New Phytologist

***New Phytologist* Supporting Information**

Article title: **Admixture may be extensive among hyperdominant Amazon rainforest tree species**
Authors: Drew A. Larson, Oscar M. Vargas, Alberto Vicentini and Christopher W. Dick
Article acceptance date: 04 August 2021

The following Supporting Information is available for this article:

Fig. S1 Schematic of the process for making redeterminations based on all available evidence.

Fig. S2 The preliminary phylogeny of the Parvifolia clade, estimated from a supermatrix of intron and exon target capture data.

Fig. S3 Results of several *Structure* analyses with alternative values of K.

Fig. S4 Results of *Structure* analyses using a SNP dataset for the clade that included *E. coriacea*, *E. wachenheimii*, *E. sagotiana*, *E. truncata*, and *E. parviflora*.

Fig. S5 Three-dimensional scatterplot for a genetic principal component analysis showing geographical structuring in samples of *E. coriacea*.

Fig. S6 The Parvifolia phylogeny without reduced representation, produced using a supermatrix of intron and exon target capture data.

Fig. S7 Comparison of the reduced-representation Parvifolia phylogenies recovered with two supermatrices.

Fig. S8 Boxplots overlayed with dot plots showing the day of the year that collections were made for specimens at the New York Botanical Garden Herbarium.

Table S1 Voucher and accession information for samples used in the study.

Table S2 Summary statistics for all SNP datasets and estimated probability of the data for all *Structure* analyses for differing values of K.

Table S3 Summary of all rooted triplet tests conducted.

Table S4 Results of tree searches and likelihood recalculations for the Parvifolia phylogenies, ordered by increasing AIC score.

Table S5 Summaries of morphological and ecological traits of species inferred to engage in admixture.

Table S6 Information for flowering specimens collected in Amazonas, Brazil and housed at the New York Botanical Garden Herbarium.

Methods S1 Paralog filtering and alignment.

Methods S2 Genotyping and SNP dataset analyses.

Methods S3 Redetermination of individuals based on all available evidence.

Methods S4 Parvifolia phylogeny supermatrix construction and phylogeny estimation.

Notes S1 Note on species of the Parvifolia clade at Reserve 1501.

Notes S2 Notes on sampling at Reserve 1501 and prioritization of morphological intermediates.

Fig. S1 Schematic of the process for making redeterminations based on all available evidence. The workflow was designed to allow specimen determinations to be updated to reflect new genetic evidence, as well as any remaining uncertainty, while minimizing any circularities that could result from using redeterminations made for some samples to update those of others. ***When considering whether a sample is sister to the rest, we only take into account other samples whose morphological determination is concordant with all available genetic evidence.

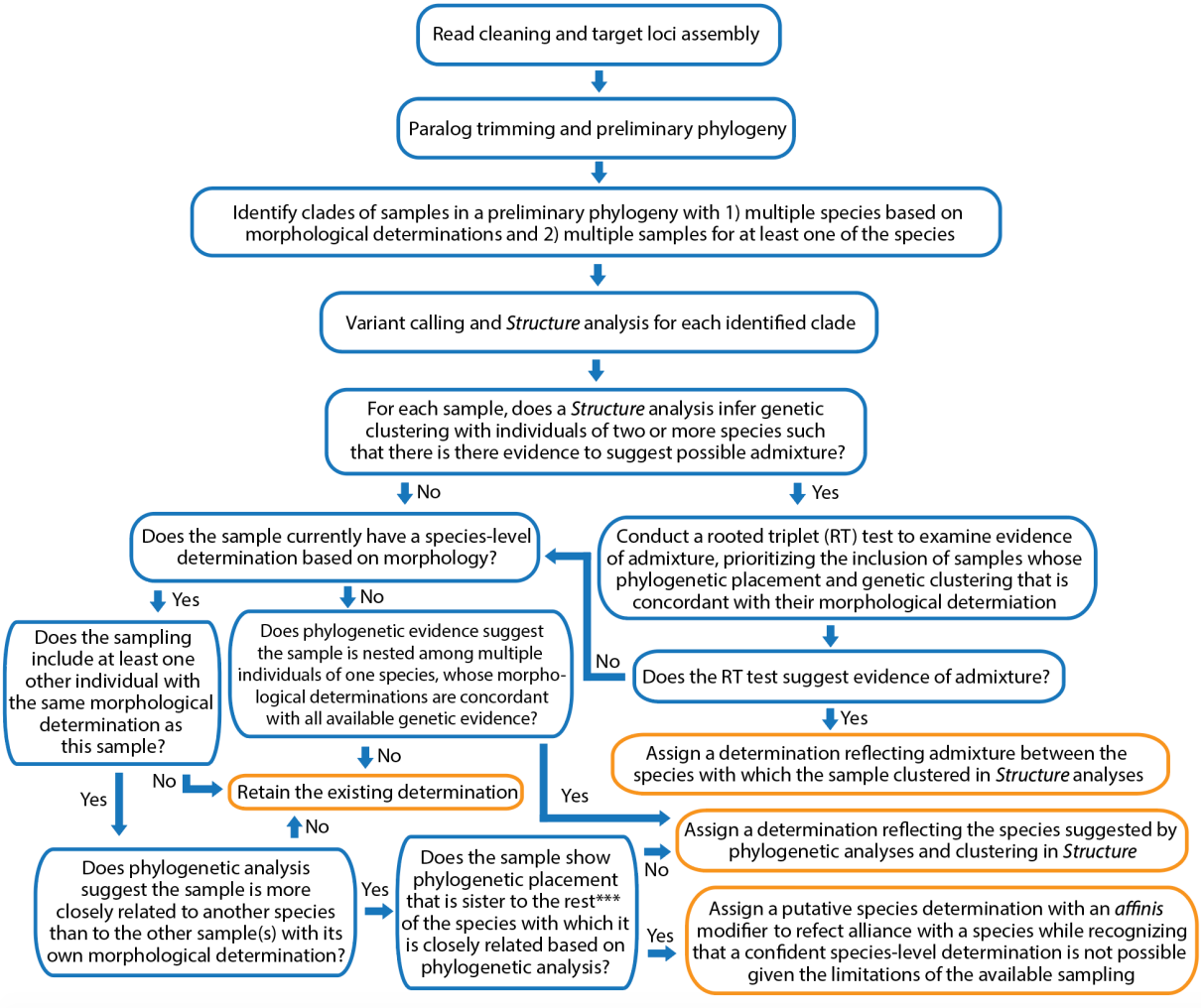

Fig. S2 The preliminary phylogeny of the Parvifolia clade estimated from a supermatrix of intron and exon target capture data. The phylogeny is shown as a cladogram and was generated with data from all 240 individuals, rooted on the genus *Napoleonaea* P.Beauv., then trimmed to include only individuals within the Parvifolia clade based on accepted taxonomy. Branch labels indicate RAxML rapid bootstrap support values. Tip labels are the accession codes used to represent each individual in all analysis files with species determinations in parentheses. In cases where a redetermination was made based on genetic evidence, the most recent morphological determination is noted on the left and the redetermination is on the right.


