasy. *Psorospermum sp2, 10, 11, 13, 15, 16, 17, 19,* and *22* refer to species recognized during the course of a total evidence analysis of the variation in the genus (Ranarivelo et al. in preparation)

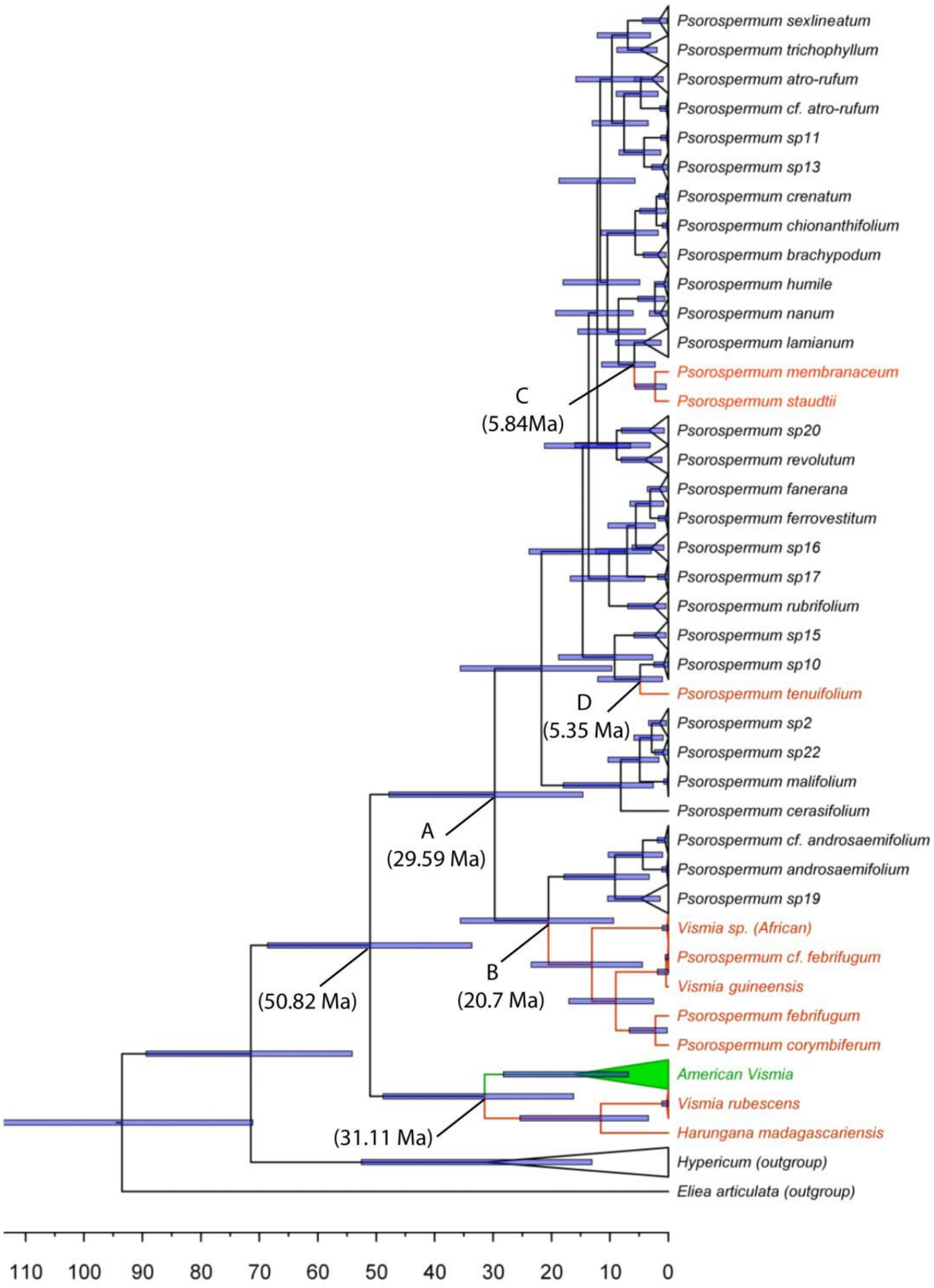


Fig 4. Time-calibrated phylogeny of Vismieae using a normal probability prior of the CC analysis. Bars represent the Highest Posterior Density (HPD) intervals of the dating analysis. Numbers represent the node ages. Dark orange: African taxa; green: American taxa; black (ingroup): Malagasy taxa. The time calibrated tree with all individuals and all areas are displayed in the Appendix (Fig S6).

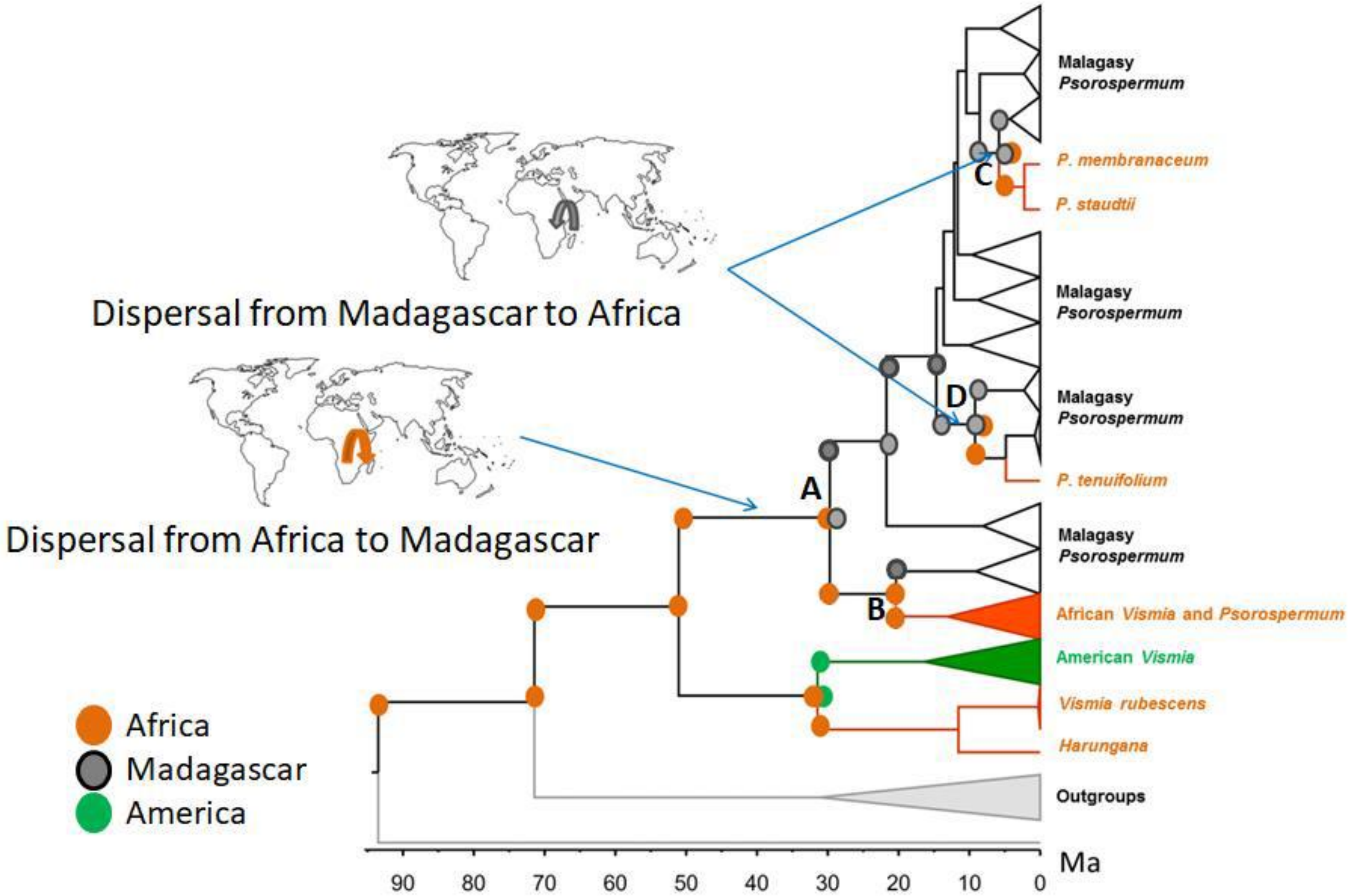


Fig. 5. Schematic diagram of the DEC-based ancestral area reconstruction. Ancestral areas are indicated by colored circles. Dispersal events are indicated by the blue arrows, dispersal direction is indicated by colored curved arrows on the maps. The time calibrated tree with all individuals and all reconstructed areas are displayed in the Appendix Fig. S8.

3.3. Ancestral State Reconstructions and character evolution

Number of anthers: “stamen fascicle with more than 10 anthers” is likely an ancestral character in Malagasy *Psorospermum* (Fig. 6A node A1). There is likely a change from “stamen with more than ten anthers” to “stamen with three anthers” occurring on the branch subtending Clade D (Fig. 6, node A2) see also discussion).

Presence of anther glands: “anther glands present” is likely ancestral in Malagasy *Psorospermum* (Fig. 6B, node B1). The anther glands may be lost at least three times in Africa and Malagasy *Vismia* and *Psorospermum,* and gained once in African *Vismia*. However, if the states “ten” and “more than ten” are coded as one state, this state is most likely ancestral in the Malagasy *Psorospermum* (node A1, Fig. 7); and when numbers of anthers is coded as “equal to ten” versus “more than ten”, the ancestral state of Malagasy *Psorospermum* is likely “three anthers” (node A1, Fig. 7).

Type of embryo: “symmetrical cotyledons” is reconstructed as the ancestral state to all Malagasy *Psorospermum* (Fig. 8), and “rolled embryo” evolved independently at least three times in the African and Malagasy *Vismia* and *Psorospermum*.

Size of embryo: the MRCA of Malagasy *Psorospermum* likely had a “large embryo”, with a reversal to “small embryo” in the Malagasy *Psorospermum* (*P. androsaemifolium* + *P. cf. androsaemifolium* + *P. sp19*) in clade B (Fig. 9A).

Color of embryo: “Green embryo” is likely the ancestral character state of Malagasy *Psorospermum* (Fig. 9B). “Yellow embryo” evolved independently at least two times in African and Malagasy *Psorospermum* and *Vismia*.

4. Discussion

4.1. Phylogenetic analyses

The topologies of ML and BI trees are similar (Figs S2 and S3), although most nodes have lower ML bootstrap values than posterior probabilities. Bootstrap proportion tends to underestimate accuracy, i.e., the probability of recovering the true clade (Felsenstein and Kishino, 1993; Sanderson and Shaffer, 2002), and is sensitive to the amount of phylogenetic signal, whereas Bayesian inference tends to perform better than ML when the data have fewer informative characters (Alfaro et al., 2003). The short branches (Fig. S4) and the lack of resolution along the spine of Clade D (Figs. 3 and S4) could reflect homoplasy, but the chloroplast gene tree and the nuclear gene tree yielded the same topology (Appendix Figs. S5 and S6). When the datasets were combined the result did not change. The basal topology in Vismieae is the same in both cpDNA and nuclear DNA analyses, as is the poor resolution within Clade D. This may indicate that both Malagasy *Psorospermum* and African Vismieae underwent rapid radiation. Under a birth-death (b-d) model (Nee 2006) with the assumption that extinction (d) and speciation (b) rates are constant through time, the lineage through time (LTT) plot suggests that diversification is slow at early stages in Vismieae but later increases at ca. 20 Mya (Fig. 10).

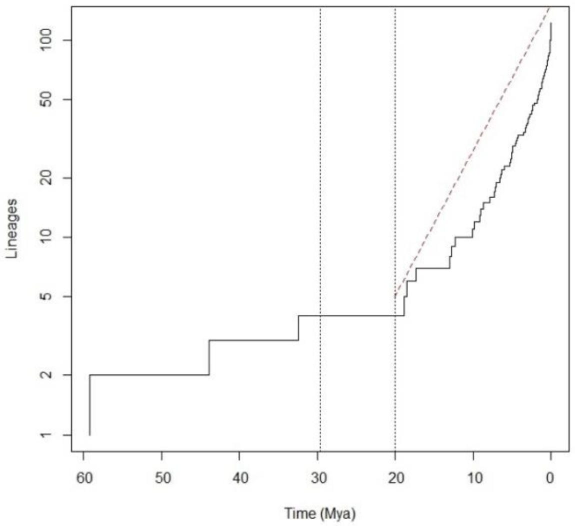


Fig. 10. Lineage through time (LTT) plot for the Africa/Madagascar taxa under a birth-death model with constant rate of extinction and speciation. Grey dotted lines represent the period of time where *Psorospermum* started to diversify in Madagascar; red dotted line indicates a rapid diversification in Vismieae after 20 Mya.

4.2. Implications for the circumscription of taxa within Vismieae

My results support recognition of the two genera *Harungana* and *Psorospermum* that were previously considered to be synonymous by Stevens (2007). The results also confirm the topologies presented by Ruhfel et al. (2011, 2013, 2016), with strong support and additional exhaustive taxon sampling. *Harungana* can be expanded to include *Vismia rubescens*, and *Psorospermum s.l.* can be expanded to include all other African and Malagasy species of Vismieae, and the name is used in this expanded sense in the discussion below. My study suggests that Malagasy *Psorospermum* do not form a monophyletic group, but some African Vismieae are embedded within them. Although sampling of African Vismieae includes only four of the six species previouslsy included in *Vismia*, all five African species of *Psorospermum* are included in this analysis, and both the cpDNA tree and the nuclear DNA tree have the same topology; the major clades (A, B, C, and D) are supported in both analyses (Figs. S5 and S6) as well as in the combined data analysis (Figs. 3, S2, S3). The absence of conflict in the gene trees suggests that lineages had enough time to be sorted into clades, so incomplete lineage sorting is not likely to explain the paraphyly of Malagasy *Psorospermum*.

4.3. Biogeographic implications

Were there recent dispersals of *Psorospermum* back to Africa? A previous study suggested an African origin of Vismieae and of the Malagasy *Psorospermum* (Ruhfel et al., 2016). My biogeographical analysis corroborates those hypotheses. The ancestral area reconstruction suggests also that *Psorospermum* may have reached Madagascar by a single dispersal event from Africa during the Oligocene (ca. 34-22 Ma) and this event was followed by diversification on Madagascar beginning ca. 20-19 Ma (Figs. 4, 5). Many Malagasy angiosperm lineages arrived in Madagascar at about this time in the Miocene (Buerki et al., 2013), and paleo-oceanographic currents were flowing from northeast Mozambique and Tanzania eastward towards Madagascar, continuing at least through the Oligocene epoch (ca. 33.9-23 Ma) (Ali and Huber, 2010), which would favor the dispersal of species from Africa to Madagascar.

Psorospermum s.l. was present in both Madagascar and Africa ca. 20 Ma (Fig. 5). Perhaps range contraction did not happen until formation of the Malagasy lineage and the African lineage in clade B (Fig. 5). Note that the program BioGeoBears includes the assumption that lineages can live in both areas after a dispersal event, unlike DIVA for example, where after dispersal there is no trace of the dispersed species in its area of origin (Matzke, 2013).

My results also suggest two recent dispersal events out of Madagascar back to Africa in the late Miocene. With its fruits being small berries (ca. 0.5 cm diameter) containing small seeds (ca. 3 x 2 mm), *Psorospermum* could perhaps have been dispersed by smaller animals, such as birds. Out-of-Madagascar dispersal events have been inferred for a number of groups of insects as well. Notably, the Comoros Islands have frequently been the recipients of flora and fauna from Madagascar, e.g. Rubiaceae (Krüger et al., 2012) and Celastraceae (Bacon et al., 2016); however, *Psorospermum* is not known from the Comoros Islands. Other cases of movement to mainland Africa include grammitid ferns (Bauret et al. 2017), chameleons (Raxworthy et al. 2002), rodents (Jansa et al. 1999), butterflies (Zakharov et al. 2004), and a large-bodied diving beetle (Bukontaite et al. 2015).

I note that use of secondary calibrations for molecular dating analysis can be inaccurate due to a large uncertainty in age estimates (Shaul and Graur, 2002; Graur and Martin, 2004; Morrison, 2010; Schenk, 2016). I applied a normal distribution for the prior, which is appropriate when approximating the mean of the calibration age (Ho and Phillips, 2009), given that there is no fossil record for *Psorospermum*. Although recent studies suggest that using a normal prior can lead to greater error than using a uniform prior (Schenk, 2016), a uniform calibration prior requires both maximum and minimum age bounds, with at least the minimum bound provided by the fossil member of the clade (Ho and Phillips, 2009). The uniform distribution also places equal probability across all ages spanning the interval between the lower and upper bounds, so the mean of the calibration age cannot be approximated. Despite the dating caveats, my results support an African origin of the Malagasy *Psorospermum* (Ruhfel et al., 2016)

4.4. Apomorphies and noteworthy character evolution

Apomorphies:

Two character states, three anthers per stamen fascicle and anthers without glands, are synapomorphies for (*Harungana madagascariensis* + *Vismia rubescens*). The clade of American *Vismia* is supported by the character state hairy staminodes, and also by the presence of hypericin glands on the inside of the carpel wall (they are not visible outside). In *Psorospermum* and African *Vismia*, the hypericin glands, when present, are visible on the outside of the carpel. However, I did not identify any synapomorphies for the clade ((*H. madagascariensis* + *V. rubescens*) + American Vismieae). One character in which *Harungana* and American *Vismia* differ from Malagasy and African *Psorospermum* is the large size of the thyrsoid inflorescence. Peduncle and pedicels are conspicuously long in *Harungana* and and American Vismieae, while they are mostly condensed in Malagasy *Psorospermum*. Large, green embryos can be placed as synapomorphies for the (clade B+ clade C + clade D).

Convergence of characters:

Convergence of characters in the Malagasy *Psorospermum* of clades B and C are noteworthy. The Malagasy species can be sorted into two groups, one with three anthers per fascicle, and one with ten anthers. *Psorospermum cerasifolium, P. malifolium* and *P. androsaemifolium* all have ten anthers (Table 7) but *P. malifolium* and *P. cerasifolium* are in clade C (Figs. 6 and 7), while *P. androsaemifolium* is sister to the African Vismieae in clade B (Figs 6 and 7). The “loss of anther glands” seems have happened independently in the Malagasy *Psorospermum* of clades B and C (Table 7 and Figs. 6 and 7). “The number of sepal glands” is also convergent in *P. malifolium* and *P. androsaemifolium* (Table 7); both have no more than four glands on their sepals, usually only two.

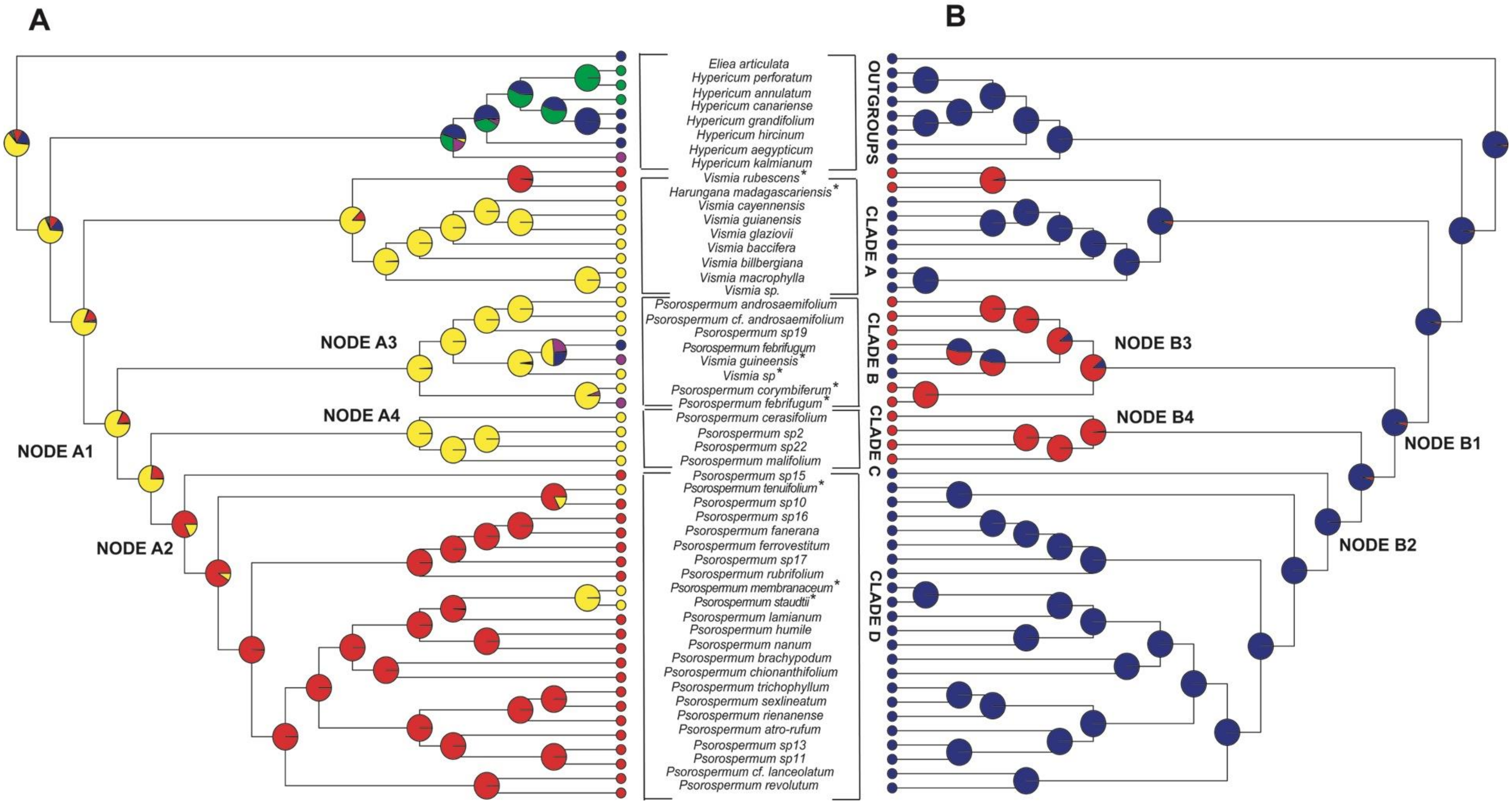


Fig. 6. Comparison of the Ancestral Character State Reconstruction of the number of anther (**A**) and anther glands (**B**). The anther number ⩽10 and anther number >10 were included in a single state. **A.** Green: 1 anther; red: 3 anthers; blue:4-5 anthers; purple:7-8 anthers; yellow: 10 anthers. **B.** Blue: anther glands present; red: anther glands absent. African species are marked with an asterisk.

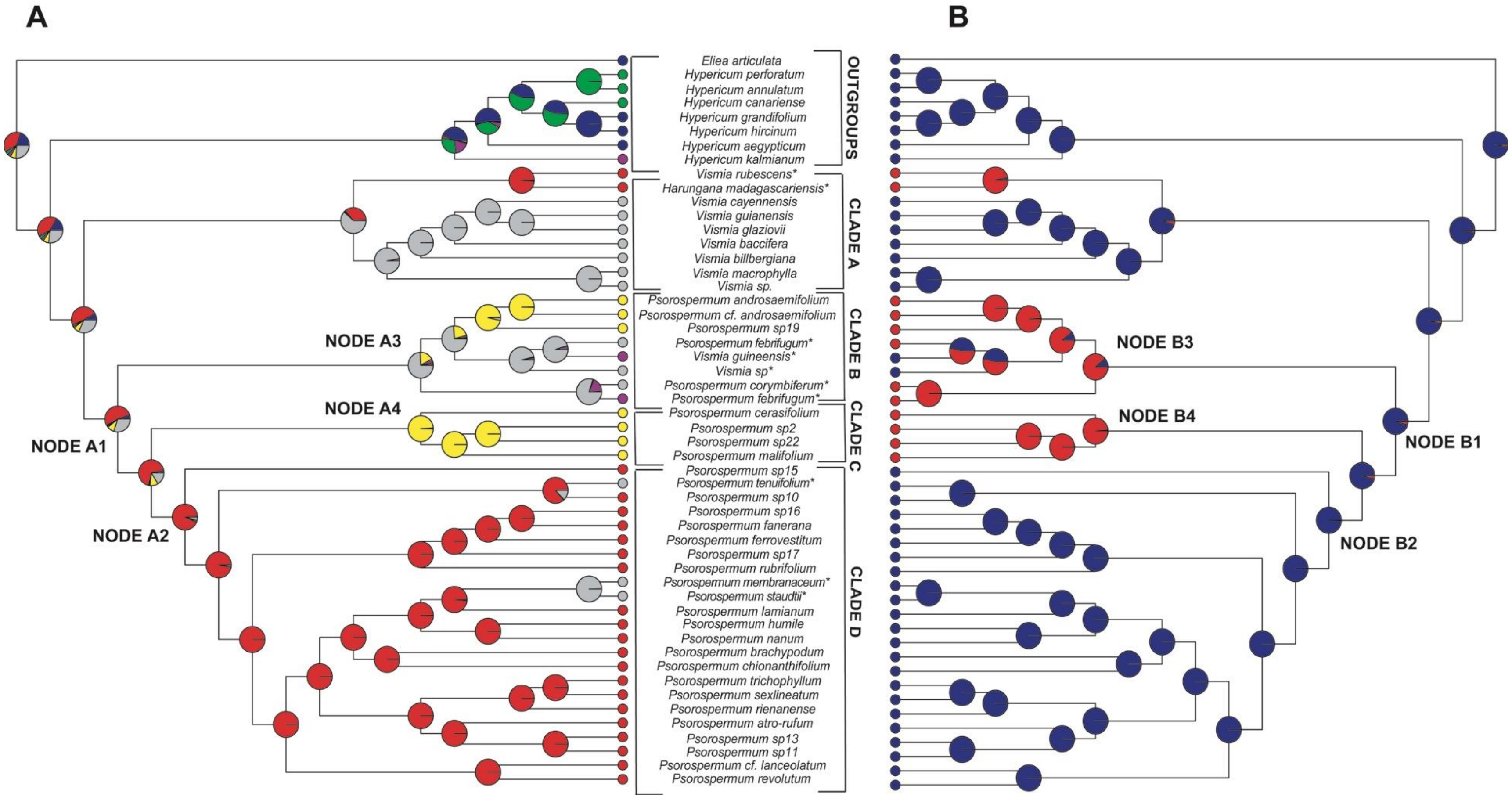


Fig. 7. Comparison of the Ancestral Character State Reconstruction of the number of anthers (**A**) and the anther glands (**B**). The anther number of =10 and anther number >10 were coded as separate states. **A.** Green: 1 anther; red: 3 anthers; blue:4-5 anthers; purple:7-8 anthers; gray = 10; yellow > 10 anthers. **B.** Blue: anther glands present; red: anther glands absent. African species are marked with an asterisk.

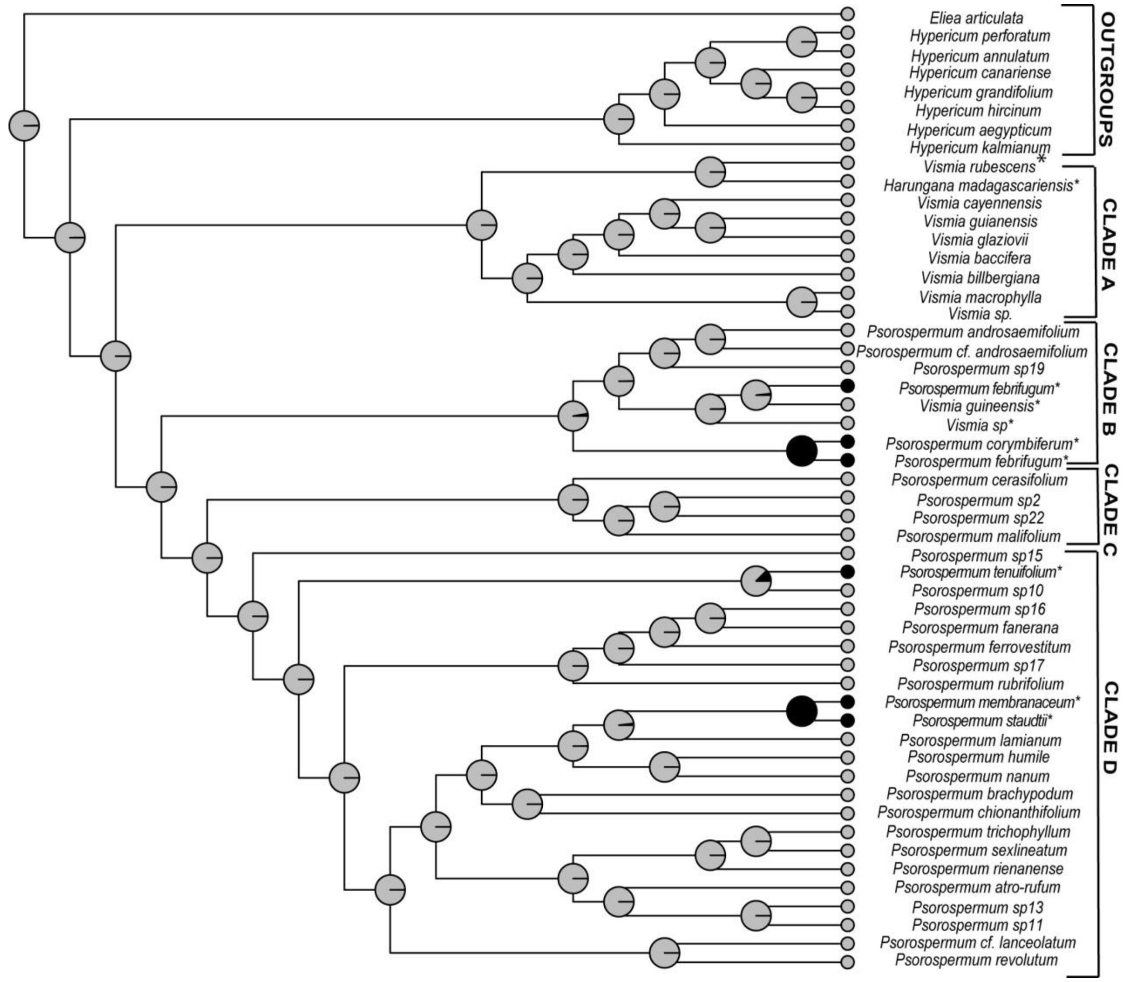


Fig. 8. Ancestral Character State Reconstruction of the type of cotyledons. Grey: symmetrical cotyledons; black: rolled cotyledons. African species are marked with an asterisk.

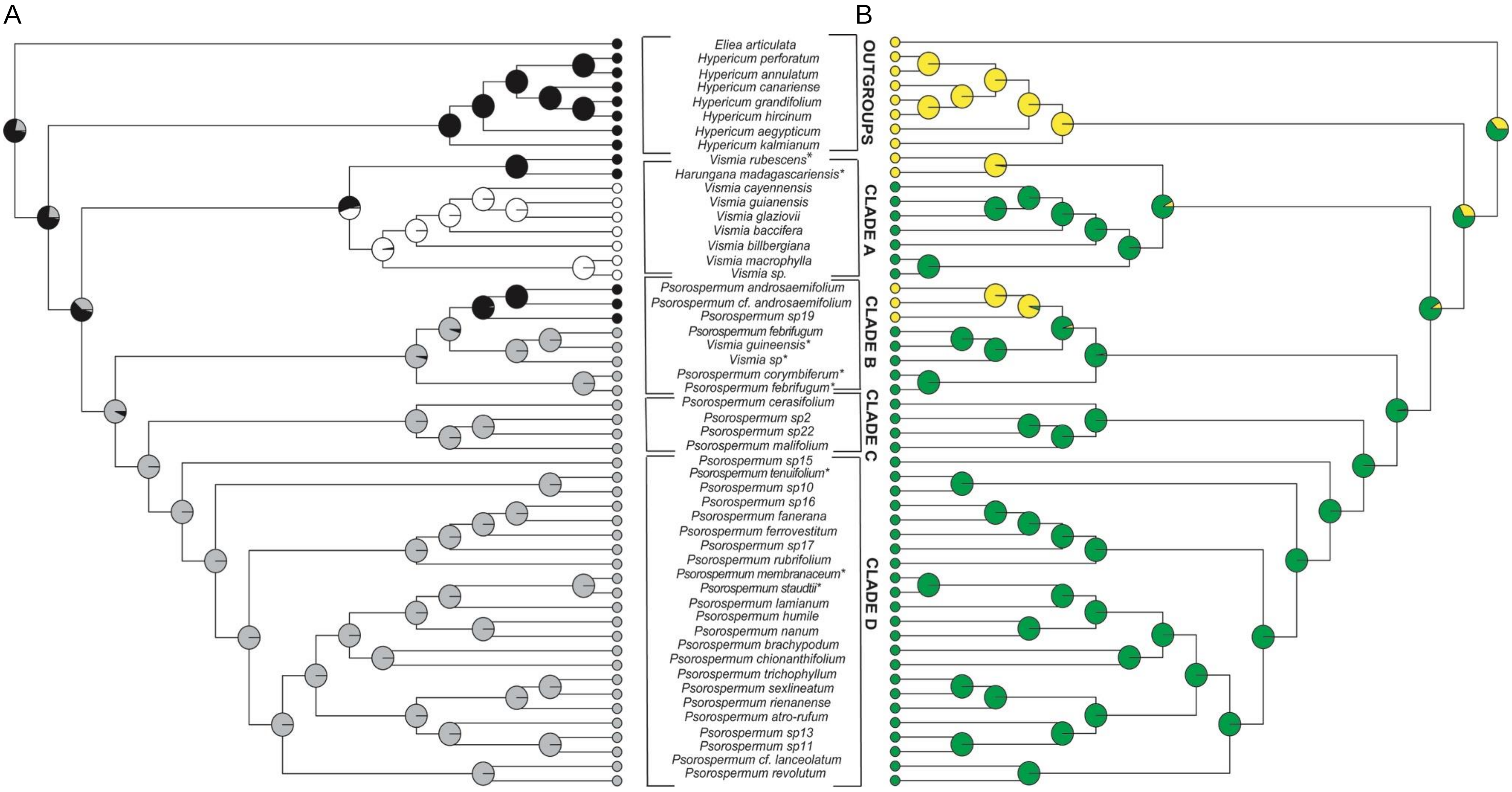


Fig. 9. Comparison of the Ancestral Character State Reconstruction of embryo size (**A**) and embryo color (**B**). A. Black: small embryo; white: intermediate, grey: large embryo. **B**. Yellow: yellow embryo, green: green embryo.

Table 7
Summary of the characters in Malagasy *Psorospermum* in clades B and C.

Clade	Species name	Number of anthers	Anther glands	Hair	Secondary veins	Number of sepal glands	Glands on the seed coat	Embryo color	Embryo type
BB	*P. sp19.*	10	absent	present	conspicuous	1-4	absent	yellow	straight
BB	*P. androsaemifolium* Bak.	10	absent	present	conspicuous	1-4	absent	yellow	straight
BB	*P. cf. androsaemifolium*	10	absent	present	conspicuous	1-4	absent	yellow	straight
BC	*P. sp22*	10	absent	absent	not conspicuous	>10	present	green	curved
BC	*P. cerasifolium* Baker	10	absent	absent	not conspicuous	>10	present	green	rolled
BC	*P. sp2*	10	absent	present	not conspicuous	2-4	present	green	rolled
BC	*P. malifolium* Baker	10	absent	present	not conspicuous	2	present	green	rolled

Pattern of evolution:

The comparison of the patterns of evolution of the characters “numbers of anthers” and “anther glands” of the Malagasy species in clades B (*P. androsaemifolium, P. cf. androsaemifolium*, *P. sp19*) and C (*P. malifolium, P. cerasifolium*) with those of clade D are noteworthy (Fig. 6). Differences in anther glands seems to match the number of anthers; species without anther glands always have ten anthers per fascicle while those with anther glands have three. Most of the species of clade D, with three anthers, occur in primary and secondary forest habitats, humid rainforests or littoral forests, while species in clades B and C, with ten or more anthers, occur in open, mostly dry areas; *P. cerasifolium* and *P. malifolium* grow mainly in open habitats and dry climates, and the widespread *P. androsaemifolium,* is also common in such habitats. However my results cannot support any conclusions about the co-evolution of those characters in clade D vs clades B and C; ASR and character optimization are subject of several issues as discussed in literature (Wortley et al., 2015; Lu et al. 2015; Stevens 2001 onwards). How the character “numbers of anthers” is coded affects interpretation of anther evolution. If the states “ten” and “more than ten” are coded as one state “ten or more anthers”, this state is most likely ancestral in the Malagasy *Psorospermum* (node A1, Fig. 6). However, the two states are kept separate, the ancestral state of Malagasy *Psorospermum* is likely “three anthers” (node A1, Fig. 7).

Little is known about the functions of the dark hypericin-containing glands in *Psorospermum*. During development acetate and malonate derived from glucose, fructose and galactose accumulate first in vacuoles and then in the periplasmic space and cell walls. These derived materials are transformed into emodin that is transformed later into hypericin and pseudohypericin (Zobayed et al., 2004). Hypericin synthesis may take place in the dark glands because emodin is present at a high concentration there but absent in other tissues (Zobayed et al., 2004). The role of hypericin glands in pollination and seed dispersal in Hypericaceae needs to be investigated. In some Myrtaceae species, for example, the secretion of the anther gland is hydrophobic and mixes with the pollen to increase levels of edible lipids for pollen-foraging insects (Beardsell et al., 1989).

5. Conclusion

This study focuses on the Malagasy *Psorospermum*. My molecular phylogeny includes exhaustive sampling of *Psorospermum* from Madagascar and mainland Africa and strongly supports *Harungana* (including *Vismia rubescens*) and *Psorospermum* (including other African *Vismia*, African and Malagasy

Psorospermum) as two genera. Additionally, this study identified several well-supported clades within *Psorospermum* that will be important for further studies, e.g., in integrative species delimitation (Ranarivelo et al. in preparation). Contrary to expectations of a monophyletic Malagasy *Psorospermum* reflecting uni-directional dispersal from Africa to Madagascar during the Oligocene, my findings strongly suggests more complex movements, including recent dispersals back to Africa. My results also provide information about the evolution of morphological characters within *Psorospermum*; the group is the subject of considerable homoplasy, and synapomorphies are scarce. One hypothesis is that lineages may have evolved rapidly, especially during the recent radiation in Madagascar (ca. 20 Ma onwards), with few mutations that translated into distinctive character states. Other types of morphological data and additional approaches (anatomy, developmental studies, gene expression) should be applied to further address these questions. Additional molecular data (e.g., nuclear low copy genes; next-generation genome sequencing) would improve the result of the molecular phylogenetic analysis.

Acknowledgements

I thank Dr. P. F. Stevens for his advice, insights and support during this work, and Dr. P. C. Hoch, Dr. E. A. Kellogg, and Dr. B. R. Ruhfel for their constructive comments and suggestions. I thank the Madagascar National Parc (MNP) and the Missouri Botanical Garden for their help during the fieldwork. Fieldwork for this research was fully funded by a Christensen Research Scholarship awarded through the International Center for Tropical Ecology at the University of Missouri – St. Louis.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Caski, F. (Eds.), Proceedings of the Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Alfaro, M.E., Zoller, S., Lutzoni, F., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20, 255–266.

Ali, J.R., Huber, M., 2010. Mammalian biodiversity on Madagascar controlled by ocean currents. Nature 463, 653–6.

Álvarez, I., Wendel, J.F., 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29, 417–434.

Bacon, C.D., Simmons, M.P., Archer, R.H., Zhao, L-C, Andriantiana, J., 2015. Biogeography of the Malagasy Celastraceae: multiple independent origins followed by widespread dispersal of genera from Madagascar. Mol. Phylogenet. Evol. 94, 365–382.

Baldwin, B.G., Sanderson, M.J., Porter, J.M., Wojciechowski, M.F., Campbell, C.S., Donoghue, M.J., 1995. The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Ann. Missouri Bot. Gard. 82, 247–277.

Bamps, P., 1966. Notes sur les Guttiferae d'Afrique tropicale. Bull. Jard. Bot. État Bruxelles 36, 425–459.

Bauret, L., Gaudeul, M., Sundue, M.A., Parris, B.S., Ranker, T.A., Rakotondrainibe, F., Hennequin, S., Ranaivo, J., Selosse, M.A., Rouhan, G., 2017. Madagascar sheds new light on the molecular systematics and biogeography of grammitid ferns: new unexpected lineages and numerous long-distance dispersal events. Mol. Phylogenet. Evol. 111, 1–17.

Beardsell, D.V., Williams, E.G., Knox, R.B., 1989. The structure and histochemistry of the nectary and anther secretory tissue of the flowers of *Thryptomene calycina* (Lindl) Stapf (Myrtaceae). Aust. J. Bot. 37, 65–80.

Bellard, C., Leclerc, C., Leroy, B., Bakkenes, M., Veloz, S., Thuillier, W., Courchamp, F., 2014. Vulnerability of biodiversity hotspots to global change. Global Ecol. Biogeogr. 23, 1376–1386.

Blattner, F.R., 1999. Direct amplification of the entire ITS region from poorly preserved plant material using recombinant PCR. BioTechniques 27, 1180–1186.

Buerki, S., Devey, S.D., Callmander, M.W., Phillipson, P.B., Forest, F., 2013. The endemic and non-endemic vascular flora of Madagascar updated. Bot. J. Linn. Soc. 171, 304–329.

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

Bukontaite, R., Ranarilalatiana, T., Randriamihaja, J.H., Bergsten, J., 2015. In or out-of-Madagascar? Colonization patterns for large-bodied diving beetles (Coleoptera: Dytiscidae). PloS ONE. 10, p.e0120777.

Callmander, M.W., Phillipson, P.B., Schatz, G.E., Andriambololonera, S., Rabarimanarivo, M., Rakotonirina, N., Raharimampionona, J., Chatelain, C., Gautier, L., Lowry, II, P.P., 2011. The endemic and non-endemic vascular flora of Madagascar updated. Plant Ecol. Evol. 144, 121–125.

Coffin, M.F., Rabinowitz, P.D., 1988. Evolution of the conjugate East African – Madagascan margins and the western Somali Basin. Special Paper 226. Geol. Soc. of Am., Boulder.

Doyle, J.J., Doyle, J.L., 1990. Isolation of plant DNA from fresh tissue. Focus 12, 13–15.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Epifano, F., Fiorito, S., Genovese, S., 2013. Phytochemistry and pharmacognosy of the genus *Psorospermum*. Phytochem. Rev. 12, 673–684.

Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1995. Constructing a significance test for incongruence. Syst. Biol. 44, 570–572.

Federman, S., Dornburg, A., Daly, D.C., Downie, A., Perry, G.H., Yoder, A.D., Sargis, E.J., Richard, A.F., Donoghue, M.J., Baden, A.L., 2016. Implication of lemuriform extinctions for the Malagasy flora. Proc. Natl. Acad. Sci. 113, 5041–5046.

Federman, S., Dornburg, A., Downie, A., Richard, A.F., Daly, D.C., Donoghue, M.J., 2015. The biogeographic origin of a radiation of trees in Madagascar: implications for the assembly of a tropical forest biome. BMC Evol. Biol. 15, 216.

Felsenstein, J., Kishino, H., 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. 42, 193–200.

Gernhard, T., 2008. The conditioned reconstructed process. J. Theor. Biol. 253, 769–778.

Graur, D., Martin, W., 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet. 20, 80–86.

Ho, S.Y.W., Phillips, M.J., 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst. Biol. 58, 367–380.

Hong-Wa, C., Besnard, G., 2013. Intricate patterns of phylogenetic relationships in the olive family as inferred from multi-locus plastid and nuclear DNA sequence analyses: a close-up on *Chionanthus* and *Noronhia* (Oleaceae). Mol. Phylogenet. Evol. 67, 367–378.

Jansa, S.A., Goodman, S.M., Tucker, P.K., 1999. Molecular phylogeny and biogeography of the native rodents of Madagascar (Muridae: Nesomyinae): a test of the single-origin hypothesis. Cladistics 15, 253–270.

Krüger, Â., Razafimandimbison, S.G., Bremer, B., 2012. Molecular phylogeny of the tribe Danaideae (Rubiaceae: Rubioideae): another example of out-of-Madagascar dispersal. Taxon 61, 629–636.

Lanfear, R., Calcott, B., Ho, S.Y., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701.

Lu, L., Wortley, A.H., Li, D.Z., Wang, H., Blackmore, S., 2015. Evolution of angiosperm pollen. 2. The basal angiosperms. Ann. Missouri Bot. Gard. 100, 227–269.

Matzke, N.J., 2013. BioGeoBEARS: biogeography with Bayesian (and likelihood) evolutionary analysis in R scripts. R package version 0.2, 1.

Meseguer, A.S., Aldasoro, J.J., Sanmartín, I., 2013. Bayesian inference of phylogeny, morphology and range evolution reveals a complex evolutionary history in St. John's wort (Hypericum). Mol. Phylogenet. Evol. 67, 379–403.

Meve, U., Liede, S., 2002. Floristic exchange between mainland Africa and Madagascar: case studies in Apocynaceae-Asclepiadoideae. J. Biogeogr. 29, 865–873.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), 14 November 2010, New Orleans, pp. 1–8.

Morrison, D.A., 2010. Counting chickens before they hatch: reciprocal consistency of calibration points for estimating divergence dates. ArXiv e-prints. 1001.3586.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853–858.

Nee, S., 2006. Birth-death models in macroevolution. Annu. Rev. Ecol. Evol. Syst. 37, 1–17.

Nürk, N.M., Madriñán, S., Carine, M.A., Chase, M.W., Blattner, F.R., 2012. Molecular phylogenetics and morphological evolution of St. John's wort (*Hypericum*; Hypericaceae). Mol. Phylogenet. Evol. 66, 1–16.

Olmstead, R.G., Sweere, J.A., 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst. Biol. 43, 467–481.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Perrier de la Bâthie, H., 1951. Hypericaceae 135e Famille. Flore de Madagascar et des Comores. Typographie Firmin-Didot et Cie, Paris.

Planet, P.J., 2006. Tree disagreement: measuring and testing incongruence in phylogenies. J. Biomed. Inform. 39, 86–102.

Poumale, H.M.P., Krebs, H.C., Amadou, D., Shiono, Y., Komguem, N.A., Ngadjui, B.T., Randrianasolo, R., 2011. Flavonol glycosides from *Psorospermum androsaemifolium*. Chin. J. Chem. 29, 85–88.

Poumale, H.M.P., Randrianasolo, R., Rakotoarimanga, J.V., Raharisololalao, A., Krebs, H.C., Tchouankeu, J.C., Ngadjui, B.T., 2008. Flavonoid glycosides and other constituents of *Psorospermum androsaemifolium* Baker (Clusiaceae). Chem. Pharm. Bull. 56, 1428–1430.

R Core Team, 2013. R: a language and environment for statistical computing. 55, 275–286.

Rabinowitz, P.D., Coffin, M.F., Falvey, D., 1983. The separation of Madagascar from Africa. Science 220, 67–69.

Rambaut, A., Drummond, A.J., 2012. LogCombiner v1. 7.4. http://tree.bio.ed.ac.uk/software/.

Rambaut, A., Drummond, A.J., 2013a. FigTree. Program distributed by the author.

Rambaut, A., Drummond, A.J., 2013b. TreeAnnotator v1. 7.0. Available as part of the BEAST package at http://beast.bio.ed.ac.uk./software/.

Rambaut, A., Suchard, M.A., Xie, D., Drummond, A.J., 2014. Tracer v1.6. http://tree.bio.ed.ac.uk/software/.

Raxworthy, C.J., Forstner, M.R.J., Nussbaum, R.A., 2002. Chameleon radiation by oceanic dispersal. Nature 415, 784–787.

Razafimandimbison, S.G., Kellogg, E.A., Bremer, B., 2004. Recent origin and phylogenetic utility of divergent ITS putative pseudogenes: a case study from Naucleeae (Rubiaceae). Syst. Biol. 53, 177–192.

Reeves, C.V., de Wit, M.J., 2000. Making ends meet in Gondwana: retracing the transforms of the Indian Ocean and reconnecting shear zones. Terra Nova 12, 272–280.

Revell, L., 2012. Phytools: an R package for the phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3, 217–223.

Ronquist, F., Teslenko, M., Van der Mark, P., Ayres, P., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.

Ruhfel, B.R., Bittrich, V., Bove, C.P., Gustafsson, M.H.G., Philbrick, C.T., Rutishauser, R., Xi, Z., Davis, C.C., 2011. Phylogeny of the clusioid clade (Malpighiales): evidence from plastid and mitochondrial genomes. Am. J. Bot. 98, 306–325.

Ruhfel, B.R., Bove, C. P., Philbrick, C.T., Davis, C.C., 2016. Dispersal largely explains the Gondwanan distribution of the ancient tropical clusioid plant clade. Am. J. Bot 103, 1117–1128.

Ruhfel, B.R., Stevens, P.F., Davis, C.C., 2013. Combined morphological and molecular phylogeny of the clusioid clade (Malpighiales) and the placement of the ancient rosid macrofossil *Paleoclusia*. Int. J. Plant Sci. 174, 910–936.

Samonds, K.E., Godfrey, L.R., Ali, J.R., Goodmand, S.M., Vence, M., Sutherland, M.R., Irwing, M.T., Krause, D.W., 2012. Spatial and temporal arrival patterns of Madagascar's vertebrate fauna explained by distance, ocean currents, and ancestor type. Proc. Natl. Acad. Sci. U.S.A. 109, 5352–5357.

Sanderson, M.J., Doyle, J.J., 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy, and confidence. Syst. Biol. 41, 4–17.

Sanderson, M.J., Shaffer, H.B., 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. 33, 49–72.

Sang, T., 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 37, 121–147.

Schenk, J.J., 2016. Consequences of secondary calibrations on divergence time estimates. PloS ONE. 11, p.e0148228.

Shaul, S., Graur, D., 2002. Playing chicken (*Gallus gallus*): Methodological inconsistencies of molecular divergence date estimates due to secondary calibration points. Gene 300, 59–61.

Simpson, G.G., 1940. Mammals and land bridges. J. Wash. Acad. Sci. 30, 137–163.

Sloan, S., Jenkins, C.N., Joppa, L.N., Gaveau, D.L.A., Laurance, W.F., 2014. Remaining natural vegetation in the global biodiversity hotspots. Biol. Conserv. 117, 12–24.

Soltis, D.E., Soltis, P.S., Doyle, J.J., 1998. Molecular Systematics of Plants II: DNA Sequencing, Volume 2. Kluwer Academic Publishers, Norwell, Massachusetts.

Stankiewicz, J., Thiart, C., Masters, J.C., de Wit, M.J., 2006. Did lemurs have sweepstake tickets? An exploration of Simpson's model for the colonization of Madagascar by mammals. J. Biogeogr. 33, 221–235.

Stevens, P.F., 2001 onwards. Angiosperm Phylogeny Website. Version 12, July 2012 [and more or less continuously updated since]. http//www.mobot.org/MOBOT/research/APweb/.

Stevens, P.F., 2007. Hypericaceae. In: Kubitzki K. (Ed.), The Families and Genera of Vascular Plants. Flowering Plants. Eudicots: Berberidopsidales, Buxales, Crossosomatales, Fabales p.p., Geraniales, Gunnerales, Myrtales p.p., Proteales, Saxifragales, Vitales, Zygophyllales, Clusiaceae alliance, Passifloraceae alliance, Dilleniaceae, Huaceae, Picramniaceae, Sabiaceae. Springer-Verlag, Berlin, pp. 194–201.

Stöver, B.C., Müller, K.F., 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinform. 11, 7.

Swofford, D.L., 2002. PAUP*4.0 b10. Sinauer Associates, Sunderland, Massachusetts.

Taberlet, P., Gielly, L., Pautou, G., Bouvet, J., 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol. Biol. 17, 1105–1109.

Tamura K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.

Torsvik, T.H., Tucker, R.D., Ashwal, L.D., Eide, E.A., Rakotosolofo, N.A., de Wit, M.J., 1998. Late Cretaceous magmatism in Madagascar: paleomagnetic evidence for a stationary Marion hotspot. Earth Planet. Sci. Lett. 164, 221–232.

Wortley, A.H., Wang, H., Lu, L., Li, D.Z., Blackmore, S., 2015. Evolution of angiosperm pollen. 1. Introduction. Ann. Missouri Bot. Gard. 100, 177–226.

Yoder, A.D., Nowak, M.D., 2006. Has vicariance or dispersal been the predominant biogeographic force in Madagascar? Only time will tell. Annu. Rev. Ecol. Evol. Syst. 37, 405–431.

Yuan, Y.M., Wohlhauser, S., Moller, M., Klackenberg, J., Callmader, M.W., Kupfer, P., 2005. Phylogeny and biogeography of *Exacum* (Gentianaceae): a disjunctive distribution in the Indian Ocean basin resulted from long distance dispersal and extensive radiation. Syst. Biol. 54, 21–34.

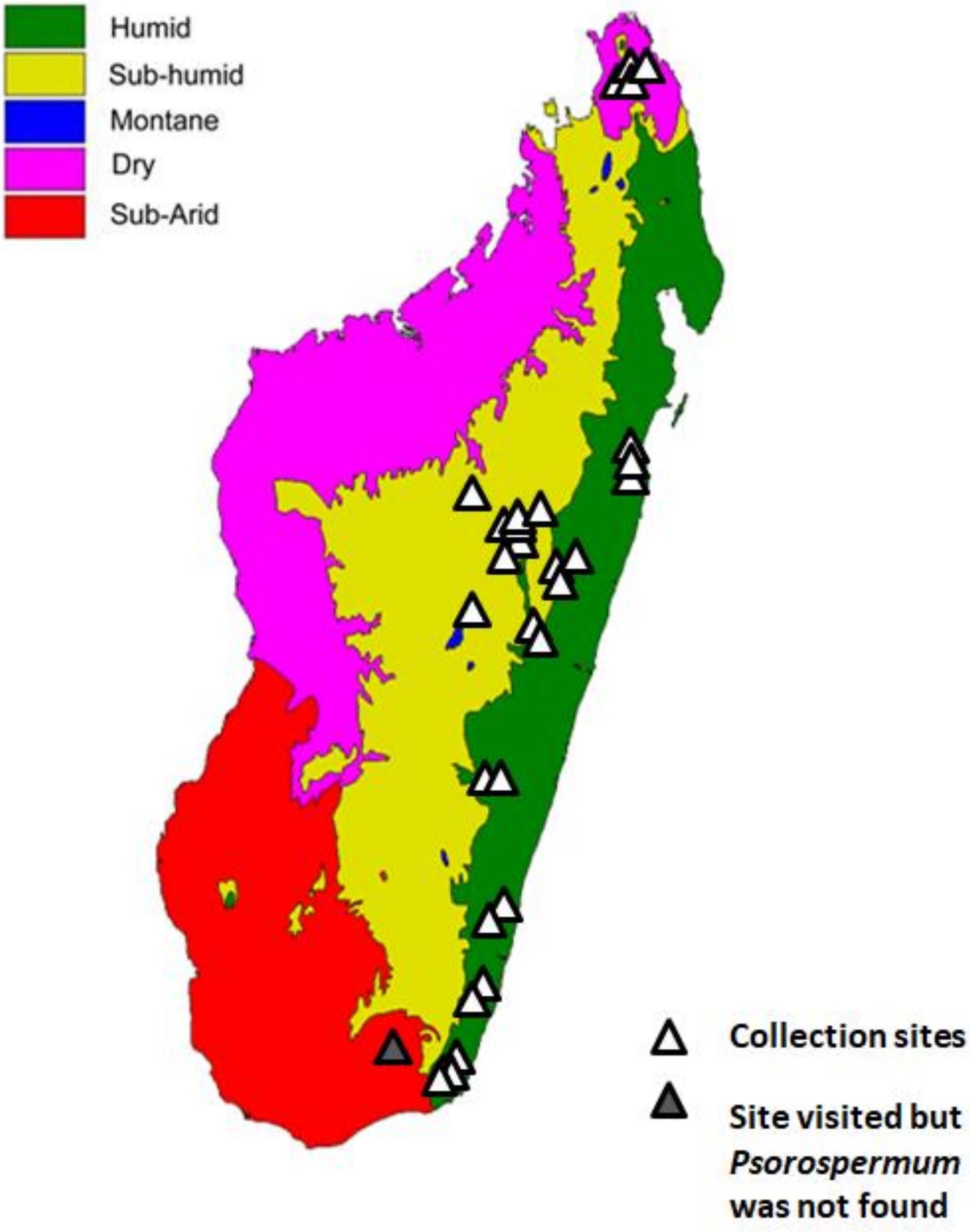
Zakharov, E.V., Caterino, M.S., Sperling, F.A., 2004. Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus *Papilio* (Lepidoptera: Papilionidae). Syst. Biol. 53, 193–215.

Zimmer, E.A., Wen, J.W., 2013. Using nuclear gene data for plant phylogenetics: progress and prospects. Mol. Phylogenet. Evol. 66, 539–550.

Zobayed, S.M.A., Afreen, F., Goto, E., Kozai, T., 2006. Plant–environment interactions: accumulation of hypericin in dark glands of *Hypericum perforatum*. Ann. Bot. 98, 793–804.

Supplementary materials

Figure. S1. Collection sites of *Psorospermum* in Madagascar.



Supplementary materials

Table S1

Statistics comparison of the markers tested

Markers	Sequences tested (n)	Total length (bp)	Variant characters (n)	Parsimony informative (n)	Parsimony informative (%)
EMB2765	10	882	69	1	0.11
psbA-trnH	10	329	26	12	3.65
trnS-trnG	8	709	49	19	2.68
trnL-trnH	10	667	97	37	5.55
ndhF	10	1166	154	65	5.57
ITS	10	742	75	32	4.31

TableS2

Primers used for EMB2765, *psbA-trnH* and *trnS-trnG*

Gene region	Primer name	Sequences	Original paper
trnH_psbA	trnH	ACT GCC TTG ATC CAC TTG GC	Hamilton et al 1999
	psbA	CGA AGC TCC ATC TAC AAA TGG	Hamilton et al 1999
trnS_trnG	trnS	GCC GCT TTA GTC CAC TCA GC	Hamilton et al 1999
	trnG	GAA CGA ATC ACA CTT TTA CCA C	Hamilton et al 1999
EMB2765	EMB2765ex9F2	TATCCAAATGAGCAGATTATGTGGGA	Wurdack and Davis (2009)
	EMB2765ex9R	TTGGTCCAYTGTGCWGCAGAAGGRT	Wurdack and Davis (2009)

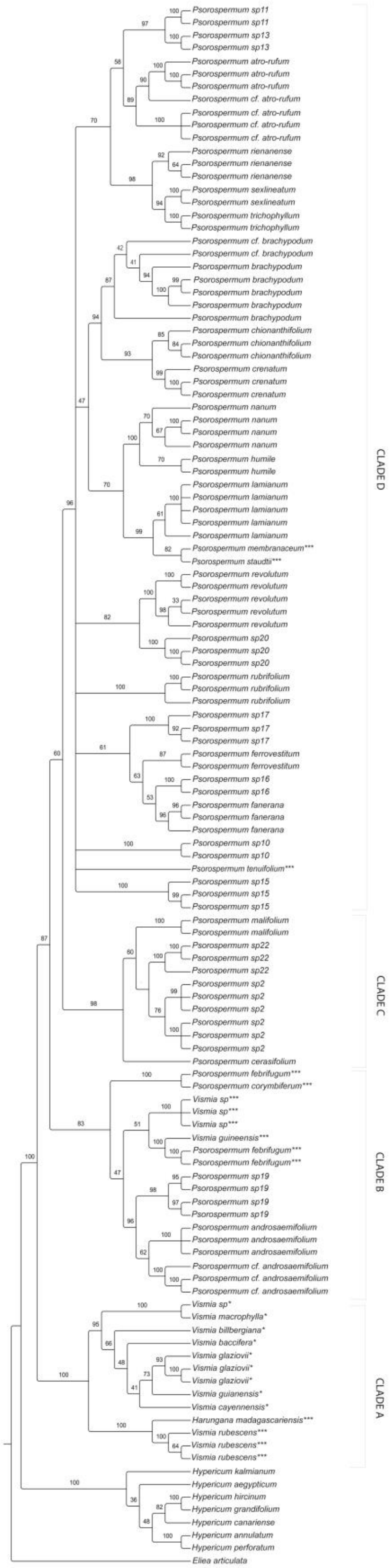


Figure S2. Maximum likelihood tree for *Psorospermum* based on the combined chloroplast DNA and internal transcribed spacer (ITS) dataset. Numbers above the branches are bootstrap support.*** = African species; *= American species.

