Report 2. A DNA barcoding toolkit for the focus species and its implementation

Executive summary

Labwork and computational analyses were undertaken to test if it was possible to develop a functional timber DNA barcoding toolkit for selected species of the mahogany family and to identify challenges to overcome for routine implementation of timber DNA barcoding in Gabon and DRC. Main goals were to 1) Build a comprehensive DNA reference dataset for a set of focus species, 2) Test and optimise DNA extraction protocols for different types of wood likely to be encountered by law enforcement authorities, 3) identify DNA barcodes informative enough to distinguish the focus species while being small enough to be sequenced from wood DNA using Polymerase Chain Reaction (PCR) amplification and Sanger sequencing, 4) evaluate the feasibility of different DNA barcoding approaches in Gabon and DRC, and 5) conclude on the feasibility of timber DNA barcoding in these countries and on requirements for its routine implementation to monitor the illegal timber trade of any species of interest (beyond the focus species).

To build the reference dataset, we used a combination of target capture sequencing of 353 nuclear genes (Angiosperms353 baits) and genome skimming to also obtain plastid regions and the ribosomal nuclear region ITS. After exclusion of samples with low recovery or dubious identity, the final reference dataset for the four focus genera (*Entandrophragma*, *Khaya*, *Lovoa* and *Swietenia*) comprised up to 351 nuclear and 107 plastid regions for 132 samples representing 22 out of the 24 species included in the four genera. DNA extraction protocols were tested and adjusted through more than 360 extractions. The reference dataset was used to identify new DNA barcodes, and PCR and sequencing of these barcodes were tested on 19 DNA samples, among which 10 (both sapwood and heartwood) yielded sequences that could be used successfully for sample identification. All the DNA work has been written up in Article 1, now submitted to Molecular Ecology Resources (Annex 5.5 of the main report), and associated data are available on GenBank and in Github at https://github.com/sidonieB/Bellot_al_Meliaceae_DNA_barcoding/tree/main. Lab visits and discussions were undertaken and showed that Gabon has already facilities in place to perform timber DNA barcoding tests in situ while building an entirely new lab may be required in DRC. Equipment lists and associated costs have been compiled.

Based on these results, and considering the fact that stakeholder consultations highlighted the need for quick and cheap timber DNA barcoding tests at export points, we identify the PCR and Sanger sequencing approach as the most readily available and feasible in the focus countries, while high-throughput approaches (both short-read and long-read based) would be more costly, more logistically complex to implement and would require more research before implementation. Although our results demonstrate that this approach can be applied for timber DNA barcoding, we identify at least five areas of future research and development to enable the deployment of DNA barcoding for the monitoring of illegal timber trade in the focus countries: 1) Meliaceae DNA barcoding tests in Gabon labs involving all relevant stakeholders including policy makers; 2) DNA barcodes and reference dataset development for all other species of interest beyond the four focus genera; 3) Strengthening botanical and DNA barcoding training in the focus countries; 4) Developing lab infrastructure for plant DNA barcoding in both countries but especially in DRC; and 5) Further streamlining DNA extraction protocols to increase success rates. These research streams could be implemented as a single large project or multiple smaller ones, but in any case their success will require collaboration between all relevant stakeholders in the focus countries and beyond (as identified in Report 1).

NB: some of the text and figures included in this report are copied from Article 1 (Annex 5.5 of the main project report)

1. Reference dataset

a. Building the reference dataset

To generate data for the reference database, leaf tissue was sampled from herbarium specimens of as many species as possible from the four focus genera. Specimens were mainly obtained from Kew herbarium (K) although a few were obtained from World Forest ID collections deposited in the Jodrell Laboratory of Royal Botanic Gardens, Kew. As a result, reference DNA data (see below) was obtained from at least one sample of all species of *Entandrophragma* (11 species), *Lovoa* (2 species), and *Swietenia* (3 species) and of seven out of eight species of *Khaya*. The missing species, *Khaya madagascariensis*, is an endemic of Madagascar that was beyond the scope of this study. Another species, *K. anthotheca*, was sampled but the different samples were not recovered as monophyletic in preliminary versions of our species trees, without a clear indication that any of the samples was indeed representing *K. anthotheca*. This is due to the complex taxonomy of this species, which makes it difficult to ascertain that samples previously identified as *K. anthotheca* indeed represent this species and not other described (e.g. *K. nyasica*) or yet undescribed species (Bouka at al., 2022). Target capture and genome skimming raw data from four samples originally assigned to this species have been deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under *Khaya sp.* so that they can be used in future studies (see Table 1 below for accession numbers), but they were not included in phylogenetic or barcoding analyses described below. Except for *Entandrophragma palustre*, *E. bussei*, *K. agboensis*, and *K. euryphylla*, which were represented by a single sample in the reference DNA dataset, all other species were represented by 2 to 18 individuals from across their distribution range (Table 1).

DNA extraction from herbarium leaf material was performed following a modification of the cetrimonium bromide (CTAB) method (Doyle & Doyle, 1987) as described in Brewer et al. (2019). DNA concentration was measured with a Quantus fluorometer (Promega UK Ltd, Southampton, UK) and DNA size was measured on an Agilent 4200 TapeStation (Agilent Technologies LDA UK Limited, Stockport, UK). DNA samples with a high proportion of DNA fragments above 500 bp were sonicated before library preparation using a Covaris ME220 Focused-ultrasonicator (Covaris Ltd, Brighton, UK). DNA library preparation was done using a NEBNext Ultra II library kit (New England BioLabs Ltd, Hitchin, UK), following the manufacturer's protocol with some modifications depending on DNA quality. Libraries were dual indexed using the NEBNext Multiplex Oligos for Illumina (New England BioLabs Ltd, Hitchin, UK). DNA libraries were pooled to reach approximately equal molarities of all libraries in a given pool. Libraries from wood DNA were pooled separately. This resulted in 10 pools of 14 to 26 libraries. A separate "genome skimming" pool was made by pooling together 2 uL from each pool made from DNA obtained from leaf tissue, while the rest of the pools was subjected to target sequence capture. Target sequence capture consisted in using the RNA bait kit Angiosperms353 (Johnson et al., 2019) to capture a set of 353 genes previously shown to be present and phylogenetically informative at the species level and below in many angiosperm genera (Slimp et al., 2021, Baker et al. 2022). The bait kit was manufactured by Arbor Daicel and target sequence capture followed their "Standard" protocol (http://www.arborbiosci.com/mybaits-manual). The hybridisation step lasted 24 hours at 62°C and was followed by PCR amplification (16-20 cycles depending on the pool). The resulting pools enriched in the 353 genes targeted by the Angiosperms353 baits were then pooled together for Illumina sequencing. The skimming and enriched pools were sequenced on an Illumina NovaSeq X platform at Macrogen Inc. (Seoul, Korea), to generate 150-bp long paired-end sequencing reads for each sample. An Angiosperms353 target sequence capture dataset obtained similarly for *Schmardaea microphylla* (accession ERR7620407; Baker et al., 2021) was retrieved from the European Nucleotide Archive to be used as an outgroup. Reads obtained from target capture sequencing and genome skimming were pooled together for each sample. Sequence data quality was assessed using FASTQC version 0.11.9 (Andrews, 2010). Illumina sequencing adapters were

removed from the reads using Trimmomatic v. 0.39 (Bolger et al., 2014) in paired-end palindrome mode with 1 seed mismatch allowed, palindrome and simple clips thresholds of 30 and 7 respectively, a minimum adapter length of 2 bp and keeping both reads. Bases with low quality at the end of the reads were then removed using a sliding window ("SLIDINGWINDOW" parameter) of 4 bp and a minimum Phred quality threshold of 30, while bases of low quality at the beginning of the reads were removed using the "LEADING" parameter and the same minimum quality threshold. Reads shorter than 40 bp after the trimming were discarded. Clean reads were first analyzed using HybPiper v. 2 (Johnson et al., 2016) to recover and assemble the regions targeted by the Angiosperms353 bait kit and regions from the chloroplast genome (expected to be recovered through genome skimming due to their high copy number). HybPiper works by aligning sequencing reads to reference sequences of the target regions in order to identify the reads corresponding to each region, and then by assembling the reads of each region separately into sequences corresponding to the target regions. We used the default Angiosperms353 reference sequences available at https://github.com/mossmatters/Angiosperms353/Angiosperms353_targetSequences.fasta, complemented with 78 coding sequences, 95 intergenic spacers and 4 ribosomal RNA sequences retrieved from the published plastome of Swietenia mahagoni (GenBank accession NC_040009). This preliminary analysis revealed many genes classified as potentially paralogous by HybPiper (i.e. more than one copy could be assembled for the gene in at least one taxon), so we used a different pipeline, CAPTUS v. 1 (Ortiz et al., 2023) to perform the final gene recovery from all our samples. CAPTUS performs de novo assembly of all reads into contigs before aligning the contigs to reference sequences in order to recover the target loci. This enables a flexible use of the multiple (potentially paralog) copies that may be assembled for a given gene: they can either be all used in downstream phylogenetic analyses (hereafter this approach is referred to as the paralog-inclusive, PI, approach), or one of the copies can be selected for downstream analyses based on its similarity to the reference sequences (hereafter paralog-exclusive, PE, approach). To explore the impact of potential paralogs on phylogenetic inferences, we followed both approaches in parallel. For these final analyses with CAPTUS, two new sets of reference sequences were created. The first set comprised only nuclear protein-coding sequences, including the above-mentioned default Angiosperms353 references, but also the sequences of the Angiosperms353 target genes obtained in four of our highest quality samples (one per genus) during the preliminary HybPiper analysis. This set of references was checked and "fixed" using the check_targetfile and fix_targetfile commands in HybPiper v. 2 so that low complexity sequences and sequences with internal stop codons were excluded. The second set comprised all the plastome regions mentioned above as well as published sequences of the Internal Transcribed Spacers (ITS1 and ITS2) and small unit (5.8S) of the nuclear ribosomal DNA region (hereafter referred to as ITS) from 13 species belonging to the four focus genera. When retrieving a sequence corresponding to a target region with CAPTUS, the user can choose to not only retrieve the part that matches the reference, but instead to retrieve any further internal section of the assembled sequence that does not match the reference. For instance if the reference only includes the exons of a gene but the recovered sequence also includes introns, we can keep the complete sequence including introns and exons. This is what we did here, using the "GE" and "MA" format options in the CAPTUS Align module, so that the retrieved sequences would contain as much information as possible.

As part of the CAPTUS pipeline, a file was created for each target region, containing the region sequence (or sequences in the PI approach) of all samples for which the region was successfully assembled. For each region, the sequences of all samples were then aligned using MAFFT v. 7 (Katoh & Standley, 2013) and the alignments were trimmed from gappy columns using ClipKIT (Steenwyk et al. 2020), both via the CAPTUS pipeline. Trimmed alignments were then further cleaned with ClAlign v. 1.1.0 (Tumescheit et al., 2022) to remove divergent, likely spurious sequences (with option "--remove_divergent_minperc 85" and option "retain-str" to prevent the outgroup to be discarded due to its genuine higher divergence) and then with TAPER v. 1.0 (Zhang

et al., 2021) using default settings to remove small mis-aligned, highly divergent stretches. Final nuclear alignments with a median ungapped sequence length < 300 bp were discarded as they would likely not contain enough phylogenetic signal to inform gene tree inferences, while all plastid alignments were kept as they would be concatenated before inferring a plastome tree. This resulted in a total of 350 (PE approach) or 343 (PI approach) Angiosperms353, one ITS and 107 plastid clean alignments. For each nuclear region alignment, a gene tree was estimated with IQ-TREE v.1.6.12 (Minh et al., 2020) based on the Maximum Likelihood phylogenetic inference method, following the best model of nucleotide substitution identified for the region by IQ-TREE's ModelFinder approach (Kalyaanamoorthy et al., 2017), and performing 1000 ultrafast bootstrap replicates. The nuclear gene trees resulting from the PE approach (except the ITS tree which was kept separate; see Results) were analyzed using Weighted Astral v. 1.16.3.4 (Zhang et al., 2018, 2022) to generate a species tree. The nuclear gene trees resulting from the PI approach (and therefore often comprising multiple potentially paralogous copies for a given sample) were analyzed using DISCO v. 1.4 (Willson et al., 2022) to be decomposed into single copy gene trees, and the latter were then analyzed with Weighted ASTRAL to generate a species tree where each sample was represented only once. All alignments corresponding to plastome regions obtained via the PE approach were concatenated using AMAS (Borowiec, 2016) and a phylogenetic tree was inferred based on the concatenated alignment using IQ-TREE with 1000 bootstrap replicates. The best partition scheme and corresponding best models of nucleotide substitutions were estimated through IQ-TREE's ModelFinder and TESTMERGE option (Chernomor et al., 2016).

The above lab work and analyses were also performed for a few samples of wood, in order to test the use of genomic data for timber species identification. This included 18 wood tissue samples representing four *Entandrophragma*, one *Lovoa*, two *Khaya* and one *Swietenia* species that were obtained from World Forest ID collections. These samples were identified at the species level at the moment of collection by a botanist, based on the morphology of the tree from which they were collected. When possible, heartwood and sapwood were sampled from the same collection. Moreover, four samples of processed heartwood obtained from a company and tentatively identified as *Entandrophragma sp.* were added to the study to explore the performance of the high throughput approaches on heavily transformed wood. DNA was extracted in a similar manner as for the leaf tissue but with further modifications described in Table 2 and in Section 2b of this report. Library pools made from DNA obtained from wood tissue were entirely subjected to target sequence capture as their low DNA content did not allow to keep a fraction for genome skimming. The hybridisation step lasted 27 hours 63°C. The rest of the protocols and analyses was as described above.

Table 1. Samples used to create the genomic reference dataset. For "tissue", L is leaf, HW is heartwood and SW is sapwood (the wood samples were not used as reference but to test if the high throughout sequencing approach could be used to identify wood samples). For "NCBI SRA accession numbers", U indicates that the data was not submitted to Genbank because the identity of the sample was uncertain, and NA indicates that there was no skimming data generated. Data from samples of uncertain identity were not included in the reference dataset, so the corresponding recovery columns also mention "U". In some cases, data were submitted to Genbank but not included in the reference dataset, this is visible in the recovery columns where L or 0 indicate that low data recovery led to the exclusion of the sample from the reference.

			NCBI SRA accession numbers			Cumulated recovery (in bp) for target regions (A353), ITS and plastome regions		
Species (tissue)	**Country**	**Voucher**	**Target capture**	**Skimming**	**Raw Gigabases**	**A353**	**ITS**	**Plastome**
Entandrophragma angolense (L)	Uganda	B. T. Styles 215	SRR31348659	NA	0.01	L	L	L
Entandrophragma angolense (L)	Angola	J. Gossweiler 1919	SRR31348535	SRR31348531	1.82	L	L	L
Entandrophragma angolense (L)	Guinea	Burgt, X. M. van der 2286	SRR31348582	SRR31348536	4.68	749320	690	145474
Entandrophragma angolense (L)	Sierra Leone	D. Small 715	SRR31348462	SRR31348358	1.40	95666	689	146513
Entandrophragma angolense (L)	Cote d'Ivoire	A. J. M. Leeuwenberg 2493	SRR31348568	SRR31348534	5.15	467783	689	146455
Entandrophragma angolense (L)	Ghana	A. E. Kitson 1230	SRR31348557	SRR31348533	4.12	497328	689	146471
Entandrophragma angolense (L)	Cameroon	Etuge, M. 6637	SRR31348580	SRR31348663	9.65	867450	689	146424
Entandrophragma angolense (L)	Gabon	F. J. Breteler 15422	SRR31348449	SRR31348441	7.28	666866	690	146428
Entandrophragma angolense (L)	Sudan	T. F. Chipp 13	SRR31348365	SRR31348436	7.02	567007	689	146371
Entandrophragma angolense (L)	Liberia	Philomena Yarwoah QRRL418	SRR31348367	SRR31348532	5.79	827518	689	146642
Entandrophragma angolense (L)	Liberia	Philomena Yarwoah UWLV282	SRR31348408	SRR31348529	5.33	860791	689	146571
Entandrophragma angolense (L)	Nigeria	Elisha, E. 1035555	SRR31348397	SRR31348418	6.08	832385	689	146563
Entandrophragma angolense (L)	Kenya	R. B. and A. J. Faden 77/899	U	U	3.55	U	U	U
Entandrophragma angolense (L)	Namibia	R. J. Rodin 2631	U	U	4.71	U	U	U
Entandrophragma angolense (SW)	Democratic Republic of the Congo	Augustin Iyokwa QYPX983	SRR31348626	SRR31348414	3.59	802080	689	146505
Entandrophragma angolense (SW)	Democratic Republic of the Congo	Augustin Iyokwa QYPX983	SRR31348363	NA	1.89	439320	689	128185
Entandrophragma bussei (L)	Tanzania	S. Bidgood et al 1156	SRR31348525	SRR31348417	7.20	837459	833	145650
Entandrophragma candollei (L)	Cameroon	B. A. Krukoff 142	SRR31348503	SRR31348415	0.01	L	L	L
Entandrophragma candollei (L)	Republic of the Congo	Liegeois P. 117	SRR31348637	NA	0.49	L	L	L
Entandrophragma candollei (L)	Sierra Leone	X. M. van der Burgt 1649	SRR31348581	SRR31348548	3.81	774836	832	146295
Entandrophragma candollei (L)	Nigeria	J. P. M. Brenan et al. 8438	SRR31348528	SRR31348370	4.15	553087	832	146340
Entandrophragma candollei (L)	Guinea	Haba, P. M. 727	SRR31348514	SRR31348416	4.24	831038	710	145794
Entandrophragma candollei (SW)	Democratic Republic of the Congo	Augustin Iyokwa LAIU104	SRR31348437	SRR31348660	1.88	747544	832	146268
Entandrophragma candollei (SW)	Democratic Republic of the Congo	Augustin Iyokwa GQPU350	SRR31348425	SRR31348569	3.65	821036	832	146134
Entandrophragma candollei (SW)	Democratic Republic of the Congo	Augustin Iyokwa GQPU350	SRR31348362	NA	1.01	367647	832	104918
Entandrophragma candollei (SW)	Cameroon	Zanguim Tchoutezou Guy Herman HMFT729	SRR31348476	NA	1.08	22361	0	453
Entandrophragma candollei (SW)	Democratic Republic of the Congo	Augustin Iyokwa KMOU215	SRR31348603	NA	3.05	396038	666	31731
Entandrophragma candollei (SW)	Democratic Republic of the Congo	Augustin Iyokwa GQPU350	SRR31348546	NA	0.11	0	0	0
Entandrophragma caudatum (L)	South Africa	J. Vahrmeyer + Joynt 177	SRR31348527	SRR31348530	2.75	577842	835	146435
Entandrophragma caudatum (L)	South Africa	F. White 10488	SRR31348610	SRR31348411	3.03	500417	835	145657
Entandrophragma caudatum (L)	South Africa	Mabatha F.W., Nkuna L.A., van Slageren M. 2272	SRR31348608	SRR31348410	5.47	843461	835	146023
Entandrophragma caudatum (L)	Mozambique	B. T. Styles 3784	SRR31348607	SRR31348409	5.82	639989	835	144948
Entandrophragma caudatum (L)	Zambia	F. White 1977	SRR31348606	SRR31348407	1.93	567221	715	146375

