Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Combining phylogenomic and morphological data reveals new patterns of diversity in the national tree of Brasil, Paubrasilia echinata

Mathew Rees
University of Edinburgh
Linda E. Neaves
Australian National University
Gwilym Peter Lewis
Royal Botanic Garden Kew
Haroldo Calvacante de Lima
Instituto de Pesquisas Jardim Botânico do Rio de Janeiro
Edeline Gagnon (edeline.gagnon@tum.de)
Technical University of Munich

Research Article

Keywords: Fabaceae, Leguminosae, Pau Brasil, Population genomics, Rad-Seq, Mata Atlantica

Posted Date: November 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2250881/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Paubrasilia echinata (Lam.) Gagnon, H. C. Lima & G. P. Lewis (“Pau Brasil”) is the national tree of Brazil and an endangered species endemic to the Brazilian Atlantic Forest. The extensive range of distribution, spanning over 2000 km distance, is matched by extensive plasticity in leaf morphology. Three morphotypes are commonly identified based on the size of the leaflets but it is unclear if they represent distinct taxa or a single polymorphic species. This study aims to clarify the taxonomic position of the three morphotypes to inform conservation decisions. A comprehensive morphometric study based on herbarium specimens from the entire distribution range of the species was coupled with genetic analyses of population structure using genotype-by-sequencing data. We found that the three morphotypes do not match separately evolving lineages. Rather, *P. echinata* is composed of five genetic lineages that are geographically structured, although we did find evidence of genetic admixture in two individuals. Leaflet size varied by over 35-fold and although morphological clustering generally matched the genetic patterns, there were some overlaps, highlighting the cryptic diversity within this group. Finally, our genetic and morphological results provide some evidence that cultivated trees from different states in Brazil seem to be most closely related to a genetic lineage from northern Brazil, which suggests that more care is needed to protect and preserve the overall genomic diversity of this highly endangered and iconic species.

Introduction

The Atlantic Forest of Brazil is one of the 36 biodiversity hotspots in the world (Habel et al. 2019), but possibly one of the most endangered with only 7-12% of its original geographical extent remaining (Morellato and Haddad, 2000; Ribeiro et al., 2009). Despite this huge habitat loss, it remains home to more than 17,150 species of plants, with many new species having been described in recent years (The Brazil Flora Group, 2022). Part of this diversity is due to the environmental heterogeneity of the Atlantic Forest which although mainly a rainforest biome, 45% of its tree species are narrow endemics in a variety of marginal habitats, including restinga, riverine forests, rocky outgroups, and high elevation mountains (Neves et al. 2017).

An example of a tree endemic to the Atlantic Forest is *Paubrasilia echinata* (Lam.) Gagnon, H. C. Lima & G. P. Lewis (Leguminosae/Fabaceae), a slow-growing, semi-deciduous, small to medium-sized tree (c. 4-20 m tall) that grows in inland ombrophilous mesic forest, as well as in a range of much drier habitats, including disturbed dry coastal cactus scrub, rocky outcrops, seasonally semi-deciduous forest (tabuleiro forest), and restinga, a type of coastal forest with well drained sandy soil (Lewis 1998; Gagnon et al. 2020). Its distribution extends over 2100 km from Rio de Janeiro in the south to Rio Grande do Norte in the north, and while previously thought to be relatively common throughout the Atlantic Forest (Lima et al. 2002), current populations are highly fragmented. This is notably due to centuries of exploitation, previously for extraction of a red dye and subsequently for the manufacture of high-quality violin bows (Varty 1998; CITES 2007). Today, interest in Pau Brasil remains high because of its cultural and commercial value, and it is sometimes cultivated as an urban street tree, as well as in monospecific or

mixed agroforestry plantations, potentially to support future wood demand from musical instrument manufacturers, especially for violin bows (Lichtenberg et al. 2019).

The population structure and levels of inbreeding in *P. echinata* has been investigated by using samples collected in seven remnant populations. Study methods have used microsatellites, random amplified polymorphic DNA or amplified fragment length polymorphism (MA Cardoso et al. 1998; SRS Cardoso et al. 2005; Lira et al. 2003; Melo et al. 2007; Cupertino et al. 2014), but none of them have actually examined the long-standing hypothesis that cryptic species or subspecies exist, a hypothesis suggested by the impressive variation in leaflet size, varying from 0.9 cm to 12 cm in length (Lewis, 1998).

Several studies have hypothesized that *P. echinata* is a species complex (Lima et al., 2002; Gagnon et al., 2016; Juchum et al., 2008; Lewis, 1998; Macedo et al., 2019; Rodrigues et al., 2018), composed of three morphotypes that are geographically and ecologically structured. The first is a common and widely distributed small-leafleted variant called variant called "folha-pequena" or "folha de Arruda" (rue-leaf in English), distributed in dry restinga forests, tabuleiro forests and rocky outcrops along the coast, in the southern state of Rio de Janeiro and the northern states of Bahia, Alagoas, Pernambuco, Paraíba,and Rio Grande do Norte (Macedo et al. 2019). The second is a medium-leafleted variant variant "folha média" or "folha de Café" (coffee-leaf), found predominantly in the state of Espírito Santo and Southern Bahia (Macedo et al. 2019). The third is an extremely rare and localized large-leafleted variant "folha de Laranja" (orange-leaf), found in restricted populations along the Rio Pardo Valley, in the state of Bahia (Lewis, 1998; Juchum et al. 2008), predominantly in ombrophyllous mesic forests. This, however, does not account for the several other wild remnant populations of Pau Brasil encountered in the more northern Brazilian states of, which display a wider variation in leaflet size; the latter have been hypothesized to belong either to the "folha de Arruda" or "folha de Café" morphotype. More work is needed to determine whether these three named morphotypes correspond to three distinct genetic and morphological lineages that could be elevated to distinct named taxa (either varieties, subspecies or species in their own right).