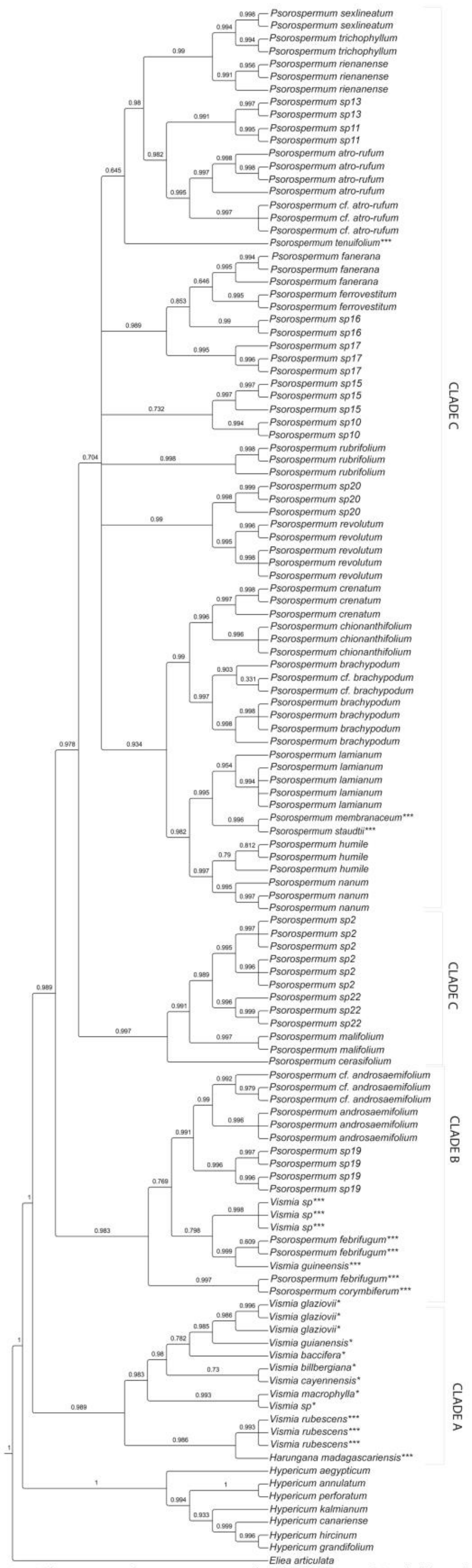

0.998 Psorospermum sexlineatum
0.994 Psorospermum sexlineatum
0.994 Psorospermum trichophyllum
0.99 Psorospermum trichophyllum
0.956 Psorospermum rienanense
0.991 Psorospermum rienanense
Psorospermum rienanense
0.98 0.997 Psorospermum sp13
0.991 Psorospermum sp13
0.995 Psorospermum sp11
Psorospermum sp11
0.982 0.998 Psorospermum atro-rufum
0.998 Psorospermum atro-rufum
0.645 0.997 Psorospermum atro-rufum
0.995 Psorospermum atro-rufum
0.997 Psorospermum cf. atro-rufum
Psorospermum cf. atro-rufum
Psorospermum cf. atro-rufum
Psorospermum tenuifolium***
0.994 Psorospermum fanerana
0.995 Psorospermum fanerana
0.646 Psorospermum fanerana
0.995 Psorospermum ferrovestitum
0.853 Psorospermum ferrovestitum
0.989 0.99 Psorospermum sp16
Psorospermum sp16
0.995 Psorospermum sp17
0.996 Psorospermum sp17
Psorospermum sp17
0.997 Psorospermum sp15
0.997 Psorospermum sp15
0.732 Psorospermum sp15
0.994 Psorospermum sp10
Psorospermum sp10
0.704 0.998 0.998 Psorospermum rubrifolium
Psorospermum rubrifolium
Psorospermum rubrifolium
0.999 Psorospermum sp20
0.998 Psorospermum sp20
Psorospermum sp20
0.99 0.996 Psorospermum revolutum
Psorospermum revolutum
0.995 Psorospermum revolutum
0.998 Psorospermum revolutum
Psorospermum revolutum
0.998 Psorospermum crenatum
0.997 Psorospermum crenatum
0.996 Psorospermum crenatum
0.996 Psorospermum chionanthifolium
Psorospermum chionanthifolium
Psorospermum chionanthifolium
0.99 0.903 Psorospermum brachypodum
0.331 Psorospermum cf. brachypodum
Psorospermum cf. brachypodum
0.997 0.998 Psorospermum brachypodum
0.998 Psorospermum brachypodum
Psorospermum brachypodum
Psorospermum brachypodum
0.978 0.934 Psorospermum lamianum
0.954 Psorospermum lamianum
0.994 Psorospermum lamianum
0.995 Psorospermum lamianum
Psorospermum lamianum
0.996 Psorospermum membranaceum***
0.982 Psorospermum staudtii***
0.812 Psorospermum humile
0.79 Psorospermum humile
0.997 Psorospermum humile
0.995 Psorospermum nanum
0.997 Psorospermum nanum
Psorospermum nanum
CLADE C
Psorospermum sp2
0.997 Psorospermum sp2
0.995 Psorospermum sp2
0.996 Psorospermum sp2
Psorospermum sp2
0.989 Psorospermum sp2
0.989 0.996 Psorospermum sp22
0.991 0.999 Psorospermum sp22
Psorospermum sp22
0.997 0.997 Psorospermum malifolium
Psorospermum malifolium
Psorospermum cerasifolium
CLADE C
0.992 Psorospermum cf. androsaemifolium
0.979 Psorospermum cf. androsaemifolium
0.99 Psorospermum cf. androsaemifolium
0.996 Psorospermum androsaemifolium
0.991 Psorospermum androsaemifolium
Psorospermum androsaemifolium
0.997 Psorospermum sp19
0.996 Psorospermum sp19
0.996 Psorospermum sp19
Psorospermum sp19
1 0.769 Vismia sp***
0.998 Vismia sp***
Vismia sp***
0.983 0.798 0.609 Psorospermum febrifugum***
0.999 Psorospermum febrifugum***
Vismia guineensis***
0.997 Psorospermum febrifugum***
Psorospermum corymbiferum***
CLADE B
0.996 Vismia glaziovii*
0.986 Vismia glaziovii*
0.985 Vismia glaziovii*
0.782 Vismia guianensis*
1 0.98 Vismia baccifera*
0.73 Vismia billbergiana*
0.983 Vismia cayennensis*
0.993 Vismia macrophylla*
Vismia sp*
0.989 Vismia rubescens***
0.993 Vismia rubescens***
0.986 Vismia rubescens***
1 Harungana madagascariensis***
CLADE A
Hypericum aegypticum
1 1 Hypericum annulatum
0.994 Hypericum perforatum
Hypericum kalmianum
0.933 Hypericum canariense
0.999 0.996 Hypericum hircinum
Hypericum grandifolium
Eliea articulata

Figure S3. Baysian inference tree for *Psorospermum* based on the combined chloroplast DNA and internal transcribed spacer (ITS) dataset. Numbers above the branches are bootstrap support.*** = African species; *= American species.

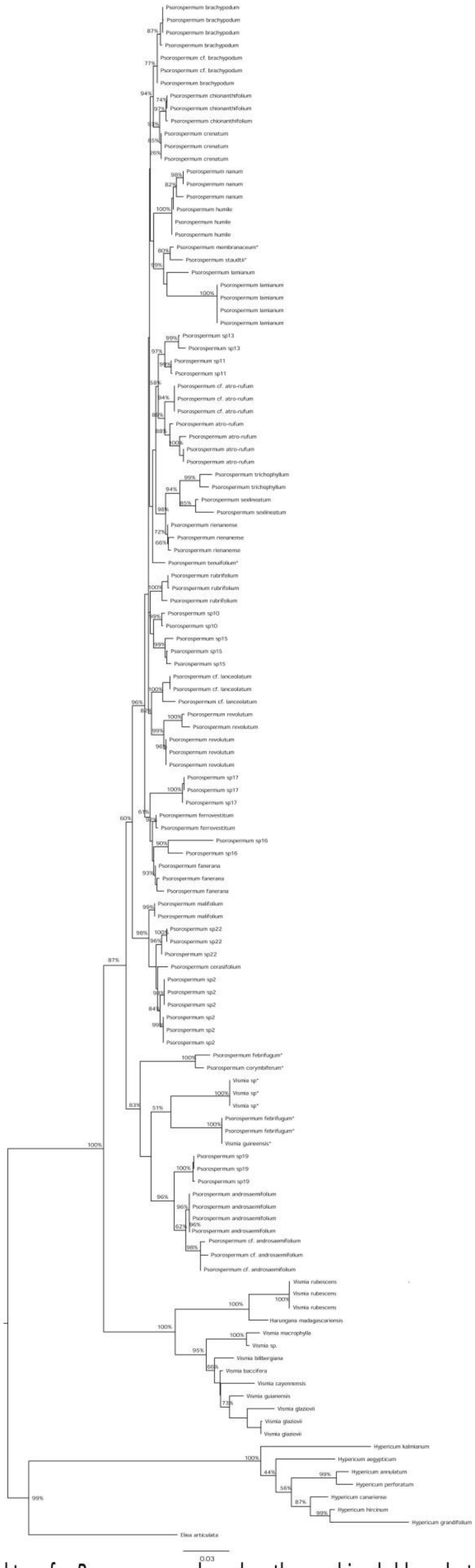


Figure S4. Maximum likelihood tree for *Psorospermum* based on the combined chloroplast DNA and internal transcribed spacer (ITS) dataset. Numbers above the branches are bootstrap support. African species are marked with an asterisk.

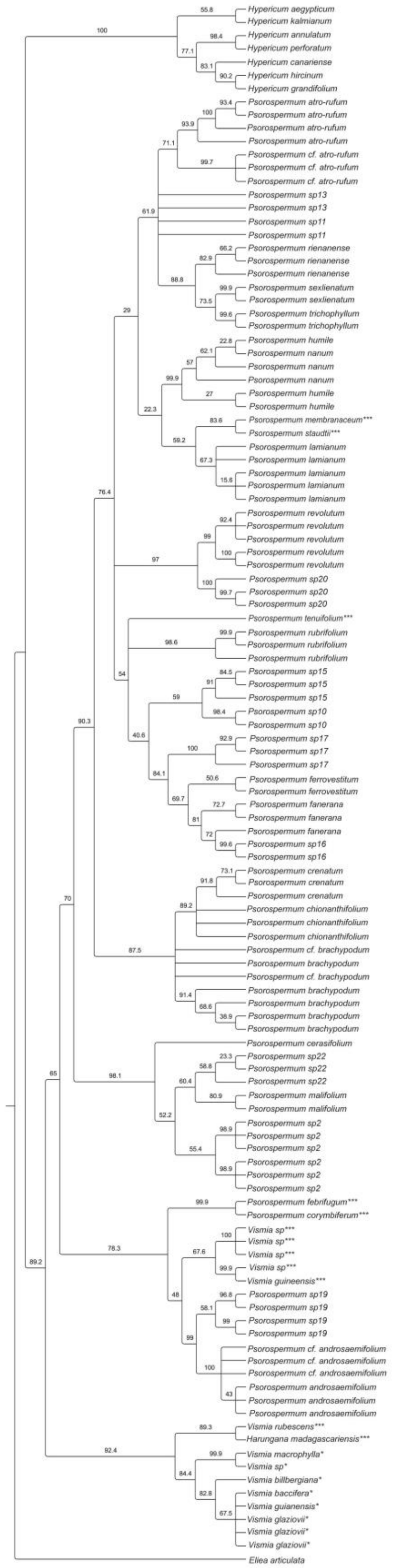


Figure S5. Maximum likelihood tree for *Psorospermum* based on the chloroplast dataset (*ndhF+trnL-trnF*). Numbers above the branches are bootstrap support.*=American; ***=African.

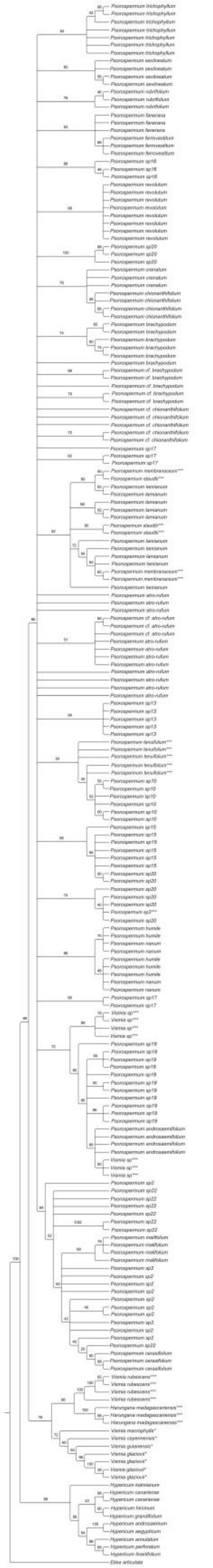


Figure S6. Maximum likelihood tree for *Psorospermum* based on the ITS gene. Numbers
. above the branches are bootstrap support.*=American; ***=African.

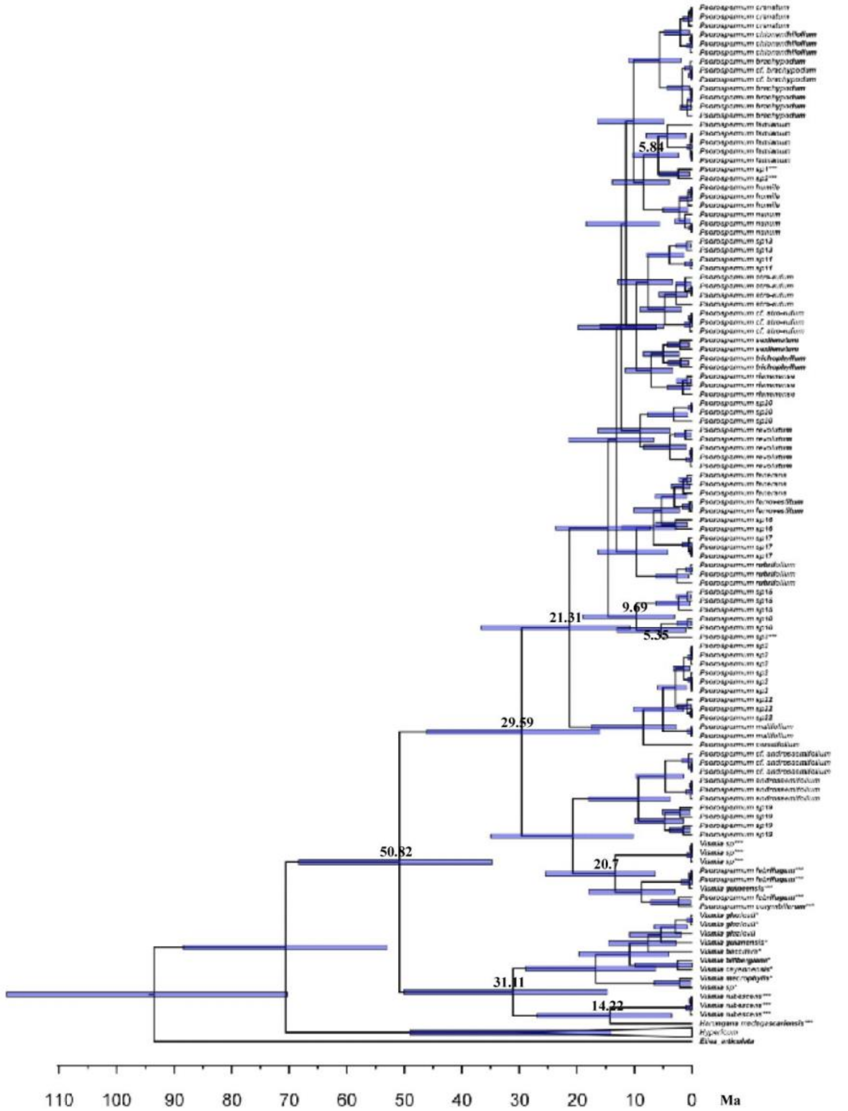


Figure S7. Time-calibrated phylogeny of Vismieae using a normal probability prior of the BC analysis. Bars represent the Highest Posterior Density (HPD) intervals of the dating analysis. Numbers represent the node ages. *** = African species; *= American species.

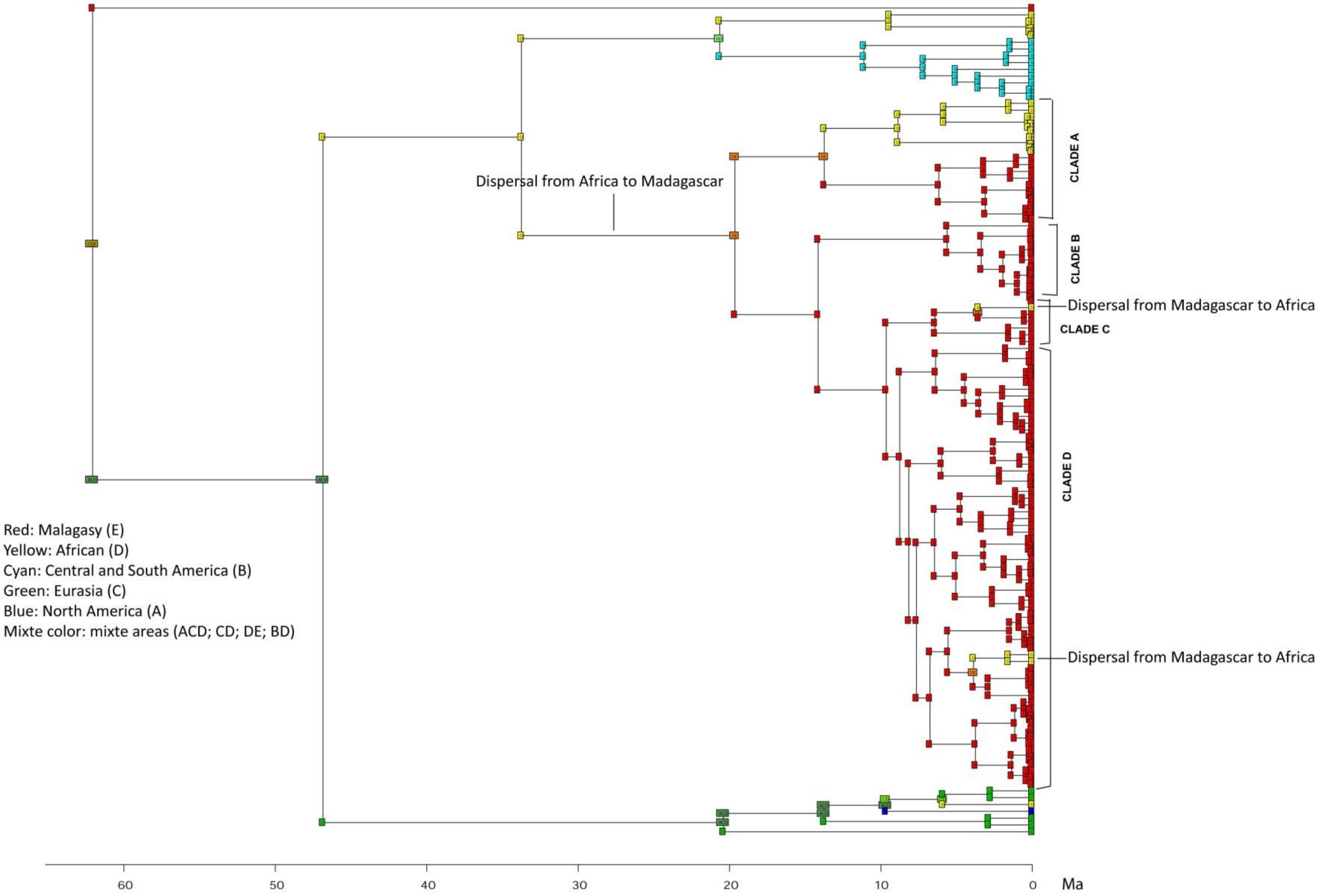

Dispersal from Africa to Madagascar
CLADE A
CLADE B
Dispersal from Madagascar to Africa
CLADE C
CLADE D
Dispersal from Madagascar to Africa
Red: Malagasy (E)
Yellow: African (D)
Cyan: Central and South America (B)
Green: Eurasia (C)
Blue: North America (A)
Mixte color: mixte areas (ACD; CD; DE; BD)
60
50
40
30
20
10
0
Ma

Figure S8. Ancestral reconstruction and biogeography inference of the Malagasy *Psorospermum* based on Dispersal-Caldogenesis Extinction model (DEC). Areas are indicated by colored boxes with letter.

Integrative taxonomy: investigating species boundaries in Malagasy *Psorospermum* using morphometrics and molecular phylogenetic methods.

HERITIANA RANARIVELO[1,2]

[1] *Department of Biology, University of Missouri – St. Louis, One University Blvd, St. Louis, MO 63121-4000, USA*
[2] *Missouri Botanical Garden, PO Box 299, St. Louis, MO 63166-0299, USA*

I investigated species hypotheses for *Psorospermum* species in Madagascar by integrating morphological species hypotheses with clustering sequences in molecular phylogenetic data. The molecular phylogenetics of *Psorospermum* based on chloroplast and nuclear DNA shows well supported clades within Malagasy *Psorospermum*. Using multivariate analysis based on Gaussian mixture models (GMM) implemented in Mclust R packages, I investigated the morphospecies within nine well supported clades in Malagasy *Psorospermum*, and compared them with the molecular clusters; individuals for each taxon tip of the clades were measured for floral and vegetative characters. GMM identifies 30 morphospecies in Malagasy *Psorospermum*. Two clades out of the 9 match the GMM morphospecies; four clades have more morphogroups than the GMM morphospecies while 2 clades have fewer morphogroups than GMM morphospecies. I recognize 27 species. My results support the hypothesis that incongruence between different species delimitation methods is explained by the particular stage of the speciation process; in this study molecular analyses can detect divergence that has not been yet fully translated into morphological characters, and were not recovered by the GMM analyses.

ADDITIONAL KEYWORDS: *Psorospermum* – species delimitation – morphometrics – Gaussian mixture models – Madagascar.

INTRODUCTION

The flora of Madagascar is unique for its endemism and diversity (Goodman & Benstead, 2005; Behrens & Branes, 2016). For almost two decades, taxonomists around the world have carried out major efforts to revise the Malagasy flora, a flora that had been largely unstudied for the previous 60 years. The discovery, description and naming of new species are still ongoing and involve considerable human and financial resources.

The main objective is to be able to protect and conserve this diversity, not only to prevent its extinction and understand its evolution but to discover its value for human beings. Thus, it is important to have confidence in species boundaries. One way to investigate species limits is to use an integrative taxonomic approach, *i.e.* evidence of morphospecies from one kind of data are hypotheses to be tested against other kinds of data (Yeates *et al.,* 2011). Since it was introduced in 2005 (Dayrat, 2005; Will *et al.*, 2005; Pante *et al.*, 2014), integrative taxonomy has been applied to delimit species boundaries in several groups of insects and animals (Tan *et al.,* 2010; Gullan *et al.*, 2010; Lumley & Sperling, 2010; Blaimer, 2012; Aguilar *et al.,* 2016). A similar approach has been applied in plants, *e.g.* Zapata and Jiménez (2012) investigated the species hypotheses in *Escallonia* Mutis (Escalloniaceae) by integrating morphological differences with geography, and Hong-Wa and Besnard (2014) included evidence from molecular phylogeny, geography and ecology to support hypotheses of species limits in the Malagasy olive genus *Noronhia* Stadtman. To my knowledge, the latter is the only study of Malagasy plants that has used an integrative taxonomic approach in species delimitation, although this approach has been broadly applied in studies of Malagasy animals, *e.g.* in ants (Blaimer, 2012); spiders (Gregorič *et al.*, 2015); ridged frogs (Zimkus *et al.*, 2016); cat-eyed snakes (Ruane et *al.*, 2016); and mouse lemurs (Hotaling *et al.*, 2016).

Here I investigate species limits in the Malagasy *Psorospermum* (Hypericaceae), a poorly known genus that grows in Africa and Madagascar, with maybe five species in Africa and 26 species in Madagascar (Perrier de la Bâthie, 1951), but the total number is uncertain. The results of this study will be directly applied to the taxonomic revision of the Malagasy species (Ranarivelo *et al.,* in preparation). Integrative taxonomic studies include methods used by different biologists *e.g.* taxonomists and molecular systematists, however, the choice of methods can be limited by reasons other than theoretical. For example, the choice can depend on the study group itself. While research in both animal and plant species delimitation often uses geographical and morphological evidence (*e.g.* Blaimer, 2012; Rato et al., 2016), species delimitation of animals tends to combine those two lines of evidence with analyses of population genetics, while plant studies often integrate such evidence with molecular phylogenetic analyses (*e.g.* Barrett *et al.,* 2011; Hong-Wa & Besnard, 2014). It also depends on the time available, while materials and infrastructure can also limit the application of multiple lines of evidence, especially when, for example, the acquisition of appropriate materials is impossible because of the political situation. A reliable source of morphological data in plant studies is herbarium collections. Although extracting good quality DNA from herbarium specimens can be difficult, voucher specimens of collected fresh materials for DNA extraction can be used to link molecular and morphological studies (see methods). Thus, at the very least, morphometrics and molecular phylogenetics must be integrated in plant species delimitation. In this study I am investigating a practical way to integrate the use of the herbarium collections with molecular techniques, and my main goal is to delimit species in Malagasy *Psorospermum*.

METHODS

Integrative taxonomy approach

My approach is based on the concept that species are the results of evolutionary process so that they can be perceived as parts of lineages (de Queiroz, 1998, 2007). Thus, I use supported clades from the molecular phylogeny of the Malagasy *Psorospermum* (Ranarivelo *et al.*, in preparation) as a frame or universe for subsequent analyses (see discussion), where herbarium specimens of the putative taxa in each clade are subject to morphometric analyses. The molecular phylogeny found a total of nine clades, and for practical use, I named each of them after one of the morphogroups it included. Fig. 1 shows a reduced-taxon phylogeny where monophyletic putatively conspecific specimens are collapsed into a single taxon name, while the maximum likelihood (ML) tree with all samples is provided in Appendix A. Nine steps enabled me to proceed from the initial recognition of morphogroups to the final delimitation of species, including the use of taxonomy, phylogeny and morphometrics as presented in the workflow below (Fig. 2):

(1) Herbarium specimens of Malagasy *Psorospermum* from MO, TAN and loans from G, K and P (abbreviations follow *Index Herbariorum*, Thiers [continuously updated]. http://sweetgum.nybg.org/science/ih/.) are placed into morphogroups based on gross overall similarity. (2) A molecular phylogenetic tree of *Psorospermum* is generated (Ranarivelo *et al.* in preparation). (3) Well-supported clades delimited in the molecular phylogeny are chosen for further detailed analyses. (4) For each clade, specimens used in the molecular phylogeny are added to the appropriate morphogroups recognized in step 1.

(5) Morphometric analyses: 29 foliar, floral and embryo characters are measured and analyzed (Appendix Table S1); preliminary morphogroups are tested with Linear Discriminant Analysis (LDA), significant characters are selected using Principal Component Analysis (PCA) (see results), and morphospecies are identified using Gaussian mixture models (GMM). Data are deposited in the Dryad Digital Repository (http//dx.doi.org/xx.xxxx/drayd.xxxxx). (For the sake of clarity, morphospecies refer to the outputs of the GMM analyses; the initial groupings of herbarium specimens that I recognized are morphogroups.)

To be able to link the identified morphospecies to the morphological characters used in the GMM analyses, three additional steps were needed. (6) Quantification of the morphological discontinuity (or "gap") between pairs of morphospecies apparent in the GMM analyses using methods described by Zapata & Jiménez (2012). (7) Assessment of the morphological character differences between pairs of morphospecies.(8) Morphogroups of individuals making up a clade were compared to the morphospecies obtained with the GMM analyses. (9) The final step is the recognition of species.

Selection of morphological characters

Characters to be used in analyzing the variation within strongly-supported clades were selected using PCA. Since the variables are projected onto two-dimensional PCA scatterplot, the cosine square values ($\cos^2$) of these characters values ranging from 0 to 1 can be used to select the variables according to the quality of their representation onto the PCA scatterplot (Abdi and Williams 2010); in other words low $\cos^2$ values are more subject to errors due to projection effects than those with high $\cos^2$ values (Appendix B),. I averaged the $\cos^2$ values of the 10 variables with highest $\cos^2$ values. I chose 10 as the number of variables to be averaged so that all the significant variables can be included. However, I retained only the variables with values higher than the calculated average. The absence of collinearity of these variables was also verified since the EM algorithm implemented in GMM performs poorly with collinear variables (Fraley & Raftery, 1998). The analysis was performed with the R packages MASS v7.3-45; FactoMineR v1.35 and caret v6.0-73.

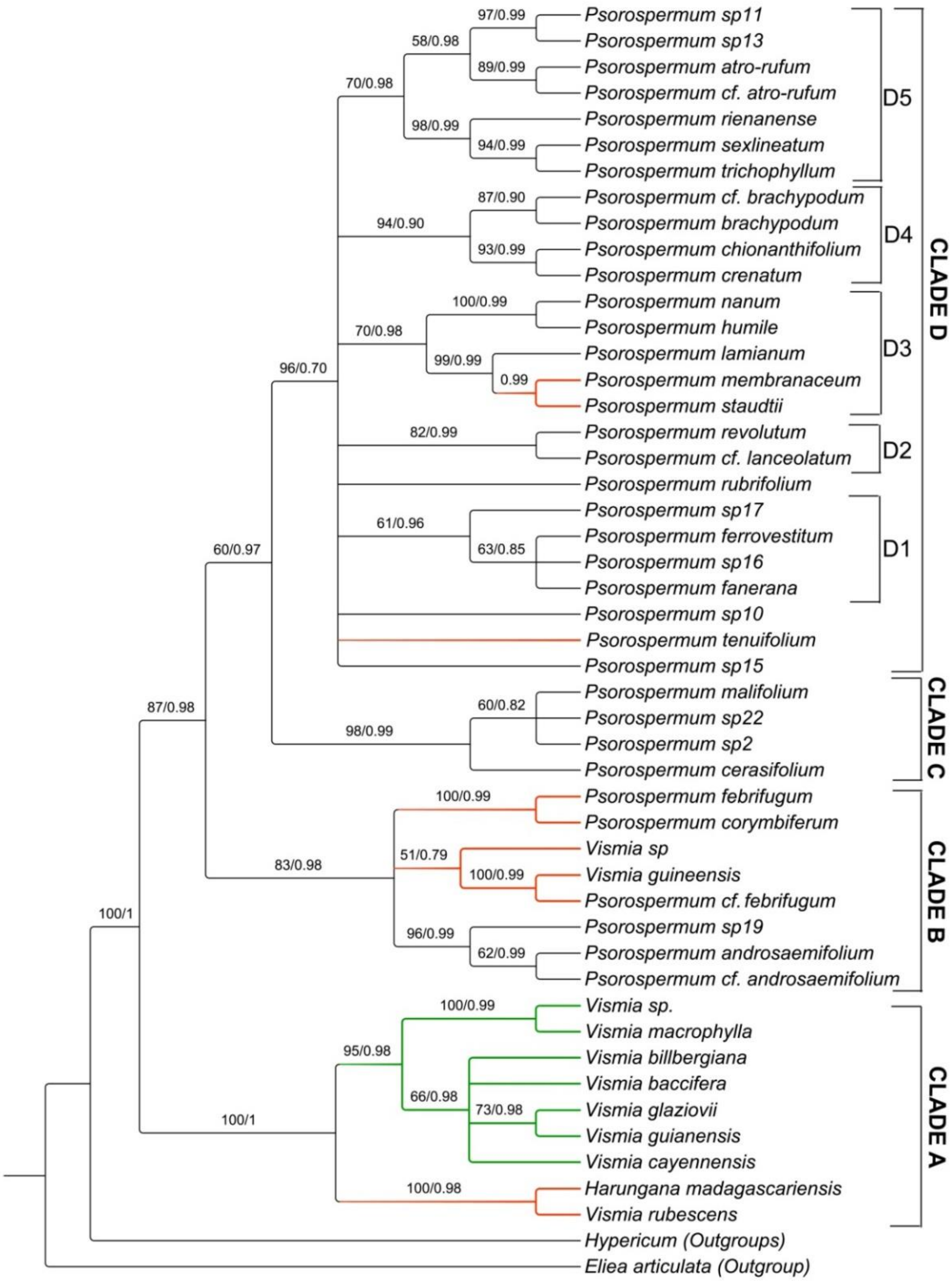


Figure 1. Maximum Likelihood tree (ML) tree inferred from analyses using combined chloroplast DNA regions (*trnL-F*, *ndhF*) and nuclear ribosomal DNA (ITS). Values above branches on the left denote maximum likelihood bootstrap support (BT %) and those on the right are Bayesian posterior probabilities (PP). In the ingroup, green branches indicate American taxa; dark orange: African; and black: Malagasy.

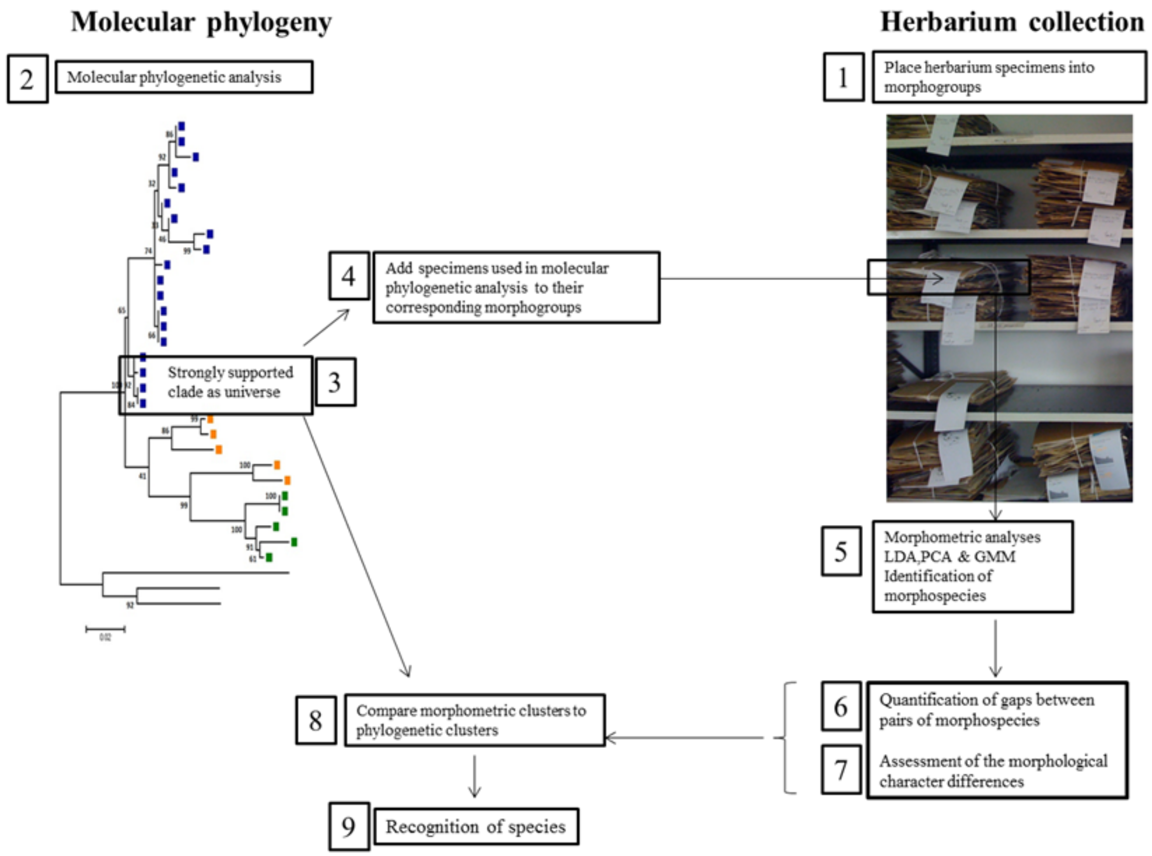

Molecular phylogeny
Herbarium collection
1
Place herbarium specimens into morphogroups
2
Molecular phylogenetic analysis
3
Strongly supported clade as universe
4
Add specimens used in molecular phylogenetic analysis to their corresponding morphogroups
5
Morphometric analyses LDA,PCA & GMM Identification of morphospecies
6
Quantification of gaps between pairs of morphospecies
7
Assessment of the morphological character differences
8
Compare morphometric clusters to phylogenetic clusters
9
Recognition of species
0.02

Figure 2. Flowchart of integrative taxonomy combining taxonomy, phylogeny and morphometrics.

Identification of morphological morphospecies

I chose GMM instead of the classical linear transformation methods such as LDA and PCA because it is a probabilistic method that can identify putative morphospecies in the dataset without any arbitrary decision from the users or any cross validation by additional discriminant analysis methods (Appendix B). (1) GMM fit the data set to an *n*-morphospecies Gaussian mixture distribution with Bayesian rules; (2) the probability density for the n-morphospecies mixture distribution is estimated with different models so that each model delimits morphospecies separately; (3) the Bayesian Information Criterion (BIC) of the models are compared, the best model has the highest BIC values and indicates the number of morphospecies included in it, and (4) morphospecies are displayed in a two-dimensional scatterplot for each pair of morphological variables used in the analysis, one variable in the x-axis and the other one in the y-axis.

GMM has been automated by the computing program packages Mclust version 5.2.2 in R (Fraley et al., 2017). Mclust has been used in GMM species delimitation in animal studies (Hausdorf and Hennig 2010, Ezard et al., 2010, Carstens and Satler 2013, Edwards and Knowles 2014; Aguilar et al., 2016, Eberle et al., 2016), but to my knowledge they it has never been used in any plant studies.

Quantifying gaps between morphospecies and assessment of the contribution of the morphological characters in separating the morphospecies

I applied the method based on the ridgeline manifold (RM) as described by Zapata and Jiménez (2012) to quantify the morphological discontinuity (gap) between pairs of morphospecies. By definition, the RM is a line in morphospace that connects the mode of one morphospecies to the mode of another (Fig. 3A), and the coordinates along that line can be calculated using the variable alpha (α), the sequence of values along the RM ranging from 0 at the mode of one group to 1 at the mode of the other (Fig. 3A) (Ray & Lindsay, 2005, Zapata & Jiménez, 2012). Consequently, α can be used to calculate the slope of the probability density function (PDF). The pattern of the slope is shown in a two-dimensional graph plot where the slope's graph shows the gap differing between the two morphospecies, the PDF represents the y-axis while α values represent the x-axis (Fig. 3B).

I also evaluated the proportion of individuals with non-overlapping phenotypes in pairs of morphospecies by calculating beta star, β*. Since each morphospecies corresponds to multivariate distributions of morphological characters, those distributions can be projected as ellipsoids in morphospace (Fig. 4A). Thus for two morphospecies, A and B, for example, the ellipsoids of species A (i.e., the proportion of phenotypes of species A that do not overlap with species B) define the ellipsoids of species B and vice versa (Fig. 4A). In theory, at a certain point the ellipsoids of both species should be equal and non-overlapping, but sharing a single point on the graph. Thus the value of that single point, β*, can be calculated using the values of the proportion of individuals within the elliptic regions of each of the two species, β (Figs. 4A and 4B). If β* is above a β value of 0.9 (90% of the phenotypes that do not overlap in the two species), this suggests strongly that the two morphospecies, with very largely discontinuous variation, are distinct. If β* is below 0.9, there is appreciable overlap of phenotypes, questioning whether the two morphospecies are indeed distinct (Fig. 4B).

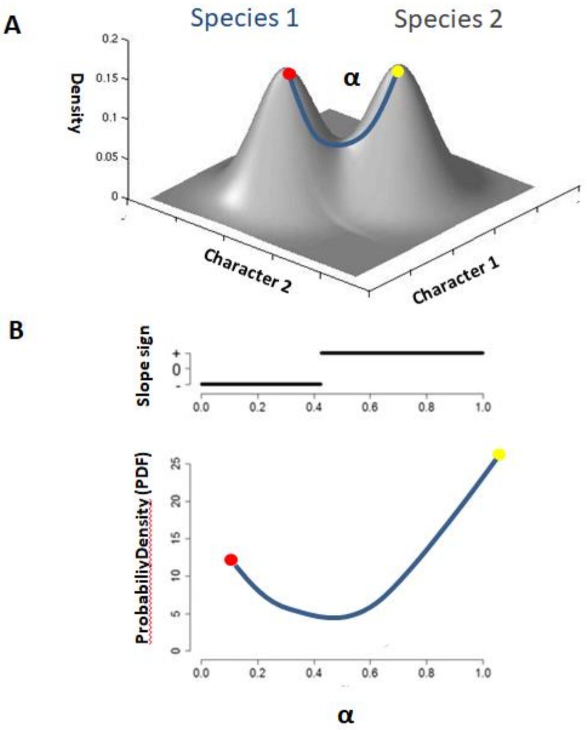


Figure 3. A. The ridgeline manifold (RM), blue line, connecting the mode of one morphospecies to the mode of another (red and yellow dots). **B.** The slope of the probability density function (PDF), blue line. Black lines indicates the change of slope sign along the PDF.

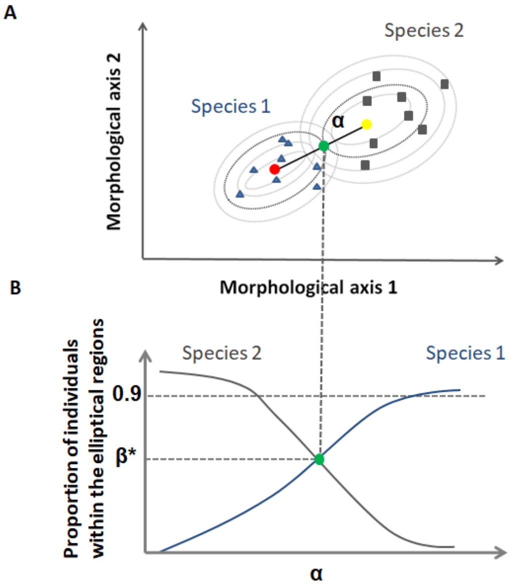


Figure 4. A. Ellipsoids of the proportion phenotypes of two morphospecies A and B, black line: RM with α values; red point: mode of species A, yellow point: mode of species B, green point: meeting point of equal ellipsoids, black filled square: individuals of morphospecies B, blue triangle: individuals of morphospecies A, **B.** Example showing that the proportion of individulas with non-overlapping phenotypes, β*, is below the cut-off 0.9. In this example the two morphospecies are not distinct.

Additionally I conducted simulation comparisons: the quartiles of the simulated univariate distribution (ST) of each character of the two species were compared, and they are represented as boxplots in the graph outputs (Fig. 5). The simulation iteration is set as N=100000. I consider that the ST of two morphospecies overlap when any of their first, second, or the third quartiles overlap (Fig. 5). Then, I plotted the quartiles of the univariate distribution (UD) of the characters from the GMM results next to the UD of ST of the two morphospecies. The null hypothesis is that the quartiles of UD do not deviate from the distribution of ST (Fig. 5). All analyses for the quantification of gaps between morphospecies and the assessment of the contribution of particular morphological characters in separating the morphospecies were conducted using the R package mvtnorm v1.0-5 in R (Genz et al., 2016).

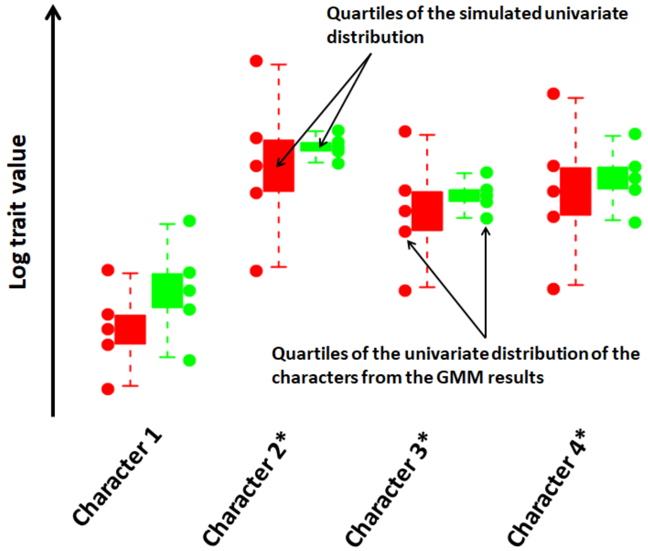

Quartiles of the simulated univariate distribution
Log trait value
Quartiles of the univariate distribution of the characters from the GMM results
Character 1
Character 2*
Character 3*
Character 4*

Figure 5. Example illustrating the details of the simulation comparisons of the overlapping characters in two morphospecies. Characters that overlap are marked with asterisks. Red: morphospecies A, green: species B.

Species recognition

This study is based on the concept that species are the results of evolutionary process (de Queiroz, 1998, 2007), the analyses are conducted within each well-supported clade where morphological characters were measured from and compared between specimens of taxa belonging to that clade. It is in this way that morphological and molecular data are integrated. Moreover, results from GMM analyses provide information of morphological differences that can distinguish morphospecies within each clade. I recognized a species if it meets the following criteria: they are units (morphospecies) identified by GMM that have evidence of a gap between them, with the proportion of individuals that have non-overlapping phenotypes, β*, above a cutoff of 0.9, and that correspond to well-supported subclades i.e. the ultimate monophyletic units within the clade.

RESULTS

Overall, a total of 314 specimens of Malagasy *Psorospermum* were used in the analyses, including 81 specimens used in the molecular phylogenetic analysis, and 29 morphological characters were measured (see Appendix Table S1). A total of 30 morphospecies are identified by GMM analysis. Below I provide the results of the analysis of the Cerasifolium clade in more detail than the rest to show how the process works, while the results for the other eight clades are more briefly summarized. The full results for all clades, including figures, are provided in the supporting information, and the summary of the results of all molecular phylogenetic clades is shown in Table 1.

Clade Cerasifolium

Clade Cerasifolium has 98% bootstrap support, and contains 6 subclades, i.e. the ultimate monophyletic units within the clade (Fig. 6).

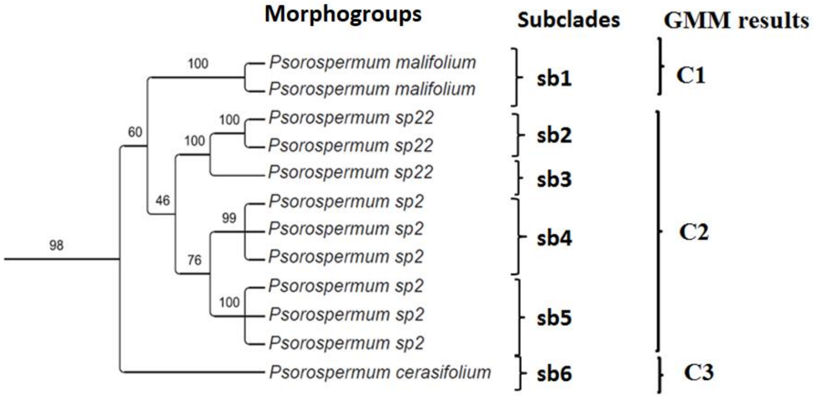
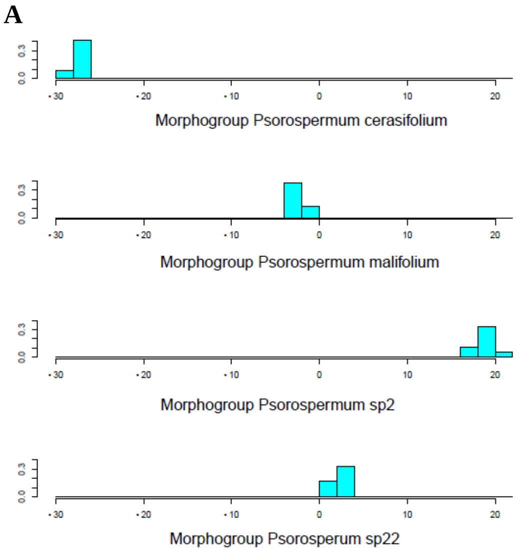


Figure 6. Expanded clade Cerasifolium retrieved from ML molecular phylogenetic of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclades; C1, C2 and C3 are the morphospecies identified by GMM.

The 12 specimens from the molecular study are assignable to four morphogroups (*Psorospermum malifolium*, *P. sp22*, *P. sp2*, and *P. cerasifolium*) (Fig. 6). These specimens were added to 17 herbarium specimens previously assigned to these 4 morphogroups for a total of 29 herbarium specimens. The four morphogroups were confirmed with LDA (Wilks's Lambda = 0.00004, p-value < 0.0001). There is no overlap of the histograms of values of the discriminant functions for the samples from the four morphogroups (Fig. 7A). Nineteen floral and foliar characters were measured (Appendix Table S1). The eight characters that significantly influence the loadings of the observations in the PCA vector scatterplot are: lamina area (size), angle of apex, and number of glands per cm^2; pedicel length; sepal length, and number of glands; filament length; and style length (Fig. 7B). These variables are used to perform the GMM analyses.

GMM analysis identified three morphospecies in this clade (Fig. 8). Fig. 8 summarizes the output of the Mclust models (table on the left), and highlights the identification of three morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in the legend box are given in Appendix Table S2); the best model is the one that has the highest BIC value and is indicated by the horizontal dashed line on the y-axis (Fig. 8). The number of morphospecies suggested by this model is indicated by the vertical dashed line on the x-axis (Fig. 8). The three morphospecies delimited by GMM are shown in two-dimensional (2-D) morphospace classification scatterplot in Fig. 9. There are eight variables in the dataset for this particular clade, and the combinations of pairs of variables are shown in 2-D scatterplots in Fig. 10.



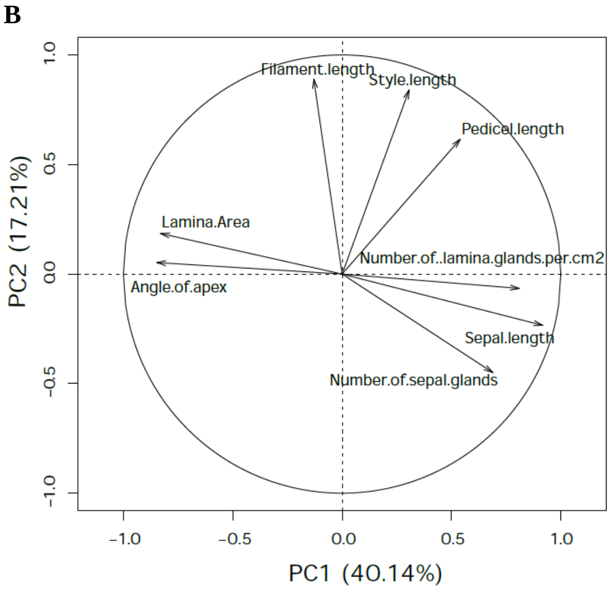


Figure 7A. Histogram of values of the first discriminant function for the samples from the 4 morphogroups.

Figure 7B. Eight morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in clade Cerasifolium. Percentage of variation explained is in parentheses.

Figure 8. Outputs of GMM analysis of the Cerasifolium clade.

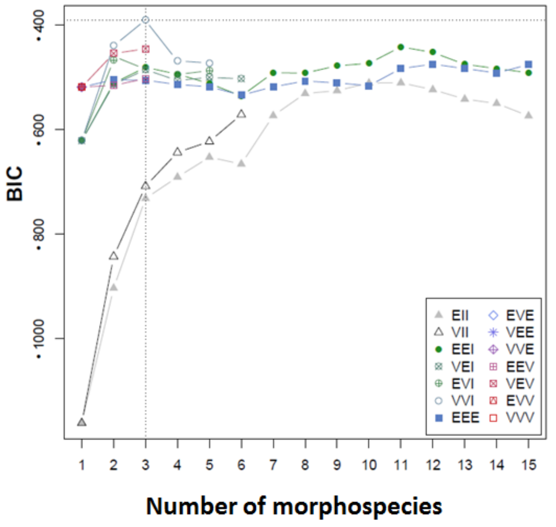
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N.ind.: number of individuals per cluster.

Mclust model with 3 morphospecies:

VVI (Diagonal, varying volume and shape)
See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
-115.85	29	48	-393.552

Clusters	1	2	3
N . ind.	8	15	6



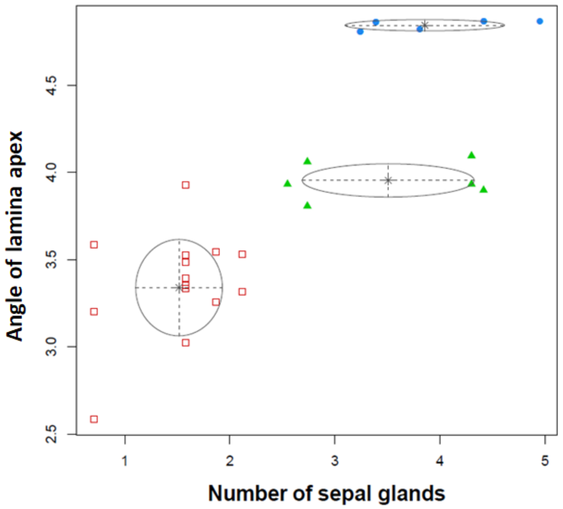


Figure 9. Bivariate scatterplot in clade Cerasifolium. The ellipses superimposed on the plot correspond to the covariances of the clusters. Filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3.

Quantification of gaps between morphospecies

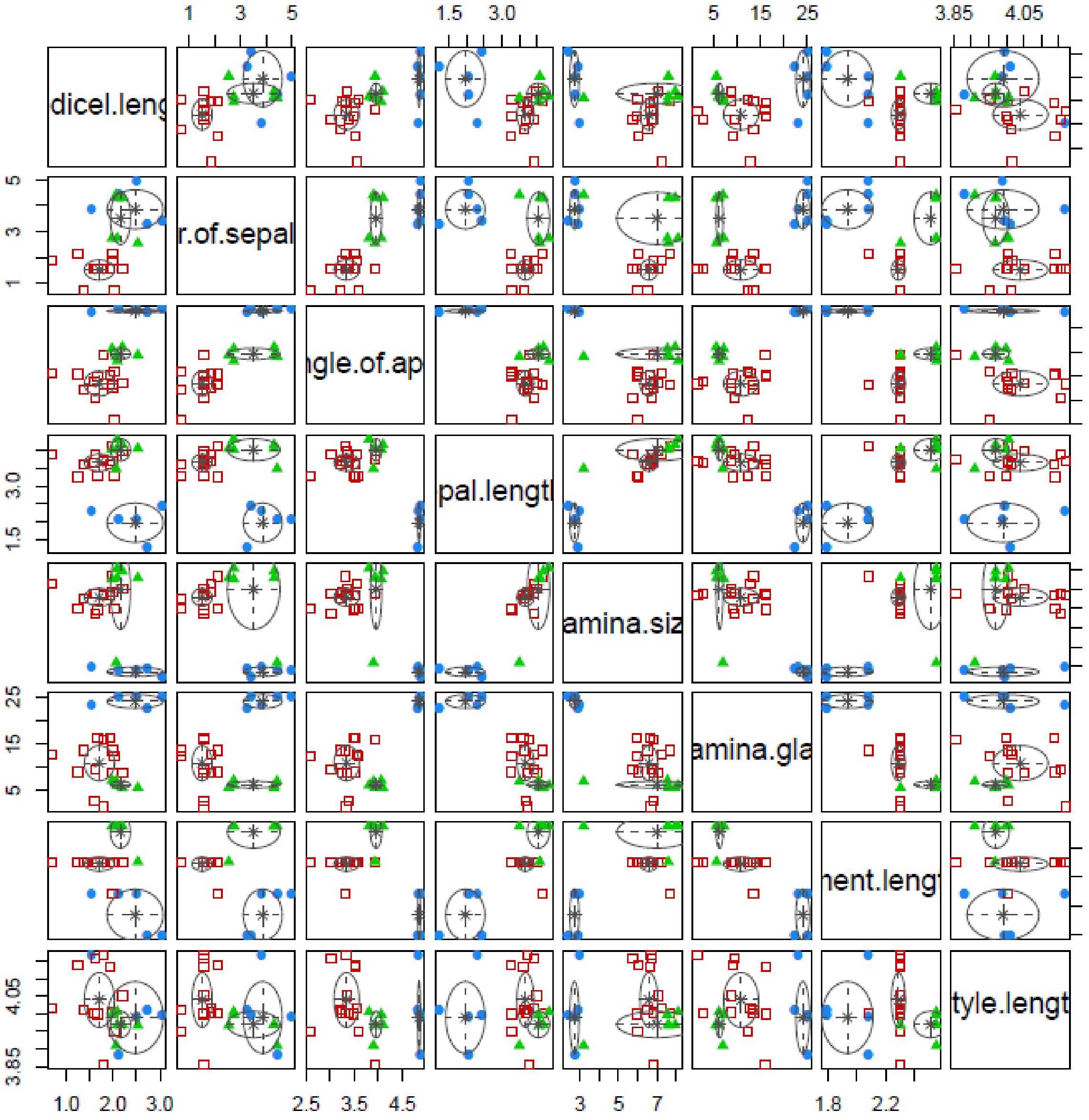
GMM identified three morphospecies: C1, C2 and C3. The RMs between the following pairs of morphospecies C1 & C2, C1 & C3, and C2 & C3, are shown in Figs. 11A, B, and C respectively while the slope of the PDFs between each pair of morphospecies is shown in Figs. 11D, E, and F respectively. All PDFs are bimodal which indicates gaps between the morphospecies. The graph line above the PDF highlights the change of slope sign and confirms the bimodal nature of the PDF in morphospace.

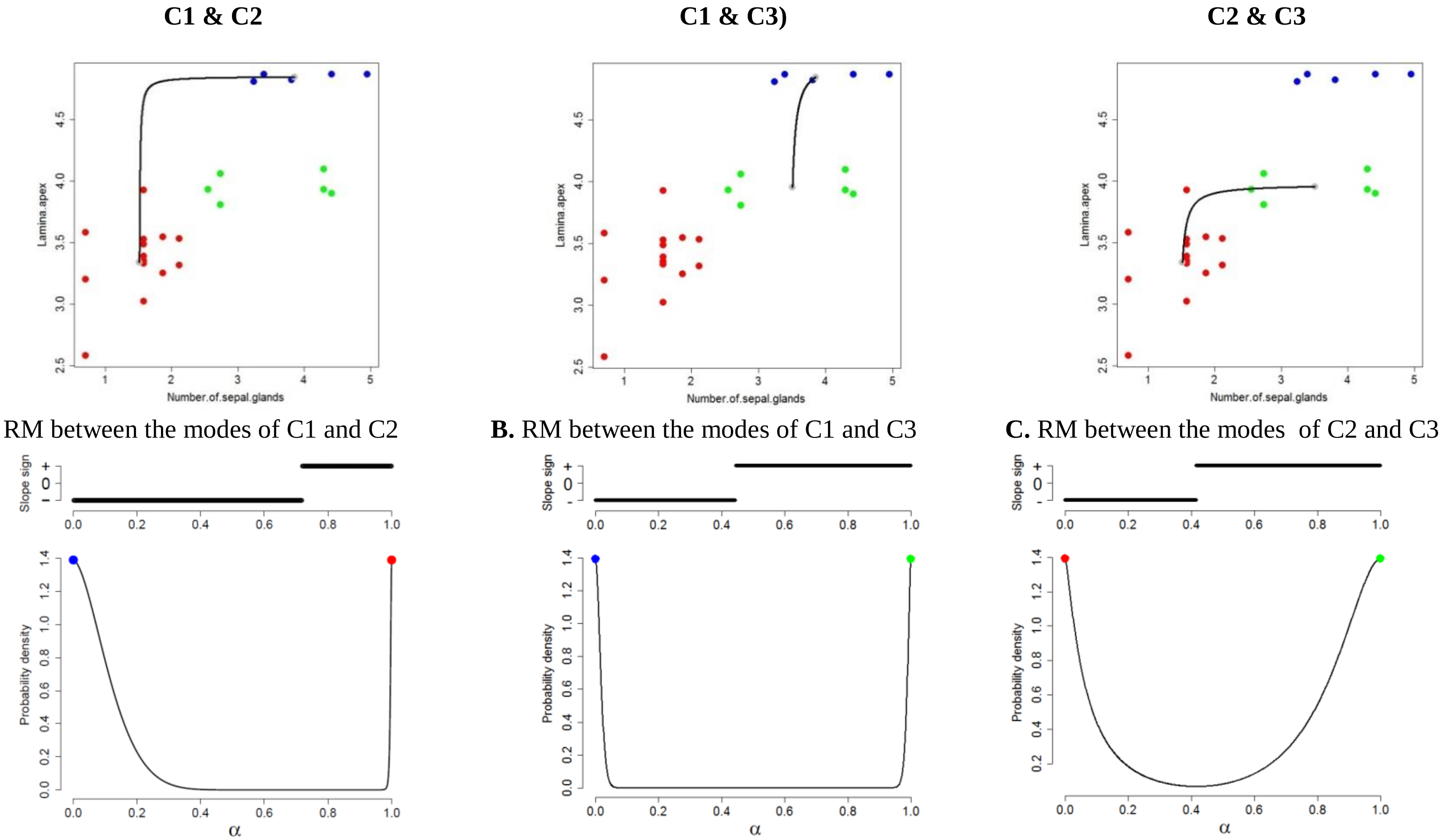
Assessment of morphological characters that separate the morphospecies

The phenotypes of C1 do not overlap with those of C2 and C3, β*values = 1 (Fig. 12A & B); however there are overlapping phenotypes between C2 and C3, β* = 0.526 (Fig. 12C) below the cutoff 0.9 of the proportion of non-overlapping phenotypes.