Entandrophragma caudatum (L)	Zimbabwe	R. B. Drummond & R. O. B. Rutherford-Smith 7543	SRR31348605	SRR31348406	2.36	615893	835	146172
Entandrophragma caudatum (L)	Malawi	P. Thopham 741	SRR31348604	SRR31348405	2.07	588766	835	146124
Entandrophragma congoense (L)	Cameroon	R. LETOUZEY 14499	SRR31348615	SRR31348413	4.55	668531	835	146448
Entandrophragma congoense (L)	Democratic Republic of the Congo	R. Dechamps 164	SRR31348488	SRR31348412	5.31	613423	835	146509
Entandrophragma cylindricum (HW)	Cameroon	Zanguim Tchoutezou Guy Herman FPBQ537	SRR31348359	NA	0.00	0	0	0
Entandrophragma cylindricum (HW)	Cameroon	Zanguim Tchoutezou Guy Herman EFTT752	SRR31348360	NA	0.00	0	0	0
Entandrophragma cylindricum (L)	Cameroon	B. T. Singles 43	SRR31348601	NA	1.17	L	L	L
Entandrophragma cylindricum (L)	Guinea	Haba, P. M. 728	SRR31348382	SRR31348542	4.25	818451	713	146422
Entandrophragma cylindricum (L)	Cote d'Ivoire	A. J. M. Leeuwenberg 2483	SRR31348658	SRR31348361	4.61	578129	713	145227
Entandrophragma cylindricum (L)	Ghana	B. A. Krukoff 40	SRR31348647	SRR31348539	2.90	801471	713	144721
Entandrophragma cylindricum (L)	Nigeria	B. O. Darmola 457194	SRR31348593	SRR31348666	7.51	315639	713	144679
Entandrophragma cylindricum (L)	Democratic Republic of the Congo	Terese B. Hart 417	SRR31348599	SRR31348404	1.79	546166	713	145666
Entandrophragma cylindricum (L)	Uganda	B. T. Styles 52	SRR31348598	NA	5.77	542193	561	13182
Entandrophragma cylindricum (SW)	Republic of the Congo	Cynel Gwenael Moundounga XRJH521	SRR31348602	NA	0.00	5246	0	0
Entandrophragma cylindricum (SW)	Cameroon	Zanguim Tchoutezou Guy Herman FPBQ537	SRR31348366	NA	0.01	14449	593	7091
Entandrophragma delevoyi (L)	Tanzania	A. A. Bullock 2071	SRR31348597	SRR31348403	6.61	592843	829	146425
Entandrophragma delevoyi (L)	Zambia	C.E. Duff 150/33	SRR31348596	SRR31348402	4.03	599455	829	146544
Entandrophragma delevoyi (L)	Zambia	W. L. Astle 865	SRR31348595	SRR31348401	3.73	614347	829	146513
Entandrophragma excelsum (L)	Republic of the Congo	Herbier M. Reynders 208	SRR31348387	SRR31348400	4.33	633548	834	146608
Entandrophragma excelsum (L)	Uganda	B. T. Styles 319	SRR31348386	SRR31348399	4.88	462068	834	146399
Entandrophragma excelsum (L)	Tanzania	J. M. Grimshaw 9373	SRR31348385	SRR31348398	5.35	609817	834	145839
Entandrophragma excelsum (L)	Malawi	T. Muller 1608	SRR31348384	SRR31348396	3.16	688689	834	145834
Entandrophragma excelsum (L)	Zambia	D.B.F. F5126	SRR31348383	SRR31348395	1.23	256738	834	144567
Entandrophragma palustre (L)	Republic of the Congo	Germain 8395	SRR31348381	NA	0.05	65490	0	0
Entandrophragma spicatum (L)	Angola	F. Crawford 456	SRR31348380	SRR31348394	6.71	870362	835	146459
Entandrophragma spicatum (L)	Namibia	D.A.H. Taylor 296	SRR31348379	SRR31348393	4.56	766934	835	146320
Entandrophragma utile (L)	Tanzania	R. I. Ludanga 1997	SRR31348373	NA	0.00	L	L	L
Entandrophragma utile (L)	Cote d'Ivoire	A. J. M. Leeuwenberg 2510	SRR31348477	SRR31348545	3.70	512484	680	146270
Entandrophragma utile (L)	Nigeria	M. G. Latilo 32980	SRR31348600	SRR31348364	3.82	683418	680	146374
Entandrophragma utile (L)	Cameroon	F. J. Breteler 2165	SRR31348377	SRR31348392	3.89	669179	680	146217
Entandrophragma utile (L)	Gabon	B. A. Krukoff 139	SRR31348376	SRR31348391	4.58	575370	679	146132
Entandrophragma utile (L)	Republic of the Congo	J. Wagemans 1497	SRR31348375	SRR31348390	4.71	540530	680	146431
Entandrophragma utile (L)	Uganda	B. T. Styles 109	SRR31348374	SRR31348389	4.70	378763	681	145595
Entandrophragma utile (SW)	Cameroon	Zanguim Tchoutezou Guy Herman ABPY935	SRR31348378	NA	0.97	70610	0	4447
Entandrophragma utile (SW)	Cameroon	Zanguim Tchoutezou Guy Herman ABPY935	SRR31348541	NA	6.08	454431	681	144075
Khaya agboensis (L)	Sierra Leone	J.S. Sawyerr F. H. K. 13601	SRR31348655	SRR31348523	3.51	506524	840	145725
Khaya agboensis (L)	Ghana	G.W.A 585	SRR31348654	SRR31348522	8.39	706872	846	146610
Khaya agboensis (L)	Cote d'Ivoire	J.J.F.E. de Wilde 3744	SRR31348653	SRR31348521	3.21	481042	824	146746

Khaya agboensis (L)	Nigeria	J.O. Amachi 38275	SRR31348652	NA	1.67	116177	0	0
Khaya agboensis (L)	Guinea	Haba P. M. 892	SRR31348651	SRR31348520	3.60	828444	846	146497
Khaya agboensis (L)	Guinea-Bissau	Douglas Latham n/a	U	U	2.76	U	U	U
Khaya euryphylla (L)	Gabon	Thomson 1	SRR31348649	NA	0.01	60297	0	0
Khaya euryphylla (L)	Cameroon	R.W.J. Keay 37426	U	U	7.76	U	U	U
Khaya grandifoliola (L)	Cameroon	M.G. Latilo & B.O. Daramol 34481	SRR31348648	SRR31348518	6.50	812986	846	146699
Khaya grandifoliola (L)	Republic of the Congo	R. Germain 4166	SRR31348646	NA	3.00	194675	0	831
Khaya grandifoliola (L)	Sudan	L. Turner 184	SRR31348645	SRR31348517	5.74	724407	846	145800
Khaya grandifoliola (L)	Uganda	B.T. Styles 263	SRR31348589	SRR31348642	5.58	669543	846	146488
Khaya grandifoliola (L)	Brazil	D.A. Folli 6785	SRR31348643	SRR31348516	7.95	678891	846	146625
Khaya grandifoliola (L)	Egypt	Min of Agrenltine n/a	SRR31348468	SRR31348515	5.11	797408	846	146636
Khaya grandifoliola (L)	Ghana	G. Vigne 1803	SRR31348467	SRR31348513	0.83	368989	748	136831
Khaya grandifoliola (L)	Cote d'Ivoire	Aubreville 63	SRR31348466	SRR31348512	5.37	737759	824	146845
Khaya grandifoliola (L)	Guinea	Aug. Chevalier 20687	SRR31348465	SRR31348511	5.02	109493	824	70633
Khaya grandifoliola (L)	Benin	H. Ern 3172	SRR31348464	SRR31348510	3.66	704543	824	146822
Khaya grandifoliola (L)	Central African Republic	John M. Fay 4164	SRR31348463	SRR31348509	4.44	763089	846	146826
Khaya grandifoliola (L)	Nigeria	B. O. Darmola 178	U	U	4.33	U	U	U
Khaya ivorensis (HW)	Cameroon	Zanguim Tchoutezou Guy Herman IMGH503	SRR31348664	NA	0.00	2004	0	6953
Khaya ivorensis (L)	Ghana	J. F. Chipp 37	SRR31348454	NA	1.21	L	L	L
Khaya ivorensis (L)	Republic of the Congo	Bouka Gaël FHEL563	SRR31348461	SRR31348508	4.39	877135	845	146165
Khaya ivorensis (L)	Gabon	B.A. Krufoff 159	SRR31348460	NA	2.42	298130	337	4419
Khaya ivorensis (L)	Republic of the Congo	L. Toussaint 680	SRR31348459	SRR31348507	1.24	414242	846	125420
Khaya ivorensis (L)	Trinidad and Tobago	F.C. Butlin 11972	SRR31348458	SRR31348506	3.32	492092	846	146303
Khaya ivorensis (L)	Brazil	G.S. Siqueira & G. Terra 14357	SRR31348457	SRR31348505	7.20	718133	846	146644
Khaya ivorensis (L)	Guinea-Bissau	M.T. Dawe 237	SRR31348456	SRR31348504	1.66	580129	846	146703
Khaya ivorensis (L)	Cote d'Ivoire	B. A. Krukoff 70	SRR31348455	SRR31348502	5.05	671634	846	145560
Khaya ivorensis (L)	Nigeria	J.P.M. Brenan, C.F.A. Onochie, E.W. Jones & P.W. Richards s.n.	SRR31348453	SRR31348501	3.20	789146	846	146153
Khaya ivorensis (L)	Equatorial Guinea	M.F. Carvalho 2829	SRR31348594	SRR31348500	3.66	536628	846	145642
Khaya ivorensis (L)	Cameroon	Etuge, M. 1486	U	U	4.08	U	U	U
Khaya ivorensis (SW)	Republic of the Congo	Bouka Gaël FHEL563	SRR31348540	NA	1.56	415450	845	33903
Khaya nyasica (L)	Uganda	K. Dawn 641	SRR31348590	NA	0.66	L	L	L
Khaya nyasica (L)	Zambia	W. L. Astle 999	SRR31348585	NA	0.26	L	L	L
Khaya nyasica (L)	Democratic Republic of the Congo	J.J. Symoens 8803	SRR31348592	SRR31348499	4.04	608467	845	146718
Khaya nyasica (L)	Tanzania	David A.H. Taylor 257	SRR31348591	SRR31348498	4.60	776564	847	145942
Khaya nyasica (L)	Mozambique	A. Gomes e Sousa 4326	SRR31348588	SRR31348641	4.49	722770	847	146438
Khaya nyasica (L)	Malawi	J.D. & E.G. Chapman 8050	SRR31348587	SRR31348640	6.52	744129	846	146989
Khaya nyasica (L)	Zimbabwe	D. F. Lovemore 327	SRR31348586	SRR31348639	2.40	611248	847	146771
Khaya nyasica (L)	Kenya	S. Mills 3044	U	U	1.37	U	U	U
Khaya senegalensis (HW)	Ghana	Emmanuel Ebanyenle CTCC962	SRR31348665	NA	0.79	61770	847	1236
Khaya senegalensis (HW)	Ghana	Emmanuel Ebanyenle NEMN471	SRR31348357	NA	2.31	356374	847	18010

Khaya senegalensis (HW)	Ghana	Emmanuel Ebanyenle NEMN471	SRR31348544	NA	0.13	156	0	0
Khaya senegalensis (L)	Uganda	B.T. Styles 255	SRR31348644	NA	0.15	61218	0	0
Khaya senegalensis (L)	Cameroon	M.G. Latilo & B.O. Daramola 34438	SRR31348584	SRR31348638	6.64	648838	847	146823
Khaya senegalensis (L)	Democratic Republic of the Congo	Mil 9717	SRR31348583	SRR31348636	8.08	779081	847	146695
Khaya senegalensis (L)	Central African Republic	J. Michael Fay & Joel Doka 5269	SRR31348579	SRR31348635	10.72	815144	847	146777
Khaya senegalensis (L)	Sudan	L. Turner 184	SRR31348578	SRR31348634	10.45	801925	848	145943
Khaya senegalensis (L)	French Guiana	Oldeman 1258	SRR31348577	SRR31348633	6.05	486757	847	146533
Khaya senegalensis (L)	Brazil	G.S. Siqueira & G. Terra 14358	SRR31348576	SRR31348632	4.71	840359	847	146338
Khaya senegalensis (L)	Senegal	NA NA	SRR31348575	SRR31348631	3.87	452630	847	145842
Khaya senegalensis (L)	Ghana	S. Kitson 702	SRR31348574	SRR31348630	3.33	631665	847	146881
Khaya senegalensis (L)	Nigeria	J.Lowe 2546	SRR31348573	NA	0.39	78986	0	1352
Khaya senegalensis (L)	Senegal	J. G. Adam 17935	SRR31348572	SRR31348629	3.84	556332	847	146896
Khaya senegalensis (L)	Gambia	J.P.Ruxton 162	SRR31348571	SRR31348628	4.63	551684	727	146743
Khaya senegalensis (L)	Mali	Bamps 2478	SRR31348570	SRR31348627	5.67	714119	847	146211
Khaya senegalensis (L)	Togo	B.T. Styles 2085	SRR31348567	SRR31348625	4.80	688468	847	146634
Khaya senegalensis (L)	Guinea	J.M. Daziel n/a	SRR31348566	SRR31348624	7.86	602559	847	146399
Khaya senegalensis (L)	Guinea-Bissau	NA 1946	SRR31348565	SRR31348623	3.05	530540	847	145813
Khaya senegalensis (L)	Sierra Leone	F.C. Deighton 5894	SRR31348564	SRR31348622	1.48	429983	847	146298
Khaya senegalensis (L)	Burkina Faso	A.J.M. Leeuwenberg 4318	SRR31348563	SRR31348621	4.87	577526	847	146770
Khaya senegalensis (L)	Ghana	N.C. McLeod 827	U	U	11.63	U	U	U
Khaya senegalensis (SW)	Ghana	Emmanuel Ebanyenle NEMN471	SRR31348368	NA	1.85	438842	847	98780
Khaya sp. cf. anthotheca (L)	Uganda	B.T. Styles 125	SRR31348372	SRR31348388	7.54	U	U	U
Khaya sp. cf. anthotheca (L)	Kenya	S.R. Semsei 51076	SRR31348657	SRR31348526	5.73	U	U	U
Khaya sp. cf. anthotheca (L)	Tanzania	B.T. Styles 261	SRR31348656	SRR31348524	6.30	U	U	U
Khaya sp. cf. anthotheca (L)	Democratic Republic of the Congo	A. Leonard 3242	SRR31348650	SRR31348519	3.82	U	U	U
Lovoa swynnertonii (L)	Uganda	B.T. Styles 245a	SRR31348562	SRR31348620	2.72	482495	837	140041
Lovoa swynnertonii (L)	Kenya	Nzano 14716	SRR31348561	SRR31348619	5.34	535786	839	143960
Lovoa swynnertonii (L)	Zimbabwe	H. Wild 2228	SRR31348560	SRR31348618	4.45	421206	824	144605
Lovoa swynnertonii (L)	Tanzania	Frontier-Tanzania 158	U	U	6.55	U	U	U
Lovoa trichilioides (L)	Liberia	W.J. Harley 1052	SRR31348559	NA	0.38	L	L	L
Lovoa trichilioides (L)	Ghana	J. F. Chipp 39	SRR31348556	NA	0.02	L	L	L
Lovoa trichilioides (L)	Cote d'Ivoire	Aubreville 210	SRR31348555	NA	0.38	L	L	L
Lovoa trichilioides (L)	Sierra Leone	Luke, W.R.Q. 15234	SRR31348558	SRR31348617	7.16	485727	766	145805
Lovoa trichilioides (L)	Cameroon	M. Cheek 11533	SRR31348554	SRR31348616	5.24	810899	766	145863
Lovoa trichilioides (L)	Zambia	J. Timberlake, M. Bingham & A. Cunningham 5825	SRR31348553	SRR31348613	5.31	781496	766	145489
Lovoa trichilioides (L)	Uganda	K. A. Lye 6240	SRR31348552	SRR31348497	8.72	532559	765	145569
Lovoa trichilioides (L)	Gabon	L.J.G. van der Maesen 5806	U	U	2.86	U	U	U
Lovoa trichilioides (L)	Tanzania	L.T. Wigg 318	U	U	0.00	U	U	U
Lovoa trichilioides (SW)	Democratic Republic of the Congo	Augustin Iyokwa SUBQ641	SRR31348662	NA	1.14	352481	765	29070
Lovoa trichilioides (SW)	Cameroon	TASSIAMBA NANFACK Stève QZRW689	SRR31348549	NA	1.45	317048	766	26869