There is some evidence that leaflet size in Pau Brasil is not due to phenotypic plasticity, as researchers from the Centro de Pesquisas do Cacau (CEPEC) have been growing seedlings from these three morphotypes side-by-side, leading to the observation that the differences in leaflet size did not appear to be related to environmental factors (Robelio de Santana, pers. comm.) Furthermore, bow makers often use the names "sartori" and "peccati" and "pernambuco", which reflect different characteristics and quality of the wood for their crafting of violin bows (Macedo et al. 2020). These correspond sometimes to specific populations, such as "peccati" wood referring specifically to trees from the “Café” morphotype in Espírito Santo. Significant differences in heartwood color have also been suggested to help distinguish the three varieties (Lewis, 1998; Lima et al. 2002). However, studies based on wood anatomy and density (Macedo et al. 2019), karyotype diversity and genome size (Van-Lume et al. 2017; Rodrigues et al. 2018) and molecular markers (Juchum et al., 2008), have not found sufficient evidence to support the description of new taxa from within *P. echinata*. Furthermore, the geographical distribution and morphological characters that help diagnose and delimit the three morphotypes based on leaflet size has never been fully described. There have been suggestions that the Arruda phenotype not only have smaller

leaflets, but also more numerous pinna and leaflets (5-10 pinnas and 12-21 leaflets, pinna approximately 4 cm in length), and the “Café” morphotype has larger leaves, but less numerous pinna and leaflets (3-5 pinnas and 3-8 leaflets, pinna approximately 7 cm in length) (Macedo et al. 2019). For the “Laranja morphotype, descriptions in the literature only mention that leaflets can reach up to 12 cm long, without additional details (Lewis, 1998; Lima et la. 2002).

To investigate whether the three known morphotypes of *P. echinata* correspond to distinct evolutionary lineages, we used an integrative approach, combining population genomics and morphometrics analyses of field-collected and herbarium samples from across the entire range of the *P. echinata* distribution, including for the first-time populations of the rare "folha de Laranja" morphotype. Specifically, we address the questions: (1) What is the population genetic structure of *P. echinata* across its entire range? (2) Do morphometric analyses of leaf traits identify three distinct morphotypes, and do they correspond to genetically distinct lineages? We use these results to evaluate whether infraspecific taxa or species should be formally described and discuss how they bring new insight into the conservation of these morphotypes.

Methods

Taxon sampling for molecular analyses

Specimens were collected in November 2014 from wild populations of *P. echinata* growing in remnant patches of the Brazilian Atlantic Forest. The "Arruda" morphotype from Bahia (five individuals/two localities) and Rio de Janeiro (four individuals/two localities), the "Café" morphotype from Espírito Santo (five individuals/one locality), and the "Laranja" morphotype from a fragment forest patch in Bahia (twelve individuals/one locality). In addition, at the "Laranja" locality, we included a sample of an individual of *P. echinata* that appeared to be morphologically closer to the "Arruda" morphotype (Lima et al. 7910). (Fig. 1, full details in Supplementary Information 1, Table S1). From all localities, we collected individuals along a transect from the centre of the forest fragment to its margin. In addition to field-collected specimens we sampled DNA from thirteen herbarium specimens, collected between 1985 and 2012. This included samples from natural populations occurring in the northern states of Brazil (Alagoas, Paraíba and Rio Grande do Norte), which we classified here as a distinct morphotype called “North”, due to the difficulty of determining whether specimens fall into the “Arruda” or “Café” morphotype. Additional localities from natural populations in Bahia and Rio de Janeiro were also sampled from herbaria, representing mostly the “Arruda” morphotype, and with one sample being the only representative of the “Café” morphotype from Bahia (*Santana et al. RB472149*). Six of the herbarium samples were from cultivated specimens, three from cultivated trees in Espírito Santo, part of a conservation programme for Pau Brasil, allowing us to trace their state of origin and assign them to morphotypes ("Café" and "Arruda"). For the other three cultivated samples we were unable to confidently trace their provenance; two were from trees grown outside the native range of *P. echinata*, in the state of São Paulo; the third (specimen *Lima et al. 6662*) was collected in the Rio de Janeiro Botanical Garden arboretum, which

cultivates several morphotypes of Pau Brasil. Our analyses outgroup comprised seven species of the genus *Cenostigma* (Supplementary Information 1, Table S1).

In summary, our full genomic dataset included 39 samples of herbarium and field-collected specimens, 12 assigned to the "Laranja" morphotype, eight to the "Café" morphotype, and 13 to the "Arruda" morphotype, plus an additional three samples from northern Brazilian populations of *P. echinata* (North group), and three cultivated specimens of unknown provenance (Supplementary Information 1, Table S1).

DNA extraction and sequencing

All DNA extractions were performed following the protocol of Arseneau et al. (2017), a modified low-salt CTAB method that yields high quality DNA for high-throughput sequencing, even from samples containing high levels of secondary metabolites. Library preparation and sequencing was carried out at the Institute of Integrative Biology and Systems (IBIS) in the University of Laval, Canada, using a modified GBS approach with single end reads and two restriction enzymes: a rare cutting enzyme PstI (CTGCAG) and a common cutting enzyme MspI (CCGG) (Poland et al., 2012). Both these restriction enzymes are sensitive to CpNpG methylation, which allows for the exclusion of repetitive elements in non-coding regions. The samples were sequenced on one lane of an Illumina HiSeq 4000 at the McGill University and Genome Quebec Innovation Centre (Montreal, Canada). An initial quality check was carried out using FastQC (Andrews, 2010). One sample (Malagodi 6323) sequenced poorly and was removed from subsequent analyses. All the generated genomic data is available at NCBI's Sequence Read Archive (SRA) under reference PRJNA689870. (https://www.ncbi.nlm.nih.gov/sra/PRJNA689870).

