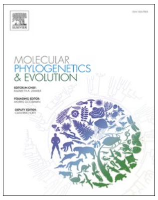
Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev



New genetic markers for Sapotaceae phylogenomics: More than 600 nuclear genes applicable from family to population levels

Camille Christe[a,b,*,1], Carlos G. Boluda[a,b,1], Darina Koubínová[a,c], Laurent Gautier[a,b], Yamama Naciri[a,b]

[a] *Conservatoire et Jardin botaniques, 1292 Chambésy, Geneva, Switzerland*
[b] *Laboratoire de botanique systématique et de biodiversité de l'Université de Genève, Department of Botany and Plant Biology, 1292 Chambésy, Geneva, Switzerland*
[c] *Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland*

ARTICLE INFO

Keywords:
Conservation
Gene capture
STR
Phylogenetics
Population genetics
Species tree
Tropical trees

ABSTRACT

Some tropical plant families, such as the Sapotaceae, have a complex taxonomy, which can be resolved using Next Generation Sequencing (NGS). For most groups however, methodological protocols are still missing. Here we identified 531 monocopy genes and 227 Short Tandem Repeats (STR) markers and tested them on Sapotaceae using target capture and NGS. The probes were designed using two genome skimming samples from *Capurodendron delphinense* and *Bemangidia lowryi*, both from the Tseboneae tribe, as well as the published *Manilkara zapota* transcriptome from the Sapotoideae tribe. We combined our probes with 261 additional ones previously published and designed for the entire angiosperm group. On a total of 792 low-copy genes, 638 showed no signs of paralogy and were used to build a phylogeny of the family with 231 individuals from all main lineages. A highly supported topology was obtained at high taxonomic ranks but also at the species level. This phylogeny revealed the existence of more than 20 putative new species. Single nucleotide polymorphisms (SNPs) extracted from the 638 genes were able to distinguish lineages within a species complex and to highlight geographical structuration. STR were recovered efficiently for the species used as reference (*C. delphinense*) but the recovery rate decreased dramatically with the phylogenetic distance to the focal species. Altogether, the new loci will help reaching a sound taxonomic understanding of the family Sapotaceae for which many circumscriptions and relationships are still debated, at the species, genus and tribe levels.

1. Introduction

The fast development of massive sequencing methods allows us now to move from a few loci using Sanger's technology to genome sequencing (Heather and Chain, 2016). However, whole genome sequencing is still expensive for taxonomical or population studies that require the use of many specimens. Additionally, genome sequencing provides many unwanted loci, as they may not adjust to the requirements of the research (e.g. non-conserved, multicopy or invariable loci), which often exceeds the computational capacity of most currently used softwares.

Gene capture therefore appears as an efficient methodology: for a cost comparable to a single genome sequencing it is possible to sequence several hundreds of pre-selected target loci for tens of specimens at once. This methodology is based on a hybridization step with specific biotinylated oligonucleotide probes complementary to the loci of interest. As biotin links to streptavidin, hybridized sequences can be retained while all non-target DNA is washed away (Moorthie et al., 2011). Whenever the specimens' DNA are previously marked with specific sequences (called barcodes), target captures of many specimens can be merged and sequenced jointly in a single sequencing lane.

Gene capture has been previously tested in plants (Nicholls et al., 2015; Stephens et al., 2015a, 2015b; Heyduk et al., 2016; Uribe-Convers et al., 2016; Sass et al., 2016; de La Harpe et al., 2019), and universal angiosperm probe kits have been recently proposed (Buddenhagen et al., 2016; Johnson et al., 2019). However, universal probes may hybridize only with genetic markers that are conserved in all Angiosperms, which may therefore exhibit low nucleotide substitution rates impeding the distinction between closely related species within a specific plant family. Then, for a microevolutionary scale study, specific probes should be

[*] Corresponding author at: Laboratoire de Phylogénie et Génetique moléculaires, Conservatoire et Jardin botaniques, 1292 Chambésy, Geneva, Switzerland.
E-mail address: camille.christe@ville-ge.ch (C. Christe).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ympev.2021.107123
Received 21 October 2020; Received in revised form 10 February 2021; Accepted 13 February 2021
Available online 19 February 2021

designed, especially when the aim is to resolve intraspecific lineages or to understand population issues.

The tree family Sapotaceae is a good indicator of forest quality in tropical areas (Gautier, 2003). Sapotaceae trees are major components of the canopy, have a slow growth rate, and form an appreciated timber which is increasingly valued at international level. As a consequence, it is often logged by local communities for regional or international trade uses. Sapotaceae displays many traits of taxonomic importance that are homoplasic, in addition to merism instability. Flowers and fruits are furthermore often absent or inaccessible in the canopy (Swenson and Anderberg, 2005; Kümpers et al., 2016). All these features make the family taxonomically difficult. Sapotaceae species are understudied in many tropical countries, particularly in Africa and more especially in Madagascar (but see Ewango et al., 2016; Gautier et al., 2013; Gautier et al., 2016; Mackinder et al., 2016; Gautier and Naciri, 2018; Rokni et al., 2019; Randriarisoa et al., 2020). Malagasy Sapotaceae were revised approximately fifty years ago, in the framework of the Flore de Madagascar et des Comores (Aubréville, 1974), when only one third of the collections now available were made. Furthermore, this revision was based essentially on herbarium samples, with most species known from a very restricted number of specimens, often lacking either flowers or fruits, and consequently with a limited understanding of species delimitations.

The two main subfamilies Sapotoideae and Chrysophylloideae are unequally represented on Madagascar, the first one with three tribes (Sideroxyleae, Tseboneae and Sapoteae), eight genera and more than 100 species, the second one with its single tribe, two genera and 11 species. Although taxonomical issues remain to be solved in all tribes, especially due to uncertain species or generic limits, the Malagasy endemic Tribe Tseboneae (Gautier et al., 2013) displays serious species delimitation problems. Recent collections indeed suggest that in the genus *Capurodendron*, more than 20 undescribed morphospecies exist, as well as many specimens falling morphologically between described species. Many Malagasy described or undescribed Sapotaceae species are critically endangered due to a continuing deforestation, and difficulties encountered in correct species identification are impeding the implementation of efficient protection measures and felling controls.

Previous attempts to reconstruct phylogenies using standard barcoding markers sometimes failed to produce supported topologies at high taxonomical ranks like the tribe level, but also at the level of closely related species (Swenson and Anderberg, 2005; Gautier et al., 2013; Armstrong et al., 2014). Due to the scarcity of herbarium material, old specimens have to be used, for which highly fragmented DNA frequently impedes a successful PCR amplification. Additionally, PCR reactions are sometimes hampered in Sapotaceae by the presence of repetitive sequences (Riet et al., 2017), and by latex and polysaccharides, which are abundant and difficult to remove (Michiels et al., 2003; McDevit and Saunders, 2009).

A combination of probes able to hybridize with markers of various substitution rates would then provide a universal set of markers for the entire Sapotaceae family. In the framework of a project on taxonomy and conservation of the tribe Tseboneae we designed specific probes that are also able to produce efficient target captures in all other tribes of the Sapotaceae family. Our strategy combined the use of two genomes (*Bemangidia lowryi* L. Gaut. and *Capurodendron delphinense* Aubrév.; Tseboneae tribe) that were sequenced for this project, with that of a published transcriptome of *Manilkara zapota* (L.) Van Royen, (Sapoteae tribe), in order to search for single or low-copy genes. This allowed us to design probes for 530 exonic markers and 227 Short tandem repeat (STR) loci. Probes for 261 additional markers suggested by Johnson et al. (2019) for the entire angiosperm group were also added to the ones we designed. Our aims were (1) to test whether we can overcome technical problems encountered with Sanger sequencing such as the use of old herbarium specimens or samples with problematic secondary metabolites (2) to test the efficiency of gene capture throughout the Sapotaceae family; (3) to test whether the analysis of the captured loci improves the phylogenetic resolution at genus and species levels; (4) to evaluate whether it is possible to get insights into the intraspecific level, more specifically within species complexes that could not be resolved using Sanger sequencing.

2. Materials and methods

2.1. Sampling

One specimen of *Capurodendron delphinense* (*Gautier 5801*; G00377582, Madagascar, 25°3′40″ S/46°52′17″ E) and one of *Bemangidia lowryi* (*Gautier 5789*; G00377560, Madagascar, 24°33′56″ S/47°11′58″ E), both from the Tseboneae tribe (Sapotaceae), were used for genome skimming in order to subsequently design probes for target capture. The transcriptome of *Manilkara zapota* from the 1KP project was used to identify putative exonic regions (Matasci et al., 2014; accession BEFC).

Probe efficiency was tested on 262 specimens belonging mostly to tribe Tseboneae, but with representatives of nearly all described tribes of the Sapotaceae family (Swenson and Anderberg, 2005, Gautier et al., 2013). All known *Capurodendron* species but four (*C. antongilense*, *C. sp. 11*, *C. nanophyllum* and *C. rufescens*) were represented by more than one accession (2–35; Table 1). Samples had two different origins: silica-gel dried leaf tissue for samples collected after 2010 (43.5%), and herbarium stored leaf fragments from older collections (56.5%, Table 1). Herbarium samples were up to 86 years-old with a mean of 29 years. All of them were most probably dried during 2–3 days at ca. 65 °C (part of them with a previous preservation in ethanol 60%) according to the main protocols used in the tropics (Forrest et al., 2019). This process is known to favor DNA fragmentation (Staats et al., 2011).

2.2. Genome skimming

Whole genome sequencing at low coverage of *Bemangidia lowryi* and *Capurodendron delphinense* was performed from fresh leaf tissue stored in silica-gel. DNA was extracted using the CTAB-chloroform protocol (see Supplementary Material 1 for more details). One Illumina TruSeq DNA Nano paired-end library was prepared for each species, with mean insert sizes of 474 bp and 598 bp, respectively, and sequenced on a single Illumina HiSeq 4000 lane (2x100 bp paired-end). Both steps were processed at iGE3, the Institute of Genetics and Genomics of Geneva (Switzerland). Adapter sequences and low-quality reads were trimmed with Trimmomatic version 0.36 (Bolger et al., 2014) using a minimum quality threshold of 20 in a 5-bp window. Several statistics such as genome size and k-mer coverage were estimated with the preqc tool from the SGA assembler (Simpson and Durbin, 2012; Simpson 2014)

2.3. Selection of target loci

Loci were selected in order to focus on three evolutionary scales: the family, tribe and species levels, targeting loci with low to supposedly high mutation rates. Target loci, consisting in orthologous low-copy nuclear protein-coding genes, were selected using the Sondovač pipeline version 1.3 (Schmickl et al., 2016). The list of retrieved loci was then shortened based on further criteria such as number of variants and the level of heterozygosity in the sequence. We also selected a set of sequences based on the information of intron–exon boundaries, in order to have enough phylogenetic information at the Tsebonae level avoiding paralogy issues and getting intronic sequences as well, as they might bear additional phylogenetic information.

We decided to combine the two paired-end libraries from *Bemangidia lowri* and *Capurodendron delphinense* as starting material to allow covering more regions and genes from the Tsebonae tribe target than when using each species separately. This combined reference was chosen to design the probes for Sondovač. The *Pouteria campechiana* (Kunth) Baehni plastome (Sangjin et al., 2016) and *Vaccinium macrocarpon* Aiton

Table 1
Specimen information of the samples used for testing the designed probes. The percentage of missing data/indels is given in an all specimens alignment containing 638 concatenated genes. LG: Gautier; RAM: Randriarisoa; RIR: Randrianaivo; RN: Réserves Naturelles; SF: Service Forestiers.