Lovoa trichilioides (SW)	Cameroon	TASSIAMBA NANFACK Stève QZRW689	SRR31348369	NA	2.30	377052	730	49535
Swietenia humilis (L)	Mexico	B.T. Styles 113	SRR31348551	SRR31348496	5.59	664196	848	146676
Swietenia humilis (L)	Mexico	T.D. Pennington & J. Sarukhan K 9219	SRR31348550	SRR31348495	3.43	815334	847	146762
Swietenia humilis (L)	Mexico	T.D. Pennington & J. Sarukhan 9456	SRR31348452	SRR31348494	6.41	783853	849	146681
Swietenia humilis (L)	Honduras	D.H. Boshier 49	SRR31348451	SRR31348493	4.70	791178	849	146671
Swietenia humilis (L)	El Salvador	E. Mayorga 3869	SRR31348450	SRR31348492	5.35	438056	730	145747
Swietenia humilis (L)	Guatemala	J.J. Castillo M. 1987	SRR31348448	SRR31348491	5.36	869733	729	146666
Swietenia humilis (L)	Nicaragua	P. P. Moreno 23530	SRR31348447	SRR31348490	7.70	600223	849	146813
Swietenia humilis (L)	Costa Rica	D.H. Boshier 21	SRR31348446	SRR31348489	4.39	874766	729	146746
Swietenia humilis (L)	Mexico	Guillermo Ibarra Manriquez 5447	U	U	7.00	U	U	U
Swietenia macrophylla (HW)	Mexico	Fabiola Lopez YGPL521	SRR31348667	NA	0.62	251634	729	15632
Swietenia macrophylla (HW)	Mexico	Fabiola Lopez TZER575	SRR31348538	NA	0.40	37804	642	7505
Swietenia macrophylla (HW)	Mexico	Fabiola Lopez YGPL521	SRR31348537	NA	0.11	0	0	0
Swietenia macrophylla (L)	El Salvador	J.M. Rosales 2127	SRR31348438	NA	0.01	L	L	L
Swietenia macrophylla (L)	Panama	R. J. Schmalzel 429	SRR31348430	NA	0.15	L	L	L
Swietenia macrophylla (L)	Mexico	Fabiola Lopez KXYH120	SRR31348445	SRR31348487	7.78	731588	849	146784
Swietenia macrophylla (L)	Mexico	R. Alvarez 237	SRR31348444	NA	5.91	635906	679	12975
Swietenia macrophylla (L)	Mexico	T.D. Pennington & J. Sarukhan K. 9629	SRR31348443	SRR31348486	10.57	633149	729	146798
Swietenia macrophylla (L)	Mexico	T.D. Pennington & J. Sarukhan K. 9411	SRR31348442	SRR31348485	2.23	604492	849	146400
Swietenia macrophylla (L)	Guatemala	J.W. Stead 182	SRR31348440	SRR31348484	8.00	744293	848	146379
Swietenia macrophylla (L)	Belize	T. Sarkinen 804	SRR31348439	SRR31348483	5.31	873542	849	146792
Swietenia macrophylla (L)	Nicaragua	P.P. Moreno 24029	SRR31348435	SRR31348482	6.82	388081	729	138666
Swietenia macrophylla (L)	Honduras	T.D. Pennington, P. E. Owen & R. Zuniga 13674	SRR31348434	SRR31348481	7.20	549724	729	145471
Swietenia macrophylla (L)	Mexico	E. Martinez S. & C. H. Ramos 26392	SRR31348433	SRR31348480	4.44	752375	729	146694
Swietenia macrophylla (L)	Trinidad and Tobago	NA T. 7010	SRR31348432	SRR31348479	4.79	577195	729	146278
Swietenia macrophylla (L)	Brazil	B. Dubs 1719	SRR31348431	SRR31348478	5.91	842145	729	146558
Swietenia macrophylla (L)	Venezuela	T. Carrillo CH. 29	SRR31348429	SRR31348475	1.67	332212	849	146777
Swietenia macrophylla (L)	Ecuador	J. Zuleta 9	SRR31348428	SRR31348474	4.60	523777	729	145867
Swietenia macrophylla (L)	Peru	T.D. Pennington, A. Daza, J. Revilla 17369	SRR31348427	SRR31348473	11.15	615745	729	12507
Swietenia macrophylla (L)	Bolivia	Israel G., Vargas C. & Claudia Jordan 6278	SRR31348426	SRR31348472	3.86	513686	721	145036
Swietenia macrophylla (L)	Costa Rica	J.F. Morales 5310	U	U	10.59	U	U	U
Swietenia macrophylla (SW)	Mexico	Fabiola Lopez TZER575	SRR31348547	NA	0.06	0	0	102
Swietenia macrophylla (SW)	Mexico	Fabiola Lopez YGPL521	SRR31348543	NA	10.95	576651	850	146715
Swietenia macrophylla (SW)	Mexico	Fabiola Lopez KXYH120	SRR31348661	NA	0.05	0	0	0
Swietenia mahagoni (L)	USA	C. L. Lundell & Amelia A. Lundell 17549	SRR31348424	SRR31348471	3.33	596474	723	146229
Swietenia mahagoni (L)	Jamaica	W.M. Harris 10821	SRR31348423	SRR31348470	10.46	599290	723	146223
Swietenia mahagoni (L)	Dominican Republic	Encarnacion, W. NA	SRR31348422	SRR31348469	3.44	849860	723	146167
Swietenia mahagoni (L)	Turks and Caicos Islands	E. Freid 08-166	SRR31348421	SRR31348614	4.68	708353	723	146176
Swietenia mahagoni (L)	Bahamas	P. Wilson 7434	SRR31348420	SRR31348612	8.68	671047	723	146596
Swietenia mahagoni (L)	Barbados	A. G. Large 1942-27	SRR31348419	SRR31348611	10.29	707514	842	146199

Swietenia mahagoni (L)	Grenada	Hawthorn W., Cable S., & Wise R. 458	SRR31348371	SRR31348609	4.69	674323	842	146639
Swietenia mahagoni (L)	Cuba	P. Acevedo-Rdgz, B. Buck, S. Hundorf & L. Montes 5633	U	U	4.63	U	U	U
Unknown (HW, glued)	NA	NA ASV2	U	U	0.39	0	0	0
Unknown (HW, glued)	NA	NA VWBO-1	U	U	0.09	0	0	0
Unknown (HW, glued)	NA	NA VWMO-1	U	U	0.10	0	0	121
Unknown (HW, glued)	NA	NA VWBO-3	U	U	0.51	0	0	0

b. Evaluation of the reference DNA dataset

As a result of the labwork and analyses described above, the final reference dataset for the four focus genera comprised up to 351 nuclear and 107 plastid regions for 132 samples representing 22 species (see Table above). In addition, data for 4 samples potentially representing *Khaya anthotheca* or close relatives were also made available but not used here. All raw sequences are available in NCBI under project number PRJNA1185931 (see Table 1 for individual accession numbers). In addition, for each gene, all the reference sequences are available as raw and clean sequence alignments at https://github.com/sidonieB/Bellot_al_Meliaceae_DNA_barcoding/tree/main.

When using all genes together, it is possible to infer the phylogenetic relationships of all samples of reference with a high level of resolution, as shown in Figure 1.1 (further discussed in Article 1).

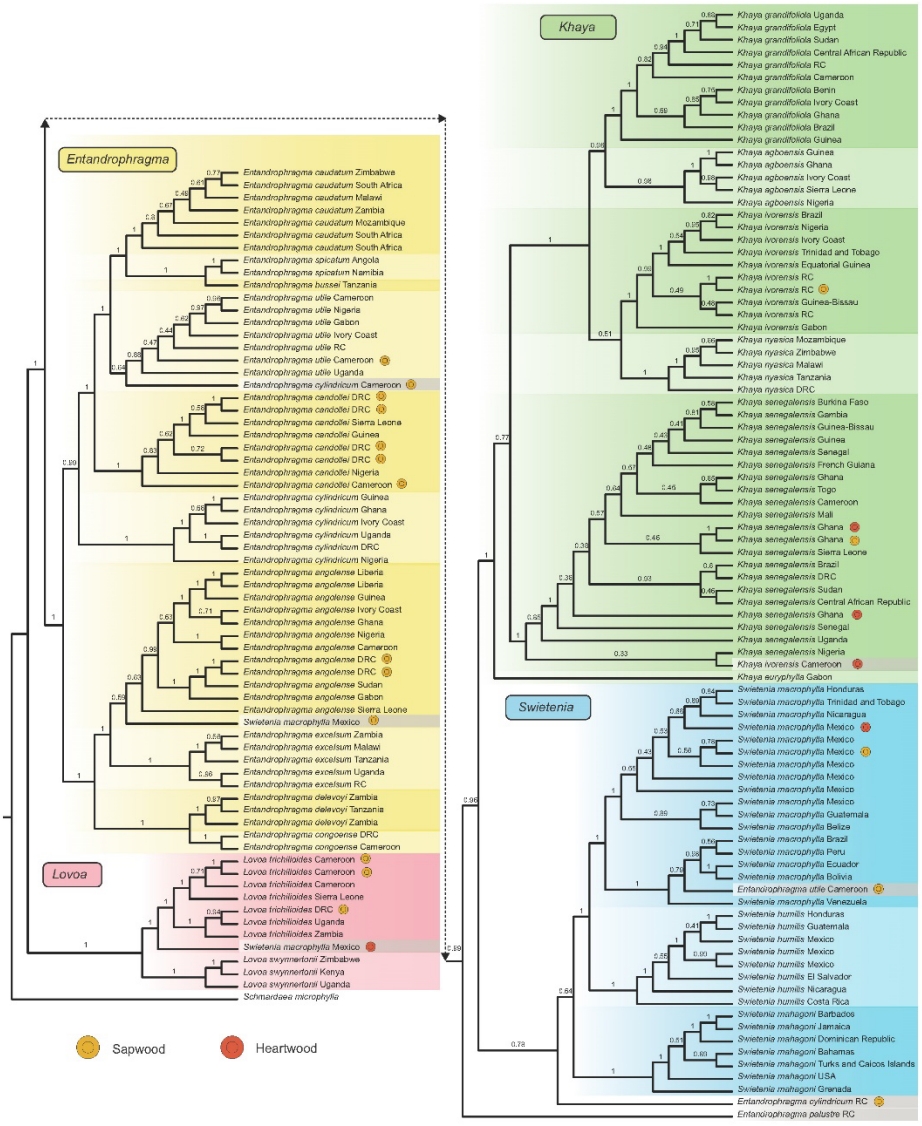


Figure 1.1. Intra and inter-specific relationships in four Meliaceae genera based on the paralog-inclusive analysis of 343 nuclear genes. Numbers on branches indicate local posterior probabilities. Circles indicate sapwood (yellow) and heartwood (brown) samples. Colors and shades delimit genera and species respectively, while grey highlights samples falling away from their clade.

On Figure 1.1, it is also visible that genomic data could be obtained for wood samples and used to identify them at the species level. Nuclear regions obtained from target capture sequencing were only slightly better at placing wood samples accurately compared to plastome or ITS data obtained through genome skimming (Figure 1.2). This was most likely due to the lower recovery for the plastome and smaller size of the ITS region (Table 1).

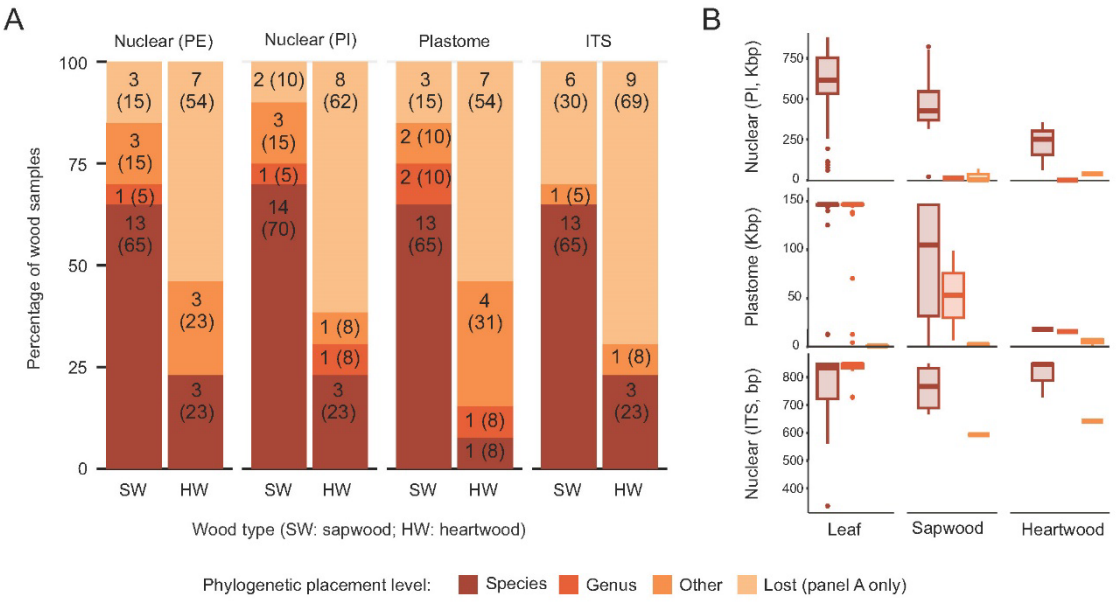


Figure 1.2. Identification of wood samples. A. Placement accuracy of the wood samples depending on the genetic regions analyzed. Numbers outside brackets are numbers of samples while numbers inside brackets are corresponding percentages. **B.** Placement accuracy of all samples depending on data recovery. Recovery is measured as the cumulated length of the DNA regions recovered in Kbp or bp. PE: paralog-exclusive, PI: paralog-inclusive. Placement categories indicate if, based on its most closely related non-wood (reference) sample, the sample was recovered in the right species (Species), the right genus (Genus), neither (Other) or if it was excluded from the phylogenetic tree due to insufficient data recovery (Lost).

Sapwood samples were less frequently lost during the data cleaning workflow, and more frequently accurately identified to species level than heartwood samples, most likely because they had better region recovery. In general, the accuracy of placement in a given tree only depended on data recovery to the extent that samples with extremely low recovery were consistently recovered outside of their species or even genus (Figure 1.3). This was mostly visible when looking at placement accuracy across wood samples: those placed at the species level had higher recovery of the considered region compared to those placed at genus level or elsewhere. Although some heartwood samples were correctly identified, it was more difficult to obtain sufficient data from these samples than for sapwood samples. For the plastome and ITS regions, we additionally observed genus-dependent performance of identification for samples with medium to high recovery, as some *Khaya* and *Swietenia* samples with good recovery could not be placed at the species level whereas this was less frequently the case for *Entandrophragma* and never the case for *Lovoa*.

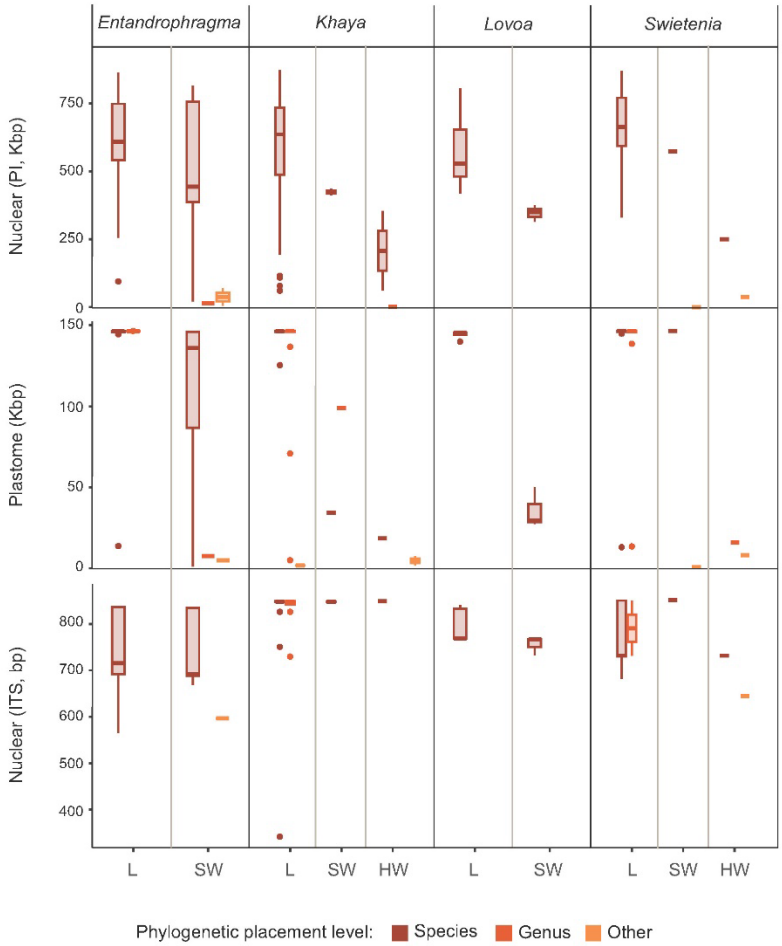


Figure 1.3. Phylogenetic placement accuracy depending on region recovery and genus. Placement categories indicate if the sample was recovered in the right species (Species), the right genus (Genus), or neither (Other). For leaf samples (L), placement is assessed based on the clade in which the sample falls, not considering wood samples. For wood samples (HW: heartwood and SW: sapwood), placement is assessed based on the most closely related leaf (reference) sample, under the rationale that this would be sufficient in a context of DNA barcoding where the whole tree does not need to be perfectly resolved to allow sample identification.

2. Timber DNA extraction protocol

a. DNA extraction protocol optimisation

As a first step to optimise a DNA extraction protocol that may work for timber, we reviewed ca. 36 research publications discussing different DNA extraction protocols for wood and/or DNA of poor quality (i.e. low concentration and/or size and/or presence of metabolites interfering with DNA amplification and sequencing)**.** In addition, we held meetings with several international experts to discuss the application of published DNA extraction protocols. Meetings were held with Steven Janssens (Meise Botanic Garden, Belgium), Samuel Vanden Abeele (University of Cambridge, United Kingdom), Céline Blanc Jolivet (Thünen Centre of Competence, Germany) and R. Odgen (University of Edinburgh). Many of the publications had also been surveyed by Jiao et al. 2020. For each publication, we recorded the species name, type of sample used, and DNA extraction protocol applied, as well as details on whether the study authors were successful at extracting DNA from heartwood. However, many of the publications did not specify the wood type (sapwood or heartwood) used, nor clearly described the effectiveness of the extraction protocol.

Main publications reviewed:

Asif, M.J. and Cannon, C.H., 2005. DNA extraction from processed wood: a case study for the identification of an endangered timber species (Gonystylus bancanus). Plant Molecular Biology Reporter, 23, pp.185-192.

Dormontt, E.E., Jardine, D.I., van Dijk, K.J., Dunker, B.F., Dixon, R.R.M., Hipkins, V.D., Tobe, S., Linacre, A. and Lowe, A.J., 2020. Forensic validation of a SNP and INDEL panel for individualisation of timber from bigleaf maple (Acer macrophyllum Pursch). Forensic Science International: Genetics, 46, p.102252.

Degen, B., Ward, S.E., Lemes, M.R., Navarro, C., Cavers, S. and Sebbenn, A.M., 2013. Verifying the geographic origin of mahogany (Swietenia macrophylla King) with DNA-fingerprints. Forensic Science International: Genetics, 7(1), pp.55-62.

Dev, S.A., Muralidharan, E.M., Sujanapal, P. and Balasundaran, M., 2014. Identification of market adulterants in East Indian sandalwood using DNA barcoding. Annals of forest science, 71(4), pp.517-522.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.

Fatima, T., Srivastava, A., Somashekar, P.V., Hanur, V.S. and Rao, M.S., 2019. Development of DNA-based species identification and barcoding of three important timbers. Bulletin of the National Research Centre, 43(1), pp.1-17.

Fatima, T., Srivastava, A., Hanur, V.S. and Rao, M.S., 2018. An effective wood DNA extraction protocol for three economic important timber species of India. American Journal of Plant Sciences, 9(02), p.139.

Ganopoulos, I., Aravanopoulos, F., Madesis, P., Pasentsis, K., Bosmali, I., Ouzounis, C. and Tsaftaris, A., 2013. Taxonomic identification of Mediterranean pines and their hybrids based on the high resolution melting (HRM) and trnL approaches: from cytoplasmic inheritance to timber tracing. PLoS One, 8(4), p.e60945.

Hanssen, F., Wischnewski, N., Moreth, U. and Magel, E.A., 2011. Molecular identification of Fitzroya cupressoides, Sequoia sempervirens, and Thuja plicata wood using taxon-specific rDNA-ITS primers. IAWA journal, 32(2), pp.273-284.

Höltken, M., A., Schröder, H., Wischnewski, N., Degen, B., Magel, E. and Fladung, M., 2012. Development of DNA-based methods to identify CITES-protected timber species: a case study in the Meliaceae family.

Jiao, L., Yin, Y., Cheng, Y. and Jiang, X., 2014. DNA barcoding for identification of the endangered species Aquilaria sinensis: comparison of data from heated or aged wood samples. Holzforschung, 68(4), pp.487-494.

Jiao, L., Liu, X., Jiang, X. and Yin, Y., 2015. Extraction and amplification of DNA from aged and archaeological Populus euphratica wood for species identification. Holzforschung, 69(8), pp.925-931.