The quartiles of univariate distribution of the morphological characters from the GMM analysis (points) fit the simulation of the non-overlapping phenotypes of C2 and C3 (boxplots). However, two characters, style length and apex angle, show overlap of the 1st, 2nd and 3rd quartiles of the GMM and the simulation in C1 and C2 (Fig. 13B); five characters, style length, sepal length, number of sepal glands, filament length and apex angle, in C1 and C3 (Fig. 13C); and three characters lamina area, filament length and style length in, C2 and C3 (Fig. 13A).

Figure 10. Pairwise two-dimensional scatterplots of the dataset clade Cerasifolium. Ellipses superimposed on the plot correspond to the covariances of the components. Filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3.





A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

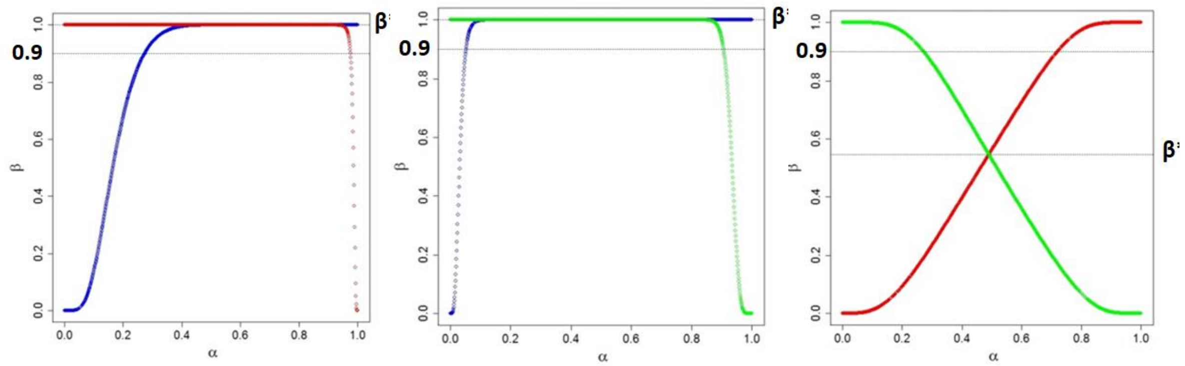
C. RM between the modes of C2 and C3

D. Slope of the PDF (C1 and C2)

E. Slope of the PDF (C1 and C3)

F. Slope of the PDF (C2 and C3)

Figure 11. Ridgeline Manifold (RM) and inference of gaps between the morphospecies. **A-C**: RM = continuous black line; **D- F** Slope of the Probability Density Function (PDFs) of the three morphospecies are C1, C2, and C3; blue, green and red points= modes of these morphospecies.Black straight lines above the slopes indicates the change of the PDFs slope sign.



A. Proportion of phenotypes that overlap in C1 and C2, β*=1

B. Proportion of phenotypes that overlap in C1 and C3, β*=1

C. Proportion of phenotypes that overlap in C2 and C3, β*=0.526

Figure 12. Estimated proportion of individuals with non-overlapping phenotypes, β*, relative to the cutoff 0.9 necessary for the two GMM morphospecies to be distinct.

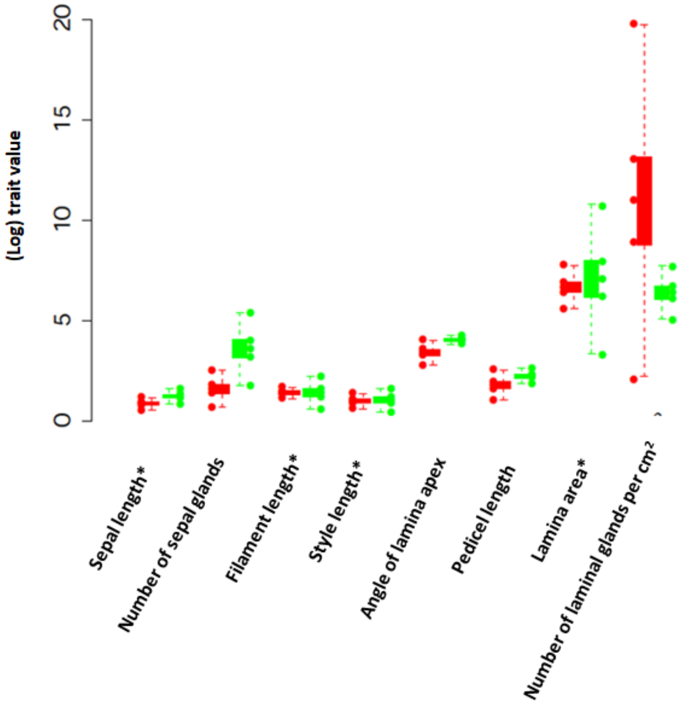


Figure 13.A. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in C2 (red boxplot) and C3 (green boxplot). Overlapping characters are marked with asterisk.

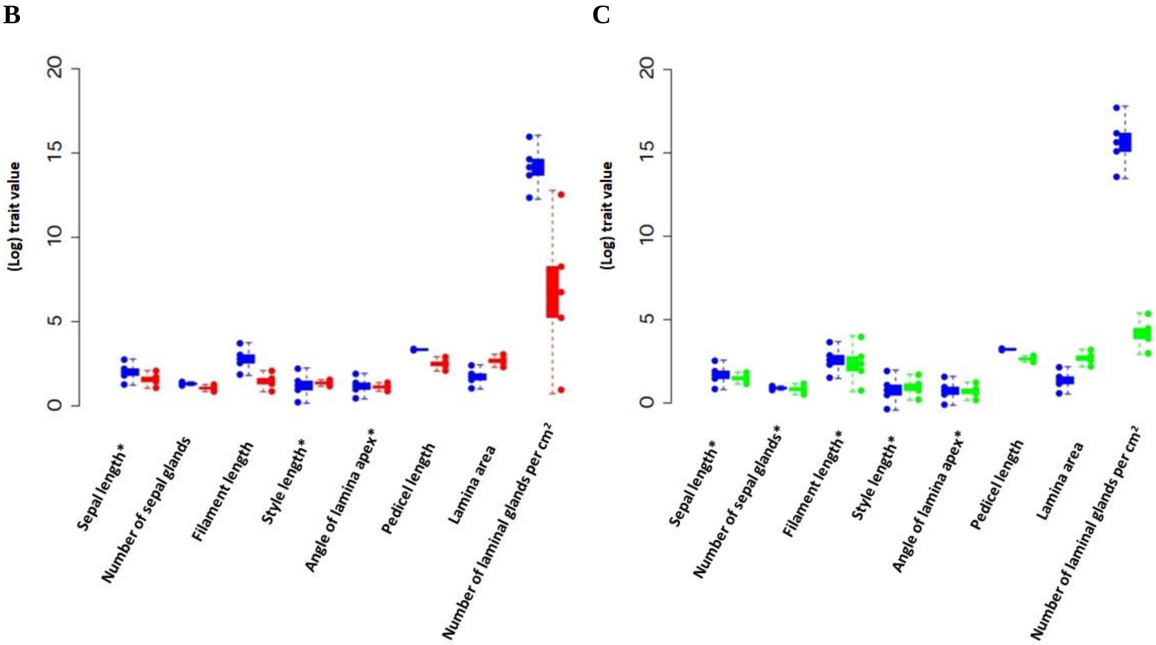


Figure 13. **B**. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in C1 (blue boxplot) and C2 (red boxplot). **C.** Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in C1 (blue boxplot) and C3 (green boxplot). Overlapping characters are marked with asterisks.

Species recognition

The three morphospecies identified by GMM are separated by gaps in the morphospace. The proportion of individuals of C1 that has non-overlapping phenotypes with C2 and C3 respectively are above the cutoff 90% (β*=1), and C1 corresponds to a subclade with 100% bootstrap support. Thus I recognize C1 as species. However, morphospecies C2 and C3 have a considerable proportion of individuals where phenotypes overlap (β*= 0.56); moreover, C2 and C3 have only 46 % and 60 % bootstrap support respectively in the phylogeny. The isotype specimen of *P. cerasifolium* Bak., *Baron 4397*, is placed in C3. I combined C2 and C3 as a single species, *P. cerasifolium* Bak.

Table 1. Summary of the GMM morphospecies analysis in Malagasy *Psorospermum*.

Molecular Phylogenetic Clades	**Morphogroups in clades**	**Number of subclades**	**Total samples used in the phylogeny**	**Total specimens measured for the GMM analyses**	**Number of morphospecies identified by GMM analyses**	**Morphospecies of the GMM compared to the Morphogroups in clades**	**Recognized as species**
Clade Atro-rufum	*P. sp11* *P. sp13* *P. atro-rufum* *P. cf. atro-rufum*	6	11	35	3	P. sp11 P. sp13 (P. atro-rufum + P. cf atro-rufum)	Yes Yes Yes
Clade Trichophyllum	*P. rienanense* *P. sexlineatum* *P. trichophyllum*	4	7	32	4	P. rienanense P. sexlineatum P. trichophyllum	Yes Yes Yes
Clade Brachypodum	*P. brachypodum* *P. cf. brachypodum*	6	7	41	3	P. brachypodum P. cf. brachypodum (1) P. cf. brachypodum (2)	Yes Yes Yes
Clade Chionanthifolium	*P. chionanthifolium* *P. crenatum*	4	6	32	4	P. chionanthifolium P. chionanthifolium 2 P. crenatum P. crenatum 2	Yes Yes Yes Yes
Clade Nanum	*P. nanum* *P. humile*	4	6	29	3	P. nanum P. humile (P. nanum + P. humile)	Yes Yes No
Clade Revolutum	*P. revolutum* *P. cf lanceolatum*	5	8	38	2	P. revolutum P. cf lanceolatum	Yes Yes
Clade Ferrovestitum	*P. sp16* *P. sp17* *P. fanerana* *P. ferrovestitum*	6	10	46	4	P. sp16 (P. sp17+P. fanerana) P. ferrovestitum	Yes Yes Yes
Clade Cerasifolium	*P. malifolium* *P. sp22* *P. sp2* *P. cerasifolium*	6	12	29	3	P. malifolium (P. sp22 + P. sp2+ P. cerasifolium)	Yes Yes
Clade Androsaemifolium	*P. androsaemifolium* *P. cf. androsaemifolium* *P. sp19*	5	10	30	4	P. cf. androsaemifolium (1) P. cf. androsaemifolium (2) P. androsaemifolium P. sp19	Yes Yes Yes Yes

Clade Atro-rufum

Clade Atro-rufum consists of four morphogroups: *Psorospermum atro-rufum*, *P. cf atro-rufum*, *P. sp13* and *P. sp11*. The GMM analysis identified 3 morphospecies: C1, C2 and C3 (Appendix S4), which correspond respectively to (*Psorospermum atro-rufum + P. cf atro-rufum)*, *P. sp13* and *P. sp11* (Fig. 14). Eight characters were used in GMM analyses: number of ovules per carpel; lamina length, width, apex angle, surface area, shape, number of secondary veins; and cotyledon width (Appendix Table S1; Fig. S3).

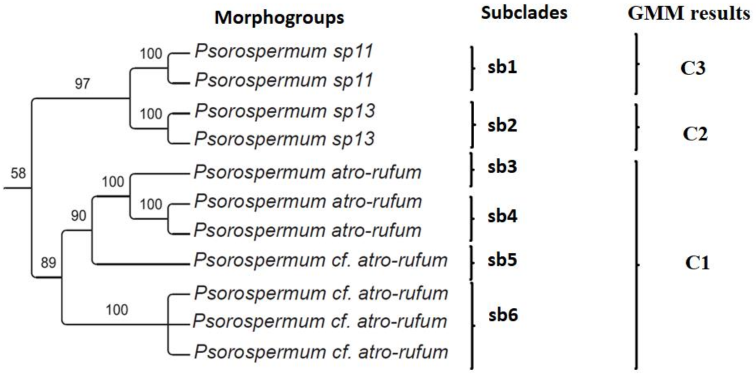


Figure 14. Expanded clade Atro-rufum retrieved from ML molecular phylogenetic analysis of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclade; C1, C2 and C3 are the morphospecies identified by GMM.

Table 2 summarizes the results of the GMM analysis in the Atro-rufum clade. There is no overlap of the characters used in the analysis except for the character, number of hairs per cm^2 on the lamina surface, which overlaps between C2 and C3 (Table 2), nevertheless β* is equal to 1 in all cases (Table 2; Fig. S7). Gaps between the morphospecies are confirmed by the analyses (Table 2; Fig. S6). The bimodal nature of PDF slopes between C1 and C2, and C1 and C3 are not conspicuous, but the graph lines of direction of the slopes (Table 2; Fig. S6) indicates gaps in both cases. The isotype specimen of *P. atro-rufum* (*Humblot 4509*) is included within C1.

Although clade Atro-rufum has only 58% bootstrap support, (*P. atro-rufum + P. cf atro-rufu*m), *P. sp13* and *P. sp11* have 89%, 100% and 100% bootstrap support respectively. Thus, three species are delimited, *P. atro-rufum* H. Perr, *P. sp11* and *P. sp13*, with the last two being potentially new species.

Table 2. Summary of the GMM analysis and morphological character assessment of the Atro-rufum clade

	C1 & C2	C1 & C3	C2 & C3
Bimodal nature of the slope	yes	yes	yes
β*	1	1	1
Overlapping characters in simulation	Number of ovules per carpel Cotyledon width	Number of ovules per carpel Cotyledon width	No overlap

Clade Trichophyllum

Clade Trichophyllum consists of three morphogroups: *P. rienanense, P. trichophyllum and P. sexlinenatum*. The GMM analysis identified 4 morphospecies C1, C2, C3, and C4 (Appendix S5; Fig. S12). C1 and C2 correspond to *Psorospermum rienanense*, C3 corresponds to *P. trichophyllum*, and *C4* correspond to *P. sexlineatum* (Fig. 15). Seven characters were used in GMM analyses: pedicel length; filament length, lamina apex angle, surface area, number of glands per cm^2, number of hairs per cm^2; and petiole length. (Appendix Table S1; Fig. S11).

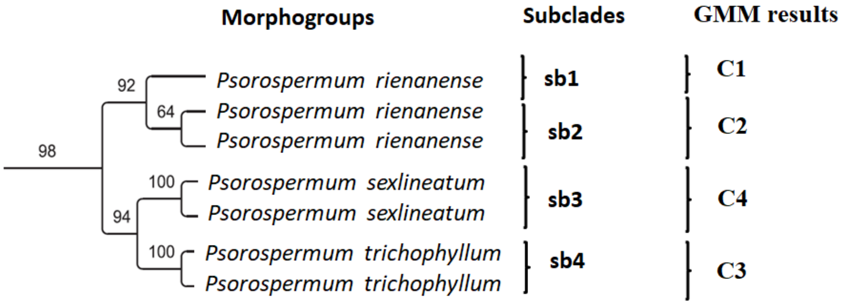


Figure 15. Expanded clade Trichophyllum retrieved from ML molecular phylogenetic of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclade; C1, C2, C3 and C4 are the morphospecies identified by GMM.

Table 3 summarizes the results of the GMM analysis in the Trichophyllum clade. C1 and C2 have β* lower than 0.9, β* = 0.49 as well as C2 and C3 and β* = 0.72 (Table 3; Fig. S16 A, D). Gaps between morphospecies are confirmed by the analyses (Table 3; Fig. S15). The graph line of direction of the slopes (Table 3; Fig. S15) indicates gap between C3 and C4, and β* = 1 for C3 and C4. *Psorospermum trichophyllum* and *P. sexlineatum* have both 100% bootstrap support. The type specimen of *P. trichophyllum* (*Baron 3016*) is placed in C3, while the type specimen of *P. sexlineatum* (*Perrier de la Bâthie 5247*) is in C4. Thus, I recognize C3 and C4 as *P. trichophyllum* Bak., *P. sexlineatum* H. Perr., respectively.

The graph line of direction of the slopes (Table 3; Fig. S15) indicates gap between C1 and C2, and β* = 0.49. However, sb2 has only 64% bootstrap support while (sb1 + sb2) has 92% (Fig. 15). The type specimen of *P. rienanense* (*Humbert 3586*) is placed in C1. Thus, I combined C1 and C2 as a single species, *P. rienanense* Bak.

Table 3. Summary of the GMM analysis and morphological character assessment of the Trichophyllum clade

	C1 & C2	C1 & C3	C1 & C4	C2 & C3	C2 & C4	C3 & C4
Bimodal nature of the PDF	Not conspicuous	Conspicuous	Not conspicuous	Not conspicuous	Not conspicuous	Not conspicuous
β*	0.49	1	1	0.72	1	1
Overlapping characters in simulation	All except Number of lamina glands/cm^2	Filament length Number of lamina hairs/cm^2 Number of lamina glands/cm^2	Filament length Number of lamina hairs/cm^2	All except number of lamina glands/cm^2	Filament length Pedicel length Lamina apex angle Number of lamina hairs/cm^2	Filament length Pedicel length Number of lamina hairs/cm^2
Characters with quantiles that deviate from the simulation	Lamina number of glands per cm^2	None	None	Lamina number of glands per cm^2	None	None

Clade Brachypodum

Clade Brachypodum consists of two morphogroups: *P. brachypodum* and *P. cf. brachypodum*. In this clade, *P. brachypodum* is polyphyletic and *P. cf. brachypodum* is paraphyletic (Fig.16). GMM identified 3 morphospecies, C1, C2 and C3 (Appendix S6). Morphogroup *P. brachypodum* corresponds to C1 and C2, while *P. cf. brachypodum* matches C3 (Fig. 16). Five characters were used in GMM analyses: number of petal glands; filament length; number of ovules per carpel; cotyledon surface area; and radicle length (Appendix Table S1; Fig. S20).

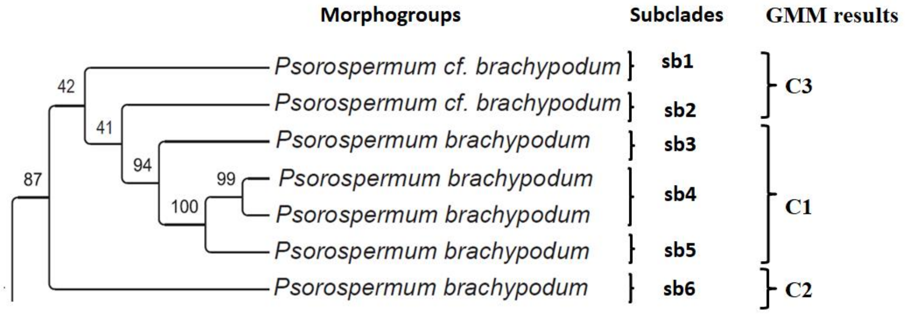


Figure 16. Expanded clade Brachypodum retrieved from the ML molecular phylogeny of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclade; C1, C2 and C3 are the morphospecies identified by GMM.

Table 4 summarizes the results of the GMM analysis in the Brachypodum clade. There is no overlap of the characters used in the analyses except for the characters number of ovules per carpel and filament length which overlaps between C1 and C2 (Table 2), and the number of petal glands, filament length and number of ovules per carpel which overlap between C2 and C3, nevertheless β* is equal to 1 in all cases (Table 4; Fig. S24). Gaps between the morphospecies are confirmed by the analyses (Table 4; Fig. S23), the bimodal nature of PDF slopes and the graph lines of direction of the slopes indicate gaps in all. Only C1 corresponds to a well-supported clade; since C1 met also all other criteria for species recognition, I recognize C1 as species *P. brachypodum* Bak. Although C2 and C3 are paraphyletic and do not correspond to any well-supported clade, all the gaps are conspicuous and β* =1 in all cases which implies that they are morphologically distinct (Fig. S24). I therefore recognize C2 and C3 as a potential new species.

Table 4. Summary of the GMM analysis and morphological character assessment of the Brachypodum clade

	C1 & C2	C1 & C3	C2 & C3
Bimodal nature of the slope	Conspicuous	Conspicuous	Conspicuous
β*	1	1	1

Overlapping characters in simulation	Filament length Number of ovules per carpel	No overlap	Number of petal glands Filament length Number of ovules per carpel

Clade Chionanthifolium

Clade Chionanthifolium is composed of two morphogroups, *P. chionanthifolium* and *P. crenatum*. Both morphogroups are monophyletic but each clade is subdivided into 2 subclades. GMM identified four morphospecies: C1, C2, C3 and C4 (Appendix S8). *P. chionanthifolium* corresponds to morphospecies C1 and C2, and the isotype of *P. chionanthifolium* (*Chapelier s.n.*) is within C1. *P. crenatum* corresponds to C3 and C4 (Fig. 17), the type specimen of *P. crenatum* (*Baron 2857*) is within C4. Five characters were used in GMM analyses: lamina length, width, length from the base to the widest point, surface area, number of secondary veins; and ratio cotyledon length:width (Appendix Table S1; Fig. S28). There is practically no overlap of the characters used in the analysis and gaps are conspicuous between the morphospecies (Table 5; Fig. S32). Thus four species are delimited, *P. chionanthifolium* Spach, *P. crenatum* Hochr., and two potentially new species.

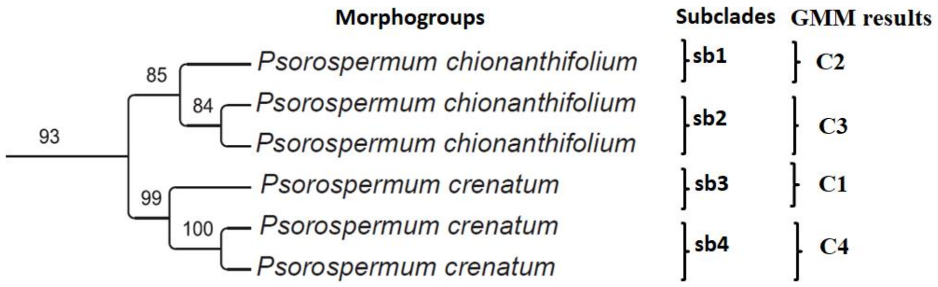


Figure 17. Expanded clade Chionanthifolium retrieved from the ML molecular phylogeny of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclade; C1, C2 and C3 are the morphospecies identified by GMM.

Table 5. Summary of the GMM analysis and morphological character assessment of the Chionanthifolium clade

	C1 & C2	C1 & C3	C1 & C4	C2 & C3	C2 & C4	C3 & C4
Bimodal nature of the slope	Conspicuous	Conspicuous	Conspicuous	Conspicuous	Not conspicuous	Not conspicuous
β*	1	1	1	1	1	1
Overlapping characters in simulation	None	None	None	None	Ratio cotyledon: radicle length	None

Clade Nanum

Clade Nanum is composed of two morphogroups, *P. nanum* and *P. humile*. Both morphogroups are monophyletic, but *P. nanum* has three subclades (Fig. 18). GMM identified three morphospecies, C1, C2 and C3 (Appendix S7). Eight specimens of *P. humile* are placed in C1, including the isotype specimen (*Perrier de la Bâthie 14016*); nine specimens of *P. nanum* are in C2, including the type specimen (*Perrier de la Bâthie 17353*); a mixture of specimens of both *P. humile* and *P. nanum* are placed into C3. Six characters are used in the GMM analyses: number of petal glands; staminode length; filament length; number of ovules per carpel; lamina apex angle, and length (Appendix Table S1; Fig. S37).

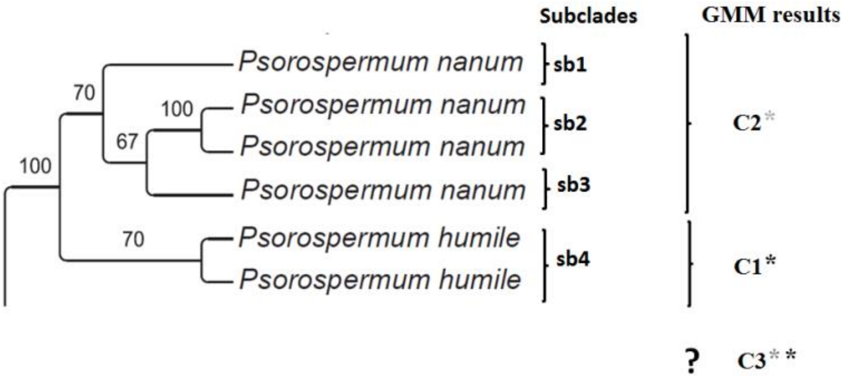


Figure 18. Expanded clade Nanum retrieved from the ML molecular phylogeny of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclade; the morphospecies identified by GMM are C1, C2 and C3. The morphospecies C3 that has mixed specimens of C1 and C2 is indicated by two asterisks in grey and black.

Interestingly, the bimodal nature of the PDF slope between C2 and C3 is inconspicuous as shown by the directional lines of the slopes (Table 6; Fig. S40 F). Moreover, the investigation of overlapping phenotypes yielded a β* value below the cutoff 0.9 indicating that there is overlap of phenotypes between morphospecies C2 and C3 (Fig. S41 C). Taking into account those results, only two species are delimited in Clade Nanum, *P. nanum* and *P. humile*. Perhaps there is overestimation of the number of morphospecies by GMM (see Discussion).

Table 6. Summary of the GMM analysis and morphological character assessment of the Nanum clade

	C1 & C2	C1 & C3	C2 & C3
Bimodal nature of the slope	Not conspicuous	Conspicuous	Not conspicuous
β*	1	1	0.8
Overlapping characters in simulation	Filament length Number of ovules per carpel	Number of ovules per carpel Number of petal glands	Filament length Number of petal glands Number of ovules per carpel

Clade Revolutum

Clade Revolutum is composed of two morphogroups: *Psorospermum revolutum* and *P. cf. lanceolatum*. Both morphogroups are monophyletic but *P. revolutum* consists of 3 subclades while *P. cf. lanceolatum* consists of 2 (Fig. 19). The GMM analysis identified 2 morphospecies, C1 and C2 (Appendix S10). Morphospecies C1 matches *P. revolutum* while C2 matches *P. cf. lanceolatum* (Fig. 19). The type specimen of *P. revolutum* (*Scott Elliot 2313*) is placed in C1. The characters used in GMM analyses are: lamina apex angle, number of glands per cm^2; petiole length; pedicel length; number of sepal glands; and number of petal glands (Appendix Table S1; Fig. S54). There is a conspicuous gap between C1 and C2 in morphospace (Fig. S57), the proportion of individuals with non-overlapping phenotypes are above the cut-off 0.9, β* =1 (Fig. S58), and both C1 and C2 correspond to subclades with 100% bootstrap support (Fig. 19). Thus, I recognize two species in clade Revolutum, *P. revolutum* Bak., and *P. cf. lanceolatum.*

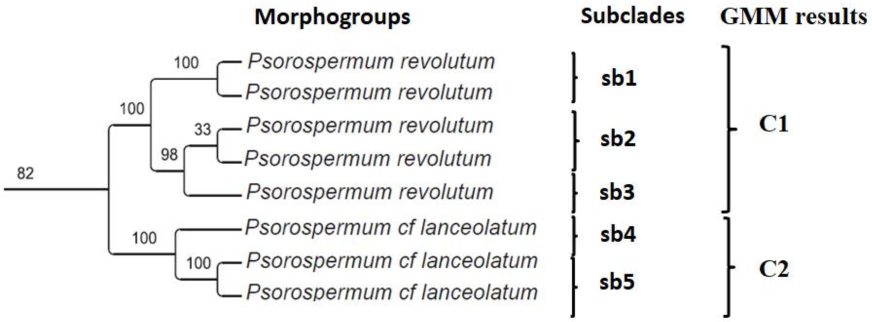


Figure 19. Expanded clade Revolutum retrieved from the ML molecular phylogeny of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclade; the two morphospecies identified by GMM are C1 and C2.

Table 7. Summary of the GMM analysis and morphological character assessment of the Revolutum Clade

	C1 & C2
Bimodal nature of the slope	Conspicuous
β*	1
Overlapping characters in simulation	Number of sepal glands

Clade Ferrovestitum

Clade Ferrovestitum is composed of four morphogroups: *Psorospermum ferrovestitum*, *P. fanerana*, *P. sp16* and *P. sp17*. All morphogroups are monophyletic but *P. fanerana* and *P. sp17* have 2 subclades each (Fig. 20). The GMM analysis identified 4 morphospecies, C1, C2, C3 and C4 (Appendix S9). The molecular and the morphological analyses in the Ferrovestitum clade seem to be congruent (Fig. 20). While C2 and C4 match *P. ferrovestitum* and *P. sp17* respectively, ten specimens of *P. fanerana* were grouped with the specimens of *P. sp16* in C1. The type specimen of *P. ferrovestitum* (*Baron s.n.*) is in C2 and that of *P. fanerana* (*Baron 5*) in C3. The characters used in GMM analyses are: petal length; lamina apex, length, width, shape, and number of hairs per cm^2 (Appendix Table S1; Fig. S45).

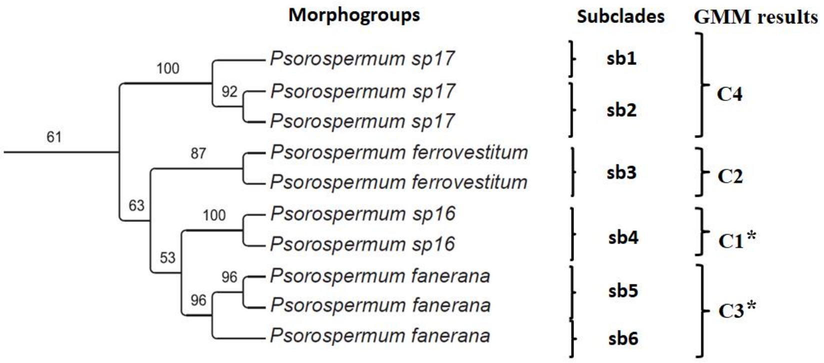


Figure 20. Expanded clade Ferrovestitum retrieved from the ML molecular phylogeny of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclades; the morphospecies identified by GMM are C1, C2, C3, and C4. The morphospecies C1 and C3, indicated by asterisks, have mixed specimens.

Gaps between the morphospecies are confirmed by the analyses (Table 8; Fig. S49). The bimodal nature of PDF slopes between C1 and C2, and between C1 and C3 are not conspicuous, but the graph lines of direction of the slopes (Fig. S49) indicates gaps in both cases. There is overlap of phenotypes between C1 and C2 and between C1 and C3, β* < 0.9 (Table 8; Fig. S50 A). Although the GMM results in this particular dataset corroborate the phylogeny; (*P. sp16* + *P. fanerana*) has only 53% support while *P. ferrovestitum* and *P. sp17* have 87% and 100% bootstrap support respectively (Fig. 20). C1 met only two of the four criteria of species recognition, therefore I recognize only three species in clade Ferrovestitum, *P. ferrovestitum* Bak., *P. fanerana* Bak. (including *P. sp16*), and *P. sp17*. The latter is a potentially new species.

Table 8. Summary of the GMM analysis and morphological character assessment of the Ferrovestitum clade

	C1 & C2	C1 & C3	C1 & C4	C2 & C3	C2 & C4	C3 & C4
Bimodal nature of the slope	Not conspicuous	Not conspicuous	conspicuous	Conspicuous	Conspicuous	Conspicuous
β*	0.55	0.78	0.98	1	1	1
Overlapping characters in simulation	All but lamina apex angle	Lamina apex angle Lamina shape Number of lamina hairs per cm^2	Petal length	Lamina shape Number of lamina hairs per cm^2	Petal length	Petal length

Clade Androsaemifolium

Clade Androsaemifolium is composed of three morphogroups: *P. androsaemifolium*, *P. cf. androsaemifolium* and *P. sp19* (Fig. 21). Each morphogroup is monophyletic, but *P. cf. androsaemifolium* has 2 subclades, as does *P. sp19*. GMM identified 4 morphospecies, C1, C2, C3 and C4 (Appendix S11). The characters used in GMM analyses are: angle cotyledon-radicle; lamina length, width, surface area, shape, and number of secondary veins (Appendix Table S1; Fig. S68 B).

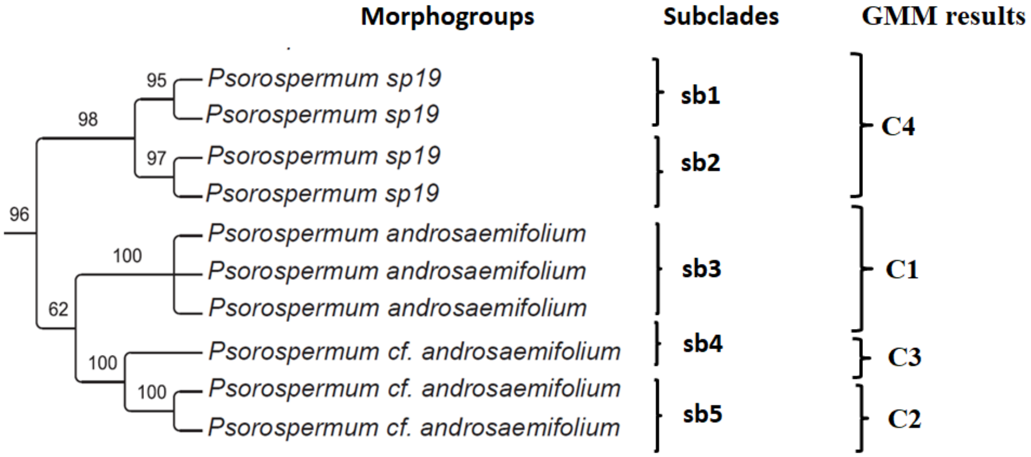

Morphogroups
Subclades
GMM results
95
98
97
96
100
62
100
100
Psorospermum sp19
Psorospermum sp19
Psorospermum sp19
Psorospermum sp19
Psorospermum androsaemifolium
Psorospermum androsaemifolium
Psorospermum androsaemifolium
Psorospermum cf. androsaemifolium
Psorospermum cf. androsaemifolium
Psorospermum cf. androsaemifolium
sb1
sb2
sb3
sb4
sb5
C4
C1
C3
C2

Figure 21. Expanded clade Androsaemifolium retrieved from the ML molecular phylogeny of *Psorospermum*. Numbers above the braches are bootstrap support values; sb=subclades; the morphospecies identified by GMM are C1, C2, C3, and C4.

Table 9 summarizes the results of GMM analysis. Three subclades seem to match the molecular results: C1 corresponds to *P. androsaemifolium* and includes the isotype (*Baron 120*); *P. cf. androsaemifolium* corresponds to two different morphospecies, C2 and C3. However *P. sp19*, which is represented by two subclades in the phylogeny, corresponds to a single morphospecies C4. Gaps are conspicuous between the morphospecies (Fig. S72). The proportion of individuals with non-overlapping phenotypes between the morphospecies in all cases are above 0.9 i.e. all β* are equal to 1 (Fig. S73). Thus four species are recognized, *P. androsaemifolium* Baker, and three potentially new ones.

Table 9. Summary of the GMM analysis and morphological character assessment of the Androsaemifolium clade

	C1 & C2	C1 & C3	C1 & C4	C2 & C3	C2 & C4	C3 & C4
Bimodal nature of the slope	Conspicuous	Conspicuous	Conspicuous	Conspicuous	Conspicuous	Conspicuous
β*	1	1	1	1	1	1
Overlapping characters in simulation	Shape	No overlap	No overlap	No overlap	No overlap	No overlap

DISCUSSION

Twenty seven species out of the thirty morphospecies identified by GMM are recognized in this study using the integrative taxonomy method; 11 are new, 13 correspond with names recognized by Perrier de la Bâthie, if only at the level of type specimens (Fig. 22 A and B), and 3 are synonyms .

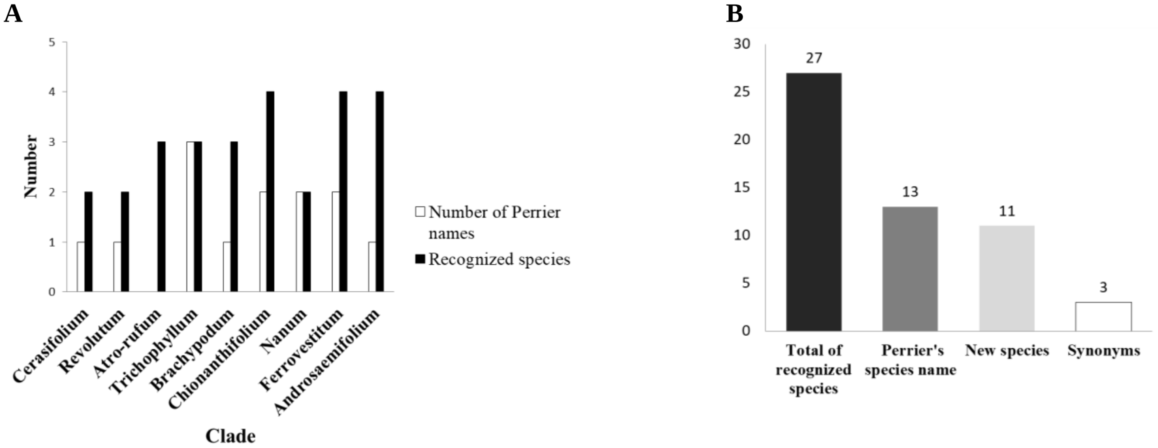


Figure 22. A. Comparison per clade of the number of species identified and the number of Perrier de la Bâthie species recognized by the integrative taxonomy analysis. **B.** Histograms of the recognized species by the integrative taxonomy analysis, the number of Perrier's species name, the number of new species, and the number of synonyms.

These results are not only useful for the ongoing revision of *Psorospermum* (Ranarivelo in preparation), but have also a direct impact on the conservation effort in Madagascar. The newly recognized species will be used as support for the program of preservation of the unique flora in several areas in Madagascar, especially those with high anthropogenic pressure where the new species occur (such as the mining areas surrounding or inside some protected areas for example). A preliminary report has already been sent to the Malagasy National Park management offices in charge of those areas.

However, since I am using a new approach, caveats need to be discussed for further improvement of the results. The mismatch between the morphogroups in the molecular analyses and the morphospecies identified by GMM analyses is striking. Only two out of the nine clades, clade Revolutum and clade Ferrovestitum, have morphogroups that match with the morphospecies identified by the GMM analysis, as summarized in Fig. 23. Overall 30 morphospecies are recognized by the GMM analyses; and I had recognized 26 morphogroups. Interestingly, all the morphogroups used in the phylogeny were completely distinct when they were tested with LDA (e.g. Fig. 7A). Perhaps GMM analyses have the advantage of being able to evaluate the probabilities of each individual belonging to a putative group, and detecting slightly overlapping characters that have a significant impact on the final output. Since all the variables are continuous, the mismatch could be due to the influence of latent variables, such as measurement errors in the datasets. GMM also might overestimate the number of morphospecies because BIC performs better in identifying the posterior probability densities rather than the morphospecies *per se* (Biernacki *et al.*, 2000). Since in GMM, a morphospecies is a mixture of components (Gaussian distributions), BIC tends to recognize more components in the mixture as morphospecies, thus overestimating the final number of morphospecies (Baudry *et al.*, 2010). However, although GMM analysis identified more morphospecies than there were morphogroups in five out of the nine clades, it also identified fewer morphospecies in two clades, Atro-rufum and Cerasifolium (Fig. 23).

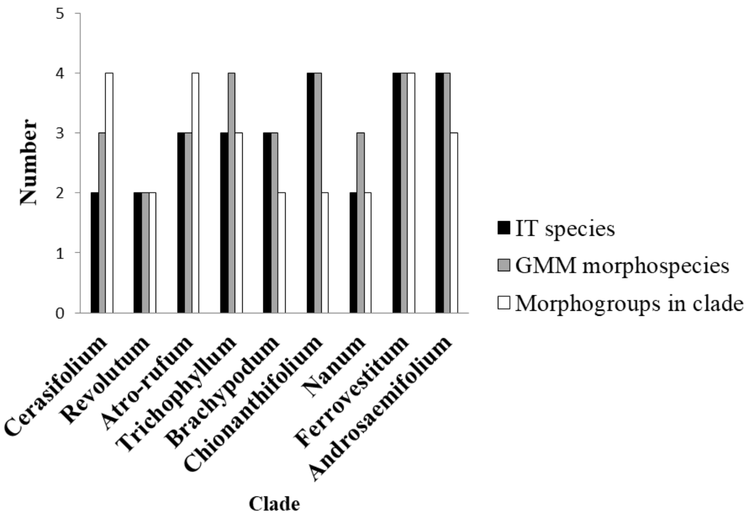


Figure 23. Summary of the integrative taxonomy of Malagasy *Psorospermum*, the number of morphogroups in the molecular phylogeny and the number of morphospecies delimited by GMM, also compared. IT species: Integrative taxonomy species.

The mismatch may have been caused by technical issues. The source of the problem may lie in the selection of the characters. Missing characters can be an issue. The 29 characters I measured (Appendix Table S1) certainly do not cover all characters; I did not include any anatomical characters for example, and there must be other informative characters that were not included. On the other hand, Scucca and Raftery (2014) argued that the identification of the morphospecies can also be corrupted by noise variables, *i.e.* characters that are not informative; however it is unlikely that they are present since the variables were selected using PCA. Additional methods for selecting variables need to be investigated.

My results suggest also that one or two characters might express visible differences between two morphogroups but do not necessarily separate them as species. In case of clade Cerasifolium, for example, the GMM analysis suggests that there are only three morphospecies and could not separate *Psorospermum* sp22 and P. sp2 (Fig. 5). However, individuals of morphogroup *P. sp2* can be separated from those of morphogroup *P. sp22* by the difference in their pedicel length (ca. 5.5-6 cm in morphogroup *P. sp2* vs. ca. 2.5-3 cm in morphogroup *P. sp22*), and their lamina area (ca. 6-8 cm2 in morphogroup *P. sp2* vs. ca. 4-5 cm2 in morphogroup *P. sp22*). Although these characters were included in the GMM analysis, morphogroup *P. sp2* and morphogroup *P. sp22* were grouped together.

The criteria for recognizing species in this study is that species are units (morphospecies) identified by GMM which have evidence of a gap between them, with the proportion of individuals that have non-overlapping phenotypes, β*, above a cutoff of 0.9, and that correspond to well-supported subclades i.e. the ultimate monophylies. My results show that the number of species recognized generally equals the number of morphospecies estimated by GMM - seven out of the nine clades - and the number of recognized species is only one fewer in the Nanum and Ferrovestitum clades. These results suggest that the output of GMM analysis alone influences the integrative analyses as a whole. It appears to confirm the theory that the morphological data is fully integrated into the molecular data, as species delimited based on phylogeny alone are overestimated (Agapow *et al*. 2004).

The number of molecular subclades (*i.e.* ultimate monophyletic units within a clade) is higher than that of the GMM morphospecies in seven out of the nine clades, only two clades have subclades that match GMM morphospecies and none has fewer (Fig. 24). The subclades that do not match morphospecies may at one time have been genetically separated, but subsequent gene flow joined them because of anthropogenic disturbance for example. This pattern might also be explained by the stage in the speciation process at which the subclades are, when outputs from different analyses then may well disagree (Sites & Marshall, 2004; de Queiroz 1998, 2007; Aguilar *et al.*, 2016); here the molecular divergences are not fully reflected by morphological and are not recovered by the GMM analyses.

Well-supported clades at deeper nodes of the phylogenetic tree indicate that after a lineage split, those new lineages subsequently had enough time to establish their distinctiveness. However, in a case of rapid radiation as in Malagasy *Psorospermum* (Ranarivelo *et al.*, in preparation) it is likely that lineages in a clade are still closely related morphospecies. Biological processes like incomplete lineage sorting, and non-biological processes like too few specimens, cannot be ignored either (Metha *et al.*, 2016). Taking these uncertainties into consideration, I chose to delimit supported clades at deeper nodes rather than at the tip of the phylogenetic tree.

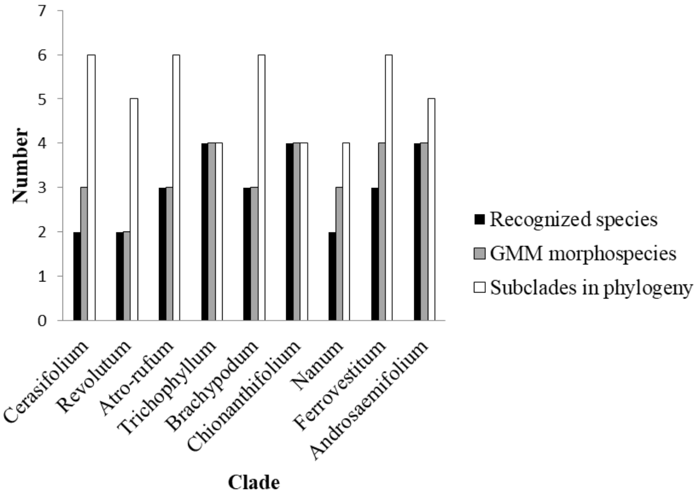


Figure 24. Comparison of the number of subclades in the molecular phylogeny clades and the morphospecies delimited by GMM.

However, I tested GMM on the subclades of 5 clades that were chosen randomly (Table 10). I analyzed them separately following the same approach I used on the clades, *i.e.* selecting characters using PCA and identifying morphospecies using GMM. In 5 subclades tested, the number of morphospecies identified by GMM analyses was always more than the number of subclades analyzed (Table 10), with the morphospecies sometimes being almost as many as the number of samples. Thus using subclades for GMM analyses may well overestimates the morphospecies. For example, the sample size may be very small and affect the output of these tests, so producing misleading results.

Table 10. Separate GMM analyses of subclades

Clade	Atro-rufum	Trichophyllum	Chionanthifolium	Cerasifolium	Nanum
Subclade analyzed	(sb1+sb2) Fig.14	(sb3+sb4) Fig. 15	(sb3+sb4) Fig.16	(sb4 +sb5) Fig. 6	(sb1 +sb2) Fig. 17
Number of characters selected using PCA	7	6	5	5	7
Sample size	22	22	20	21	16
GMM morphospecies	3	3	7	13	14

GMM morphospecies in initial analysis	2	2	2	1	1

I used GMM instead of multivariate analysis methods commonly used in systematics, such as LDA or PCA, because of GMM's advantage in identifying morphospecies without the necessity of manually choosing morphospecies from, *e.g.*, PCA results. PCA provides visualization of a multivariate data set where patterns of morphospecies of the individuals are shown in a two- or three-dimensional scatterplot. However, problems often arise in separating the morphospecies when gaps are not obvious, and the delimitation of morphospecies then relies on the arbitrary decision of the user. Alternatively, a technique like LDA can discriminate morphospecies or validate arbitrary morphospecies delimited from PCA. However LDA requires *a priori* delimitation of morphospecies (Fischer, 1936; Xanthopoulos *et al*., 2012). Moreover, GMM is a method based on probability clustering and instead of linearly mapping the characters on the component axes, it evaluates the probability of each observation belonging to each morphospecies. Thus, I used LDA only to test the preliminary morphogroups that were based on gross overall similarity, PCA to select the morphological variables to be used in the GMM analyses. I used also other criteria such as gaps between morphospecies and the proportion of individuals with non-overlapping species between the morphospecies, β*, to further support the recognition of the morphospecies.

CONCLUSION

I used a new approach to delimit species in Malagasy *Psorospermum*, and the results will guide the revision of the genus and the new species to be described. Herbarium specimens are valuable resources in retrieving morphological information that can be combined with molecular data. The GMM method is a straightforward tool that helps in the delimitation of morphospecies directly from the morphological dataset. The caveats above suggest areas for future investigation that may improve methods of species delimitation, such as the choice of morphological characters and how to delimit the clades in the molecular phylogeny to be analyzed.

ACKNOWLEDGEMENTS

I thank Dr. P. F. Stevens for his advice, insights and support during this work, and Dr. P. C. Hoch, Dr. E. A. Kellogg, Dr. B. Ruhfel for their comments and constructive criticism. My special thanks to Dr. Ivan Jiménez for insightful discussion of this project. I thank the Missouri Botanical Garden and the University of Missouri – St. Louis for logistic support. This study was funded by the Whitney R. Harris World Ecology Center - University of Missouri St. Louis and the Systematics Research Fund (SRF) - Linnean Society and the Systematics Association. I thank also the following herbaria, MO, K, G, P and TAN, for making their collections available.

REFERENCES

Abdi H, William LJ. 2010. Principal component analysis. *WIREs Computational Statistics* **2:** 433–459.

Agapow PM, Bininda-Emonds OR, Crandall KA, Gittleman JL, Mace GM, Marshall JC, Purvis A. 2004. The impact of species concept on biodiversity studies. *The Quarterly Review of Biology*. **79:**161-79.

Aguilar C, Wood PL Jr, Belk MC, Duff MH, Sites JW Jr. 2016. Different roads lead to Rome: integrative taxonomic approaches lead to the discovery of two new lizard lineages in the *Liolaemus montanus* group (Squamata: Liolaemidae). *Biological Journal of the Linnean Society.* doi:10.1111/bij.12890.

Baudry J-P, Raftery AE, Celeux G, Lo K, Gottardo R. 2010. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* **9:** 332–353.

Barrett CF, Freudenstein JV 2011. An integrative approach to delimiting species in a rare but widespread mycoheterotrophic orchid. *Molecular Ecology* **20:** 2771–2786.

Behrens K, Barnes K. 2016. *Wildlife of Madagascar.* Princeton University Press, New Jersey.

Biernacki C, Celeux G, Govaert G. 2000. Assessing a mixture model for morphospeciesing with the integrated completed likelihood,. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22:** 719–725.

Blaimer BB. 2012. Untangling complex morphological variation: taxonomic revision of the subgenus *Crematogaster* (*Oxygyne)* in Madagascar, with insight into the evolution and biogeography of this enigmatic ant clade (Hymenoptera: Formicidae). *Systematic Entomology* **37:** 240–260.

Carstens BC, Satler JD. 2013. The carnivorous plant described as *Sarracenia alata* contains two cryptic species. *Biological Journal of the Linnean Society* **109:** 737–746.

Dayrat B. 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society* **85:** 407–415.

Eberle J, Warnock RC, Ahrens D. 2016. Bayesian species delimitation in Pleophylla chafers (Coleoptera) - the importance of prior choice and morphology. *BMC Evolutionary Biology* **16:** 94.

Edwards DL, Knowles LL. 2014. Species detection and individual assignment in species delimitation: can integrative data increase efficacy? *Proceeding of the Royal Society B* **281:** 20132765.

Ezard TH, Pearson PN, Purvis A. 2010. Algorithmic approaches to aid species' delimitation in multidimensional morphospace. *BMC Evolutionary Biology* **10:** 175.

Fisher RA. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7:** 179–188.

Fraley C, Raftery AE. 1998. How many morphospecies? Which clustering method? Answers via model-based morphospecies analysis. *The Computer Journal* **41:** 578–588.

Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* **97:** 611–631.

Fraley C, Raftery AE, Scrucca L. 2012. Package 'mclust'. Normal mixture modeling for model-based morphospeciesing, classification, and density estimation.

Fraley C, Raftery AE, Scrucca L, Murphy TB, Fop M, Scrucca ML. 2017. Package 'mclust'.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. 2008. mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-2, URL http://CRAN. R-project. org/package= mvtnorm.

Goodman SM, Benstead JP. 2005. Updated estimates of biotic diversity and endemism for Madagascar. *Oryx* **39:** 73–77.

Gregorič M, Blackledge TA, Agnarsson I, Kuntner M. 2015. A molecular phylogeny of bark spiders reveals new species from Africa and Madagascar (Araneae: Araneidae: *Caerostris*). *Journal of Arachnology* **43:** 293–312.

Gullan PJ, Kaydan, M, Hardy NB. 2010. Molecular phylogeny and species recognition in the mealybug genus *Ferrisia* Fullaway (Hemiptera: Pseudococcidae). *Systematic Entomology* **35:** 329–339.

Hausdorf B, Hennig C. 2010. Species delimitation using dominant and codominant multilocus markers. *Systematic Biology* **59:** 491–503.

Hong-Wa C, Besnard G. 2014. Species limits and diversification in the Madagascar olive (*Noronhia*, Oleaceae). *Botanical Journal of the Linnean Society* **174:** 141–161.

Hotaling S, Foley ME, Lawrence NM, Bocanegra J, Blanco MB, Rasoloarison R, Kappeler PM, Barrett MA, Yoder AD, Weisrock DW. 2016. Species discovery and validation in a cryptic radiation of endangered primates: coalescent-based species delimitation in Madagascar's mouse lemurs. *Molecular Ecology* **25:** 2029–2045.

Karanovic T, Djurakic M, Eberhard SM. 2016. Cryptic apecies or inadequate taxonomy? Implementation of 2D geometric morphometrics based on integumental organs as landmarks for delimitation and description of copepod taxa. *Systematic Biology* **65:** 304-327.

Lumley LM, Sperling FA. 2010. Integrating morphology and mitochondrial DNA for species delimitation within the spruce budworm (*Choristoneura fumiferana*) cryptic species complex (Lepidoptera: Tortricidae). *Systematic Entomology* **35:** 416–428.

Mehta RS, Bryant D, Rosenberg NA. 2016. The probability of monophyly of a sample of gene lineages on a species tree. *Proceedings of the National Academy of Sciences* **113:** 8002–8009.

Pante E, Schoelinck C, Puillandre N. 2014. From integrative taxonomy to species description: one step beyond. *Systematic Biology* **64:** 152–160.

Perrier de la Bâthie H. 1951. *Hypericaceae 135^{e} Famille. Flore de Madagascar et des Comores*. Firmin-Didot et Cie, Paris.

de Queiroz K. 1998. The general lineage concept of species, species criteria, and the process of speciation. In: Howard DJ, Berlocher SH, eds. *Endless forms: species and speciation*. New York: Oxford University Press, pp. 57–75.

de Queiroz K. 2007. Species concepts and species delimitation. *Systematic Biology* **56:** 879–886.

Rato C, Harris DJ, Carranza S, Machado L, Perera A. 2016. The taxonomy of the *Tarentola mauritanica* species complex (Gekkota: Phyllodactylidae): Bayesian species delimitation supports six candidate species. *Molecular Phylogenetics and Evolution* **94:** 271–278.

Ray S, Lindsay BG. 2005. The topography of multivariate normal mixtures. *Annals of Statistics* **33:** 2042–2065.

Ruane S, Burbrink FT, Randriamahatantsoa B, Raxworthy CJ. 2016. The cat-eyed snakes of Madagascar: phylogeny and description of a new species of *Madagascarophis* (Serpentes: Lamprophiidae) from the tsingy of Ankarana. *Copeia* **104:** 712–721.

Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8:** 289–317.

Scrucca L, Raftery AE. 2015. Improved initialisation of model-based morphospeciesing using Gaussian hierarchical partitions. *Advances in Data Analysis and Classification* **9:** 447–460.

Sites JW Jr, Marshall JC. 2004. Operational criteria for delimiting species. *Annual Review of Ecology, Evolution, and Systematics* **35:** 199–227.

Tan DS, Ang Y, Lim GS, Ismai, MRB, Meier R. 2010. From 'cryptic species' to integrative taxonomy: an iterative process involving DNA sequences, morphology, and behaviour leads to the resurrection of *Sepsis pyrrhosoma* (Sepsidae: Diptera). *Zoologica Scripta* **39:** 51–61.

Thiers B. [continuously updated]. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/science/ih/.

Will KW, Mishler BD, Wheeler QD. 2005. The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology* **54:** 844–851.

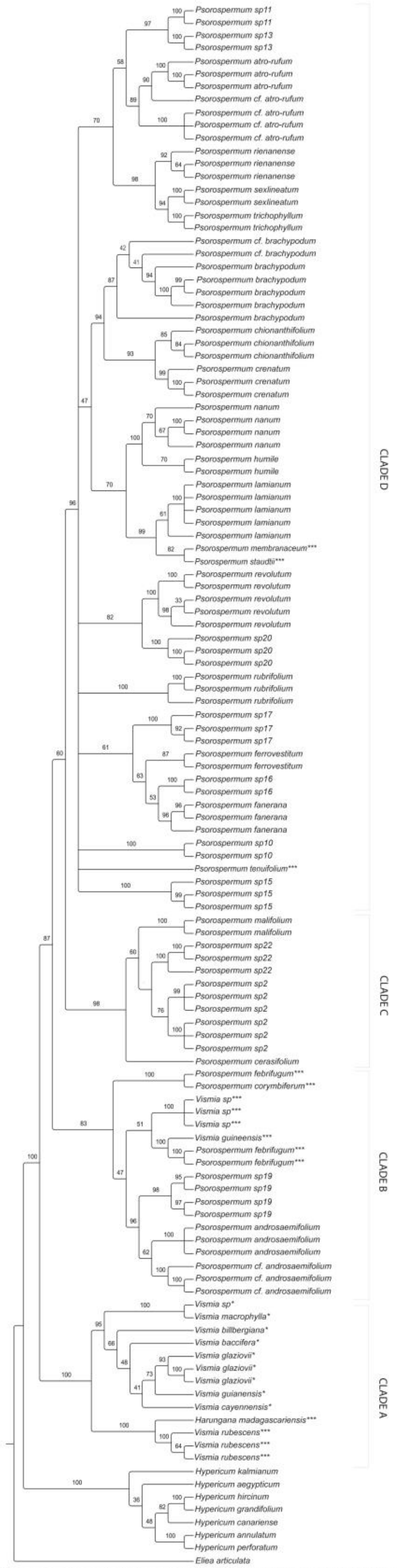
Xanthopoulos P, Pardalos PM, Trafalis TB. 2013. Linear discriminant analysis. In *Robust Data Mining.* Springer, New York, pp. 27–33.

Yeates DK, Seago A, Nelson L, Cameron SL, Joseph LEO, Trueman JW. 2011. Integrative taxonomy, or iterative taxonomy? *Systematic Entomology* **36:** 209–217.

Zapata F, Jiménez I. 2012. Species delimitation: inferring gaps in morphology across geography. *Systematic Biology* **61:** 179–194.

Zimkus BM, Lawson LP, Barej MF, Barratt CD, Channing A, Dash KM, Dehling JM, Du Preez L, Gehring PS, Greenbaum E, Gvoždík V. 2017. Leapfrogging into new territory: how Mascarene ridged frogs diversified across Africa and Madagascar to maintain their ecological niche. *Molecular Phylogenetics and Evolution* **106:** 254–269.

APPENDIX A



Maximum likelihood tree for *Psorospermum* based on the combined chloroplast DNA and internal transcribed spacer (ITS) dataset. Numbers above the branches are bootstrap support.*** = African species; *= American species.

APPENDIX B

Use of the cosine square in Principal Component Analysis:

When variables are projected to the PCA scatterplot, they are standardized and have unit lengths, therefore they can be contained inside a circle of unit length, the correlation circle. Since the variables can be mapped as projected vectors inside the correlation circle, they are confidently represented in the scatterplot surface plane when their vectors are close to that correlation circle. The principal components (PCs) explain most of variance in the dataset and the observations are positioned according to the coordinates of the variables (the variable points), thus $\cos^2$ corresponds to the square of the cosine of the angle between the straight line going through the origin and the variable point in the morphospace and the straight line going through the origin and the projection of that point onto the scatterplot (Abdi & William, 2010) (Fig. B1). The closer $\cos^2$ is to 1, the closer the projection of the variable to the PC scatterplot surface plane. Thus, variables with higher $\cos^2$ values are better represented than those with lower values. I averaged the $\cos^2$ values of the 10 variables with highest $\cos^2$ values, and retained only the variables with values higher than the calculated average, and the collinearity of these variables was also verified since the EM algorithm implemented in GMM performs poorly with collinear variables (Fraley & Raftery, 1998). The analysis was performed with the R packages MASS v7.3-45; FactoMineR v1.35 and caret v6.0-73.