Code	Tribe	Species (alphabetically ordered)	Collector code	Year	Origin	Missing data/indels	BioSample n°
104b	Tseboneae	*Bemangidia lowri*	SF, s.n., Barcode [P00568786]	2011	Herbarium	5–40%	SAMN17142034
121	Tseboneae	*B. sp.*	LG 5790	2011	Silica gel	<5%	SAMN17141888
122	Tseboneae	*B. sp.*	Razakamalala 3976	2007	Herbarium	<5%	SAMN17141889
128	Tseboneae	*Capurodendron androyense*	LG 6328	2017	Silica gel	<5%	SAMN17141894
70	Tseboneae	*C. androyense*	Rakotomalaza 1719	1998	Herbarium	<5%	SAMN17141843
79	Tseboneae	*C. androyense*	Rogers 474	2004	Herbarium	<5%	SAMN17141850
125	Tseboneae	*C. androyense*	LG 6372	2017	Silica gel	<5%	SAMN17141891
126	Tseboneae	*C. androyense*	LG 6376	2017	Silica gel	<5%	SAMN17141892
127	Tseboneae	*C. androyense*	LG 6387	2017	Silica gel	<5%	SAMN17141893
138	Tseboneae	*C. androyense*	LG 6361	2017	Silica gel	<5%	SAMN17141903
139	Tseboneae	*C. androyense*	LG 6346	2017	Silica gel	<5%	SAMN17141904
140	Tseboneae	*C. androyense*	LG 6358	2017	Silica gel	<5%	SAMN17141905
141	Tseboneae	*C. androyense*	LG 6343	2017	Silica gel	<5%	SAMN17141906
143	Tseboneae	*C. androyense*	LG 6370	2017	Silica gel	5–40%	SAMN17141907
144	Tseboneae	*C. androyense*	LG 6371	2017	Silica gel	<5%	SAMN17141908
145	Tseboneae	*C. androyense*	LG 6374	2017	Silica gel	<5%	SAMN17141909
149	Tseboneae	*C. androyense*	RIR 2954	2017	Silica gel	<5%	SAMN17141913
150	Tseboneae	*C. androyense*	SF 22,230	1962	Herbarium	>80%	SAMN17141914
75	Tseboneae	*C. ankaranense* (Type)	Humbert 25,489	1951	Herbarium	<5%	SAMN17141846
1	Tseboneae	*C. ankaranense*	RAM 40	2017	Silica gel	<5%	SAMN17141776
119	Tseboneae	*C. ankaranense*	RN 6118	1954	Herbarium	5–40%	SAMN17141886
147	Tseboneae	*C. ankaranense*	LG 6241	2016	Silica gel	<5%	SAMN17141911
177	Tseboneae	*C. ankaranense*	SF 18,545	1958	Herbarium	<5%	SAMN17141939
92	Tseboneae	*C. antongiliense* (Type)	SF 8961	1954	Herbarium	>80%	SAMN17141862
91	Tseboneae	*C. apollonioides*	SF 9014	1954	Herbarium	>80%	SAMN17141861
117	Tseboneae	*C. apollonioides*	SF 21,804	1964	Herbarium	<5%	SAMN17141884
178	Tseboneae	*C. apollonioides*	Raharimalala 2262	1990	Herbarium	>80%	SAMN17141940
179	Tseboneae	*C. apollonioides*	SF 8672	1953	Herbarium	<5%	SAMN17141941
106	Tseboneae	*C. bakeri s.l.*	LG 5553	2010	Herbarium	>80%	SAMN17141874
107	Tseboneae	*C. bakeri s.l.*	LG 5780	2011	Silica gel	<5%	SAMN17141875
205	Tseboneae	*C. bakeri s.l.*	SF 28,059	1967	Herbarium	<5%	SAMN17141962
206	Tseboneae	*C. bakeri s.l.*	SF 8936	1954	Herbarium	5–40%	SAMN17141963
2	Tseboneae	*C. bakeri var. antalahaense*	RN 7056	1955	Herbarium	>80%	SAMN17141777
111	Tseboneae	*C. bakeri var. antalahaense*	RIR 1844	2011	Herbarium	5–40%	SAMN17141879
3	Tseboneae	*C. bakeri var. bakeri*	Razakamalala 2491	2005	Herbarium	5–40%	SAMN17141778
4	Tseboneae	*C. bakeri var. bakeri*	LG 6390	2017	Silica gel	<5%	SAMN17141779
5	Tseboneae	*C. bakeri var. bakeri*	LG 6392	2017	Silica gel	5–40%	SAMN17141780
72	Tseboneae	*C. costatum*	Leandri 2038	1952	Herbarium	<5%	SAMN17141844
99	Tseboneae	*C. costatum*	LG 5864	2012	Silica gel	<5%	SAMN17141868
115	Tseboneae	*C. costatum*	SF 6789	1952	Herbarium	5–40%	SAMN17141883
6	Tseboneae	*C. delphinense*	Ramison 471	2007	Herbarium	<5%	SAMN17141781
7	Tseboneae	*C. delphinense*	Randriatafika 722	2006	Herbarium	<5%	SAMN17141782
98	Tseboneae	*C. delphinense*	LG 5801	2011	Silica gel	<5%	SAMN17141867
8	Tseboneae	*C. gracilifolium*	LG 6318	2017	Silica gel	5–40%	SAMN17141783
151	Tseboneae	*C. gracilifolium*	LG 5736	2011	Silica gel	<5%	SAMN17141915
156	Tseboneae	*C. gracilifolium*	Messmer 607	1998	Herbarium	<5%	SAMN17141920
181	Tseboneae	*C. gracilifolium*	SF 9438	1953	Herbarium	>80%	SAMN17141942
182	Tseboneae	*C. gracilifolium*	RIR 2972	2017	Silica gel	<5%	SAMN17141943
9	Tseboneae	*C. greveanum*	RAM 28	2017	Silica gel	<5%	SAMN17141784
10	Tseboneae	*C. greveanum*	RIR 2974	2017	Silica gel	<5%	SAMN17141785
11	Tseboneae	*C. greveanum*	Ranaivojaona 267	2000	Herbarium	<5%	SAMN17141786
74 g	Tseboneae	*C. greveanum*	Jongkind 3623	1997	Herbarium	<5%	SAMN17142036
86	Tseboneae	*C. ludiifolium*	SF 28,097	1967	Herbarium	>80%	SAMN17141857
184	Tseboneae	*C. ludiifolium*	SF,s.n., P04609609	–	Herbarium	<5%	SAMN17141945
216	Tseboneae	*C. ludiifolium*	RIR 3063	2018	Silica gel	<5%	SAMN17141972
217	Tseboneae	*C. ludiifolium*	RIR 3126	2018	Silica gel	<5%	SAMN17141973
218	Tseboneae	*C. ludiifolium*	RIR 3162	2018	Silica gel	<5%	SAMN17141974
219	Tseboneae	*C. ludiifolium*	RIR 3014	2018	Silica gel	<5%	SAMN17141975
243	Tseboneae	*C. ludiifolium*	RIR 3166	2018	Silica gel	5–40%	SAMN17141995
87	Tseboneae	*C. madagascariense*	SF 27,524	1967	Herbarium	>80%	SAMN17141858
89	Tseboneae	*C. madagascariense*	SF 18,033	1957	Herbarium	<5%	SAMN17141859
94	Tseboneae	*C. madagascariense*	SF 5407	1952	Herbarium	<5%	SAMN17141863
185	Tseboneae	*C. madagascariense*	SF 16,962	1956	Herbarium	<5%	SAMN17141946
12	Tseboneae	*C. mandrarense*	Razafindraibe 165	2006	Herbarium	<5%	SAMN17141787
13	Tseboneae	*C. mandrarense*	Ratovoson 1473	2008	Herbarium	<5%	SAMN17141788
15	Tseboneae	*C. mandrarense*	LG 6332	2017	Silica gel	<5%	SAMN17141790
16	Tseboneae	*C. mandrarense*	LG 6336	2017	Silica gel	<5%	SAMN17141791
17	Tseboneae	*C. mandrarense*	LG 6337	2017	Silica gel	<5%	SAMN17141792
18	Tseboneae	*C. mandrarense*	LG 6339	2017	Silica gel	<5%	SAMN17141793
19	Tseboneae	*C. mandrarense*	LG 6341	2017	Silica gel	<5%	SAMN17141794
20	Tseboneae	*C. mandrarense*	LG 6349	2017	Silica gel	<5%	SAMN17141795
21	Tseboneae	*C. mandrarense*	LG 6350	2017	Silica gel	<5%	SAMN17141796
22	Tseboneae	*C. mandrarense*	LG 6351	2017	Silica gel	<5%	SAMN17141797

(*continued on next page*)

Table 1 (*continued*)

Code	Tribe	Species (alphabetically ordered)	Collector code	Year	Origin	Missing data/indels	BioSample n°
23	Tseboneae	*C. mandrarense*	LG 6356	2017	Silica gel	<5%	SAMN17141798
24	Tseboneae	*C. mandrarense*	LG 6366	2017	Silica gel	5–40%	SAMN17141799
25	Tseboneae	*C. mandrarense*	LG 6378	2017	Silica gel	<5%	SAMN17141800
26	Tseboneae	*C. mandrarense*	LG 6379	2017	Silica gel	5–40%	SAMN17141801
27	Tseboneae	*C. mandrarense*	RIR 2956	2017	Silica gel	<5%	SAMN17141802
29	Tseboneae	*C. mandrarense*	RIR 2959	2017	Silica gel	5–40%	SAMN17141803
30	Tseboneae	*C. mandrarense*	RIR 2960	2017	Silica gel	<5%	SAMN17141804
31	Tseboneae	*C. mandrarense*	RIR 2961	2017	Silica gel	5–40%	SAMN17141805
32	Tseboneae	*C. mandrarense*	RIR 2962	2017	Silica gel	5–40%	SAMN17141806
33	Tseboneae	*C. mandrarense*	RIR 2964	2017	Silica gel	<5%	SAMN17141807
34	Tseboneae	*C. mandrarense*	RIR 2966	2017	Silica gel	<5%	SAMN17141808
35	Tseboneae	*C. mandrarense*	RIR 2967	2017	Silica gel	<5%	SAMN17141809
37	Tseboneae	*C. mandrarense*	RIR 2970	2017	Silica gel	<5%	SAMN17141811
38	Tseboneae	*C. mandrarense*	RIR 2980	2017	Silica gel	<5%	SAMN17141812
39	Tseboneae	*C. mandrarense*	RIR 2981	2017	Silica gel	<5%	SAMN17141813
77	Tseboneae	*C. mandrarense*	Phillipson 5603	2002	Herbarium	<5%	SAMN17141848
110	Tseboneae	*C. mandrarense*	RIR 1785	2011	Herbarium	<5%	SAMN17141878
113	Tseboneae	*C. mandrarense*	RIR 1187	2005	Herbarium	<5%	SAMN17141881
158	Tseboneae	*C. mandrarense*	Randrianasolo 204	1991	Herbarium	<5%	SAMN17141922
159	Tseboneae	*C. mandrarense*	RIR 1764	2009	Herbarium	<5%	SAMN17141923
160	Tseboneae	*C. mandrarense*	McPherson 17,358	1998	Herbarium	<5%	SAMN17141924
161	Tseboneae	*C. mandrarense*	SF 22,286	1962	Herbarium	5–40%	SAMN17141925
162	Tseboneae	*C. mandrarense*	SF 6692	1952	Herbarium	40–80%	SAMN17141926
163	Tseboneae	*C. mandrarense*	Andriamihajar 1532	2004	Herbarium	<5%	SAMN17141927
183	Tseboneae	*C. mandrarense*	Andrianjafy 1679	2006	Herbarium	<5%	SAMN17141944
40	Tseboneae	*C. microphyllum*	LG 6382	2017	Silica gel	<5%	SAMN17141814
41	Tseboneae	*C. microphyllum*	LG 6393	2017	Silica gel	<5%	SAMN17141815
120	Tseboneae	*C. microphyllum*	LG 5794	2011	Silica gel	5–40%	SAMN17141887
186	Tseboneae	*C. microphyllum*	SF 22,411	1963	Herbarium	<5%	SAMN17141947
81	Tseboneae	*C. nanophyllum* (Type)	SF28521	1968	Herbarium	<<5%	SAMN17141852
42	Tseboneae	*C. nodosum*	Ranirison 454	2004	Herbarium	5–40%	SAMN17141816
43	Tseboneae	*C. nodosum*	RAM 6	2017	Silica gel	<5%	SAMN17141817
176	Tseboneae	*C. nodosum*	RAM 26	2017	Silica gel	<5%	SAMN17141938
36	Tseboneae	*C. perrieri*	RIR 2968	2017	Silica gel	<5%	SAMN17141810
45	Tseboneae	*C. perrieri*	RIR 2976	2017	Silica gel	<5%	SAMN17141819
46	Tseboneae	*C. perrieri*	Noyes 1044	1992	Herbarium	<5%	SAMN17141820
47	Tseboneae	*C. perrieri*	Razakamalala 5177	2010	Herbarium	<5%	SAMN17141821
114	Tseboneae	*C. perrieri*	RIR 969	2003	Herbarium	<5%	SAMN17141882
44	Tseboneae	*C. perrieri var. oblongifolium*	Rakotonasolo 1601	2015	Herbarium	<5%	SAMN17141818
48	Tseboneae	*C. perrieri var. oblongifolium*	Ramananjanahary 51	2004	Herbarium	<5%	SAMN17141822
49	Tseboneae	*C. perrieri var. oblongifolium*	Razakamalala 1809	2004	Herbarium	<5%	SAMN17141823
190	Tseboneae	*C. perrieri var. oblongifolium*	PerrierBâthie 1105	1974	Herbarium	5–40%	SAMN17141951
50	Tseboneae	*C. pervillei*	RIR 2397	2013	Herbarium	<5%	SAMN17141824
76	Tseboneae	*C. pervillei*	Labat 3557	2005	Herbarium	<5%	SAMN17141847
164	Tseboneae	*C. pervillei*	Ramananjanahary 244	2004	Herbarium	<5%	SAMN17141928
165	Tseboneae	*C. pervillei*	Razakamalala 1677	2004	Herbarium	<5%	SAMN17141929
191	Tseboneae	*C. pervillei*	Randrianarivelo 307	2005	Herbarium	<5%	SAMN17141952
192	Tseboneae	*C. pervillei*	RIR 953	2003	Herbarium	<5%	SAMN17141953
193	Tseboneae	*C. pseudoterminalia* (Type)	RN 9157	1957	Herbarium	40–80%	SAMN17141954
73	Tseboneae	*C. rubrocostatum*	Luino 21	2012	Herbarium	<5%	SAMN17141845
100	Tseboneae	*C. rubrocostatum*	LG 5936	2012	Silica gel	<5%	SAMN17141869
194	Tseboneae	*C. rubrocostatum*	Chauvet 187	1961	Herbarium	40–80%	SAMN17141955
195	Tseboneae	*C. rubrocostatum*	Andriamihajarivo 782	2005	Herbarium	<5%	SAMN17141956
197	Tseboneae	*C. rufescens*	SF 9186	1954	Herbarium	>80%	SAMN17141957
51	Tseboneae	*C. cf. rufescens*	Randrianjanaka 41	1993	Herbarium	5–40%	SAMN17141825
249	Tseboneae	*C. cf. rufescens*	RIR 3010	2018	Silica gel	>80%	SAMN17142001
248	Tseboneae	*C. cf. rufescens*	RIR 3002	2018	Silica gel	>80%	SAMN17142000
58	Tseboneae	*C. sahafariense* (Type)	Ratovoson 1217	2007	Herbarium	<5%	SAMN17141831
57	Tseboneae	*C. sahafariense*	Rakotonandra. 1207	2007	Herbarium	<5%	SAMN17141830
84	Tseboneae	*C. sahafariense*	SF 23,087	1963	Herbarium	<5%	SAMN17141855
52	Tseboneae	*C. sakalavum*	LG 4670	2004	Herbarium	<5%	SAMN17141826
95	Tseboneae	*C. sakalavum*	LG 5570	2011	Herbarium	>80%	SAMN17141864
97	Tseboneae	*C. sakalavum*	LG 5825	2012	Herbarium	<5%	SAMN17141866
148	Tseboneae	*C. sakalavum*	LG 6179	2015	Herbarium	<5%	SAMN17141912
78	Tseboneae	*C. schatzii*	Schatz 3786	1999	Herbarium	>80%	SAMN17141849
112	Tseboneae	*C. schatzii*	RIR 123	1997	Herbarium	>80%	SAMN17141880
215	Tseboneae	*C. aff. schatzii*	RIR 3064	2018	Silica gel	<5%	SAMN17141971
61	Tseboneae	*C. suarezense*	Razafimandimbison 274	1998	Herbarium	<5%	SAMN17141834
62	Tseboneae	*C. suarezense*	Randrianasolo 632	2007	Herbarium	<5%	SAMN17141835
63	Tseboneae	*C. suarezense*	Andrianantoa. 1043	1997	Herbarium	<5%	SAMN17141836
154	Tseboneae	*C. suarezense*	RAM 46	2017	Silica gel	<5%	SAMN17141918
155	Tseboneae	*C. suarezense*	RAM 36	2017	Silica gel	<5%	SAMN17141919
66	Tseboneae	*C. tampinense s.l.*	Ramandimbimana 260	2012	Herbarium	<5%	SAMN17141839
109	Tseboneae	*C. tampinense s.l.*	RIR 1848	2011	Herbarium	<5%	SAMN17141877
167	Tseboneae	*C. tampinense s.l.*	Razakamalala 4269	2009	Herbarium	5–40%	SAMN17141931
170	Tseboneae	*C. tampinense s.l.*	Ludovic 993	2004	Herbarium	40–80%	SAMN17141933