Jiao, L., Yu, M., Wiedenhoeft, A.C., He, T., Li, J., Liu, B., Jiang, X. and Yin, Y., 2018. DNA barcode authentication and library development for the wood of six commercial Pterocarpus species: the critical role of Xylarium specimens. Scientific Reports, 8(1), p.1945.

Jiao, L., Lu, Y., He, T., Li, J. and Yin, Y., 2019. A strategy for developing high-resolution DNA barcodes for species discrimination of wood specimens using the complete chloroplast genome of three Pterocarpus species. Planta, 250, pp.95-104.

Jiao, L., Lu, Y., He, T., Guo, J. and Yin, Y., 2020. DNA barcoding for wood identification: Global review of the last decade and future perspective. IAWA Journal, 41(4), pp.620-643.

Jolivet, C. and Degen, B., 2012. Use of DNA fingerprints to control the origin of sapelli timber (Entandrophragma cylindricum) at the forest concession level in Cameroon. Forensic Science International: Genetics, 6(4), pp.487-493.

Kannangara, S., Karunarathne, S., Ranaweera, L., Ananda, K., Ranathunga, D., Jayarathne, H., Weebadde, C. and Sooriyapathirana, S., 2020. Assessment of the applicability of wood anatomy and DNA barcoding to detect the timber adulterations in Sri Lanka. Scientific Reports, 10(1), p.4352.

Kistler, L., 2012. Ancient DNA extraction from plants. Ancient DNA: Methods and Protocols, pp.71-79.

Lee, S.Y., Ng, W.L., Mahat, M.N., Nazre, M. and Mohamed, R., 2016. DNA barcoding of the endangered Aquilaria (Thymelaeaceae) and its application in species authentication of agarwood products traded in the market. PloS one, 11(4), p.e0154631.

Lendvay, B., Hartmann, M., Brodbeck, S., Nievergelt, D., Reinig, F., Zoller, S., Parducci, L., Gugerli, F., Büntgen, U. and Sperisen, C., 2018. Improved recovery of ancient DNA from subfossil wood–application to the world's oldest Late Glacial pine forest. New Phytologist, 217(4), pp.1737-1748.

Lowe, A.J., Wong, K.N., Tiong, Y.S., Iyerh, S. and Chew, F.T., 2010. A DNA method to verify the integrity of timber supply chains; confirming the legal sourcing of merbau timber from logging concession to sawmill. Silvae Genetica, 59(1-6), pp.263-268.

Lu, Y., Jiao, L., He, T., Zhang, Y., Jiang, X. and Yin, Y., 2020. An optimized DNA extraction protocol for wood DNA barcoding of Pterocarpus erinaceus. IAWA Journal, 41(4), pp.644-659.
Ng, K.K.S., Lee, S.L., Tnah, L.H., Nurul-Farhanah, Z., Ng, C.H., Lee, C.T., Tani, N., Diway, B., Lai, P.S. and Khoo, E., 2016. Forensic timber identification: a case study of a CITES listed species, Gonystylus bancanus (Thymelaeaceae). Forensic Science International: Genetics, 23, pp.197-209.
Ng, C.H., Ng, K.K.S., Lee, S.L., Tnah, L.H., Lee, C.T. and Zakaria, N.F., 2020. A geographical traceability system for Merbau (Intsia palembanica Miq.), an important timber species from peninsular Malaysia. Forensic Science International: Genetics, 44, p.102188.
Nuroniah, H.S., Gailing, O. and Finkeldey, R., 2017. Development of a diagnostic DNA marker for the geographic origin of Shorea leprosula. Holzforschung, 71(1), pp.1-10.
Phong, D.T., Van Tang, D., Hien, V.T.T., Ton, N.D. and Van Hai, N., 2014. Nucleotide diversity of a nuclear and four chloroplast DNA regions in rare tropical wood species of Dalbergia in Vietnam: a DNA barcode identifying utility. Asian Journal of Applied Sciences, 2(2).
Rachmayanti, Y., Leinemann, L., Gailing, O. and Finkeldey, R., 2009. DNA from processed and unprocessed wood: factors influencing the isolation success. Forensic Science International: Genetics, 3(3), pp.185-192.
Rachmayanti, Y., Leinemann, L., Gailing, O. and Finkeldey, R., 2006. Extraction, amplification and characterization of wood DNA from Dipterocarpaceae. Plant Molecular Biology Reporter, 24, pp.45-55.
Rossi, F., Crnjar, A., Comitani, F., Feliciano, R., Jahn, L., Malim, G., Southgate, L., Kay, E., Oakey, R., Buggs, R. and Moir, A., 2021. Extraction and high-throughput sequencing of oak heartwood DNA: Assessing the feasibility of genome-wide DNA methylation profiling. Plos one, 16(11), p.e0254971.
Schroeder, H., Cronn, R., Yanbaev, Y., Jennings, T., Mader, M., Degen, B. and Kersten, B., 2016. Development of molecular markers for determining continental origin of wood from white oaks (Quercus L. sect. Quercus). PloS one, 11(6), p.e0158221.
Tang, X., Zhao, G. and Ping, L., 2011. Wood identification with PCR targeting noncoding chloroplast DNA. Plant molecular biology, 77, pp.609-617.
Tereba, A., Woodward, S., Konecka, A., Borys, M. and Nowakowska, J.A., 2017. Analysis of DNA profiles of ash (Fraxinus excelsior L.) to provide evidence of illegal logging. Wood science and technology, 51, pp.1377-1387.
Tnah, L.H., Lee, S.L., Ng, K.K.S., Faridah, Q.Z. and Faridah-Hanum, I., 2010. Forensic DNA profiling of tropical timber species in Peninsular Malaysia. Forest ecology and management, 259(8), pp.1436-1446.
Tanaka, S. and Ito, M., 2020. Species identification of Indonesian agarwood using a DNA-barcoding method. Journal of natural medicines, 74, pp.323-330.
Tsumura, Y., Kado, T., Yoshida, K., Abe, H., Ohtani, M., Taguchi, Y., Fukue, Y., Tani, N., Ueno, S., Yoshimura, K. and Kamiya, K., 2011. Molecular database for classifying Shorea species (Dipterocarpaceae) and techniques for checking the legitimacy of timber and wood products. Journal of plant research, 124, pp.35-48.
Watanabe, U. and Abe, H., 2017. Sequencing and quantifying plastid DNA fragments stored in sapwood and heartwood of Torreya nucifera. Journal of Wood Science, 63(3), pp.201-208.
Yu, M., Jiao, L., Guo, J., Wiedenhoeft, A.C., He, T., Jiang, X. and Yin, Y., 2017. DNA barcoding of vouchered xylarium wood specimens of nine endangered Dalbergia species. Planta, 246, pp.1165-1176.

Following this preliminary research, we tested the effect of various parameters on the quality of the DNA extracted from wood samples by performing a total of more than 360 extractions. These were performed on 8 heartwood samples representing 5 species of *Khaya*, *Swietenia* and *Entandrophragma*, 5 processed (glued) heartwood samples corresponding to unknown species, and 11 sapwood samples representing 8 species of the four focus genera. Multiple extraction attempts were performed per sample, often with more than one extraction being performed per attempt, and extracts obtained during the same attempt were then pooled (see Table below).

Two first tests were informal tests that did not include replicates and mainly consisted in estimating a quantity for parameters originally used without indication of quantities. Three other tests were then performed formally; they were all conducted together for a total of 48 combinations of the three parameters and their 2 or 3 tested values, performed on 4 different wood samples, with each combination being repeated 4 times per sample (3 times including wood and once without wood to serve as negative control). Negative controls were not included in the result plots. Results of these tests are provided below.

Test 1 (informal): Freezing the wood chips before grinding them

Test: Wood chips from one sample were put in the grinding mill jar and submitted to no freezing or to freezing during 15 min, 30 min, 1h or overnight before being grinded.
Results: There was no observable difference in texture or required grinding time
Conclusion: The final protocol does not need to include a freezing step, except in cases where there must be a delay between the making of wood chips and their grinding

Test 2 (informal): Amount of wood to input in a single extraction

Test: Weight different volumes of wood powder and establish how much one can put in a 2 mL tube while still allowing to add 1.5 mL of solution
Results: 300mg (i.e. 1/3 of the 2mL tube) was the maximum powder weight that could fit
Conclusion: The final protocol should recommend to use ca. 300 mg of wood powder per extraction

Test 3: Shaking after the addition of SEVAG

Test: Four samples were submitted to either no shaking, or shaking during 10 min on an orbital shaker, using 18 replicates per sample per treatment. NB: the 18 replicates per sample per treatment differed by their parameter value for centrifugation time (test 4) and precipitation time (test 5), so when accounting for those differences too, there were 3 replicates per samples having experienced the exact same protocol. One negative control (without wood) was also included for each group of 3 replicates for each sample.
Results: Results are plotted on Figure 2.1; DNA yield is measured by its concentration in ng/uL.
Conclusion: Shaking is either better or not worse than not shaking so recommend shaking

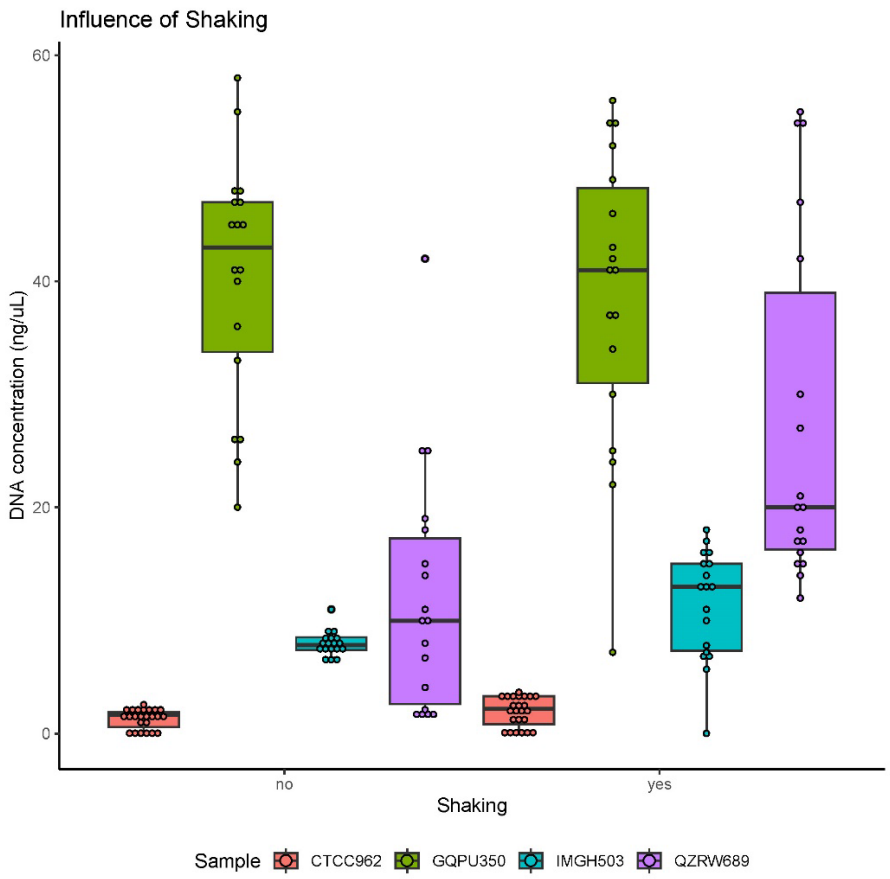


Figure 2.1. Relationship between DNA concentration and shaking at the SEVAG step.

Test 4: Centrifugation time after the addition of SEVAG

Test: Four samples were submitted to either 5 or 15 min of centrifugation after the addition of SEVAG, using 18 replicates per sample per treatment. NB: the 18 replicates per sample per treatment differed by their parameter value for shaking (test 3) and precipitation time (test 5), so when accounting for those differences too, there were 3 replicates per samples having experienced the exact same protocol. One negative control (without wood) was also included for each group of 3 replicates for each sample.
Results: Results are plotted on Figure 2.2; DNA yield is measured by its concentration in ng/uL.
Conclusion: 15 min of centrifugation is either better or not worse than 5 min so recommend 15 min.

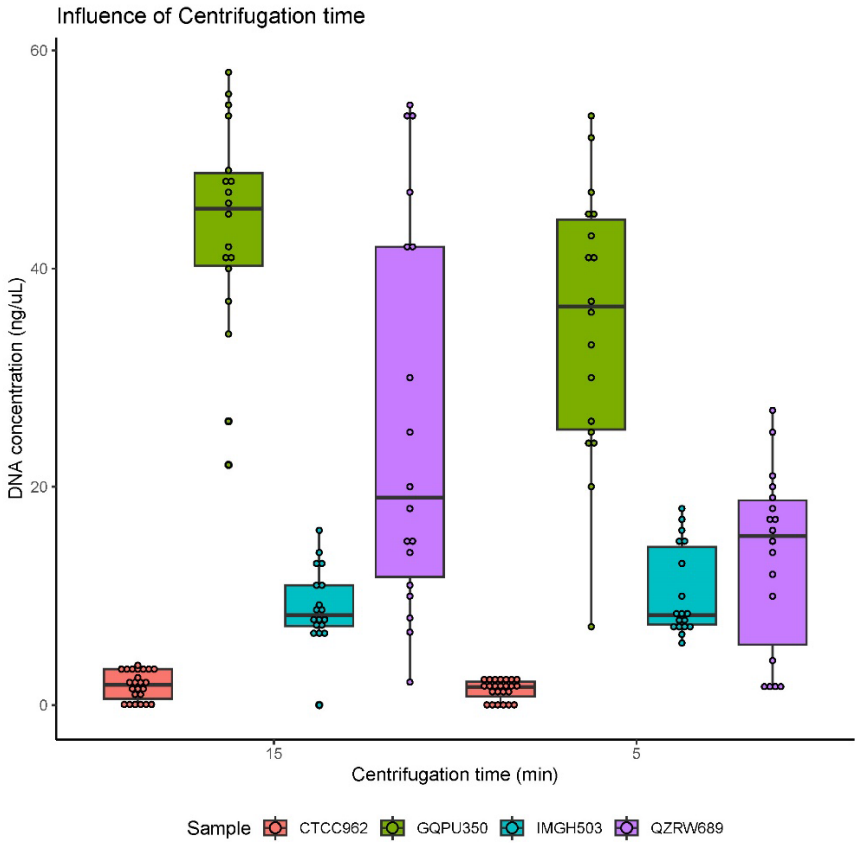


Figure 2.2. Relationship between DNA concentration and centrifugation time.

Test 5: DNA precipitation time

Test: Four samples were submitted to DNA precipitation during either a few minutes, 1 hour or overnight (ca. 12 hours), using 12 replicates per sample per treatment. NB: the 12 replicates per sample per treatment differed by their parameter value for shaking (test 3) and centrifugation time (test 4), so when accounting for those differences too, there were 3 replicates per samples having experienced the exact same protocol. One negative control (without wood) was also included for each group of 3 replicates for each sample.
Results: Results are plotted on Figure 2.3; DNA yield is measured by its concentration in ng/uL.
Conclusion: Precipitating at least 1 h seems to yield better or not worse results than not precipitating. Increasing to 12 hours does not seem better nor worse. Recommend 1h at least, and overnight if it makes the process easier.

NB: Further informal comparisons on heartwood DNA extractions that were precipitated 1 vs 11 days show some increase in concentration for the DNAs that were precipitated the longest, but this requires further testing to be demonstrated.

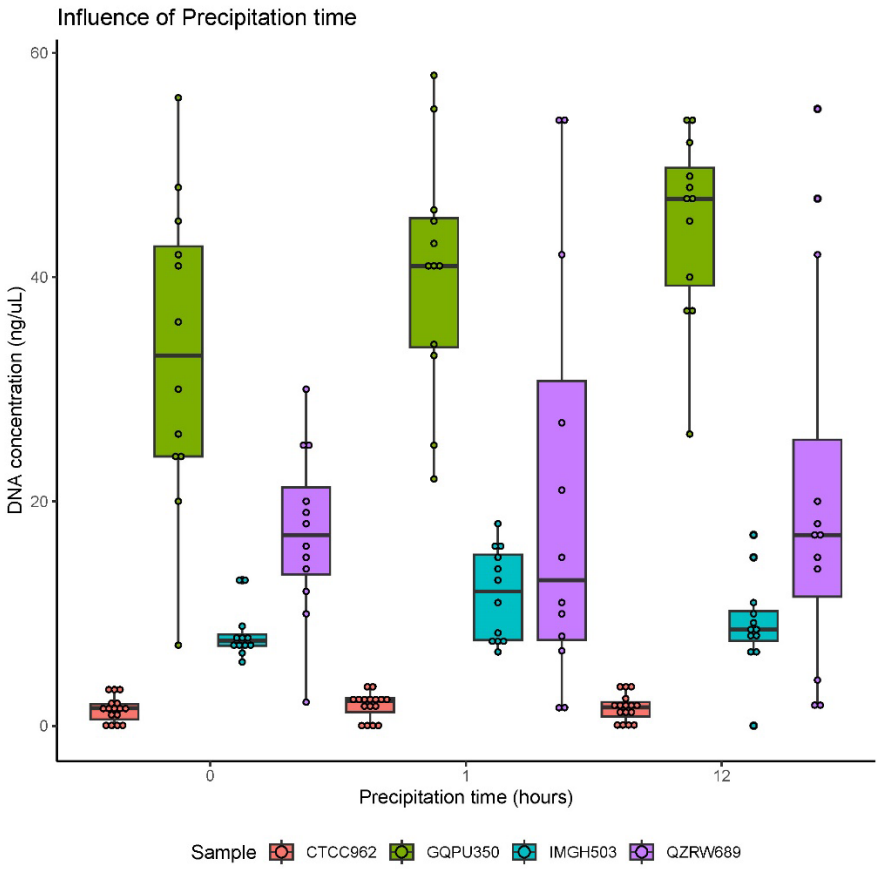


Figure 2.3. Relationship between DNA concentration and DNA precipitation time.

Other parameters explored

The use of glycogen and sodium acetate during the DNA precipitation was experimented without formal comparisons due to time limitation. Looking at the results in Table 2 below, the DNA extracts obtained with these reagents appeared to contain similar or marginally more DNA than those obtained without using these reagents. In conclusion, the use of these reagents is advised when they are easy to procure but does not appear mandatory to obtain DNA from wood.

A cleaning step before extraction (see next section) and the use of Proteinase K during incubation were also tested on some samples. Although no formal comparison can be done, most samples that underwent the cleaning step were clear and were successful in the PCR and Sequencing tests (see below). These two steps are included in the protocol but they can be considered optional as some samples that were extracted without them also worked in downstream sequencing tests (Table 2).