Data assembly

All analyses were run on the Crop Diversity Server, from the James Hutton Institute (Dundee, Scotland, UK). The software ipyrad v. 0.9.84 (Eaton and Overcast 2020) was used to build de novo assemblies of RAD-tags, by specifying the data type as being "ddrad". As the three morphotypes of *P. echinata* have been confirmed all to be diploid (Rodrigues et al. 2018), we set the maximum number of alleles in the individual consensus sequence at two. We used all other default settings provided by ipyrad, with the following modifications: 1) the first 10 bases of each locus were trimmed to account for the heterogeneity in per base sequence content highlighted in our Fast-QC results; 2) a minimum depth of 10 was specified for statistical base calling and majority rule consensus to avoid potential artefacts created by the SNPs called in shallow loci in the herbarium samples; 3) a minimum of 20 specimens had to share a RAD-Tag. These analyses were initially carried out on the full dataset with the outgroup (46 samples), and on a second dataset that excluded the outgroup (giving a total 39 samples).

After demultiplexing the raw reads, samples recovered, on average, 7.48 million reads (max: 32,783,012; min: 44,495; Supplementary Information 2). The dataset containing the outgroup resulted in the retention of 7,165 loci after filtering, with 41,496 SNPs (26.14% missing sites); the dataset with the outgroup

pruned resulted in the retention of 7,123 loci, with 31,439 SNPs (18.45% missing sites) (full details in Supplementary Information 2).

Population genomics and phylogenetic analyses

To understand the population structure of *P. echinata*, the maximum likelihood algorithm in ADMIXTURE v.1.3.0 (Alexander et al. 2009) was used to determine the number of genetic clusters present in the data, using the dataset pruned of the outgroup. Prior to running the analysis, BCFTOOLS v.1.3.1 (Danecek et al. 2021) was used to remove loci with more than two variants and PLINK v.1.90b6.21 (Weeks 2010) was used to reduce linkage disequilibrium, by pruning the data matrix using a threshold of $r^2 < 0.2$. The best number of clusters (K) was inferred using the cross-validation procedure implemented in ADMIXTURE, by testing values from 1 to 10, with 10 replicates for each K; the optimal value of K was determined by using the lowest average of the standard error of the cross-validation error estimate. Results for the analyses were visualised using custom R scripts, based on the package POPHELPER v.2.3.1 (Francis 2017).

In addition, a discriminant analysis of principal components (DAPC) (Jombart et al. 2010) was performed, using the final datasets without any outgroup (39 samples). The R package VCFR v. 1.8.0 (Knaus and Grünwald 2017) was used to convert variant call format (VCF) files to genind objects used by ADEGENET v.2.1.8 (Jombart 2008). The function "find.clusters()" was used to choose the optimal number of genetic clusters from k=1-39, selecting the K with the lowest BIC (Bayesian Information Criterion) value. To select how many principal components should be retained to maintain a high discriminant power without overfitting the data, a cross-validation method ("xvalDapc()" function), with 1000 repetitions, was used, as well as the method-based calculation of the a.score ("optim.a.scores()" function).

Finally, phylogenetic analysis of the entire dataset, including the outgroup taxa, was carried out using maximum likelihood (ML) and quartet inference methods. ML phylogenetic analyses were conducted with RaxML v.8.2.12 (Stamatakis 2014) using 1000 rapid bootstrap analyses, followed by 10 rapid hill climbing ML searches from random starting trees, under a GTRGAMMA substitution model. Phylogenies based on quartets was computed with TETRAD v.0.9.14 (Eaton et al. 2017), with 200 non-parametric bootstrap replicates to form a consensus tree by joining all quartets together into a supertree using QMC (Snir and Rao 2012). Results of both phylogenetic analyses were viewed in FigTree v.1.4.4 (https://github.com/rambaut/figtree/releases/tag/v1.4.4) and plotted with the package GGTREE v. 3.2.0 (Yu et al. 2017), using R v.4.1.0 (R Core Team 2021).

Morphometric analyses

The morphological analysis used a total of 79 herbarium specimens, deposited in five different herbaria (K, MT, MO, NY, RB, Supplementary Information 3). When accounting for duplicate specimens, they represented 56 individual collections, distributed across 42 localities of the entire extant range of *P.*

echinata, from the states of São Paulo north to Rio Grande do Norte. Localities of herbarium specimens were carefully screened to determine which specimens came from wild remnant populations; if they were cultivated or part of a conservation programme, we attempted to trace information about their provenance. In total, 39 specimens came from wild remnant populations, which were assigned to four categories: the "Laranja" morphotype (six specimens), the "Café" morphotype (Espírito Santo, three specimens), the "Arruda" morphotype (individuals with small-leaflets from Bahia and Rio de Janeiro, 22 specimens), and a group of specimens growing in states of northern Brazil (Alagoas, Pernambuco, Paraíba, and Rio Grande do Norte, 8 specimens). An additional nine specimens came from cultivated trees for which we were able to trace their provenance and assign them to the vernacular groups described above, and eight specimens came from cultivated trees for which we were unable to trace their provenance, and hence did not assign them any morphotype (Supplementary Information 3). Nineteen of the herbarium samples were also included in the genetic analyses (indicated in colour in Fig. 2c).

Measurements of morphological characters on herbarium specimens were made using an architect ruler and a dissecting microscope. Two specimens that were only available as high-resolution digital images (*Lima et al. 7905*) were measured using the software imageJ (Rueden et al., 2017). A total of 17 quantitative and qualitative variables were initially selected to perform the morphometric analysis (described in Supplementary Information 3). For traits related to leaflet size, a total of seven leaflets were measured per specimen (the two largest, two smallest and three leaflets selected at random). For all other traits three leaves were measured for each specimen, taking care to only select mature leaves. Duplicates specimens representing the same field collection were merged for each morphological variable (Supplementary Information 3).

As the total amount of missing data was low in the dataset (0.6%), we performed an imputation by PCA (Brown et al. 2012), using the function "imputepca()" in the R package MISSMDA v. 1.14 (Josse and Husson, 2016). We first checked if there was any bias in our dataset between the collection date and leaflet measurements by calculating the correlation of the month of collection and a subset of quantitative leaf variables. As there did not appear to be any significant trends, we then proceeded to select five variables with an absolute correlation value less than 0.6, using the function "cor()" in R: leaflet length, petiole length, rachis length, pinna length and ratio of leaflet length to width (Table 1, Supplementary Information 3). The distribution of the data was visualised and checked for normality using qqplots and the Shapiro-wilk test, resulting in the square root and log transformation of four of the five variables (Table 1). To avoid infinite values with log transformation, zero values were replaced by half the measurement precision (0.05mm). Variables were subsequently standardised using the function "scale()".