EsbrL733 (Eschweilera bracteosa > Eschweilera truncata)
EstrL711 (Eschweilera truncata)
EswaL839 (Eschweilera wachenheimii > Eschweilera truncata)
EstrL802 (Eschweilera truncata)
EstrL793 (Eschweilera truncata)
EstrL772 (Eschweilera truncata)
EstrL838 (Eschweilera truncata)
EstrL698 (Eschweilera truncata)
EsteL690 (Eschweilera tessmannii > Eschweilera truncata)
EsroL620 (Eschweilera truncata)
EswaL686 (Eschweilera wachenheimii)
EswaL832 (Eschweilera wachenheimii)
EswaL779 (Eschweilera wachenheimii)
EswaL695 (Eschweilera wachenheimii)
EscoL796 (Eschweilera coriacea > Eschweilera coriacea × wachenheimii)
EscoL824 (Eschweilera coriacea [noted possible hybrid] > Eschweilera coriacea × wachenheimii)
EstrL882 (Eschweilera truncata > Eschweilera parviflora × wachenheimii)
EstrL891 (Eschweilera truncata > Eschweilera parviflora × wachenheimii)
EspaL068 (Eschweilera parviflora)
EsmiL332 (Eschweilera micrantha > Eschweilera aff. parviflora)
EscoL771 (Eschweilera coriacea)
EscoL834 (Eschweilera coriacea)
EscoL780 (Eschweilera coriacea)
EsspL785 (Eschweilera sp. > Eschweilera coriacea)
EscoL770 (Eschweilera coriacea)
EsspL781 (Eschweilera sp. > Eschweilera coriacea)
EscoL010 (Eschweilera coriacea)
EscoL954 (Eschweilera coriacea)
EscoL938 (Eschweilera coriacea)
EscoL932 (Eschweilera coriacea)
EscoL301 (Eschweilera coriacea)
EscoL545 (Eschweilera roseocalyx > Eschweilera coriacea)
EswaL271 (Eschweilera wachenheimii > Eschweilera aff. sagotiana)
EsmiL817 (Eschweilera micrantha > Eschweilera aff. sagotiana)
EssaL335 (Eschweilera sagotiana)
EspeL815 (Eschweilera pedicellata)
EspeL819 (Eschweilera pedicellata)
EspeL809 (Eschweilera pedicellata)
EspeL812 (Eschweilera pedicellata)
EspeL806 (Eschweilera pedicellata)
EspeL017 (Eschweilera pedicellata)
EsovL015 (Eschweilera ovata)
EsalL857 (Eschweilera albiflora)
EsalL846 (Eschweilera albiflora)
EsalL914 (Eschweilera albiflora)
EspeL866 (Eschweilera pedicellata > Eschweilera albiflora)
EspsL792 (Eschweilera pseudodecolorans)
EspsL799 (Eschweilera pseudodecolorans)
EspsL715 (Eschweilera pseudodecolorans)
EsmiL807 (Eschweilera micrantha > Eschweilera pseudodecolorans)
EspsL818 (Eschweilera pseudodecolorans)
EsdeL070 (Eschweilera decolorans)
EsgrL066 (Eschweilera grandiflora)
EssuL020 (Eschweilera subglandulosa)
EscyL797 (Eschweilera cyathiformis)
EscyL820 (Eschweilera cyathiformis)
LechL653 (Lecythis chartacea > Eschweilera cyathiformis)
EscyL813 (Eschweilera cyathiformis)
EscyL822 (Eschweilera cyathiformis)
EsrhL578 (Eschweilera rhododendrifolia)
EsatL833 (Eschweilera atropetiolata)
EsatL643 (Eschweilera atropetiolata)
EsatL831 (Eschweilera atropetiolata)
EsatL791 (Eschweilera atropetiolata)
EsatL716 (Eschweilera atropetiolata)
EsalL245 (Eschweilera alata)
EslaL810 (Eschweilera laevicarpa)
EslaL671 (Eschweilera laevicarpa)
EslaL827 (Eschweilera laevicarpa)
EsspL756 (Eschweilera sp. nov. > Eschweilera laevicarpa)
EslaL783 (Eschweilera laevicarpa)
EslaL814 (Eschweilera laevicarpa)
EslaL013 (Eschweilera laevicarpa)
EscoL885 (Eschweilera coriacea > Eschweilera aff. laevicarpa)
EsbrL778 (Eschweilera bracteosa)
EsbrL816 (Eschweilera bracteosa)
EsbrL794 (Eschweilera bracteosa)
EsbrL800 (Eschweilera bracteosa)
EscoL803 (Eschweilera collina)
EscoL828 (Eschweilera collina)
EscoL670 (Eschweilera collina)
EscoL825 (Eschweilera collina)
EscoL241 (Eschweilera collina)
EsroL664 (Eschweilera romeu-cardosoi)
EsmiL707 (Eschweilera micrantha > Eschweilera aff. romeu-cardosoi)
EsamL886 (Eschweilera amazoniciformis > Eschweilera aff. romeu-cardosoi)
EsmiL821 (Eschweilera micrantha)
EsmiL823 (Eschweilera micrantha)
EspaL386 (Eschweilera parvifolia)
EspaL868 (Eschweilera parvifolia)
EsteL936 (Eschweilera tessmannii > Eschweilera aff. parvifolia)
EsteL704 (Eschweilera tessmannii)
EsraL805 (Eschweilera rankiniae)
EsraL830 (Eschweilera rankiniae)
EsraL835 (Eschweilera rankiniae)
EsraL826 (Eschweilera rankiniae)
EscaL836 (Eschweilera carinata)
EsroL539 (Eschweilera rotundicarpa)
EscoL535 (Eschweilera correae)
EsspL532 (Eschweilera sp. nov.)
EsbiL047 (Eschweilera biflava)
EsdoL543 (Eschweilera donosoensis)
EsspL533 (Eschweilera sp. nov.)
EsjuL040 (Eschweilera juruensis)
EsteL302 (Eschweilera sp.)
EspiLA07 (Eschweilera pittieri)
EspiL529 (Eschweilera pittieri)
EsruL940 (Eschweilera rufifolia)
EschL261 (Eschweilera chartaceifolia)

Fig. S3 Results of *Structure* analyses for alternative values of K when the run resulting in the best estimated probability of the data was not selected as the optimal K due to *a prior* taxonomic information. Each individual is labeled with a unique code used throughout all analyses and asterisks indicate samples from focal species collected at Reserve 1501. A) *Structure* results for the clade that included *E. romeu-cardosoi*, *E. carinata*, *E. micrantha, E. parvifolia*, E. *rankiniae*, and *E. tessmannii*. The estimated ln probability of the data was -5156.6, -5047.0, -5084.1, and -5171.0 for K=3, 4, 5, and 6 respectively. The optimal K was determined to be 6 because for lower values of K, the addition of more clusters tended to result in clustering that was increasingly concordant with the accepted taxonomy of Parvifolia clade based on previous morphological analyses. B) *Structure* results for the clade that included *E. atropetiolata*, *E. cyathiformis*, and *E. rhododendrifolia*. The estimated ln probability of the data was -3510.8.0, -3559.0, and -3483.7 for K=2, 3, and 4 respectively. The optimal K was determined to be K=3 because, though K=4 had a better scoring probability of the data, no individual was inferred to have more than 0.4% ancestry corresponding to the fourth cluster and K=3 was concordant with our *a priori* expectation given the current taxonomy of the Parvifolia clade.