Table 2. DNA extractions from wood

Accepted Species Name	Sample ID	Material	DNA extract code (*samples used in the three formal tests)	Number of individual extracts before pooling	Buffer (*Lowe et al., 2015)	Glycogen (y: yes; n: no)	Sodium acetate (y: yes; n: no)	Precipitation (days)	Proteinase K (y: yes; n: no)	Pre-extraction cleaning (y: yes; n: no)	Pooled volume (*50 per extraction, not pooled)	Color after pooling	Concentration after pooling (ng/uL)	Number of regions successfully sequenced (NT: not tested)
Entandrophragma candollei	GQPU350	Sapwood	EA4	1	CTAB	n	n	1	n	n	50	NA	15	NT
Entandrophragma candollei	GQPU350	Sapwood	*GQPU350	18	CTAB	n	n	up to 1	n	n	800; 800	clear brown	47; 49	0
Entandrophragma candollei	GQPU350	Heartwood	350H	6	CTAB	n	y	14	n	y	50*	clear	1.75	7
Entandrophragma candollei	GQPU350	Sapwood	350S	7	CTAB	n	y	8	n	y	50*	clear	1.11	6
Entandrophragma angolense	QYPX983	Sapwood	EC3	1	CTAB	n	n	1	n	n	50	NA	45	NT
Entandrophragma cylindricum	EFTT752	Heartwood	JLE5	3	CTAB	y	y	1	y	n	250	black	7.3	NT
Entandrophragma cylindricum	FPBQ537	Heartwood	ECH	3	BoTAB, DTT*	y	y	1	y	n	150	black	2.2	0
Entandrophragma cylindricum	FPBQ537	Heartwood	JLE1	3	CTAB	y	y	1	y	n	300	black	2.26	NT
Entandrophragma cylindricum	FPBQ537	Heartwood	JLE13	2	CTAB	y	y	1	y	n	100	brown	2.66	0
Entandrophragma cylindricum	FPBQ537	Heartwood	JLE29	6	CTAB	n	n	11	n	n	250	brown	2.16	NT
Entandrophragma cylindricum	FPBQ537	Sapwood	ECS	3	BoTAB, DTT*	y	y	1	y	n	150	black	4	0
Entandrophragma cylindricum	FPBQ537	Sapwood	JLE7	2	CTAB	y	y	1	y	n	200	black	4.5	NT
Entandrophragma cylindricum	FPBQ537	Sapwood	JLE17	6	CTAB	n	y	1	y	n	300	black	6.1	NT
Entandrophragma cylindricum	FPBQ537	Heartwood	537H	6	CTAB	n	y	14	n	y	50*	clear brown	0.13	0
Entandrophragma cylindricum	FPBQ537	Sapwood	537S	7	CTAB	n	y	8	n	y	50*	clear brown	0.68	0
Entandrophragma utile	ABPY935	Sapwood	JLE32	5	CTAB	n	n	11	n	n	250	clear	5.8	NT
Khaya ivorensis	IMGH503	Heartwood	*IMGH503	18	CTAB	n	n	up to 1	n	n	800; 800	black	8.4; 8.5	0
Khaya ivorensis	IMGH503	Heartwood	JLE6	2	CTAB	y	y	1	y	n	170	brown	39	NT
Khaya ivorensis	IMGH503	Heartwood	JLE16	1	CTAB	y	y	1	y	n	50	brown	5.3	NT
Khaya ivorensis	FHEL563	Sapwood	JLE33	5	CTAB	n	n	11	n	n	250	clear	1	NT
Khaya senegalensis	CTCC962	Heartwood	*CTCC962	18	CTAB	n	n	up to 1	n	n	800; 800	clear	2.3; 2.5	0
Khaya senegalensis	CTCC962	Heartwood	JLE23	6	CTAB	a-c y; d-f n	y	1	y	n	150; 150	clear	0.9; 0.478	NT
Khaya senegalensis	NEMN471	Sapwood	KSS	3	BoTAB, DTT*	y	y	1	y	n	150	clear	15	1
Khaya senegalensis	NEMN471	Sapwood	JLE8	2	CTAB	y	y	1	y	n	200	clear but cloudy	13	2
Khaya senegalensis	NEMN471	Heartwood	KSH	3	BoTAB, DTT*	y	y	1	y	n	150	clear	0.077	NT
Khaya senegalensis	NEMN471	Heartwood	JLE2	3	CTAB	y	y	1	y	n	250	clear	0.6	NT
Khaya senegalensis	NEMN471	Heartwood	JLE14	2	CTAB	y	y	1	y	n	100	clear	1.47	NT

Khaya senegalensis	NEMN471	Heartwood	JLE25	8	CTAB	a-d y; e-h n	y	1	y	n	200; 200	clear	0.5; 0.6	NT
Khaya senegalensis	NEMN471	Heartwood	JLE28	6	CTAB	n	n	11	n	n	300	clear	2.6	NT
Khaya senegalensis	NEMN471	Heartwood	471H	10	CTAB	n	y	14	n	y	50*	clear	0.48	7
Khaya senegalensis	NEMN471	Sapwood	471S	10	CTAB	n	y	8	n	y	50*	clear	2.33	5
Lovoa trichilloides	QZRW689	Sapwood	*QZRW689	18	CTAB	n	n	up to 1	n	n	600; 600	clear brown, A viscous	23; 10	NT
Lovoa trichilloides	QZRW689	Sapwood	JLE9	3	CTAB	y	y	1	y	n	300	clear brown	12	NT
Lovoa trichilloides	SUBQ641	Sapwood	JLE10	3	CTAB	y	y	1	y	n	300	clear	18	6
Swietenia macrophylla	KXYH120	Sapwood	JLE31	6	CTAB	n	n	11	n	n	300	clear brown	23	NT
Swietenia macrophylla	TZER575	Heartwood	SM2H	3	BoTAB, DTT*	y	y	1	y	n	150	clear	0	NT
Swietenia macrophylla	TZER575	Heartwood	JLE3	2	CTAB	y	y	1	y	n	150	clear	0.11	NT
Swietenia macrophylla	TZER575	Heartwood	JLE15	1	CTAB	y	y	1	y	n	50	clear	0	NT
Swietenia macrophylla	TZER575	Heartwood	JLE22	23	CTAB	a-e y; f-w n	y	1	y	n	250; 1000	clear	0.24; 0.2	NT
Swietenia macrophylla	TZER575	Heartwood	JLE34	7	CTAB	n	n	11	n	n	350	clear	0.4	NT
Swietenia macrophylla	TZER575	Sapwood	SM2S	3	BoTAB, DTT*	y	y	1	y	n	150	clear	2.4	NT
Swietenia macrophylla	TZER575	Sapwood	JLE19	11	CTAB	n	y	1	y	n	500	clear brown	7.6	NT
Swietenia macrophylla	YGPL521	Heartwood	SM1H	3	BoTAB, DTT*	y	y	1	y	n	150	clear	0	NT
Swietenia macrophylla	YGPL521	Heartwood	JLE4	2	CTAB	y	y	1	y	n	180	clear	0.14	NT
Swietenia macrophylla	YGPL521	Heartwood	JLE24	18	CTAB	a-e y; f-r n	y	1	y	n	200; 500	clear	0.28; 0.17	2
Swietenia macrophylla	YGPL521	Heartwood	JLE37	6	CTAB	n	n	11	n	n	300	clear	0.33	NT
Swietenia macrophylla	YGPL521	Sapwood	SM1S	3	BoTAB, DTT*	y	y	1	y	n	150	clear	5.8	NT
Swietenia macrophylla	YGPL521	Sapwood	JLE12	2	CTAB	y	y	1	y	n	200	clear	16	4
Swietenia macrophylla	YGPL521	Sapwood	JLE20	7	CTAB	n	y	1	y	n	300	clear	18	NT
Swietenia macrophylla	YGPL521	Heartwood	521H	10	CTAB	n	y	14	n	y	50*	clear	1.56	8
Swietenia macrophylla	YGPL521	Sapwood	521S	8	CTAB	n	y	8	n	y	50*	clear	0.1	0
Unknown	ASV2	Heartwood, glued	JLE21	10	CTAB	a-e y; f-j n	y	1	y	n	250; 250	clear	0.12; 0.12	NT
Unknown	ASV2	Heartwood, glued	JLE27	6	CTAB	n	n	11	n	n	300	clear	0.3	NT
Unknown	VWBO-1	Heartwood, glued	JLE30	6	CTAB	n	n	11	n	n	300	clear	0.81	NT
Unknown	VWBO-3	Heartwood, glued	JLE36	9	CTAB	n	n	11	n	n	450	clear	1.54	NT
Unknown	VWMO-1	Heartwood, glued	JLE35	9	CTAB	n	n	11	n	n	450	clear	0.27	NT
Unknown	ASV1	Heartwood, glued	JLE26	6	CTAB	n	n	11	n	n	300	clear	0	NT

b. Final DNA extraction protocol for timber

(Also provided as Supplementary Methods of Article 1)

The protocol is mainly based on Doyle & Doyle 1987, Inglis et al., 2018 and Lu et al., 2020, with adaptations by L. Csiba, S. Bellot and C. Quintero-Berns (RBG Kew). Another lysis buffer containing boric acid (BoTAB; Lowe et al., 2015) was experimented on a few samples (Table 2).

1. Sample preparation and grinding

1a. Cut or use chisel to break down the wood tissue in small (5 mm) chips, and put them into a stainless steel mill jar for grinding

1b. Put two metal beads in the jar and put all at -80°C if not proceeding quickly to step 1d. If proceeding to 1d after freezing, avoid letting the chips and jar warm to room temperature.

1c. Set up as many labelled 2 mL microcentrifuge tubes as needed depending on the quantity of chips.

1d. Grind the wood chips into a fine powder, for instance by grinding 4 times 6 min at 25 Hz. Adjust the time or number of grinding rounds until powder is obtained. Proceed immediately to the next step.

1e. Using a clean spatula, spoon ca. 300 mg (about a third of a 2 mL tube) of wood powder from the jar into a labelled tube and proceed immediately to the next step.

2. [Optional] Cleaning

2a. Add 1500 µL Sorbitol Buffer, 0.35 M + 15 µL Mercaptoethanol

2b. Centrifuge at 5000 rcf for 5 mins

2c. Discard supernatant

2d. Repeat steps 2a to 2c

3. Sample lysis with CTAB

3a. Warm [745 µl * (number of samples + 2)] of CTAB in 55°C water bath / oven

3b. Just before proceeding with 3c, add [3µl * (number of samples + 2)] of 2-mercaptoethanol to the heated CTAB: this is the extraction buffer

3c. Add 750 µl of this extraction buffer to each tube containing wood powder

3d. Vortex to make sure the CTAB is in contact with all the powder, dissolving any clump

3e. Incubate tubes at 55°C for 5 hours – shaking constantly and/or vortexing regularly

3f. [Optional] 30 min before the end of the incubation, add 15 µL of proteinase K originally prepared at a concentration of 20 mg/mL and vortex

4. SEVAG Addition

4a. Transfer 750 µl of SEVAG to each tube. Close tubes tightly and vortex thoroughly

4b. Agitate tubes horizontally on an orbital shaker for 10 mins

4c. Centrifuge tubes at 13,000 rpm for 15 mins

4d. Transfer the supernatant to a new labelled tube

4e. Repeat steps above

5. DNA Precipitation

4a. Add equal volume of chilled isopropanol to each tube

5b. [Optional] Add 65 uL (i.e. ca. $1/10^{th}$ of the volume already in the tube) of sodium acetate (NaOAc 3M pH 5.2) + 10 µL of glycogen

5c. Ensure lids are firmly closed and gently invert tubes several times

5d. Keep in the freezer (-20^{o}C) at least 1 hour and overnight if convenient

6. DNA Elution and Resuspension

6a. Centrifuge sample tubes at 13,000 rpm for 20 minutes

6b. Discard the supernatant over a waste container

6c. Add 750 µl of 70% ethanol to each tube; vortex or flick to detach the pellet from tube

6d. Centrifuge tubes at 13,000 rpm for 5 minutes

6e. Discard supernatant, being extra careful to not remove DNA pellet

6f. Repeat ethanol wash (steps 6c-6e)

6g. Open the tubes and allow the pellet to dry completely until no ethanol remains

6h. Heat an aliquot of Molecular Grade Water in 65^{o}C water bath / oven

6i. Dissolve the ethanol pellet in 50 µl of water (increase to 75 µl or 100 µl if necessary due to viscosity of the DNA)

7. *[Optional] Cleaning*

Not usually done as a lot of DNA can be lost, but consider doing it if the DNA extract is dark brown or black and/or viscous and/or if the final DNA pellet is dark, shiny and does not easily come off the tube wall at the previous step. Use SPRI beads, not columns.

Solutions used in the protocol

- Sorbitol Buffer 0.35 M (from Inglis et al., 2018)

  - 100 mM Tris-HCl pH 8.0
  - 0.35 M Sorbitol
  - 5 mM EDTA pH 8.0
  - 1% (w/v) Polyvinylpyrrolidone (average molecular weight 40,000; PVP-40)

- 2X CTAB (Cetrimonium bromide)

  - 100 mM Trizma® TRIS base, NH2C(CH2OH), C4H11NO3, FW 121.1
    - 100 mL/L 1 M pH 8.0 stock solution
  - 20 mM EDTA, C10H14N2O8Na2·2H2O, FW 372.2
    - 80 mL/L 0.25 M pH 8.0 stock solution
  - 1.4 M Sodium Chloride, NaCl, FW 58.44, 81.82 g/L
  - 2 % (w/v) CTAB, C19H42BrN, FW 364, 20.00 g/L
  - 2 % (w/v) PVP, Av.Mol.Wt. 40,000 / Av.Mol.Wt. 36,000, 20.00 g/L
  - Molecular Grade H_2O up to 1000 mL

- SEVAG
  - 24 units Chloroform, Trichloromethane, CHCl3, FW 119.4
  - 1 unit Isoamyl alcohol, 3-Methyl butanol, (CH3)2CHCH2CH2OH, FW 88.15

References

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.

Inglis PW, Pappas MdCR, Resende LV, Grattapaglia D (2018) Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. PLoS ONE 13(10): e0206085. https://doi.org/10.1371/journal.pone.0206085

Lowe, A. J., Jardine, D. I., Cross, H. B., Degen, B., Schindler, L., Hoeltken, A. M. (2015). A method of extracting plant nucleic acids from lignified plant tissue. International Patent Number WO/2015/070279.

Lu, Y., Jiao, L., He, T., Zhang, Y., Jiang, X. and Yin, Y., 2020. An optimized DNA extraction protocol for wood DNA barcoding of *Pterocarpus erinaceus*. IAWA Journal, 41(4), pp.644-659.

3. Identifying and testing new DNA barcodes

The reference DNA dataset described in 1) was used to identify DNA regions that could serve as DNA barcodes for distinguishing species of the four focus genera. Requirements for a functional DNA barcode in our context included its ability to distinguish species from each other (i.e. phylogenetic informativeness at the species level) but also the possibility to sequence the DNA barcode using PCR and Sanger sequencing, which are more readily available and cheaper methods to implement in the focus countries. The whole search for suitable DNA barcodes and their testing on the focus species are described in detail in Article 1 (Annex 5.5 of the main report) and only summarized below.

a. Candidate DNA barcodes

For each gene alignment from the reference dataset described in 1, a gene tree was estimated with IQ-TREE v.1.6.12 (Minh et al., 2020) based on the Maximum Likelihood phylogenetic inference method, following the best model of nucleotide substitution identified for the region by IQ-TREE's ModelFinder approach (Kalyaanamoorthy et al., 2017), and performing 1000 ultrafast bootstrap replicates. Gene trees were rooted on *Schmardaea* or, when not present in the tree, on *Lovoa*, using the pxrr program of the Phyx toolkit (Brown et al., 2017). Gene trees that could not be rooted on either of the two genera were discarded from downstream analyses as lacking one of the focus genera would make them sub-optimal for DNA barcoding. The gene trees were then imported in R Studio and a custom script (available at https://github.com/sidonieB/Bellot_al_Meliaceae_DNA_barcoding/tree/main) was written to analyze the informativeness and recovery rate of each gene across the four focus genera.

We estimated the potential for DNA barcoding of each nuclear and plastid gene sequenced in this study based on two criteria. The first criterion was the rate of recovery of the gene among our samples (as a better recovery would enable a more informed primer design); this was estimated by calculating the average species' percentage of individuals for which the gene was recovered. The second criterion was the gene's propensity to recover species as monophyletic (indicative of the information contained by the gene that can allow distinguishing samples of different species from each-other); this was estimated by calculating the average percentage of individuals recovered in the largest clade for their species.

As shown on Figure 3.1, among the 174 best genes (i.e. with a value superior to 70% for both criteria), only 2 were plastid genes while the rest were nuclear regions, including ITS. Nuclear and plastid genes had similar performances for the first criterion but nuclear genes were better at recovering species as monophyletic (criterion 2).

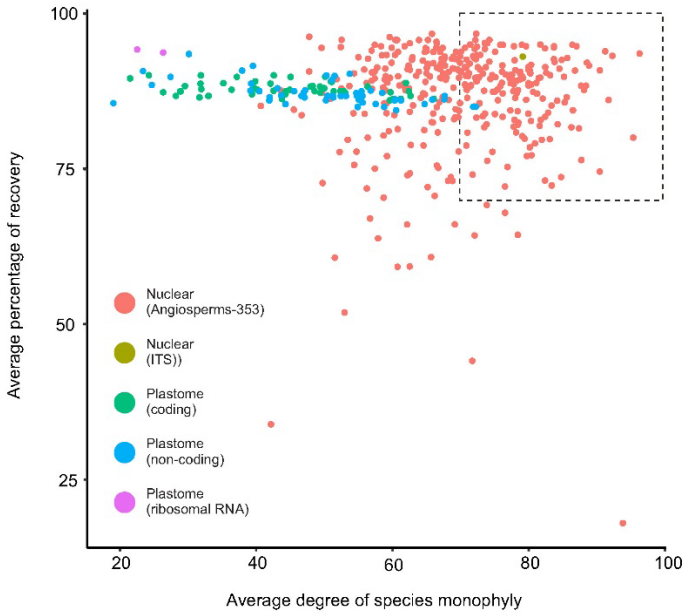


Figure 3.1. Potential of single DNA regions to serve as DNA barcodes in four Meliaceae genera. The dashed box encloses the genes displayed on panel

All the best genes were potentially paralogous (i.e. CAPTUS assembled more than one sequence for that gene in at least one sample), except one: the Angiosperms353 gene “g5816” (Figure 3.2). This gene was therefore selected for the search of new DNA barcode regions (see below). In addition, two other genes were selected randomly among the other best nuclear genes: Angiosperms353 genes “g6420” and “g6968”. Both genes had intermediate levels of paralogy, with 53% and 58% of the samples having more than one copy for these genes respectively (Figure 3.2).