To examine the distribution of specimens in the morphometric space, based on their assigned morphotypes, a principal component analysis (PCA) was carried out on the five traits, using the GGPLOT and GGFORTIFY packages in R (Tang et al., 2016). Subsequently, a Ward hierarchical clustering analysis was also performed on a Euclidean distance matrix calculated from the morphometric dataset, using the functions "hclust()" and "rect.hclust()". Because Ward clustering does not give any indication as to how

many clusters should be considered, a k-means analysis was run using the function "kmeans()". Both the Silhouette method (function "fviz_nbclust()", from the package FACTOEXTRA v.1.0.7 (Kassambara and Mundt, 2020)) and the Calinski criterion (function "cascadeKM()" from the package VEGAN v.2.5-7 (Oksanen et al., 2013)) were used to assess the optimal value of K from 1 to 10. These clusters were mapped back onto the hierarchical clustering analysis.

Finally, to understand the morphological variation of different leaf traits amongst the genetic lineages detected in our population structure analyses, boxplots of our five morphometric traits were produced, based on their untransformed values. We assigned individuals in our dataset to one of the genetic lineages detected in our population structure analyses; to reduce potential bias or error in our dataset, we removed cultivated samples of unknown provenance, as well as admixed individuals in our analyses. This left us with 46 samples from wild populations and cultivated trees of known provenance from the wild. We then assigned specimens to the five groups based on the results of the ADMIXTURE analysis (Fig. 2c): "Arruda-RJ", "Arruda-BA", "Café", "Laranja" and "North". A nonparametric Kruskal-Wallis test was carried out to determine whether there were significant differences amongst the medians of each group, followed by pairwise comparisons using the Dunn-

Table 1. Results from morphometric analyses. (a) Five leaf traits selected for morphometric analysis, with selected variable transformation. (b) Range of measurements of leaflet traits according to genetic clusters, based on merged, untransformed data, with cultivated specimens of unknown origin removed, as well as two admixed specimens.

a. Five leaflet traits selected for morphometric analysis, with selected variable transformation					
Leaf trait	**Petiole length (cm)**	**Rachis length (cm)**	**Pinna length (cm)**	**Leaflet length (cm)**	**Leaflet length/width**
Transformation	Log	Square-root	Log	Log	None
b. Range of measurements of leaf traits based on genetic clusters					
Leaf trait	**Petiole length (cm)**	**Rachis length (cm)**	**Pinna length (cm)**	**Leaflet length (cm)**	**Leaflet length/width**
Arruda – BA (12 specimens)	Mean: 0.892 Med: 0.830 SD: 0.465	Mean: 6.423 Med: 5.915 SD: 2.834	Mean: 5.940 Med: 6.000 SD: 1.240	Mean: 1.366 Med: 1.230 SD: 2.528	Mean: 1.801 Med: 1.785 SD: 0.118
Arruda –RJ (10 specimens)	Mean: 1.151 Med: 1.150 SD: 0.369	Mean: 6.092 Med: 5.200 SD: 2.942	Mean: 5.257 Med: 5.385 SD: 1.197	Mean: 1.456 Med: 1.435 SD: 2.528	Mean: 1.852 Med: 1.880 SD: 0.160
Café (7 specimens)	Mean: 0.439 Med: 0.430 SD: 0.082	Mean: 3.174 Med: 3.230 SD: 1.237	Mean: 4.787 Med: 4.480 SD: 1.282	Mean: 3.033 Med: 2.640 SD: 0.818	Mean: 1.530 Med: 1.520 SD: 0.121
Laranja (7 specimens)	Average: 0.991 Med: 0.650 SD: 1.108	Mean: 0.587 Med: 0.150 SD: 1.005	Mean: 8.077 Med: 8.330 SD: 3.696	Mean: 7.336 Med: 7.210 SD: 2.528	Mean: 1.721 Med: 1.670 SD: 0.177
North (10 specimens)	Mean: 1.025 Med: 0.915 SD: 0.459	Mean: 3.797 Med: 3.350 SD: 2.160	Mean: 5.310 Med: 4.235 SD: 2.249	Mean: 1.769 Med: 1.855 SD: 0.520	Mean: 1.574 Med: 1.575 SD: 0.110

test, to see which groups were significantly different from each other (alpha=0.05).

Results

Population genomic and phylogenetic analyses

For the ADMIXTURE analysis, the lowest cross-validation error was found for a value of k = 5 (average 4.3), followed by k = 4 (0.45) and K = 6 (0.46; Supplementary Information 1, Fig. S2). The five genetic clusters consisted of the following groups: (1) a cluster from Southern Bahia (BA1) containing all but one of the specimens assigned to the "Laranja" morphotype; (2) a cluster from Espírito Santo, containing all the "Café" morphotypes; (3) a cluster composed of individuals from populations in the Northern provinces of Brazil (Rio Grande do Norte, Paraíba, Alagoas) and cultivated individuals in the south (Espírito Santo, Rio de Janeiro, and São Paulo); (4) a cluster containing individuals of the "Arruda" morphotype from Bahia (referred hereafter as the "Arruda-BA" group) ; and (5) a cluster containing individuals of the "Arruda" morphotype from Rio de Janeiro (referred hereafter as the "Arruda-RJ" group; Fig. 2b). In addition, two individuals collected in Southern Bahia, *Lima et al. 7910* and *Santana et al. RB472149*, which were identified as belonging to the "Arruda" and "Café" morphotypes, displayed admixtures with the "Laranja" and "Arruda-BA" genotypes from Southern Bahia (Fig. 2b).