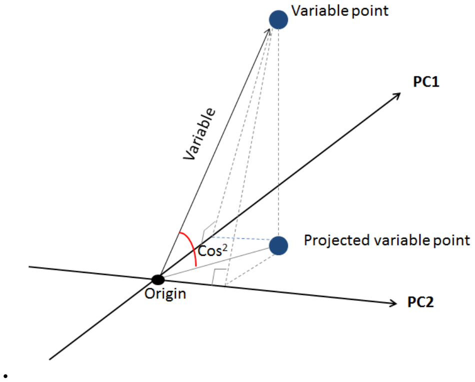


Figure B1. Association between the Principal Components, the observations and the variables

General Mixture Models GMM analyses

In the GMM analyses, models are the statistical tools for estimating the probability densities in the data set. Since GMM assumes that the data are generated by a finite mixture of probability distributions and that each morphospecies has a different multivariate probability density distribution (Fraley *et al.*, 2012, Fraley & Raftery, 2002; Scrucca *et al.*, 2015), each morphospecies is considered a multidimensional Gaussian with its own mean and covariance (Scrucca *et al.*, 2015). Thus the probability density function model in a mixture of K Gaussians is equal to the sum of the prior probability (W) of each individual Gaussian (k) multiplied by the covariance and mean of the morphospecies.:

$$p(x) = \sum_{k=1}^{K} Wk * p(x|mean\, k, covk)$$

However, in a multidimensional morphospace, each identified morphospecies has its own volume, shape, form and orientation. GMM tools incorporate models that estimate those geometric features (Fraley *et al.*, 2012; Scrucca *et al.*, 2015). The models parameters are: the scalar determining the volume of the morphospecies, k (λ_k), the matrix of suitable dimension I, and the matrix of eigenvectors, D_k (rotation), determining the orientation of the morphospace of the morphospecies, and is proportional to the diagonal matrix of eigenvalues (stretching) determining its shape A_k (Fraley *et al.*, 2012; Scrucca *et al.*, 2015), and the models are calculated as:

$$\Sigma= \lambda_k I_k D_k A_k$$

The models vary depending on the parameters included (I, λ , D, A). For example, the simplest model is $\Sigma= \lambda$ I, meaning that the morphospace of the morphospecies recognized are all spherical and have equal volumes (Appendix S3).

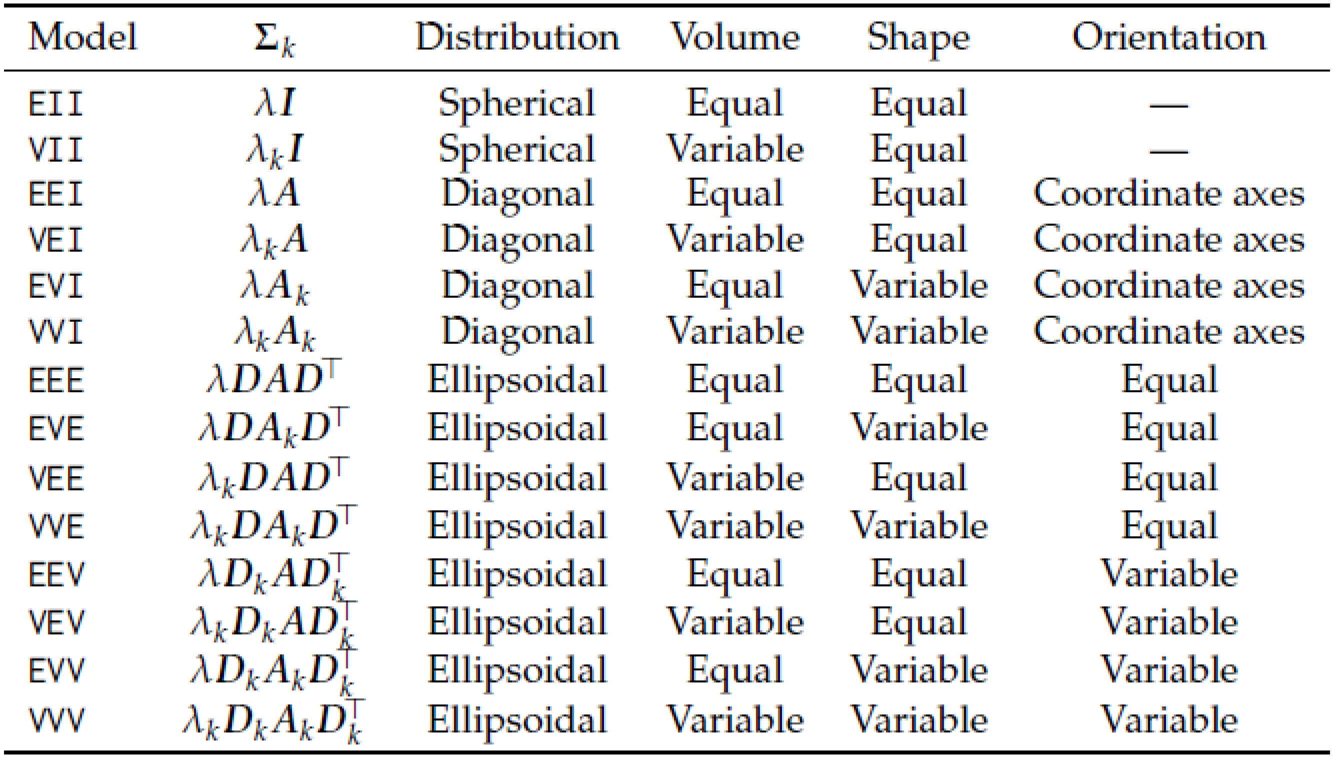
Appendix Table S1: List of morphological characters measured in the integrative taxonomy of *Psorospermum* and the morphological characters selected by PCA per clade for the General Mixture Model analysis are marked with "X". Tricho: Trichophyllum; Brachy: Brachypodum; Chio: Chionanthifolium; Revo: Revolutum; ferro: Ferrovestitum; Cerasi: Cerasifolium; Andro: Androsaemifolium

	Clade Atro-rufum	**Clade Tricho**	**Clade Brachy**	**Clade Chio**	**Clade Nanum**	**Clade Revo**	**Clade Ferro**	**Clade Cerasi**	**Clade Andro**
Pedicel length		yes				X			
Sepal length						X		X	
Number of sepal glands			X					X	
Petal length						X	X		
Number of petal glands					X				
Staminode length					X				
Filament length		X	X		X			X	
Number of anthers									
Number of anther glands									
Ovary length									
Style length								X	
Number of ovules per carpel	X		X		X				
Lamina angle of apex	X	X			X	X		X	

Lamina length	X			X	X		X		X
Lamina width	X			X			X		X
Lamina length from the base to the widest point				X					
Lamina surface area	X			X	X		X	X	X
Lamina shape	X								X
Number glands per cm^2 on the lamina surface		X				X		X	
Number of secondary veins	X			X			X		X
Number of hairs per cm^2 on the lamina surface		X							
Petiole length		X				X			
Cotyledon length			X						
Cotyledon width	X								
Cotyledon surface area									
Radicle length			X						
Radicle width									
Ratio cotyledon:radicle				X					X

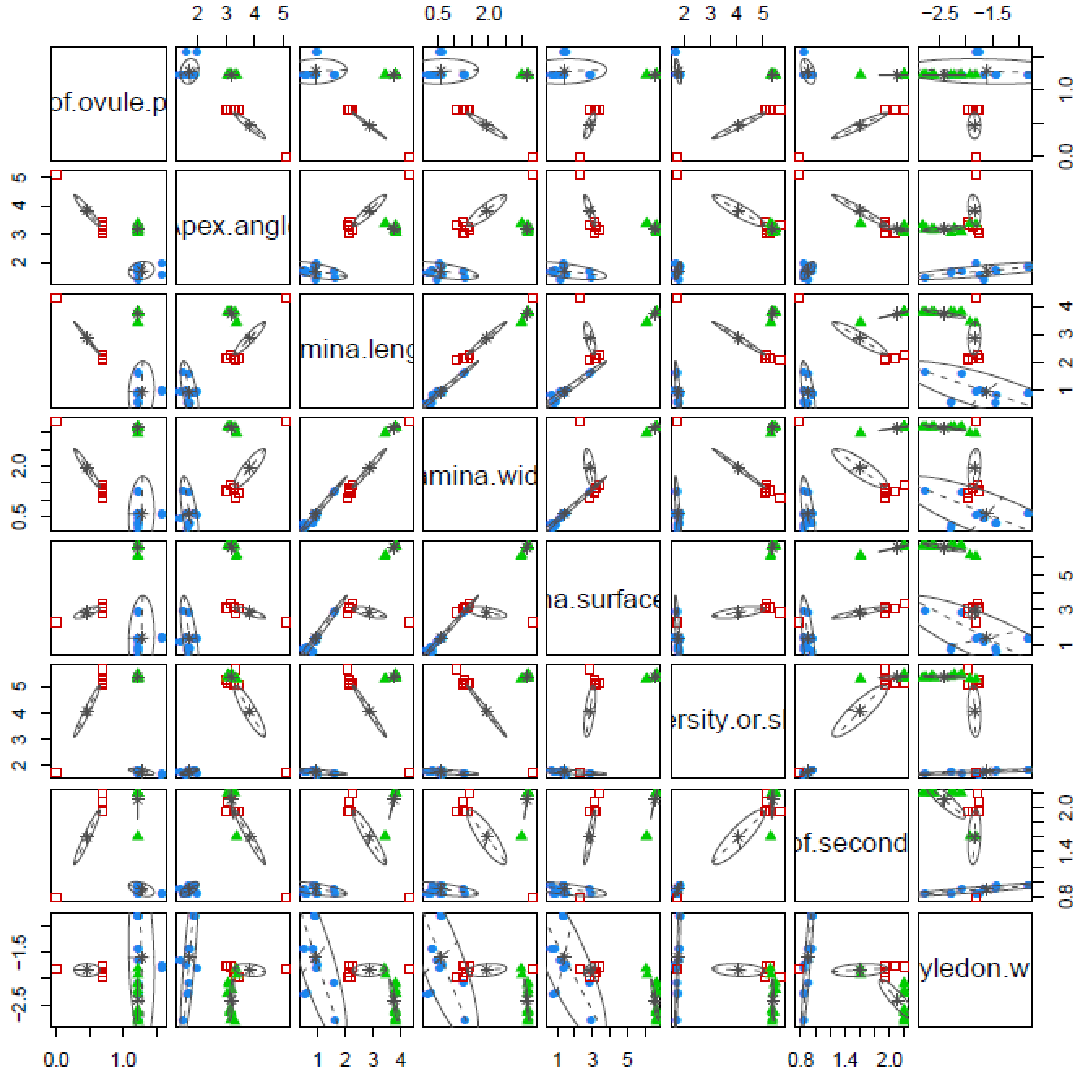
length									
Angle of cotyledon to radicle									

Appendix Table S2. Different types of models and their parameters in Mclust. I did not create this table, it is retrieved directly from Scrucca *et al.* (2015).

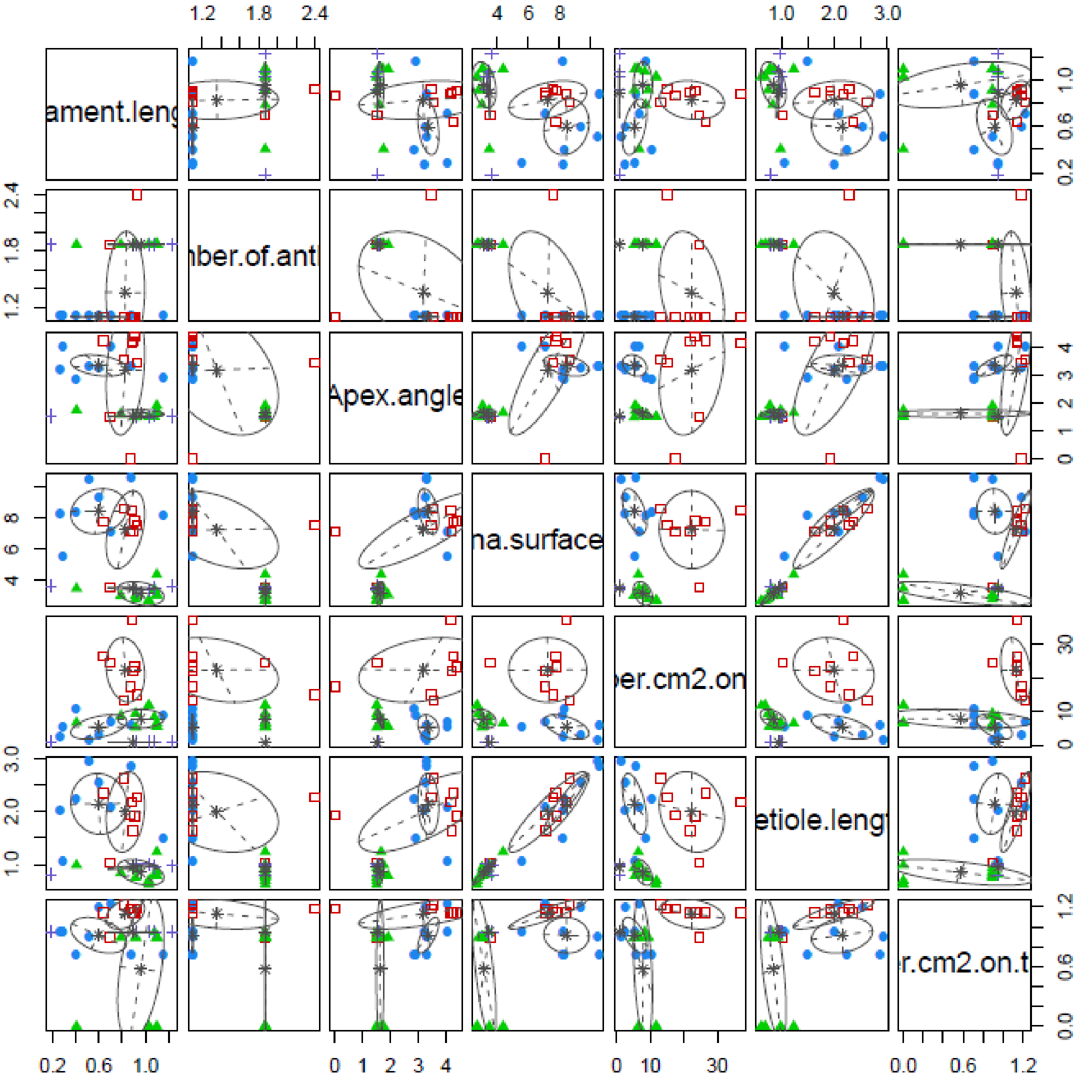


Model	Σ_k	Distribution	Volume	Shape	Orientation
EII	λI	Spherical	Equal	Equal	—
VII	$\lambda_k I$	Spherical	Variable	Equal	—
EEI	λA	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k A$	Diagonal	Variable	Equal	Coordinate axes
EVI	λA_k	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k A_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda D A D^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda D A_k D^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k D A D^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k D A_k D^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda D_k A D_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k D_k A D_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda D_k A_k D_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k D_k A_k D_k^\top$	Ellipsoidal	Variable	Variable	Variable

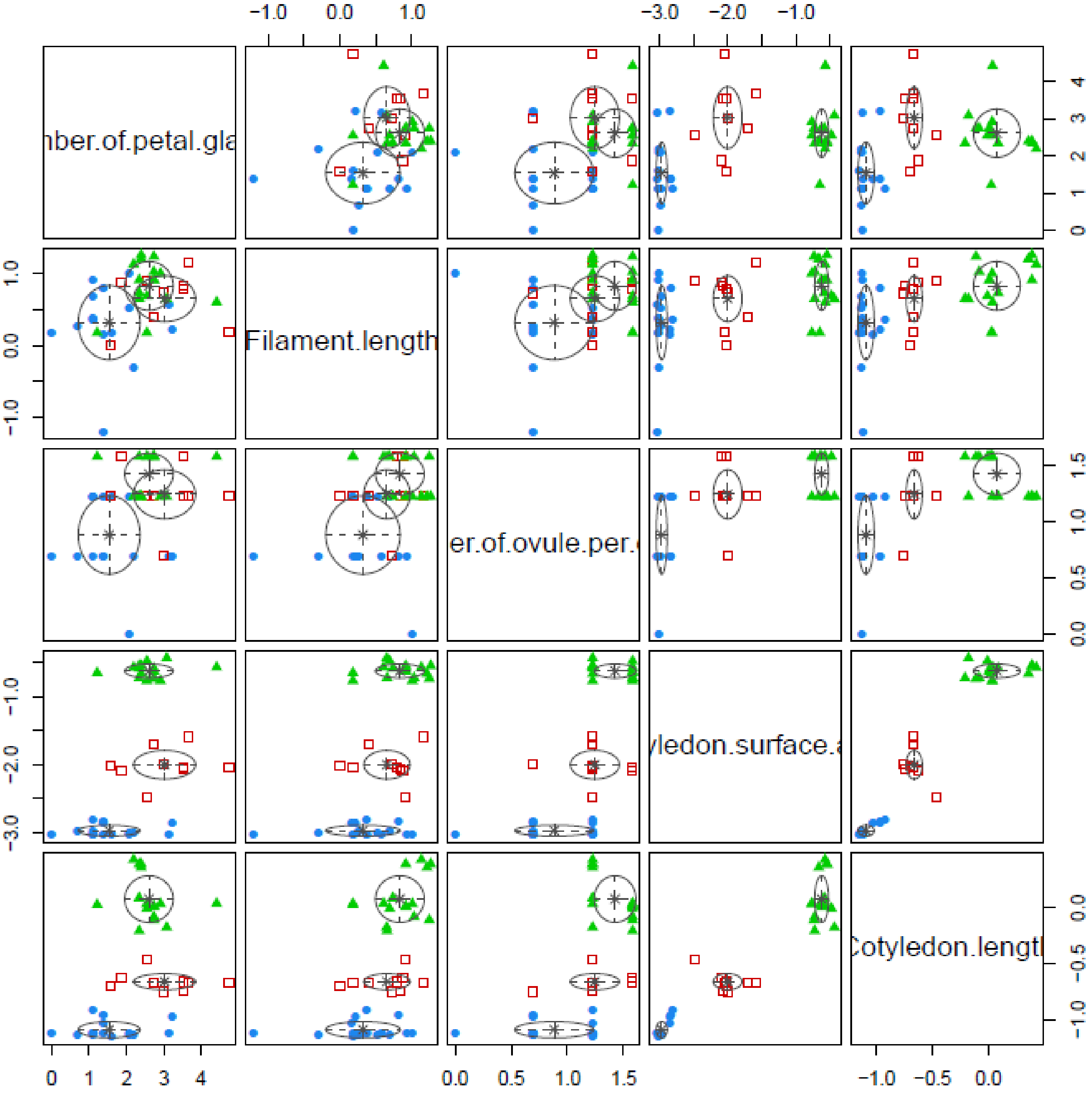
Appendix S4: Pairwise two dimensional scatterplots of the dataset of the Atro-rufum clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



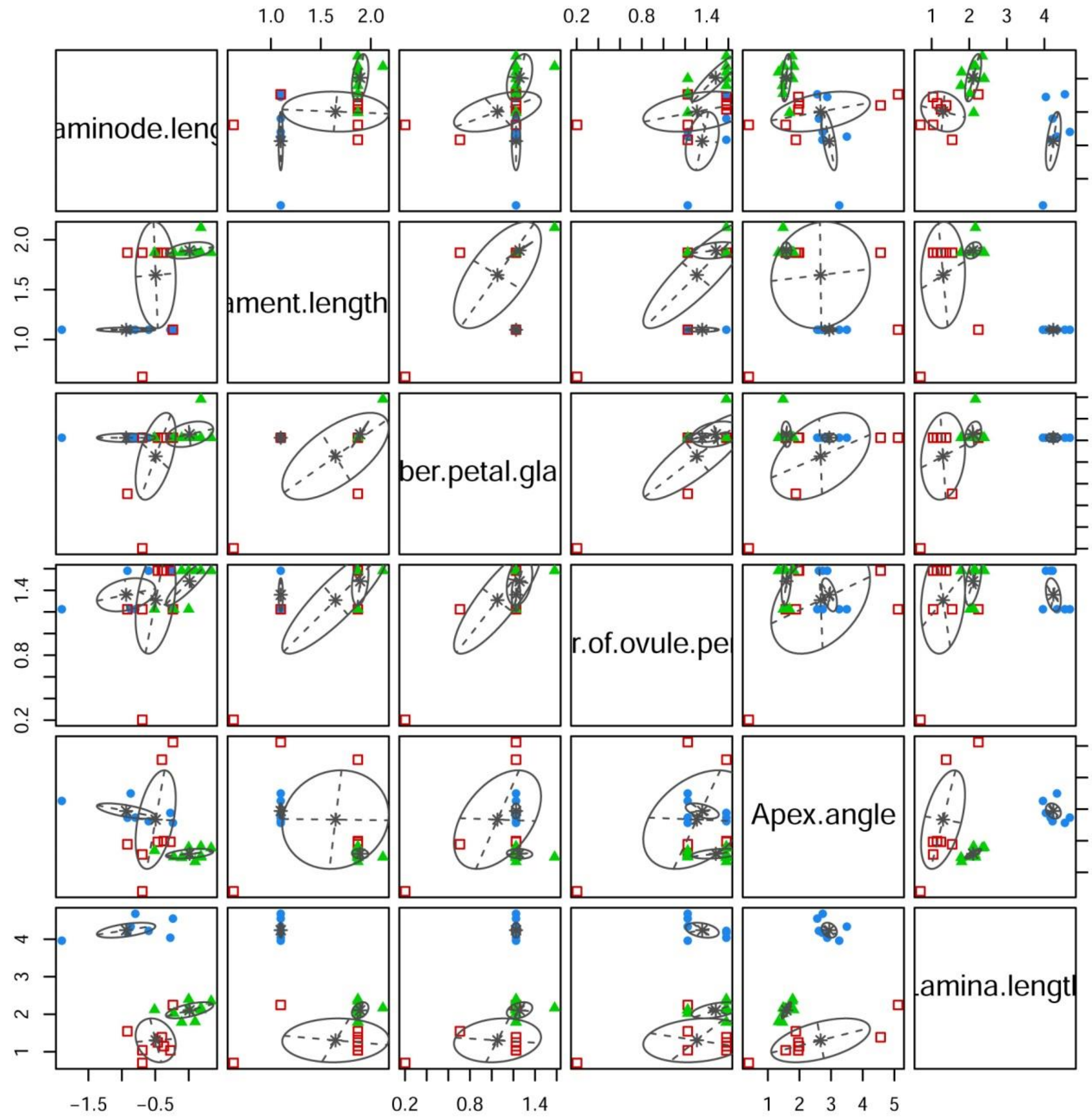
Appendix S5: Pairwise two dimensional scatterplots of the dataset of the Trichophyllum clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



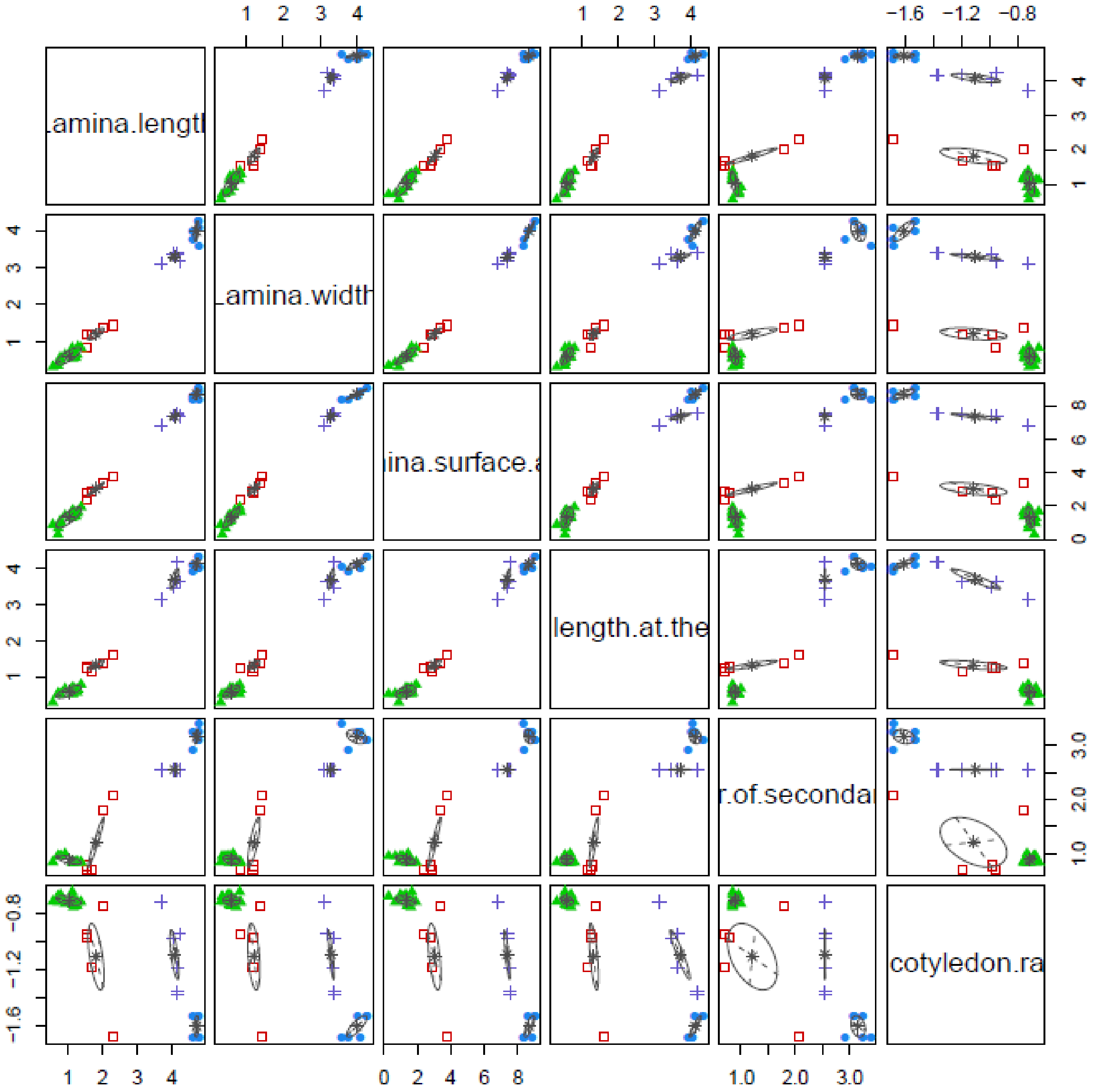
Appendix S6: Pairwise two dimensional scatterplots of the dataset of the Brachypodum clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



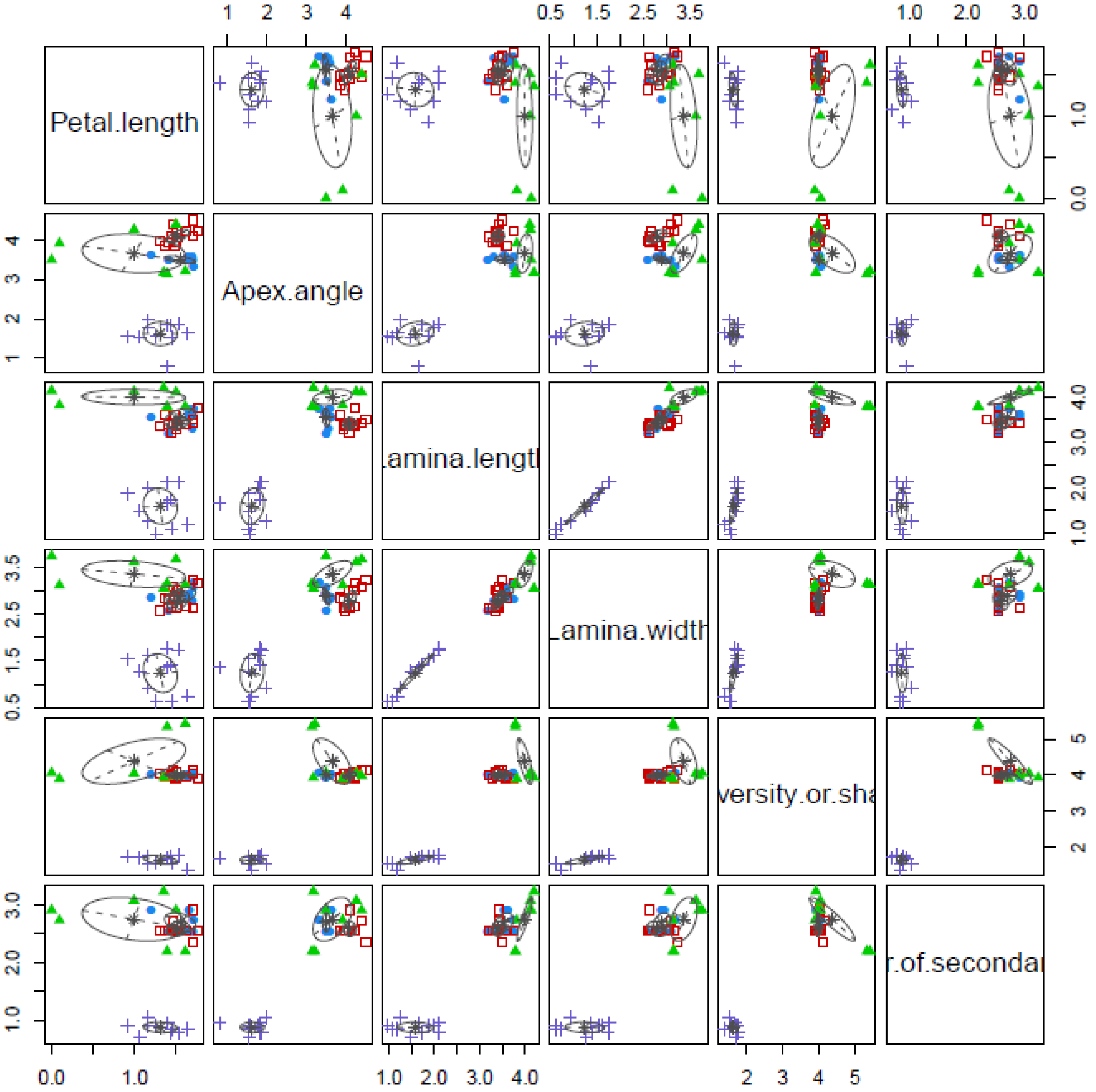
Appendix S7: Pairwise two dimensional scatterplots of the dataset of the Nanum clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



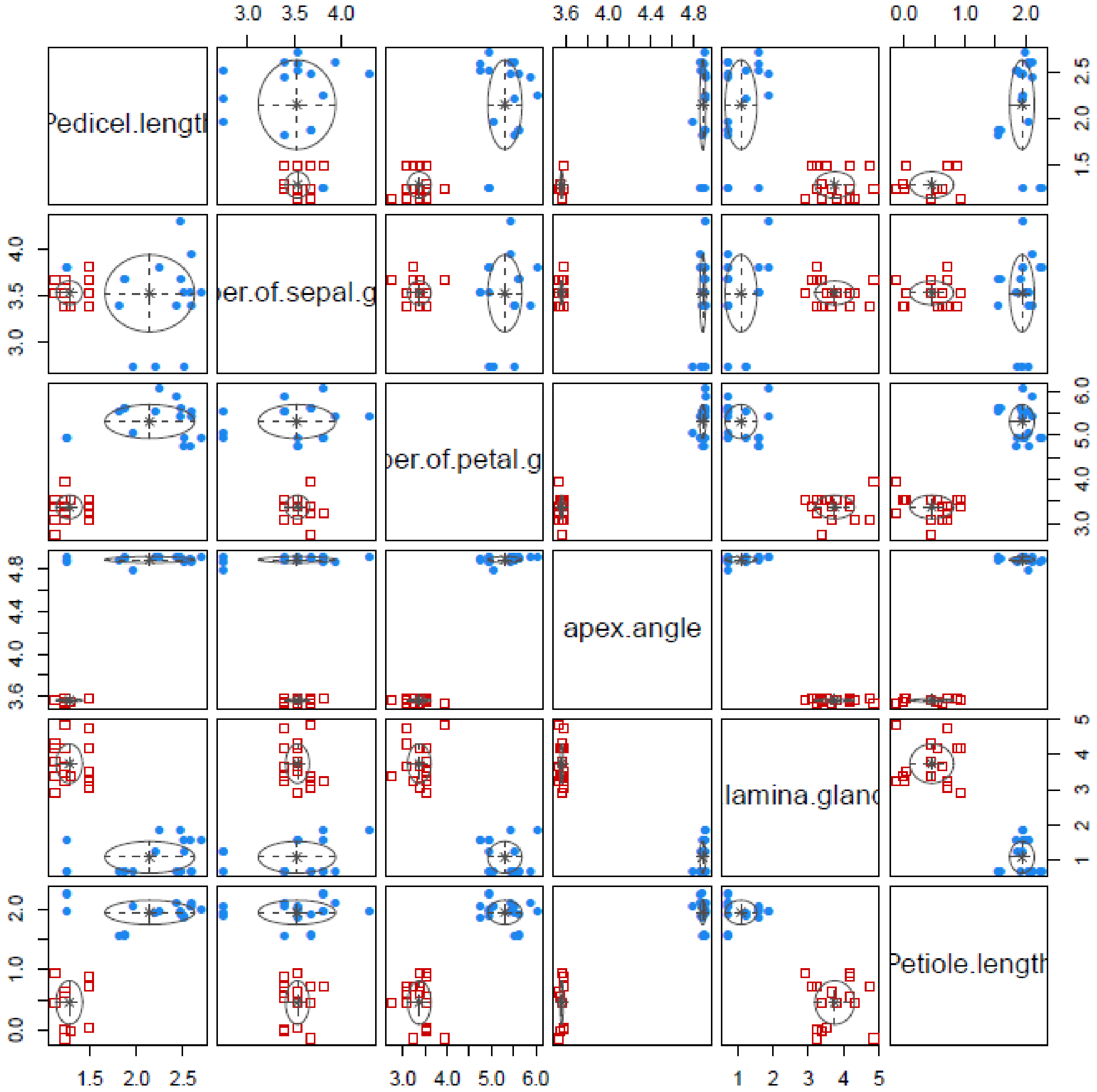
Appendix S8: Pairwise two dimensional scatterplots of the Chionanthifolium clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



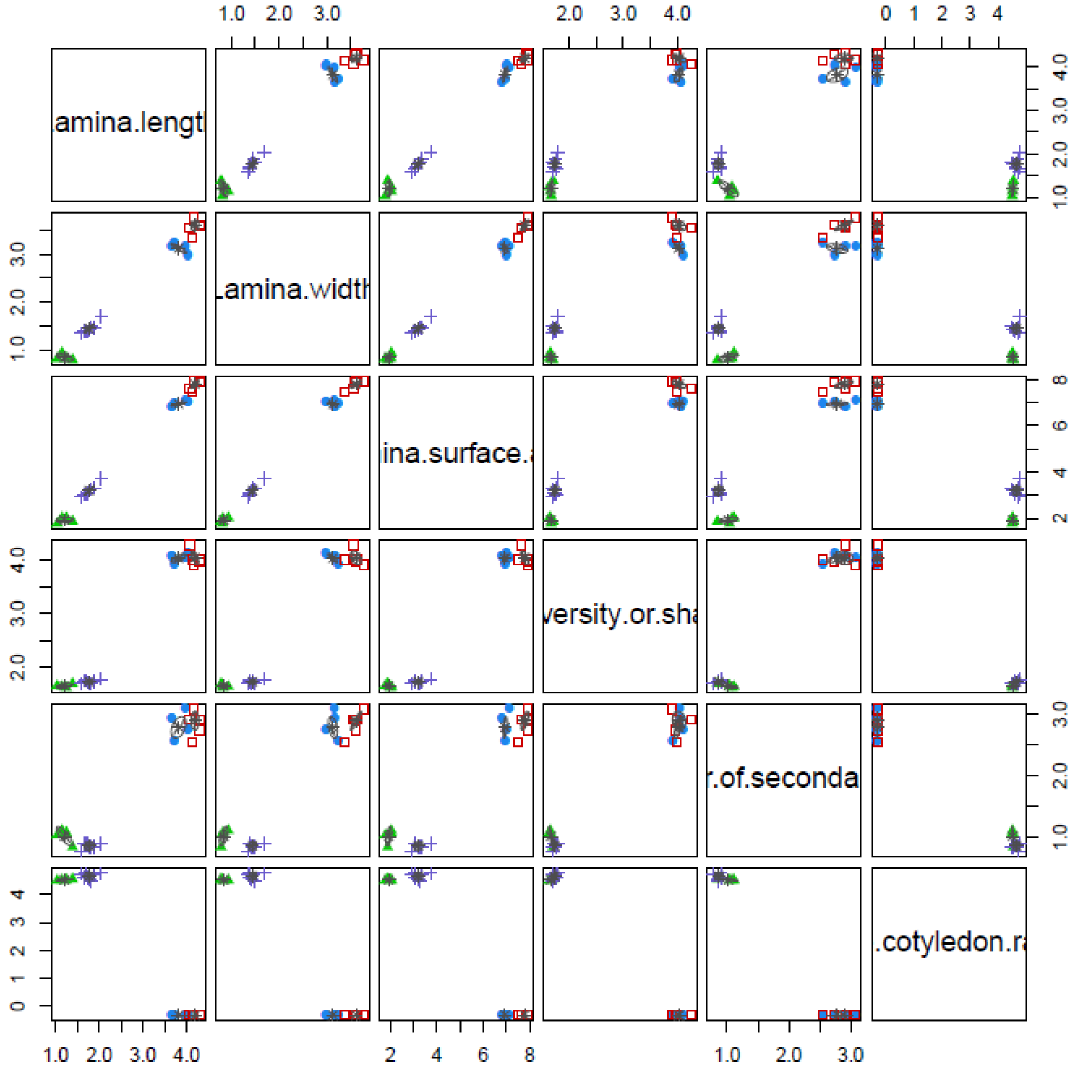
Appendix S9: Pairwise two dimensional scatterplots of the Ferrovestitum clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



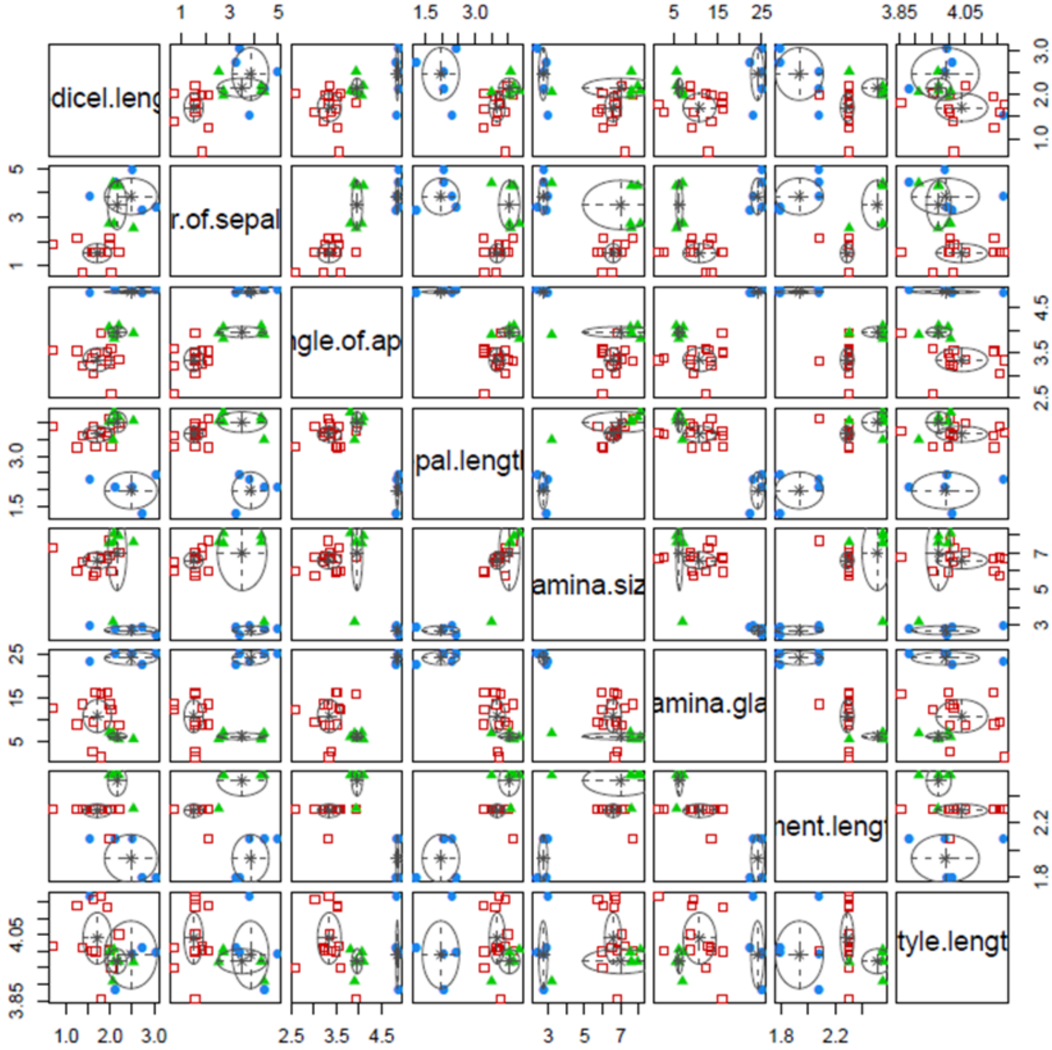
Appendix S10: Pairwise two dimensional scatterplots of the Revolutum clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



Appendix S11: Pairwise two dimensional scatterplots of the Androsaemifolium clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



Appendix S12: Pairwise two dimensional scatterplots of the dataset of the Cerasifolium clade. Ellipses superimposed on the plot correspond to the covariances of the components. Individuals in different morphospecies are indicated by different symbols and colors.



SUPPLEMENTARY MATERIALS

CLADE ATRO-RUFUM

GMM morphospecies in the Atro-rufum clade

A total of 35 herbarium specimens are used in this particular dataset. Clade Atro-rufum has 58% bootstrap, and the 11 specimens from the molecular study contained are assignable to 4 morphogroups (*Psorospermum atro-rufum, P. cf. atro-rufum, P. sp11* and *P. sp19*) (Fig. S1).

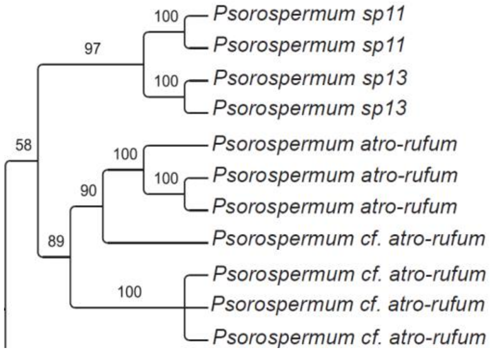


Fig. S1. Expanded clade Atro-rufum retrieved form the ML molecular phylogenetic analysis of *Psorospermum*. Numbers above the braches are bootstrap support values

These specimens were added to the 24 herbarium specimens that made up these 4 morphogroups. The four morphogroups were tested with LDA (Wilks's Lambda = 5e-^{7}, p-value <0.0001). There is no overlap of the values of the discriminant functions of the 4 morphogroups (Fig. S2).

The best model of the GMM analysis yielded three morphospecies. Fig. S5 summarizes the output of the Mclust models (table on the left), and highlights the identification of three morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model, VEV (ellipsoidal, equal shape), is the one that has the highest BIC value, and it is indicated by the horizontal dotted line on the y-axis (Fig. S5); the number of morphospecies suggested by that model is indicated by the vertical dotted line on the x-axis (Fig. S5). Eight characters significantly influence the loadings of the observations in the PCA vector scatterplot: number of ovules per carpel, lamina length, width, apex angle, surface area, shape, secondary veins, and cotyledon width (Fig. S3; Appendix Table S1 for complete list). These variables are used to perform the GMM analyses. Morphospecies in two-dimensional pairwise comparison of variables are shown in a two-dimensional (2-D) morphospace scatterplot and an example is given in Fig. S4. All pairwise combinations of the eight variables are shown in 2-D scatterplots in Appendix S4.

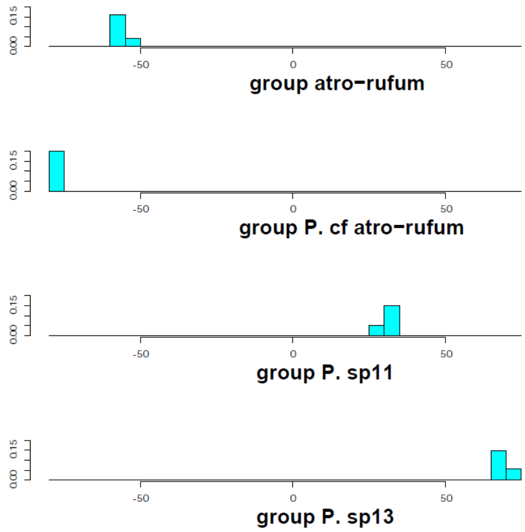


Figure S2. Histogram of values of the first discriminant function for the samples from the four morphogroups in the Atro-rufum clade.

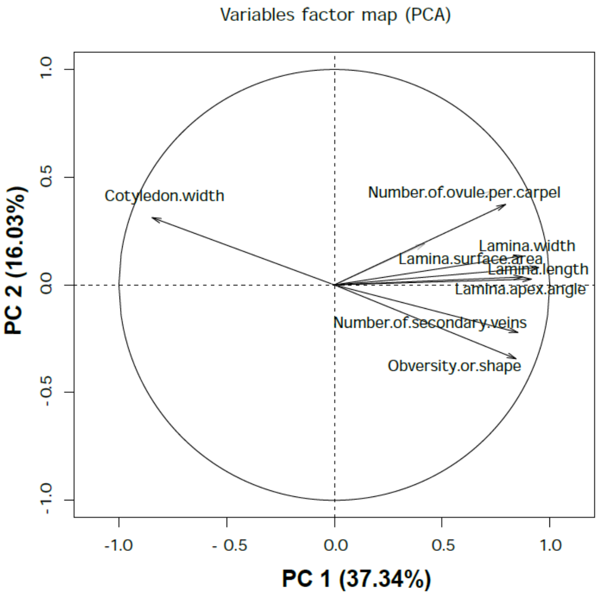


Figure S3. Eight morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in the Atro-rufum clade.

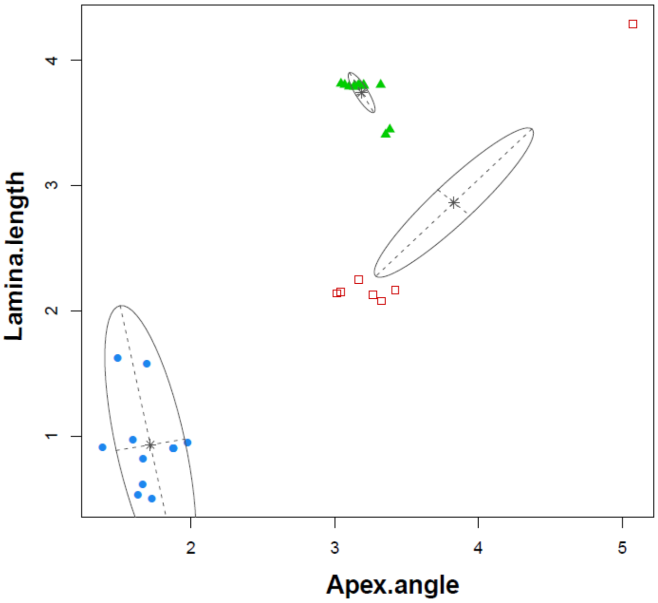


Figure S4. Bivariate scatterplot in the Atro-rufum clade. The ellipses superimposed on the plot correspond to the covariances of the morphospecies. Individuals in different morphospecies are indicated by different symbols and colors, filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3.

Figure S5. Outputs of GMM analysis of the Atro-rufum clade.

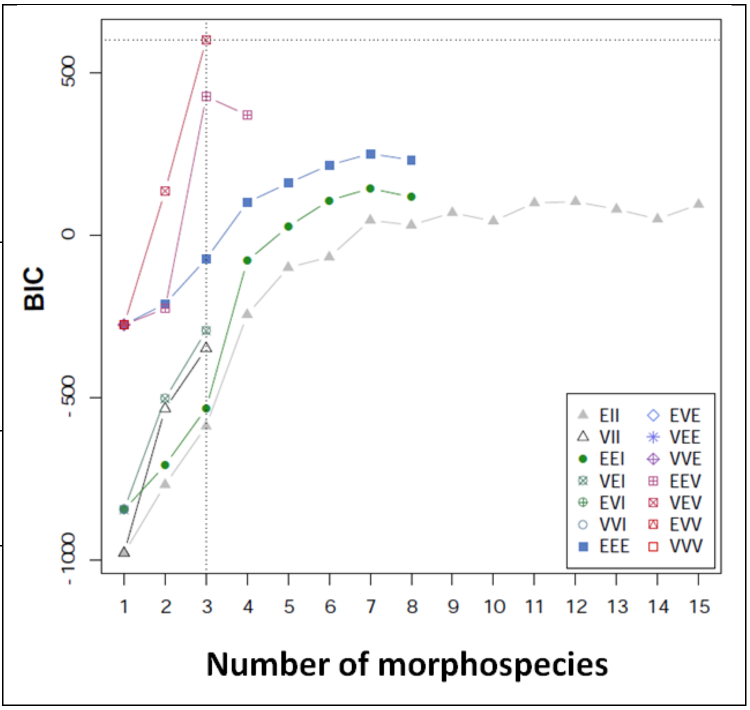
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N. ind.: number of individuals per morphospecies.

Mclust model with 3 morphospecies:

Ellipsoidal, Equal shape (VEV).
See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
514.60	35	120	602.56

Morphospecies	1	2	3
N . ind.	13	9	13



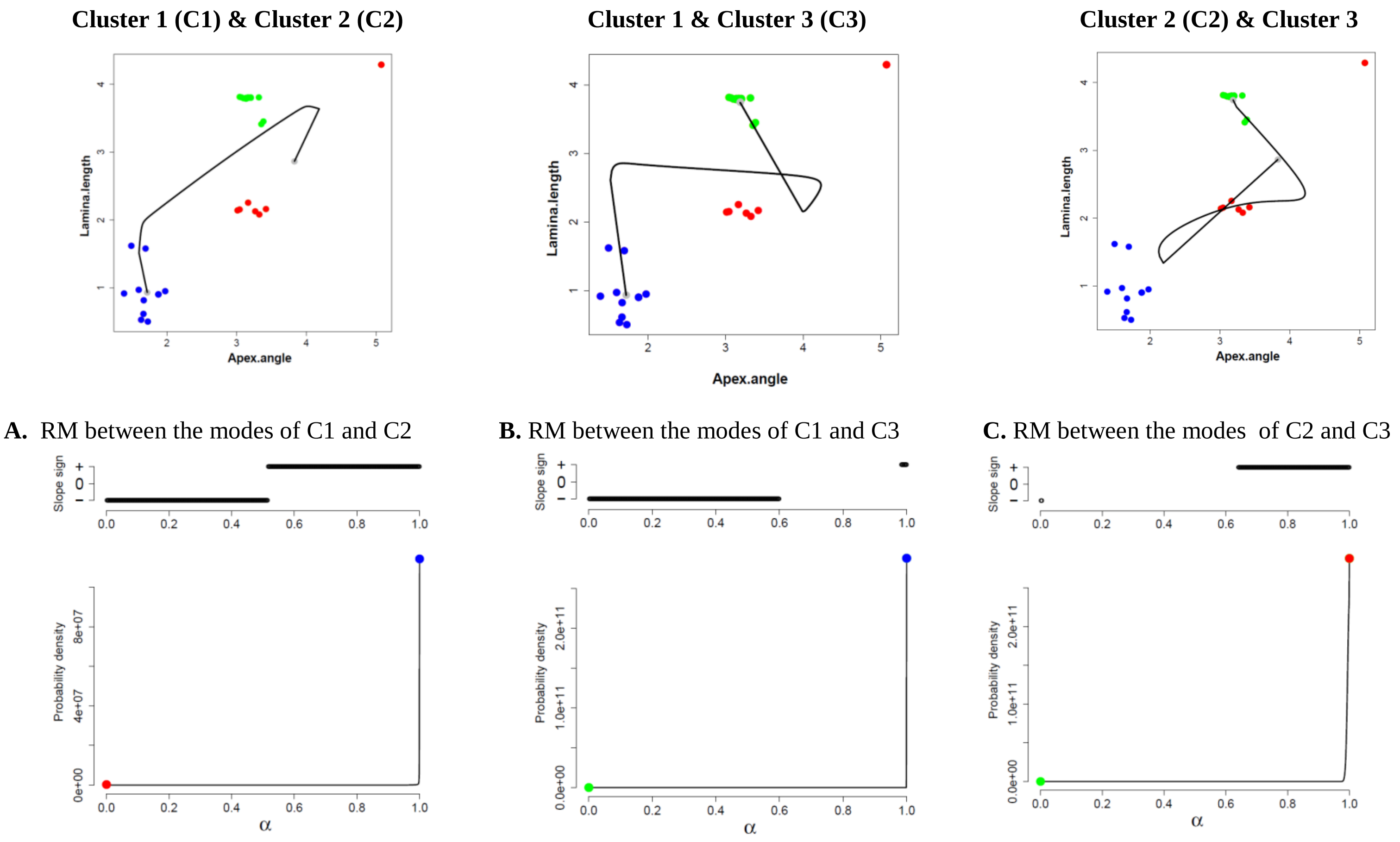
Quantification of gaps between morphospeciesin the Atro-rufum clade

The RM and the slopes of the PDFs of the three identified morphospecies are shown in Fig. S6.A-C and Fig. S6.D-F respectively. The PDFs are bimodal which indicates gaps between the morphospecies. The graph line above the PDF highlights the change of slope sign, and confirms the bimodal nature of the PDF in morphospace.

Assessment of morphological characters that separate the morphospecies in the Atro-rufum clade

The phenotypes of the three morphospecies having β* values above the cutoff value 0.9 of the proportion of tolerance regions of overlapping phenotypes (β*= 1), Fig. S7.

Fig. S8 shows that the quantiles of univariate distribution of the morphological characters from the GMM analysis (points) fit the simulation of the non-overlapping phenotypes of all morphospecies, the character cotyledon width overlaps in C1 and C2, and the number of ovules per carpel and cotyledon width overlap in C1 and C2, and C1 and C3 respectively.



Cluster 1 (C1) & Cluster 2 (C2)

Cluster 1 & Cluster 3 (C3)

Cluster 2 (C2) & Cluster 3

A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

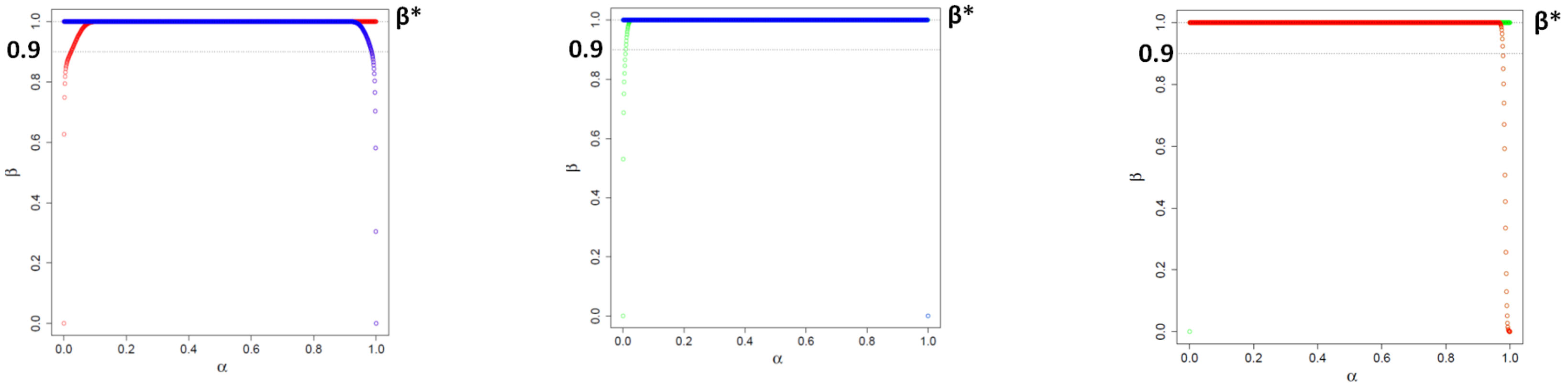
C. RM between the modes of C2 and C3

D. Slope of the PDF (C1 and C2)

E. Slope of the PDF (C1 and C3)

F. Slope of the PDF (C2 and C3)

Figure S6. Ridgeline manifold (RM) and inference of gaps between the morphospecies of the Atro-rufum clade. **A-C**: RM = continuous black line; **D- F**; blue, green and red points= modes of the morphospecies.



A. β*=1 in C1 and C2, **B.** β*=1 in C1 and C3, **C.** β*=1 in C2 and C3,

Figure S7. Value of the proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Atro-rufum clade compared to the cutoff 0.9 necessary for the two GMM morphospecies to be distinct. C1: blue, c2: red: C3: green.

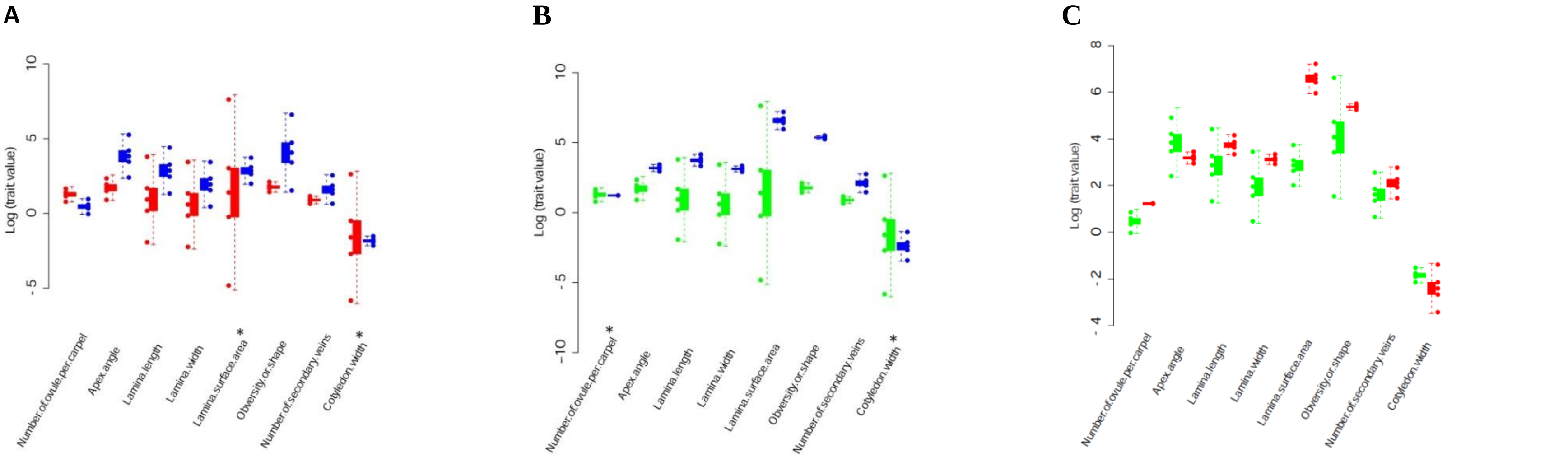


Figure S8. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in the Atro-rufum clade**.** Morphospecies = **C;** C1 (blue boxplot),C2 (red boxplot) and C3 (green boxplot). Overlapping characters are marked with asterisks.

CLADE TRICHOPHYLLUM

GMM morphospecies in the Trichophyllum clade

Clade Trichophyllum has 98% bootstrap support, and the 7 specimens from the molecular study are assignable to 3 morphogroups (*Psorospermum rienanense*, *P. sexlineatum,* and *P. trichophyllum*) (Fig. S9).