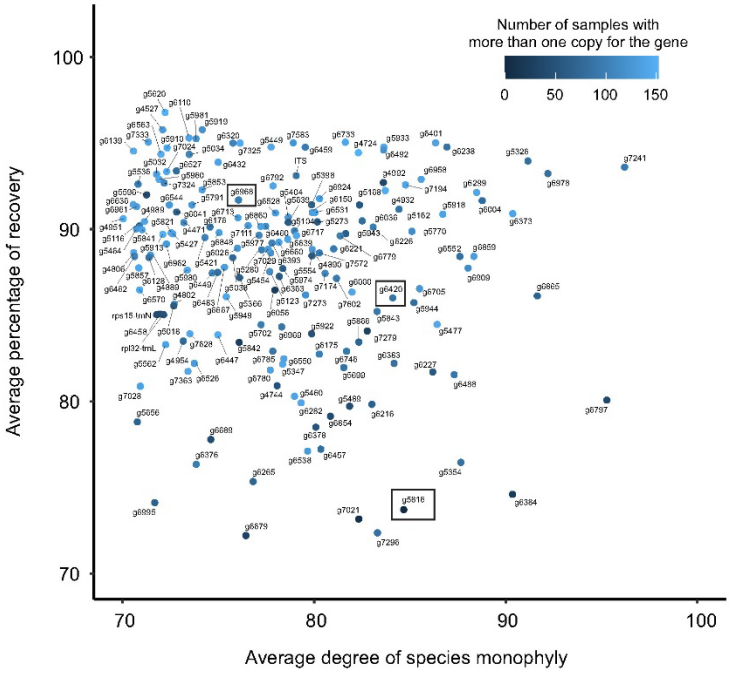


Figure 3.2. Most promising genes for DNA barcoding in the focus genera. Boxes indicate the three genes in which new DNA barcodes were designed for this study.

The untrimmed alignments of the three genes were screened using Geneious Prime 2024 to look for relatively short but variable regions to sequence flanked by regions more conserved in which PCR primers could be designed that would likely work in all focus genera. This resulted in the identification of ten regions to amplify and their corresponding primer pairs. Primer design failed for one of the regions (which we called 5816_r3). The other nine regions (5816_r1, 5816_r2, 5816_r4, 6420_r1, 6420_r2, 6968_r1, 6968_r2, 6968_r3a and 6968_r3b) and their primers are described in Figure 3.3 and Table 3. The standard DNA barcode ITS and the plastid intergenic spacer trnL-trnF were also described and analysed for comparison.

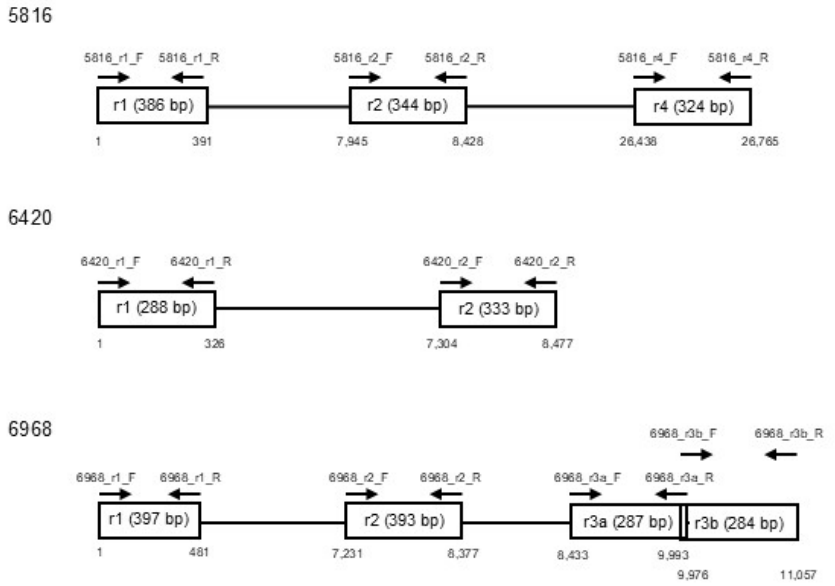


Figure 3.3. Schematic representation of the nine new barcodes and corresponding primers. Arrows represent primers designed in this study. Coordinates indicate the position of the regions in the reference alignments. The size of an individual region is the median length of the sequences in the reference alignments obtained from the target capture sequencing data. Scale is not respected.

The new barcodes have median lengths of 284-397 bp, against 368-425 bp for ITS and trnL-trnF. When looking at all the genera together, the new barcodes have more informative sites than trnL-trnF, and two (6968_r2 and 6968_r3a) have more than the ITS region (139-145 against 136). These relative differences are also found when looking at individual genera (Table 1). Phylogenetic trees inferred from the barcoding regions only show a high degree of resolution, confirming their high potential as new barcodes (shown in Article 1).

Table 3 Characteristics of the nine new and two traditional DNA barcodes. F: Forward primer; R: Reverse primer; ML: Median lengths of the region in the reference alignments created from the target capture sequencing data; IS: number of informative sites in the reference alignments. Abbreviations under each region indicate how promising the region is for DNA barcoding in the focus genera based on tests performed in this study: IN: intermediate, LI: less informative, NW: not working with the current protocol, WW: working well with the current protocol (see text of next section for further explanations).

Region	Primers (bold: newly designed)		All genera	*Entandrophragma*	*Khaya*	*Lovoa*	*Swietenia*
5816_r1 (NW)	**F: ATTCTATGCTGTCTTTTTCTGATTC**	ML:	386	386	386	387	387
	R: AGTTGATAGTAACTTAAGTTGAC	IS:	99	45	2	0	10
5816_r2 (WW)	**F: TGTTGGCTTCTATGGATTCTGC**	ML:	344	344	356	306	332
	R: CCAGCAATGAAATCAGTCCCTTG	IS:	108	50	7	0	7
5816_r4 (IN)	**F: GCAACAAGTTCTACCATAGTCACC**	ML:	324	324	328	328	315
	R: CACAAGACAAACCCGGTTCAC	IS:	65	32	2	0	8
6420_r1 (WW)	**F: TGCTTATATGCTTCATAATCTAGCTC**	ML:	288	291	288	291	287
	R: AGCACATTATGAGCAGGCATATC	IS:	61	29	6	11	5
6420_r2 (WW)	**F: GAGGCTAGAGGACTTGATGC**	ML:	333	333	333	333	333
	R: AGCTGAATAATTGAAGGGTCC	IS:	82	38	21	17	4
6968_r1 (LI)	**F: CGGGACAGCAGGCTAATGAAG**	ML:	397	397	397	389	391
	R: CCTCTGCTTTTGCCAGACCC	IS:	103	51	24	17	5
6968_r2 (IN)	**F: CCTCGTGCTGCTGCTGC**	ML:	393	393	393	392	390
	R: AAGCTGCCCCCTATTACAGG	IS:	139	92	25	19	11
6968_r3a (NW)	**F: GCAGGAGCATTGCAGGTAG**	ML:	287	343	287	337	287
	R: GCGAAAAGAAAGCAGGCTCTG	IS:	145	91	15	23	6
6968_r3b (IN)	**F: AGCCTGCTTTCTTTTCGCAAG**	ML:	284	284	275	283	288
	R: GTGGTGGGAAAAAGCCTATCC	IS:	122	78	27	18	12
ITS (IN)	F (17SE = AB101): ACGAATTCATGGTCCGGTGAAGTGTTCG	ML:	425	325	431	377	332
	R (ITS2): GCTGCGTTCTTCATCGATGC	IS:	136	76	22	42	15
trnL-trnF (LI)	F (trnLe): GGTTCAAGTCCCTCTATCCC	ML:	368	368	368	368	368
	R (trnFf): ATTTGAACTGGTGACACGAG	IS:	17	5	0	3	3

b. Performance of the new barcodes on wood samples

The PCR amplification and Sanger sequencing of the nine new barcodes, of the ITS region (using primers 17SE and ITS-2 described in Sun et al., 1994 and White et al., 1990 respectively), and of the plastid intergenic spacer trnL-trnF (using primers e and f described in Taberlet et al., 1991) were tested on 19 wood DNA extracts. These samples included 9 heartwood DNAs from *E. cylindricum* (3 samples from 1 collection), *E. candollei* (1 sample), *K. ivorensis* (1 sample), *K. senegalensis* (2 samples from 2 different collections), and *S. macrophylla* (2 samples from 1 collection), and 10 sapwood DNAs from *E. cylindricum* (2 samples from 1 collection), *E. candollei* (2 samples from 1 collection), *K. senegalensis* (3 samples from 1 collection), *S. macrophylla* (2 samples from 1 collection) and *L. trichilioides* (1 sample). Heartwood and sapwood samples available for the same species (i.e. for *E. cylindricum*, *E. candollei*, *K. senegalensis* and *S. macrophylla*) always came from the same collection. The list of samples used is provided in Table 2 with indications on how the DNA was extracted for each sample, while the full DNA extraction protocol is provided in Section 2b. High molecular weight DNAs from leaf tissue of *E. angolense* and *K. senegalensis* that had been successfully used to prepare DNA libraries and generate Illumina data for the DNA reference dataset (i.e. samples EA6 and KS126 ; Table 1) were used as positive controls, while water was used as negative control. PCR was done by mixing 2 µL of the sample DNA with 8.5 µL of water, 10 µL of TBT (5X), 2.5 µL of DMSO, 1 µL of each primer (10 µM) and 25 µL of Taq polymerase (2X 'Dream Taq' Thermo Sci., 4.0 mM MgCL2). The mix was then incubated in a thermocycler with the following programme: 2 min at 94°C + 28*(1 min at 94°C + 1 min at 52°C + 1 min at 72°C) + 7 min at 72°C. PCR products were analyzed by electrophoresis in a 1% Agarose gel with SYBRTM Safe dye (Invitrogen) dye and photographed using a UVP GelStudio imaging system (Analytikjena). PCR products that showed clear bands in the electrophoresis were purified using the Macherey-Nagel NucleoSpin® Purification Kit (52355 Düren, Germany). Clean products were sequenced using a 3730sl DNA Analyzer (Applied Biosystems).

Sequences obtained by Sanger sequencing were cleaned and corrected based on the observation of sequencing chromatograms in Geneious Prime 2024. When available, the Forward and Reverse reads were joined into a single sequence. All clean sequences were then checked for non-plant contamination by comparing them with the "Core nucleotide database" of GenBank, NCBI using the "blastn" algorithm (Altschul et al., 1990), as implemented in the dedicated online portal (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) but done through Geneious Prime. Sequences that matched plants or returned no results at all (keeping in mind that the new regions were not yet present in GenBank for the focus genera or even Meliaceae) were kept while four sequences were discarded as they matched bacteria or animal genes. The sequences were then aligned to the set of reference sequences available for each gene from the reference dataset. The resulting alignments were examined in Geneious Prime 2024. When multiple sequences of a same DNA sample were available for a same region, their consensus (including Ns and ambiguities as needed if there was any gap or conflict between the sequences) was used so that final alignments contained each sample only once. Gene trees were then generated using IQ-TREE from each region alignment, with automatic selection of the nucleotide substitution model and 1000 ultrafast bootstrap replicates, as described above. All alignments and phylogenetic trees are available at https://github.com/sidonieB/Bellot_al_Meliaceae_DNA_barcoding/tree/main.

As shown in Figure 3.4 and in the last column of Table 2, PCR and Sanger sequencing of at least one region among the nine new barcodes regions, ITS and trnL-trnF was successful in four out of the nine heartwood and six out of the ten sapwood samples tested, representing seven out of the eleven collections/wood-type combinations and four out of the six species tested (Table 2).

Between two and eight regions could be successfully sequenced from the heartwood samples, and between one and six regions could be sequenced from the sapwood samples (Figure 3.4; Table 2). The two species that did not work in any of the tests were *E. cylindricum* (three heartwood and two sapwood samples from a single collection were tried) and *K. ivorensis* (one heartwood sample tried).

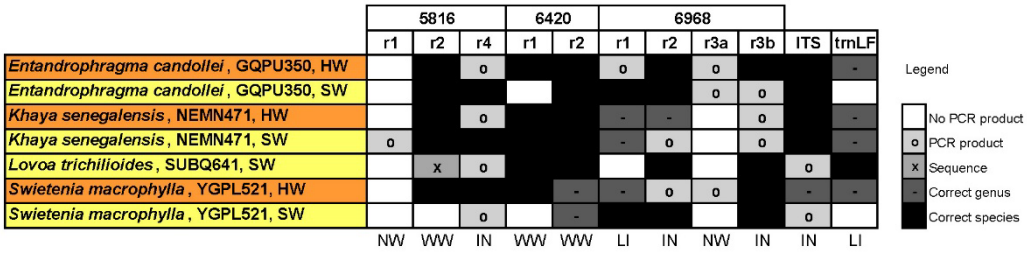


Figure 3.4. Identification of wood samples using PCR and Sequencing of DNA barcodes. Summary of the success of the PCR, sequencing and phylogenetic analysis of the nine new barcodes, the ITS and the trnL-trnF regions for four species of the focus genera. The heatmap depicts the most advanced test stage successfully achieved with at least one sample of the species for each wood type available (HW: heartwood; SW: sapwood), for each region. "Sequence" means that a sequence was produced but did not place in the correct genus in the phylogenetic tree. Abbreviations under each region indicate how promising the region is for DNA barcoding in the focus genera: IN: intermediate, LI: less informative, NW: not working with the current protocol, WW: working well with the current protocol (see text for further explanations).

Three regions (5816_r2, 6420_r1 and 6420_r2) worked well with the current protocols, showing high rates of success in terms of PCR amplification and sequencing coupled with accurate phylogenetic placement at the species level. In contrast, four intermediate regions (5816_r4, 6968_r2, 6968_r3b, ITS) were less easily sequenced but when they were obtained, they could also accurately identify the samples at the species level, while two less informative regions (6968_r1, trnL-trnF) were easily sequenced but only seldomly succeeded at identifying the samples beyond the genus level (Figure 5; Figure S4). Finally, two regions did not work well with the current protocols (5816_r1, 6968_r3a), with almost no sequence generated (Figure 3.4, Table 3).

Comparing the DNA extractions from which barcode sequences could be obtained with those from which no sequence was obtained showed that regions longer than 300 bp could be sequenced from wood DNA samples, including when a lot of the DNA in these samples appeared heavily degraded (< 100 bp). Nevertheless, many of the samples that showed no success at all had highly fragmented DNA (Figure 3.5; Table 3). There seemed to be a relationship between sequencing success and DNA purity when the latter was assessed from the DNA color in that no region could be amplified from any of the dark brown or black DNA extracts (Figure 3.5; Table 3). However, a clear DNA was not a guarantee for successful DNA barcode sequencing, even in the presence of large DNA fragments. The relationship between success and DNA purity was less clear when purity was measured based on absorbance ratios 260/280 and 260/230 since barcodes could be sequenced from DNA extracts with ratios well inferior to 2 (Figure 3.5; Table 3).

Figure 3.5. Relationship between DNA quality and PCR and Sanger sequencing results.

Table 3. PCR and Sequencing of the DNA barcodes from wood DNA

For tissue, HW is heartwood, SW is sapwood. For size range, A: >1500 bp; B: 500-1500 bp; C: 100-500 bp, D: <100 bp.

Accepted Species Name	Sample ID	Tissue	DNA extract code	Color	Concentration after pooling (ng/uL)	Size range	Main size	260/280	260/230	5816_r1	5816_r2	5816_r4	6420_r1	6420_r2	6968_r1	6968_r2	6968_r3a	6968_r3b	ITS	trnL-trnF	Number of regions successfully sequenced
Entandrophragma candollei	GQPU350	SW	GQPU350	clear brown	47	A(B)	A	1.54	0.77	N	N	N	NT	NT	N	N	N	N	NT	NT	**0**
Entandrophragma candollei	GQPU350	HW	350H	clear	1.75	C	C	1.05	0.26	N	**species (99)**	**P species (69)**	**species (97)**	**species (99)**	**P species (98)**	**species (100)**	**P**	**species (100)**	**species (40)**	**genus (77)**	**7**
Entandrophragma candollei	GQPU350	SW	350S	clear	1.11	(B)C	C	1.19	0.3	N	**species (99)**	N	**species (99)**	**species (65)**	**P**	**P**	**species (40)**	N	**6**		
Entandrophragma cylindricum	FPBQ537	HW	ECH	black	2.2	(B)C	C	1.24	0.67	N	N	N	NT	NT	N	N	N	N	NT	NT	**0**
Entandrophragma cylindricum	FPBQ537	HW	JLE13	brown	2.66	(B)C	C	1.23	0.64	N	N	N	N	N	N	N	N	N	N	N	**0**
Entandrophragma cylindricum	FPBQ537	SW	ECS	black	4	(A)B	B	1.23	0.77	N	N	N	NT	NT	N	N	N	N	NT	NT	**0**
Entandrophragma cylindricum	FPBQ537	HW	537H	clear brown	0.13	D	D	1.34	0.55	N	N	N	N	N	N	N	N	N	N	N	**0**
Entandrophragma cylindricum	FPBQ537	SW	537S	clear brown	0.68	(A?)D	D	1.24	0.58	N	N	N	N	N	N	N	N	N	N	N	**0**
Khaya ivorensis	IMGH503	HW	IMGH503	black	8.4	(ABC?)D	D	2.36	2.13	N	N	N	NT	NT	N	N	N	N	NT	NT	**0**
Khaya senegalensis	CTCC962	HW	CTCC962	clear	2.3	B,C	B	0.9	0.26	N	N **species**	N	NT	NT	N	N	N	N	NT	NT	**0**
Khaya senegalensis	NEMN471	SW	KSS	clear	15	A(B)	A	1.93	1.13	**P**	**(74)**	**P**	NT	NT	**P**	**P**	**P**	**P**	NT	NT	**1**
Khaya senegalensis	NEMN471	SW	JLE8	clear but cloudy	13	B,C	C	1.5	0.59	N	N **species**	**P**	**species (22)**	**P species (43)**	**P genus (99)**	**P genus (31)**	**P**	**P**	**S species (80)**	N **genus (63)**	**2**
Khaya senegalensis	NEMN471	HW	471H	clear	0.48	C	C	0.83	0.19	N	**species (72)**	**P species (70)**	**species (35)**	**species (84)**	**genus (99)**	**P species (74)**	**CS**	**P species (100)**	**species (80)**	**genus (92)**	**5**
Khaya senegalensis	NEMN471	SW	471S	clear	2.33	(B)C	C	0.75	0.12	N	**P**	N **species (95)**	**species (99)**	N	N	**species (97)**	**species (97)**	**genus (40)**	**genus (53)**	**8**	
Lovoa trichilloides	SUBQ641	SW	JLE10	clear	18	(A)B	B	1.9	1.41	N	**other (47)**	**P**									**6**
Swietenia macrophylla	YGPL521	HW	JLE24	clear	0.28	(C),D	D	1.62	0.37	N	**S**	**S**	N	**P genus (48)**	N **species (31)**	**P species (27)**	N **CS**	**P species (97)**	**P**	N	**2**
Swietenia macrophylla	YGPL521	SW	JLE12	clear	16	(A)BC	B	2.27	0.79	N	N **species (87)**	**P species (38)**	N **species (20)**								**4**
Swietenia macrophylla	YGPL521	HW	521H	clear	1.56	(C),D	D	2.84	0.26	N											**8**
Swietenia macrophylla	YGPL521	SW	521S	clear	0.1	NA	NA	2.16	0.21	N	N	N	N	N	N	N	N	N	N	N	**0**

c. Tutorials

The steps to design primers to amplify DNA regions and the analysis that can be performed to identify a wood sample based on a DNA sequence have been written up so that they can be repeated by the partners and their colleagues after the project. The resulting tutorials (in French) are pasted below.