The DAPC analysis also recovered five distinct genetic clusters, with the k-means clustering algorithm in ADEGENET retrieving k = 5 clusters as the best score (BIC = 283.7). The clusters were the same as those recovered in the phylogenetic and ADMIXTURE analyses (except for the admixed individuals, with *Lima et al. 791*0 clustering with "Arruda-BA", and *Santana et al. RB472149* clustering with the "Laranja" clade). Results from the cross-validation method and the alpha-optimal method both indicated that the first five principal components should be retained to perform the DAPC, representing 62.8% of accumulated variance in the data. Projection of the DAPC along the first and second principal components show that the "Arruda-RJ" and "Café" clusters are more distant from the "North", "Arruda-BA" and "Laranja" clusters (Fig. 3a).

Both the maximum likelihood and quartet phylogenies recovered five groups, with 100% bootstrap support for all clades (Fig. 2a, Supplementary Information 1, Fig. S1). Relationships amongst the five clusters had low support in the ML analysis, whereas in the quartet analysis, the "Arruda" morphotype from Bahia ("Arruda-BA") was sister to the "Arruda" morphotype from Rio de Janeiro ("Arruda-RJ"), with 89% bootstrap support and the “Laranja” clade from Bahia was sister to the “North” clade with 76% bootstrap support (Supplementary Information 1, Fig. S1). Low support was consistent with very short branches near the base of the tree followed by long branches for each of the five genetic clusters. Of the two admixed individuals detected in our analysis in the ML likelihood analysis, *Lima et al. 7910* was found to be sister to the "Arruda-BA" clade, with bootstrap support above 95%, whereas *Santana et al. RB472149* was found to be sister to this clade + *Lima et al. 7910*, albeit with poor bootstrap support, below 60% (Fig. 2a). In the Quartet analysis, *Lima et al. 7910* was also found to be sister to the "Arruda-BA" clade, with 100% bootstrap support, whereas *Santana et al. RB472149* was found to be sister to the "Laranja" clade, with 74% bootstrap support (Supplementary Information 1, Fig. S1).

Morphometric analyses

The first and second axis of the PCA explained 38.9% and 27.6% respectively of the variance observed in the data. Leaflet length and rachis length were the two variables that provided the strongest contribution

to explaining the variance in PC1, whereas average pinna length contributed most to explaining the variance in PC2 (Fig. 3b). While there were no clearly segregated clusters of individuals that appeared in the PCA, the specimens representing the larger "Café" and "Laranja" morphotypes were all located on the left side of the graph, whereas specimens from all "Arruda" morphotypes clustered together with most cultivated specimens for which we could not determine their geographical origin. Specimens from the Northern Brazilian states overlapped with both the "Café" and "Arruda" morphotypes, as well as with cultivated specimens of unknown origin.

For the hierarchical clustering analysis, both the Calinski and Silhouette methods determined the ideal number of clusters in our morphometric dataset to be two (K = 2). The first of these clusters contained all the specimens from Southern Bahia attributed to the "Laranja" morphotype (Fig. 2c). This cluster also contained all the "Café" morphotypes, as well as five out of the nine specimens representing collections of non-cultivated individuals from the states of northern Brazil (Pernambuco, Alagoas, Rio Natal and Paraíba). Only a single specimen of the "Arruda" morphotype was also found to be nested within this cluster (*Lima et al. 7596*, Fig. 2c). The second cluster contained 21 of the 22 specimens of the "Arruda" morphotype, including all the specimens from the state of Rio de Janeiro. In addition, four out of nine specimens from the north-east of Brazil (Paraíba and Alagoas) were found within this cluster, as a sister clade to the "Arruda" morphotype. All other cultivated trees of uncertain provenance, including from Brasília, São Paulo, Minas Gerais, Rio de Janeiro, and Espírito Santo also clustered with this group.

Boxplots and results from the Kruskall-Wallis tests of the morphometric dataset, based on the genetic clusters retrieved, are shown in Fig. 4 and Table 2. Results from the Kruskall-Wallis test show that all morphometric variables, except pinna length, had significantly low p-values, below 0.05 (Table 2). For subsequent pairwise comparison tests, only the variable "leaflet length" supported at least half of the pairwise comparisons as having significant p-values. Notably, for this variable, the "Laranja" group was significantly different from all except the "Café" group. The "Café" group was only significantly different from the "Arruda-BA" and "Arruda-RJ" groups. The "North" group was significantly different from the "Laranja" and "Café" groups, but not the "Arruda-BA and "Arruda-RJ" groups (Table 2).

Table 2

Results of Kruskall-Wallis test, and pairwise Dunn comparison test. Threshold for p-values: 0.01. * test result significant (p-value below 0.01). NS: non-significant (p-value above 0.01). Rows coloured in grey contain comparisons between the Laranja group and other genetic lineages detected in this study.

Leaf trait		**Petiole length (cm)**	**Rachis length (cm)**	**Pinna length (cm)**	**Leaflet length (cm)**	**Leaflet length/width**
Result Kruskall-Wallis test		**** p = 0.00478**	**** p = 3.13e-06**	p = 0.0546	**** p = 3.25e-06**	**** p = 0.0001**
Group 1	**Group2**					
Arruda BA	Arruda RJ	**Ns**	**Ns**	–	**Ns**	**Ns**
Arruda BA	Café	**Ns**	**Ns**	–	****	****
Arruda BA	Laranja	**Ns**	*****	–	****	**Ns**
Arruda BA	North	**Ns**	**Ns**	–	**Ns**	****
Arruda RJ	Café	****	**Ns**	–	****	****
Arruda RJ	Laranja	**NS**	*****	–	****	**NS**
Arruda RJ	North	**NS**	**Ns**	–	**Ns**	****
Café	Laranja	**NS**	**Ns**	–	**Ns**	**Ns**
Café	North	**NS**	**Ns**	–	**Ns**	**Ns**
Laranja	North	**NS**	**Ns**	–	****	**NS**

Discussion

Whereas previous studies have examined the levels of inbreeding of populations of *P. echinata* using microsatellites, random amplified polymorphic DNA or amplified fragment length polymorphism from seven localities (Cardoso et al. 1998; Lira et al. 2003; Cardoso et al. 2005), our study is the first to use genomic DNA to examine the overall population structure of this tree species throughout the Atlantic Coast of Brazil, using representative specimens from all three putative morphotypes of *P. echinata*, including the rare "Laranja" and "Café" morphotypes from Southern Bahia and Espírito Santo. We expected to find three geographically structured genetic lineages corresponding to three morphotypes, but our analyses revealed a more complex picture.