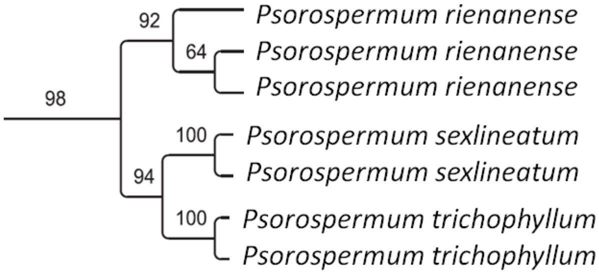


Fig. S9. Expanded clade Trichophyllum retrieved from the ML molecular phylogenetic analysis of *Psorospermum*. Numbers above the branches are bootstrap support values

These specimens were added to 25 herbarium specimens previously assigned to these 3 morphogroups. A total of 32 herbarium specimens were used in this particular dataset. The three morphogroups were confirmed with LDA (Wilks's Lambda = 0.004, p-value = 0.00017). There is no overlap of the values of the discriminant functions (Fig. S10).

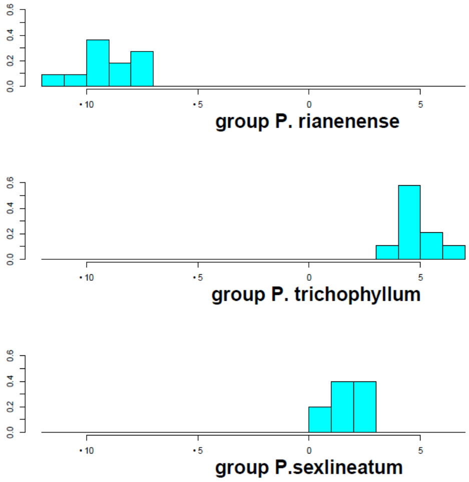


Figure S10. Histogram of values of the first discriminant function for the samples from the three morphogroups in the Trichophyllum clade.

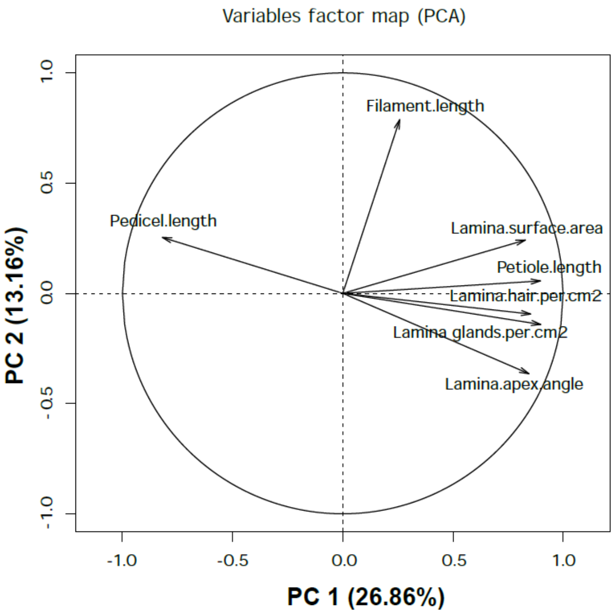


Figure S11. Seven morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in the Trichophyllum clade.

The best model of the GMM analysis yielded four morphospecies. Fig. S12 summarizes the output of the Mclust models (table on the left), and highlights the identification of four morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model, VEV (Ellipsoidal, Equal shape), is the one that has the highest BIC value, it is indicated by the horizontal dotted line on the y-axis (Fig. S12); the number of morphospecies suggested by that model is indicated by the vertical dotted line on the x-axis (Fig. S12). Seven characters influence significantly the loadings of the observations in the PCA vector scatterplot: pedicel length, filament length, lamina apex angle, surface area, number of glands per cm^2, number of hairs per cm^2, and petiole length (Fig. S11; Appendix Table S1 for complete list). These variables are used to perform the GMM analyses. Morphospecies in two-domensional pairwise comparison of variables are shown in a two-dimensional (2-D) morphospace scatterplot and an example is given in Fig. S13. All pairwise combinations of the seven variables are shown in 2-D scatterplots in Appendix S5.

Figure S12. Outputs of GMM analysis of the Trichophyllum clade.

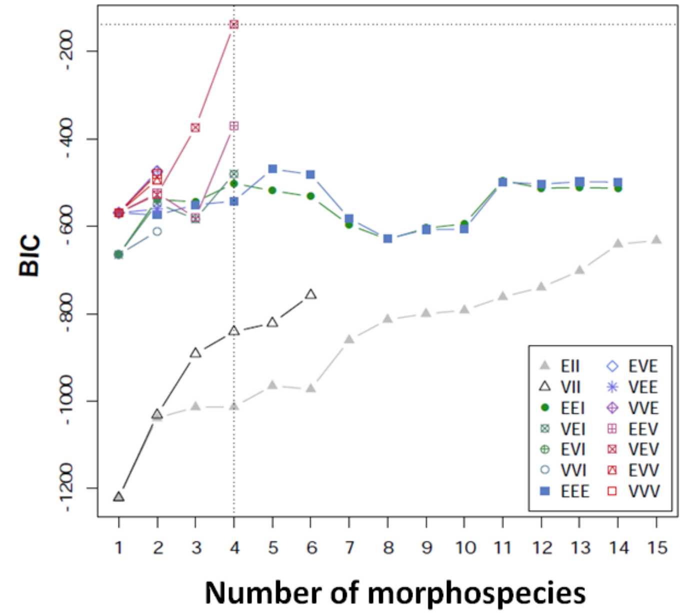
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N.ind.: number of individuals per morphospecies.

Mclust model with 4 morphospecies:

VEV (Ellipsoidal, Equal shape).
See Appendix Table S2 for the
details of the models in the legend.

Log likelihood	n	df	BIC
147.28	32	125	-338.64

Morphospecies	1	2	3	4
N . ind.	8	8	11	5



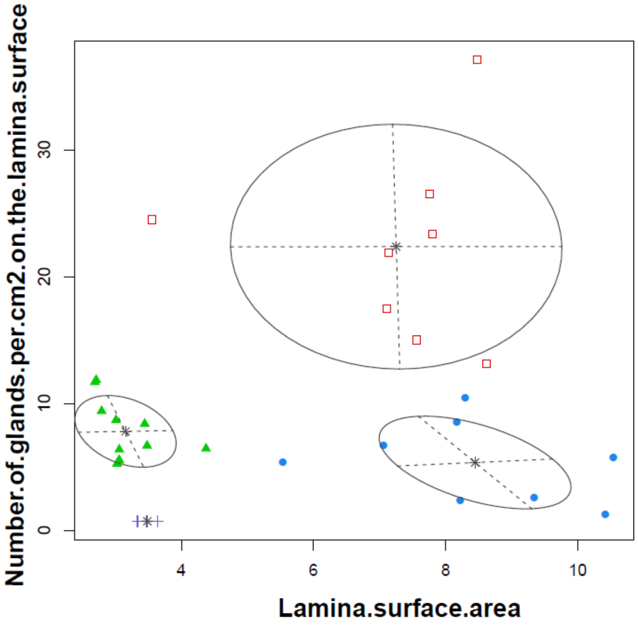
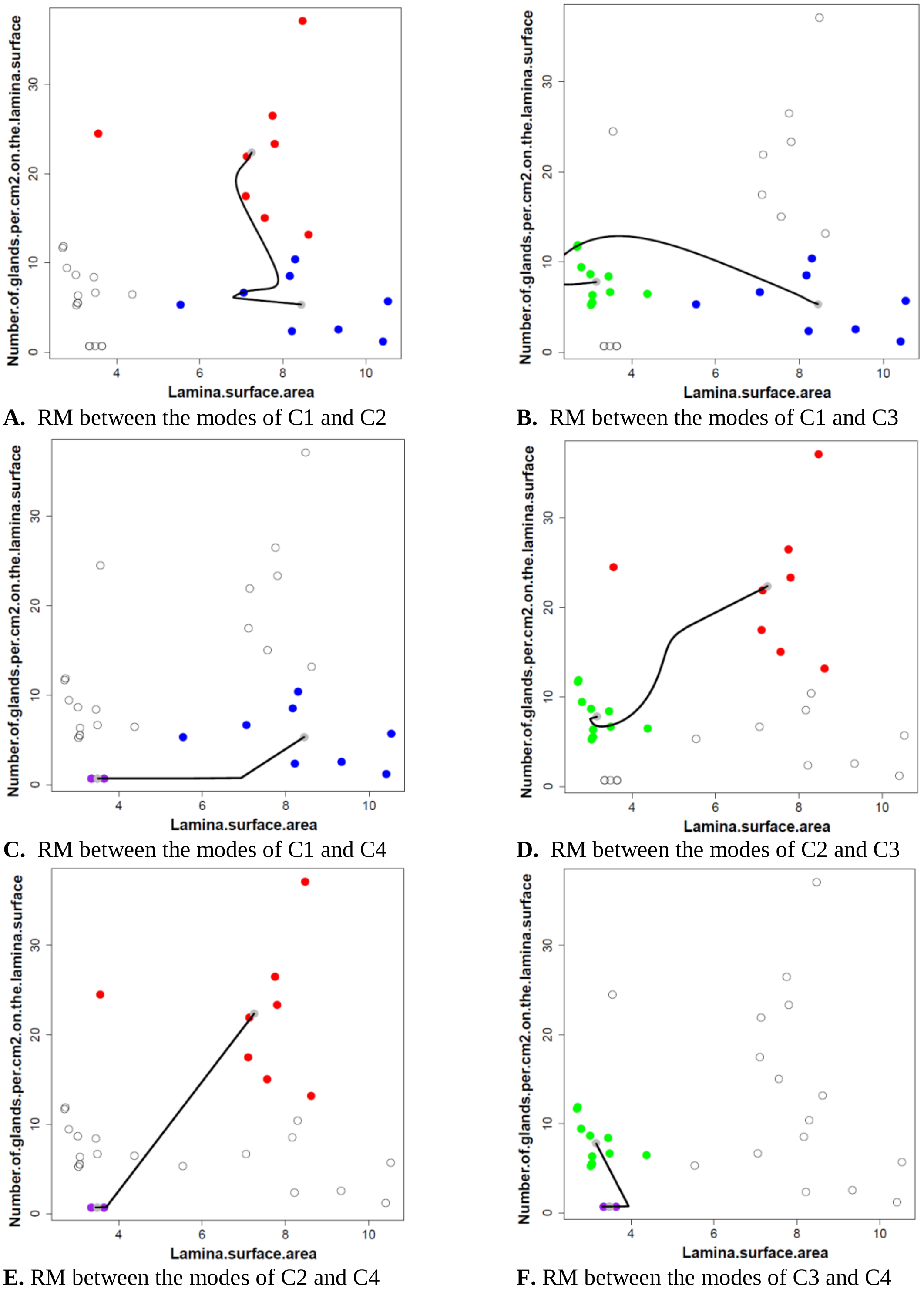


Figure S13. Bivariate scatterplot in the Trichophyllum clade. The ellipses superimposed on the plot correspond to the covariances of the morphospecies. Individuals in different morphospecies are indicated by different symbols and colors, filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3, purple cross:4.



A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

C. RM between the modes of C1 and C4

D. RM between the modes of C2 and C3

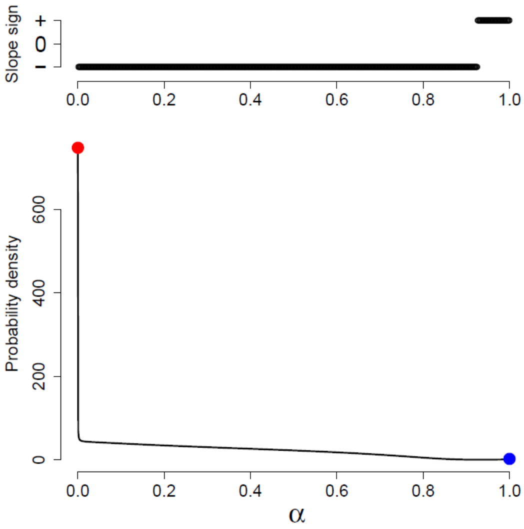
E. RM between the modes of C2 and C4

F. RM between the modes of C3 and C4

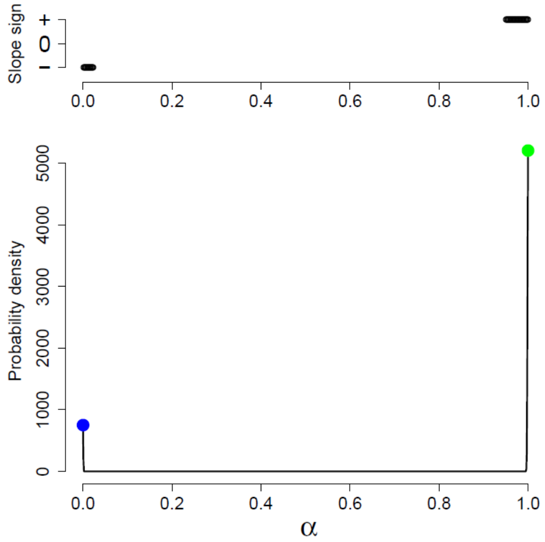
Figure S14. Ridgeline Manifold (RM) between the morphospecies in the Trichophyllum clade. RM = continuous black line; blue, red, green and purple points: morphospecies C1, C2, C3, and C4.

Quantification of gaps between morphospecies in the Trichophyllum clade

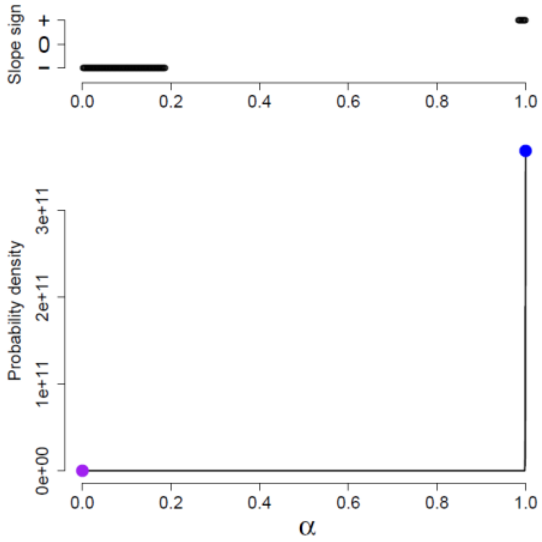
The RM and the slopes of the PDFs of the four identified morphospecies are shown in Fig. S14.A-F and Fig. S15A-F respectively. The PDFs are bimodal which indicates gaps between the morphospecies. The graph line above the PDF highlights the change of slope sign and confirms the bimodal nature of the PDF in morphospace.



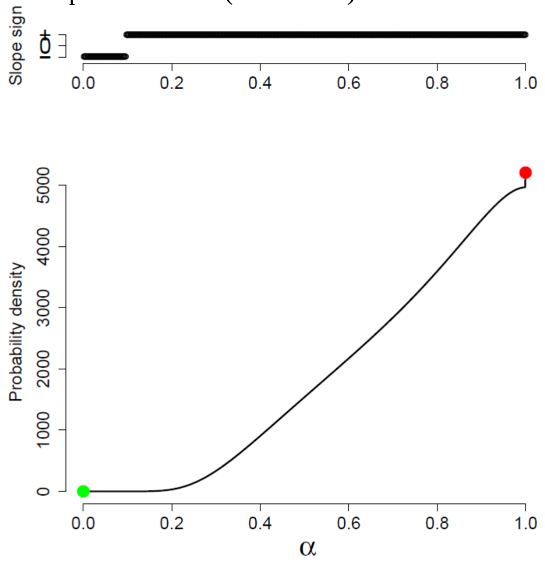
A. Slope of the PDF (C1 and C2)



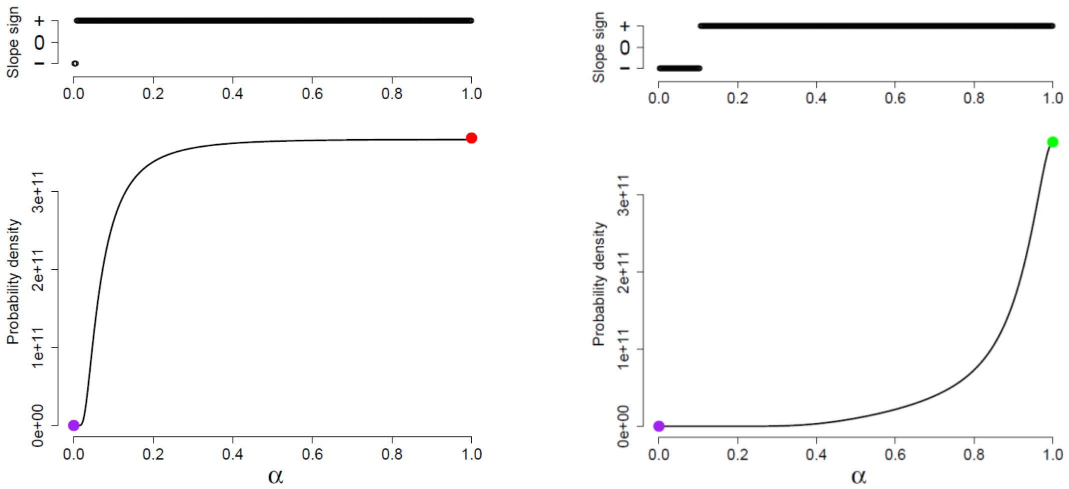
B. Slope of the PDF (C1 and C3)



C. Slope of the PDF (C1 and C4)



D. Slope of the PDF (C2 and C3)



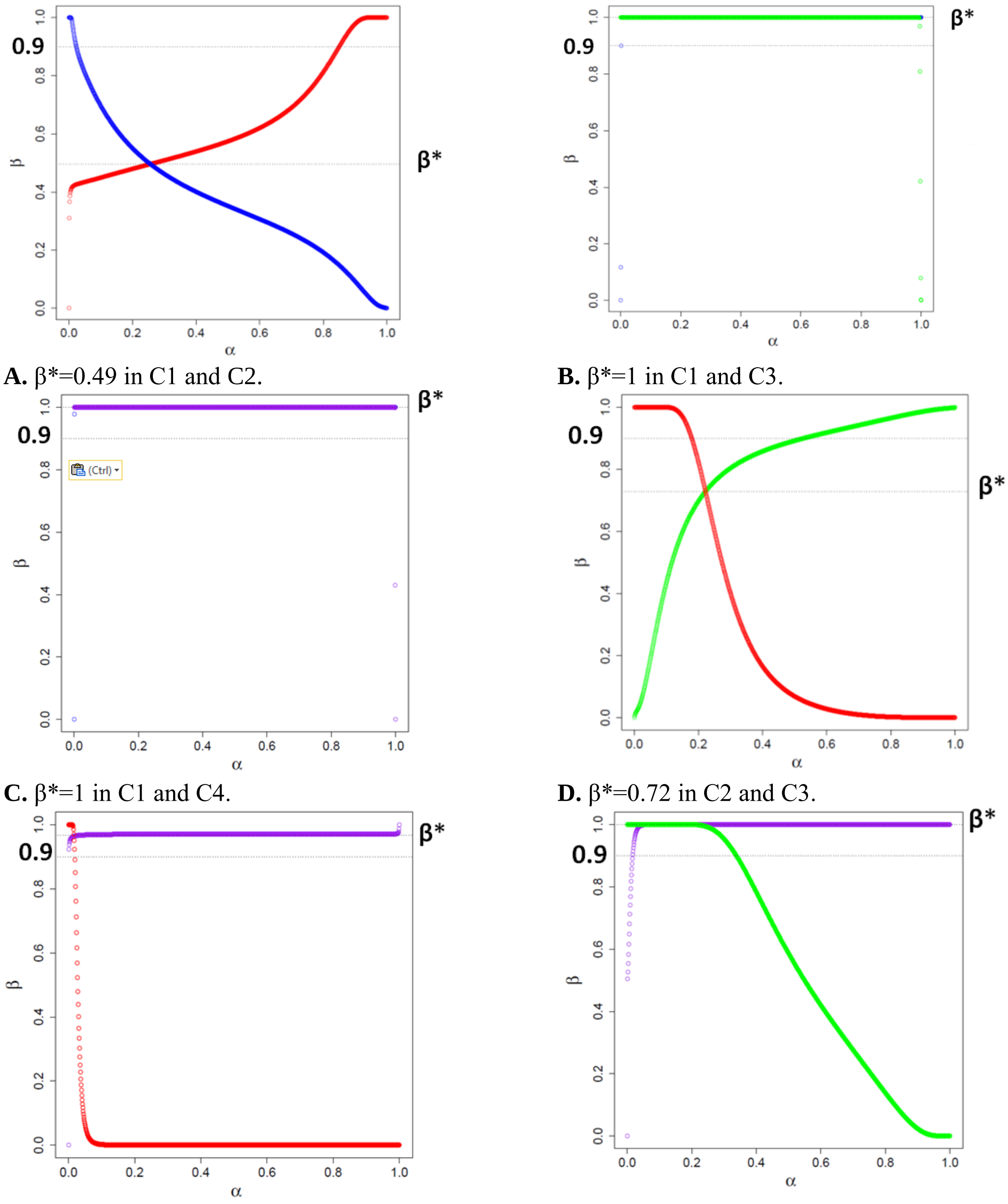
E. Slope of the PDF (C2 and C4)

F. Slope of the PDF (C3 and C4)

Figure S15. Probability density function (PDF) slope in the Trichophyllum clade. Blue Red green purple circle= modes of the morphospecies C1, C2, C3, and C4.

Assessment of morphological characters that separate the morphospeciesin the Trichophyllum clade

The phenotypes of C1 overlap with those of C2 (β*values = 0.49), Fig. S16.A, while the phenotypes the phenotypes of C2 overlap with those of (β*values =0.72), Fig. S16D. The characters in C1 and C2 overlap in the two morphospecies (Fig. S17). Moreover, the quartiles of univariate distribution of the morphological characters (points), number of gland per cm^2, deviates from the simulation of the non-overlapping phenotypes in C1 and C2 (Fig. S17A), and in C2 and C3 (Fig. S17D).



A. β*=0.49 in C1 and C2.

B. β*=1 in C1 and C3.

C. β*=1 in C1 and C4.

D. β*=0.72 in C2 and C3.

E. β*=0.96 in C2 and C4.

F. β*=1 in C3 and C4

Figure S16. Value of proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Trichophyllum clade compared to the cutoff 0.9 necessary for the four GMM morphospecies to be distinct. C1: blue, C2: red; C3: green, C4: purple.

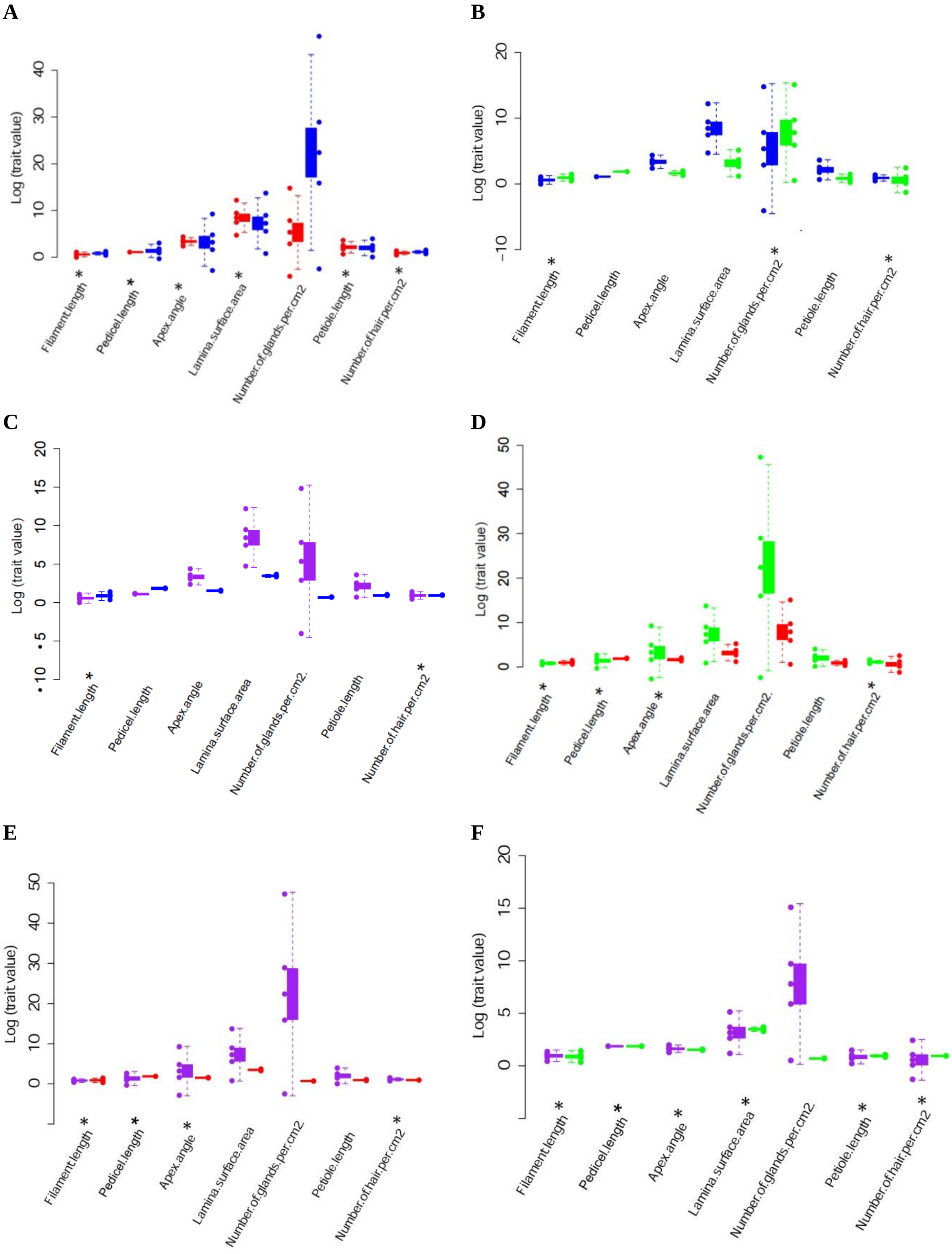


Figure S17. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analyses (points) and the simulated non-overlapping phenotypes in the Trichophyllum clade. Morphospecies (C), C1 (blue boxplot), C2 (red boxplot), C3 (green boxplot). C4 (purple boxplot). **A.** C1&C2, **B**. C1&C3, **C.** C1&C4, **D.** C2&C3, **E.** C2&C4, **F.** C3&C4. Overlapping characters are marked with asterisks.

CLADE BRACHYPODUM

GMM morphospecies in the Brachypodum clade

Clade Brachypodum has 87% bootstrap support, and the **7** specimens from the molecular study contained are assignable to 2 morphogroups (*Psorospermum brachypodum* and *P. cf. brachypodum*) (Fig. S18).

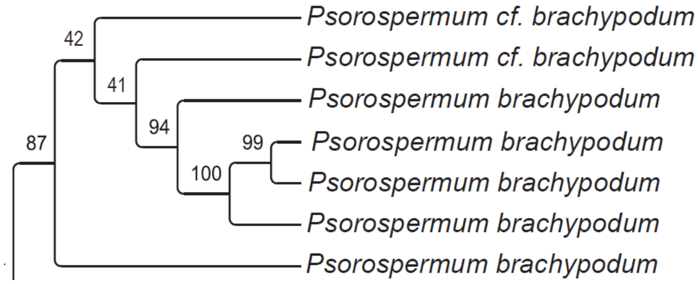


Figure S18. Expanded clade Brachypodum retrieved from the ML molecular phylogenetic analysis of *Psorospermum*. Numbers above the braches are bootstrap support values.

These specimens were added to 34 herbarium specimens previously assigned to these 2 morphogroups. A total of 41 herbarium specimens are used in this particular dataset. The two morphogroups were confirmed with LDA (Wilks's Lambda = 0.06699, p-value < 0.0001). There is no overlap of the values of the discriminant functions (Fig. S19).

The best model of the GMM analysis yielded three morphospecies. Fig. S21 summarizes the output of the Mclust models (table on the left), and highlights the identification of three morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model, EVI (diagonal, equal volume, varying shape), is the one that has the highest BIC value, it is indicated by the horizontal dotted line on the y-axis (Fig. S21); the number of morphospecies suggested by that model is indicated by the vertical dotted line on the x-axis (Fig. S21). Morphospecies in two-dimensional pairwise comparison of variables are shown in a two-dimensional (2-D) morphospace scatterplot and an example is given in Fig. S22. Five characters influence significantly the loadings of the observations in the PCA vector scatterplot: number of petal glands; filament length; number of ovules per carpel; cotyledon surface area; and radicle length (Fig. S20; Appendix Table S1 for complete list). These variables are used to perform the GMM analyses. All pairwise combinations of the five variables are shown in 2-D scatterplots in Appendix S6.

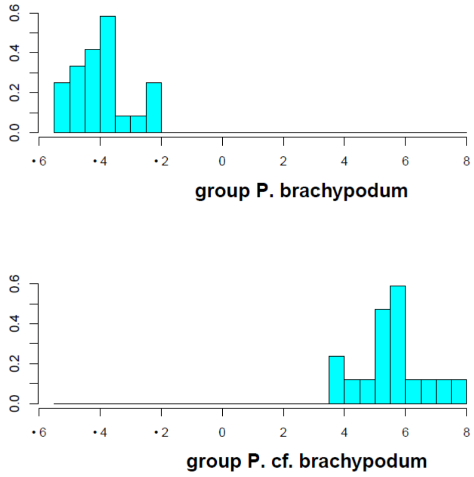


Figure S19. Histogram of values of the first discriminant function for the samples from the two morphogroups in the Brachypodum clade.

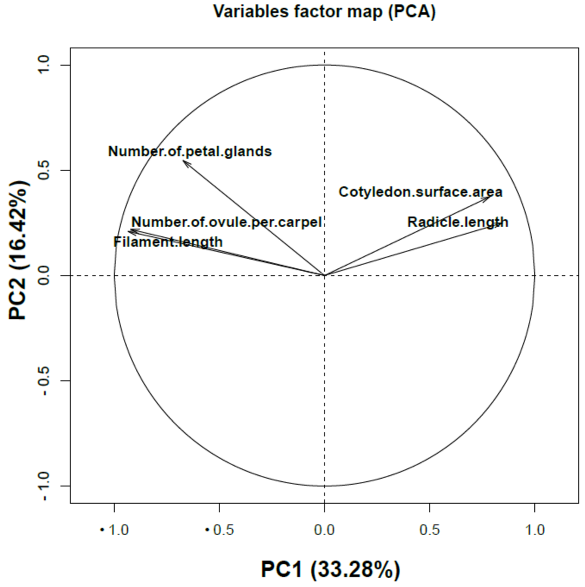


Figure S20. Five morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in the Brachypodum clade.

Figure S21. Outputs of GMM analysis of the Brachypodum clade.

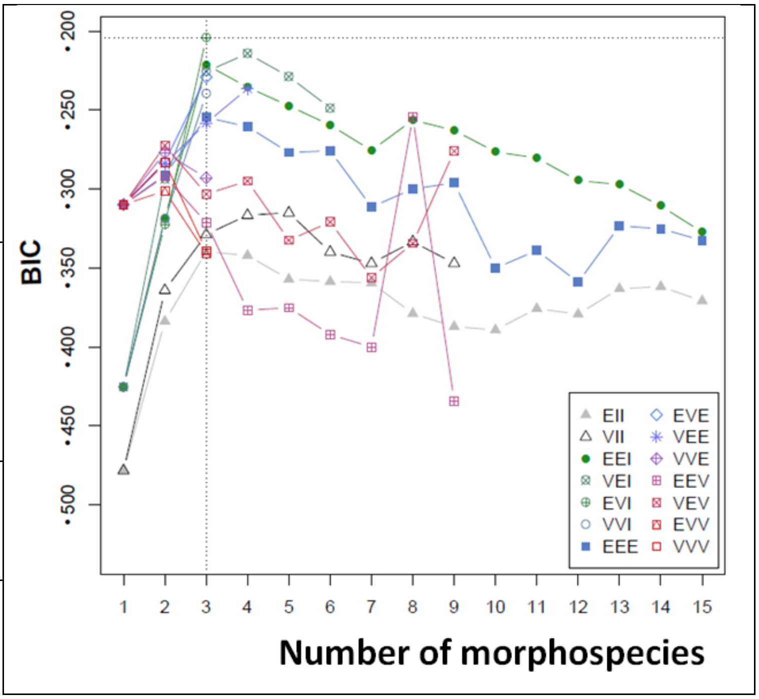
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N. ind.: number of individuals per morphospecies.

Mclust model with 3 morphospecies:

Diagonal, Equal volume, Varying shape (EVI).
See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
-46.38	41	30	-204.17

Morphospecies	1	2	3
N . ind.	16	9	16



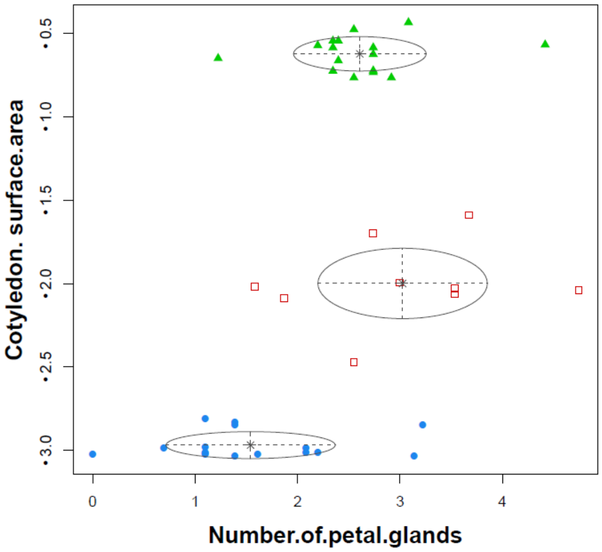


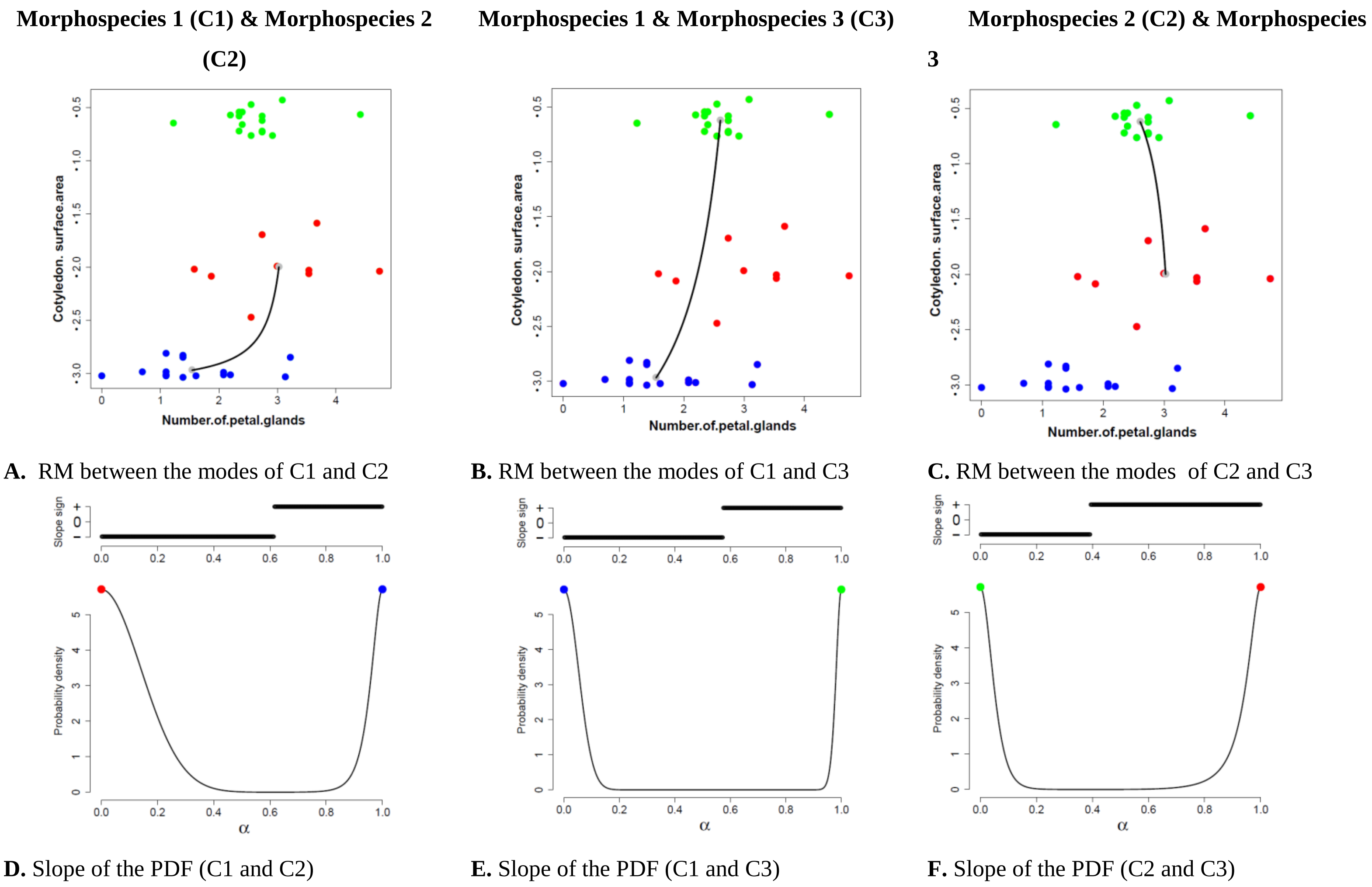
Figure S22. Bivariate scatterplot in the Brachypodum clade. The ellipses superimposed on the plot correspond to the covariances of the morphospecies. Individuals in different morphospecies are indicated by different symbols and colors, filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3.

Quantification of gaps between morphospecies in the Brachypodum clade

The RM and the slopes of the PDFs of the three identified morphospecies are shown in Fig. S23.A-C and Fig. S23.D-F respectively. The PDFs are conspicuously bimodal which indicates gaps between the morphospecies. The graph line above the PDF highlights the change of slope sign and confirms the bimodal nature of the PDF in morphospace.

Assessment of morphological characters that separate the morphospeciesin the Brachypodum clade

The phenotypes of the morphospecies do not overlap, β* values = 1 (Fig. S24. A-C). However, the quartiles of univariate distributions of the morphological characters are more ambiguous. Those from the GMM analysis (points) and the phenotypes (boxplots), overlap between C1 and C2, as well as C2 and C3. (Fig. S25).



Morphospecies 1 (C1) & Morphospecies 2 (C2)

Morphospecies 1 & Morphospecies 3 (C3)

Morphospecies 2 (C2) & Morphospecies 3

A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

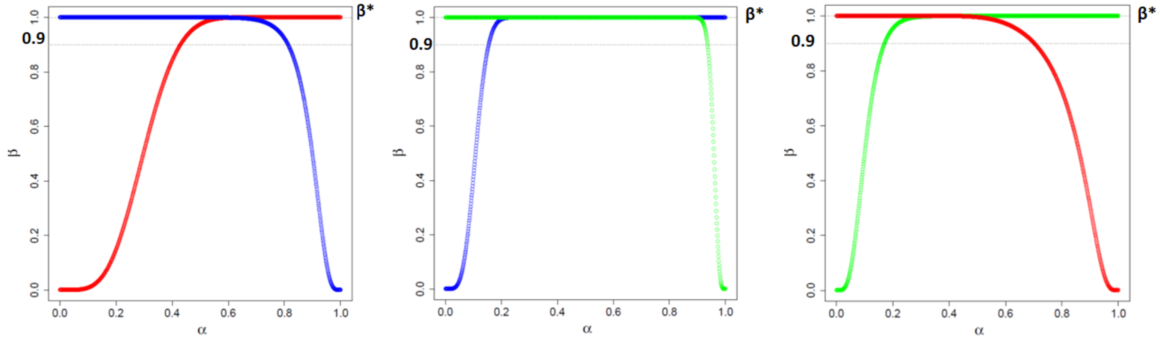
C. RM between the modes of C2 and C3

D. Slope of the PDF (C1 and C2)

E. Slope of the PDF (C1 and C3)

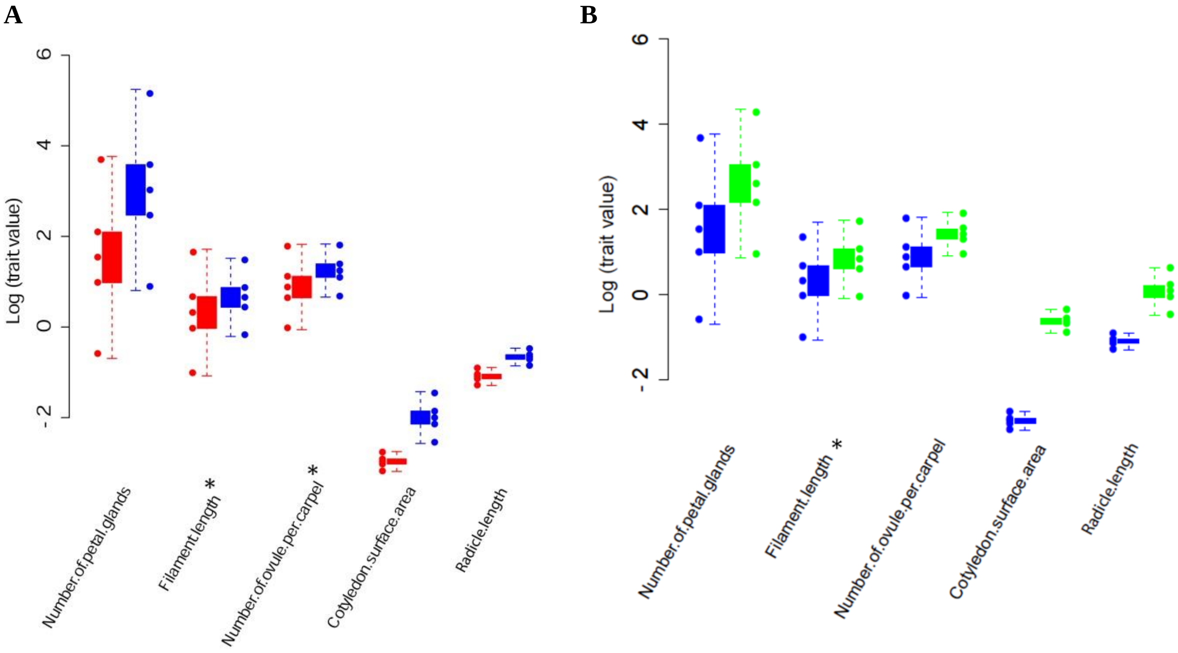
F. Slope of the PDF (C2 and C3)

Figure S23. Ridgeline Manifold (RM) and inference of gaps between the morphospecies in the Brachypodum clade. **A-C**: RM = continuous black line; **D- F**; blue, green and red points= modes of the morphospecies.



A. β*=0.99 in C1 and C2 **B.** β*=1 in C1 and C3 **C.** β*=0.99 in C2 and C3

Figure S24. Value of the proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Brachypodum clade compared to the cutoff 0.9 necessary for the three GMM morphospecies to be distinct.



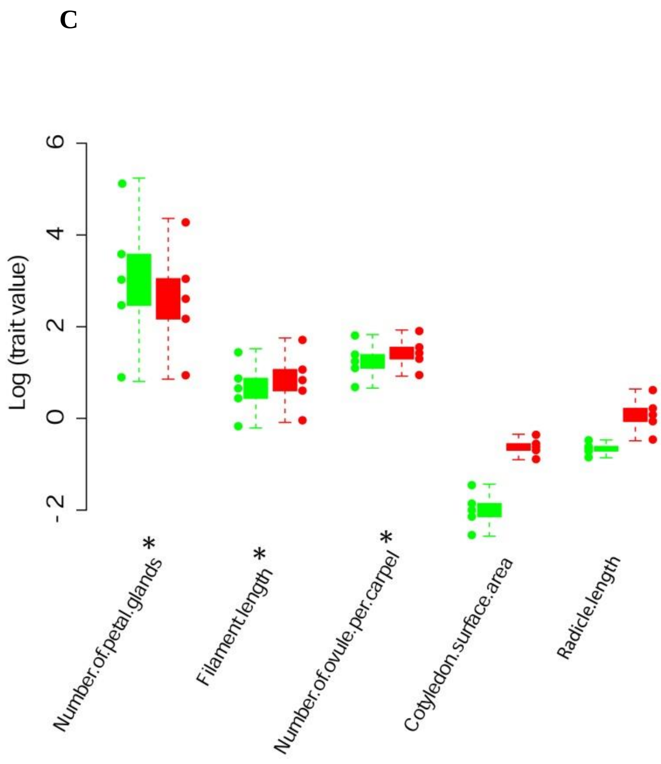


Figure S25. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in the Brachypodum clade. **A.** C1 (blue boxplot) and C2 (red boxplot); **B.** C1 (blue boxplot) and C3 (green boxplot); **C.** C2 (red boxplot) and C3 (green boxplot).

CLADE CHIONANTHIFOLIUM

GMM morphospecies in the Chionanthifolium clade

Clade Chionanthifolium has 93% bootstrap support, and the 6 specimens from the molecular study contained are assignable to 2 morphogroups (*Psorospermum chionanthifolium* and *P. crenatum*) (Fig. S26). These specimens were added to 26 herbarium specimens previously assigned to these 2 morphogroups. A total of 32 herbarium specimens are used in this particular dataset.

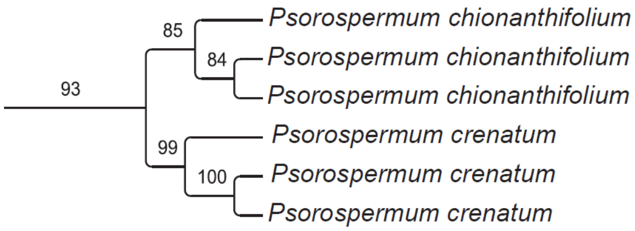

93
85
84
99
100
Psorospermum chionanthifolium
Psorospermum chionanthifolium
Psorospermum chionanthifolium
Psorospermum crenatum
Psorospermum crenatum
Psorospermum crenatum

Figure S26. Expanded clade Chionanthifolium retrieved from the ML molecular phylogenetic analysis of *Psorospermum*. Numbers above the braches are boostrap support values.

The 2 morphogroups were confirmed with LDA (Wilks's Lambda = 0.0014731, p-value < 0.001). There is no overlap of the values of the discriminant functions (Fig. S27).

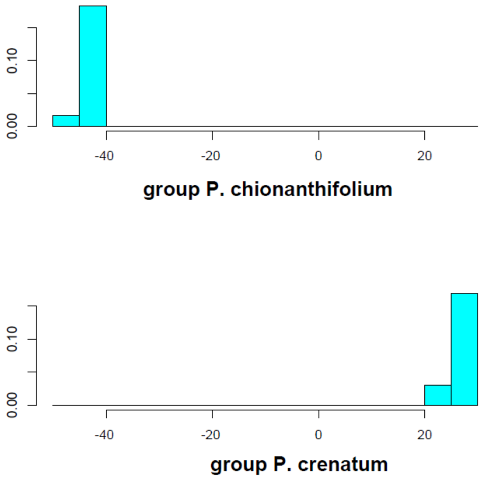


Figure S27. Histogram of values of the first discriminant function for the samples from the two morphogroups in the Chionanthifolium clade.

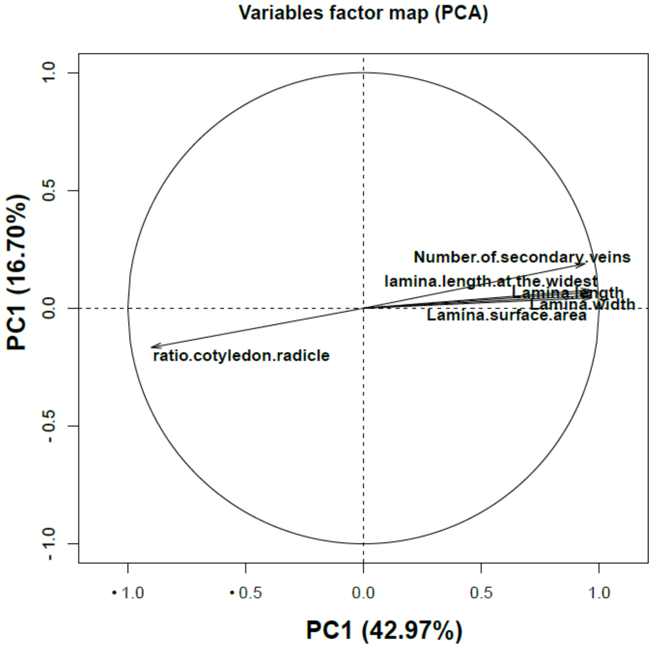


Figure S28. Seven morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in the Chionanthifolium clade

The best model of the GMM analysis yielded four morphospecies. Fig. S29 summarizes the output of the Mclust models (table on the left), and highlights the identification of four morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model, VEV (ellipsoidal, equal shape), is the one that has the highest BIC value, it is indicated by the horizontal dotted line on the y-axis (Fig. S29); the number of morphospecies suggested by that model is indicated by the vertical dotted line on the x-axis (Fig. S29). Six characters influence significantly the loadings of the observations in the PCA vector scatterplot: lamina length, width, length from the base to the widest point, surface area, number of secondary veins; and ratio cotyledon length:width (Fig. S28; Appendix Table S1 for complete list). These variables are used to perform the GMM analyses). Morphospecies in two-dimensional pairwise comparison of variables are shown in a two-dimensional (2-D) morphospace scatterplot and an example is given in Fig. S30. And all pairwise combinations of the six variables are shown in 2-D scatterplots in Appendix S7.

Figure S29. Outputs of GMM analysis of the Chionanthifolium clade.

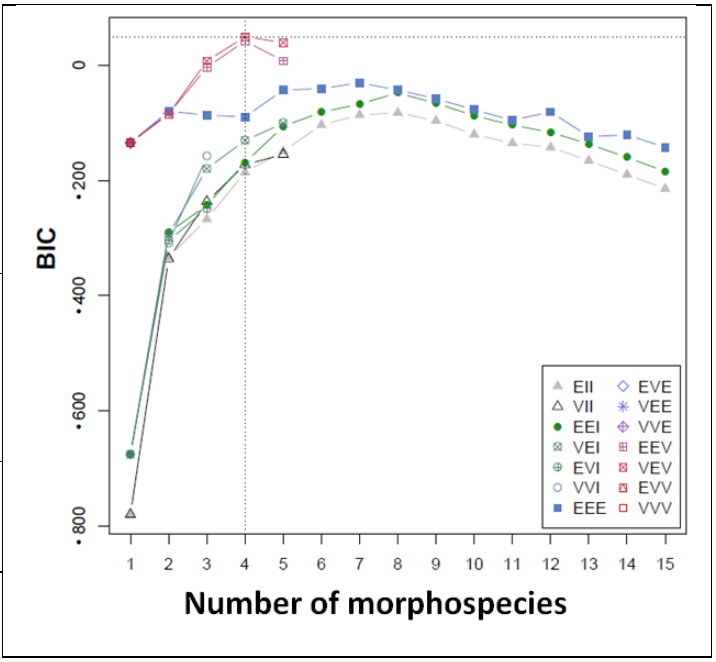
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N. ind.: number of individuals per morphospecies.

Mclust model with 4 morphospecies:

Ellipsoidal, Equal shape (VEV).
See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
191.26	32	96	49.85

Morphospecies	1	2	3	4
N . ind.	6	5	15	6



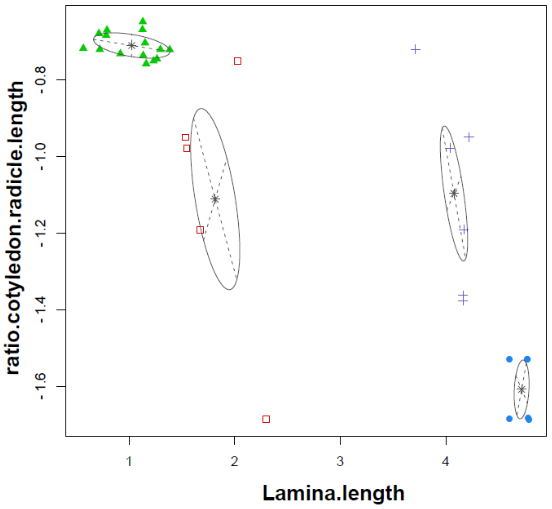


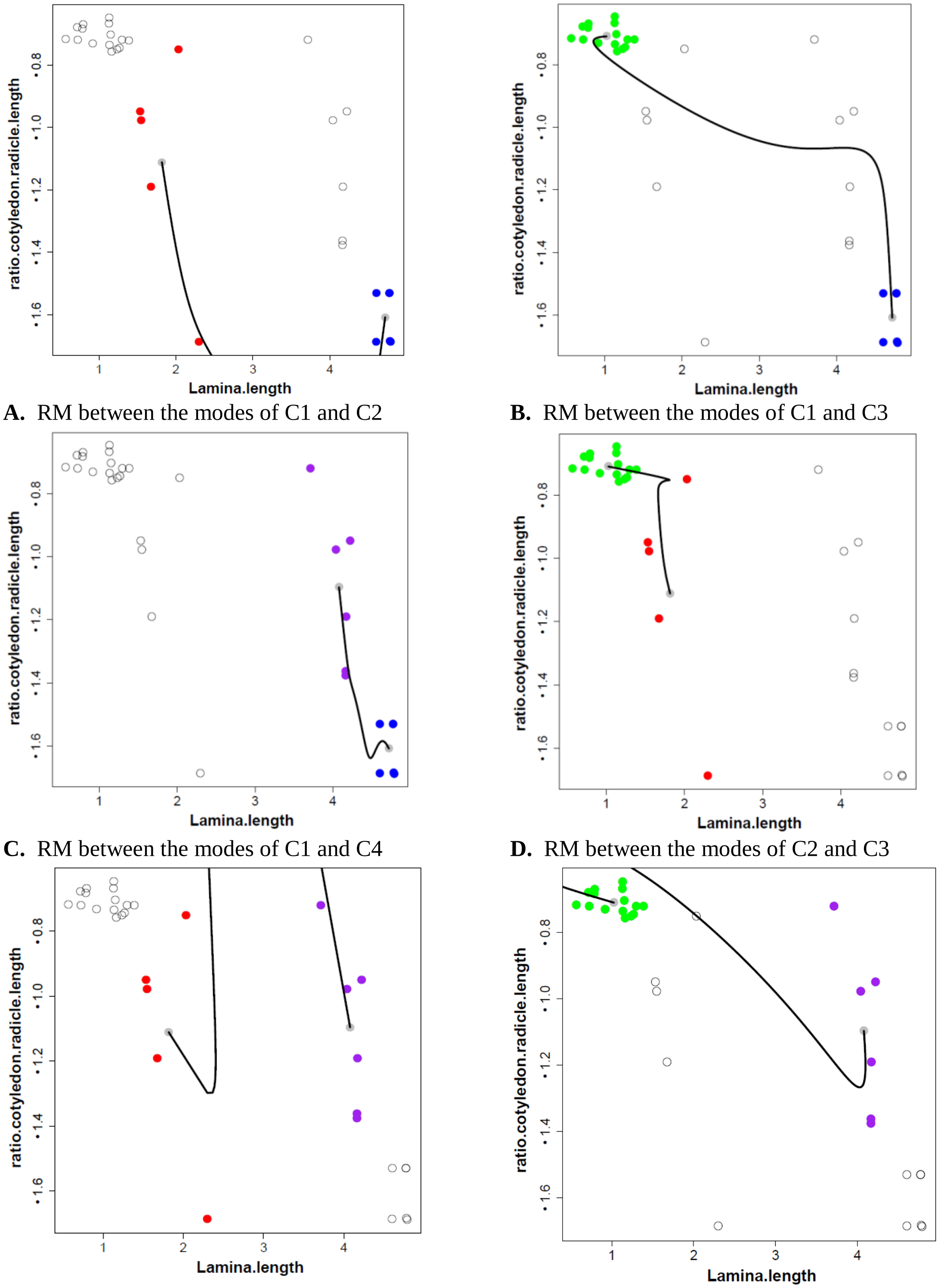
Figure S30. Bivariate scatterplot in the Chionanthifolium clade. The ellipses superimposed on the plot correspond to the covariances of the morphospecies. Individuals in different morphospecies are indicated by different symbols and colors, filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3, purple cross:4.

Quantification of gaps between morphospecies in the Chionanthifolium clade

The RM and the slopes of the PDFs of the three identified morphospecies are shown in Fig. S31.A-C and Fig. S32.E-F respectively. The PDFs are bimodal which indicates gaps between the morphospecies. The graph line above the PDF highlights the change of slope sign and confirms the bimodal nature of the PDF in morphospace.

Assessment of morphological characters that separate the morphospecies in the Chionanthifolium clade

The phenotypes of the 4 morphospecies do not overlap, having β* values = 1 (Fig. S33). The quantiles of univariate distribution of the morphological characters from the GMM analysis (points) fit the simulation of the non-overlapping phenotypes of all morphospecies, only the character, ratio of the cotyledon to the radicle, between morphospecies C2 and C4, and the character number of secondary veins, between morphospecies C2 and C3 show overlap (Fig. S34.E).



A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

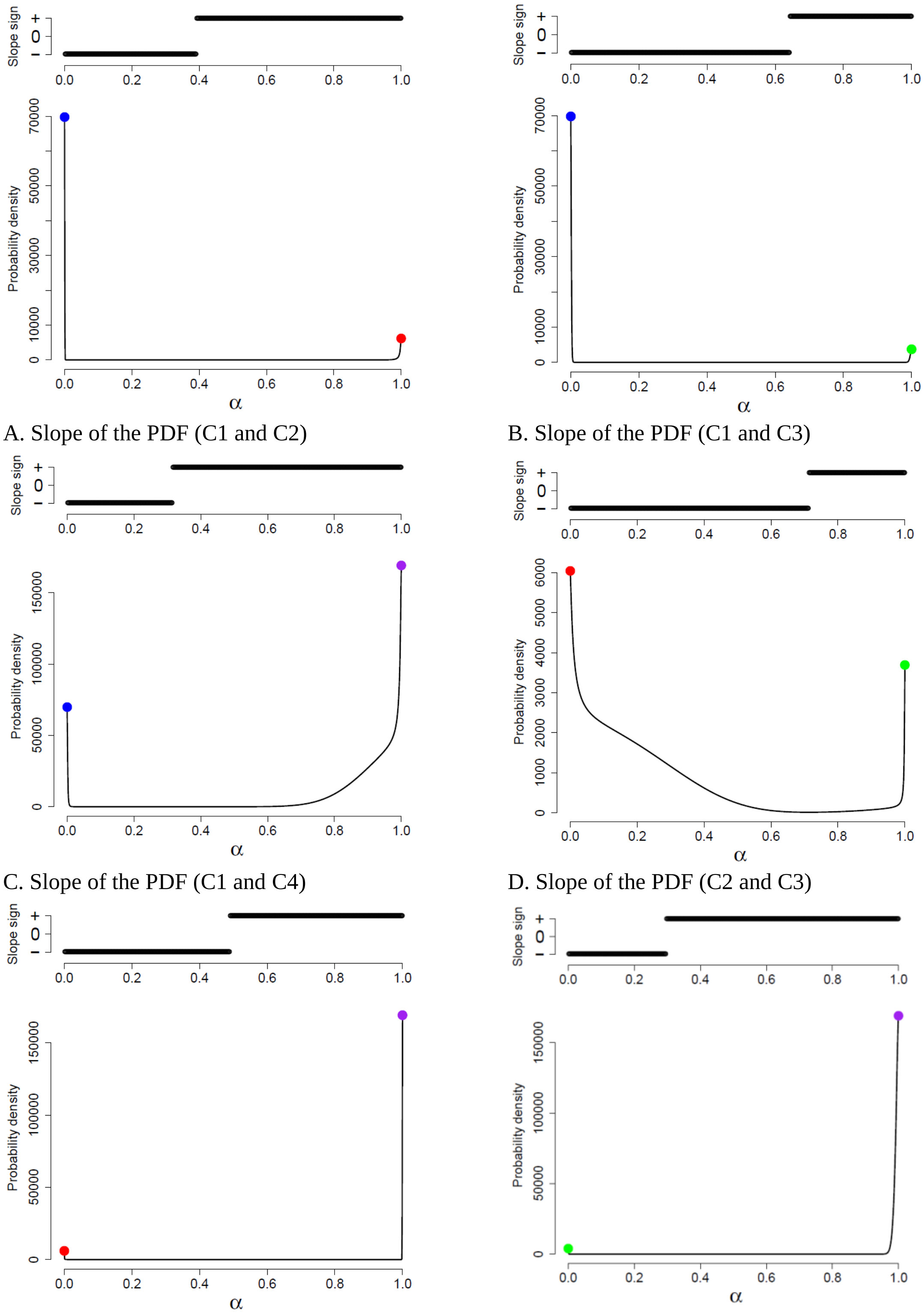
C. RM between the modes of C1 and C4

D. RM between the modes of C2 and C3

E. RM between the modes of C2 and C4

F. RM between the modes of C3 and C4

Figure S31. Ridgeline manifold (RM) between the morphospecies. RM = continuous black line; blue, red, green and purple points: morphospecies 1, 2, 3, 4.