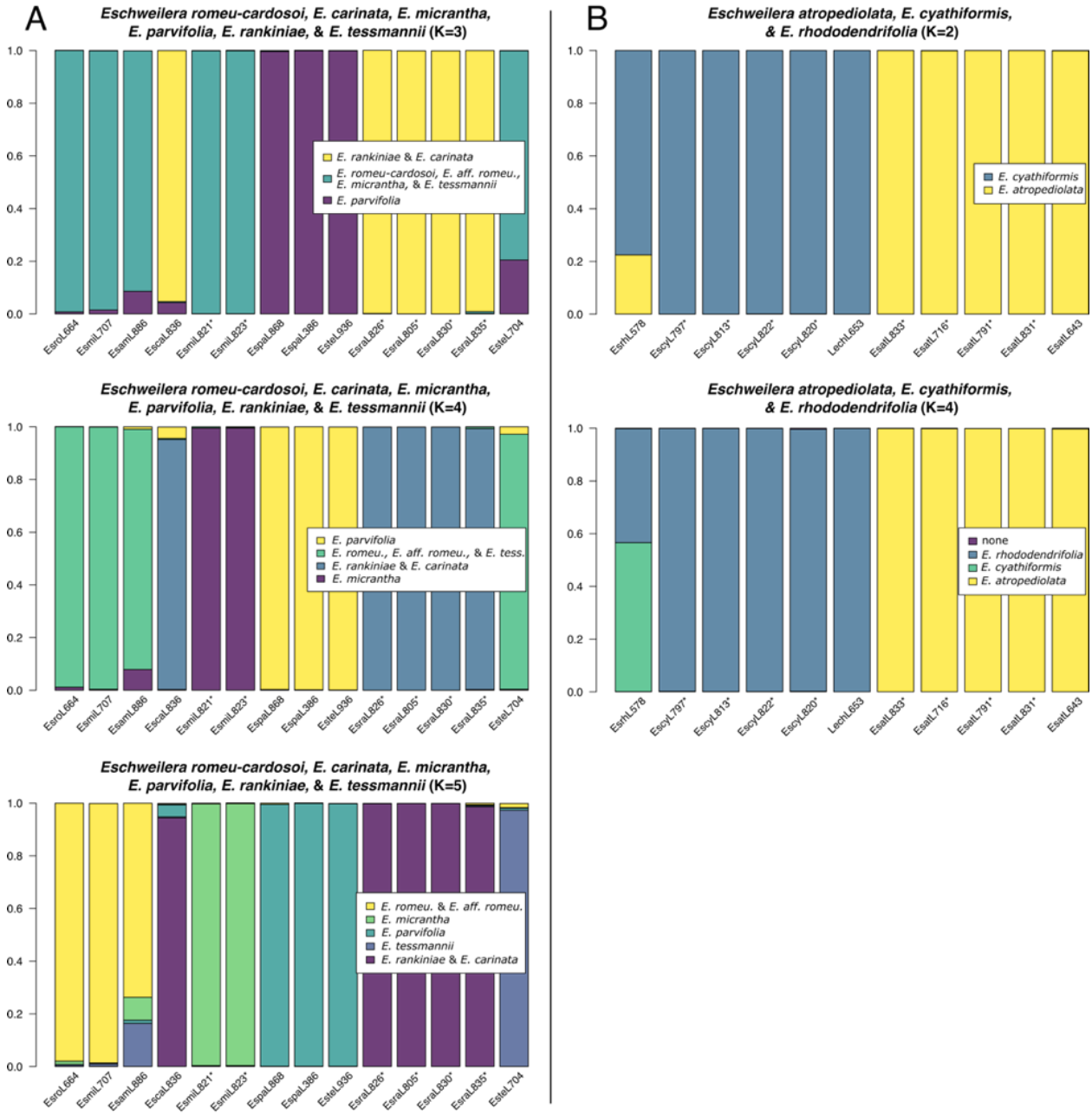
A
Eschweilera romeu-cardosoi, E. carinata, E. micrantha, E. parvifolia, E. rankiniae, & E. tessmannii (K=3)
E. rankiniae & E. carinata
E. romeu-cardosoi, E. aff. romeu., E. micrantha, & E. tessmannii
E. parvifolia
Eschweilera romeu-cardosoi, E. carinata, E. micrantha, E. parvifolia, E. rankiniae, & E. tessmannii (K=4)
E. parvifolia
E. romeu., E. aff. romeu., & E. tess.
E. rankiniae & E. carinata
E. micrantha
Eschweilera romeu-cardosoi, E. carinata, E. micrantha, E. parvifolia, E. rankiniae, & E. tessmannii (K=5)
E. romeu. & E. aff. romeu.
E. micrantha
E. parvifolia
E. tessmannii
E. rankiniae & E. carinata
EsroL664
EsmiL707
EsmiL686
EscaL636
EsmiL821*
EsmiL823*
EspaL868
EspaL386
EsteL936
EsraL826*
EsraL805*
EsraL830*
EsraL835*
EsteL704
B
Eschweilera atropediolata, E. cyathiformis, & E. rhododendrifolia (K=2)
E. cyathiformis
E. atropediolata
Eschweilera atropediolata, E. cyathiformis, & E. rhododendrifolia (K=4)
none
E. rhododendrifolia
E. cyathiformis
E. atropediolata
EsrhL578
EscyL797*
EscyL813*
EscyL822*
EscyL820*
LechL653
EsatL833*
EsatL716*
EsatL791*
EsatL831*
EsatL643

Fig. S4 Results of *Structure* analyses using a SNP dataset for the clade that included *E. coriacea*, *E. wachenheimii*, *E. sagotiana*, *E. truncata*, and *E. parviflora*. Each individual is labeled with a unique code used throughout all analyses and asterisks indicate samples from focal species collected at Reserve 1501. The estimated probability of the data was -11138.1 and -11061.8 for K=6 (A) and 7 (B) respectively. While the run with K=7 had the better score, the run with K=6 resulted in three individuals having nearly complete inferred ancestry from a single cluster that corresponded to *E. sagotiana*.