(*continued on next page*)

Table 1 (*continued*)

Code	Tribe	Species (alphabetically ordered)	Collector code	Year	Origin	Missing data/indels	BioSample n°
171	Tseboneae	*C. tampinense s.l.*	Rabenantoandro 1148	2002	Herbarium	>80%	SAMN17141934
172	Tseboneae	*C. tampinense s.l.*	Rabevohitra 4431	2003	Herbarium	5–40%	SAMN17141935
199	Tseboneae	*C. tampinense s.l.*	Ludovic 1829	2012	Herbarium	<5%	SAMN17141958
210	Tseboneae	*C. tampinense s.l.*	RIR 3106	2018	Silica gel	<5%	SAMN17141966
223	Tseboneae	*C. tampinense s.l.*	RAM 113	2018	Silica gel	5–40%	SAMN17141978
224	Tseboneae	*C. tampinense s.l.*	RIR 3108	2018	Silica gel	<5%	SAMN17141979
225	Tseboneae	*C. tampinense s.l.*	RIR 3150	2018	Herbarium	<5%	SAMN17141980
226	Tseboneae	*C. tampinense s.l.*	RIR 3084	2018	Silica gel	<5%	SAMN17141981
227	Tseboneae	*C. tampinense s.l.*	RIR 3089	2018	Silica gel	<5%	SAMN17141982
229	Tseboneae	*C. tampinense s.l.*	RAM 146	2018	Silica gel	<5%	SAMN17141983
246	Tseboneae	*C. tampinense s.l.*	RIR 3128	2018	Silica gel	5–40%	SAMN17141998
250	Tseboneae	*C. tampinense s.l.*	RIR 3095	2018	Silica gel	<5%	SAMN17142002
74 t	Tseboneae	*C. tampinense s.l.*	Ludovic 719	2000	Herbarium	<5%	SAMN17142037
90	Tseboneae	*C. tampinense var. analamaza* (Type)	SF 11,522	1954	Herbarium	40–80%	SAMN17141860
64	Tseboneae	*C. tampinense var. analamazaotrense*	Rakotondrafara 955	2014	Herbarium	>80%	SAMN17141837
65	Tseboneae	*C. tampinense var. analamazaotrense*	Antilahimena 2323	2003	Herbarium	>80%	SAMN17141838
209	Tseboneae	*C. tampinense var. analamazaotrense*	RAM 175	2018	Silica gel	<5%	SAMN17141965
211	Tseboneae	*C. tampinense var. analamazaotrense*	RAM 161	2018	Silica gel	<5%	SAMN17141967
212	Tseboneae	*C. tampinense var. analamazaotrense*	RAM 156	2018	Silica gel	<5%	SAMN17141968
82	Tseboneae	*C. terminalioides*	SF 28,499	1968	Herbarium	>80%	SAMN17141853
208	Tseboneae	*C. terminalioides*	Humbert 13,185	1933	Herbarium	40–80%	SAMN17141964
174	Tseboneae	*C. cf. terminalioides*	SF 42-R-224	1954	Herbarium	<5%	SAMN17141936
80	Tseboneae	*C. sp. 1*	Vasey 103	1994	Herbarium	>80%	SAMN17141851
104c	Tseboneae	*C. sp. 1*	LG 5520	2010	Herbarium	<5%	SAMN17142035
59	Tseboneae	*C. sp. 3*	Ramandimbimanana 388	2012	Herbarium	5–40%	SAMN17141832
83	Tseboneae	*C. sp. 4*	SF 27,345	1966	Herbarium	<5%	SAMN17141854
102	Tseboneae	*C. sp. 4*	LG 6036	2013	Herbarium	<5%	SAMN17141871
169	Tseboneae	*C. sp. 4*	Rabehevitra 940	2004	Herbarium	<5%	SAMN17141932
187	Tseboneae	*C. sp. 5*	Ranirison 1089	2006	Herbarium	<5%	SAMN17141948
188	Tseboneae	*C. sp. 5*	Ranirison 1095	2006	Herbarium	<5%	SAMN17141949
85	Tseboneae	*C. sp. 6*	SF 18,129	1957	Herbarium	40–80%	SAMN17141856
201	Tseboneae	*C. sp. 6*	SF 742-R182	1954	Herbarium	>80%	SAMN17141960
213	Tseboneae	*C. sp. 6*	RIR 3175	2018	Silica gel	<5%	SAMN17141969
152	Tseboneae	*C. sp. 7*	SF 5-R-82	1953	Herbarium	>80%	SAMN17141916
204	Tseboneae	*C. sp. 7*	Rabesandratana 4190	1994	Herbarium	>80%	SAMN17141961
60	Tseboneae	*C. sp. 9*	Antilahimena 343	1996	Herbarium	<5%	SAMN17141833
105	Tseboneae	*C. sp. 11*	LG 5544	2010	Silica gel	<5%	SAMN17141873
108	Tseboneae	*C. sp. 11*	LG 6024	2013	Silica gel	<5%	SAMN17141876
230	Tseboneae	*C. sp. 11*	RAM 125	2018	Silica gel	<5%	SAMN17141984
231	Tseboneae	*C. sp. 11*	RIR 3012	2018	Silica gel	<5%	SAMN17141985
232	Tseboneae	*C. sp. 11*	RIR 3049	2018	Silica gel	<5%	SAMN17141986
233	Tseboneae	*C. sp. 11*	RIR 3122	2018	Silica gel	<5%	SAMN17141987
235	Tseboneae	*C. sp. 11*	RIR 3020	2018	Silica gel	<5%	SAMN17141988
56	Tseboneae	*C. sp. 12*	Birkinshaw 438	1997	Herbarium	<5%	SAMN17141829
54	Tseboneae	*C. sp. 15*	Razakamalala 2609	2005	Herbarium	<5%	SAMN17141828
53	Tseboneae	*C. sp. 16*	Ranirison 1029	2005	Herbarium	<5%	SAMN17141827
157	Tseboneae	*C. sp. 17*	Guittou 184	2005	Herbarium	<5%	SAMN17141921
153	Tseboneae	*C. sp. 18*	RAM 75	2017	Silica gel	5–40%	SAMN17141917
200	Tseboneae	*C. sp. 19*	Ratovoson 43	1999	Herbarium	<5%	SAMN17141959
146	Tseboneae	*C. sp. 20*	LG 6276	2016	Silica gel	<5%	SAMN17141910
14	Tseboneae	*C. sp. 21*	LG 6329	2017	Silica gel	<5%	SAMN17141789
103	Tseboneae	*C. sp. 22*	LG 5395	2010	Silica gel	<5%	SAMN17141872
67	Tseboneae	*C. sp. 23*	RAM 25	2017	Silica gel	<5%	SAMN17141840
68	Tseboneae	*C. sp. 23*	RAM 50	2017	Silica gel	<5%	SAMN17141841
166	Tseboneae	*C. sp. 23*	Andrianjafy 428	2004	Herbarium	<5%	SAMN17141930
189	Tseboneae	*C. sp. 23*	Randrianaivo 1359	2006	Herbarium	<5%	SAMN17141950
96	Tseboneae	*C. sp.*	LG 5762	2011	Silica gel	40–80%	SAMN17141865
118	Tseboneae	*C. sp.*	SF 5878	1952	Herbarium	40–80%	SAMN17141885
175	Tseboneae	*C. sp.*	Guittou 207	2005	Herbarium	<5%	SAMN17141937
221	Tseboneae	*C. sp.*	RIR 3098	2018	Silica gel	<5%	SAMN17141977
241	Tseboneae	*C. sp.*	RIR 3160	2018	Silica gel	>80%	SAMN17141993
242	Tseboneae	*C. sp.*	RIR 3066	2018	Silica gel	>80%	SAMN17141994
131	Isonandreae	*Diploknema butyracea*	J.F.Dobremez 2591	1974	Herbarium	5–40%	SAMN17141897
236	Chrysophylleae	*Donella fenerivensis*	RIR 3081	2018	Silica gel	<5%	SAMN17141989
237	Chrysophylleae	*D. sp.*	RIR 3091	2018	Silica gel	<5%	SAMN17141990
239	Sapoteae	*Faucherea littoralis nom. ined.*	RAM 140	2018	Silica gel	<5%	SAMN17141991
251	Sapoteae	*F. littoralis nom. ined.*	RIR 3097	2018	Silica gel	<5%	SAMN17142003
244	Sapoteae	*F. sp.*	RIR 3068	2018	Silica gel	<5%	SAMN17141996
280	Sapoteae	*F. sp.*	RIR 3023	2018	Silica gel	40–80%	SAMN17142031
247	Sapoteae	*Faucherea. sp.*	RIR 3065	2018	Silica gel	40–80%	SAMN17141999
134	Glueminea	*Gluema ivorensis*	D. Ouvattara	2012	Herbarium	<5%	SAMN17141900
270	Glueminae	*G. ivorensis*	Jongkind 12,344	2014	Herbarium	<5%	SAMN17142021
275	Glueminae	*G. ivorensis*	MBGtransec t487	2007	Herbarium	<5%	SAMN17142026
255	Glueminae	*G. korupensis* (Paratype)	vdBurgt 758A	2005	Herbarium	40–80%	SAMN17142006
257	Glueminae	*G. korupensis* (Type)	vdBurgt 732	2005	Herbarium	<5%	SAMN17142008
272	Glueminea	*Inhambanella guereensis*	Aké Assi 10,149	1968	Herbarium	<5%	SAMN17142023