A. Slope of the PDF (C1 and C2)

B. Slope of the PDF (C1 and C3)

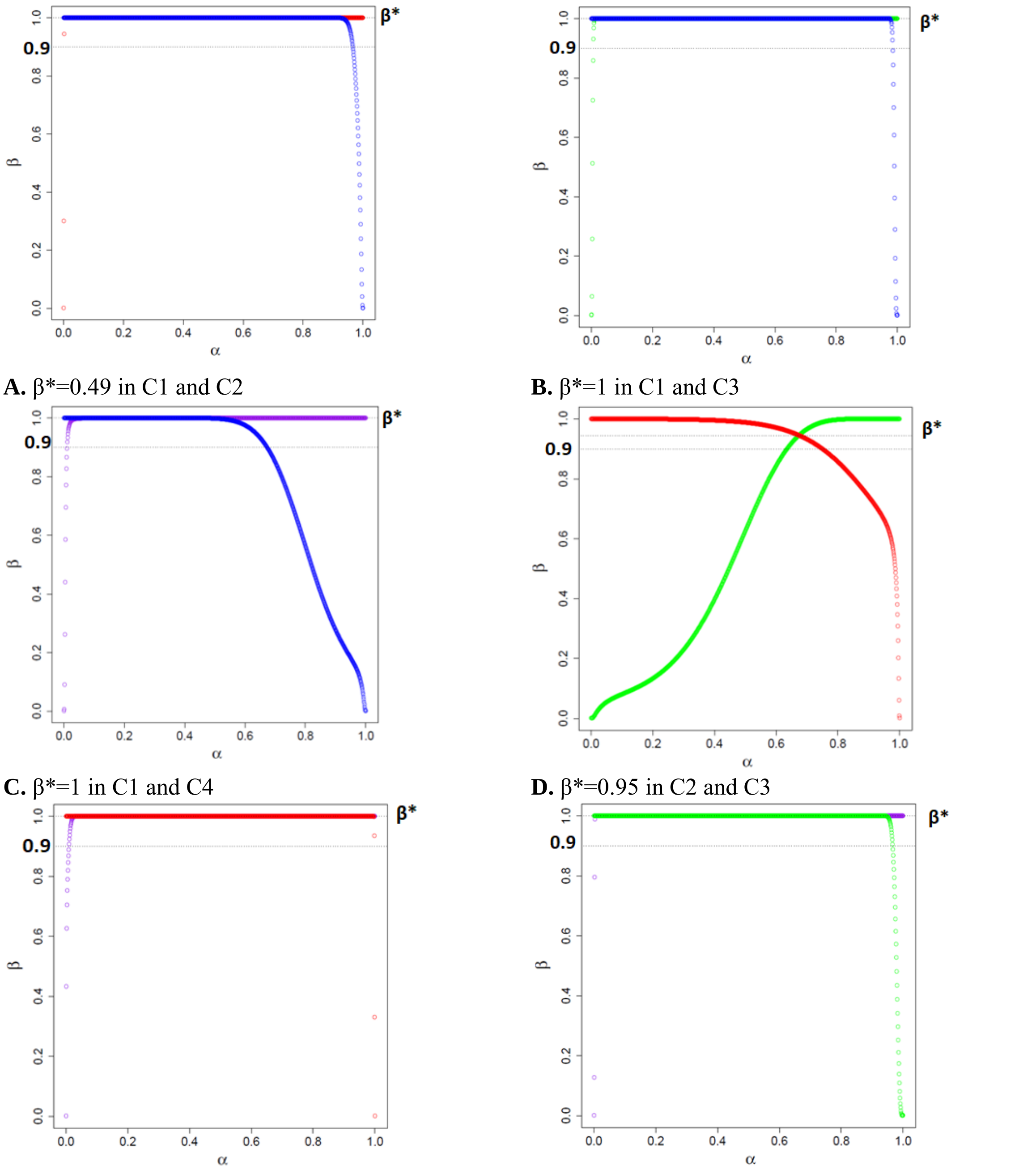
C. Slope of the PDF (C1 and C4)

D. Slope of the PDF (C2 and C3)

E. Slope of the PDF (C2 and C4)

F. Slope of the PDF (C3 and C4)

Figure S32. Inference of gaps between the morphospecies. Lines above the slopes indicates the length and the direction of the slopes. PDF= Probability Density Functions PDF.



A. β*=0.49 in C1 and C2

B. β*=1 in C1 and C3

C. β*=1 in C1 and C4

D. β*=0.95 in C2 and C3

E. β*=1 in C2 and C4

F. β*=1 in C3 and C4

Figure S33. Values of the proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Chionanthifolium clade compared to the cutoff 0.9 necessary for the four GMM morphospecies to be distinct.

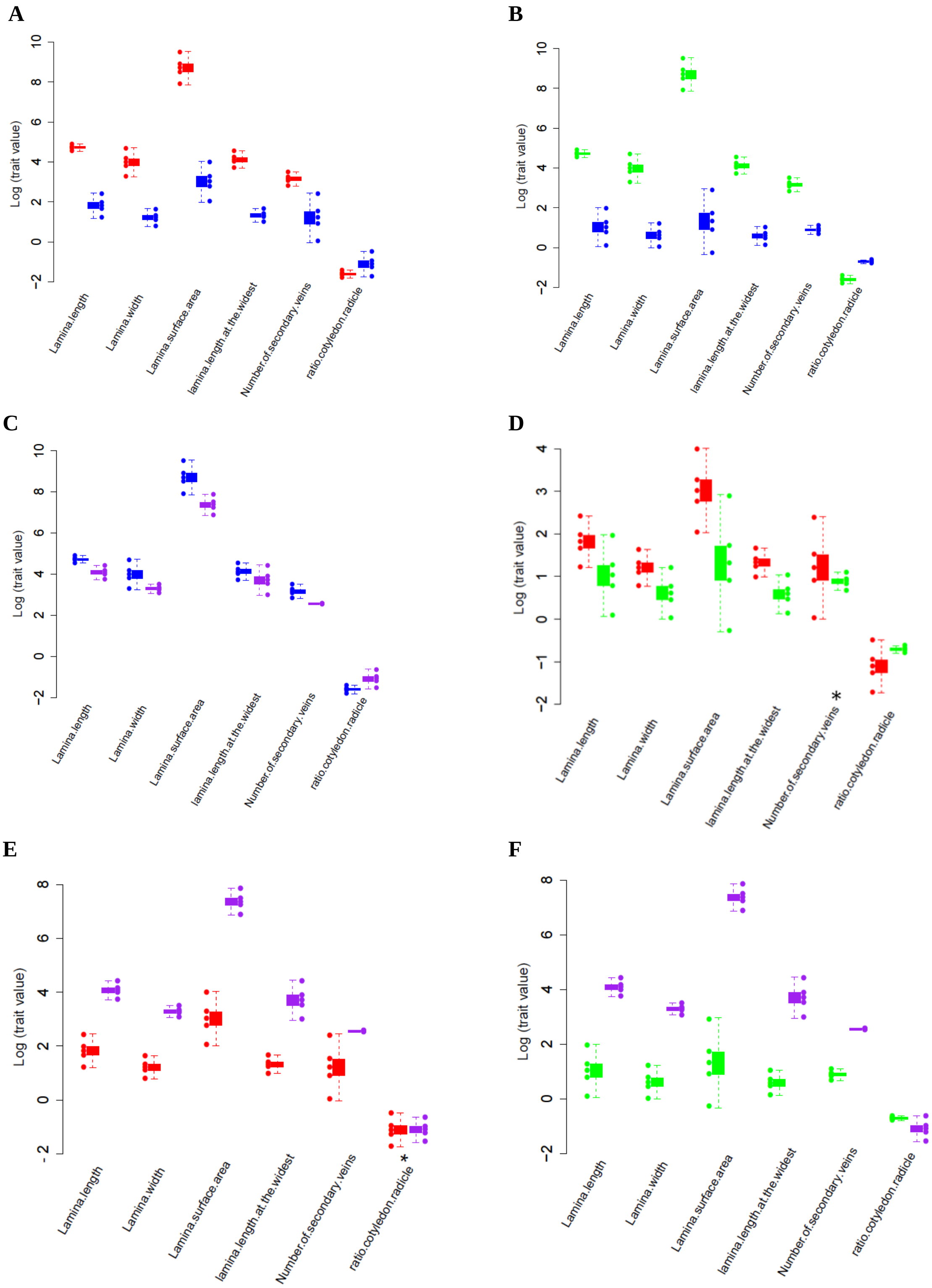


Figure S34. Comparison of the five quantiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in the Chionanthifoliium clade. C1 (blue boxplot), C2 (red boxplot), C3 (green boxplot). C4 (purple boxplot). **A.** C1&C2, **B**. C1&C3, **C.** C1&C4, **D.** C2&C3, **E.** C2&C4, **F.** C3&C4. Overlapping characters are marked with asterisks.

CLADE NANUM

GMM morphospecies in the Nanum clade

Clade Nanum has 93% bootstrap support, and the 6 specimens from the molecular study contained are assignable to 2 morphogroups (*Psorospermum nanum* and *P. humile*) (Fig. S35). These specimens were added to 23 herbarium specimens previously assigned to these 2 morphogroups. A total of 29 herbarium specimens are used in this particular dataset.

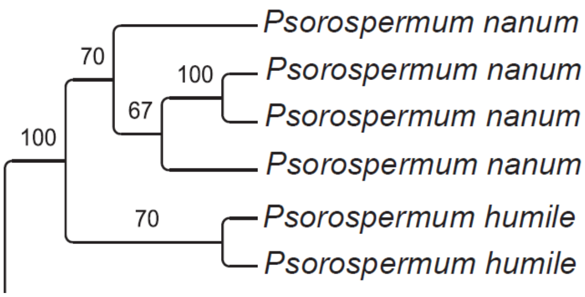


Figure S35. Expanded clade Nanum retrieved form ML molecular phylogenetic of *Psorospermum*. Numbers above the braches are bootstrap support values.

The 2 morphogroups were confirmed with LDA (Wilks's Lambda = 0.12531, p-value =0.004). There is no overlap of the values of the discriminant functions (Fig. S36).

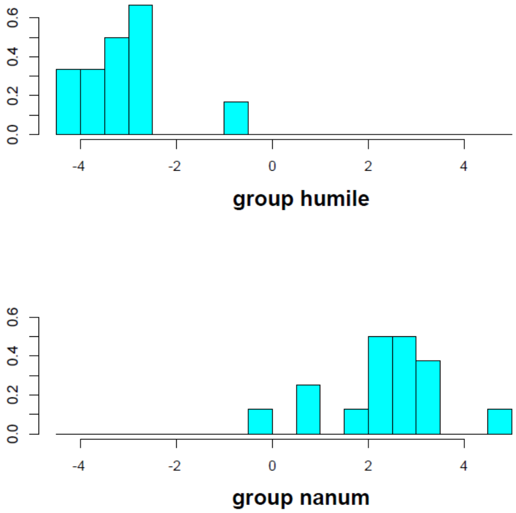


Figure S36. Histogram of values of the first discriminant function for the samples from the two morphogroups in the Nanum clade.

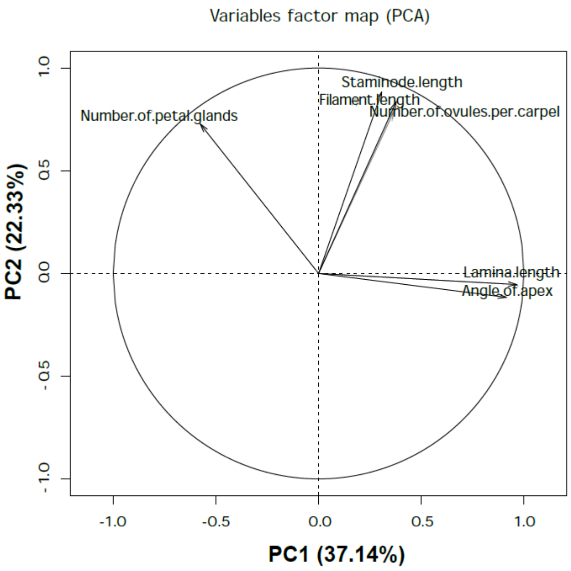


Figure S37. Six morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in the Nanum clade.

The best model of the GMM analysis yielded three morphospecies. Fig. S38 summarizes the output of the Mclust models (table on the left), and highlights the identification of three morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model, VEV (ellipsoidal, equal shape), is the one that has the highest BIC value, it is indicated by the horizontal dotted line on the y-axis (Fig. S38); the number of morphospecies suggested by that model is indicated by the vertical dotted line on the x-axis (Fig. S38). Six characters that influence significantly the loadings of the observations in the PCA vector scatterplot are: number of petal glands, staminode length, filament length, number of ovules per carpel, lamina apex angle, and length (Fig. S37; Appendix Table S1 for complete list). These variables are used to perform the GMM analyses. Morphospecies in two-dimensional pairwise comparison of variables are shown in a two-dimensional (2-D) morphospace scatterplot and an example is given in Fig. S39. All the pairwise combinations of the six variables are shown in 2-D scatterplots in Appendix S8.

Figure S38. Outputs of GMM analysis of the Nanum clade.

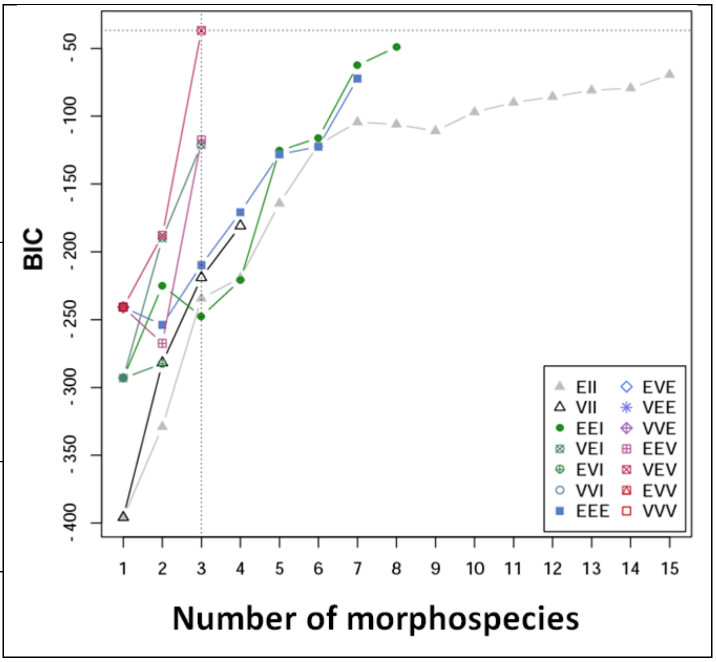
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N. ind.: number of individuals per morphospecies.

Mclust model with 3 morphospecies:

Ellipsoidal, equal shape (VEV).
See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
103.22	28	73	-36.80

Morphospecies	1	2	3
N . ind.	8	9	11



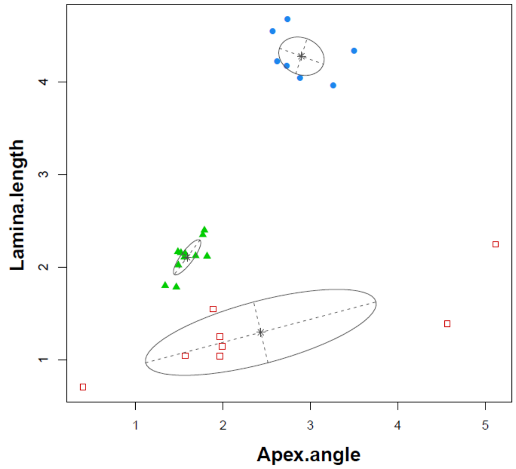
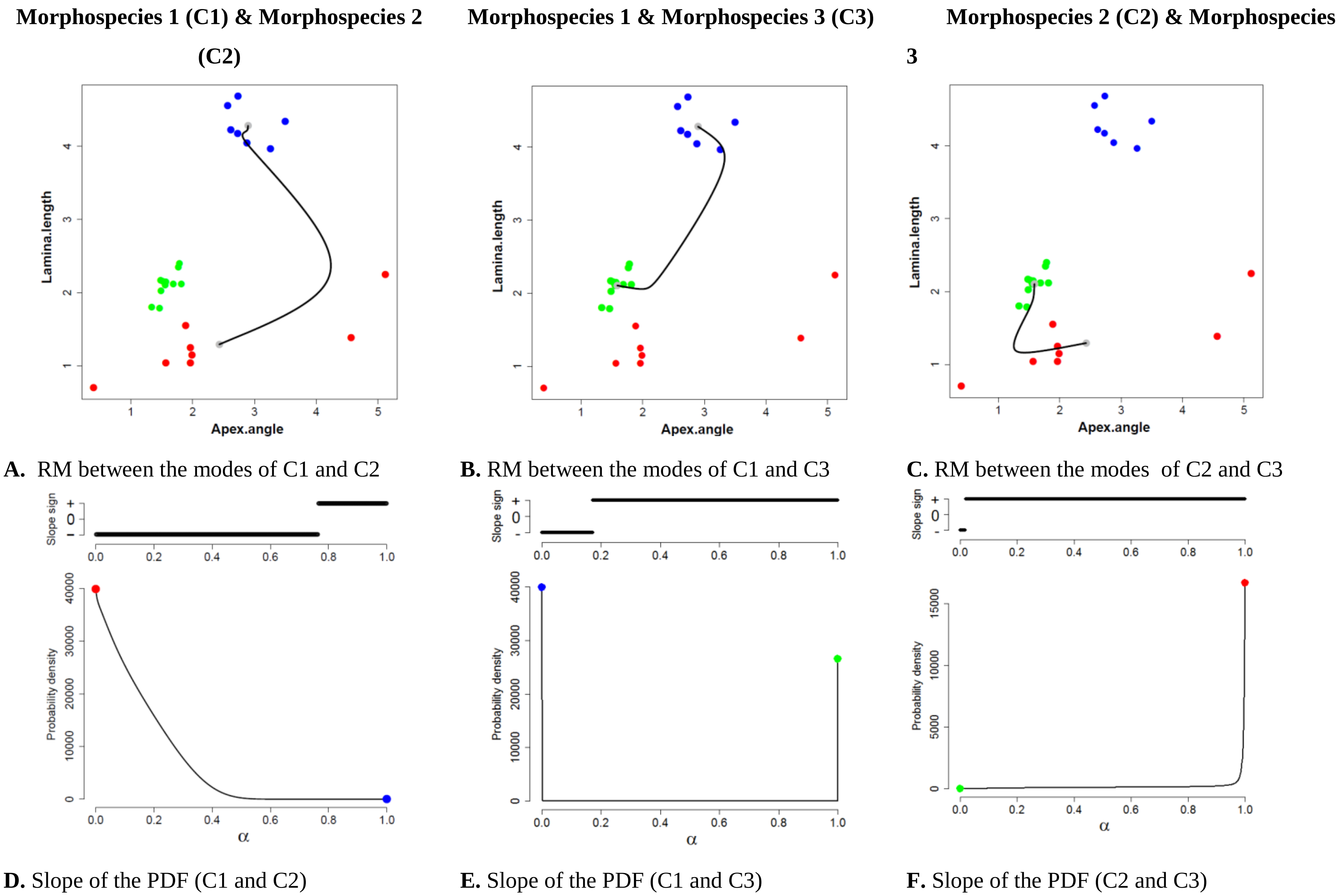


Figure S39. Bivariate scatterplots in the Nanum clade. The ellipses superimposed on the plot correspond to the covariances of the morphospecies. Individuals in different morphospecies are indicated by different symbols and colors, filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3.

Quantification of gaps between morphospecies in the Nanum clade

The RM and the slopes of the PDFs of the three identified morphospecies are shown in Fig. S40.A-C and Fig. S40.D-F respectively. The PDFs are bimodal which indicates gaps between the morphospecies, but the bimodal nature of the PDFs of C1 and C2, as well as C2 and C3 are not conspicuous (Fig. S40.D and Fig. S40.F).



Morphospecies 1 (C1) & Morphospecies 2 (C2)

Morphospecies 1 & Morphospecies 3 (C3)

Morphospecies 2 (C2) & Morphospecies 3

A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

C. RM between the modes of C2 and C3

D. Slope of the PDF (C1 and C2)

E. Slope of the PDF (C1 and C3)

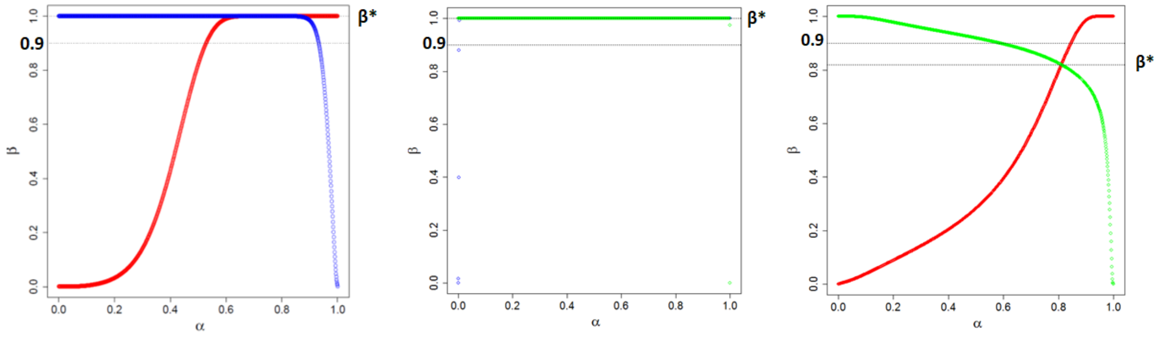
F. Slope of the PDF (C2 and C3)

Figure S40. Ridgeline Manifold (RM) and inference of gaps between the morphospecies in the Nanum clade. **A-C**: RM = continuous black line; **D- F**; blue, green and red points= modes of the morphospecies.

However, the graph line above the PDF highlights the change of slope sign and confirms the bimodal nature of the PDF in morphospace.

Assessment of morphological characters that separate the morphospecies in the Nanum clade

The phenotypes of C1 do not overlap with those of C2 and C3 with β*values = 1 (Fig. S41.A & B); however there are overlapping phenotypes between C2 and C3, with a β* value (Fig. S41.C) below the cutoff 0.9 of the proportion of the individuals with non-overlapping phenotypes (β*values = 0.8). The quartiles of univariate distribution of the morphological characters from the GMM analysis (points) and simulation of the non-overlapping phenotypes of C2 and C3 (boxplots), show some overlap (Fig. S42). The number of ovules per carpel and filament length simulate poorly the non-overlap in C1 versus C2, the number of petal glands and the number of ovules per carpel overlap in C1 versus C3, and number of ovules per carpel, the number of petal glands, and filament length overlap in C2 versus C3.



A. β*=1 in C1 and C2 **B.** β*=1 in C1 and C3 **C.** β*=0.8 in C2 and C3

Figure S41. Values of the proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Nanum clade compared to the cutoff 0.9 necessary for the three GMM morphospecies to be distinct.

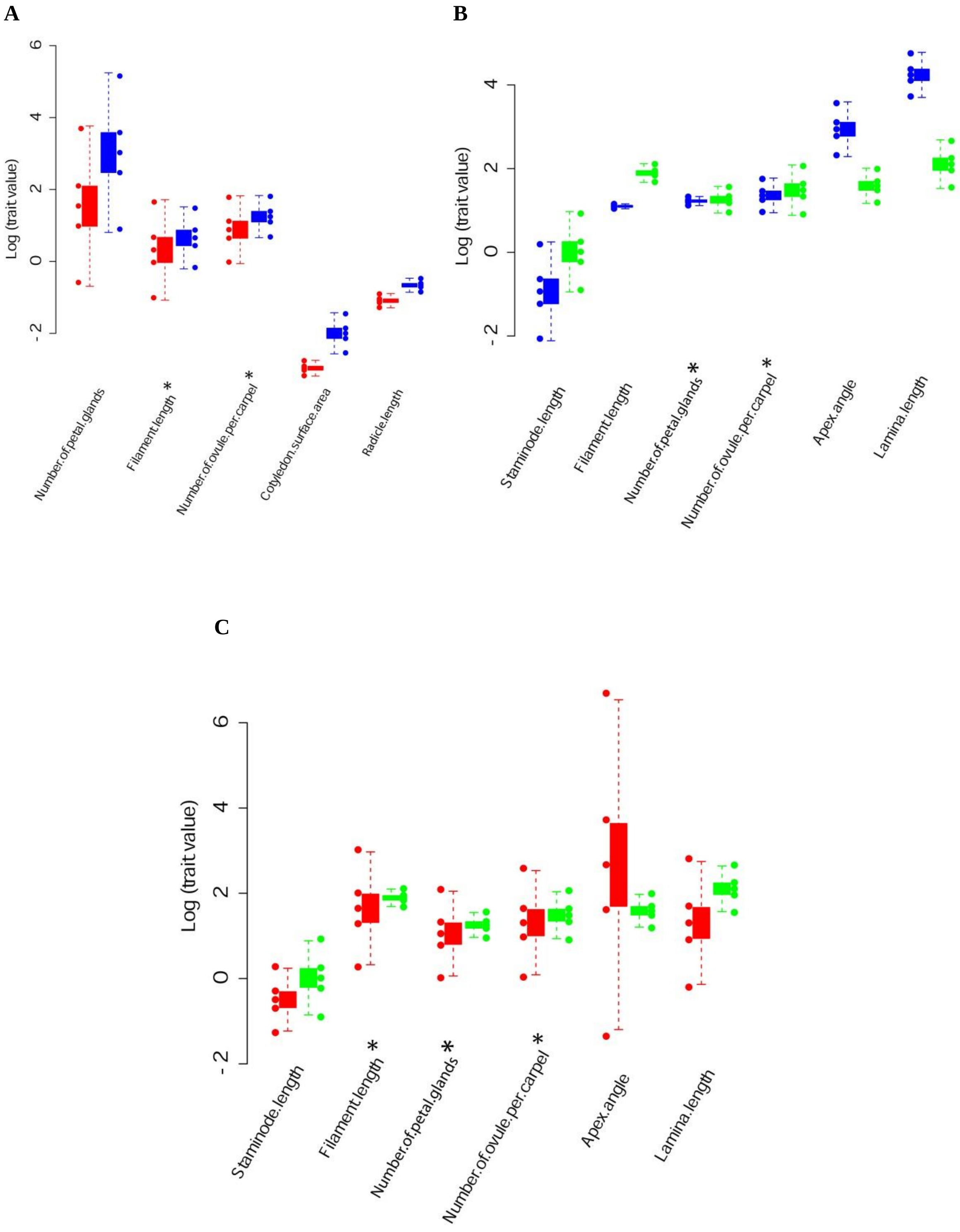


Figure S42. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in the Nanum Clade **A**. C1 (blue boxplot) and C2 (red boxplot); **B.** A. C1 (blue boxplot) and C3 (green boxplot); **C.** C2 (red boxplot) and C3 (greenboxplot).

CLADE FERROVESTITUM

GMM morphospecies in the Ferrovestitum clade

Clade Ferrovestitum has 61% bootstrap support, and the 10 specimens from the molecular study contained are assignable to 4 morphogroups (*Psorospermum sp17, Psorospermum sp16, P. ferrovestitum* and *P. fanerana*) (Fig. S43). These specimens were added to 38 herbarium specimens previously assigned to these 4 morphogroups. A total of 48 herbarium specimens are used in this particular dataset.

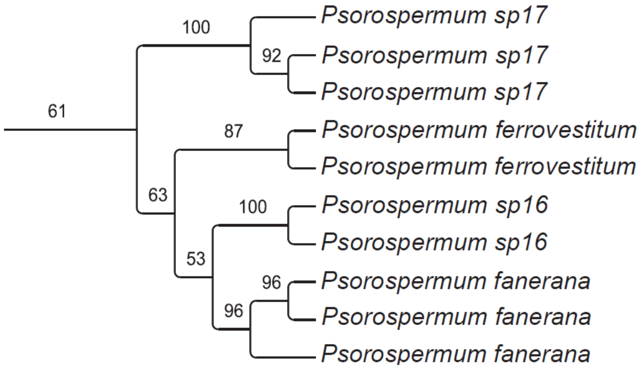


Fig. S43. Expanded clade Ferrovestitum retrieved from the ML molecular phylogenetic analysis of *Psorospermum*. Numbers above the braches are bootstrap support values.

The 4 morphogroups were tested with LDA (Wilks's Lambda = 0.0056, p-value < 4.974e^{-14}). There is overlap in the histograms of the discriminant functions between *P. sp16*, *P. ferrovestitum* and *P. fanerana* (Fig. S44).

The best model of the GMM analysis yielded four morphospecies. Fig. S46 summarizes the output of the Mclust models (table on the left), and highlights the identification of four morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model, VVV (ellipsoidal, varying volume, shape and orientation), is the one that has the highest BIC value, it is indicated by the horizontal dotted line on the y-axis (Fig. S46); the number of morphospecies suggested by that model is indicated by the vertical dotted line on the x-axis (Fig. S46). Six characters influence significantly the loadings of the observations in the PCA vector scatterplot are: petal length; lamina apex, length, width, shape, and number of hairs (Fig.

S45; Appendix Table S1 for complete list). These variables are used to perform the GMM analyses. Morphospecies are shown in a two-dimensional (2-D) morphospace scatterplot and an example is shown in Fig. S47. These variables are used to perform the GMM analyses. All the pairwise combinations the six variables are shown in 2-D scatterplots in Appendix S9.

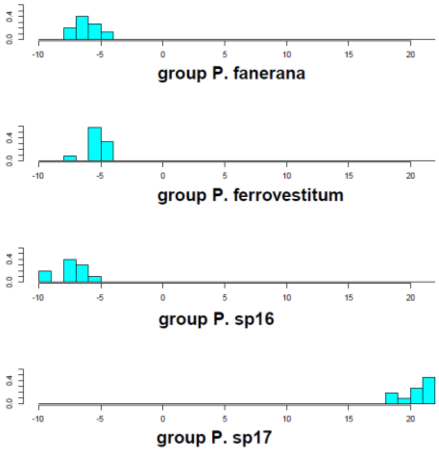


Figure S44. Histogram of values of the first discriminant function for the samples from the four morphogroups in the Ferrovestitum clade.

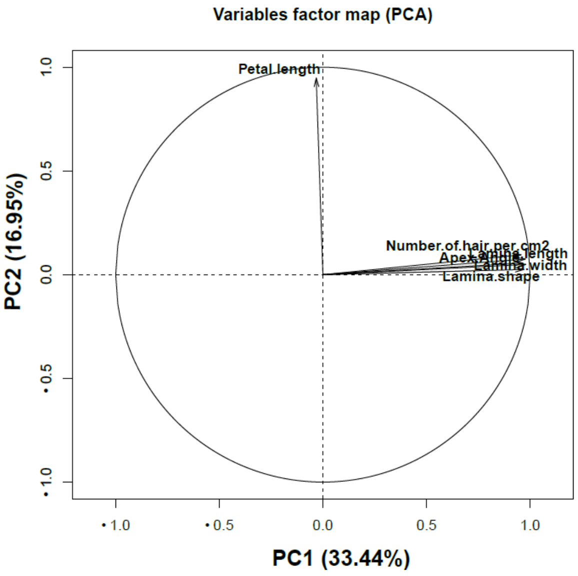


Figure S45. Six morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in the Ferrovestitum clade.

Figure S46. Outputs of GMM analysis of the Ferrovestitum clade.

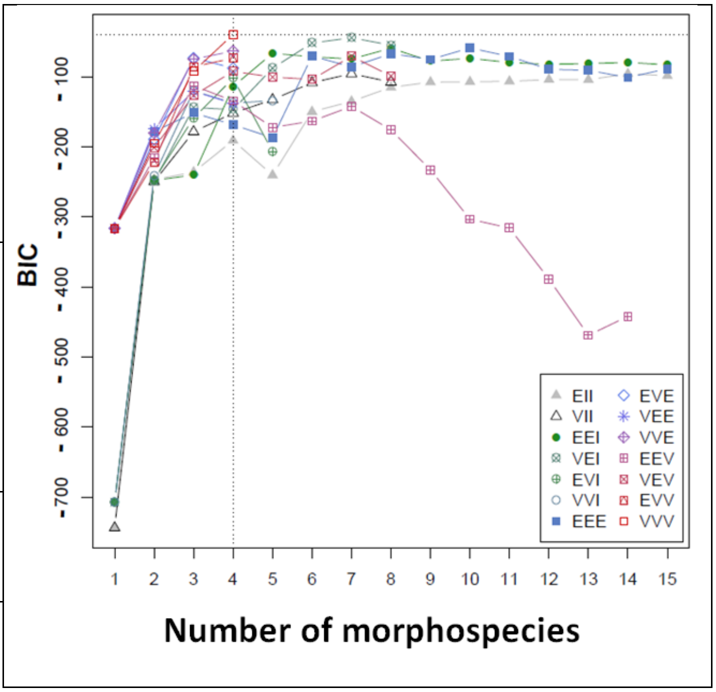
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N. ind.: number of individuals per morphospecies.

Mclust model with 4 morphospecies:

Ellipsoidal, varying volume, shape and orientation (VVV).
See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
195.05	48	111	-39.59

Morphospecies	1	2	3	4
N . ind.	12	18	7	11



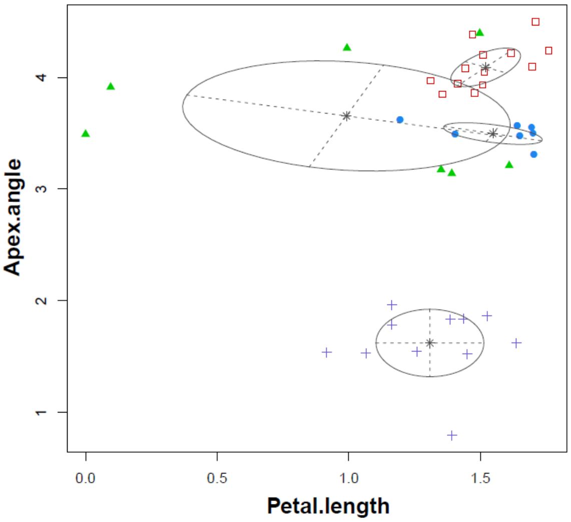


Figure S47. Bivariate scatterplot of the Ferrovestitum clade. The ellipses superimposed on the plot correspond to the covariances of the morphospecies. Individuals in different morphospecies are indicated by different symbols and colors, filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3, purple cross:4.

Quantification of gaps between morphospecies in the Ferrovestitum clade

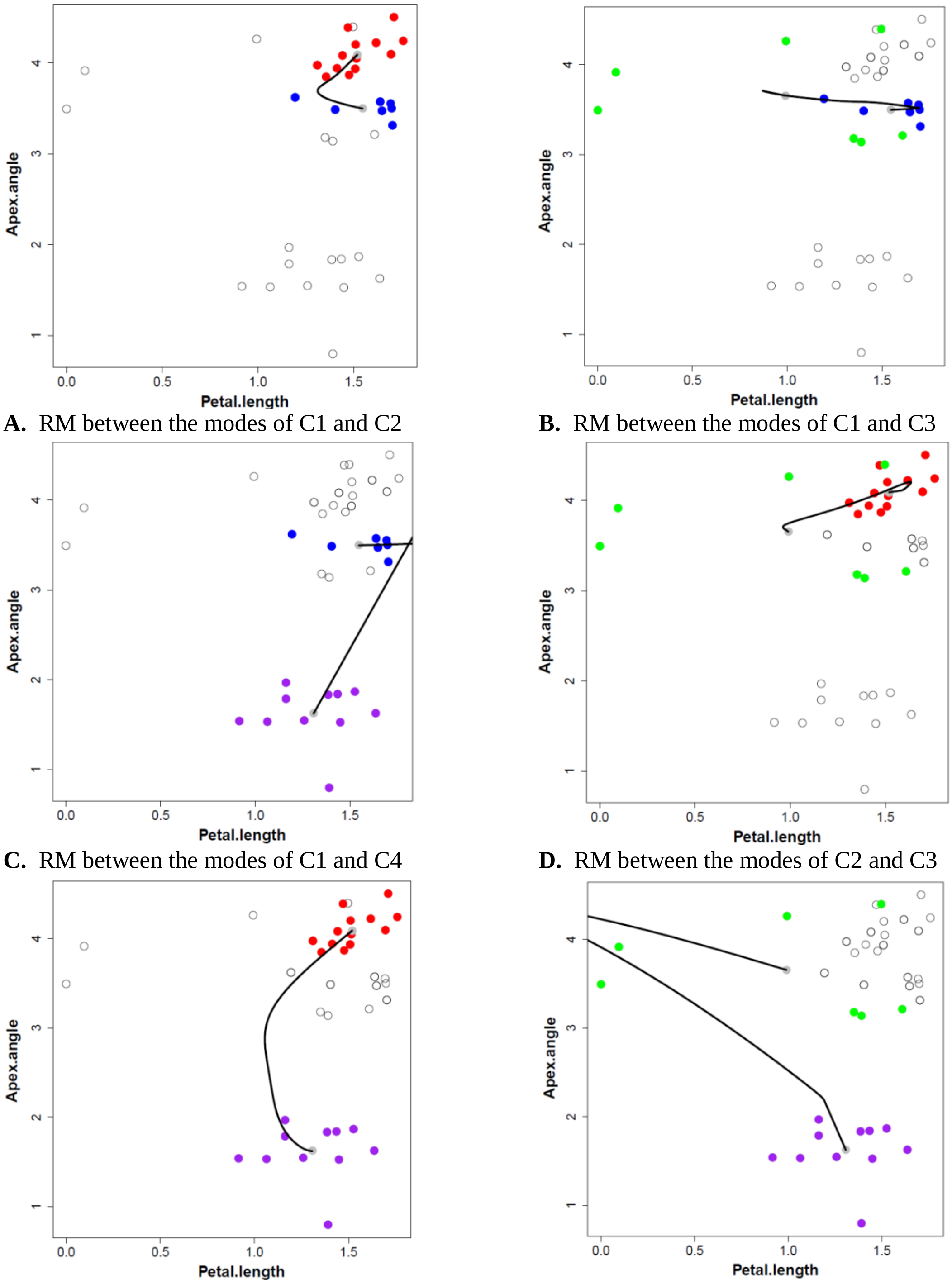
The RM and the slopes of the PDFs of the four identified morphospecies are shown in Fig. S48.A-F and Fig. S49.A-F respectively. The PDFs of C2 and C3, C2 and

C4, and C3 and C4 are bimodal which indicates gaps between the morphospecies. The bimodal nature of the slopes between C1 and C2; C3 and C4 respectively are inconspicuous, however the graph lines of change of slope sign above the PDFs confirm that there are gaps.

Assessment of morphological characters that separate the morphospeciesin the Ferrovestitum clade

The phenotypes of C1 overlap with those of C2 and C3 with β*values below the 0.9 cutoff of the proportion of individuals with overlapping phenotypes, respectively equal to 0.55 and 0.78 (Fig. S50.A & B); There are no overlapping phenotypes between C1 and C4; C2 and C3, C2 and C4; C3 and C4. β* = 1 (Fig. S50.C-F).

The simulations of the non-overlapping phenotypes (boxplots Fig. S51.A-F**)** indicates that petal length overlap in all cases, and the number of hair per cm^2 on the lamina surface overlap as well in C1 and C3, and C2 and C3. The quartiles of univariate distribution of the petal length and angle of apex deviate from the simulation of univariate distribution the non-overlapping phenotypes in all morphospecies pairwise comparisons but C2 and C3; the quartiles of univariate distribution of shape and number of hair per cm^2 on the lamina surface also deviate from the non-overlapping phenotypes in C1 and C3; and the quartiles of univariate distribution of all six characters deviate from the simulation of univariate distribution the non-overlapping phenotypes in C1 and C2 (Fig. S51).



A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

C. RM between the modes of C1 and C4

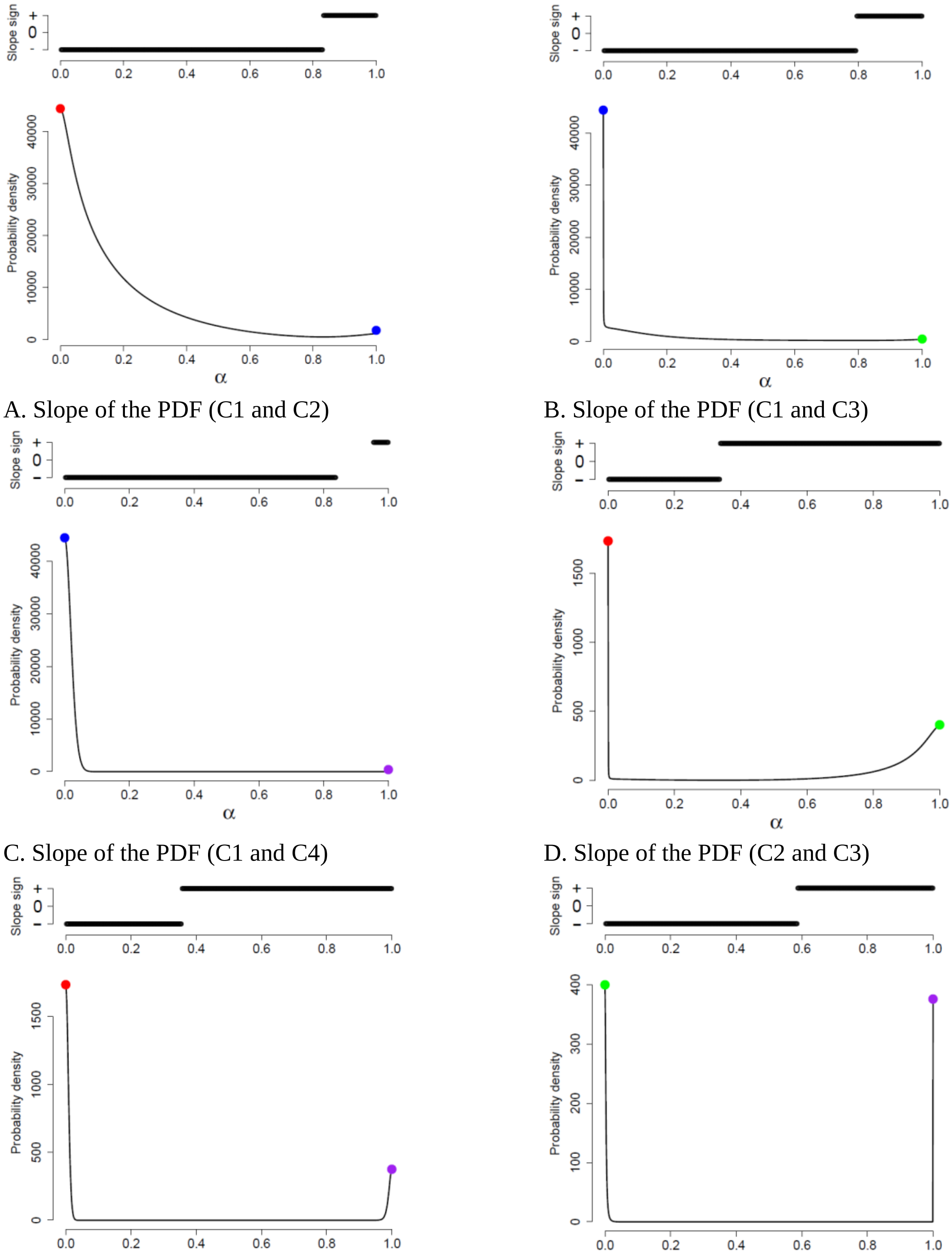
D. RM between the modes of C2 and C3

E. RM between the modes of C2 and C4

E. RM between the modes of C3 and C4

Figure S48. Ridgeline Manifold (RM) between the morphospecies in the Ferrovestitum clade. RM = continuous black line; color of points correspond to the morphospecies, blue, red, green and purple are respectively C1,C2,C3,and C4. Only the pairs of morphospecies analyzed are colored.

.



A. Slope of the PDF (C1 and C2)

B. Slope of the PDF (C1 and C3)

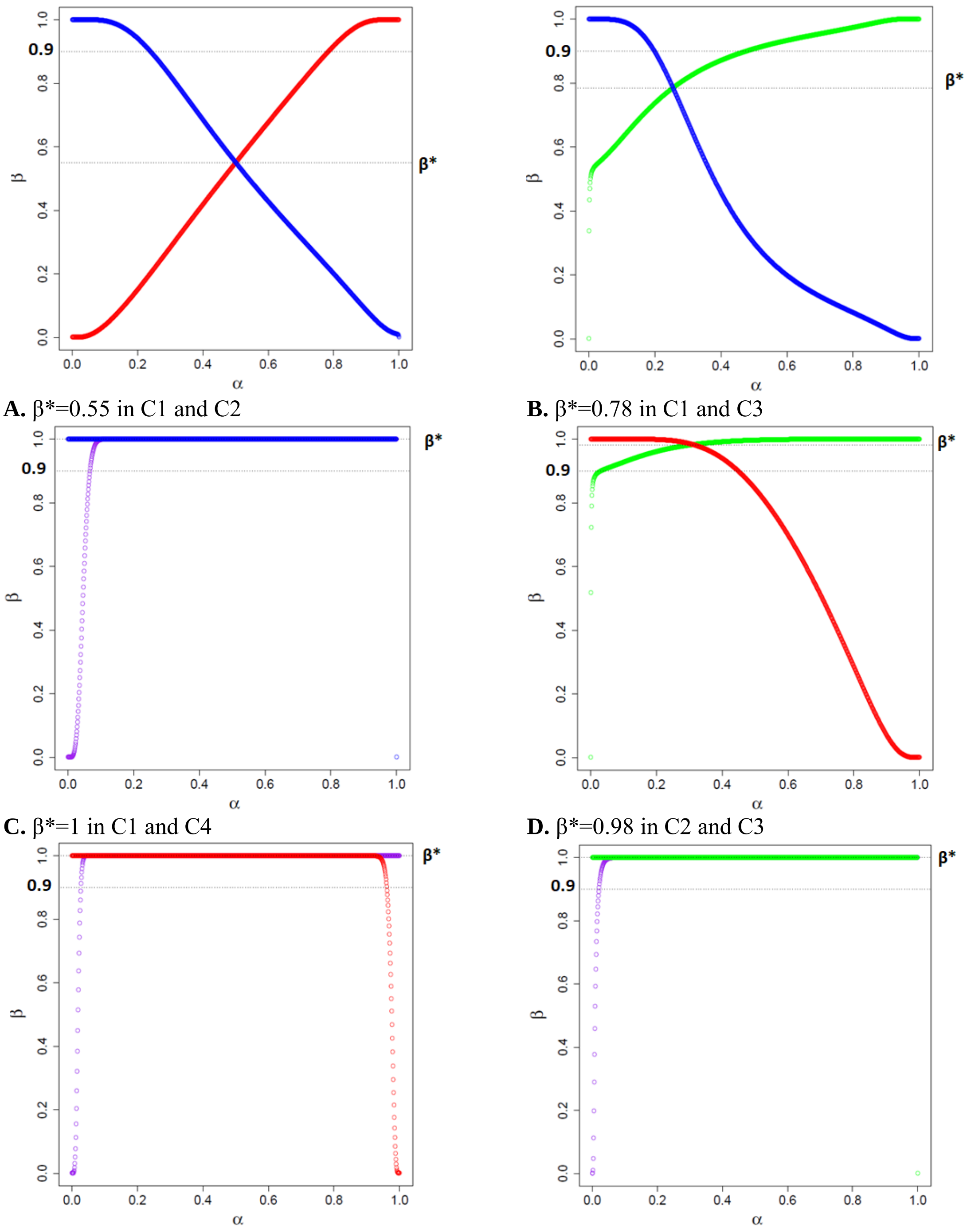
C. Slope of the PDF (C1 and C4)

D. Slope of the PDF (C2 and C3)

E. Slope of the PDF (C2 and C4)

F. Slope of the PDF (C3 and C4)

Figure S49. Inference of gaps between the morphospecies in the Ferrovestitum clade. Lines above the slopes indicates the length and the direction of the slopes.



A. β*=0.55 in C1 and C2

B. β*=0.78 in C1 and C3

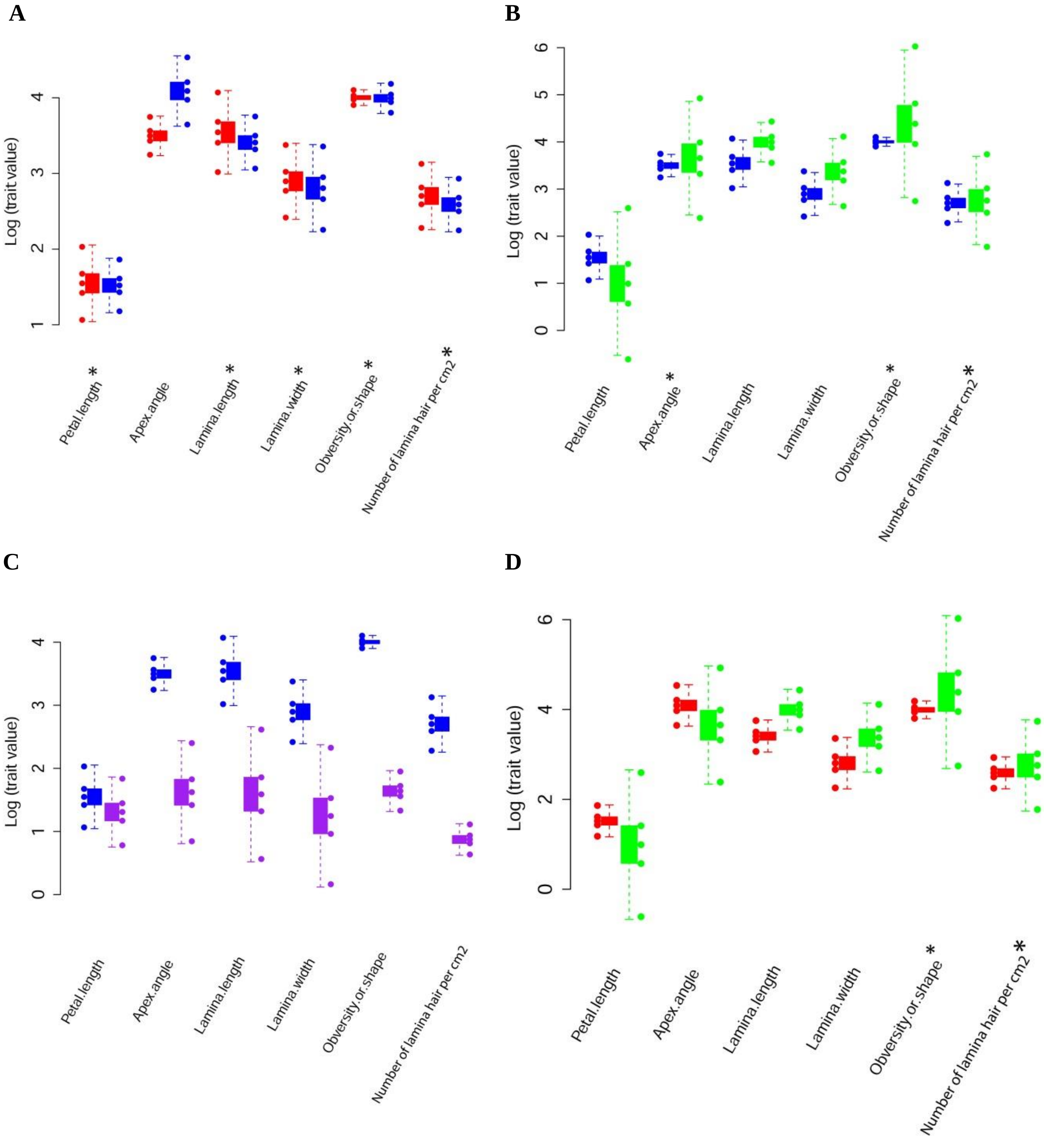
C. β*=1 in C1 and C4

D. β*=0.98 in C2 and C3

E. β*=1 in C2 and C4

F. β*=1 in C3 and C4

Figure S50. Values of the proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Ferrovestitum clade compared to the cutoff 0.9 necessary for the four GMM morphospecies to be distinct.


A
B
C
D
Log (trait value)
1
2
3
4
0
5
6
Petal.length
Apex.angle
Lamina.length
Lamina.width
Obversity.or.shape
Number of lamina hair per cm2
*

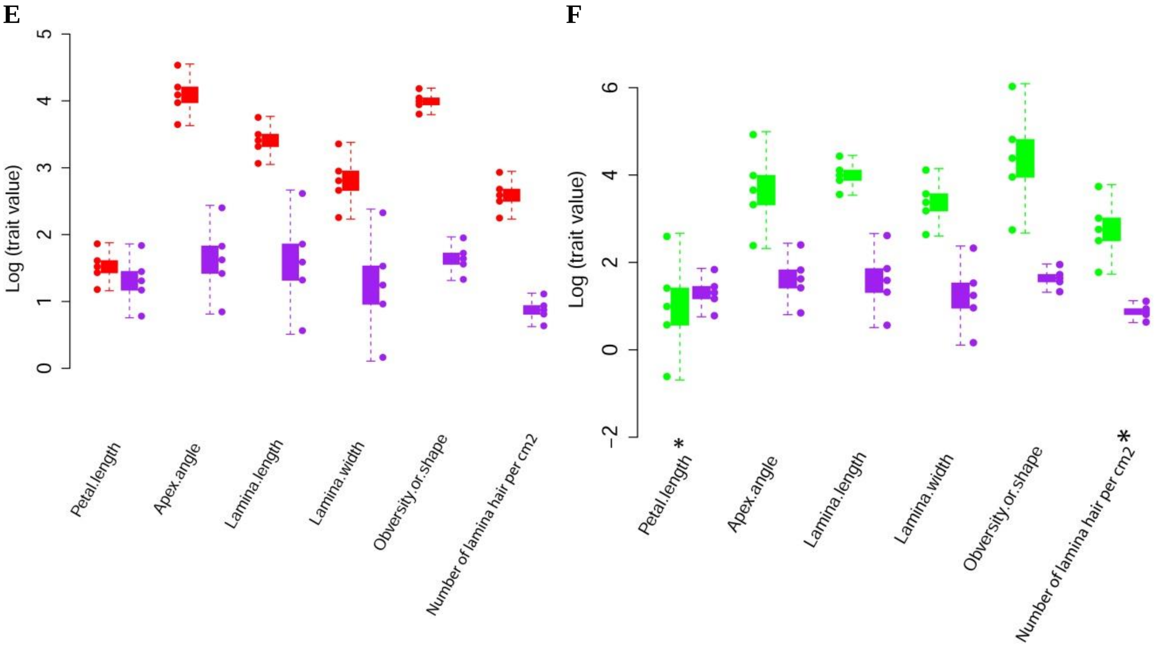


Figure S51. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in clade Ferrovestitum. C1 (blue boxplot), C2 (red boxplot), C3 (green boxplot); C4 (purple boxplot). **A.** C1&C2, **B**. C1&C3, **C.** C1&C4, **D.** C2&C3, **E.** C2&C4, **F.** C3&C4. Overlapping characters are marked with asterisks.

CLADE REVOLUTUM

GMM morphospecies in the Revolutu clade

Clade Revolutum has 82% bootstrap support and the 8 specimens from the molecular study contained are assignable to 2 morphogroups, *Psorospermum revolutum* and *P. cf lanceolatum*, (Fig. S52).

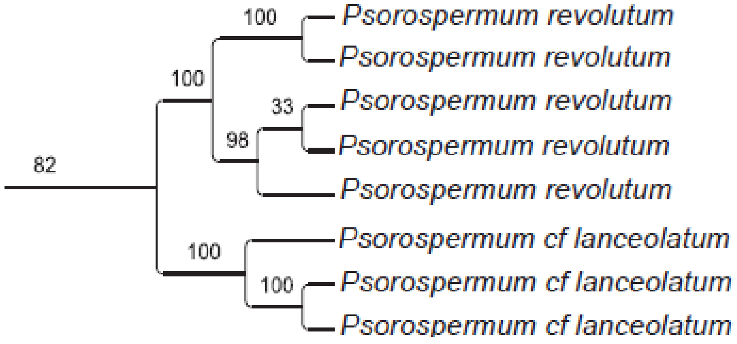


Figure. S52. Expanded clade Revolutum retrieved from the ML molecular phylogeny of *Psorospermum*. Numbers above the braches are bootstrap support values

These specimens were added to 30 herbarium specimens previously assigned to these 2 morphogroups. A total of 38 herbarium specimens are used in this particular dataset. The two morphogroups were confirmed with LDA (Wilks's Lambda = 0, p-value < 0.00001). There is no overlap of the histograms of the value of the discriminant functions (Fig. S53).

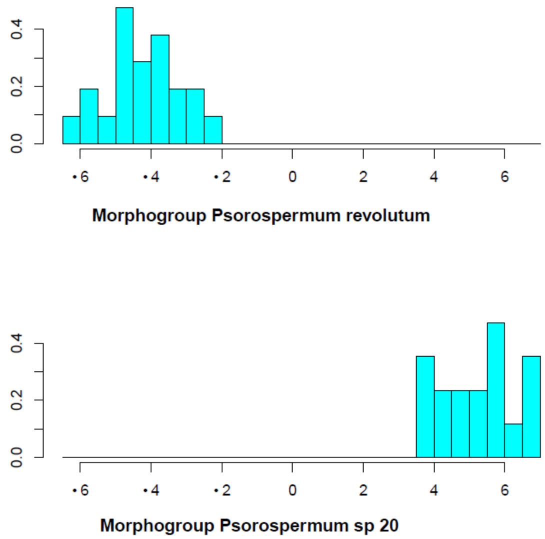


Figure S53. Histogram of the discriminant function's values for the samples from the two morphogroups in the Revolutum clade.

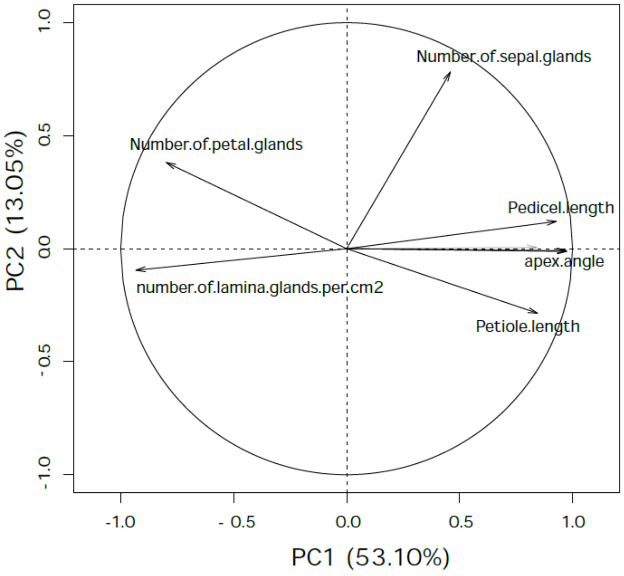


Figure S54. Principal Component Analysis (PCA) vector scatterplot of the dataset in the Revolutum clade.

The best model of GMM analysis yielded 2 morphospecies. Fig. S55 summarizes the output of Mclust models (table on the left), and highlights the identification of two morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model is indicated by the junction of the vertical and horizontal dotted lines and the number of morphospecies is indicated by the vertical dotted line (Fig. S55). Six characters influence significantly the loadings of the observations in the PCA vector scatterplot are: lamina apex angle, number of glands per cm^2; petiole length; pedicel length; number of sepal glands; and number of petal glands (Fig. S54; Appendix Table S1 for complete list). These six variables are used to perform the GMM analyses. Morphospecies are shown in a two-dimensional (2-D) morphospace scatterplot and an example is shown in Fig. S56. All the pairwise combinations of the six variables are shown in 2-D scatterplots in Appendix S10.