First, the recovery of five distinct genetic clusters, instead of three, was confirmed by the ADMIXTURE, DAPC, and phylogenetic analyses (Figs. 2a-b, 3a, S1). Of these five groups, only the Laranja group corresponded to our initial hypothesis, although it did contain an individual that showed admixture with another genetic lineage, based on both the phylogenetic and ADMIXTURE analysis (Fig. 2a,b, Supplementary Information 1, Fig. S1). While specimens from the "Café" morphotype formed a single group, the specimen from Southern Bahia that appeared to also have medium-sized leaflets (*Santana et al. RB472149*) was actually found to be a hybrid between the "Laranja" group, and the "Arruda" morphotype from Bahia (Fig. 2a,b, Supplementary Information 1, Fig. S1). Furthermore, the "Arruda" group split into two geographical clusters, one containing specimens from Rio de Janeiro, and the other containing specimens from coastal Atlantic forests of the state of Bahia; one individual also showed indication of admixture (*Lima et al. 7910*). The fifth clade recovered contained all specimens from the north-eastern region of Brazil, as well as three cultivated specimens from various geographic locations, including São Paulo, Espírito Santo, and the Rio de Janeiro Botanical Garden. This was contrary to the initial hypothesis, which predicted based on morphology that they should cluster with other individuals from the "Arruda" or "Café" morphotype.

Considering the morphometric analysis, the k-means analysis combined with the Ward dendrogram (Fig. 2c) and the PCA analysis (Fig. 3b) also did not recover evidence to support the existence of three distinct morphotypes. On the one hand, the "Laranja", "Café" and "Arruda" morphotypes did group together in the majority in these analyses. However, samples from the Northern states of Brazil disrupted this pattern, clustering with both the "Arruda" and "Café" individuals in the clustering dendrogram (Fig. 2c) and overlapping with these groups in the PCA analysis (Fig. 3b). Both samples that clustered with the "Café" and "Arruda" morphotypes were sampled genetically and found to belong to the same genotype (Fig. 2). Furthermore, one admixed individual between the "Laranja" and "Arruda-BA" groups (*Santana et al. RB472149)* appeared to have an intermediate leaf morphology between these two groups that was easily confused with the "Café" morphotype (Fig. 3b). Finally, both the boxplots and Kruskall-Wallis analysis (Fig. 4, Table 2) revealed significant differences amongst the mean values of these five groups for four of our analysed traits, but subsequent pairwise analyses did not find unequivocal evidence for clear morphometric distinction amongst these groups. For example, for leaflet length, our results show that the "Laranja" morphotype has significantly larger leaflets than the "North" and "Arruda" morphotypes, but not the "Café" morphotype (Table 2). Our results allowed us to compare and find differences with previous claims about leaflet and pinnae length, and the number of pinnae and number of leaflets per pinnae for the previously proposed morphotypes. We found that specimens that we assigned as belonging to the “Laranja” morphotype (excluding admixed individual *Santana et al. RB472149*) have average leaflets sizes that vary between c. 5 to 12 cm, with a reduced number of pinna (2–3), and leaflets per pinnae (c. 4–6), whereas specimens assigned to the café morphotypes have leaflets c. 2.3 to 4.4 cm in length, and a higher average number of pinnae (2–4) and leaflets per pinnae (5–6). (Fig. 4, Table 1, Supplementary Information 3). Specimens assigned to the Arruda phenotype, from Rio de Janeiro, Bahia and Northern Brasil, had average leaflet sizes that varied between c. 1-2.84 cm, and had a much higher number of pinnae (4–9) and leaflets per pinnae (9–20)(Fig. 4, Table 1, Supplementary Information 3).

In summary, while our genetic and morphometric analyses did recover some genetic and morphological clusters that correspond notably to the "Café" and "Laranja" morphotypes, our analyses also show that the small-leafleted "Arruda" group separates into two genetic clusters that are not each other's closest relatives. Furthermore, the existence of a fifth genetic lineage whose leaflet size and morphology overlap with the "Arruda" and "Café" groups, explains why three infraspecific taxa of *P. echinata* have not been formally recognised to date. Pau Brasil has a more complicated genetic history than the three morphological leaflet variants would suggest. Equally, the evidence presented does not support the description of five distinctive taxonomic units, as it would be difficult to distinguish them based on leaflet morphology alone, and any taxonomic categories proposed would, in part, be cryptic with respect to morphology. Nevertheless, our results do provide much information about the phylogeography of Pau Brasil and bring novel information regarding the conservation of the species.

Implications for phylogeography and local adaptation

The strong geographical structuring found amongst different genetic and morphological clusters shows commonalities with other major biogeographical patterns and species community turnovers found in other plant and animal taxa of the Brazilian Atlantic Forest, recently reviewed in Peres et al. (2020). For example, the recovery of a genetically distinct population north of the São Francisco river, in Northern Brazil, is a pattern that has been observed in other species diversity patterns, including tree species composition (Prance 1982), passerine birds (Silva et al. 2004), and amphibians (Trevisan et al. 2020), as well as in epiphytes endemic to the Atlantic Forest (Menini Neto et al. 2016). Also concordant with known areas of endemism and diversity are the more southern populations of Pau Brasil, with the "Laranja" and "Arruda-BA" clusters agreeing with the "Central" area of endemism in the province of Bahia, and the more southern populations of the "Café" and "Arruda-RJ" groups mirroring the "Serra do Mar" area of endemism, typical of a marked turnover in species composition that occurs south of the River Doce (Carnaval and Moritz 2008; Peres et al. 2020).