(*continued on next page*)

Table 1 (*continued*)

Code	Tribe	Species (alphabetically ordered)	Collector code	Year	Origin	Missing data/indels	BioSample n°
254	Glueminae	*I. guereensis*	H.Téré,sn	2013	Herbarium	<5%	SAMN17142005
273	Glueminae	*I. henriquezii*	Goldsmith 176/62	1962	Herbarium	<5%	SAMN17142024
274	Glueminae	*I. henriquezii*	Goldsmith 178/62	1962	Herbarium	<5%	SAMN17142025
135	Isonandreae	*Isonandra compta*	Kostermans 27,571	1979	Herbarium	<5%	SAMN17141901
259	Gluemeae	*Lecomtodoxa biraudii*	Dauby 2149	2010	Herbarium	<5%	SAMN17142010
133	Gluemeae	*L. heitzana*	Dauby 2566	2012	Herbarium	>80%	SAMN17141899
258	Gluemeae	*L. heitzana*	MBGtransect 662	2007	Herbarium	5–40%	SAMN17142009
260	Gluemeae	*L. heitzana*	Issembe 12	1998	Herbarium	>80%	SAMN17142011
261	Gluemeae	*L. heitzana*	Valkenburg 2518	2003	Herbarium	>80%	SAMN17142012
265	Gluemeae	*L. klaineana*	Louis 1839	1985	Herbarium	<5%	SAMN17142016
266	Gluemeae	*L. klaineana*	vdBurgt 727	2005	Herbarium	<5%	SAMN17142017
267	Gluemeae	*L. klaineana*	McPherson 16,835	1997	Herbarium	<5%	SAMN17142018
276	Gluemeae	*L. klaineana*	Parmentier-Mambo 4803	2008	Herbarium	<5%	SAMN17142027
268	Gluemeae	*L. plumosa*	vdBurgt 1132	2008	Herbarium	<5%	SAMN17142019
269	Gluemeae	*L. plumosa*	vdBurgt 935	2007	Herbarium	<5%	SAMN17142020
256	Gluemeae	*L. plumosa* (Type)	vdBurgt 771	2005	Herbarium	<5%	SAMN17142007
262	Gluemeae	*L. saint-aubinii*	MBGtransect 241	2005	Herbarium	<5%	SAMN17142013
277	Gluemeae	*L. saint-aubinii*	Dauby 1944	2009	Herbarium	<5%	SAMN17142028
132	Isonandreae	*Madhuca insignis*	Sooryaprakash HSS 3502	2003	Herbarium	>80%	SAMN17141898
101	Sapoteae	*Mimusops capuronii*	LG 6027	2013	Silica gel	<5%	SAMN17141870
69	Sapoteae	*M. capuronii*	RAM 69	2017	Silica gel	5–40%	SAMN17141842
240	Sapoteae	*M. cf. capuronii*	RIR 3055	2018	Silica gel	>80%	SAMN17141992
245	Sapoteae	*M. sp.*	RIR 3007	2018	Silica gel	5–40%	SAMN17141997
281	Sapoteae	*M. sp.*	RIR 3071	2018	Silica gel	<5%	SAMN17142032
282	Sapoteae	*M. sp.*	RIR 3073	2018	Silica gel	5–40%	SAMN17142033
220	Sapoteae	*cf. M. sp.*	RIR 3178	2018	Silica gel	<5%	SAMN17141976
279	Gluemeae	*Neolemonniera batesii*	LisowskiM-580	1997	Herbarium	<5%	SAMN17142030
278	Gluemeae	*N. batesii*	Peguy 3001	2000	Herbarium	<5%	SAMN17142029
264	Gluemeae	*N. clitandrifolia*	Jongkind 1777	1994	Herbarium	<5%	SAMN17142015
271	Gluemeae	*N. clitandrifolia*	AkéAssi 7913	1965	Herbarium	5–40%	SAMN17142022
263	Gluemeae	*N. clitandrifolia*	vdBurgt 1447	2010	Herbarium	<5%	SAMN17142014
253	Gluemeae	*N. ogouensis*	Thomas 7689	1988	Herbarium	>80%	SAMN17142004
130	Sideroxyleae	*Sideroxylon gerrardianum*	RAM 149	2018	Silica gel	<5%	SAMN17141896
129	Sideroxyleae	*S. gerrardianum*	RIR 3087	2018	Silica gel	<5%	SAMN17141895
124	Tseboneae	*Tsebona macrantha*	LG 5509	2010	Silica gel	<5%	SAMN17141890
137	Tseboneae	*T. macrantha*	RIR 3149	2018	Silica gel	<5%	SAMN17141902
214	Tseboneae	*T. macrantha*	RIR 3131	2018	Silica gel	<5%	SAMN17141970

mitochondrion references (Fajardo et al., 2014) were matched to the combination of paired-end genome skim reads to remove most of the reads of organelle origin as they would amplify far too much compared to nuclear loci. Paired-end reads were then merged into longer sequences when possible with the FLASH program (Magoč and Salzberg, 2011) of the Sondovač pipeline. We increased the maximum overlapping size to 120 bp following the program warnings and considering that our DNA samples had many short fragments due to degradation. The *Manilkara zapota* transcriptome (accession BEFC of Matasci et al., 2014;) was chosen from the three Sapotaceae transcriptomes available from the 1KP initiative (the two others being *Sideroxylon reclinatum* Michx., OXYP and *Synsepalum dulcificum* (Schumach. and Thonn.) Daniell, WRPP) because it was, among the former three, the closest genome to the Tsebonae tribe (Swenson and Anderberg, 2005; Gautier et al., 2013). *Manilkara zapota* transcripts sharing >90% sequence similarity were removed in order to keep unique transcripts only. Following this step, transcripts with more than 85% sequence similarity with the combined set of genomic reads were kept. At this step, we use SPAdes version 3.9 instead of Geneious for *de novo* assembly of genome skim sequences in order to design the probes following Uribe Convers (https://uribeconvers.wordpress.com). We obtained a list of contigs that comprised exons with baits longer than 120 bp and a total locus length higher than 720 bp, cleaned from sequences sharing >90% sequence similarity with other sequences within the list of contigs or with the two organelle genomes.

After getting this first list of contigs with Sondovač, we created a BLAST (Altschul et al., 1990) database for each contig and run a query against the *Manilkara* transcriptome in order to check whether some of the contigs could be found in similar *Manilkara* transcripts. Then the two species reads were mapped separately on the corresponding *Manilkara* transcript with Bowtie2 version 2.3 (Langmead and Salzberg, 2012) and the bam files indexed with Picard tools version 1.119 (Broad Institute, 2019). We searched manually the resulting bam files with IGV version 2.0 (Robinson et al., 2017) for non-overlapping reads that might be interpreted as the presence of introns. It provided information about the intron–exon boundaries in absence of a good genome assembly, and therefore helped avoiding, for a set of probes, that the designed baits would lay between two exons. This maximized the chance to get information in the intronic part of the genes. These were taken preferentially as they represent longer transcripts. The length of the contig, the number of variants existing for both Tsebonae species and for *Capurodendron* only, and the number of sites being heterozygote was calculated from the vcf file produced with UnifiedGenotyper from GATK version 3.8 (DePristo et al., 2011) using the EMIT_ALL_SITE output mode. The final decision as to whether including a contig in the bait set relied on a ratio of heterozygote/SNP below 0.3 and SNP/length above 0.03. We discarded as well all sequences with unusual high coverage. Finally, we extracted the corresponding sequence in *C. delphinense* for bait design.

Probes for 261 of the 353 single-copy genes orthologous across all angiosperms (Johnson et al., 2019), were added to our dataset. For each gene, the closest sequence to *Manilkara* transcriptome was retained using BLAST, selecting only blast hits with percentage identity higher than 70%. If no sequence was found in *Manilkara*, the two other Sapotaceae transcriptomes (*Sideroxylon reclinatum* and *Synsepalum dulcificum*) were examined and the longest sequence was kept whenever a match was found.

STR loci were additionally selected using the method of Kistler et al. (2017), and probes for the STR flanking regions were designed based on the reads of *C. delphinense* only. From all the blocks found using the script BaitSTR, only blocks with a number of TC repetitions higher than 10 were selected.

Finally, all sets of probes were checked for redundancy using BLAST, both within and between datasets. Probes were subsequently produced by Arbor Bioscience (Ann Arbor, MI, U.S.A.).

2.4. Library preparation

DNA was extracted using the CTAB method with chloroform, including sorbitol washes to remove mucilaginous substances (Russell et al., 2010; Souza et al., 2012) with some modifications (Supplementary Material 1). From one to six DNA extractions per specimen were performed, to ensure a minimum amount of 250 ng per specimen. DNA extraction was quantified using a Qubit® Fluorimeter version 3.0 (Invitrogen, Thermo Fisher Scientific, Waltham, MA, U.S.A.).

DNA fragment sizes of all samples were analyzed with a 2200 TapeStation (Agilent, Santa Clara, CA, U.S.A.). We targeted 500 bp fragments on average, and DNA samples with higher sizes were fragmented with a Bioruptor® sonicator machine (Diagenode, Seraing, Belgium), using cycles of ultrasounds lasting 30 s on/off. As the optimal number of cycles depends on the original DNA size, it was determined testing 10 specimens at different DNA fragmentation stage, using 2, 4, 6, 10, 12, 14 and 16 cycles, and visualizing the fragmentation intensity on electrophoresis gels. From 0 to 14 cycles were then applied to our samples depending of the sample original size distribution. Fragmentation appeared to be unpredictable for samples with gelatinous substances, as those substances decrease the fragmentation efficiency (see Bioruptor® manual). In that case the desired size was ensured testing the fragment sizes after each sonication cycle.

From 250 to 1000 ng of DNA per sample were used for library preparation, targeting preferentially 1000 ng whenever possible (55% of the samples). Library construction was performed with dual-indexed primers (Kircher et al., 2012) with the KAPA HyperPrep Kit (Roche, Basel, Switzerland), additionally using Bst Polymerase for Large Fragments (New England Biolabs, Ipswich, MA, U.S.A.), the P5-P7 adapter mix and NGS P5 and NGS P7 indexed primers (Microsynth, Balgach, Switzerland), following the KAPA HyperPrep Kit protocol adapted from VanBuren et al. (2018) with some small modifications (Supplementary Material 2). The washing steps were done with Sera-Mag™ Speed Beads Carboxylate-Modified Magnetic Particles (GE Healthcare, Little Chalfont, Buckinghamshire, U.K.) in a PEG/NaCl buffer, following the protocol of modified by Faircloth and Glenn (2011). As some of the DNA extractions contained gelatinous substances that impeded the beads migration towards the magnet, the washing was improved by increasing the migration time and magnet power. For very viscous samples, the magnetic wash was replaced by tube centrifugation during 1–2 min. at ~9000 rpm. For highly fragmented DNA samples, washing PEG ratios were increased until a maximum of 2.4X, to retain fragments as small as 75 bp. Libraries were quantified with Qubit® Fluorimeter version 3.0 (Invitrogen, Thermo Fisher Scientific, Waltham, MA, U.S.A.) using the high sensitivity reaction buffer. For specimens with low DNA quantities, library preparation was repeated up to three times trying to reach a total of 50 ng of DNA.

2.5. Target capture and sequencing

Dual-indexed libraries were pooled in equimolar ratios in 8 tubes, using 50 ng of DNA library per specimen when possible, with a minimum of 15 ng for specimens with low DNA concentrations. The number of specimens per tube ranged from 10 to 51. Pooling was done according to library concentrations, fragment sizes and expected probes specificity, for example pooling in separate tubes outgroups and ingroups. Target capture was performed following myBait® Custom Target Capture Kits protocol (Arbor Biosciences, Ann Arbor, MI, U.S.A.). Incubation time and temperature were adapted for each of the 8 tubes, ranging from 16 h at 65 °C for the best samples, to 44 h at 62 °C for old fragmented outgroups. After the capture, two independent PCRs were performed for each tube, with 10 cycles each in order to minimize PCR duplicates. PCR products were quantified with the Qubit® Fluorimeter version 3.0 and DNA size distributions measured using a 2200 TapeStation machine. Reactions with signals of PCR duplicates (abnormal or asymmetric curve) were discarded when possible. Tubes containing old herbarium specimens showed the presence of primer-dimers, visible as a peak at ~140 bp. These dimers could not be removed using washing steps as dimers and target sequences displayed overlapping sizes. Because dimers attach well to the Illumina plate as they are good competitors in the sequencing process, old samples were pooled and sequenced with a density of 26–40 specimens per lane, against the 71–102 used for the fresh samples.