Figure S55. Outputs of GMM analysis of the Revolutum clade.

n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N. ind.: number of individuals per morphospecies.

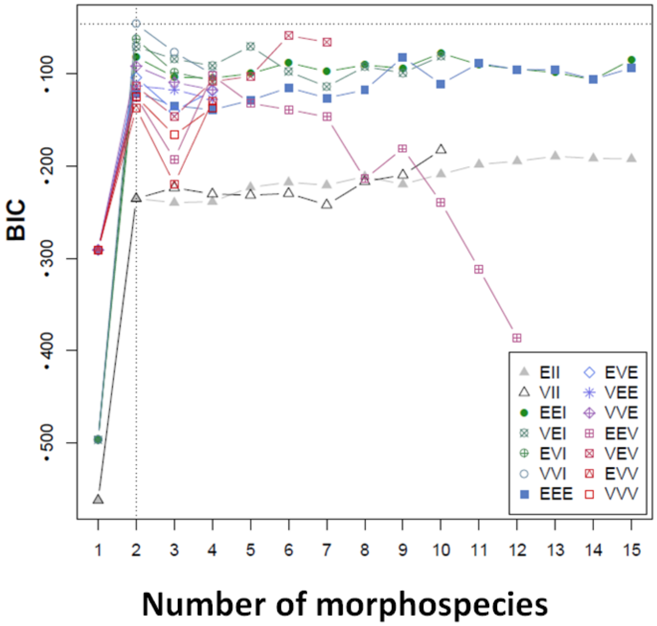
Mclust model with 2 morphospecies:

Diagonal, equal shape (VVI).

See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
46.50	38	33	-27.03

Morphospecies	1	2
N . ind.	21	17



The bimodal slope of the PDF is shown in Fig. S57.B, and emphasized with the change of slope sign in the graph line above the PDF. There is a gap between the two morphospecies, and the phenotypes of C1 do not overlap with those of C2 (β*values = 1), and β* is above the cutoff 0.9 of the proportion of individuals with non-overlapping phenotypes (Fig. S58).The quartiles of univariate distribution of the characters fit the simulation of the non-overlapping phenotypes of the two morphospecies except the character number of sepal glands (Fig. S59).

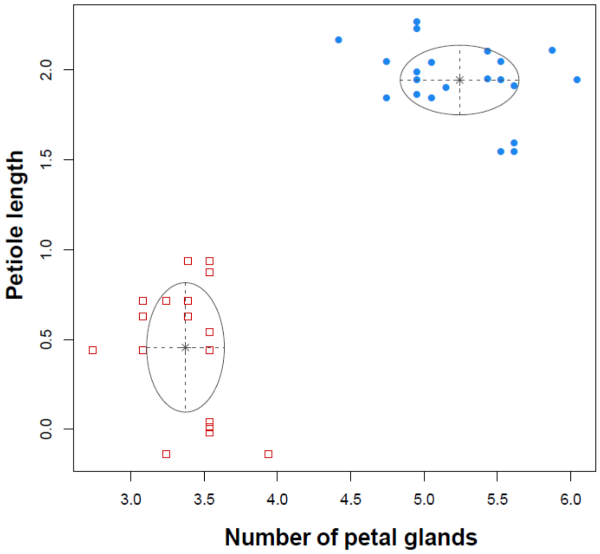
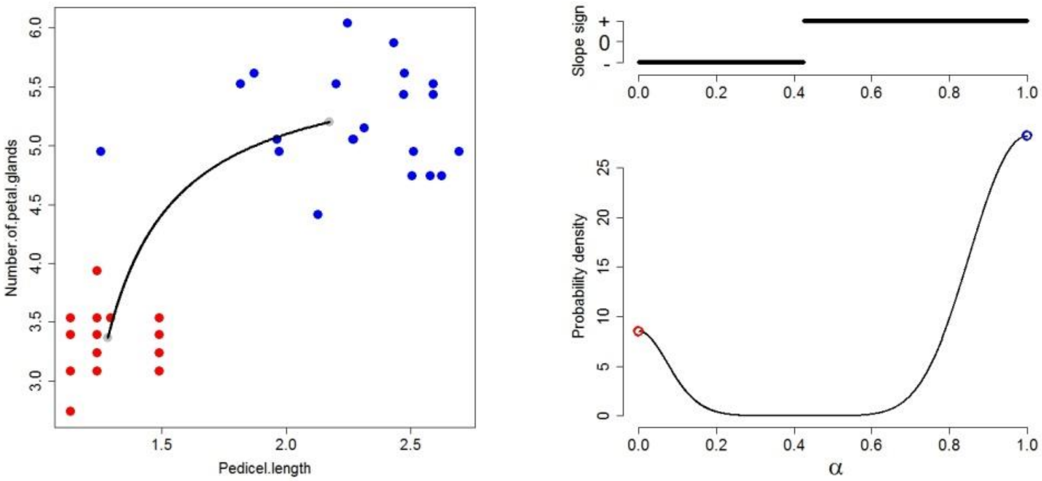


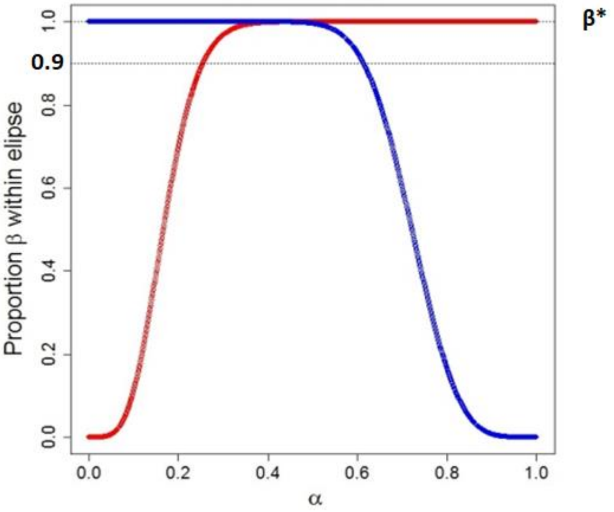
Figure S56. Bivariate plot of the Revolutum clade. Red: morphogroup *Psorospermum sp 20*; blue: morphogroup *Psorospermum revolutum*



A. RM (black line) between the modes of C1 and C2

B. Slope of the PDF and slope sign at various points along the RM

Figure S57. Ridgeline Manifold (RM) and the probability density function (PDF) slope in the Revolutum clade. **A.** RM = Ridgeline Manifold; **B.** Red and blue circle= modes of the two morphospecies.



C. β*=1 in C1 and C2

Figure S58. Value of the proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Revolutum clade compared to the cutoff 0.9 necessary for the two GMM morphospecies to be distinct.

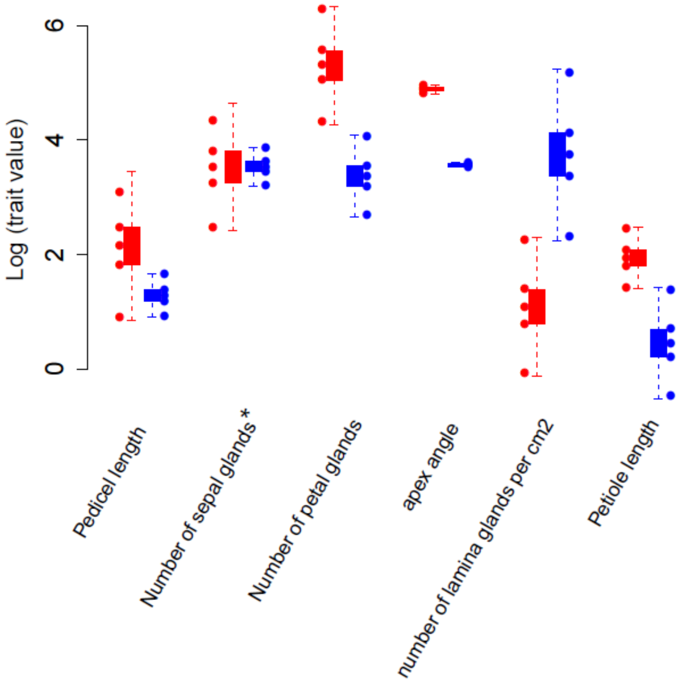


Figure S59. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analyses (points) and the simulated non-overlapping phenotypes in C1 (red boxplot) and C2 (blue boxplot). Overlapping characters are marked with asterisks.

CLADE CERASIFOLIUM

GMM morphospecies in the Cerasifolium

Clade Cerasifolium has 98% bootstrap support, and the 12 specimens from the molecular study contained are assignable to 4 morphogroups (*Psorospermum malifolium*, *P. sp22*, *P. sp2*, and *P. cerasifolium*) (Fig. S60). A total of 29 herbarium specimens are used in this particular dataset.

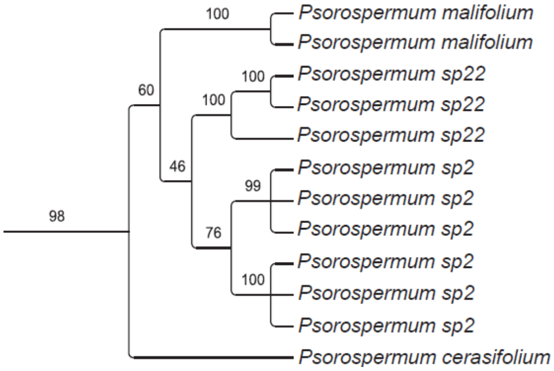
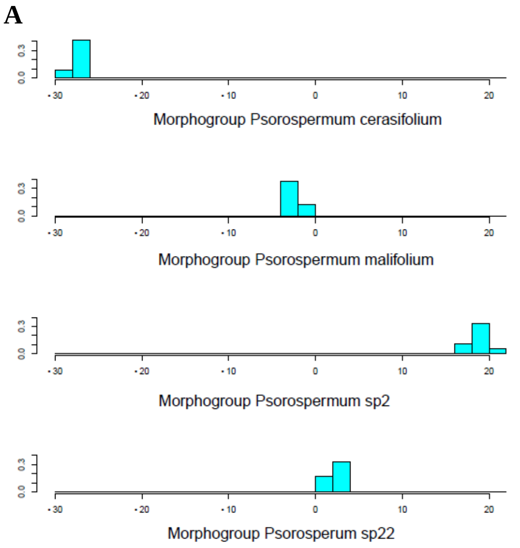


Figure S60. Expanded clade Cerasifolium retrieved from the ML molecular phylogenetic analysis of *Psorospermum*. Numbers above the braches are bootstrap support values.

These specimens were added to 17 herbarium specimens previously assigned to these 4 morphogroups. The four morphogroups were confirmed with LDA (Wilks's Lambda = 0.00004, p-value < 0.0001). There is no overlap of the values of the discriminant functions (Fig. S61.A).



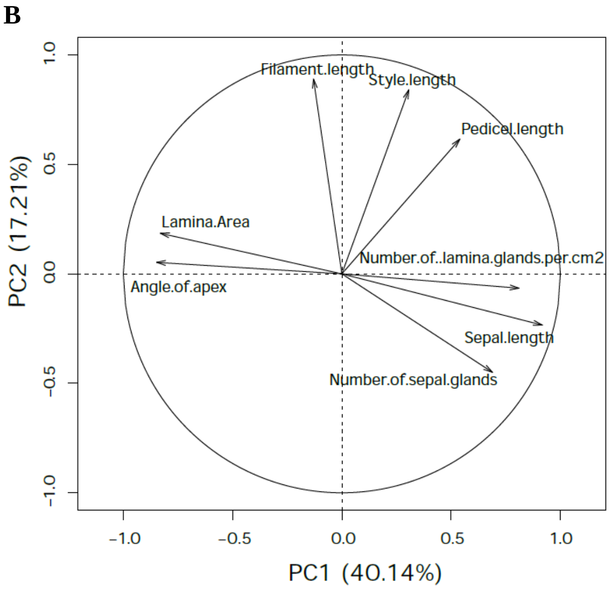


Figure S61. A. Histogram of values of the first discriminant function for the samples from the four morphogroups in the Cerasifolium clade.

Figure S61. **B**. Eight morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in the Cerasifolium clade

The best model of the GMM analysis yielded three morphospecies. Fig. S62 summarizes the output of the Mclust models (table on the left), and highlights the identification of 3 morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model is the one that has the highest BIC value, it is indicated by the horizontal dotted line on the y-axis (Fig. S62); the number of morphospecies suggested by that model is indicated by the vertical dotted line on the x-axis (Fig. S62). The eight characters that influence significantly the loadings of the observations in the PCA vector scatterplot are: lamina area (size), angle of apex, and number of gland dots per cm^2; pedicel length; sepal length, number of sepal glands; filament length; and style length (Fig. S61.B; Appendix Table S1 for complete list).These eight variables are used to perform the GMM analyses. Morphospecies are shown in a two-dimensional (2-D) morphospace and an example is shown in Fig. S63. All the pairwise combinations of the eight variables are shown in 2-D scatterplots in Appendix S12.

Figure S62. Outputs of GMM analysis of the Cerasifolium clade.

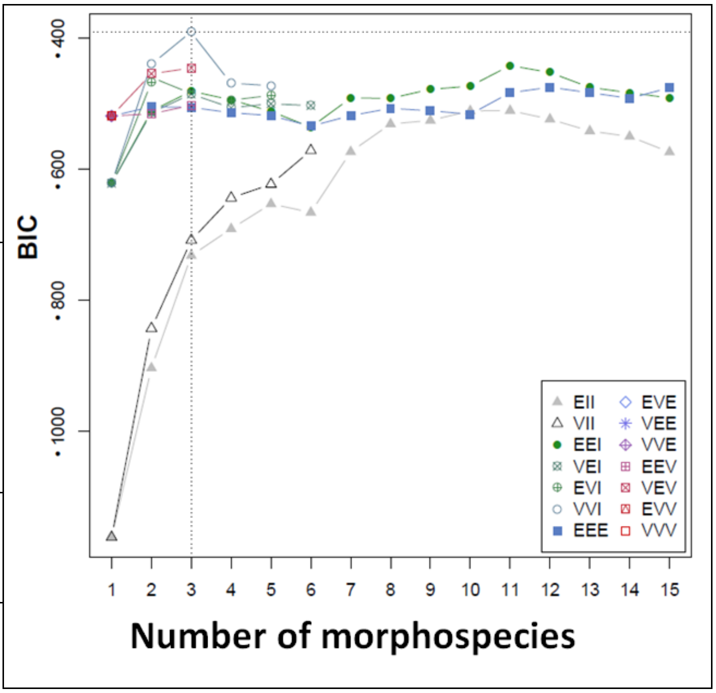
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N. ind.: number of individuals per morphospecies.

Mclust model with 3 morphospecies:

Diagonal, varying volume and shape (VVI).
See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
-115.85	29	48	-393.552

Morphospecies	1	2	3
N . ind.	8	15	6



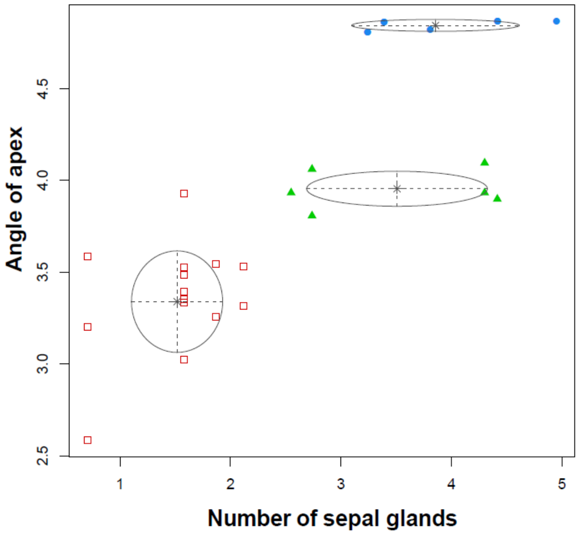


Figure S63. Bivariate scatterplot of the Cerasifolium clade. The ellipses superimposed on the plot correspond to the covariances of the morphospecies. Individuals in different morphospecies are indicated by different symbols and colors, filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3.

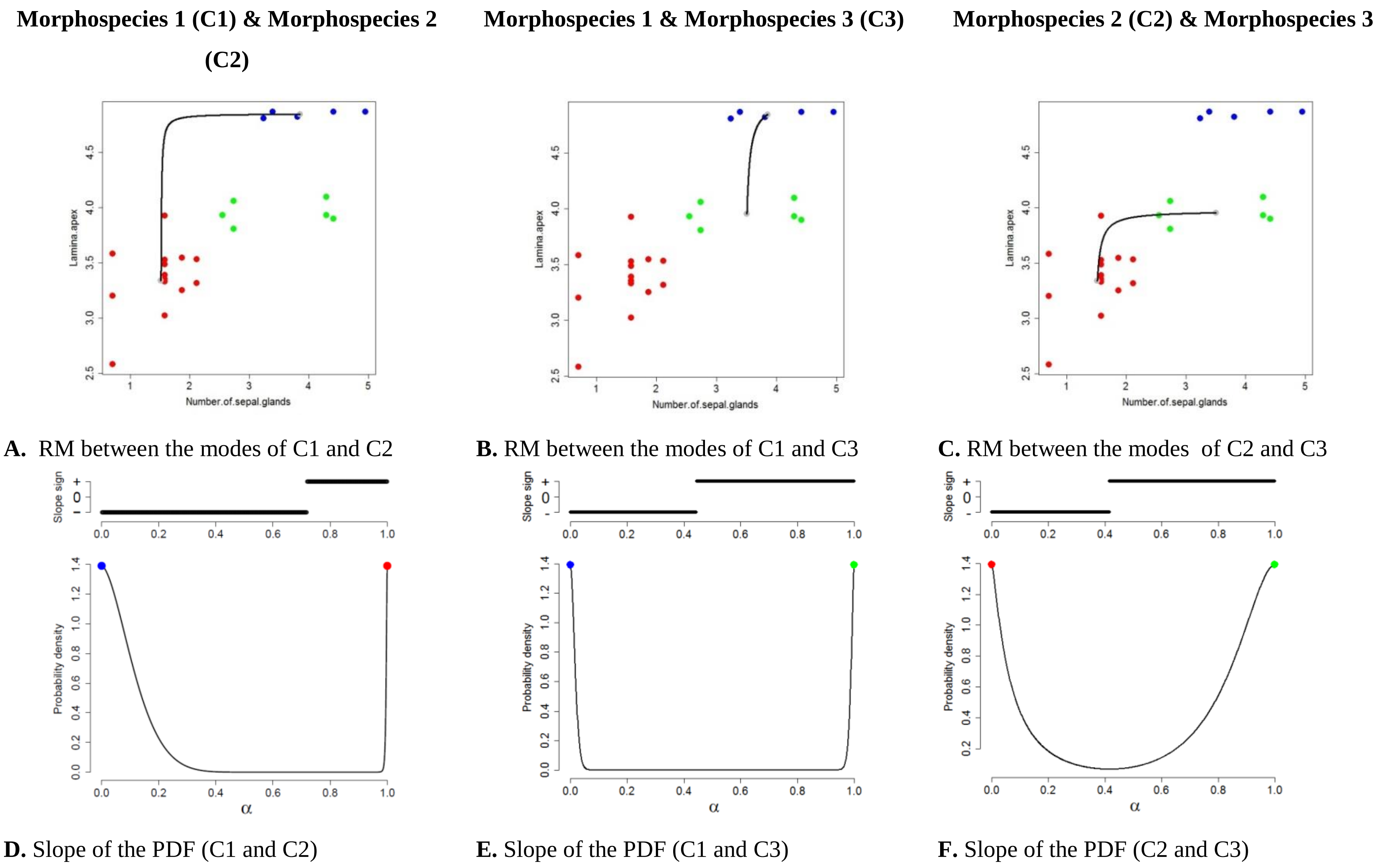
Quantification of gaps between morphospeciesin the Cerasifolium clade

The RM and the slopes of the PDFs of the three identified morphospecies are shown in Fig. S64.A-C and Fig. S64.D-F respectively. The PDFs are bimodal which indicates gaps between the morphospecies. The graph line above the PDF highlights the change of slope sign and confirms the bimodal nature of the PDF in morphospace.

Assessment of morphological characters that separate the morphospecies

The phenotypes of C1 do not overlap with those of C2 and C3 with β*values = 1 (Fig. S65.A & B); however there are overlapping phenotypes between C2 and C3, β* = 0.526 (Fig. S65.C) is below the cutoff 0.9 of the proportion of individuals with non-overlapping phenotypes.

The quartiles of univariate distribution of the morphological characters from the GMM analysis (points) fit the simulation of the non-overlapping phenotypes of C2 and C3 (boxplots). However, two characters, style length and apex angle, show overlap of the 1st, 2nd and 3rd quartiles of the GMM and the simulation in C1 and C2 (Fig. 66B); five characters, style length, sepal length, number of sepal glands, filament length and apex angle in C1 and C3 (Fig. 66C); and three characters lamina area, filament length and style length in C2 and C3 (Fig. 66A).


Morphospecies 1 (C1) & Morphospecies 2 (C2)

Morphospecies 1 & Morphospecies 3 (C3)

Morphospecies 2 (C2) & Morphospecies 3

A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

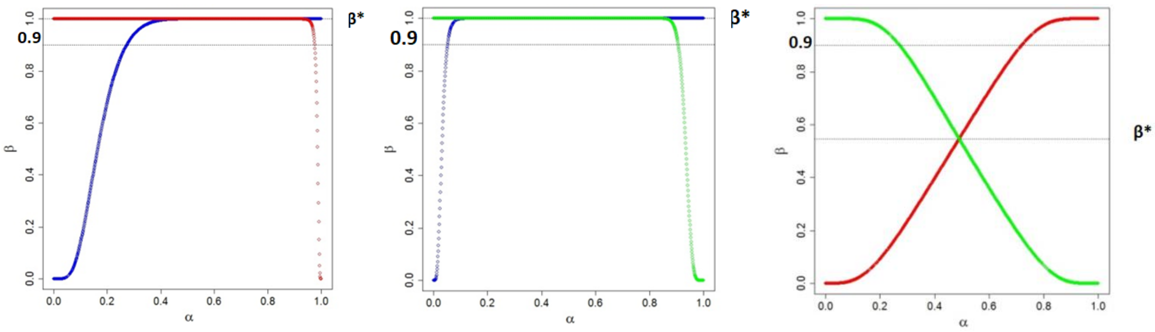
C. RM between the modes of C2 and C3

D. Slope of the PDF (C1 and C2)

E. Slope of the PDF (C1 and C3)

F. Slope of the PDF (C2 and C3)

Figure S64. Ridgeline Manifold (RM) and inference of gaps between the morphospecies. **A-C**: RM = continuous black line; **D- F**; blue, green and red points= modes of the morphospecies.



A. β*=1 in C1 and C2

B. β*=1 in C1 and C3

C. β*=0.526 in C2 and C3

Figure S65. Values of the proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Cerasifolium clade compared to the cutoff 0.9 necessary for the tthree GMM morphospecies to be distinct.

A

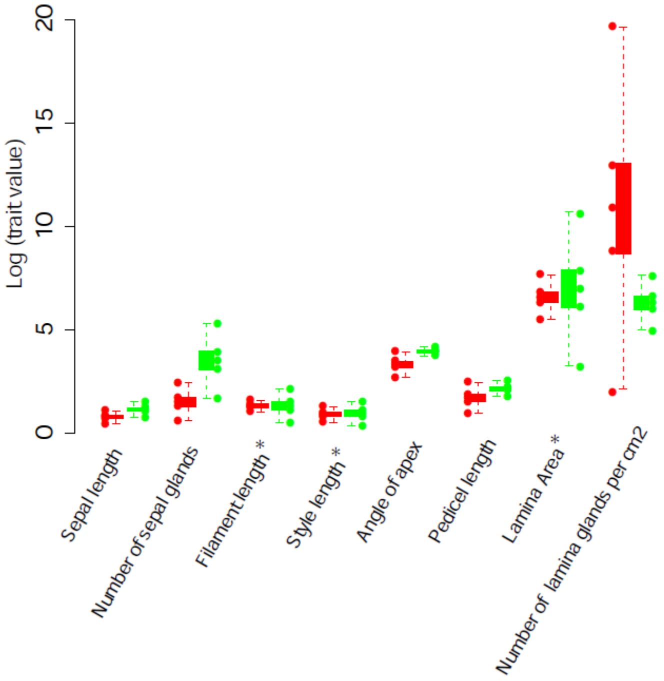


Figure S66 A. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in C2 (red boxplot) and C3 (green boxplot). Overlapping characters are marked with asterisk.

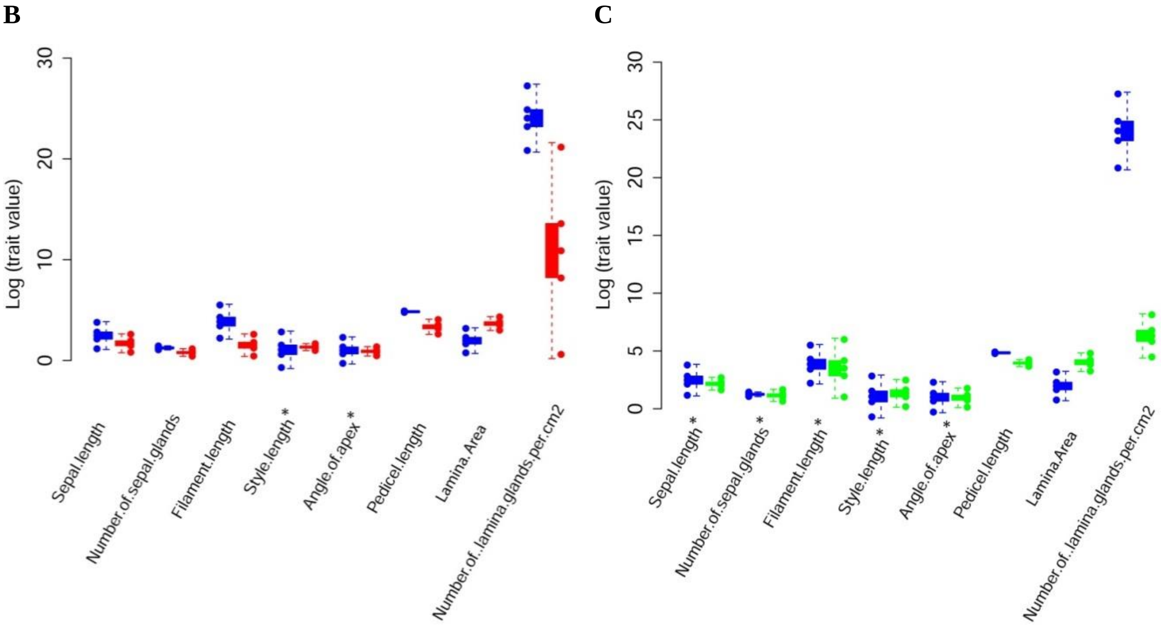


Figure S66 B. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in C1 (blue boxplot) and C2 (red boxplot). **C.** Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in C1 (blue boxplot) and C3 (green boxplot). Overlapping characters are marked with asterisks.

CLADE ANDROSAEMIFOLIUM

GMM morphospecies in the Androsaemifolium clade

Clade Androsaemifolium has 96% bootstrap support, and the 10 specimens from the molecular study contained are assignable to 3 morphogroups (*Psorospermum androsaemifolium*, *P. cf. androsaemifolium*, and *P. sp19*) (Fig. S67). These specimens were added to 20 herbarium specimens previously assigned to these 3 morphogroups. A total of 30 herbarium specimens are used in this particular dataset.

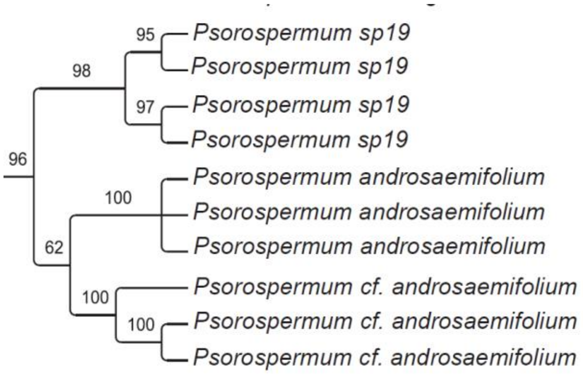

95
98
97
96
100
62
100
100
Psorospermum sp19
Psorospermum sp19
Psorospermum sp19
Psorospermum sp19
Psorospermum androsaemifolium
Psorospermum androsaemifolium
Psorospermum androsaemifolium
Psorospermum cf. androsaemifolium
Psorospermum cf. androsaemifolium
Psorospermum cf. androsaemifolium

Fig. S67. Expanded clade Androsaemifolium retrieved form the ML molecular phylogenetic analysis of *Psorospermum*. Numbers above the braches are bootstrap support values.

The morphogroups were confirmed with LDA (Wilks's Lambda = 0.014, p-value = 0.0001). There is no overlap of the values of the discriminant functions (Fig. S68 A).

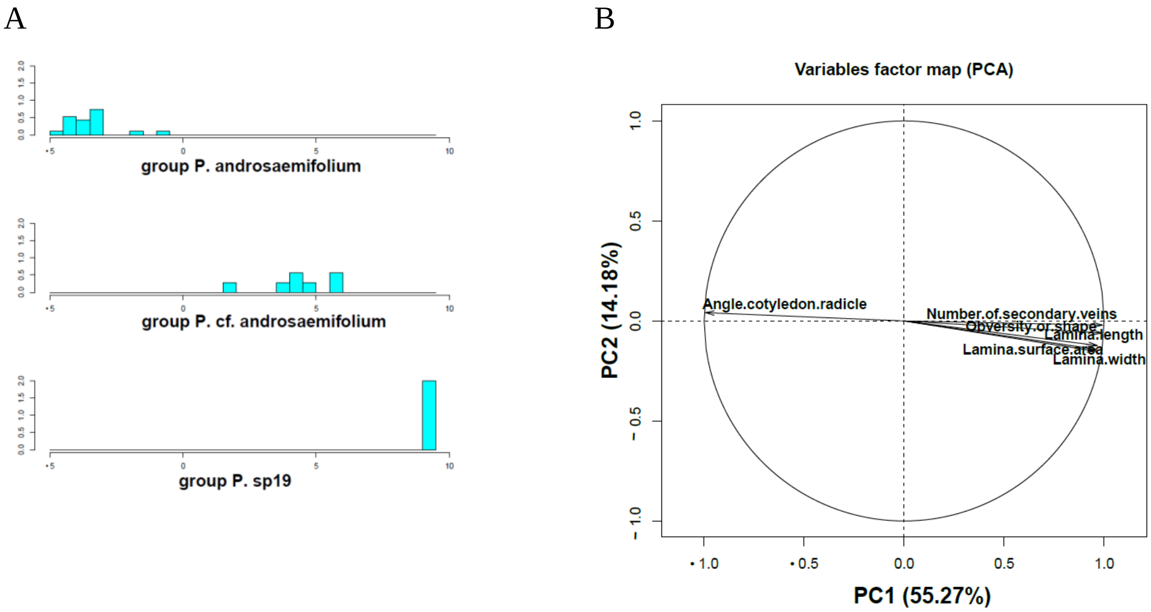


Figure S68. A. Histogram of values of the first discriminant function for the samples from the 3 morphogroups in the Androsaemifolium clade. **B.** Seven morphological characters that influence most significantly the loadings of the observations in the PCA scatterplot in the Androsaemifolium clade.

The best model of the GMM analysis yielded four morphospecies. Fig. S69 summarizes the output of the Mclust models (table on the left), and highlights the identification of four morphospecies in the dataset (graph on the right). The models are represented in different colors and shapes in the graph (details of the model names in legend box are given in Appendix Table S2); the best model, EEV (Ellipsoidal, Equal Volume), is the one that has the highest BIC value, it is indicated by the horizontal dotted line on the y-axis (Fig. S69); the number of morphospecies suggested by that model is indicated by the vertical dotted line on the x-axis (Fig. S69). Six characters influence significantly the loadings of the observations in the PCA vector scatterplot: angle cotyledon-radicle, lamina length, width, surface area, shape, and number of secondary veins (Fig. S68.B; Appendix Table S1 for complete list).These six variables were used to perform the GMM analyses. Morphospecies are shown in a two-dimensional (2-D) morphospace scatterplot and an example is shown in Fig. S70. All the combinations of pairs of variables are shown in 2-D scatterplots in Appendix S11.

Figure S69. Outputs of GMM analysis of the Androsaemifolium clade.

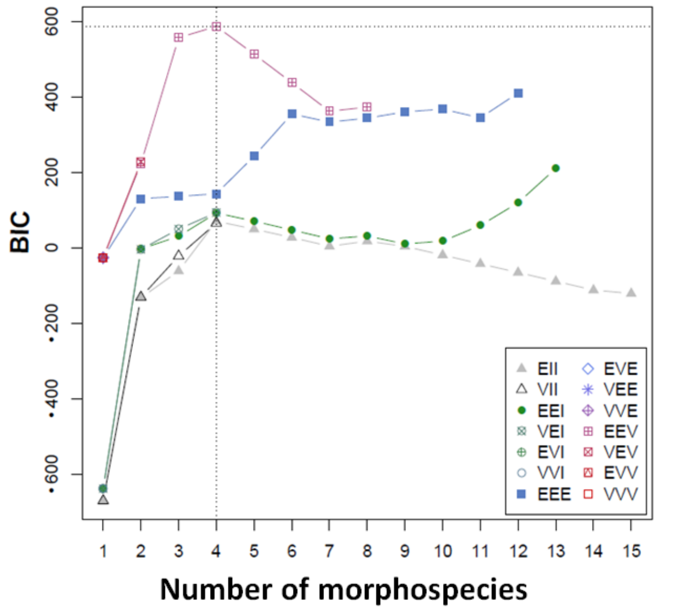
n: number of individuals; df: degree of freedom; BIC: Bayesian Information Criterion; N. ind.: number of individuals per morphospecies.

Mclust model with 4 morphospecies:

Ellipsoidal, Equal shape (EEV).
See Appendix Table S2 for the details of the models in the legend.

Log likelihood	n	df	BIC
578.29	30	120	586.68

Morphospecies	1	2	3	4
N . ind.	7	9	4	8



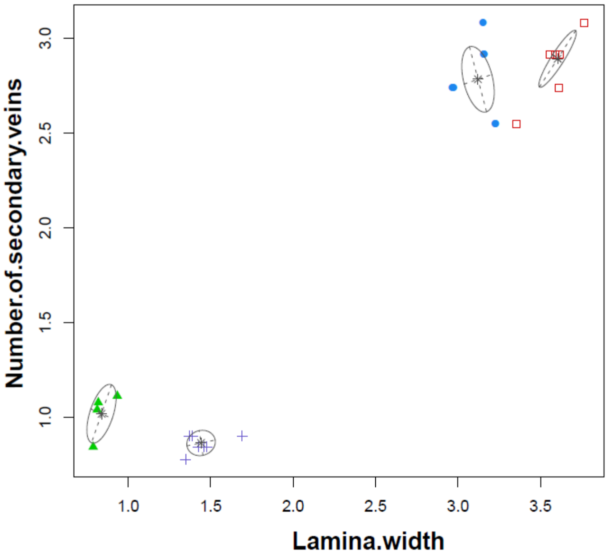


Figure S70. Bivariate scatterplot in the Androsaemifolium clade. The ellipses superimposed on the plot correspond to the covariances of the morphospecies. Individuals in different morphospecies are indicated by different symbols and colors, filled blue round: morphospecies 1; empty red square: morphospecies 2; green triangle: morphospecies 3, purple cross: 4.

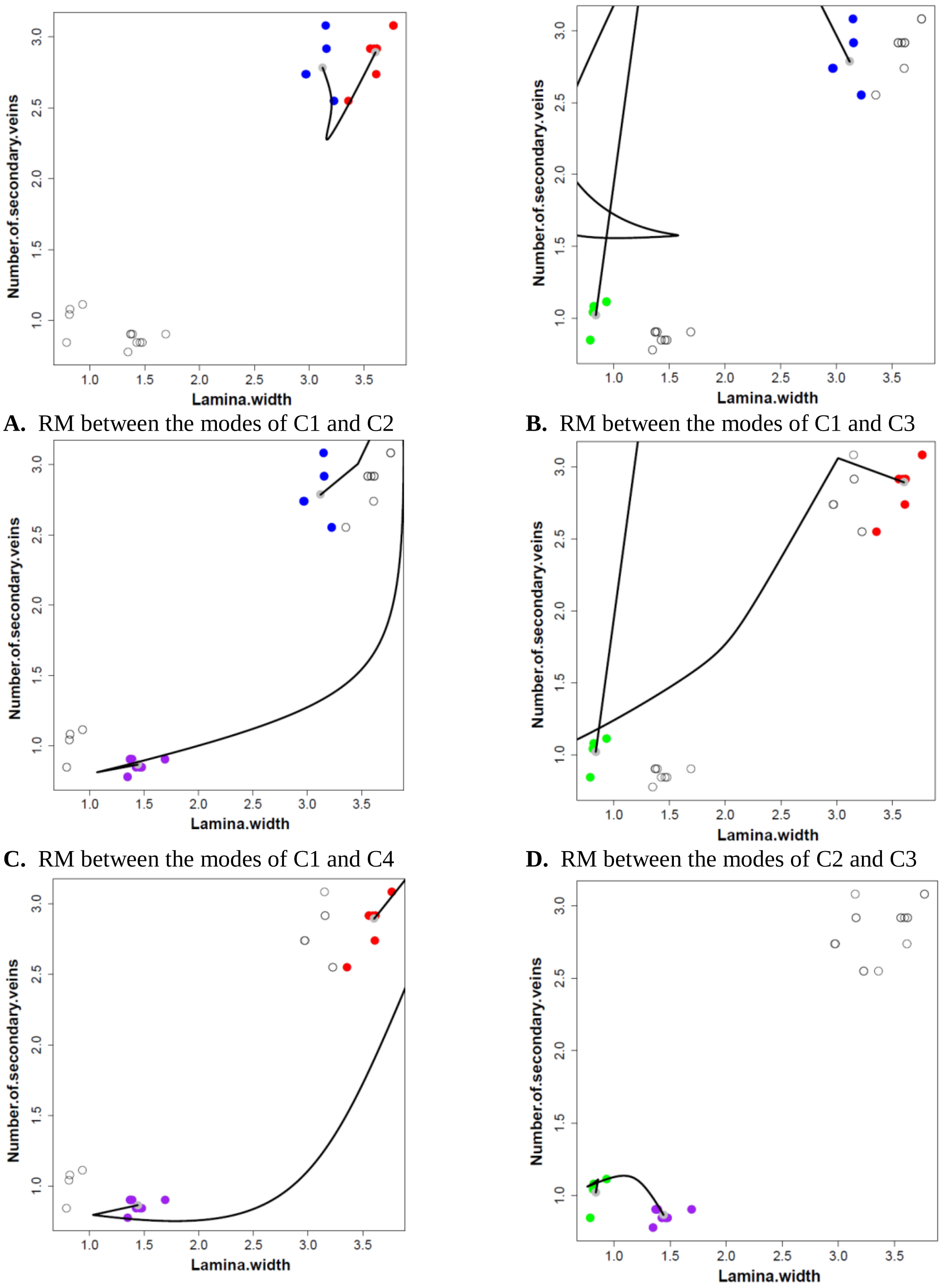
Quantification of gaps between morphospecies in the clade Androsaemifolium

The RM and the slopes of the PDFs of the three identified morphospecies are shown in Fig. S71.A-F and Fig. S72.A-F respectively. The PDFs are bimodal which indicates gaps between the morphospecies. The graph line above the PDF highlights the change of slope sign and confirms the bimodal nature of the PDF in morphospace (Fig. S72.A-F.)

Assessment of morphological characters that separate the morphospecies

The phenotypes of the 4 morphospecies do not overlap having β* values = 1, i.e. above the cutoff value 0.9 of the proportion of individuals with non-overlapping phenotypes (Fig. S73).

Fig. S74 shows that the quartiles of univariate distribution of the morphological characters from the GMM analysis (points) fit the simulation of the non-overlapping phenotypes of all morphospecies. Only two characters, lamina shape and the angle of the cotyledon to the radicle, overlap in the comparison of C1versus C2.



A. RM between the modes of C1 and C2

B. RM between the modes of C1 and C3

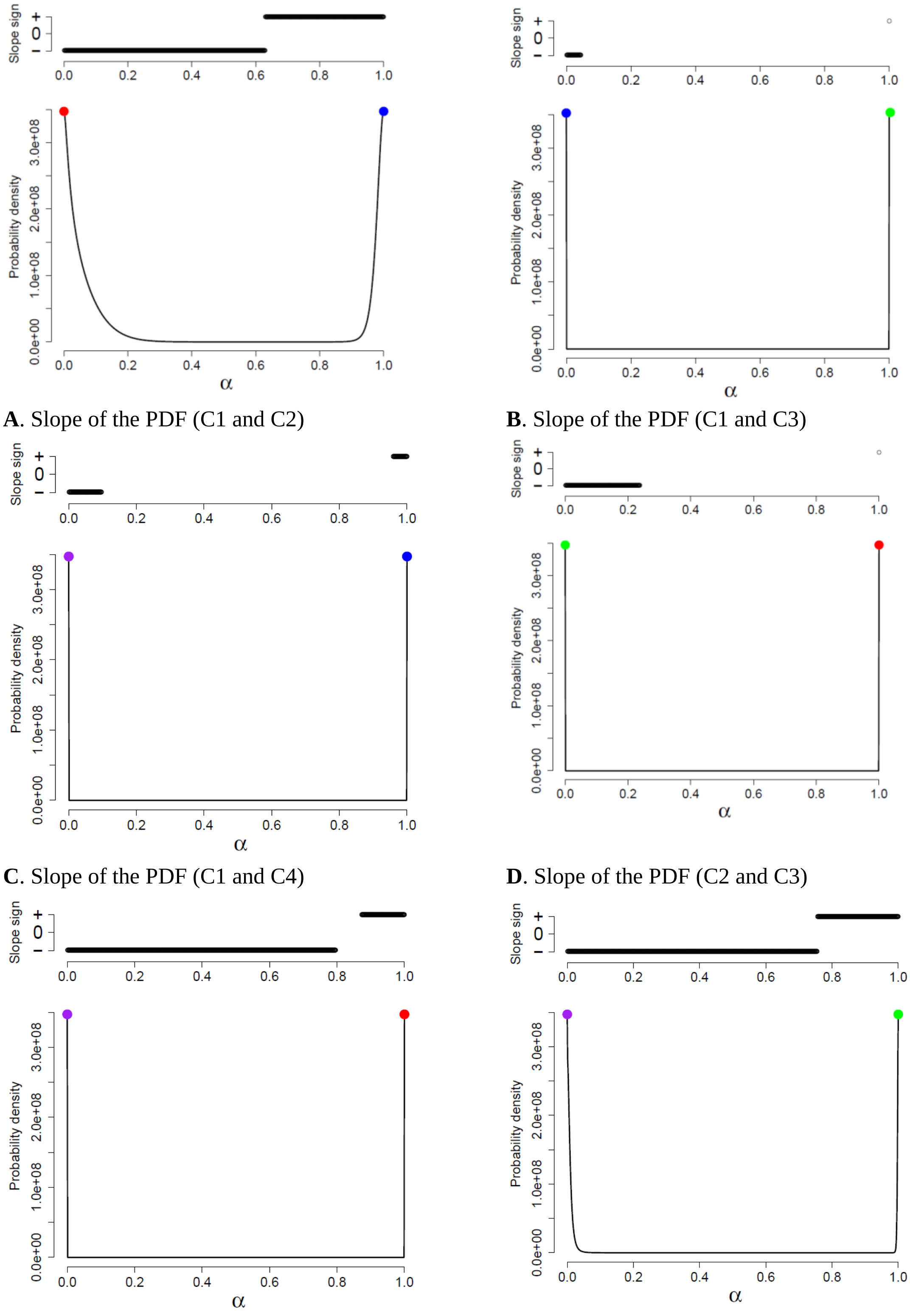
C. RM between the modes of C1 and C4

D. RM between the modes of C2 and C3

E. RM between the modes of C2 and C4

E. RM between the modes of C3 and C4

Figure S71. RM between the morphospecies in the Androsaemifolium clade. RM = continuous black line. Color of points correspond to the morphospecies, blue, red, green and purple are respectively C1,C2,C3,and C4. Only the pairs of morphospecies analyzed are colored.



A. Slope of the PDF (C1 and C2)

B. Slope of the PDF (C1 and C3)

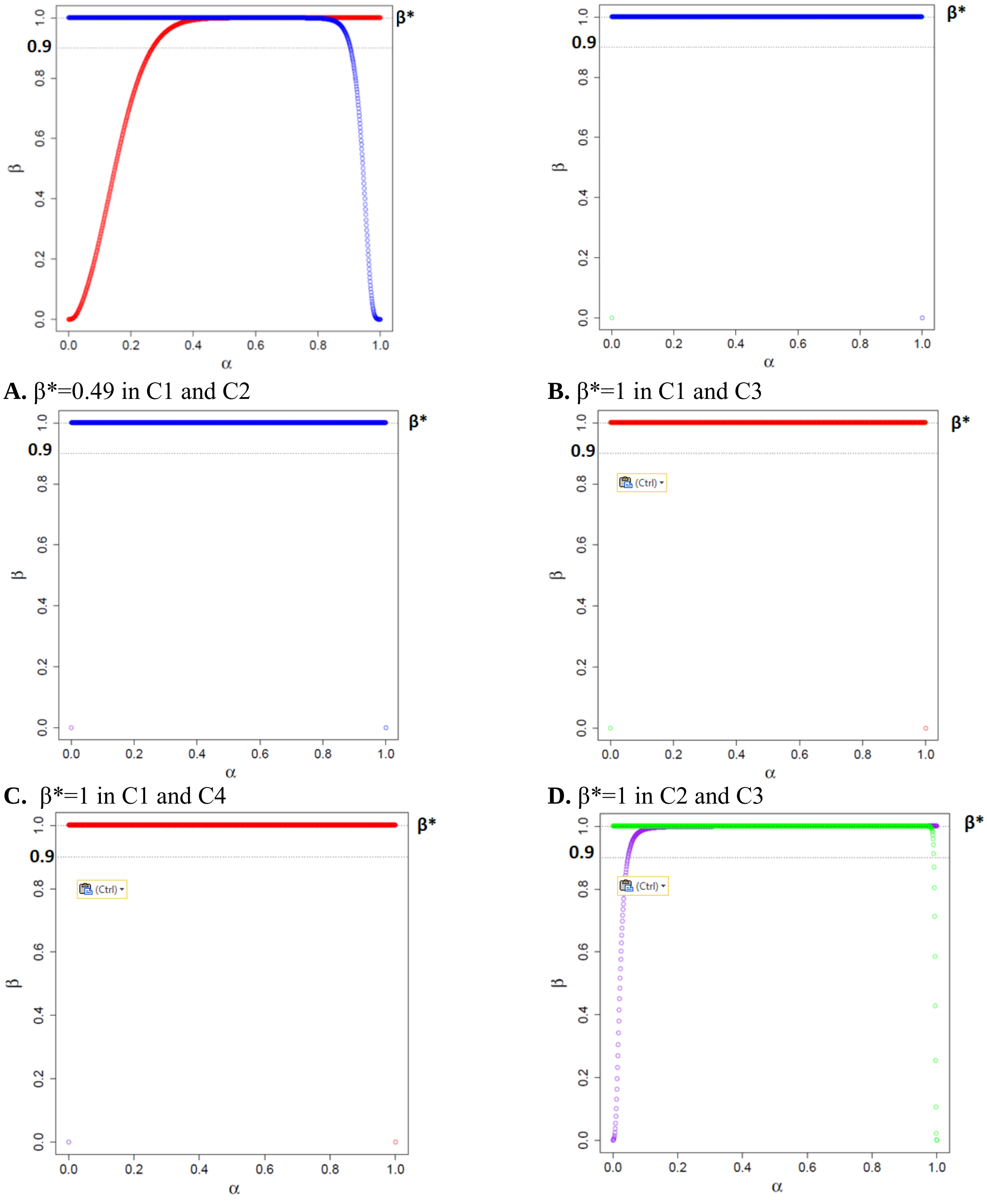
C. Slope of the PDF (C1 and C4)

D. Slope of the PDF (C2 and C3)

E. Slope of the PDF (C2 and C4)

F. Slope of the PDF (C3 and C4)

Figure S72. Inference of gaps between the morphospecies in the Androsaemifolium clade. Lines above the slopes indicates the length and the direction of the slopes. Colored points are the modes of the morphospecies, blue: C1, red: C2, green: C3, and purple C4.



A. β*=0.49 in C1 and C2

B. β*=1 in C1 and C3

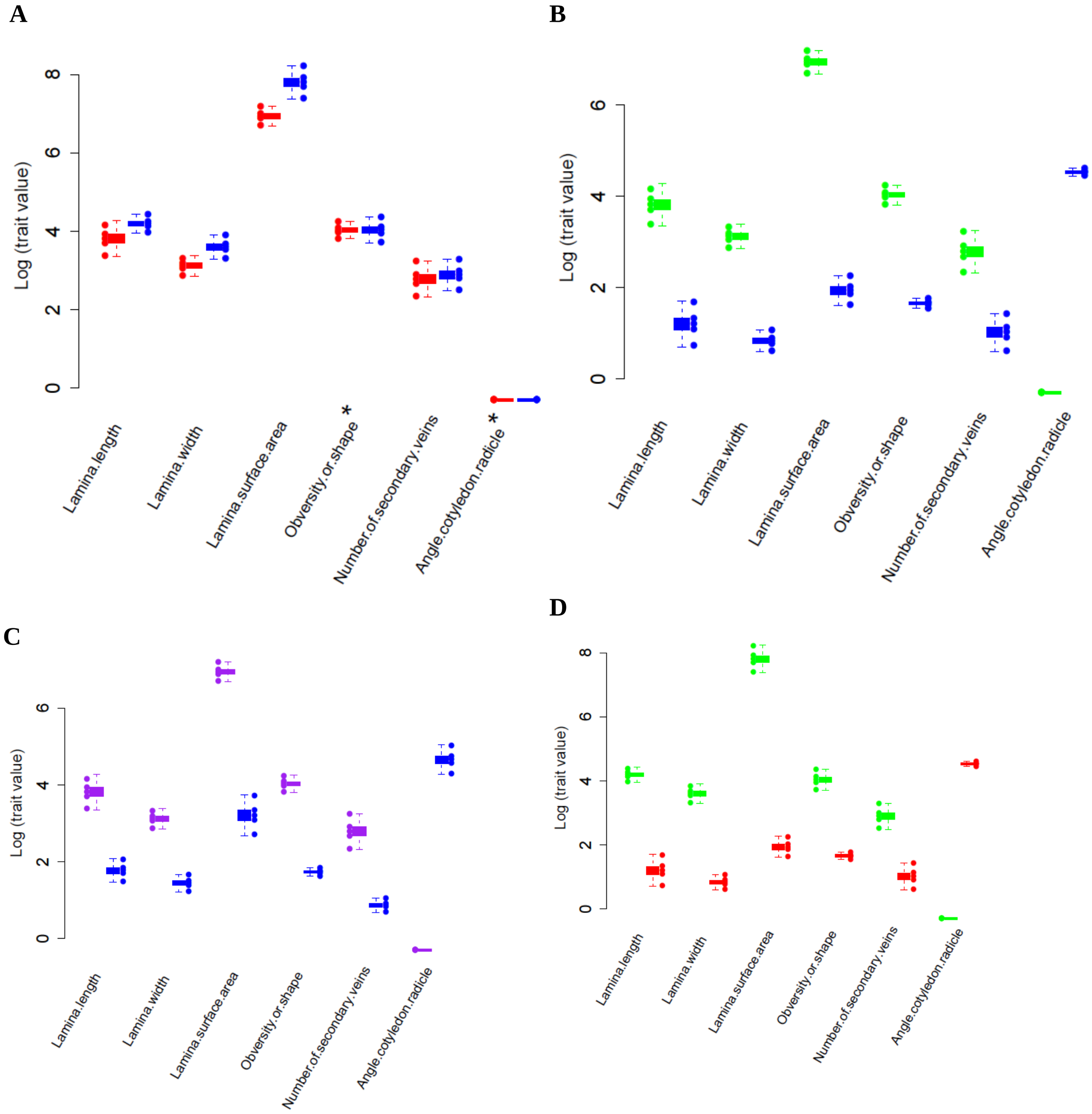
C. β*=1 in C1 and C4

D. β*=1 in C2 and C3

E. β*=1 in C2 and C4

F. β*=1 in C3 and C4

Figure S73. Values of the proportion of the individuals with non-overlapping phenotypes of the morphospecies, β*, in the Androsaemifolium clade compared to the cutoff 0.9 necessary for the three GMM morphospecies to be distinct. Blue: C1, red: C2, green: C3, and purple C4.


A
Log (trait value)
0
2
4
6
8
Lamina.length
Lamina.width
Lamina.surface.area
Obversity.or.shape *
Number.of.secondary.veins
Angle.cotyledon.radicle *
B
Log (trait value)
0
2
4
6
Lamina.length
Lamina.width
Lamina.surface.area
Obversity.or.shape
Number.of.secondary.veins
Angle.cotyledon.radicle
C
Log (trait value)
0
2
4
6
Lamina.length
Lamina.width
Lamina.surface.area
Obversity.or.shape
Number.of.secondary.veins
Angle.cotyledon.radicle
D
Log (trait value)
0
2
4
6
8
Lamina.length
Lamina.width
Lamina.surface.area
Obversity.or.shape
Number.of.secondary.veins
Angle.cotyledon.radicle

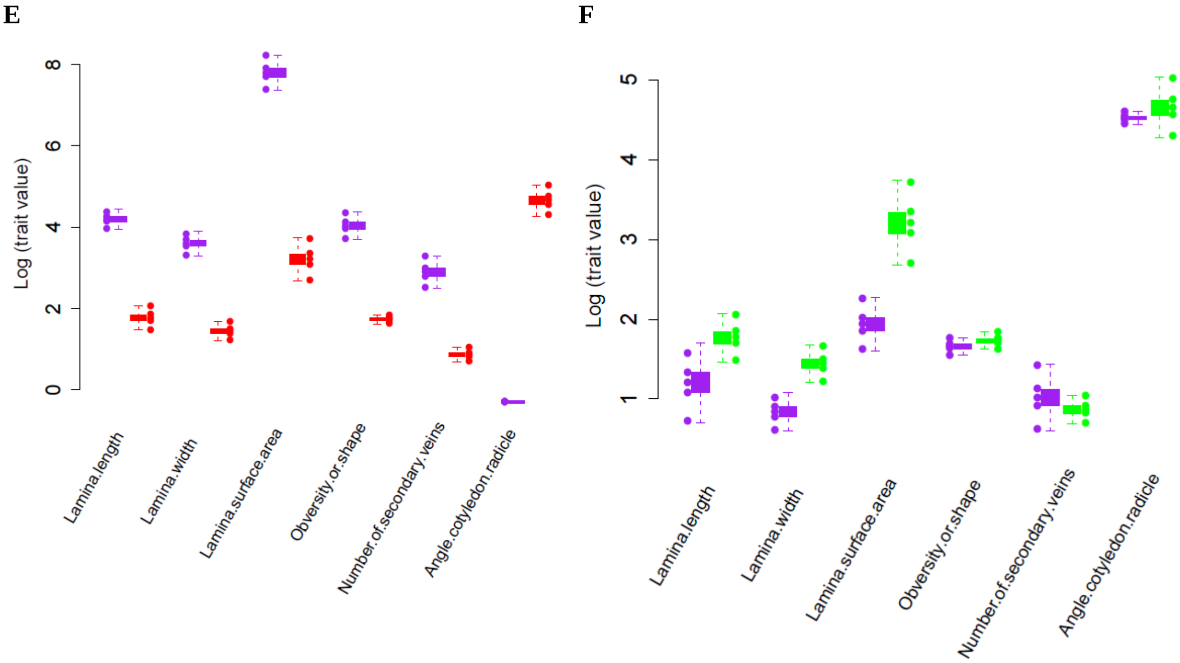


Figure S74. Comparison of the quartiles of the univariate trait distributions of the morphological characters from the GMM analysis (points) and the simulated phenotypes in the Androsaemifolium clade. C1: blue boxplot, C2: red boxplot; C3: green boxplot; C4: purple boxplot. Overlapping characters are marked with asterisks.

Towards the taxomic revision of the Malagasy *Psorospermum* (Hypericaceae): what can we retrieve from old literature?

Heritiana Ranarivelo[1,2]
[1]Department of Biology, University of Missouri–St. Louis, One University Blvd, St. Louis, MO
63121-4400, USA.
[2]Missouri Botanical Garden, PO Box 299, St. Louis, MO 63166-0299, USA.
Author for correspondence (hsrq98@mail.umsl.edu)

Abstract– Psorospermum Spach (Hypericaceae), an Afro-Malagasy genus, has 26 species in Madagascar. They were described by Perrier de la Bâthie in the series *Flore de Madagascar et des Comores* in 1951. Malagasy *Psorospermum* has not been revised since then. To help understand Perrier's work in the context of 21st century approaches to species delimitation, I conducted comparative morphometric analyses of two datasets. The first dataset consists of a matrix of 19 characters that Perrier mentioned and their measurements for 152 of the specimens he examined. The second dataset consists of a matrix of 33 characters and 314 specimens, including the 19 characters and 152 specimens of Perrier. I conducted morphometric analyses, Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), and Gaussian Mixture Model (GMM) to identify groupings in these datasets and I compared the results with recent species delimitation of *Psorospermum* that integrates molecular phylogenies and morphology. My results confirm that the characters and specimens used by Perrier do not provide enough information to delimit all the species he recognized. Additionally, my results highlight the importance of herbarium collections and the observation of additional characters to improve the output. However, species delimitation using GMM methods performs better within a well-defined universe rather than using methods to analyze large datasets as here.

Keywords– Linear Discriminant Analysis, Principal Component Analysis, Morphometrics, integrative taxonomy, Species delimitation, Perrier de la Bâthie.

Psorospermum (Hypericaceae) is a poorly known genus that occurs in Madagascar and mainland Africa. The exact number of species is uncertain; there might be as many as 50 (Stevens 2007), with at least 26 species endemic to the island of Madagascar alone (Perrier de la Bâthie 1951). Attention needs to be paid to Malagasy *Psorospermum*: as with most Malagasy plant genera, *Psorospermum* has not been revised for nearly 65 years. The last treatment was by Perrier de la Bâthie (1951) published by the Muséum national d'Histoire naturelle, Paris, in the series *Flore de Madagascar et des Comores*. This treatment was one of Perrier's later published volumes; he began work on the series in 1936 and left it unfinished at the time of his death in 1958 (*Humbert* 1958).

Perrier's treatments were mostly based on herbarium specimens, but he himself collected ca. 20,000 specimens in different areas of Madagascar from 1896 to 1932 (*Humbert* 1958). From 1927 to 1936, prior to beginning the *Flore de Madagascar et des Comores*, Perrier wrote preliminary notes for the descriptions of some of those specimens (*Humbert* 1958), but to my knowledge there is no record of the duplicates he examined, nor of any measurements he might have made of them; at that time, all or most of the duplicates of the specimens were likely held in P (abbreviations follow the *Index Herbariorum*, Thiers, [continuously updated], http://sweetgum.nybg.org/science/ih/). Thus it was possible to build a data matrix, the PB dataset, by measuring the characters Perrier mentioned in his 1951 revision and using the specimens that he himself cited and utilized. (See methods.)

A study of species delimitation of Malagasy *Psorospermum* was conducted using techniques integrating morphometric methods and molecular data (Ranarivelo et al. in preparation), where twenty-seven species were recognized. Gaussian Mixture Models (GMM) combined with Principal Component Analysis (PCA) were used to analyze the data, and additional characters and specimens were added to those mentioned by Perrier using an integrative taxonomy approach (Ranarivello et al. in preparation). I built a data matrix of all the characters and specimens I examined including the characters Perrier mentioned, the ALL dataset. Using the two datasets (PB and ALL), comparative morphometric analyses were conducted to address the questions:
(1) What characters or group of characters used by Perrier best explain the variation within the PB dataset? (2) How many species can be recognized from the PB dataset using his characters? (3) Do these species match the species Perrier described? (4) How do additional characters and specimens affect the groupings? (5) How do Perrier's species correspond to the species delimited by the integrative taxonomic approach?