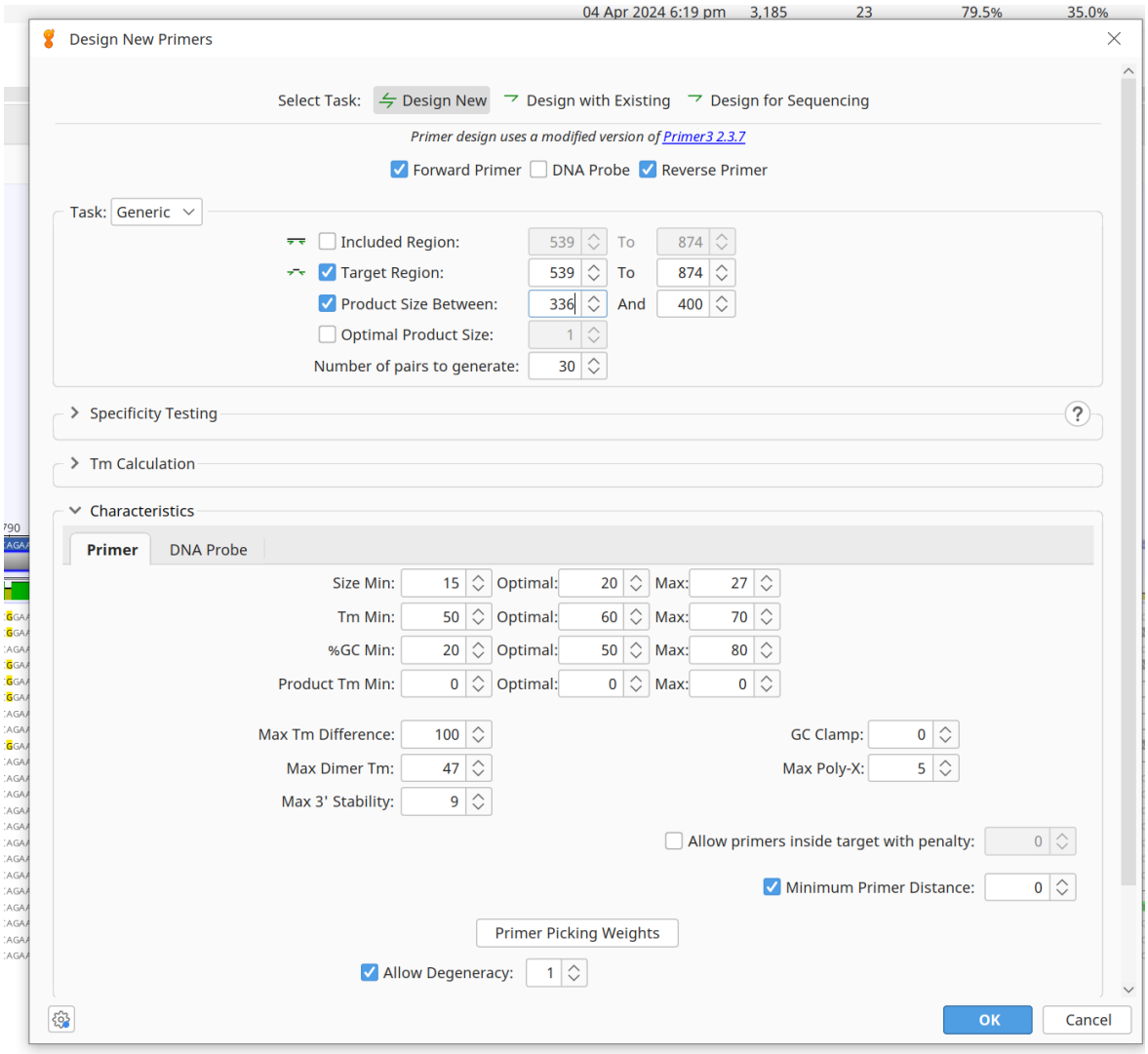
Tutoriel 1 – Design d'amorces

- Regarder dans l'alignement complet quelles peuvent etre les régions cibles intéressantes (variables) et si elles ont des régions conservées autour qui faciliteront le design d'amorces
- Anoter les regions cibles de maniere approximative sur le consensus
- Classer les sequences par espece pour plus de facilité (click droit sur une séquence > sort > by name)
- Sélectioner une séquence bien complete (longue) pour chaque espece (presser sur Ctrl pour sélectioner plusieurs séquences a la fois)
- Extraire les séquences (click droit > Extract Regions > Extract Region as alignment > ok)
- Faire le design d'amorces sur le petit alignement ainsi créé (voir capture d'écran et criteres ci-dessous)
    - selectionner et annoter les regions a amplifier (excluant les régions ou on veut chercher les amorces) sur le petit alignement en se basant sur la sélection faite sur le grand alignement (si tout va bien les regions annotées sur le grand alignement le seront déja sur le petit, sinon il faudra les retrouver)
    - pour une region a amplifier : sélectioner la région, cliquer sur Primers > Design New Primers. Dans la fenetre de design :
        - Garder selectioné « Forward » et « Reverse » primers
        - Déselectioner « included region »
        - Garder « Target region » selectioné (les valeurs devraient correspondre a la position de la region a amplifier)
        - Sélectioner « Product size » et changer les valeurs : le minimum doit correspondre a la taille de la région a amplifier sans compter les gaps, le maximum doit permettre le design d'amorces dans des régions flanquantes pas trop éloignées, par exemple on peut ajouter environ 100 bp a la taille minimum, donc si par exemple le minimum est 275 bp, alors on peut tenter un maximum de 375 bp et modifier par la suite si aucune bonne amorce n'est identifiee
        - Adapter « Numbers of pairs to generate", par exemple essayer 10 et augmenter si aucune bonne amorce n'est trouvée (NB si trop de paires sont données sur un petit écran on ne voit plus l'alignement, donc éviter les trop grands nombres)
        - Changer les valeurs de size dans « Characteristics » min : 15, opt : 20, max : 27
        - Changer les valeurs de Tm dans « Characteristics » min : 50, opt : 60, max : 70
    - Une fois les amorces proposées, choisir celles qui nous conviennent le mieux (parfois il faudra modifier un peu les amorces proposées, ou bien recommencer la recherche avec des parametres un peu differents)
    - Pour chaque amorce choisie, annoter l'amorce (sélectioner la région, click droit quand la souris est sur la region sélectionée, add annotation). Utiliser l'annotation « Misc Feature »
    - Une fois l'amorce F et l'amorce R choisies pour une region, faire une copie du document et sur la copie, supprimer les amorces non-choisies toutes a la fois (Delete all primer binds). Proceder éventuellement a la recherche d'amorces pour une autre région du meme gene en répetant les étapes précedentes.

- Une fois toutes les amorces choisies pour toutes les régions sur le petit alignement, utiliser copier + Ctrl+F pour trouver les amorces sur l'alignement avec tous les individus et annoter alors les amorces

sur cet alignement avec l'annotation Primer Bind. Ainsi l'alignement de départ contiendra toutes les amorces annotees. Verifier que les regions a amplifier seront bien de la taille prevue pour la plupart des individus de l'alignement complet.

- Une fois satisfait, exporter le tableau des annotations de l'alignement complet, qui doit normalement contenir la sequence de chaque amorce (toujours dans l'ordre 5'>3', donc reverse-complementée pour les amorces reverse)
- Commander les amorces



Criteres de design pour cette étude :

- **Product size** (region a amplifier = region entre les amorces, a l'exclusion des amorces): viser environ 300 bp, regler la taille minimum correspondant a notre selection and la taille maximum a environ 400 bp

- **Primer size**: 15-27 bp

- **Primer sequence:**

Particulierement important vers la "tete de fleche" de l'amorce, c'est a dire du coté de l'extrémité 3'

  - Séquence conservée le plus possible entre especes
  - Le 3' end devrait idéalement etre un G ou un C
  - Preferer un équilibre GC/AT, avec un peu plus de GC si possible mais pas un contenu GC extrement haut non plus
  - Eviter les répetitions (TTTT, AAA etc)

- **Melting temperature**:
  - Idealement moins de 5 degrés C de difference entre les amorces d'une meme paire
  - min 50C, max 70C, optimal 60C

- **Autres aspects**
  - Eviter les paires d'amorces ou les deux amorces sont complémentaires car elles pourraient s'attacher entre elles
  - Eviter les amorces dont la sequence est repetée dans le génome et pourrait s'attacher a d'autres régions que la région cible

Explications approfondies (en anglais):

Vocabulaire:

https://www.bioinformatics.nl/molbi/SCLResources/sequence_notation.htm#:~:text=Forward%2C%20reverse%2C%20(%2B)%20and%20(%2D),the%20end%20of%20a%20gene.

Design d'amorce (voir parametres en fin de page): **https://sharebiology.com/primer-designing-demonstration-step-by-step/**

Primer3Plus **https://www.primer3plus.com/index.html** est un logiciel gratuity en ligne pour le design d'amorces (aussi utlisé par Geneious)

Details sur la temperature (extrait de la Primer3Plus Help page): "The **annealing temperature** in a PCR reaction is usually chosen 6-10°C below the melting temperature of the primers […]. The idea behind this reduction in temperature is to increase the fraction of primers bound to target. While at the melting temperature 50% of the primers are bound to target, at the reduced annealing temperature 95-98% should be bound.[…] Ideally, primers should not be matched on melting temperature (PRIMER_OPT_TM) but on the **fraction of primers bound at annealing temperature** (PRIMER_OPT_BOUND)."

Tutoriel 2 - Identification d'échantillons par PCR et séquencage Sanger

Laboratoire (résumé) :

- Extraction ADN
- Evaluation de la quantité (nanodrop) et de la qualité de l'ADN (ratios nanodrop, electrophorese)
- Purification de l'ADN (optionelle)
- PCRs : amplification des régions cibles
- Sequencage Sanger : production de séquences ADN pour les régions amplifiées

➔ Résultat pour un gene comportant plusieurs regions cibles :
- Séquences avec chromatogrames (pics) indiquant la qualité de chaque nucléotide lu
- En général, deux séquences (une reverse R et une forward F) sont produites pour chaque région amplifiée

Analyses

1. **Programmes nécessaires :**

Visualiser et corriger les séquences :

Geneious (idéalement mais payant) : https://www.geneious.com/

Ou bien Ugene (gratuit) : https://ugene.net/

Ou bien AliView (gratuit, moins complet ?) : https://ormbunkar.se/aliview/

Blast (en ligne) : https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Aligner les séquences :

MAFFT en ligne ou en plug-in de Geneious

Arbre phylogenetique:

IQ-TREE (en ligne) : http://iqtree.cibiv.univie.ac.at/

Ou bien Raxml en plug-in de Geneious

Figtree (visualisation): https://github.com/rambaut/figtree/releases (télécharger le *.zip si c'est pour Windows)

2. **Correction des séquences :**

- Ouvrir Geneious (ou équivalent)
- Creer un nouveau dossier
    - File > New > Folder
- Importer et formater la sequence a corriger
    - File > Import > Files... [choisir le fichier et importer]
    - Sélectioner la séquence pour qu'elle soit visible dans le paneau du bas
    - Changer le code couleur pour qu'il corresponde aux nucléotides (icone maison sur la droite General > Colors > choisir le meme code que celui utilisé pour les pics, souvent ACGT)
    - Zoomer/dézoomer selon les besoins en utilisant la loupe a droite. Le chromatograme peut aussi etre remonté ou descendu.

- Vérifier si les nucléotides sont bien codés en comparant le nucléotide au pic. Corriger le nucléotide si besoin :
    - Cliquer sur Allow Editing (au dessus de la séquence, panneau du bas)
    - Selectioner le nucléotide a corriger et remplacer par la lettre appropriée.
    - Si on a une ambiguité, remplacer par le code correspondant (par exemple si on a deux pics qui se chevauchent correspondant a A ou G, mais certainement pas a C ni T, on peut mettre K). Le code pour les ambiguités se trouve ici : https://www.bioinformatics.org/sms/iupac.html
- Enregistrer
- Extraire la séquence en excluant le début et la fin si ceux-ci sont seulement constitués de NNN (attention ! si il y a des NNN dans la séquence il faut les garder):
    - Sélectioner la partie de la séquence a garder
    - Cliquer sur Extract (au dessus de la séquence, panneau du bas) > OK (changer le nom si besoin)

3. Assemblage eventuel des séquences F et R pour chaque région :

Une fois les séquences corrigées, si l'on a une séquence F et une séquence R pour une région donnée, il faut les assembler en une seule séquence :

- Selectioner les deux séquences
- Cliquer sur Align / Assemble (en haut) > Pairwise Align
- Cliquer « Automatically determine direction » puis « Ok » (les autres options par défaut devraient suffire mais on peut raffiner si l'alignment ne se fait pas bien)

➔ Un nouveau document est créé

- Sélectioner le nouveau document et vérifier que l'alignement a l'air normal
- Combiner les deux séquences pour avoir la séquence la plus longue possible avec le moins d'ambiguités possibles. Pour cela, deux manieres de faire sont possibles :
    - Soit utiliser le consensus : Editer le consensus si besoin (ou bien choisir un Threshold dans Display > Consensus > Options de maniere a ce que le consensus reflete les conflits eventuels entre les séquences).
    - Soit éditer une des deux séquences en completant les nucléotides manquants avec ceux de l'autre séquence
- Extraire la séquence finale avec un nouveau nom (par exemple en ajoutant « consensus FR » au nom originel :
    - Selectioner la séquence en cliquant sur son nom
    - Extract > [editer le nom] > OK

4. Blaster les séquences sur NCBI pour vérifier si ce n'est pas une contamination

- Aller sur https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
- Copier la séquence (seulement les nucléotides, pas le nom) et la coller dans le site internet blast
- Garder toutes les options de défaut (standard databases) et cliquer sur BLAST en bas de la page
- Si les résultats n'indiquent pas des plantes, ou bien si les plantes indiquées ne sont pas de la famille attendue (par exemple Meliaceae), il faudra se méfier car il est possible que la séquence soit issue d'une contamination et non pas de l'échantillon lui-meme. Ceci dit, s'il n'y a pas déja des séquences de la famille attendue dans Genbank, la famille n'apparaitra pas dans les résultats meme si la séquence provient bien de l'échantillon.
- Si on a une contamination averée (par exemple champignon) ce n'est pas la peine de continuer.

5. Alignement des régions avec l'alignement de référence :

- Importer les séquences de reference : cliquer sur File > Import > Files et sélectioner le fichier contenant les séquences puis cliquer sur Import

NB : Les séquences de référence peuvent etre sous forme d'une liste de séquences ou bien déja sous la forme d'un alignement

- Sélectioner le document importé (liste de séquences ou bien alignement) ainsi que les séquences finales corrigées générées a l'étape 3
- Aligner le tout : cliquer sur Align / Assemble > Multiple Align
  Sélectioner la methode favorite (a décider selon le cas, apres essais, en cas de doute utiliser l'option Geneious et cliquer « Automatically determine direction » puis « Ok »)

Alternative si les options d'alignement dans Geneious ne sont pas satisfaisantes :

- Créer un document contenant les séquences de référence ainsi que les séquences finales corrigées :
  - Sélectioner le document importé (liste de séquences ou bien alignement) ainsi que les séquences finales corrigées générées a l'étape 3
  - Export
- Utiliser MAFFT en ligne : https://mafft.cbrc.jp/alignment/server/

6. Vérification de l'alignement et combinaison des régions d 'un meme échantillon en une seule séquence

- Une fois l'alignement terminé, cliquer dessus pour le visualiser dans le panneau du bas
- Examiner si les séquences de notre échantillon a identifier ont l'air bien alignées
- Si on a plusieurs séquences pour un meme échantillon (par exemple région 1 et région 2), on les combine en une seule séquence :
  - Cliquer sur Allow Editing
  - On utilise la séquence la plus a gauche (par exemple région 1) comme point de départ pour combiner les séquences : on va y ajouter les autres séquences
  - Selectioner tous les nucléotides de la deuxieme séquence (par exemple région 2) : cliquer au début, faire défiler jusque a la fin, appuyer sur majuscule et puis cliquer en meme temps a la fin de la séquence, si c 'est bien sélectioné, tout doit etre en vert
  - Copier les nucléotides (ctrl-C) puis les coller (ctrl-V) dans la premiere séquence au meme niveau de l'alignement (cela va créer des gaps ---- entre les deux séquences/régions)
  - Faire de meme avec une troisieme séquence qui sera copiée dans la premiere séquence, etc selon le nombre de séquences a combiner
- Enregistrer (Save)
- Une fois la séquence combinée terminée, on peut supprimer les autres séquences :
  - Clic droit (2 doigts) sur le nom puis cliquer sur Delete Selected Bases

- Optionel pour protocole final : vérifier que l'alignement est bien et éventiellement enlever des régions non désirées, par exple le début et la fin
  - Sélectioner la région a enlever (attention a sélectioner sur toute la hauteur de l'alignement = toutes les séquences !)
  - Appuyer sur la touche Delete ou ←
  - Enregistrer sous un nouveau nom : Cliquer sur File > Save As

7. Arbre phylogénétique et identification de l'échantillon

- Sélectionner le nouvel alignment fait ci-dessus contenant la sequence de notre échantillon (par exemple « Nucleotide alignment trimmed »)
- Cliquer sur Tree dans la barre du haut
- Changer les options si besoin (idéalement il faut utiliser une autre methode, pas le Neighbor Joining ni le UPGMA. Les autres méthodes, par exemple maximum likelihood sont disponibles via des « plug-in » a installer. Si on n'utilise pas Geneious, il faut faire l'arbre séparément : exporter l'alignement et puis l'importer dans le programe désiré – voire ci-dessous)
- Sélectionner l'Outgroup si besoin (Schmardea)
- Cliquer sur « Resample tree » et sélectioner 100 bootstrap replicates (minimum, 1000 c'est mieux)
- Cocher la case Create Consensus Tree
- Cliquer sur OK

Alternative : exporter l'alignement et faire l'arbre dans un programme séparé comme IQ-TREE :

- Sélectioner l'alignment
- Cliquer sur File > Export > Documents (choisir l'endroit ou on veut l'exporter)
- Choisir le format fasta et cliquer sur Export
- Décocher « Include sequence description » et garder les autres options d'exportation par défaut > Ok
- Aller sur le site internet de IQ-Tree : http://iqtree.cibiv.univie.ac.at/
- Importer l'alignement dans le serveur : Alignment file > Browse
- Garder les options d'analyse par défaut sauf :
    - Cocher DNA
    - Cocher No pour le SH-aLRT branch test
    - Indiquer son email
- Cliquer sur Submit Job
- Une fois le résultat obtenu, on peut le visualiser en utilisant FigTree

4. Towards a deployment of timber DNA barcoding in the focus countries – synthesis

Three main DNA barcoding approaches were discussed between partners and stakeholders during this project:

- A cheap and quick approach using a few DNA barcodes that can be amplified from DNA using PCR and sequenced using Sanger sequencing. This is the approach for which we identified 9 potential new DNA barcodes in this project.