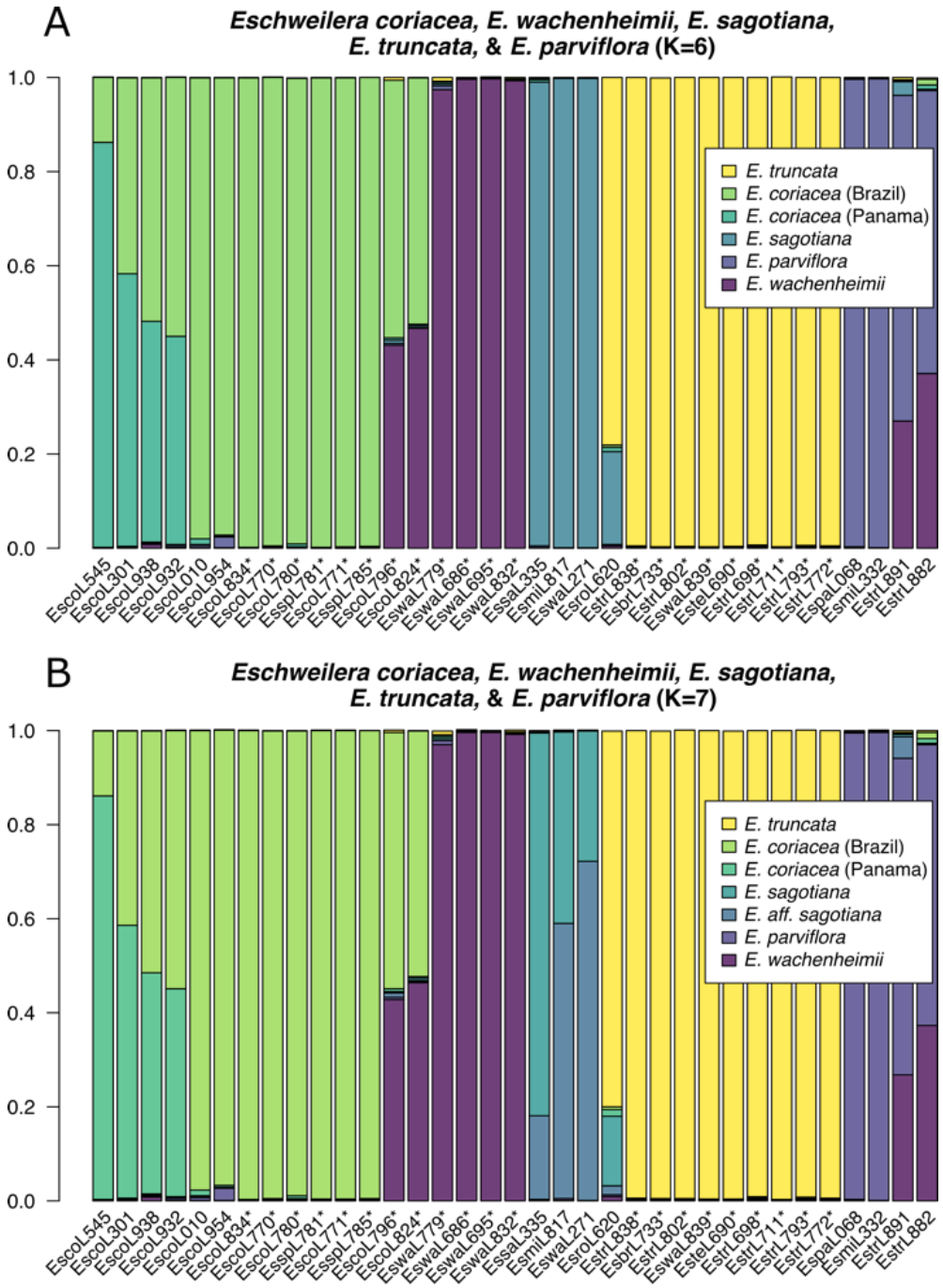

Fig. S5 Evidence of geographic structure among individuals of *E. coriacea*. The three-dimensional scatterplot shows the results of a genetic principal component analysis (PCA) that included all individuals of *E. coriacea* with no evidence of admixture. The x-, y-, and z-axes correspond to the first three principal components of the PCA, respectively. Each point represents an individual and colors correspond to the country in which it was collected.

Plot in file: Figure_S5.html

Fig. S6 The Parvifolia phylogeny without reduced representation, produced using a supermatrix of intron and exon target capture data. Branch labels and coloration indicate concordance (1, blue) and conflict (0, red) with the results using an exon-only supermatrix with the same taxa and data filtering strategy. The tip labels represent taxon identities after redeterminations based on all available evidence. The phylogeny was rooted on an outgroup consisting of five members of the Integrifolia clade of *Eschweilera*.

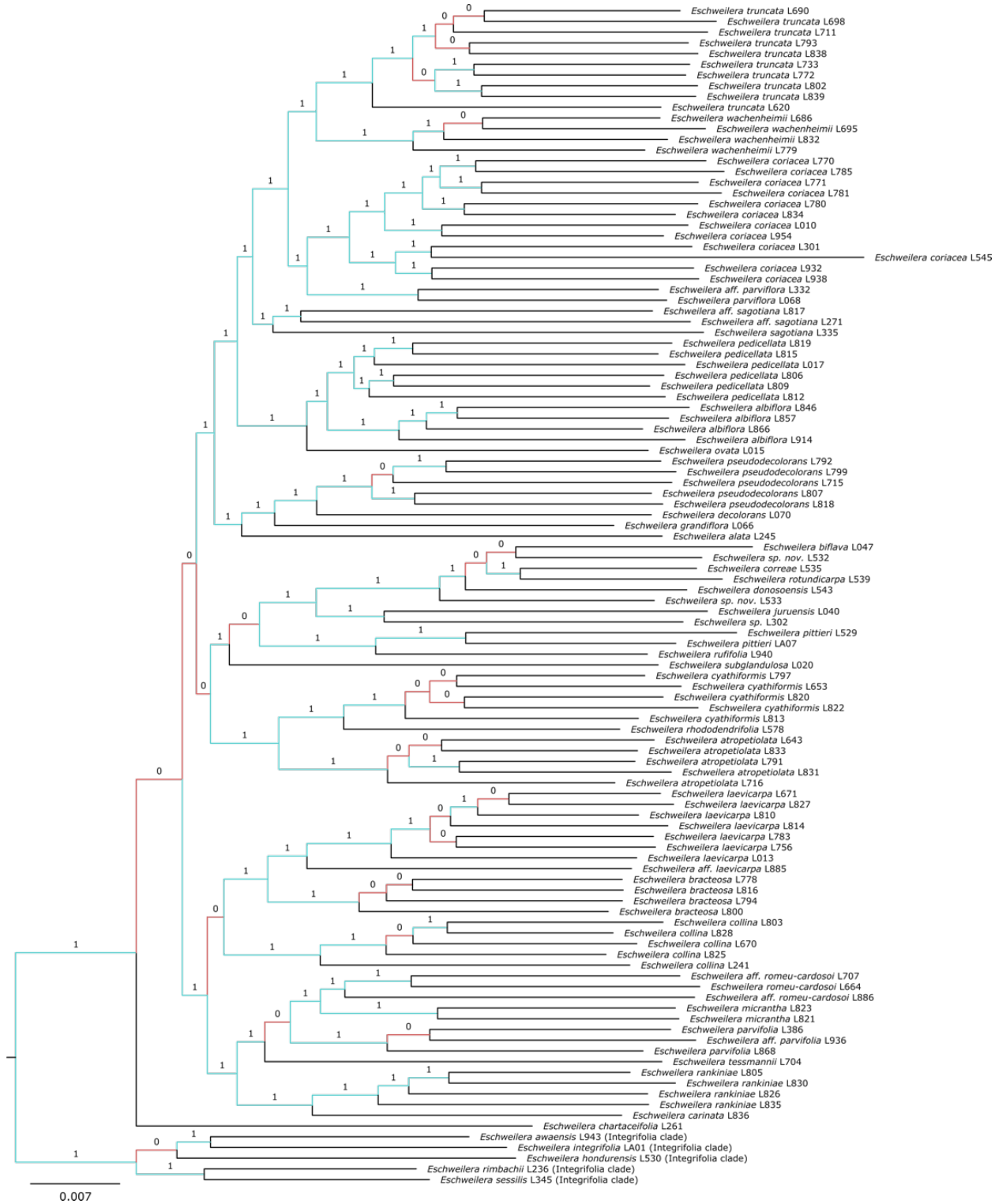
Eschweilera truncata L690
Eschweilera truncata L698
Eschweilera truncata L711
Eschweilera truncata L793
Eschweilera truncata L838
Eschweilera truncata L733
Eschweilera truncata L772
Eschweilera truncata L802
Eschweilera truncata L839
Eschweilera truncata L620
Eschweilera wachenheimii L686
Eschweilera wachenheimii L695
Eschweilera wachenheimii L832
Eschweilera wachenheimii L779
Eschweilera coriacea L770
Eschweilera coriacea L785
Eschweilera coriacea L771
Eschweilera coriacea L781
Eschweilera coriacea L780
Eschweilera coriacea L834
Eschweilera coriacea L010
Eschweilera coriacea L954
Eschweilera coriacea L301
Eschweilera coriacea L545
Eschweilera coriacea L932
Eschweilera coriacea L938
Eschweilera aff. parviflora L332
Eschweilera parviflora L068
Eschweilera aff. sagotiana L817
Eschweilera aff. sagotiana L271
Eschweilera sagotiana L335
Eschweilera pedicellata L819
Eschweilera pedicellata L815
Eschweilera pedicellata L017
Eschweilera pedicellata L806
Eschweilera pedicellata L809
Eschweilera pedicellata L812
Eschweilera albiflora L846
Eschweilera albiflora L857
Eschweilera albiflora L866
Eschweilera albiflora L914
Eschweilera ovata L015
Eschweilera pseudodecolorans L792
Eschweilera pseudodecolorans L799
Eschweilera pseudodecolorans L715
Eschweilera pseudodecolorans L807
Eschweilera pseudodecolorans L818
Eschweilera decolorans L070
Eschweilera grandiflora L066
Eschweilera alata L245
Eschweilera biflava L047
Eschweilera sp. nov. L532
Eschweilera correae L535
Eschweilera rotundicarpa L539
Eschweilera donosoensis L543
Eschweilera sp. nov. L533
Eschweilera juruensis L040
Eschweilera sp. L302
Eschweilera pittieri L529
Eschweilera pittieri LA07
Eschweilera rufifolia L940
Eschweilera subglandulosa L020
Eschweilera cyathiformis L797
Eschweilera cyathiformis L653
Eschweilera cyathiformis L820
Eschweilera cyathiformis L822
Eschweilera cyathiformis L813
Eschweilera rhododendrifolia L578
Eschweilera atropetiolata L643
Eschweilera atropetiolata L833
Eschweilera atropetiolata L791
Eschweilera atropetiolata L831
Eschweilera atropetiolata L716
Eschweilera laevicarpa L671
Eschweilera laevicarpa L827
Eschweilera laevicarpa L810
Eschweilera laevicarpa L814
Eschweilera laevicarpa L783
Eschweilera laevicarpa L756
Eschweilera laevicarpa L013
Eschweilera aff. laevicarpa L885
Eschweilera bracteosa L778
Eschweilera bracteosa L816
Eschweilera bracteosa L794
Eschweilera bracteosa L800
Eschweilera collina L803
Eschweilera collina L828
Eschweilera collina L670
Eschweilera collina L825
Eschweilera collina L241
Eschweilera aff. romeu-cardosoi L707
Eschweilera romeu-cardosoi L664
Eschweilera aff. romeu-cardosoi L886
Eschweilera micrantha L823
Eschweilera micrantha L821
Eschweilera parvifolia L386
Eschweilera aff. parvifolia L936
Eschweilera parvifolia L868
Eschweilera tessmannii L704
Eschweilera rankiniae L805
Eschweilera rankiniae L830
Eschweilera rankiniae L826
Eschweilera rankiniae L835
Eschweilera carinata L836
Eschweilera chartaceifolia L261
Eschweilera awaensis L943 (Integrifolia clade)
Eschweilera integrifolia LA01 (Integrifolia clade)
Eschweilera hondurensis L530 (Integrifolia clade)
Eschweilera rimbachii L236 (Integrifolia clade)
Eschweilera sessilis L345 (Integrifolia clade)
0.007