The tubes containing captured loci were pooled equimolarly in four tubes according to DNA sizes and the presence/absence of primer-dimers (ranging from 26 to 102 specimens per tube), and sequenced in four separate lanes on a HiSeq4000 Illumina machine (100 bp pair-end reads).

2.6. Resulting sequences and mapping

Illumina reads were demultiplexed at the sequencing facility and read quality was checked with FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were trimmed with Trimmomatic version 0.38 (Bolger et al., 2014), removing reads with a quality threshold lower than 20 in a 5-bp window. Then HybPiper version 1.3.1 (Johnson et al., 2016) was used to map the reads against the probes. Consensus “exons” sequences of successfully mapped loci were obtained and aligned using MAFFT version 7 (Katoh and Standley 2013). The bam files produced by HybPiper were indexed with Picard tools version 1.119 (Picard Toolkit, 2019), and UnifiedGenotyper from GATK version 3.8 was used for SNP calling for each individual using the -EMIT_ALL_SITES output mode and turning of the downsampling option as (-dt).

The produced VCF files were split by gene with vcftools version 0.1.16, then compressed with bgzip and indexed with tabix in order to output the consensus IUPAC sequences with bcftools setting -M missing option to “-“ in order to avoid replacing missing data by the reference (baits reference for each gene). These sequences were not used in the present paper.

For STR markers, lobSTR with the option index files (Gymrek et al., 2012) were generated for the trimmed reads mapped to the *Capurodendron delphinense* STR reference blocks with BWA-mem. Samtools version 1.8 and Picard tools version 1.119 were subsequently used for indexing them. Then the module allelotype from lobSTR package was used to summarize read supports for STR genotypes. The used options were modified according to Kistler et al. (2017) (–realign –filter-clipped –min-read-end-match 10 –filter-mapq0 –max- repeats-in-ends 3) and Illumina version 2.0.3 was used as noise model. The generated VCF files were used as input for calling genotype with SONiCS (Kedzierska et al., 2018) using a minimum coverage of 25 to estimate STR genotype. This program was run in order to avoid including incorrect alleles due to polymerase slippage.

2.7. Phylogenetic and Principal component analyses

The consensus sequences of all genes showing no paralogy signal according to Hybpiper pipeline (638 genes) were selected for the phylogeny (Supplementary Table 1). Different datasets were created with 12 thresholds of missing data/indels (1%, 2%, 3%, 5%, 10%, 20%, 30%, 40%, 50%, 80%, 90%, and 97%) removing specimens that exceeded the targeted value (Supplementary Table 3). As a first approach, a SNP phylogenetic tree was reconstructed based on each dataset. In a second step, the datasets containing only specimens with a maximum of 5%, 40%, and 80% of missing data/indels were selected for Astral tree reconstruction. They were selected as to represent different optima ratios between sequence length and specimens’ number. Bioedit version 7.2 (Hall, 1999) and trimAl version 1.4 (Capella-Gutiérrez et al., 2009) were used to remove specimens with more than a given level of missing data/indel positions, AMAS version 1.0 (Borowiec, 2016) was used to concatenate the loci, and SNP-sites version 2.5.1 (Page et al., 2016) was run to extract SNPs from multi-FASTA alignments, ignoring indels and ambiguous sites.

The phylogenetic trees were reconstructed using RAxML version 8.2.4 (Stamatakis, 2014) on concatenated SNPs and RAxML associated to Astral-II (Mirarab et al., 2014; Mirarab and Warnow, 2015) on unlinked sequences. For RAxML we used the ASC_GTRGAMMA substitution model, with the Lewis ascertainment bias correction when working with SNPs, and 100 rapid bootstrap replicates. Both analyses were computed at the Genetic Diversity Centre (GDC, ETH Zurich). The resulting trees were visualized with FigTree version 1.4 (Rambaut, 2009).

Principal components analysis (PCA) was computed on the *Capurodendron mandrarense* lineage and all genes with the package smartPCA from EIGENSOFT program (Patterson et al., 2006) using plink formatted merged VCF files.

3. Results

3.1. Genome skimming

For *Bemangidia lowryi*, a total of 83,715,821 raw sequences were obtained. However, only 81,251,589 were retained after the quality control. K-mer coverage was estimated to fluctuate between 20x and 40x. Low contamination signals by endophytic and environmental microorganisms were detected according to the GC content bias (Supplementary Fig. 1). For *Capurodendron delphinense*, much less reads were obtained (51,775,143). Their number reduced to 50,569,253 after quality control. The coverage estimation ranged from 2x to 20x, but no contamination was detected according to the GC content bias (Supplementary Fig. 1).

3.2. Probe design

A total of 20,060 probes containing 90 bp each were designed to hybridize with 1241 loci (information and probe sequences in Supplementary Table 1). These loci comprised 80 complete genes with 1 to 12 putative exons (302 exons; total length 261952 bp, mean length 3272 +/− 895 bp), 451 exons (total length 454114 bp, mean length 1007+/−229 bp), 261 exons corresponding to 261 monocopy genes from a pool suggested by Johnson et al. (2019) for the entire angiosperm lineage (total length 16,6547 bp, mean length 635+/−267) and 227 regions containing STR and flanking regions (total length 152,118 bp, mean length 699+/−370 bp).

3.3. Library preparation

The library construction was highly affected by fragmentation in herbarium samples and viscous substances in recently collected specimens. Polysaccharides are one of the main components of the Sapotaceae latex (Fosu and Quainoo, 2013) and they frequently result in viscous or gelatinous DNA extractions (Porebski et al., 1997). Alternative DNA extraction methods were tested, as the DNeasy® Plant mini kit (Qiagen, Venlo, the Netherlands), or methods based on precipitation in high salt concentrations (Fang et al., 1992), but none of them gave significant improvement. Viscosity of the samples affected the results of DNA sonication and especially the Sera-Mag beads magnetic migration, causing a low yield at the library construction step or even a total failure. Viscosity mainly affected samples collected less than five years ago, indicating that polysaccharides may be fragmented after long storage periods.

3.4. Capture efficiency

From the 262 sequenced specimens, 31 samples were discarded (231 were kept) because they did not pass the read quality control or contained more than 80% of missing data/indel positions: 11 were older than 1970; 15 were collected from 1990 to 2014; the remaining 5 were collected in 2018 but contained high quantities of polysaccharides in DNA extraction (Table 1).

The mean number of reads per specimen for the target loci was 5.3 million, with 44 specimens with less than 1 million reads. Mapped reads to exon dataset (1014 exons, 882,613 bp, 86% of probes) represented 68.97% +/− 11.17 of the retained reads after quality-filtering, Mapped reads to STR dataset (227 regions, 152,118 bp, 14% of probes) represented 10.01% +/− 2.22 of the retained reads after quality-filtering.

Exon capture efficiency was high (Supplementary Table 2; Fig. 1), with an average number of captured genes per specimen of 92.2%. This percentage increased to 98.6% when the worst 25 samples were not taken into account. Capture efficiency is uncorrelated with the phylogenetic distance to the reference genomes, with an average of 784 (over 792) captured genes for the Tseboneae ingroup, and 788 for the farthest outgroup Chrysophylleae. The percentage of captured genes for silica-gel samples ($n = 114$) was 97.5%, while it was 87.8% for herbarium samples ($n = 148$, $p < 0.001$). When the analysis is restrained to years from which both silica-gel and herbarium specimens were collected (2010–2018), the average percentage of captured genes is 90.8% for herbarium samples ($n = 24$) and 97.5% for silica-gel preserved samples ($n = 114$). We therefore obtain a highly significant improvement of 7% ($p < 0.001$) when samples are stored in silica-gel in the field compared to herbarium preserved specimens (Supplementary Table 2, Fig. 2).

According to HybPiper and paralogs detection, out of the 792 captured genes, 638 showed a null probability to contain paralogs (80.5%), 99 may contain from one to five putative paralogs (12.5%) and 56 show signals of six or more paralogous copies (7.1%; Supplementary Table 1).

3.5. Phylogenetic reconstructions

Phylogenetic reconstructions were performed with three different datasets, containing only specimens with less than 5% (189 specimens), 40% (222 specimens), and 80% of missing data/indel positions (231 specimens; Supplementary Table 3). All three datasets showed similar topologies and the main difference was found for branch supports in the nodes connecting the main *Capurodendron* lineages. The latter were higher when specimens with a high percentage of missing data were excluded. RAxML and Astral phylogenetic trees using less than 5% missing data (Fig. 1) showed the same topology at the backbone and for supported clades. All Sapotaceae tribes are retrieved as monophyletic with bootstrap/PP values of 100/1, respectively. The Chrysophylleae tribe, belonging to the subfamily Chrysophylloideae, appears as the earliest diverging lineage from Sapotaceae. For the remaining specimens (subfamily Sapotoideae), the tribe Sideroxyleae is placed as sister of the remaining lineages, and Sapoteae is found as sister to Isonandreae. *Inhambanella* forms a clade sister to Sapoteae + Isonandreae. Finally, *Gluema* and related genera, all from continental Africa and consisting of species with dehiscent fruits, appears as a monophyletic clade sister to *Inhambanella* + Sapoteae + Isonandreae.

The uncollapsed phylogenetic tree (not illustrated) shows a good resolution at the species level, with 97.7% of the species being supported with bootstrap/PP values of 100/1, respectively. Only *Capurodendron androyense*, *C. bakeri*, *C. mandrarense*, *C. tampinense*, *Gluema ivorensis* and *Inhambanella guereensis* were found to be paraphyletic or polyphyletic. Within Tseboneae, the topology *Tsebona* (*Bemangidia* + *Capurodendron*) is found, and *Capurodendron* is divided into two main clades, one solely composed of four accessions belonging to *C. madagascariense* and the second one comprising all the other species.

Within *Capurodendron*, eleven highly supported lineages were found, each of them comprising one to 16 well-supported species. The *mandrarense* lineage was selected to test the suitability of our markers within species complexes as it comprises four morphologically, geographically

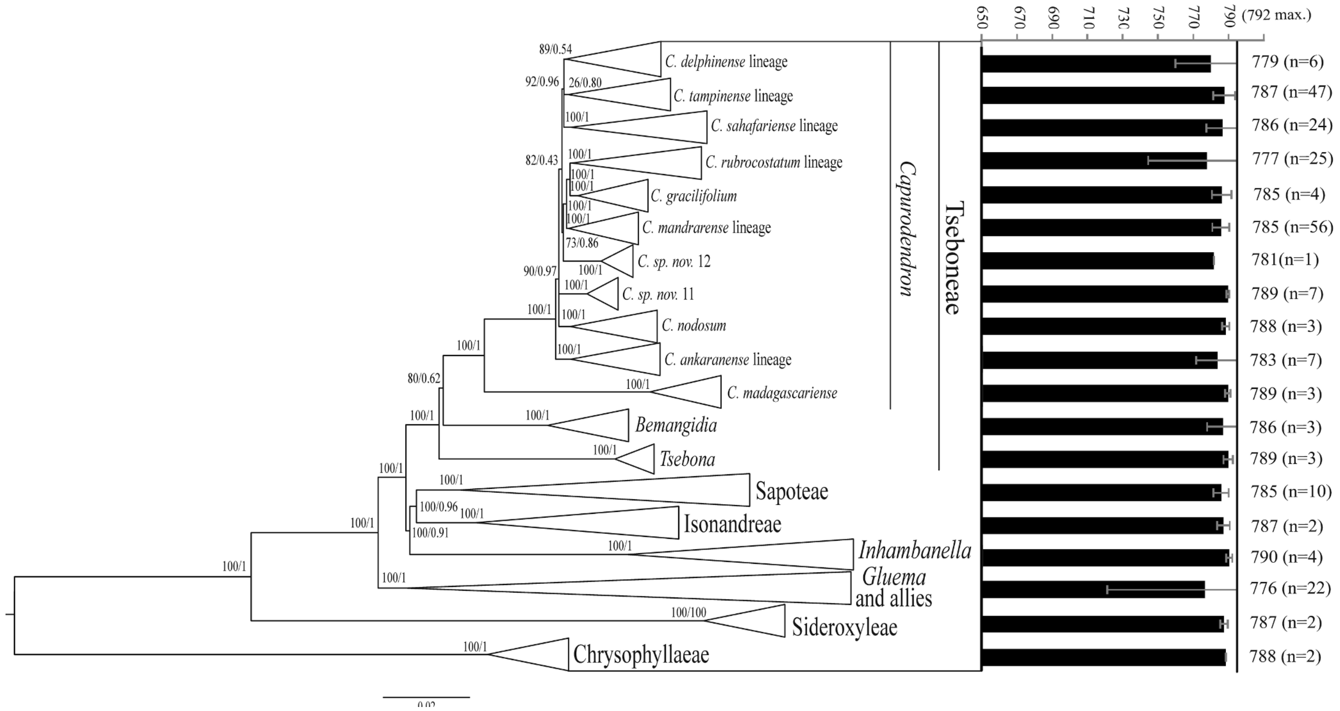


Fig. 1. RaxML and Astral tree from 638 protein coding genes and 189 specimens with less than 5% of missing data/indels. Branch numbers indicate respectively bootstrap (RaxML using 289 558 SNPs) and branch posterior probability (Astral using DNA sequences), respectively. The major clades have been collapsed at tribe, genus or infrageneric level. Bars in the right margin indicate the average number of captured genes per clade and their standard deviation (including specimens with more than 5% missing data/indels, for which lineage assignation was done from 40% and 80% missing data/indels tree).