- A more expensive and time-consuming approach where the DNA is submitted to shotgun sequencing using high throughput sequencing techniques relying on the Illumina sequencing technology, possibly preceded by a capture of hundreds of target genes via hybridisation with existing baits. This is the approach that was used to create the DNA reference dataset of the focus species.

- An intermediate approach where the DNA is submitted to random or targeted sequencing using the portable sequencing technology developed by Oxford Nanopore Technology (minION). This approach was discussed but not tested.

Together with D. Bourobou, S. Bellot and C. Quintero-Berns visited the IRET lab in Libreville, Gabon. This is a tropical ecology lab led by Dr Etienne Okomo Okue where people currently work on tropical diseases and great apes genetics. They have 3 small rooms with all the equipment needed to do DNA extraction, PCR, RT-PCR, and visualisation of PCR products and cycle sequencing. The PCR products are then sent for Sanger sequencing to a lab at 26 km from there (easier than having to rely on DHL to send to other country). An ambition is to acquire their own sequencing machine (a Japanese equipment that is convenient in contexts with electricity access instability).

IRET Lab building in Libreville (left) and inside the IRET lab (right) © C. Quintero-Berns, RBG Kew

Together with D. Bourobou, S. Bellot and C. Quintero-Berns also visited the WAVE lab, also in Gabon, in the vicinity of Libreville (directed by Dr Claude Gnacadja). The lab was built with funds from the Melissa Gates foundation in 2020 to support a large international project on cassava farming improvement. The lab has 5 relatively small rooms, each dedicated to a step, from sample entry and storage to DNA extraction (mainly CTAB), to PCR, to PCR products visualisation. There is a liquid nitrogen grinder. There is a RT-PCR machine to study SNPs in real time. The lab may be used for other projects.

The lab is visited by many people, mainly PhD students from Gabon or neighbouring countries involved in the cassava project. Around 20 people tend to be involved in the activities of the lab in

a way or another, even though most are not on site as this is not their main job (there is no lab technician that would be there all the time). Other labs are developing at the same site (soil lab, agrochemistry lab). The lab has a bioinformatic unit to process data from the cassava projects.

Inside the WAVE Lab © C. Quintero-Berns, RBG Kew

As revealed by these visits and by discussions with D. Bourobou and the lab directors, the lab capacity and expertise necessary to develop plant (including timber) DNA barcoding projects are high in Gabon. Obtaining chemicals is not an issue although that may depend on providers. They have custom fee exonerations. An upscale of lab infrastructures in Gabon, and of the number of people to manage them, might be needed if projects with thousands of samples are developed given the relatively small size and already high intensity of use of the existing labs and equipment. **If such upscaling was to happen in the context of developing a timber DNA barcoding activity, the following additional equipment may be required:** steel mill grinding jars and possibly additional mills, a centrifuge with adaptors for 96-well plates, pipette sets, a fluorometer, magnetic racks, a pH meter, a fridge-freezer and an electricity generator.

Information provided by Dr J. Lisingo shows that so far**, there is no functional genetics lab in the University of Kisangani (DRC),** but multiple researchers there, and especially Dr J. Lisingo, are very keen to develop one that could both support research projects on plant genetics (including timber DNA barcoding) and be used for education purposes.

J. Lisingo and colleagues at the university of Kisangani have compiled the following list of the equipment required to build such a lab and preliminary estimates of the necessary funding:

Equipment	**Quantity**	**Cost**
Autoclave de table	1	1,300.00 GBP
Centrifugeuse	1	328.00 GBP
Centrifugeuse	1	2,300.00 GBP
Centrifugeuse	1	2,400.00 GBP
Hotte d'aspiration	1	2,600.00 GBP
Incubateur (etuve)	1	1,500.00 GBP
Distillateur	1	2,500.00 GBP
Réfrigérateur	1	- GBP
Congélateur -40°C	1	1,950.00 GBP
Agitateur orbital ou magnétique	1	1,850.00 GBP
Agitateur vortex	2	210.00 GBP
Balance	2	352.00 GBP
Bain-marie universel avec couvercle	1	1,030.00 GBP
machine à glace	1	1,470.00 GBP
Micropipettes à volume unique	6	900.00 GBP
Micropipettes à volume variable	5	1,100.00 GBP

Micropipettes électroniques	2	1,300.00 GBP
Micropipettes électroniques	1	1,000.00 GBP
Pissettes	10	150.00 GBP
Béchers forme basse	10	150.00 GBP
Béchers de mesure	5	75.00 GBP
Fioles coniques	5	75.00 GBP
Tubes à essai plastique	10	- GBP
Portoirs pour tube	5	- GBP
Portoirs pour microtubes	2	- GBP
portoirs pour microtubes	5	- GBP
Portoirs universels	5	- GBP
Portoirs pour tubes centrifuger	2	- GBP
Boites pour tube centrifuger	2	- GBP
Flacons de laboratoire transparents	8	- GBP
Bouteilles à col large	5	- GBP
Bouteilles à col étroit	5	- GBP
Entonnoirs	5	- GBP
Pot à echantillons gradués	100	756.00 GBP
Bonbonnes avec robinet	5	75.00 GBP
Cristallisoirs avec bec	10	75.00 GBP
cuillère à échantillon	20	150.00 GBP
lunette de sécurité	5	- GBP
Gant de protection	2	- GBP
Blouse de laboratoire	10	- GBP
Support en metal pour sac à déchet	4	- GBP
Thermomètre	2	- GBP
minuteur du temps	2	- GBP
mortier porcelaine	2	- GBP
pilon procelaine	2	- GBP
Tube à centrifuger	2	- GBP
Boîte cryogénique	20	- GBP
Glacière	2	- GBP
Seau à glace	2	- GBP
portoir de micropipettes	2	- GBP
Micropipette multicanaux	3	- GBP
Total (to be finalised)		**25,596.00 GBP**

In the current context, it appears from the observations made above that Meliaceae timber DNA barcoding using PCR and Sanger sequencing of a few DNA barcodes could already be trialled in Gabon if using the facilities of the IRET and/or WAVE labs as well as the Sanger sequencing arrangements in place. A more routine analysis of thousands of samples may require expanding the lab infrastructure in Gabon. For the DRC, a reasonable investment in terms of infrastructure could enable similar trials as well as training activities. In both countries, the routine test of thousands of timber samples will require dedicated staff and the corresponding funding to recruit and train them.

Stakeholder consultations (Report 1) have revealed that the most interesting supply chain stage for timber identification tests would be the export points. Identification campaigns in forest concessions are also of interest. The type of tissue available for testing at export points will mainly include heartwood, but it will also often include sapwood, while both wood types and potentially even leaf material will be available for samples taken in forest concessions. Our tests on heartwood and sapwood show that PCR and Sanger sequencing of samples taken at these stages of the supply chain can work, but that tests may fail. Reasons for failure were difficult to statistically identify given the limited time frame of this project, which limited our ability to perform multifactorial tests accounting for all possible factors on enough replicates. However, essential factors emerged as likely to increase chances of success: 1) the wood sample must be grinded in a very fine powder, 2) if performed in small tubes as in our protocols, many extractions (at least 6-10, corresponding to ca. 2-3 g of wood tissue, ideally more) are likely to be required to provide a

sufficiently high amount of DNA, 3) a pre-extraction cleaning step can help increase DNA purity, which is likely to increase success rates; 4) although multiple extracts may then be pooled, re-concentrating the DNA following pooling is unlikely to help for its use in a PCR, and it may even be deleterious by increasing the concentration in other molecules interfering with the PCR, 5) DNA in concentrations as low as 0.48 ng/uL can be amplified by PCR and yield sequences enabling sample identification, as long as there are fragments longer than the target DNA barcodes, and 6) when there is a sufficient amount of DNA but the PCR fails, and especially if the DNA has a dark colour, it may be helpful to clean it using magnetic beads (the latter was not trialled here but previous trials on herbarium DNA shows that it can improve downstream success when the DNA appears impure based on a dark colour or viscous texture). All of the above appears to be feasible with only a few adjustments to the equipment of the labs in Gabon and DRC mentioned above.

The minION approach could be an interesting alternative to sequence individual DNA barcodes where labs cannot be set-up for PCR and Sanger sequencing, for instance in DRC. However, it would still require the DNA to be longer than the DNA barcodes and it may require higher amounts of input DNA. Moreover, its costs and potential logistical issues with obtaining and conserving the necessary reagents are so that further examination and testing would need to be done before it can be recommended for deployment.

High throughput sequencing using short reads may be successful on samples with highly degraded DNA where PCR and Sanger sequencing has failed, as found here with the unknown glued heartwood samples. Although this method is unlikely to be easy to implement in the focus countries without a significantly higher investment, it may not be impossible in the medium term, as it was discussed with stakeholders from Gabon that having more projects using high throughput sequencing could contribute to building a case for acquiring a high throughput sequencing facility in country. Since sapwood may be available at most supply chain points in the exporting countries, the need for high throughput sequencing of hundreds of genetic regions (either after target enrichment of nuclear regions or just randomly so that regions recovered would mainly be ITS and the plastome) using Illumina or even the minION may not be high. High throughput sequencing may nevertheless be useful for importing countries that have the facilities already in place, and for tests on important samples on which the PCR and Sanger sequencing of a few barcodes has failed. Such tests could also be performed by the focus countries if partnerships with foreign labs can be put in place.

Regardless the approach, a common bottleneck in the efficiency of timber sample DNA testing is the ability to reduce samples in a fine powder, which can take a lot of time. A wide deployment of timber DNA barcoding will therefore require the appropriate equipment mentioned above as well as enough dedicated staff.

Finally, as also discussed in Report 1, timber species that need monitoring in Gabon and DRC include dozens of species not studied in this project, notably in the Fabaceae family. Lists of species to monitor have been compiled for both countries, and DNA reference datasets for these species and their close relatives and look-alikes will be required if DNA barcoding is to be deployed with the purpose of monitoring the illegal timber trade in Gabon and DRC.

In conclusion, this pilot project has reached its objectives of developing a DNA barcoding toolkit for the focus species in the focus countries, including a reference dataset, DNA extraction protocols and DNA barcodes. Nevertheless, larger projects will be necessary to go from this proof of concept to the efficient implementation of DNA barcoding for the monitoring of timber in the forest concessions and/or export points by the relevant control authorities. Combining insights from our stakeholder consultations and laboratory and computational work, we identify the following as main

points of action for future projects to enable the implementation of a DNA-based control of timber in the focus countries:

1) A large test of Meliaceae DNA barcoding in Gabon labs, involving all relevant stakeholders to identify and overcome any overlooked logistical or socio-political challenges and to produce the legal framework in which tests could ultimately be performed so that they can be used by law enforcement authorities. This project would require buying the equipment mentioned above for Gabon labs and leveraging the capacity of existing labs.

2) The testing of the performance of the new DNA barcodes as well as of traditional barcodes such as ITS in all other species of interest for timber monitoring in the focus countries, and based on these results, the assembly of a multi-individual reference dataset for the most relevant barcodes for all these species.

3) A strengthening of botanical training and DNA barcoding data generation and analysis training in the focus countries, for instance by developing dedicated university courses and standardising existing resources.

4) Further investigating possibilities for plant DNA PCR and Sanger sequencing in DRC and developing a dedicated plant genetics and DNA barcoding lab, for instance at the University of Kisangani, where interest is high.

5) Further streamlining of the DNA extraction protocols to increase success rate so that testing can be more cost- and time-efficient.

5) References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215, 403–410.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (21 June 2024, date last accessed).

Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3, <https://CRAN.R-project.org/package=gridExtra>.

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., Botigué, L. R., Brewer, G., Carruthers, T., Clarkson, J. J., Cook, J., Cowan, R. S., Dodsworth, S., Epitawalage, N., Françoso, E., Gallego, B., Johnson, M. G., Kim, J. T., Leempoel, K., … Forest, F. (2022). A Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life. Systematic Biology, 71(2), 301–319. https://doi.org/10.1093/sysbio/syab035

Baker, W. J., Dodsworth, S., Forest, F., Graham, S. W., Johnson, M. G., McDonnell, A., Pokorny, L., Tate, J. A., Wicke, S., & Wickett, N. J. (2021). Exploring Angiosperms353: An open, community toolkit for collaborative phylogenomic research on flowering plants. American Journal of Botany, 108(7), 1059–1065. https://doi.org/10.1002/ajb2.1703

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. PeerJ, 2016(1). https://doi.org/10.7717/peerj.1660

Bouka, G. U. D., Doumenge, C., Ekué, M. R. M., Daïnou, K., Florence, J., Degen, B., Loumeto, J. J., McKey, D., & Hardy, O. J. (2022). Khaya revisited: Genetic markers and morphological analysis reveal six species in the widespread taxon K. anthotheca. Taxon, 71(4), 814–832. https://doi.org/10.1002/tax.12720

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., Biggs, N., Cowan, R. S., Davies, N. M. J., Dodsworth, S., Edwards, S. L., Eiserhardt, W. L., Epitawalage, N., Frisby, S., Grall, A., Kersey, P. J., Pokorny, L., Leitch, I. J., Forest, F., & Baker, W. J. (2019). Factors Affecting Targeted Sequencing of 353 Nuclear Genes From Herbarium Specimens Spanning the Diversity of Angiosperms. Frontiers in Plant Science, 10. https://doi.org/10.3389/fpls.2019.01102

Brown, J. W., Walker, J. F., & Smith, S. A. (2017). Phyx: Phylogenetic tools for unix. Bioinformatics, 33(12), 1886–1888. https://doi.org/10.1093/bioinformatics/btx063

Chernomor, O., Von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Systematic Biology, 65(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin, 19, 11–15.

Jiao, L., Lu, Y., He, T., Guo, J., & Yin, Y. (2020). DNA barcoding for wood identification: Global review of the last decade and future perspective. IAWA Journal, 41(4), 620–643. https://doi.org/10.1163/22941932-bja10041

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

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., Eiserhardt, W. L., Epitawalage, N., Forest, F., Kim, J. T., Leebens-Mack, J. H., Leitch, I. J., Maurin, O., Soltis, D. E., Soltis, P. S., Wong, G. K. S., Baker, W. J., & Wickett, N. J. (2019). A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. Systematic Biology, 68(4), 594–606. https://doi.org/10.1093/sysbio/syy086

Kalyaanamoorthy, S., Minh, B., Wong, T., Haeseler, A. von, & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods, 587–589.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Monthe, F. K., Duminil, J., Kasongo Yakusu, E., Beeckman, H., Bourland, N., Doucet, J. L., Sosef, M. S. M., & Hardy, O. J. (2018). The African timber tree Entandrophragma congoense (Pierre ex De Wild.) A.Chev. is

morphologically and genetically distinct from Entandrophragma angolense (Welw.) C.DC. Tree Genetics and Genomes, 14(5). https://doi.org/10.1007/s11295-018-1277-6
Monthe, F. K., Migliore, J., Duminil, J., Bouka, G., Demenou, B. B., Doumenge, C., Blanc-Jolivet, C., Ekué, M. R. M., & Hardy, O. J. (2019). Phylogenetic relationships in two African Cedreloideae tree genera (Meliaceae) reveal multiple rain/dry forest transitions. Perspectives in Plant Ecology, Evolution and Systematics, 37, 1–10. https://doi.org/10.1016/j.ppees.2019.01.002
Ortiz, E. M., Höwener, A., Shigita, G., Raza, M., Maurin, O., Zuntini, A., Forest, F., Baker, W. J., & Schaefer, H. (2024). A novel phylogenomics pipeline reveals complex patterns of reticulate evolution in Cucurbitales. BioRxiv. https://doi.org/10.1101/2023.10.27.564367
Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633
Posit team. (2024). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.
POWO. (2025, January 12). Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet; https://powo.science.kew.org/.
R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
Raza, M., Ortiz, E. M., Schwung, L., Shigita, G., & Schaefer, H. (2023). Resolving the phylogeny of Thladiantha (Cucurbitaceae) with three different target capture pipelines. BMC Ecology and Evolution, 23(1). https://doi.org/10.1186/s12862-023-02185-z
Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3(2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x
Slimp, M., Williams, L. D., Hale, H., & Johnson, M. G. (2021). On the potential of Angiosperms353 for population genomic studies. Applications in Plant Sciences, 9(7). https://doi.org/10.1002/aps3.11419
Slowikowski, K. (2024). ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2”. R package version 0.9.6, <https://CRAN.R-project.org/package=ggrepel>.
SRA. (2025). Sequence Read Archive (SRA) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2009 - [cited 2025 01 13]. Available from: https://www.ncbi.nlm.nih.gov/sra/.
Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X. X., & Rokas, A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biology, 18(12). https://doi.org/10.1371/journal.pbio.3001007
Sun, Y., Skinner, D. Z., Liang, G. H., & Hulbert, S. H. (1994). Phylogenetic analysis of Sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. Theoretical and Applied Genetics, 89, 26–32.
Taberlet, P., Gielly, L., Pautou, G., & Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Molecular Biology, 17(5), 1105–1109. https://doi.org/10.1007/BF00037152
Tumescheit, C., Firth, A. E., & Brown, K. (2022). ClAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. PeerJ. https://doi.org/10.7717/peerj.12983
Wang, L. G., Lam, T. T. Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C. W., Jones, B. R., Bradley, T., Zhu, H., Guan, Y., Jiang, Y., & Yu, G. (2020). Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Molecular Biology and Evolution, 37(2), 599–603. https://doi.org/10.1093/molbev/msz240
White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In M. A. Innis, D. H. Gelfand, J. J. Sninsky, & T. J. White (Eds.), PCR protocols - a guide to methods and applications (pp. 315–322). Academic Press.
Wickham, H. (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12). http://www.jstatsoft.org/
Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software , 40(1). http://www.jstatsoft.org/
Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag.
Wickham, H., François, R., Henry, L., Muller, K., & Vaughan, D. (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, <https://CRAN.R-project.org/package=dplyr>.
Wickham, H., Vaughan, D., & Girlich, M. (2024). tidyr: Tidy Messy Data. R package version 1.3.1, <https://CRAN.R-project.org/package=tidyr>.

Willson, J., Roddur, M. S., Liu, B., Zaharias, P., & Warnow, T. (2022). DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. Systematic Biology, 71(3), 610–629. https://doi.org/10.1093/sysbio/syab070

Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., Dai, Z., Lam, T. T., Guan, Y., & Yu, G. (2022). Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. IMeta, 1(4). https://doi.org/10.1002/imt2.56

Yu, G. (2022). Data Integration, Manipulation and Visualization of Phylogenetic Trees (1st edition). Chapman and Hall/CRC.

Zhang, C., & Mirarab, S. (2022). Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. Molecular Biology and Evolution, 39(12). https://doi.org/10.1093/molbev/msac215

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics, 19. https://doi.org/10.1186/s12859-018-2129-y

Zhang, C., Zhao, Y., Braun, E. L., & Mirarab, S. (2021). TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. Methods in Ecology and Evolution, 12(11), 2145–2158. https://doi.org/10.1111/2041-210X.13696