Fig. S7 Comparison of the reduced-representation Parvifolia phylogenies recovered with the two supermatrices. The Parvifolia phylogeny constructed using intron and exon target capture data is on the left and the exon-only Parvifolia phylogeny is on the right. Both phylogenies are presented as cladograms and the blue lines connecting tip labels indicate the same species in either tree.

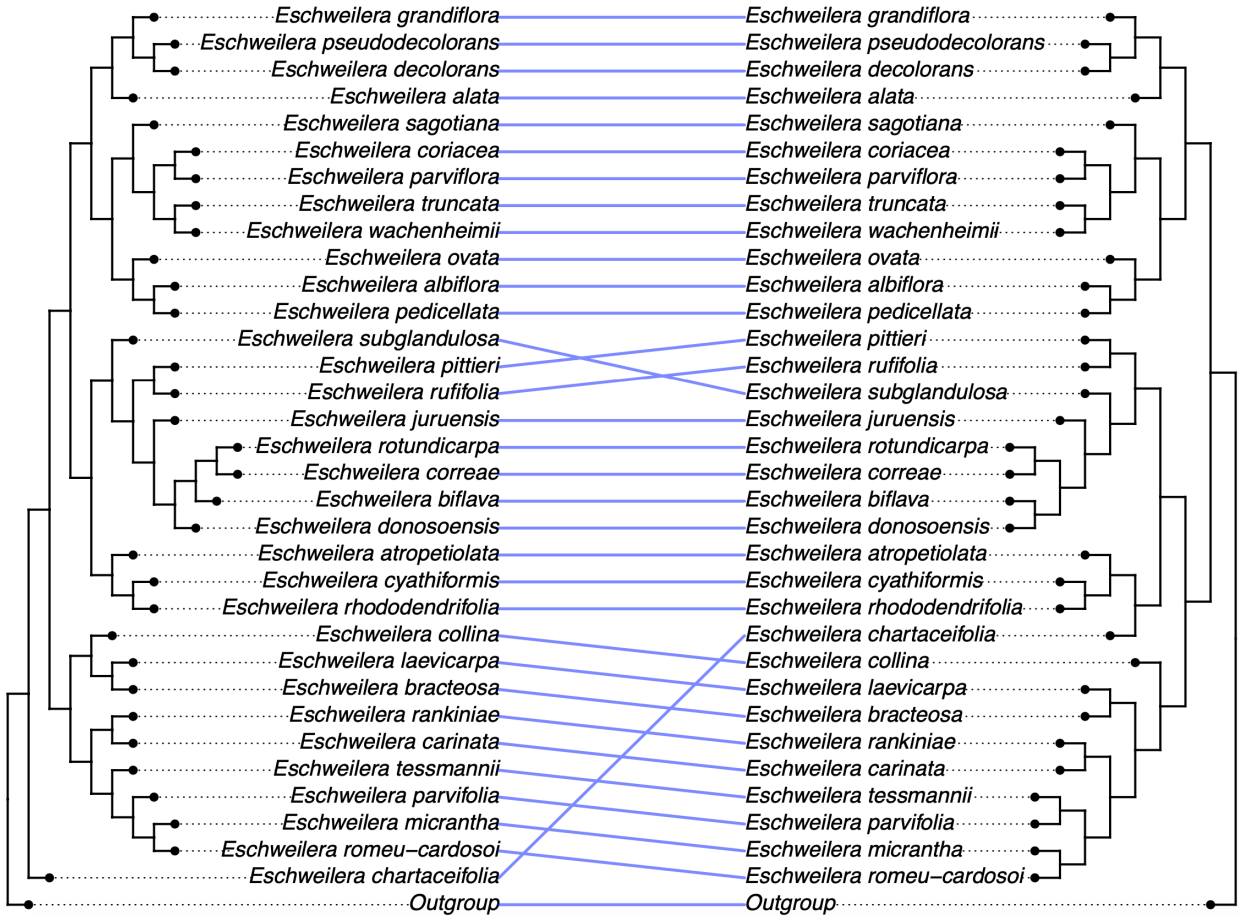

Fig. S8 Boxplots overlayed with dot plots showing the day of the year that collections of *Eschweilera coriacea* (n=35), *E. parviflora* (n=18), and *E. wachenheimii* (n=9) were made from Amazonas, Brazil for specimens housed at the New York Botanical Garden Herbarium (NY). Each dot represents a collection from a unique individual with open flowers or flower buds. The upper and lower limits of boxes represent the third and first quartiles respectively. Whiskers show the minimum and maximum values that occur within 1.5 times the length of the interquartile range. Points beyond the whiskers could be considered outliers. The darker midlines represent medians. On the y-axis, day 1 represents January 1st and day 300 represents October 27th in non-leap years.