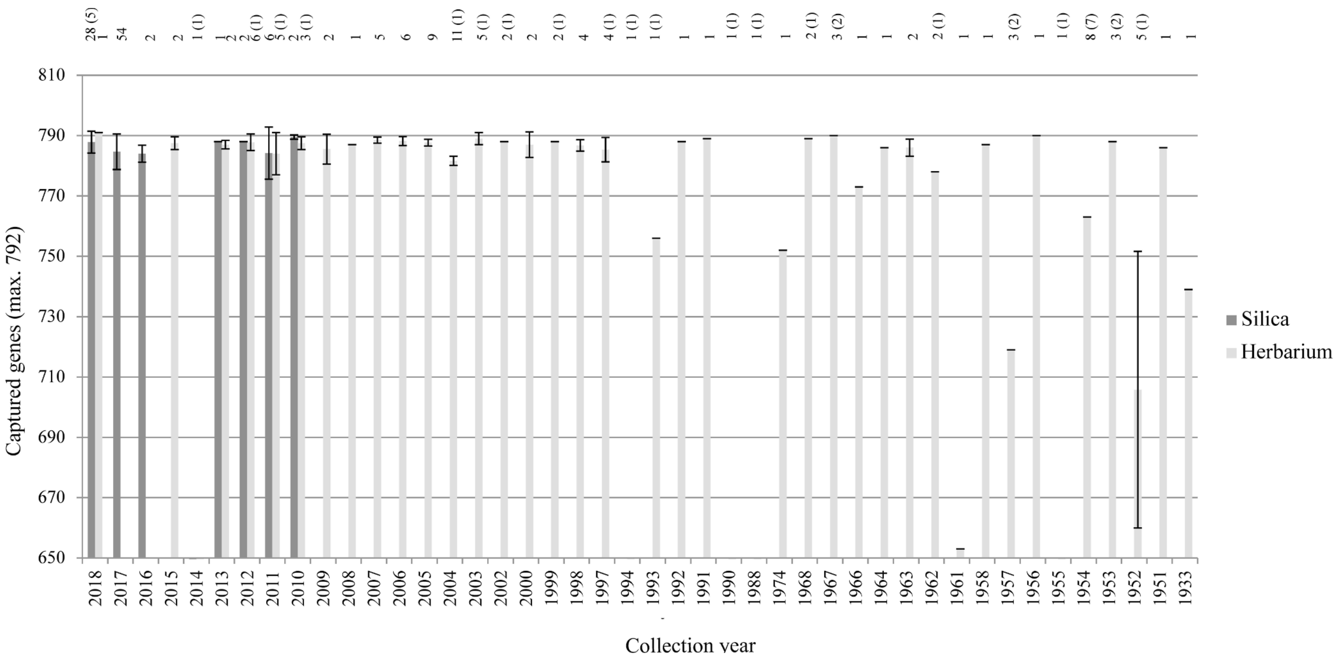


Fig. 2. Number of retrieved genes according to the kind of sampling storage (silica-gel or herbarium) and collection year for *Capurodendron* genus. Values in the upper part indicate the number of analyzed samples for each year, with the number of failed specimens between brackets.

and ecologically resembling species. It contains two monophyletic clades composed by *C. nanophyllum* and *C. microphyllum*, respectively and a third clade sister to *C. microphyllum* composed by intermixed specimens of *C. androyense* and *C. mandrarense*. The *C. mandrarense* specimens can be split into two morpho-groups, one similar to the species type specimen and the other sharing characteristics with *C. greveanum*, a genetically well isolated species from the *C. rubrocostatum* lineage. The PCA analysis performed on the entire *madrarense* lineage (Fig. 3) displays axes with a low percentage of information, (PC1 = 3.87%, PC2 = 2.88%), as expected for populations under a speciation process. Four clusters are found within the *C. mandrarense* lineage. Cluster 1 is composed entirely by *C. androyense*, from the extreme south of Madagascar. Cluster 2 contains both *C. mandrarense* and *C. androyense* morphs, all spreading in the south-

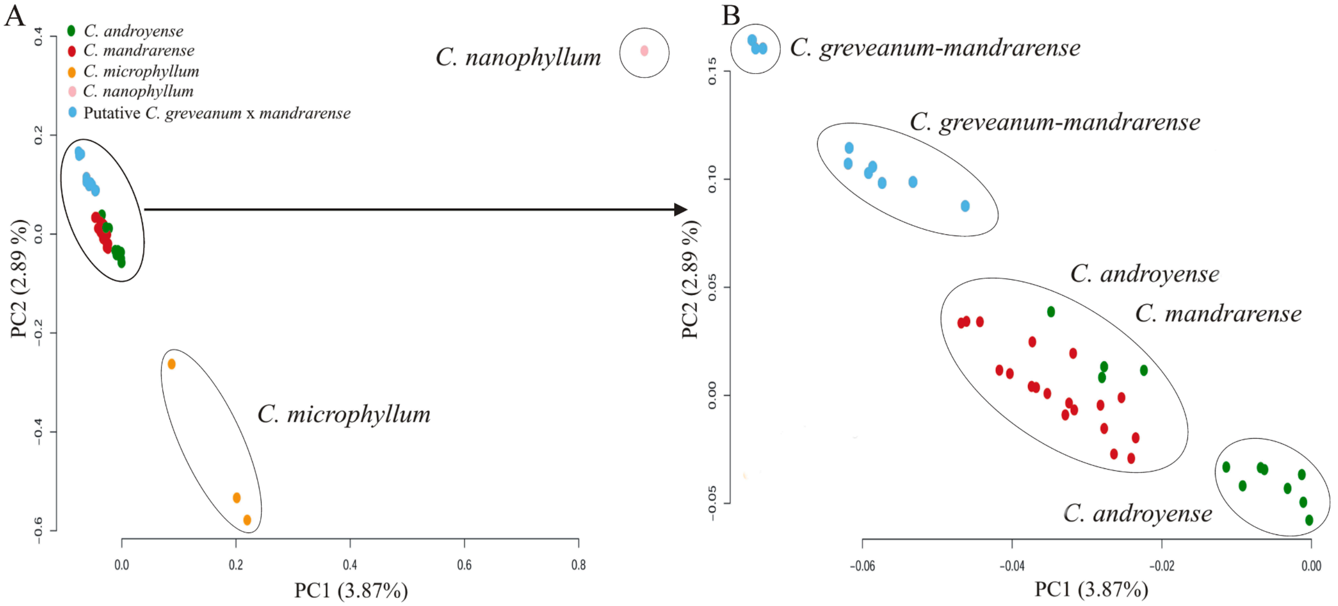


Fig. 3. Principal Components Analysis of the Capurodendron mandrarense lineage using 638 loci.

western dry regions of Madagascar. Finally, groups 3 and 4 contain specimens sharing morphological traits of both *C. mandrarense* and *C. greveanum*; it is found only on the north-western part of the area of the group, corresponding to the southernmost distribution area of *C. greveanum*. When the analyzed specimens of *C. greveanum* are included in the PCA, they form a well-defined group found very far from the *mandrarense* lineage (data not shown).

3.6. STR markers

Out of the 227 STR loci tested, 192 (84.6%) were successfully obtained for the reference species *Capurodendron delphinense*. This yield decreases considerably for the remaining species, with an average of 78 (34.4%) for the entire *Capurodendron* genus, and 10 (4.4%) for Sapotaceae out of the Tseboneae tribe. STR capture is not related to the number of reads, but it is strongly dependent of the phylogenetic distance to the reference species used to design the probes (Fig. 4).

4. Discussion

4.1. Target capture quality

Except for the oldest herbarium samples, old and fresh specimens provided similar levels of captured loci, and gene capture worked well even in highly fragmented genomic DNA (~75 bp average). However, the number of missing data in specimen increases with samples containing fewer and more fragmented DNA. This is an expected phenomenon, which can be solved by increasing the DNA concentration and washing away the small fragments in the first step of the library preparation. However, washing steps might be inefficient, as much DNA is lost, especially when DNA size fragment have similar sizes than adapter dimers. If the DNA quantity is very limited, more efficient library preparation methods may be used, as those performed in a single tube (Carøe et al., 2018), therefore avoiding washing steps. To deal with small fragments and competition with dimers, those samples have to be

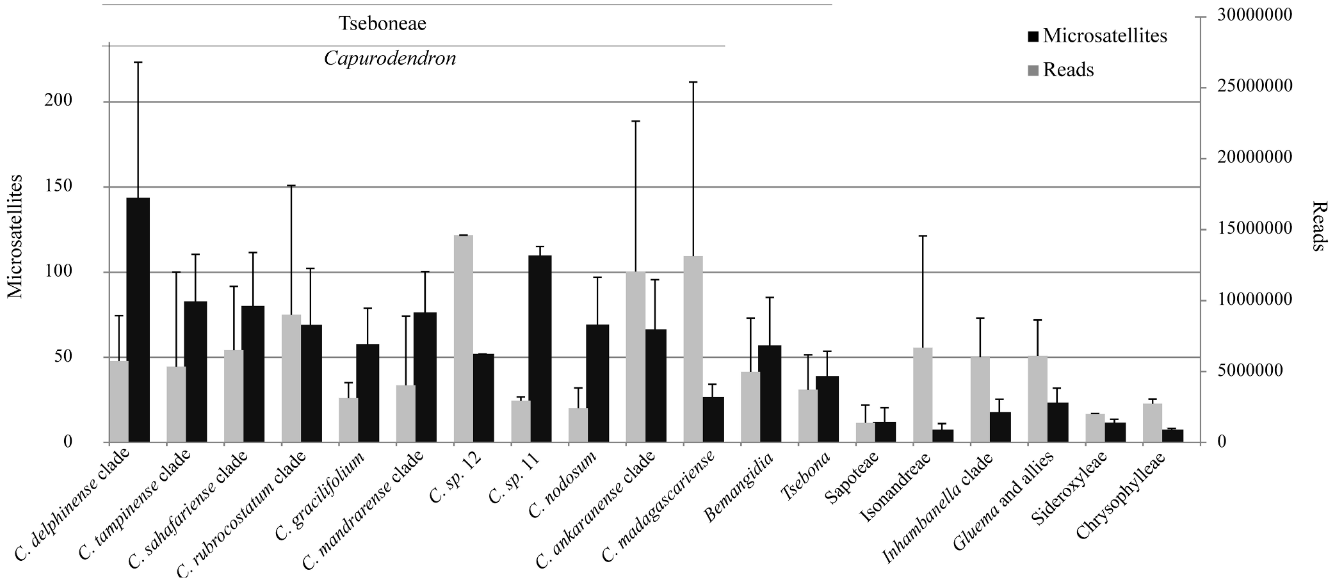


Fig. 4. Average number of captured STR (black bars) and reads (gray bars) obtained for each clade. Vertical lines on bars indicate the standard deviation.

sequenced at a lower density per lane. We obtained good results gathering 2.5 times less specimens with dimers in a single lane compared to specimens without.

Storing plant samples in silica-gel from the field is a common practice to avoid DNA fragmentation. Our results show a significant 7% improvement in loci number when silica-gel is used to preserve DNA compared to a sampling on herbarium specimens. This low, but still, relevant increase in efficiency may be related to our sequencing method, which produces reads of 100 bp and which is consequently less affected by small DNA fragments. Silica-gel storing may have a higher positive impact when longer sequences are sought for.