Previous authors have hypothesised that *P. echinata* could have survived and prospered during glacial periods due to its preference for dry soils and an arid climate, surviving in patches of dry forest separated by a mosaic of wet forest that created natural isolation of the populations (Cunha and Lima, 1992; Lima et al., 2002). This hypothesis of forest refugia playing an important role in explaining overall biogeographic distribution patterns and high biological endemism in the Atlantic Forest is supported by studies which have suggested that climatic oscillation in the recent Pleistocene period has led to differences in the stability of different areas of the Atlantic Forest, with the Central Bahian and Northern regions being far more paleoclimatically stable in contrast to the forest South of the Rio Doce (Carnaval and Moritz 2008; Carnaval et al. 2014), which would have been more unstable during the Pleistocene. Other studies have pointed to the possible role of the expansion of the Atlantic shelf which would have contributed to the complex dynamics of the Atlantic Forest during the last Ice Age (Leite et al. 2016). Others have suggested that some of the biogeographical patterns could be linked to even earlier events prior to the late Pleistocene (Grazziotin et al. 2006; Thomé et al. 2010; Raposo do Amaral et al. 2013; Bacci et al. 2021).

Another possible explanation of the genetic and morphological differentiations of local populations of *P. echinata* could be the result of local adaptation to different soil types and climatic conditions across the Atlantic Forest Biome, including the differing environments in the restinga and tabuleiro forests. This is re-enforced by observation in common-garden plots in CEPEC showing that these different leaflet morphotypes are stable and not the result of genetic plasticity reacting to changes in environmental conditions (pers. comm. R. Santana). The adaptive significance of morphological variation in leaflet size and shape of compound leaves is usually attributed to optimizing thermal regulation of leaves in habitats with different environmental conditions (Leigh et al. 2017). Compound leaves are commonly found on large trees of the rainforest, as well as on understory plants (Malhado et al. 2010), as also seen in the ombrophilous forests where the "Laranja" and "Café" groups of Pau Brasil are found. In such environments with higher levels of humidity, larger leaflets can potentially reduce transpiration and minimize latent heat loss (Meinzer et al. 1995). In contrast, plants with smaller leaves or leaflets typically occur in drier areas (McDonald et al. 2003), and on soils that are nutrient poor (Fonseca et al. 2000) and/or saline (Ball et al. 1988), conditions encountered in coastal areas and dune-like ecosystems, such as restinga forests (Scarano 2002). However, studies have shown that such relationships are not straightforward (Leigh et al. 2017), and further investigation is desirable before we can conclude that leaflet size in *P. echinata* is an adaptation to local environments.

Finally, if we hypothesise that natural populations of Pau Brasil are locally adapted, the expectation is that other morphological traits adaptated to climatic and edaphic conditions would also vary across the different genetic lineages of the species. Macedo et al. (2019) did find differences in the wood anatomy of the "Laranja" and "Café" morphotypes compared with the “Arruda” morphotype, which in their study was defined as including specimens from populations in Rio de Janeiro, Bahia, and Northern Brazil. Notably, different populations of the "Arruda" morphotype possessed wood traits that suggested adaption to higher drought and increased temperature. These traits included a higher percentage of axial parenchyma and wider vessels in samples from Rio de Janeiro, which would guarantee higher conductance safety, whereas populations from the states of Rio Grande de Natal and Pernambuco had smaller solitary vessels and sparse axial parenchyma, which would help reduce the occurrence of embolism (Macedo et al. 2019).

While we cannot be certain of the extent to which phylogeographic history and local adaptation has shaped the population diversity of *P. echinata* in the Brazilian Atlantic Forest, future studies should focus on expanding population-level sampling which might detect candidate genes for local adaptation to these different environments (Capblancq and Forester 2021).

Implications for conservation

Previous studies have suggested that the populations of Pau Brasil could be divided into three genetically distinct and geographically distributed forms which should represent priority areas for conservation, corresponding to Northern Brazil, Southern Bahia, and Rio de Janeiro, suggesting that at the very least a single population from each area should be targeted for conservation (Cardoso et al., 2005). Our results,

based on increased sampling, suggest that such an approach would not protect the current morphological and genetic diversity of Pau Brasil and that additional populations from Espírito Santo and Southern Bahia should also be a focus for conservation.

Special attention should especially be given to the southern populations in Bahia and Espírito Santo, particularly with regards to populations belonging to the "Café" and "Laranja" morphotype, which are hyper-localised in distribution, and is currently not included in any protected area, contrarily to the other four lineages. This is even more important given that our analyses showed evidence in two individuals for genetic introgression between the "Laranja" and "Arruda-BA" genotypes. One of these putative hybrids was collected from the margin of the "Laranja" forest patch, and hybridization is assumed to have been between trees planted in the neighbouring cacao plantation. The second hybrid is an herbarium specimen, collected from the same region as the "Laranja" forest patch, but it is unclear whether hybridization has resulted from the planting of different genotypes close to each other (Fig. 2).

In addition, our morphometric analyses show that most specimens from cultivated trees group with the second cluster composed of specimens from the "Arruda" morphotypes and "North" clade. When considering the results from the population genomics analyses, where the cultivated specimens grouped with members from the "North" clade, one explanation is that cultivated trees in Brazil may be primarily derived from this genotype. As Pau Brasil is a rare and protected species in the wild, most of the cultivated seed stock in horticultural nurseries may be of similar geographical origin. It would not be surprising to find that further genetic sampling of cultivated trees in urban environments, natural plantations, and potentially also in nature reserves and protected areas are possibly propagated from a similar genetic stock, which based on our study seems to be from populations in Northern Brazil. This would make sense given that this area was once the historical centre for exploitation and export of Pau Brasil, and "Pernambouc" wood is an alternative common name for the species *Paubrasilia echinata*, recognising the state of Pernambuco in Northern Brazil (Macedo et al. 2020).