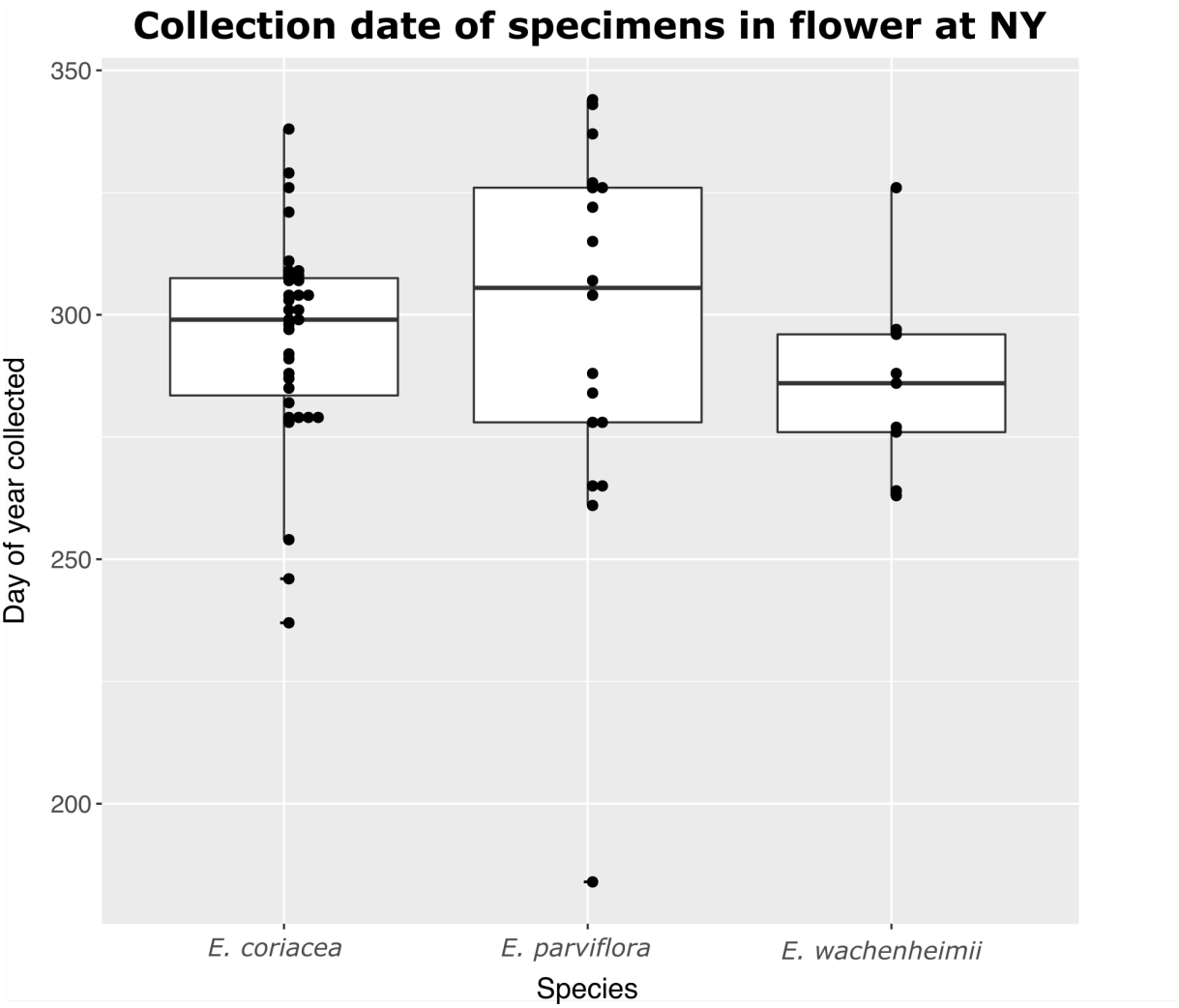

Table S1 Voucher and accession information for samples used in the study. National Center for Biotechnology Information SRA accession numbers for all samples are listed, as are summarized results used to make redeterminations based on all available evidence. “Not applicable” is abbreviated as “n.a.”.

Table S2 Summary statistics for all SNP datasets and estimated probability of the data for all *Structure* analyses for differing values of K. “Not applicable” is abbreviated as “n.a.”.

Table S3 Summary of all rooted triplet tests conducted.

Table S4 Results of tree searches and likelihood calculations for the Parvifolia phylogenies, ordered by increasing AIC score.

Table S5 Morphological and ecological traits of species inferred to engage in admixture.

References

Mori SA, Lepsch-Cunha N. **1995**. *The Lecythidaceae of a central Amazonian moist forest*. Bronx, New York: New York Botanical Garden Press.

Mori SA, Smith NP, Cornejo X, Prance GT. **2010**. The Lecythidaceae pages. *The Lecythidaceae pages*. [WWW document] URL http://sweetgum.nybg.org/science/projects/lp/. [accessed 9 July 2020].

Table S6 Information for specimens with flowers or flower buds collected in Amazonas, Brazil and housed at the New York Botanical Garden Herbarium.

Tables S1-S6 in file: SI_tables_S1-6.xlsx

Methods S1 Paralog filtering and alignment.

We applied a tree-based approach to filtering the gene assemblies produced by HybPiper. While the probe set we used for target capture sequencing was meant to recover the same genes in all samples (i.e., orthologs), in cases where two or more similar sequences match a probe, both can be enriched during library preparation. The settings we used to assemble the data in HybPiper were such that when more than one contig is recovered at similar read coverage (less than 10x difference) for a given gene target and both assemblies were at least 85% the length of the reference sequence, the program returned whichever had the greatest percent identity to the reference. As an additional step to limit the inclusion of paralogs in the

ortholog groups (i.e., orthogroups), we used *a priori* knowledge of the major clades of Lecythidaceae to split orthogroups, while retaining nearly all data for downstream analysis.

We first aligned exon and amino acid sequences for each of the 343 orthogroups recovered by HybPiper. All alignments were generated using MAFFT v7.271 and the option --maxiterate 1000. Amino acid sequences were aligned with the L-INS-i algorithm and nucleotide sequences were aligned using the less computationally intensive FFT-NS-i algorithm. A phylogeny for each was estimated using RAxML v8.2.11 with the GTRCAT model of evolution for nucleotide alignments and the PROTCATWAG model for amino acid alignments.

Then, we visually inspected orthogroup alignments for evidence of paralogy issues such as regions with many mis-matched bases, regions with lots of gaps, and instances of nearly identical sequences occurring in some taxa but not others in a biologically implausible way based on previous phylogenetic studies. In cases where visual inspection of an orthogroup alignment suggested paralogy issues, we found that phylogenies estimated from amino acid data tended to clearly reflect these issues. Many such amino acid phylogenies contained long branches (i.e. relatively many inferred substitutions per site), subtending "clades" which were extremely unlikely to occur due to biological processes, based on our understanding of the Lecythidaceae phylogeny. For example, the Neotropical Lecythidaceae (sometimes referred to as subfamily Lecythidoideae) has been strongly supported as a clade, and gene trees in which these species do not form a monophyletic group may be suspect. While phylogenies built using the corresponding exon and/or intron data usually also contained the same biologically dubious relationships, the branch lengths were more variable than those of the amino acid trees, possibly because there was a high proportion of gaps inferred for these nucleotide alignments. Therefore, we chose to examine individual amino acid phylogenies, to identify instances of potential paralog issues.
Amino acid trees without apparent paralog issues were used to determine the range amino acid substitutions per site inferred for the branch separating the Neotropical Lecythidaceae from the other members of the family. In nearly all cases where no issue was detected, the branch in question had a branch length less than 0.25 substitutions per site and this value was therefore used as the upper limit of expected branch lengths for genuinely orthologous sequences in amino acid trees. Using the methods of Yang & Smith (2014), we then cut any internal branches in the amino acid trees longer that 0.25 and retained any subtree with at least 10 taxa as a separate orthogroup. Any terminal branch in amino acid trees longer than 0.15 substitutions per site was also cut, to reduce paralogs occurring in any single sample. Following this procedure, there were 661 orthogroups. The corresponding nucleotide orthogroups (both exon and intron data) were then split to match the results of the amino acid orthogroup pruning. Nucleotide sequences for each of the 661 resulting orthogroups were aligned separately with the L-INS-i algorithm in MAFFT.