These markers also allow us to overcome the difficulties encountered with the standard loci (Saddhe and Kumar, 2018), such as the multicopy problems (Nieto Feliner and Rosselló, 2007), low resolution at population level (Starr et al., 2009; Caetano Wyler and Naciri, 2016; Gutiérrez-Larruscain et al., 2018), paralogy, hybridization and chloroplast capture in the case of plastid markers, or the presence of NuCp (Arthofer et al., 2010; Naciri and Manen, 2010; Hollingsworth et al., 2011; Naciri and Linder, 2015; Caetano Wyler and Naciri, 2016). Some of our loci could potentially be used as barcodes for identification purposes as they allow to bypass several of the previous issues. This would imply selecting the ones that are the more appropriate, being aware that they could be different from one taxonomic group to the other within Sapotaceae.

Sanger sequencing is sometimes hampered by the presence of repetitive sequences that blocks PCR reactions (Riet et al., 2017), as we observed it in *Gluema* and related lineages, and by latex and polysaccharides, which are abundant and difficult to remove in Sapotaceae (Michiels et al., 2003; McDevit and Saunders, 2009). All these problems are solved with the target-capture methodology using our probes, combined with the next generation sequencing methodology.

4.2. Markers suitability in Sapotaceae

Capture efficiency was high in all members of the Sapotaceae family, whether they were closely related or not to the reference genomes (Fig. 1). The oldest divergence estimations between the three genomes used for the probe design ranges between 20 mya (Rose et al., 2018) and 55 mya (Armstrong et al., 2014). The designed probes however showed a high performance in capturing genes from lineages with a much older divergence, such as the tribe Chrysophylleae, which diverged most probably during the upper Cretaceous (100–66 mya; Armstrong et al., 2014). Three subfamilies are currently described in Sapotaceae: Sapotoideae, Chrysophylloideae, and Sarcospermatoideae (Swenson and Anderberg, 2005). Our study only contains two members of the Chrysophylloidea, but they did not show any decrease in target capture efficiency, suggesting that the designed probes also work well in this subfamily. The last subfamily, Sarcospermatoideae that ranges from India to Southern China and Malaysia, contains a single small genus sister to the remaining Sapotaceae, and none of the 18 accepted species has been tested here. It is however probable that probes are able to provide a great portion, if not all, of the 792 markers.

Captured loci not only work well at high divergence times, but they are especially powerful at the species level with 97.7% of them being highly supported in the phylogenetic tree.

4.3. Future implications for the Sapotaceae taxonomy

The 638 genetic markers used here provide a highly supported topology of the backbone of the subfamily Sapotoideae, allowing the circumscription at tribe, genera and species ranks with high confidence (Fig. 1). At the upper taxonomic level, previous studies based on traditional sequencing (L. Gautier et Y. Naciri, unpublished results) were not able to clarify the monophyly of the subtribe Gluemineae, including *Inhambanella* (i.e. sensu Pennington, 1991), but our results show that it constitutes two monophyletic groups, with *Inhambanella* more closely related to Sapoteae + Isonandreae than to the rest of Gluemineae. Gluemineae, excluding *Inhambanella*, comprises two early divergent clades, reported in previous phylogenies as independent non-sister lineages (L. Gautier and Y. Naciri, unpublished results). Our reconstruction of a monophyletic clade matches better with morphological characters, especially with the dehiscent fruits and ballistic-dispersed seeds, only present in *Gluema* and the other related continental African genera *Lecomtedoxa* and *Neolemonniera*.

The resulting phylogenetic tree confirms that captured genes exhibit a good phylogenetical signal from family down to population level, and are able to deal with species complexes and with radiations such as the ones that seem to have occurred in *Capurodendron*. We expect that phylogenetic reconstructions using the coalescence as in *Beast (Bouckaert et al., 2019) or STACEY (Jones et al., 2015; Jones, 2015), may produce even better supported trees, as they are able to deal with incomplete lineage sorting. Our preliminary phylogeny has validated more than 20 new species of Sapotaceae in the genus *Capurodendron* solely. With these promising results, future research will be able to establish a clear species concept in this family generally considered as taxonomically challenging. Current research focused on Glueminae and Sapoteae using our probes (Gautier et al., in prep.; Randriarisoa et al., in prep.) confirms its efficiency in other Sapotaceae tribes and the presence of still undescribed species.

4.4. Detection of intraspecific variation

Loci variability is high enough to furthermore separate lineages within a species complex, as it is the case in the *mandrarense* lineage. This lineage comprises several morphospecies that inhabit the driest parts of Madagascar. However, the species limits are not clear, especially in *C. mandrarense*, a widely distributed species partially sympatric with *C. androyense* and *C. greveanum*. The PCA was able to detect geographical structure within those species, as well as putative hybrids versus well delimited putative species (Fig. 3). For example, the western and southern populations of *C. androyense,* although morphologically very similar, are genetically differentiated. Additionally, the western population shares genetic affinity with *C. mandrarense*, indicating a possible hybrid component. In *C. mandrarense* several groups emerge, gathering specimens from the southeastern region, the inland highlands, or those that share phenotypical characters with *C. greveanum*. The latter specimens may correspond to hybrids between *C. mandrarense* and *C. greveanum*. All our data confirm that captured loci are suitable to study intraspecific population structure. Additionally, we only tested the protein-coding exons here, but target capture can provide also long fragments of introns, especially if recently collected silica-gel samples are processed and a sequencing methodology producing reads longer than 100 bp is used. As intronic regions are more variable than exonic ones (Sang, 2002), the designed probes can also be used to obtain more variable sequences than those utilized here.

The number of captured STR loci decreases dramatically with the phylogenetic distance to the reference species (Fig. 4). This suggests that a great proportion of the selected STR loci are specific to *C. delphinense*, something known for long as the ascertainment bias (Hutter et al., 1998). Surprisingly the loci retrieved across all other species, seem to be randomly distributed, as no locus was consistently obtained for specific lineages (Supplementary Fig. 1). A likely explanation may be related to the used bioinformatics pipeline, which may provide biased results towards the reference species, especially at the filtering level, as percentage of mapping reads is consistent with the number of STR included in the probes set and not related to the phylogenetic signal. Potential solutions will be investigated in a future article focusing on species delimitation in species complexes using the full potential of our probes. However, it is already advised to use several distantly related species to design baits for STR markers. This is expected to increase the capture efficiency on non-focal species.

Target capture holds great promises for biodiversity genomic studies. It alleviates several obstacles by the combination of new methods that are particularly suitable for both the type of questions and the type of samples used in this field. Efficient library preparation, reduced representation genome sequencing and high throughput sequencing, that was first developed for small fragments, tackle particularly well many old challenges that scientists studying non-model organisms had to face. Challenging source of tissues such as those found in herbarium and historical collections or samples with high amount of secondary metabolites can now be part of the sample design. Design of custom probes, based on the target organisms and the taxonomic levels of the study, as well as inclusion of universal probes allows to target the correct loci for a high number of study samples in order to concentrate the sequencing effort on the loci that can give the highest information.

With good practices but relatively low investment, target capture keeps its promises to get valuable information, at different taxonomic level, in a family with few existing genetic resources, challenging source of tissues but with a high conservation importance. We provide here a set of probes able to retrieve 792 nuclear genes, even on difficult material. From those loci, at least 638 are paralog-free and informative for the entire Sapotaceae family from the tribe to the population level. We are confident that those new markers will contribute to a better understanding of the Sapotaceae taxonomy and their conservation.

CRediT authorship contribution statement

Camille Christe: Data curation, Software, Formal analysis, Methodology, Writing - original draft, Writing - review & editing. **Carlos G. Boluda:** Formal analysis, Methodology, Software, Writing - original draft, Writing - review & editing. **Darina Koubínová:** Formal analysis, Writing - review & editing. **Laurent Gautier:** Conceptualization, Funding acquisition, Writing - review & editing, Supervision. **Yamama Naciri:** Conceptualization, Funding acquisition, Writing - review & editing, Supervision.

Acknowledgements

This work was supported by a grant from the Swiss National Foundation attributed to YN and LG (n° 31003A-166349), a grant from the Franklinia Foundation attributed to LG (n° 2019-20) and by two grants from the Schmidheiny Foundation attributed to YN in 2016 and 2018. We thank the iGE3 platform for their help in the sequencing process, and Richard Randrianaivo, Aina Randriarisoa, as well as the Malagasy Government and the Malagasy local people for their help in sampling Sapotaceae specimens. Data produced and analyzed in this paper were generated in collaboration with the Genetic Diversity Centre (GDC), ETH Zurich. We thank Patrik Mráz for helping us trying to estimate the samples C contents. We are grateful to the herbaria MO, P, TAN and TEF for allowing us to study herbarium specimens and to perform destructive sampling. Computational resources used by DK were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Data availability

Following files are deposited at Zenodo and they are freely accessible (DOI:10.5281/zenodo.4434519): Sapotaceae baits, the reference target sequences for the baits for the genes and for the STR separately, the list of target genes and STR, cleaned and aligned sequences for the 791 loci, as well as RAxML and Astral trees for the different missing data threshold presented in this paper. Raw sequences of all individuals are accessible in the NCBI BioProject PRJNA691138 and BioSample number for each individual is listed in Table 1 as well as in Supplementary Table 2.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2021.107123.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Armstrong, K.E., Stone, G.N., Nicholls, J.A., Valderrama, E., Anderberg, A.A., Smedmark, J., et al., 2014. Patterns of diversification amongst tropical regions compared: a case study in Sapotaceae. Front. Genet. 5, 362. https://doi.org/10.3389/fgene.2014.00362.

Arthofer, W., Schüler, S., Steiner, F.M., Schlick-Steiner, B.C., 2010. Chloroplast DNA-based studies in molecular ecology may be compromised by nuclear-encoded plastid sequence. Mol. Ecol. 19 (18), 3853–3856. https://doi.org/10.1111/j.1365-294X.2010.04787.x.

Aubréville, A., 1974. Sapotaceae. In: Humbert, H., Leroy, J.-F. (Eds.), Flore de Madagascar et des Comores: plantes vasculaires: Vol. fam 164: S. Tananarive : Imprimerie officielle; Retrieved from https://www.biodiversitylibrary.org/item/36318.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Borowiec, M.L., 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. e1660-e1660 PeerJ 4. https://doi.org/10.7717/peerj.1660.

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al., 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15 (4), e1006650 https://doi.org/10.1371/journal.pcbi.1006650.

Broad Institute. http://broadinstitute.github.io/picard/.

Buddenhagen, C., Lemmon, A.R., Lemmon, E.M., Bruhl, J., Cappa, J., Clement, W.L., et al., 2016. Anchored phylogenomics of angiosperms I: Assessing the robustness of phylogenetic estimates. BioRxiv 86298. https://doi.org/10.1101/086298.

Caetano Wyler, S., Naciri, Y., 2016. Evolutionary histories determine DNA barcoding success in vascular plants: seven case studies using intraspecific broad sampling of closely related species. BMC Evol. Biol. 16 (1), 103. https://doi.org/10.1186/s12862-016-0678-0.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25 (15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.H.S., Samaniego, J.A., et al., 2018. Single-tube library preparation for degraded DNA. Methods Ecol. Evol. 9 (2), 410–419. https://doi.org/10.1111/2041-210X.12871.

de La Harpe, M., Hess, J., Loiseau, O., Salamin, N., Lexer, C., Paris, M., 2019. A dedicated target capture approach reveals variable genetic markers across micro- and macro-evolutionary time scales in palms. Mol. Ecol. Resour. 19 (1), 221–234. https://doi.org/10.1111/1755-0998.12945.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. May, 43(5), 491–498. https://doi.org/10.1038/ng.806.

Ewango, C.E.N., Kenfack, D., Sainge, M.N., Thomas, D.W., van der Burgt, X.M., 2016. *Gambeya korupensis* (Sapotaceae: Chrysophylloideae), a new rain forest tree species from the Southwest Region in Cameroon. Kew Bull. 71 (2), 28. https://doi.org/10.1007/s12225-016-9633-x.

Fajardo, D., Schlautman, B., Steffan, S., Polashock, J., Vorsa, N., Zalapa, J., 2014. The American cranberry mitochondrial genome reveals the presence of selenocysteine (tRNA-Sec and SECIS) insertion machinery in land plants. Gene 536 (2), 336–343. https://doi.org/10.1016/j.gene.2013.11.104.

Fang, G., Hammar, S., Grumet, R., 1992. A quick and inexpensive method for removing polysaccharides from plant genomic DNA. Biotechniques 13 (1), 52–54,56.

Forrest, L.L., Hart, M.L., Hughes, M., Wilson, H.P., Chung, K.-F., Tseng, Y.-H., Kidner, C. A., 2019. The limits of hyb-seq for herbarium specimens: impact of preservation techniques. Front. Ecol. Evol. 7, 439. https://doi.org/10.3389/fevo.2019.00439.