First, I ran morphometric analyses using the PB dataset. Linear Discriminant Analysis (LDA) was used to test the species recognized by Perrier. PCA was used to select the variables that best explain variation in the PB dataset, and GMM was used to identify putative species in this particular dataset. Second, I ran the same analyses (LDA, PCA and GMM) using only Perrier specimens but with additional characters, the PB33 dataset. Third I ran the same morphometric analyses using the ALL dataset. (See materials and methods for details.) Species delimited by both datasets are compared to one another and to the species delimited by Ranarivelo et al.

Psorospermum is usually a shrub or small tree that grows in various habitats, rainforests, deciduous forests, or woodland areas, in Madagascar. It is easily recognized by its opposite entire or crenulate leaves, hairy buds, lack of stipules, and presence of yellow exudate. The lamina can be glabrous or have stellate hairs, and black hypericin glands are usually visible on both sides, appearing as black dots. Hypericin is a bioactive compound common also in *Hypericum* (St. John's Wort species) where it accumulates in dark glands in different parts of the plant, e.g., leaves, petals or anthers. Hypericin is commonly used in the pharmaceutical industry, usually as an antidepressant (Zobayed et al. 2006). The flowers are 5-merous; peduncles, pedicels and sepals are often hairy and have hypericin glands as well, and those on the sepals are visible as stripes running the length of the sepals; the sepals are often persistent in young fruits. The petals are white with hairs on their abaxial surface and, depending on the species, hypericin gland stripes and dots are present as well. The stamens are in phalanges; the number of anthers of each phalange ranging from 3 to 15 depending on the species. The anthers are white and in some species there is a black gland dot on their tips; the filaments have thin unicellular hairs. There are five glabrous yellow staminodes. The gynoecium has one or two ovules per carpel, and there are five yellow stigmata. The fruits are spherical greenish berries and the seed coat often has hypericin gland dots.

Perrier's treatment of the Malagasy *Psorospermum* was based on observation of 209 herbarium specimens that he placed into 26 species. However, 12 of the species were described from fewer than three herbarium specimens and seven of these were described from a single specimen (Fig. 1). Perrier mentioned a total of 8 foliar and 11 floral characters in his identification key and species descriptions (Table 1). There is no evidence that Perrier measured all of the characters; he reported measurements of a few foliar characters, such as lamina length and width, but there is no record of measurements of individual specimens. However, 17 of his characters are quantifiable by measurements and counts; two characters (lamina margin and position of glands on the petals) are categorical.

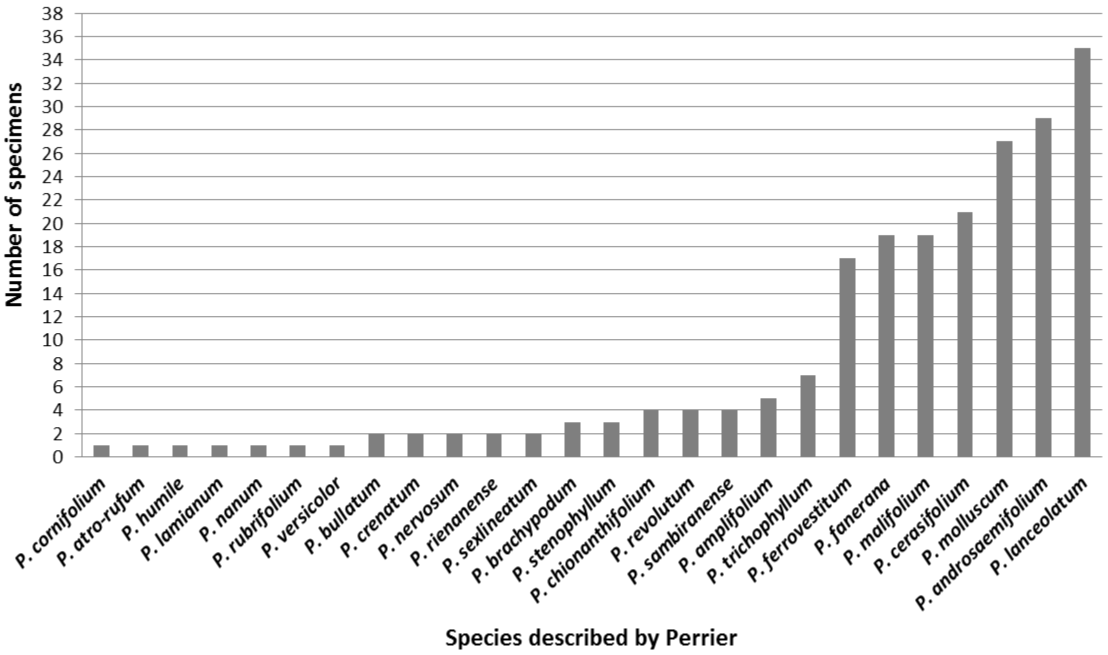


FIG 1. Number of specimens per species described in Perrier de la Bâthie (1951).

MATERIALS AND METHODS

***Datasets and characters*–** The PB dataset includes 152 individuals of the 209 specimens observed by Perrier; the missing specimens are very old collections that could not be borrowed from the herbaria in which they are housed. Twenty-five of the 26 species recognized by Perrier de la Bâthie (1951). are included in the dataset,and 19 characters that he mentioned are measured for these specimens. The PB33 dataset also includes 152 individuals but 33 characters. The only species that is not included is *P. cornifolium* which was described from a single specimen (*Commerson s.n.*). The ALL dataset has a total of 314 specimens and 33 characters. Characters were classified as either measures, transformed variables, or count variables (Table 1). The only transformed variable is lamina shape, which is the ratio of the length to the widest point and the total length of the lamina, which captures an important difference between lamina shapes. Two characters were categorical: lamina margin, and the position of glands on the petals (Table 1). The software ImageJ (Rasband 2012) was used to measure the area of the lamina, the angle of the apex of the lamina, the area of the cotyledon, and the angle formed by the cotyledon and radicle. Other embryo characters, such as length and width, were measured directly through the microscope. Under the microscope, the hypericin glands appear as black dots and are easily distinguishable from similar dots caused by parasites or disease; fungi can appear as black dots, but can be easily removed by scratching the leaves with a needle. For further details of the characters and their measurements, see Ranarivelo et al. All data have been submitted to Dryad Digital Repository (Dryad Digital Repository http://dx.doi.org/xx.xxx/dryad.xxxx).

TABLE 1. List of the morphological characters measured.

	Character measured	Type of variable	Mentioned in Perrier de la Bâthie (1951)
	Foliar characters		
1	Lamina apex	Measure	Yes
2	Lamina length (L)	Measure	Yes
3	Lamina width	Measure	Yes
4	Lamina area	Measure	Yes
5	Lamina length from the base to the widest point (LW)	Measure	No
6	Lamina shape (LW/L x 100)	Measure	Yes
7	Number glands per cm^2 on the abaxial lamina surface	Count	No
8	Petiole length	Measure	Yes
9	Number of secondary veins	Count	Yes
10	Number of hair per cm^2 on the abaxial lamina surface	Count	Yes
11	Type of lamina margin	Categorical	Yes
	Floral characters		
12	Number of flowers per cyme	Count	Yes
13	Pedicel length	Measure	Yes
14	Sepal length	Measure	No
15	Petal length	Measure	Yes
16	Number of sepal glands	Count	No
17	Length of petal glands	Measure	Yes
18	Number of glands on petals per cm^2 square	Count	No
19	Staminode length	Measure	No
20	Filament length	Measure	Yes
21	Number of anthers	Measure	Yes
22	Number of anther glands	Count	Yes

23	Style length	Measure	Yes
24	Ovary length	Measure	No
25	Number of ovules per carpel	Count	No
	Embryo characters		
26	Cotyledon area	Measure	No
27	Cotyledon length	Measure	No
28	Cotyledon width	Measure	No
39	Radicle length	Measure	No
30	Radicle width	Measure	No
31	Cotyledon-radicle angle	Measure	No

Character states of two categorical variables

Character	**Character states**	**Mentioned in Perrier de la Bâthie (1951)**
32 Lamina margin	Crenulate	Yes
	Entire	Yes
33 Position of glands on petals	Absent	Yes
	Tip	Yes
	Base	Yes
	Tip and base	Yes
	All over	Yes

Morphometric analyses– I used LDA to cross-validate Perrier and Ranarivelo et al.'s species, and I calculated the accuracy of the discrimination of those species by LDA. Values of the variables are $\log_{10}$ transformed, and count variables are transformed as square root +1/2 to make data normally distributed and to avoid conflicts of units (Whitlock and Schluter 2014). The outputs show the overall accuracy of the LDA classifications and their sensitivity, i.e., how accurately the group has been predicted. Only values of sensitivity > 0.95 are considered significant (see results). The R package caret (Kuhn 2012) was used for the analyses. I used GMM analyses to identify clusters in the PB, PB33, and ALL datasets. To select the morphological characters to use in the GMM analyses, I used PCA, which is commonly used in studies that examine morphological character variation in the context of species delimitation (e.g., Díaz 2013; Hong-Wa and Besnard 2014; Layton and Kellogg 2014; Pierre et al. 2014). Some of the criticism concerning the use of PCA points out a major handicap: it can handle only continuous variables, not categorical ones, and recognition of groups from the scatterplots is subjective. The package I used for the analysis, FactoMineR (Lê et al. 2008) in R version 3.2.5 (R Development Core Team 2013), was built to handle both continuous and categorical variables. The outcome of PCA analysis with FactoMineR consists of two graphs. One graph (e.g., Fig. 5) represents the groupings of the individuals in the reduced dimensional space displayed as a two-dimensional scatterplot with the principal component axes (Lê et al. 2008). The second graph (e.g., Fig. 6) shows the projection of the variables in a reduced dimensional space (a vector map). On the vector map the variables appear in the scatterplot as arrows of different lengths. The directions of the arrows show the loading of the species on the first and second principal components. This method allows display of the projection of the variables and the clustering of the observations simultaneously.

RESULTS

COMPARISON OF THE IDENTIFIED CLUSTERS BY ANALYSIS OF THE PB DATASET AND THE SPECIES RECOGNIZED BY PERRIER.

Linear Discriminant Analysis (LDA) of the PB dataset– The overall accuracy of the classification by LDA is equal to 69.53%, which indicates a weak classification. Only five species recognized by Perrier have sensitivity values over 95% (Fig. 3, asterisks), i.e., the LDA analysis correctly identified those species 95% of the time. *Psorospermum androsaemifolium* and *P. ferrovestitum* have 29 and 17 specimens respectively in Perrier de la Bâthie (1951); the three other species each have fewer than 5 specimens: *P. revolutum* has 4 specimens, while *P. nanum*, and *P. sexlineatum* have 2 specimens each (Fig. 3). However, most of the species recognized by Perrier overlap in the LDA 2-dimensional scatterplot (Fig. 4). Linear discriminant functions LD1 and LD2 discriminate mostly 3 groups, (*Psorospermum cerasifolium* and *P. malifolium*), (*P. androsaemifolium* and *bullatum*), and the rest of the species.

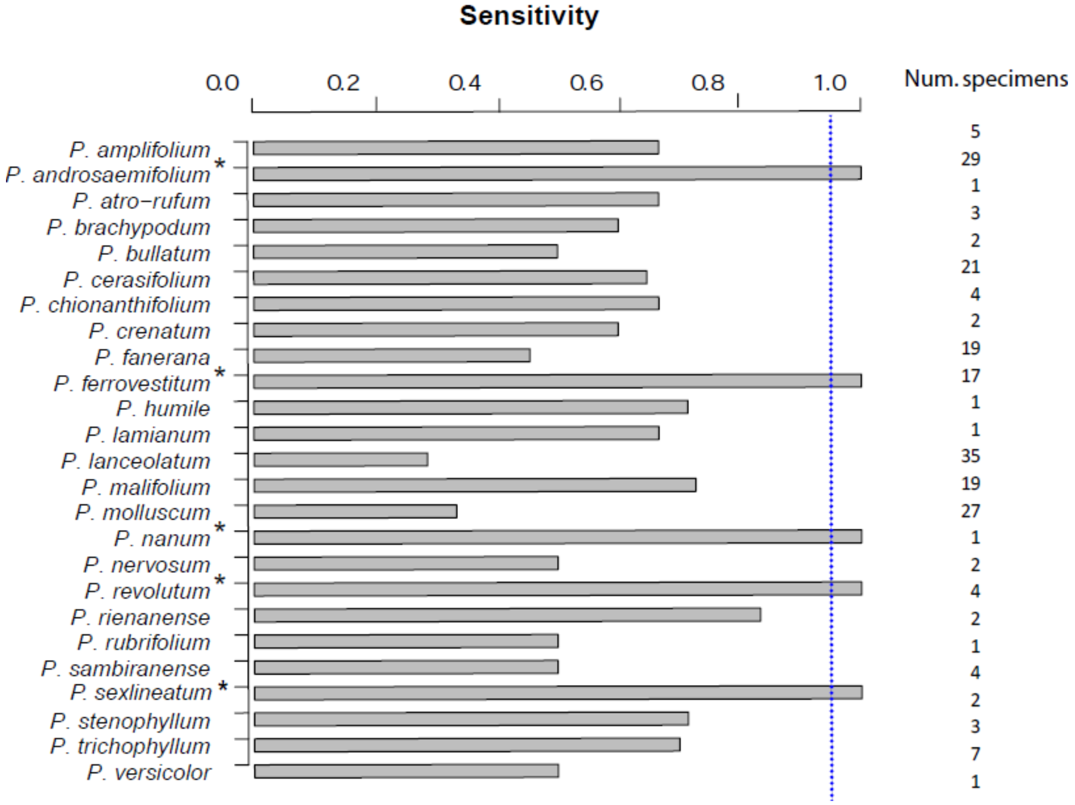


FIG 3. Histogram of the sensitivity values of Perrier's species tested with the linear discriminant cross-validation analysis using the PB dataset. Species confirmed by LDA analysis are marked with asterisks. The blue dotted line corresponds to sensitivity value 0.95. The species recognized by LDA are marked with asterisks. The numbers next to the histograms (Num. specimens) are the numbers of specimens cited by Perrier.

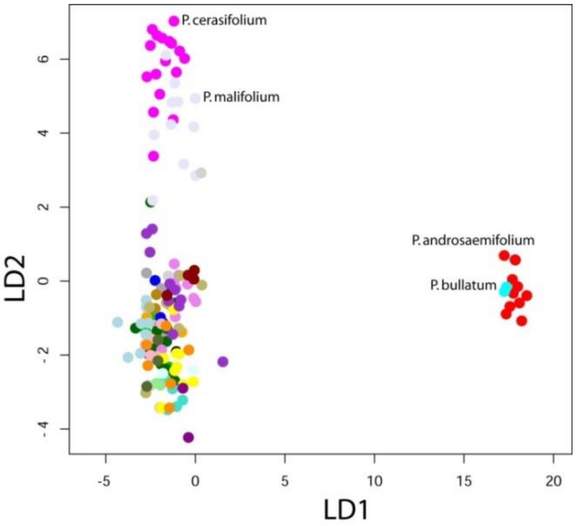


FIG 4. Scatterplot of the linear discriminant functions LD1 and LD2. Points correspond to individuals and color to Perrier's species.

Variation of characters and patterns of grouping in PB dataset– The first and second axes of the PCA account for 27.63% and 18.22% of the variation respectively (Fig. 5). I defined clusters as groups of individuals (represented by points) separated by gaps in the two-dimensional scatterplots. Although some groups are tentatively recognized, distinct clusters can hardly be recognized (Fig. 5). (Note that individual dots are all colored black so as not to prejudice the interpretation of what might be gaps.) The seven characters that contribute most to the variation in the data sets are mapped in Fig. 6: petal length, number of anthers, style length, lamina length, surface area, and width, and petiole length.

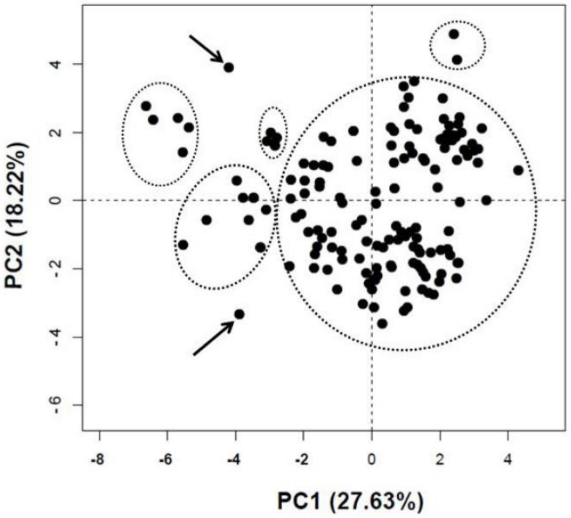


FIG 5. Two-dimensional PCA scatterplot showing putative clusters in the PB dataset. Dotted circles and arrows indicate possible delimitation of putative species.

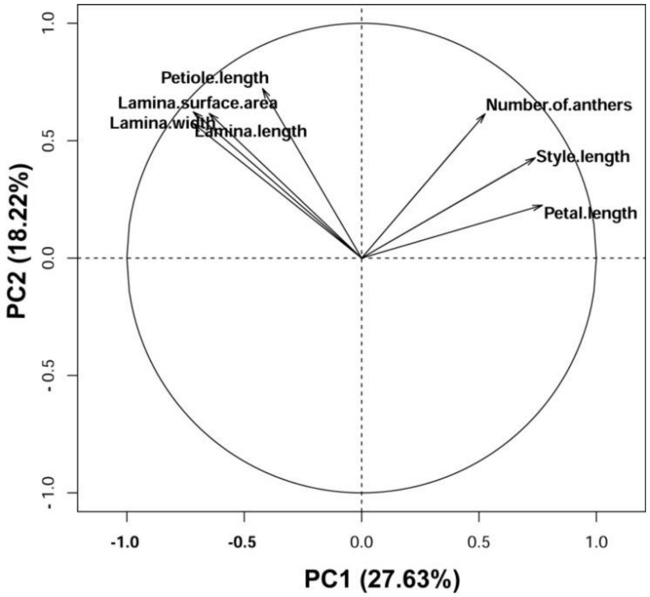


FIG 6. Vector factor map showing the seven significant variables in the PB dataset.

***Identification of putative species in PB dataset*–** The results of the Gaussian Mixture Model analysis using the seven characters from PCA yielded two putative species (Fig. 7). All pairwise scatterplots are shown in Appendix 1.

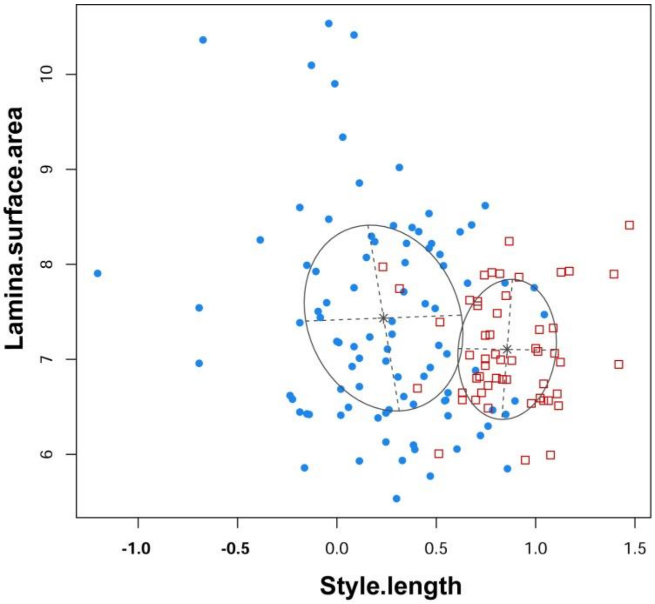


FIG 7. Scatterplot classification of the two putative species identified by GMM analyses of PB dataset. Putative species 1: filled blue circles, putative species 2: empty red squares.

COMPARISON OF IDENTIFIED CLUSTERS FROM THE ANALYSIS OF PB33 DATASET AND THE SPECIES RECOGNIZED BY PERRIER.

***Linear Discriminant Analysis of the PB33 dataset*–** In this analysis, the dataset consists of all 33 characters and only the 152 specimens observed by Perrier. The overall accuracy of the classification by LDA is equal to 81.81%. Fourteen species recognized by Perrier have sensitivity values over 95% (Fig. 8), i.e., the LDA analysis correctly identified those species 95% of the time. *Psorospermum androsaemifolium. P. cerasifolium* and *P. ferrovestitum* have 29 and 21 and 17 specimens respectively in Perrier de la Bâthie (1951); the eleven other species each have fewer than 5 specimens (Fig. 8). Most of the species recognized by Perrier overlap in the LDA scatterplot (Fig. 9). Linear discriminant functions LD1 and LD2 seems to discriminate mostly 3 groups, (*Psorospermum cerasifolium* and *androsaemifolium*), (*P. rianense*, *P. fanerana*, *P. ferrovestitum* and *P. trichophyllum*), and all other species.

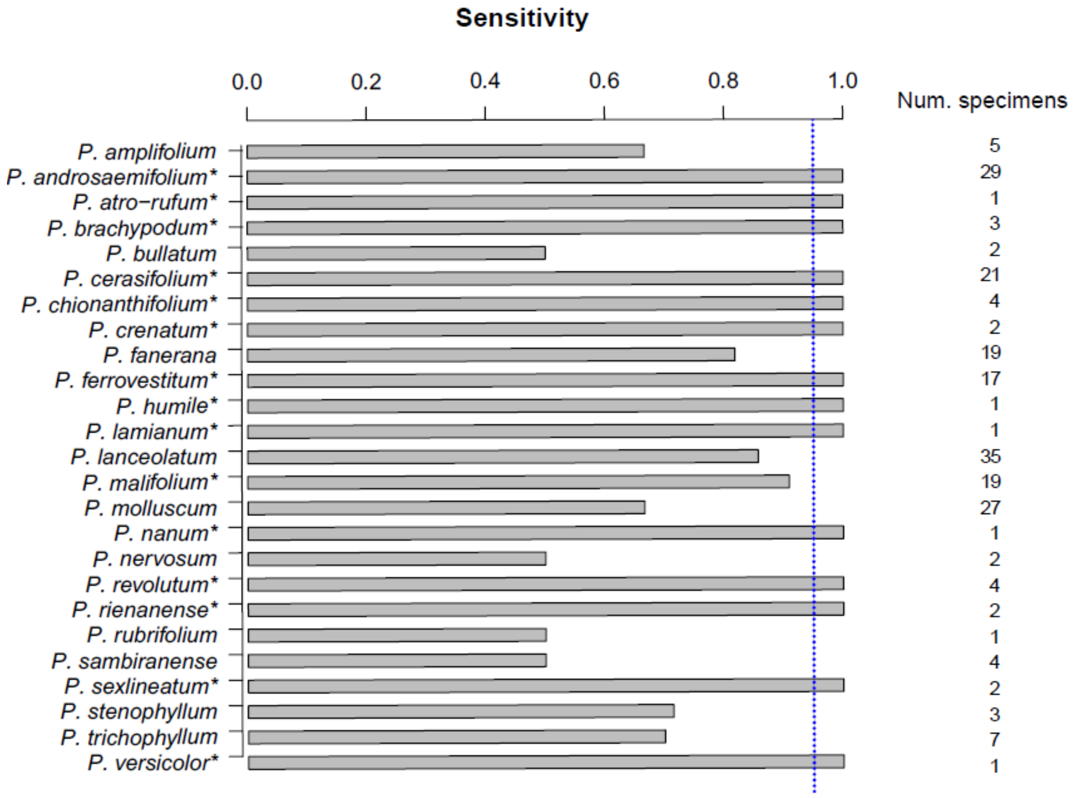


FIG 8. Histogram of the sensitivity values of Perrier's species tested with the linear discriminant cross-validation analysis using the PB33 dataset. Species confirmed by LDA analysis are marked with asterisks. The blue dotted line corresponds to sensitivity value 0.95. The species recognized by LDA are marked with asterisks. The numbers next to the histograms (Num. specimens) are the numbers of specimens cited by Perrier.

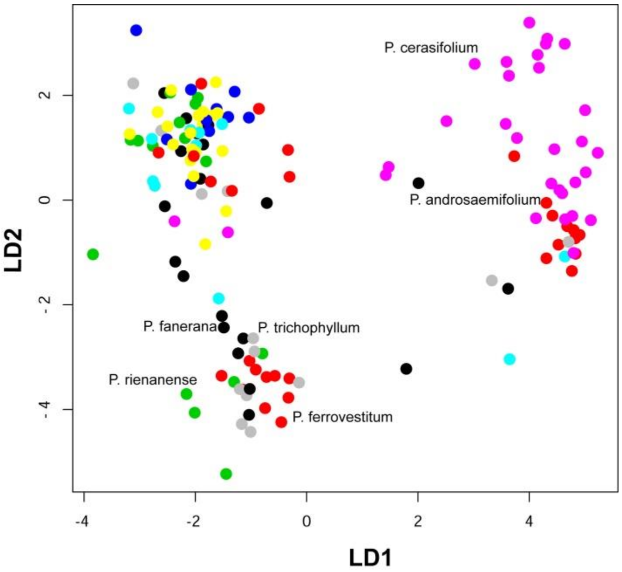


FIG 9. Scatterplot of the linear discriminant functions LD1 and LD2. Points correspond to individuals and color to Perrier's species.

Variation of characters and patterns of grouping in PB33 dataset– The first and second axes of the PCA explain 20.06% and 14.13% of the variance for the first and second axes, respectively (Fig. 10). Distinct clusters cannot be recognized (Fig. 10); perhaps there are two clusters but gaps are not obvious making clusters difficult to separate. The five characters that contribute most to the variation in the data sets are mapped in Fig. 11: number of anthers; number of anther glands; lamina width, length from the base to the widest width, and surface area.

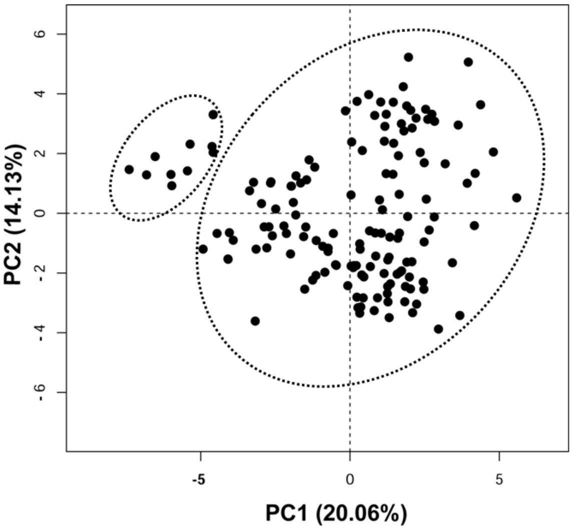


FIG 10. Two-dimensional PCA scatterplot showing putative clusters in the PB33 dataset.

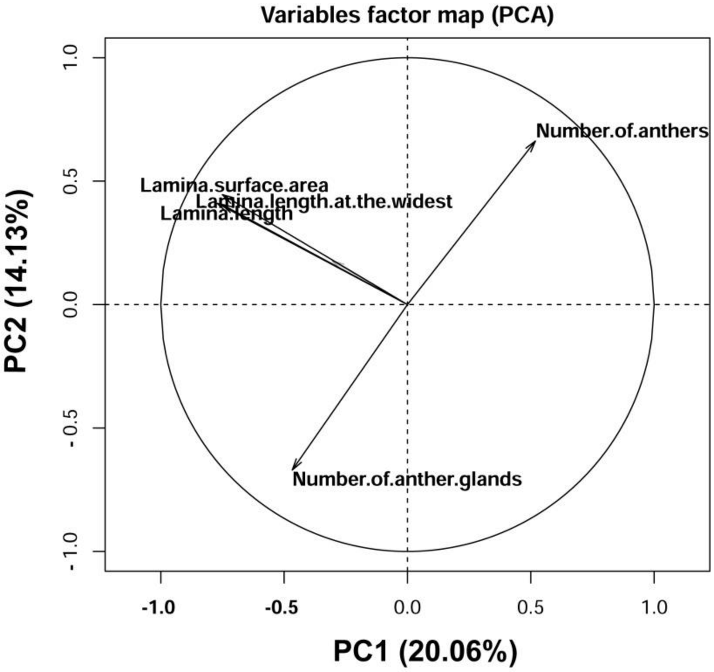


FIG 11. Vector factor maps of the seven significant variables in the PB33 dataset.

Identification of putative species in PB33 dataset– The results of the Gaussian Mixture Model analysis using the seven characters from PCA yielded six putative species (Fig. 12). All pairwise scatterplots are shown in Appendix 2.

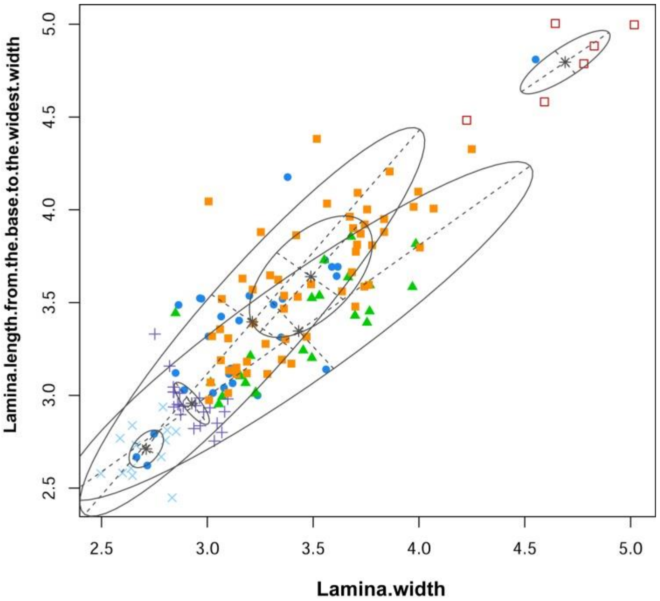


FIG 12. Scatterplot classification of the six putative species identified by GMM analyses of PB33 dataset. Putative species are in different symbols and color.

COMPARISON OF IDENTIFIED CLUSTERS FROM THE ANALYSIS OF ALL DATASETS AND THE INTEGRATIVE TAXONOMY SPECIES.

The taxonomic revision of *Psorospermum* is still in progress; the 27 species recognized from the integrative taxonomy analysis (Ranarivelo et al. in preparation) are named as ITX followed by a number ranging from 1 to 27 (e.g., ITX1,2….27). Detailed specimen lists and identification are provided in Appendix 4.

Linear Discriminant Analysis of the ALL dataset**–** The overall accuracy of the identification by LDA is equal to 85.20% and no species has a sensitivity value lower than 0.8. Nineteen species out of the 27 have sensitivity values over 0.95. LDA recognizes 13 of Perrier's named species. Four of them include all the specimens Perrier had included in the species, in six of them some specimens were included in other species. The number of specimens of each species is shown in Fig. 13. The LDA plot shows overlap of the ITX species but specimens of each species seem to cluster conspicuously (Fig. 14).

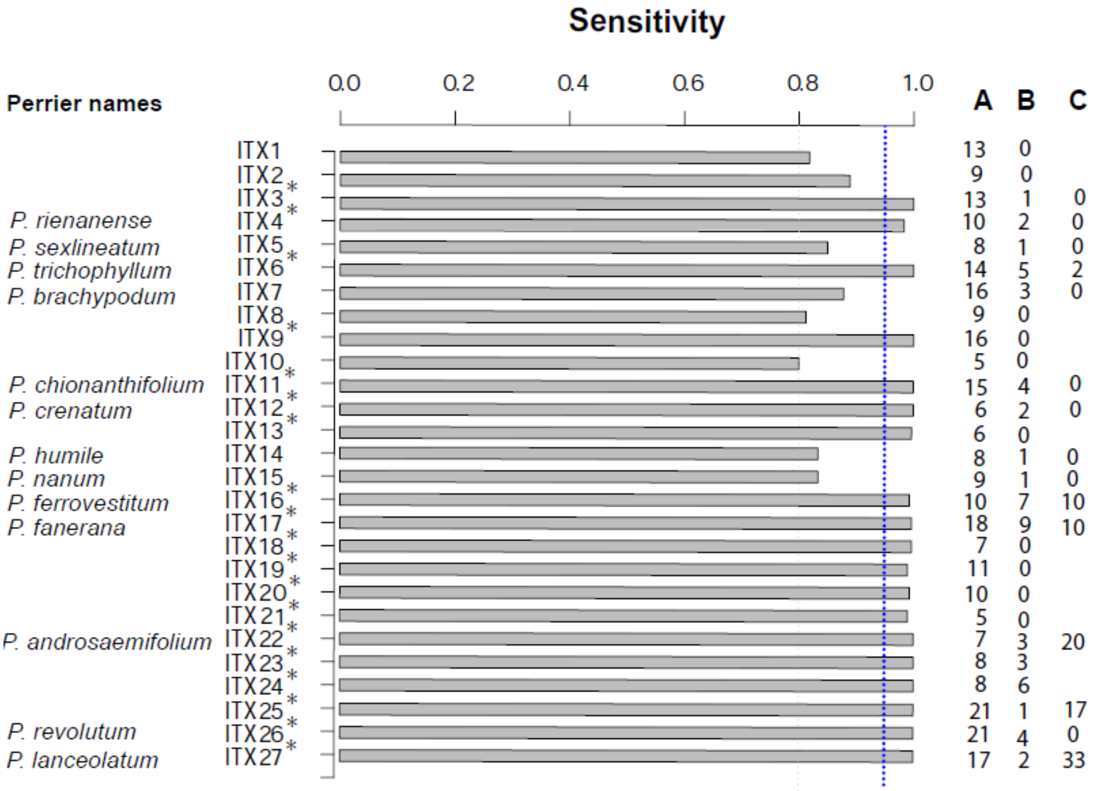


FIG 13. Histogram comparing the sensitivity of each morphospecies tested with the linear discriminant cross-validation analysis in the ALL dataset. Species confirmed by LDA analysis are marked with asterisks. The dotted line corresponds to sensitivity value 0.95. Column **A**: number of specimens of the species identified by the integrative taxonomy (ITX species); Column **B**: number of Perrier specimens in the ITX species; column **C**: number of Perrier's specimens placed elsewhere by the integrative taxonomy, so not in Perrier's species.

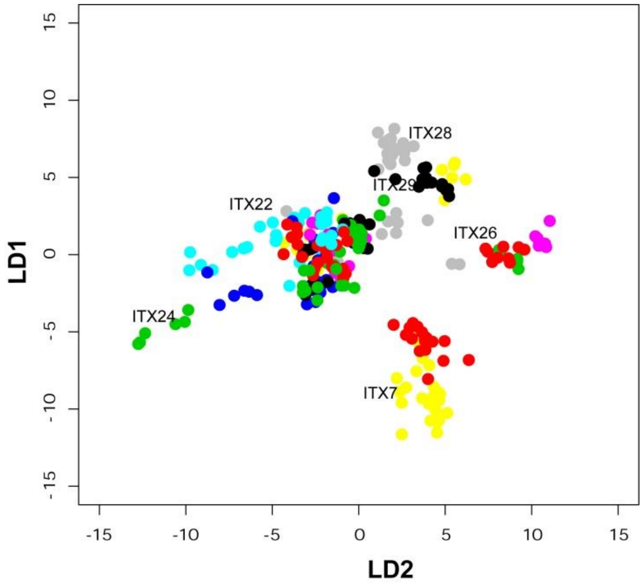


FIG 14. Scatterplot of the linear discriminant functions LD1 and LD2. Points correspond to individuals and color corresponds to the species identified by the integrative taxonomy.

Variation of characters and patterns of grouping with additional samples and additional characters (ALL dataset)– The results of the PCA showed no conspicuous gaps when additional characters and specimens were added (Fig. 15); distinct clusters cannot be recognized. The five characters that contribute most to the variation in the data set are mapped in Fig. 16, they are lamina length, surface area and number of secondary veins, cotyledon surface area and radicle length.

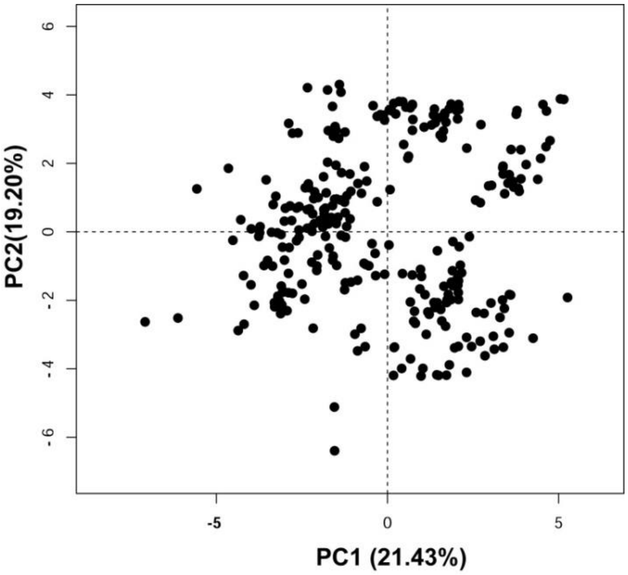


FIG 15. Two-dimensional PCA scatterplot showing putative clusters in the ALL dataset (A).

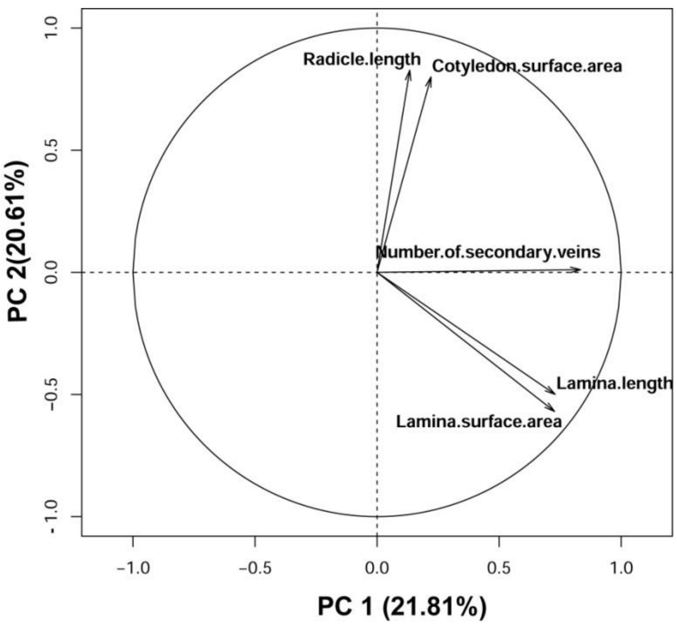


FIG 16. Vector factor maps showing the five significant variables in the ALL dataset.

Identification of clusters in ALL dataset– The Gaussian Mixture Model analysis using the six characters from the PCA mentioned above, applied to the ALL dataset, yielded 13 clusters (Fig. 17). All pairwise scatterplots are shown in Appendix 3.

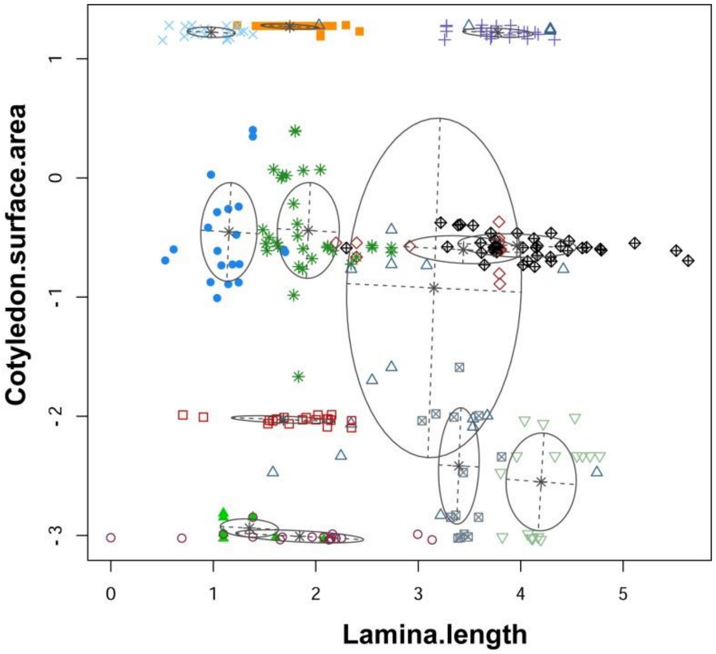


FIG 17. Scatterplot classification of the thirteen clusters identified by GMM analyses of ALL dataset. Putative species are in different symbols and colors.

DISCUSSION

What can we retrieve from old literature?– The *Flore de Madagascar et des Comores* has been cited as both outdated and in need of urgent revision (e.g., Callmander et al. 2011). The study presented here shows statistical results confirming the above statement, at least for the genus *Psorospermum*. However, I emphasize that although the PB dataset used for this study is a matrix that was built with the same characters mentioned by Perrier in his descriptions (and identification key), and the measurements of those characters were taken from the specimens he observed, those characters were not recorded in the same way that he did. To my knowledge, Perrier did not produce a data matrix.

Table 3 compares the results from the different species delimitation analyses. According to the cross-validation of LDA, there is higher confidence for the species delimited by integrative taxonomy: 19 of the 27 species are validated by LDA (Fig. 13, asterisks), and all have a sensitivity value higher than 0.8, while only 5 out of the 26 species delimited by Perrier are validated by LDA (Fig. 3, asterisks). Overall, additional samples and characters seem to improve the output of all analyses, the LDA analyses of PB33 dataset recognized more species compared to the PB dataset (Fig. 8, asterisks). However, GMM analyses of PB and ALL datasets yielded fewer species than either Perrier's study or the integrative taxonomy.

TABLE 3. Summary of the multivariate analyses using three datasets (PB, PB33 and ALL datasets).

Morphometric method	PB dataset	PB33 dataset	ALL dataset
number of species in the original studies*	26	26	27
Species recognized by LDA	5	14	19
Number of species recognized by LDA with Perrier's names	5	14	13
Species recognized by PCA	Uncertain	Uncertain	Uncertain
Species recognized by GMM	2	6	13

Asterisk * indicates that those results are not the output of PB, PB33 and ALL datasets, but of Perrier de la Bâthie (1951) and Ranarivelo et al. (in preparation).

My results here suggest that GMM might be difficult to apply on a large scale when delimiting species because the GMM analyses with the three datasets identified few morphospecies: only two when using PB dataset (vs. 26 species described by Perrier), six when using PB33 dataset and 13 when using ALL dataset (vs. 27 species delimited by Ranarivelo et al.). However, none of the 13 species apparent in the GMM analyses of the ALL dataset above correspond to species identified in Perrier's 1951 work. The GMM model behaves differently depending on the scale of the analysis. In the integrative taxonomy approach where analyses were carried out using subsets of the specimens (Ranarivelo et al. in preparation), thirteen of the 27 species recognized by the integrative taxonomic approach bear Perrier's names. However, in four of these names Perrier cited only a single specimen, at least 50% of the total specimens used by Perrier are now placed in other species (Fig. 13), and three names are in synonymy with Perrier species.

Although GMM of ALL dataset led to the recognition of a number of species, it could be questioned whether GMM or PCA adequately captured how Perrier de la Bâthie decided on the limits of the species. To the extent that PCA and GMM represent desirable approaches to taxonomic decision-making today, it is unclear what of value can be obtained by scrutinizing the limits of Perrier's species. So as a historical exercise this study failed. Indeed, although Perrier mentioned 19 characters in his key and descriptions, it is unknown whether he took other characters into account when making his decisions, or weighted some characters more heavily than others. Such uncertainty remains true of much of taxonomy today.

ACKNOWLEDGEMENTS. I thank Dr. P. F. Stevens for his advice, insight and support during this work, and Dr. P. C. Hoch, Dr. E. A. Kellogg, and Dr. B. R. Ruhfel for their constructive comments. I thank the Madagascar National Parc (MNP), the Missouri Botanical Garden in Madagascar, the NGO Fanamby, Letsara Rokiman and Jacky Andriatiana for their help during the fieldwork. Fieldwork for this research was fully funded by the Christensen Research Scholarship awarded through the International Center for Tropical Ecology at the University of Missouri – St. Louis. I also thank Dr. Peter P. Lowry II for facilitating the loans of the type specimens at P, and the following herbaria: MO, G, P, TAN, for making their collections available.

LITERATURE CITED

Callmander, M. W., P. B. Phillipson, G. E. Schatz, S. Andriambololonera, M. Rabarimanarivo, N. Rakotonirina, J. Raharimampionona, and C. Chatelain, L. Gautier and P.P. Lowry II. 2011. The endemic and non-endemic vascular flora of Madagascar updated. *Plant Ecology and Evolution* 144: 121–125.

Díaz, D. M. V. 2013. Multivariate analysis of morphological and anatomical characters of *Calophyllum* (Calophyllaceae) in South America. *Botanical Journal of the Linnean Society* 171: 587–626.

Hong-Wa, C. and G. Besnard. 2014. Species limits and diversification in the Madagascar olive (*Noronhia*, Oleaceae). *Botanical Journal of the Linnean Society* 174: 141–161.

Kuhn, M. 2012. "Caret" package (R Package Version 5.15-023). Vienna, Austria: R Foundation for Statistical Computing.

Layton, D. and E. A. Kellogg. 2014. Morphological, phylogenetic, and ecological diversity of the new model species *Setaria viridis* (Poaceae: Paniceae) and its close relatives. *American Journal of Botany* 101: 539–557.

Lê, S., J. Josse, and F. Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* 25: 1–18.

Pierre, A.H., G. Le Moguédec, P. P. Lowry, II, and J. Munzinger. 2014. Multivariate morphometric analysis and species delimitation in the endemic New Caledonian genus *Storthocalyx* (Sapindaceae). *Botanical Journal of the Linnean Society* 176: 127–146.

Perrier de la Bâthie, H. 1951. Hypericaceae 135e Famille. *Flore de Madagascar et des Comores*. Paris: Typographie Firmin-Didot et Cie.

R Core Team, 2013. R: a language and environment for statistical computing. 55: 275–286.

Ranarivelo, H. S. In preparation. Integrative taxonomy: investigating the species boundaries in Malagasy *Psorospermum* using morphometrics and molecular phylogenetic methods. To be submitted to the *Botanical Journal of the Linnean Society*.

Ranarivelo, H. S. In preparation. Systematics and Biogeography of the African-Malagasy *Psorospermum* (Hypericaceae) with emphasis on the Malagasy species. To be submitted to *Molecular Phylogenetics and Evolution*.

Rasband, W. S. 2012. ImageJ: Image processing and analysis in Java. *Astrophysics Source Code Library* 1 p.06013.

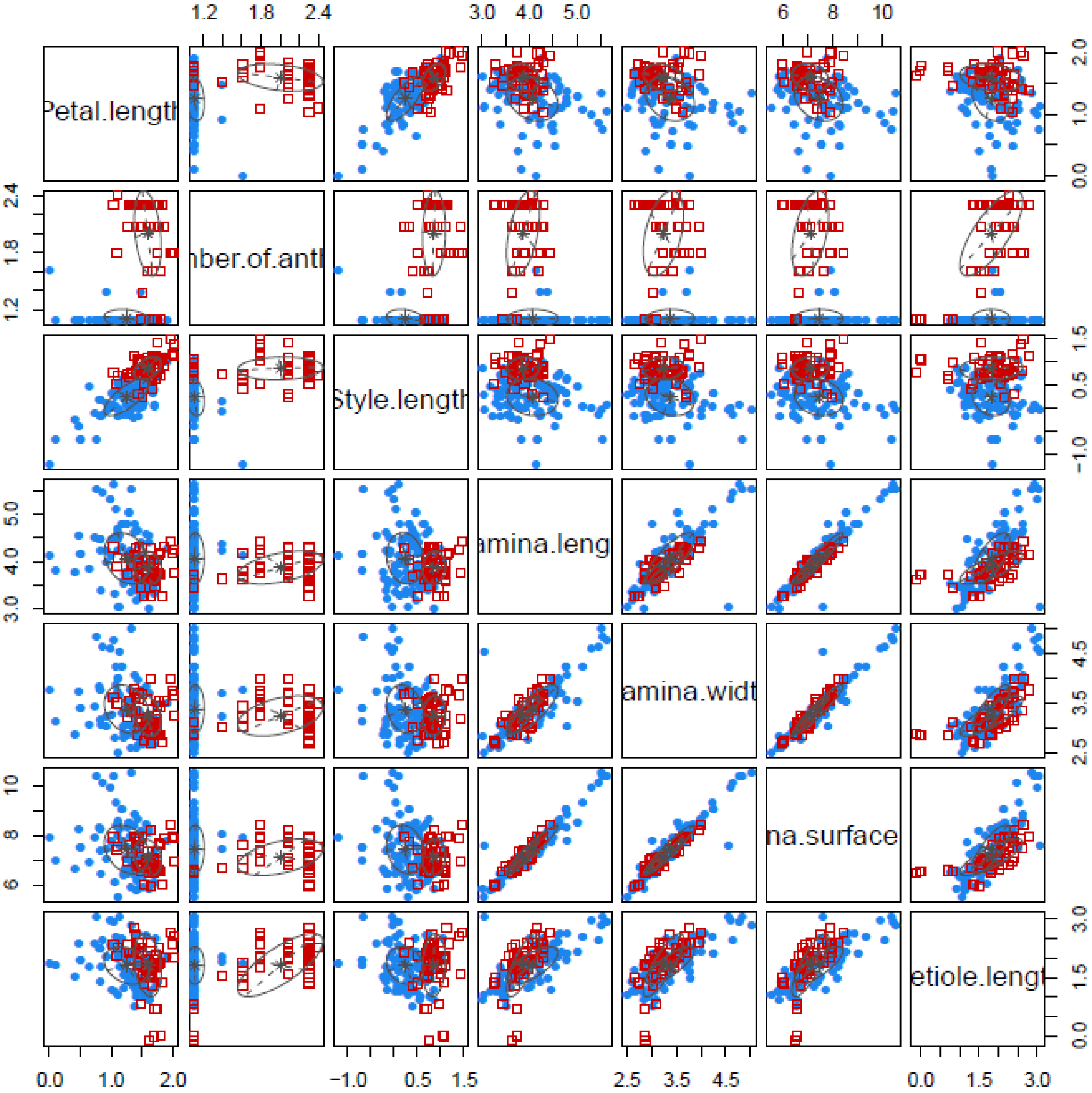
Stevens, P. F. 2007. Hypericaceae. Pp. 194–201 in *The Families and Genera of Vascular Plants. Flowering Plants. Eudicots: Berberidopsidales, Buxales, Crossosomatales, Fabales p.p., Geraniales, Gunnerales, Myrtales p.p., Proteales, Saxifragales, Vitales, Zygophyllales, Clusiaceae alliance, Passifloraceae alliance, Dilleniaceae, Huaceae, Picramniaceae, Sabiaceae*, ed. K. Kubitzki. Berlin: Springer-Verlag.

Thiers, B. [continuously updated]. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/science/ih/

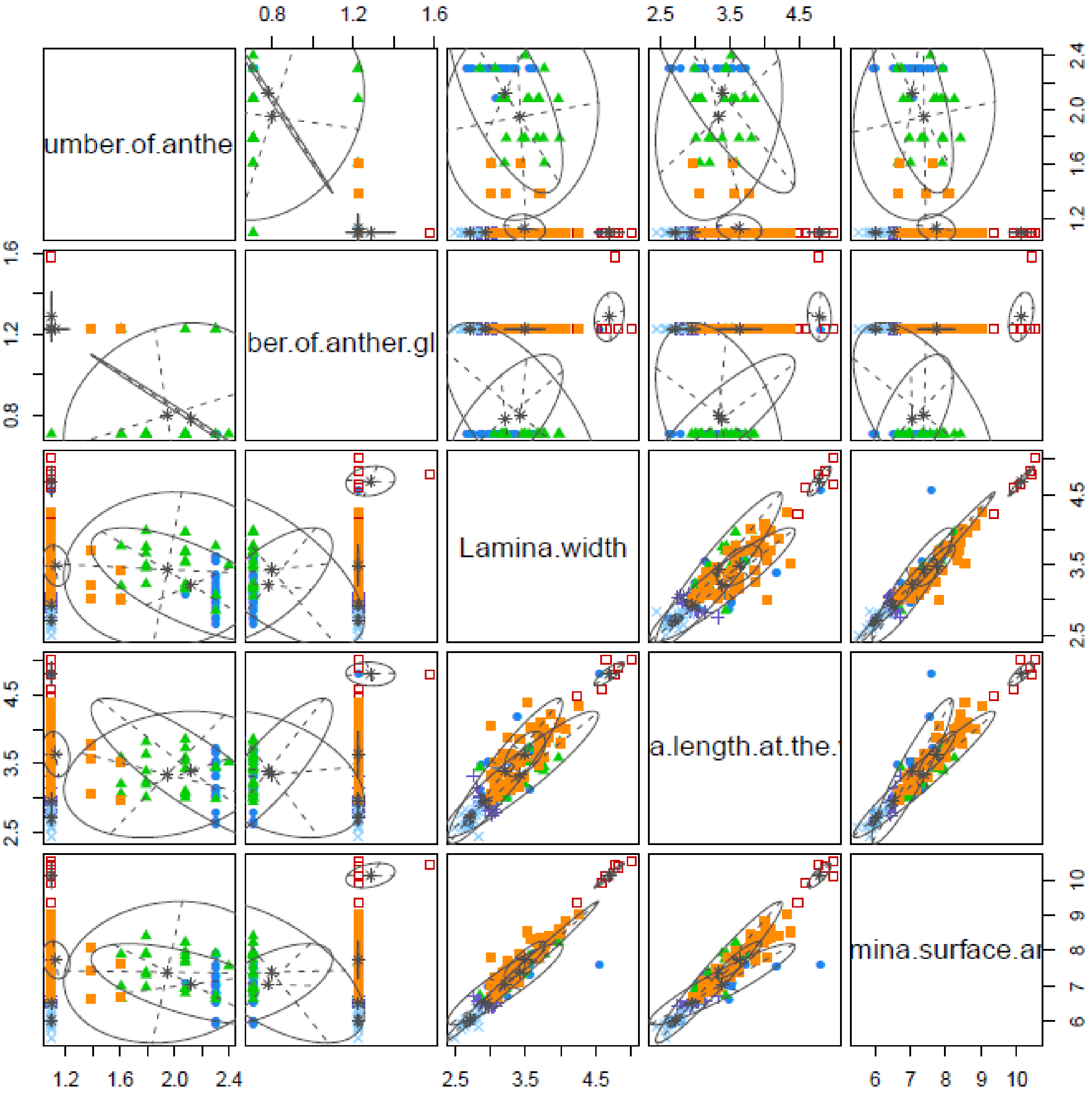
Whitlock, M. C. and D. Schluter. 2014. *The Analysis of Biological Data*. Ed. 2. New York: W. H. Freeman.

Zobayed, S. M. A., F. Afreen, E. Goto, and T. Kozai. 2006. Plant–environment interactions: accumulation of hypericin in dark glands of *Hypericum perforatum*. *Annals of Botany* 98: 793–804.

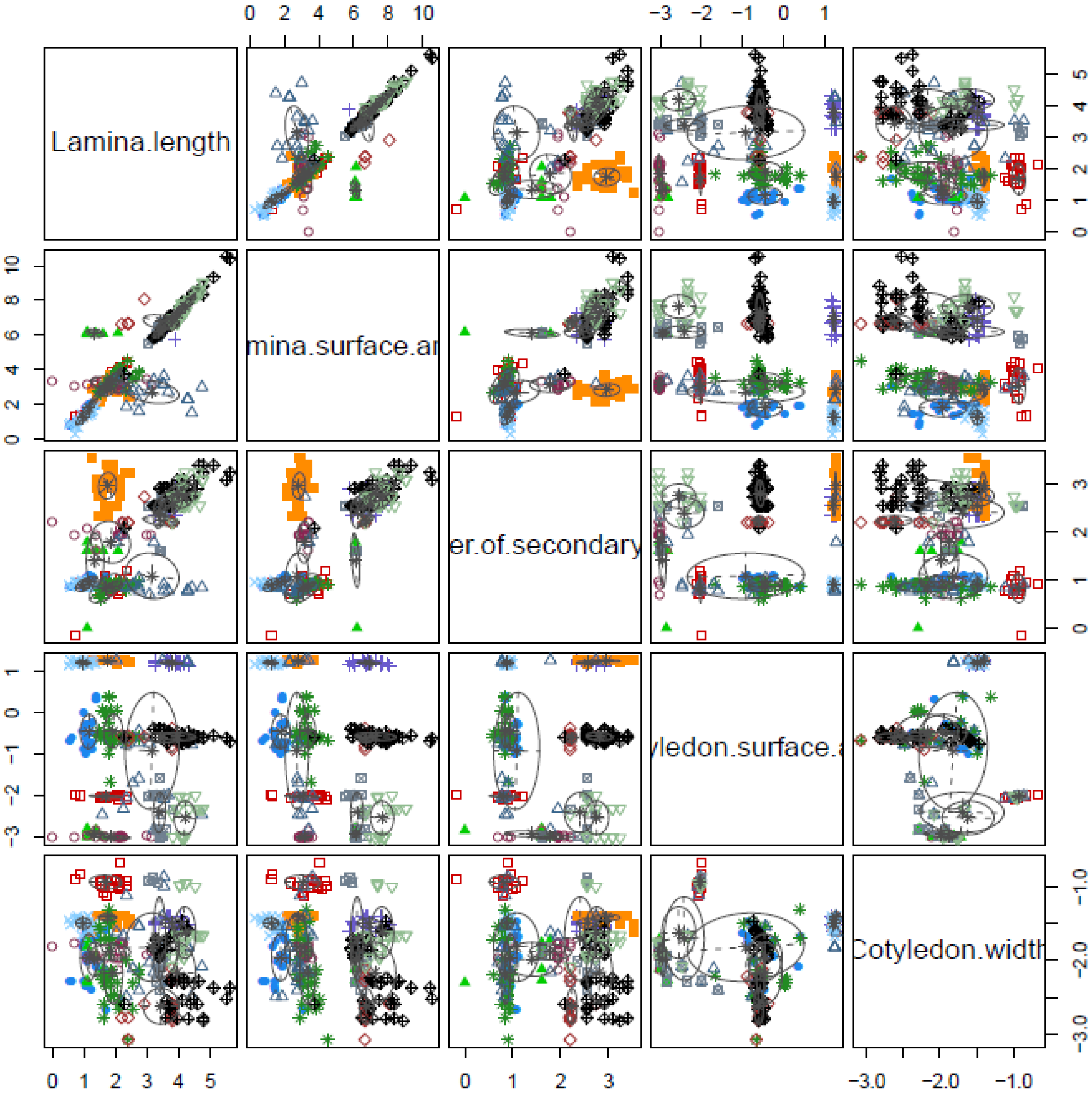
Appendix 1. Pairwise two dimensional scatterplot classification of the PB dataset. Ellipses superimposed on the plot correspond to the covariances of the clusters. Individuals in different morphospecies are indicated by different symbols and colors. Filled blue circle: morphospecies 1; empty red square: morphospecies 2.



Appendix 2. Pairwise two dimensional scatterplots classification of the PB33 dataset. Ellipses superimposed on the plot correspond to the covariances of the clusters. Individuals in different morphospecies are indicated by different symbols and colors.



Appendix 3. Pairwise two dimensional scatterplots classification of the ALL dataset. Ellipses superimposed on the plot correspond to the covariances of the clusters. Individuals in different morphospecies are indicated by different symbols and colors.



Appendix 4. Species delimited by the integrative taxonomy (Ranarivelo et al. in preparation).

Morphospecies of the GMM compared to the Phylogeny morphogroups	**Recognized as species**
P. sp11	ITX2
P. sp13	ITX3
(P. atro-rufum + P. cf atro-rufum)	ITX1
P. rienanense	ITX4
P. sexlineatum	ITX5
P. trichophyllum	ITX6
P. brachypodum	ITX7
P. cf. brachypodum (1)	ITX8
P. cf. brachypodum (2)	ITX9
P. chionanthifolium1	ITX10
P. chionanthifolium 2	ITX11
P. crenatum 1	ITX12
P. crenatum 2	ITX13
P. nanum	ITX14
P. humile	ITX15
(P. nanum + P. humile)?	
P. revolutum	ITX26
P. cf lanceolatum	ITX27
P. sp16	ITX18
P. sp17	ITX19
P. fanerana	ITX17
P. ferrovestitum	ITX16
P. malifolium	ITX24
(P. sp22 + P. sp2+ P. cerasifolium)	ITX25
P. cf. androsaemifolium (1)	ITX20
P. cf. androsaemifolium (2)	ITX21
P. androsaemifolium	ITX22
P. sp19	ITX23