Methods S2 Genotyping and SNP dataset analyses.

Exon sequences from a single sample (i.e. *Eschweilera coriacea*; EscoL834; Table S1) were used as a reference "genome" because we were able to recover sequence data for 343 loci for this sample and *E. coriacea* was the most extensively sampled species in our dataset. The custom script *make_reference_genome_from_exonerate_exons.py* was used to generate the reference from exon data using the *exonerate_results.fasta* file generated for this sample by HybPiper. For each target locus, exons were concatenated with a 400 "N" spacer between each to produce a single "pseudo-contig", meant to preserve linkage among exons of the same gene while reducing the possibility that read mapping errors could be caused by concatenating exon sequences in a non-biological way. Raw, untrimmed reads from target capture sequencing were concatenated with reads from whole genome shotgun sequencing of unenriched libraries. Then, SAM files were generated from raw reads for each of the 109 members of the Parvifolia clade (and one member of *Eschweilera* from outside the Parvifolia clade; *Eschweilera integrifolia*; sample EsinLA01; Table S1) using the script *generate_sam_files_from_raw_reads_parvifolia.py*. Next, the custom script *GATK_sam_to_halplotypeCaller_parvifolia.py* was run for each sample in a Docker v18.09.7 container, which executed the following commands using The Genome Analysis Toolkit (GATK) v4.1.0.0 and Picard v2.18.25: 1) **samtools view** 2) **SortSam** 3) **MarkDuplicates** 4) **AddOrReplaceReadGroups** 5) **samtools index** 6) **HaplotypeCaller**. This resulted in a Genomic Variant Call Format (GVCF) file for each sample that was combined with the GATK command **CombineGVCFs**. Finally, variant calling was conducted using the command **GenotypeGVCFs** with the options "--include-non-variant-sites" and "--annotate-with-num-discovered-alleles true" and non-SNP variants were removed with the command **SelectVariants** and the options "--exclude-non-variants true", "--exclude-filtered true" and "--select-type-to-include SNP". This resulted in a Variant Call Format (VCF) file with SNP data for 110 samples that could later be subset to address specific questions regarding the population structure of species of the Parvifolia clade. The **SelectVariants** command was also used to exclude sample EsinLA01 (which is not a member of the Parvifolia clade) in order to confirm that its inclusion did not affect the total number of polymorphic sites identified, which was 148,310.

Subsets of samples were selected for analysis in *Structure* v2.3.4 based on the results of the preliminary phylogeny (Figure S2). For each taxon subset, variants were filtered using the programs plink (http://pngu.mgh.harvard.edu/purcell/plink/) and plink2 (www.cog-genomics.org/plink/2.0/) to obtain a subset of SNPs for that subset of individuals that were in approximate linkage equilibrium. This was accomplished using the following commands where X.plink.txt and X.pops.txt were files that specified the following respectively: 1) the samples to include in the subset and 2) *a priori* population membership based on previous morphological determination and the results of the preliminary phylogeny:

plink --vcf ../Genotypes_110_parvifolia_clade_EscoL834_ref_SNP_only.vcf --keep X.plink.txt --allow-extra-chr --make-bed

plink2 --allow-extra-chr --bfile plink --set-all-var-ids @_#_\$r_\$a --out prefilter --new-id-max-allele-len 286 --max-alleles 50 --make-bed

plink2 -bfile prefilter --indep-pairwise 50kb 1 0.0001 --allow-extra-chr -out LD_STEP

plink2 --allow-extra-chr --bfile prefilter --out final --new-id-max-allele-len 286 --max-alleles 50 --make-bed --extract LD_STEP.prune.in

plink --recode structure --allow-extra-chr -bfile final

python convert_plinkRecode_to_structure.py plink.recode.strct_in X.pops.txt

All *Structure* analyses were run for $1x10^6$ MCMC generations after $1x10^5$ generations of burnin. A full list of all parameters used in *Structure* runs is available in the Dryad repository submission accompanying this article. For each SNP dataset, *Structure* was run separately with the maximum number of populations (K) from 1 to 7, when we expected five or fewer clusters, or from 1 to 10 if the subset was expected to contain six or more clusters. Results of *Structure* analyses were formatted with the custom script *f2R.py* and visualized using custom R scripts.

The optimal K for each subset was determined by comparing the "Estimated Ln Prob of Data" (herein referred as the score) in the output files of each run (Table S2) with *a priori* taxonomic information (Table S1). In most cases, the run with the best score matched or nearly matched our *a priori* expectation of the number of species based on morphological determinations and preliminary phylogenetic analysis (Table S2). Exceptions to this occurred for taxa subsets where there were multiple species with only a single representative in the analysis and when there were multiple inferred clusters within one species. Though some of the species we included in our *Structure* analyses were represented by a single individual, we believe the results of these analyses are robust because 1) the loci targeted by our sequencing were selected *a priori* based on phylogenetic informativeness; 2) the results of these tests were concordant with our understanding of the Lecythidaceae phylogeny; 3) our SNP datasets averaged only 28.17% missing data and averaged 229 SNPs with no missing data for any individual; 4) results from runs with overlapping of individuals corroborate one another with respect to the ancestry of individual samples (Table S3).

Methods S3 Redetermination of individuals based on all available evidence.

Closely related species of Lecythidaceae are known to be difficult to identify, especially if reproductive material is unavailable and determinations must be based made on vegetative characters alone. In cases were the most recent morphological determination did not agree with the results of genetic analyses (i.e. the preliminary phylogeny, *Structure* analyses, and any relevant RT tests) all available evidence was used to determine the most likely species identity for the individual if it was shown to be a member of the Parvifolia clade in the preliminary phylogeny (Fig. S1). For samples that fell outside the Parvifolia clade in preliminary phylogeny and were inferred to belong to a genus other than that of their most recent morphological determination, we assigned a genus-level redetermination (i.e. *Lecythis sp.*); such redeterminations had no effect on downstream analyses for this study, but were made in order to provide more accurate taxonomy for these samples in online databases and to facilitate their use in other studies. With one exception (i.e. *Eschweilera roseocalyx*, discussed below), a redetermination based on genetic evidence was made for individuals in the Parvifolia clade if a *Structure* analysis showed that the individual clustered most closely with individuals other than of its most recent morphological determination and these results were corroborated by phylogenetic analyses. Individuals were assigned a "*species affinis*" (i.e. aff.), designation to indicate alliance with a species other than that of their morphological determination if they met the following criteria: 1) The sample showed no significant evidence of admixture; 2) the sample had a species-level morphological determination and was not the only sample with that morphological determination; 3) phylogenetic analysis showed the sample was more closely related to another species than it was to other samples with its own morphological determination and 4) in phylogenetic analysis, the sample was inferred to diverge earlier than all other individuals of the species with which it was inferred to be most closely related, ignoring any individuals whose morphological and genetic determinations disagreed (Fig. S1).