Fosu, R., Quainoo, A.K., 2013. Comparism of the Proximate composition of Shea (*Vitellaria paradoxa*) and Rubber Latex (*Hevea brasilienesis*). Int. J. Res. Stud. Biosci. 1 (2), 14–16.

Gautier, L., 2003. Sapotaceae. In: Goodman, J.P., Bensted, S. M. (Ed.). The Natural History of Madagascar. The University of Chicago Press, Chicago, pp. 342–346.

Gautier, L., Farias do Valle, M., Boluda, C.G, W. Stauffer, F., Naciri, Y., n.d. The Gluemeae: a new tribe to accommodate the maverick African rainforest Sapotaceae with dehiscent fruits. Mol. Phyl. Evol. (in preparation).

Gautier, L., Lachenaud, O., van der Burgt, X., Kenfack, D., 2016. Five new species of *Englerophytum* K. Krause (Sapotaceae) from central Africa. Candollea 71 (2), 287–305.

Gautier, L., Naciri, Y., 2018. Three Critically Endangered new species of *Capurodendron* (Sapotaceae) from Madagascar. Candollea 73 (1), 121–129. https://doi.org/10.15553/c2018v731a13.

Gautier, L., Naciri, Y., Anderberg, A.A., Smedmark, J.E.E., Randrianaivo, R., Swenson, U., 2013. A new species, genus and tribe of Sapotaceae, endemic to Madagascar. Taxon 62 (5), 972–983. http://www.jstor.org/stable/taxon.62.5.972.

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

Jones, G., 2015. STACEY: Species delimitation and phylogeny estimation under the multispecies coalescent. BioRxiv 10199. https://doi.org/10.1101/010199.

Jones, G., Aydin, Z., Oxelman, B., 2015. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics 31 (7), 991–998. https://doi.org/10.1093/bioinformatics/btu770.

Gutiérrez-Larruscain, D., Santos-Vicente, M., Anderberg, A.A., Rico, E., Martínez-Ortega, M.M., 2018. Phylogeny of the *Inula* group (Asteraceae: Inuleae): Evidence from nuclear and plastid genomes and a recircumscription of *Pentanema*. Taxon 67 (1), 149–164. https://doi.org/10.12705/671.9.

Gymrek, M., Golan, D., Rosset, S., Erlich, Y., 2012. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 22 (6), 1154–1162. https://doi.org/10.1101/gr.135780.111.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107 (1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.

Heyduk, K., Trapnell, D.W., Barrett, C.F., Leebens-Mack, J., 2016. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. Biol. J. Linn. Soc. 117 (1), 106–120. https://doi.org/10.1111/bij.12551.

Hollingsworth, P.M., Graham, S.W., Little, D.P., 2011. Choosing and using a plant DNA barcode. PLoS ONE 6 (5), e19254. https://doi.org/10.1371/journal.pone.0019254.

Hutter, C.M., Schug, M.D., Aquadro, C.F., 1998. Microsatellite variation in *Drosophila melanogaster* and *Drosophila simulans*: A reciprocal test of the ascertainment bias hypothesis. Mol. Biol. Evol. 15 (12), 1620–1636. https://doi.org/10.1093/oxfordjournals.molbev.a025890.

Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., et al., 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 4 (7) https://doi.org/10.3732/apps.1600016.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., et al., 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-Medoids clustering. Syst. Biol. 68 (4), 594–606. https://doi.org/10.1093/sysbio/syy086.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780. https://doi.org/10.1093/molbev/mst010.

Kedzierska, K.Z., Gerber, L., Cagnazzi, D., Krützen, M., Ratan, A., Kistler, L., 2018. SONiCS: PCR stutter noise correction in genome-scale microsatellites. Bioinformatics 34 (23), 4115–4117. https://doi.org/10.1093/bioinformatics/bty485.

Kircher, M., Sawyer, S., Meyer, M., 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40 (1), e3 https://doi.org/10.1093/nar/gkr771.

Kistler, L., Johnson, S.M., Irwin, M.T., Louis, E.E., Ratan, A., Perry, G.H., 2017. A massively parallel strategy for STR marker development, capture, and genotyping. Nucleic Acids Res. 45 (15), e142 https://doi.org/10.1093/nar/gkx574.

Kümpers, B.M.C., Richardson, J.E., Anderberg, A.A., Wilkie, P., Ronse De Craene, L.P., 2016. The significance of meristic changes in the flowers of Sapotaceae. Bot. J. Linn. Soc. 180 (2), 161–192. https://doi.org/10.1111/boj.12363.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359. https://doi.org/10.1038/nmeth.1923.

Mackinder, B., Harris, D.J., Gautier, L., 2016. A reinstatement, recircumscription and revision of the genus *Donella* (Sapotaceae). Edinburgh J. Bot. 73 (3), 297–339. https://doi.org/10.1017/S0960428616000160.

Magoč, T., Salzberg, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27 (21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507.

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., et al., 2014. Data access for the 1,000 Plants (1KP) project. GigaScience 3 (1), 17. https://doi.org/10.1186/2047-217X-3-17.

McDevit, D.C., Saunders, G.W., 2009. On the utility of DNA barcoding for species differentiation among brown macroalgae (Phaeophyceae) including a novel extraction protocol. Phycol. Res. 57 (2), 131–141. https://doi.org/10.1111/j.1440-1835.2009.00530.x.

Michiels, A., Van den Ende, W., Tucker, M., Van Riet, L., Van Laere, A., 2003. Extraction of high-quality genomic DNA from latex-containing plants. Anal. Biochem. 315 (1), 85–89. https://doi.org/10.1016/S0003-2697(02)00665-6.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30 (17), i541–8. https://doi.org/10.1093/bioinformatics/btu462.

Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31 (12), i44–52. https://doi.org/10.1093/bioinformatics/btv234.

Moorthie, S., Mattocks, C.J., Wright, C.F., 2011. Review of massively parallel DNA sequencing technologies. HUGO J. 5 (1–4), 1–12. https://doi.org/10.1007/s11568-011-9156-3.

Naciri, Y., Linder, H.P., 2015. Species delimitation and relationships: The dance of the seven veils. Taxon 64 (1), 3–16. https://doi.org/10.12705/641.24.

Naciri, Y., Manen, J.F., 2010. Potential DNA transfer from the chloroplast to the nucleus in *Eryngium alpinum*. Mol. Ecol. Resour. 10 (4), 728–731. https://doi.org/10.1111/j.1755-0998.2009.02816.x.

Nicholls, J.A., Pennington, R.T., Koenen, E.J.M., Hughes, C.E., Hearn, J., Bunnefeld, L., et al., 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). Front. Plant Sci. 6, 710. https://doi.org/10.3389/fpls.2015.00710.

Nieto Feliner, G., Rosselló, J.A., 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Mol. Phylogenet. Evol. 44 (2), 911–919. https://doi.org/10.1016/j.ympev.2007.01.013.

Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., Harris, S.R., 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genomics 2 (4), e000056. https://doi.org/10.1099/mgen.0.000056.

Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLOS Genetics 2 (12), e190. https://doi.org/10.1371/journal.pgen.0020190.

Pennington, T., 1991. The genera of the Sapotaceae. Royal Botanic Gardens, Kew.

Porebski, S., Bailey, L.G., Baum, B.R., 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol. Biol. Rep. 15 (1), 8–15. https://doi.org/10.1007/BF02772108.

Rambaut, A., 2009. FigTree v1.3.1. Computer program available from: <http://tree.bio.ed.ac.uk/software/figtree/> (accessed October 2019).

Randriarisoa, A., Naciri, Y., Armstrong, K., Boluda, C.G., Daffrevile, S., Gautier, L., n.d. A genus sinks, another one emerges. Taxon (in preparation).

Randriarisoa, A., Naciri, Y., Gautier, L., 2020. *Labramia ambondrombeensis* (Sapotaceae), a Critically Endangered new species from Madagascar. Candollea 75 (1), 83–87. https://doi.org/10.15553/c2020v751a8.

Riet, J., Ramos, L.R.V., Lewis, R.V., Marins, L.F., 2017. Improving the PCR protocol to amplify a repetitive DNA sequence. Genet. Mol. Res.: GMR 16 (3). https://doi.org/10.4238/gmr16039796.

Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A., Mesirov, J.P., 2017. Variant review with the integrative genomics viewer. Cancer Res. 77 (21), e31–e34. https://doi.org/10.1158/0008-5472.CAN-17-0337.

Rokni, S., Wursten, B., and Darbyshire, I., 2019 (16 C.E.). Synsepalum chimanimani (Sapotaceae), a new species from the Chimanimani Mountains of Mozambique and Zimbabwe, with notes on the botanical importance of this area. PhytoKeys, 133, 115–132. https://doi.org/10.3897/phytokeys.133.38694.

Rose, J.P., Kleist, T.J., Löfstrand, S.D., Drew, B.T., Schönenberger, J., Sytsma, K.J., 2018. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. Mol. Phylogenet. Evol. 122, 59–79. https://doi.org/10.1016/j.ympev.2018.01.014.

Russell, A., Samuel, R., Rupp, B., Barfuss, M.H.J., Šafran, M., Besendorfer, V., Chase, M. W., 2010. Phylogenetics and cytology of a pantropical orchid genus *Polystachya* (Polystachyinae, Vandeae, Orchidaceae): Evidence from plastid DNA sequence data. Taxon 59 (2), 389–404. http://www.jstor.org/stable/25677598.

Saddhe, A.A., Kumar, K., 2018. DNA barcoding of plants: Selection of core markers for taxonomic groups. Plant Sci. Today 5 (1), 9–13. https://doi.org/10.14719/pst.2018.5.1.356.

Sang, T., 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 37 (3), 121–147. https://doi.org/10.1080/10409230290771474.

Sangjin, J., Hoe-Won, K., Young-Kee, K., Se-Hwan, C., Ki-Joong, K., 2016. The first complete plastome sequence from the family Sapotaceae, *Pouteria campechiana* (Kunth) Baehni. Mitochondrial DNA Part B 1 (1), 734–736. https://doi.org/10.1080/23802359.2016.1233469.

Sass, C., Iles, W.J.D., Barrett, C.F., Smith, S.Y., Specht, C.D., 2016. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. PeerJ 4, e1584. https://doi.org/10.7717/peerj.1584.

Schmickl, R., Liston, A., Zeisek, V., Oberlander, K., Weitemier, K., Straub, S.C.K., et al., 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). Mol. Ecol. Resour. 16 (5), 1124–1135. https://doi.org/10.1111/1755-0998.12487.

Simpson, J.T., 2014. Exploring genome characteristics and sequence quality without a reference. Bioinformatics 30 (9), 1228–1235. https://doi.org/10.1093/bioinformatics/btu023.

Simpson, J.T., Durbin, R., 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 22 (3), 549–556. https://doi.org/10.1101/gr.126953.111.

Souza, H.A.V., Muller, L.A.C., Brandão, R.L., Lovato, M.B., 2012. Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis* (Leguminosae), a tree from the Brazilian Cerrado. Genet. Mol. Res.: GMR 11 (1), 756–764. https://doi.org/10.4238/2012.March.22.6.

Staats, M., Cuenca, A., Richardson, J.E., Vrielink-van Ginkel, R., Petersen, G., Seberg, O., Bakker, F.T., 2011. DNA damage in plant herbarium tissue. PLoS ONE 6 (12), e28448. https://doi.org/10.1371/journal.pone.0028448.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30 (9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Starr, J.R., Naczi, R.F.C., Chouinard, B.N., 2009. Plant DNA barcodes and species resolution in sedges (Carex, Cyperaceae). Mol. Ecol. Resour. 9 Suppl s1, 151–163. https://doi.org/10.1111/j.1755-0998.2009.02640.x.

Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., Malmberg, R.L., 2015a. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. Mol. Phylogenet. Evol. 85, 76–87. https://doi.org/10.1016/j.ympev.2015.01.015.

Stephens, J.D., Rogers, W.L., Mason, C.M., Donovan, L.A., Malmberg, R.L., 2015b. Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. Am. J. Bot. 102 (6), 910–920. https://doi.org/10.3732/ajb.1500031.

Swenson, U., Anderberg, A.A., 2005. Phylogeny, character evolution, and classification of Sapotaceae (Ericales). Cladistics 21 (2), 101–130. https://doi.org/10.1111/j.1096-0031.2005.00056.x.

Uribe-Convers, S., Settles, M.L., Tank, D.C., 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). PLoS ONE 11 (2), e0148203. https://doi.org/10.1371/journal.pone.0148203.

VanBuren, R., Paris, M., Wai, J., Zhang, J., Huang, L., Zhou, H., et al., 2018. Sexual recombination and selection during domestication of clonally propagated pineapple. Cell. https://doi.org/10.2139/ssrn.3155832.