A recently described species, *Eschweilera roseocalyx*, was included our sampling (i.e. EscoL545, Table S1). This species was described by Batista et al. (2017) based on a single individual found in a cloud forest in Panama's Chagres National Park. Our sampling included only one individual from Panama that was determined to be *E. coriacea* based on morphology; results from the preliminary phylogeny showed this individual was not a member of the Parvifolia clade and was likely mis-identified. Later phylogenomic analyses conducted in this study produced results that were compatible with an interpretation of *E. roseocalyx* as a geographically restricted ecotype of *E. coriacea*. Because of this compatibility and the absence of population-level data available elsewhere in the literature, we treat this individual as a member of *E. coriacea*. Species delimitation is outside the scope of the present work and we do not wish for our treatment of *E. roseocalyx* here to be interpreted as evidence against the taxonomic validity of any species. Subsequent studies may show this to be an example of peripatric

speciation whereby the widespread *E. coriacea* has given rise to a reproductively distinct endemic species. However, given the available evidence and the focal questions the present work seeks to address, we feel that treating this individual as *E. coriacea* is more justifiable than treating it as a separate species in our results and discussion.

Methods S4 Parvifolia phylogeny supermatrix construction and phylogeny estimation.

Orthogroup alignments used for the preliminary phylogenies were subset to include only members of the Parvifolia clade and five members of the Integrifolia clade as outgroups (Table S1). Samples with evidence of recent admixture were excluded from this analysis, since admixture violates the assumptions of tree-based phylogenetic inference. We choose not to re-align orthogroups at this step, allowing sequences from taxa outside the Parvifolia clade to inform the final alignment of this dataset. This decision was due mainly to the fact that intron sequence recovery was highly variable among samples, which is to be expected because the probes used for the target capture sequencing protocol are meant to capture exons. The intronic sequences recovered are expected to be those adjacent to the targeted exon regions, captured because they are located close enough to the targeted exons to be enriched during library preparation, but outside the region for which the probes were specifically designed. This can result in greater variability in recovery during sequencing and assembly. The resulting alignments will likely have more missing data, especially if some intronic regions are only recovered in small number of taxa, which could increase alignment error. Poor alignment of sequences can severely impact model parameterization, likelihood calculations, and the inferred topology of phylogenies, and we therefore chose to include all available data in generating our orthogroup alignments.

To further reduce possible non-orthology in our estimation of the Parvifolia phylogeny, we employed a second tree-based filtering protocol to generate the final supermatrices. To accomplish this additional filtering, for each orthogroup (i.e. each of the 661 the resulted from the first round of trimming for the preliminary phylogeny), a tree was estimated separately for each intron and exon alignment using IQ-TREE with a GTR+G model of evolution after filtering out columns with less than 30% occupancy with the pxclsq command in phyx. For each unrooted tree, the average length and standard deviation of internal branch lengths and terminal branch lengths was calculated using the custom script *trim_trees_based_on_branch_distributions.py*. For this procedure, internal branch lengths were compared to other internal branches and terminal branch lengths were compared to other terminal branches. This allowed us, in an automated way, to determine branch lengths that were much longer than those of comparable branches within an orthogroup, which could be indicative of orthology issues. Any internal branch longer than average plus two standard deviations was cut using the methods of Yang & Smith (2014) and any terminal branches longer than average plus two standard deviations was also cut from the

resulting subtrees. Any subtree without at least 27 tips representing members of the Parvifolia clade and five outgroups (approximately 25% occupancy) was excluded from further analysis. The resulting pruned trees were used to generate the final alignments by subsetting the orthogroup alignments to remove any sequences cut during the trimming procedure. Any sequence that would have contained more than 75% missing data in the final alignment was also excluded. Finally, sites with less than 30% occupancy were removed from each filtered alignment with the pxclsq command in phyx. Two supermatrices were produced with the pxcat command in phyx, one that included only exon data (668,353 aligned sites, 382 partitions, 29.74% missing data), and a second that included both exon and intron data (2,085,546 aligned sites, 765 partitions, 36.3% missing data).

For each supermatrix, 1) a maximum likelihood (ML) tree was estimated using RAxML v8.2.11 with a separate GTRCAT model of evolution specified for each partition with the -q option, 2) an ML tree was estimated using IQ-TREE v1.6.9 with a separate GTR+G model for each partition using the -q option and 3) a second ML tree was estimated with IQ-TREE and the same settings but allowing for partition-specific scaled evolutionary rates with the -spp option. To allow for direct comparisons, likelihoods and information criteria scores for each tree were recalculated with IQ-TREE with the following criteria 1) a tree topology fixed to that of the original result 2) a separate GTR+G model for each partition with and without partition-specific rates in separate analyses and 3) re-estimated branch lengths for the tree. The Akaike information criterion (AIC) score for each result was compared to determine the best scoring phylogeny for each supermatrix, which we considered as representing the best topological hypothesis for that dataset.

Notes S1 Note on species of the Parvifolia clade at Reserve 1501.

There were five species known to occur in the 100-ha Lecythidaceae plot at Reserve 1501 that are now recognized as members of the Parvifolia clade, but that we did not include in our focal sampling either due to their rarity in the plot or past uncertainty in their phylogenetic placement. These non-focal species, all of which were represented by at least one individual in our overall sampling, were *E. carinata* S.A.Mori, *E. grandiflora* (Aubl.) Sandwith, *E. parviflora* (Aubl.) Miers, *E. romeu-cardosoi* S.A.Mori, and *E. tessmannii* R.Knuth.

Notes S2 Notes on sampling at Reserve 1501 and prioritization of morphological intermediates.

We employed a mixture of planned and opportunistic sampling (i.e. adjusting sampling plans in the field and collecting obtainable specimens as they are discovered, rather than in a strictly randomized way) while making new collections at Reserve 1501. Various logistical challenges can make sampling from tropical trees difficult, even in established plots. For example, tree tags may fall off, making the

target tree difficult or impossible to locate. Trees may have died since the last census, species may grow in patchy distributions so that sampling direct relatives may be a concern, and some trees may simply be too large or tall to safely collect a specimen. These logistical realities mean that opportunistically sampling can be a useful way to collect specimens in tropical forests.

In the vast majority of cases, we sampled trees based solely on the species name assigned during previous censuses of the plot in order to achieve a sampling rate of 4-8 individuals per focal species. We chose target trees to visit based on their previous species determination, that they were a minimum distance of 100m from any other sampled individual of their species, and that they were along a route that would facilitate collection of multiple samples during the day of field work. If, for whatever reason, a collection could not be safely obtained from a target tree, a substitute was chosen using the same criteria outlined above.

There were three trees we encountered in the Reserve 1501 plot that appeared to us to be possible hybrids between *Eschweilera coriacea* and another species (Table S1). These trees displayed morphological traits, including branching architecture, that appeared to be intermediate between species. We intentionally prioritized these three samples for collection and genetic analysis. While one of these specimens did ultimately show genetic evidence of admixture between *E. coriacea* and *E. wachenheimii*, the other two showed no evidence of admixture. However, a fourth sample from Reserve 1501, thought to be *E. coriacea* in the field, did show strong evidence of admixture in later analyses. Outside of Reserve 1501, neither individual ultimately found to be *E. parviflora* × *wachenheimii* based on genetic evidence was suspected of being admixed prior to genetic analysis. Thus, it should be noted that most individuals with genetic evidence of admixture in our study were not suspected to be admixed based on morphology.