In order to verify these hypotheses, more specimens from cultivated urban trees, as well as from Cacao plantations, are needed for inclusion in future genetic analyses to confirm the hypothesis presented by the results of our analyses. The full extent to which hybridization occurs between different cultivated specimens of *P. echinata* is unknown and should be further investigated. We noticed that the majority of the “Laranja” populations included in our morphometric analyses have smaller leaflet sizes than *Lima et al.* 7905, the population that was included the genetic analyses (Supplementary Information 3). This could be a sign that these other populations might be experiencing some sort of introgression with other genetic lineages, as our results have shown that admixed individuals have smaller leaflet lengths. The effects of introducing novel alleles from geographically distant lineages could lead to outbreeding depression (Cardoso et al., 2005). While conservation of *P. echinata* as an urban street tree, or as a tree in cacao plantations, is one solution to the long-term survival of the species, it would be preferable to plant selected genotypes that representative local populations. Planting local genotypes might lead to more resilient trees adapted to local habitats.

Finally, if the distinctive large-leafleted populations are in patches of forest that represent past glacial refugia, or are local adaptations to ecological conditions, it is quite likely that the areas where they occur could also harbour several undescribed and localized endemic plant and animal taxa. Such localities would benefit from conservation and protection from further deforestation.

Challenges using herbarium specimens for morphometric analyses

Because we did not use the same specimens in the morphometric and genetic analyses, the comparison of the groups between the two datasets was complicated. Because *P. echinata* occurs in a highly fragmented and modified habitat, it is not always clear, especially on herbarium specimen labels, whether the specimens were collected in the wild or from introduced or cultivated trees. Nevertheless, many of the specimens in our study were collected by the co-authors (37 out of 56), giving confidence to their status as wild or cultivated.

Factors other than genotype can lead to phenotypic differences in leaflet size. In addition to natural variation amongst individuals within a species, leaves and leaflets from individuals growing in the shade tend to be larger than those growing in full sun (Nicotra et al. 2011) and freshly flushed leaves and leaflets can be smaller than fully mature ones. We attempted to account for this variation by using as many samples as possible for our morphometric analyses, and we measured several leaflets per herbarium specimen and found no correlation between leaflet size and collection month of the sample. We also examined other morphological traits, including fruit and flower morphology and size, but were unable to locate sufficient fertile herbarium specimens to include these structures in our morphometric analyses. Collection and observation of flowering and fruiting traits is not straightforward, given that phenology is partly reliant on sporadic rains in the dry season (Borges et al. 2009). Another reported trait variation is the colour of the heartwood (Cunha and Lima, 1992; Lewis, 1998), but this is not recorded on most herbarium specimens.

Conclusion

By combining morphological and genetic data from across the extant distribution of *Paubrasilia echinata*, we find evidence for at least five genetically distinct lineages, within the fragmented Atlantic Forest of Brazil. These genetic lineages are highly geographically structured, and while some of them seem to correspond to the previously defined "Laranja", "Café", and "Arruda" leaflet morphotypes, individuals from Northern Brazil were found to have morphometric leaf traits that overlap with both the "Café" and "Arruda" morphotypes. Further investigation into understanding the morphological differences amongst sub-populations of *P. echinata* should refrain from using the three traditionally defined morphotypes related to leaflet size, but rather focus on maximizing conservation of all five lineages detected in this study. The two lineages in most urgent need of further research and conservation are the rarest and less well-known "Café" and "Laranja" groups, that represent distinct evolutionary lineages and occur in different types of forests than the other populations. Whether these morphotype should be formally recognised as a distinct infraspecific taxon is an ongoing question. More work is notably needed

on the "Laranja" morphotype to understand how widespread the apparently rare genetic lineage is, and whether hybridization with a second evolutionary lineage from Bahia is a natural phenomenon or the result of human introduction of different genotypes of *P. echinata* across the geographical range of the species. A better understanding of the original source of individual cultivated trees in urban environments and cacao plantations, and their relationship to wild specimens from fragmented populations, will help to guide conservation strategies for Pau Brasil.

Statements And Declarations

Acknowledgements

We would like to thank Dr. Royce Steeves for advice on DNA sequencing, Dr. Anik Dutta for advice on bioinformatic analyses, as well as undergraduate research assistant Carlo Cormier for carrying out DNA extractions. We would also like to thank collaborators and partners at the Centro de Pesquisas do Cacau (CEPEC), in Ilhéus, Bahia, Brazil, including Robelio de Santana.

Funding

This work was, in part, made possible thanks to a grant from the Department of Biology of the University of Moncton, New Brunswick, as well as a postdoctoral fellowship to EG from the Fonds de recherche du Québec en Nature et technologies (FQRNT). Field work was made possible by a travel grant from the Biodiversity Research Centre of Quebec.The authors acknowledge the Research/Scientific Computing teams at The James Hutton Institute and NIAB for providing computational resources and technical support for the "UK"s *Crop Diversity* Bioinformatics HPC" (BBSRC grant BB/S019669/1), use which has contributed to the results reported within this paper.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Authors contribution

EG and HL carried out field work and conceptualised the paper; EG, HL and GPL selected herbarium specimens for analysis; EG carried out some of the DNA extractions and arranged for sequencing; MR carried out the morphometric measurements, and analysed the morphological and genetic data, under the guidance and supervision of EG and LN. MR and EG wrote the manuscript, with revisions from HdL, LN and GPL.

Data availability

The datasets and notebooks used in this paper are available in Supplementary Information, as well as on this online webpage: https://github.com/MyosotisMatt/Paubrasil

References

1. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664. https://doi.org/10.1101/gr.094052.109
2. Andrews S (2010) FASTQC. A quality control tool for high throughput sequence data. [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
3. Arseneau JR, Steeves R, Laflamme M (2017) Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues. Mol Ecol Resour 17(4):686–693. https://doi.org/10.1111/1755-0998.12616
4. Bacci LF, Reginato M, Bochorny T et al (2021) Biogeographic breaks in the Atlantic Forest: evidence for Oligocene/Miocene diversification in *Bertolonia* (Melastomataceae). Bot J Linn Soc 199:128–143. https://doi.org/10.1093/botlinnean/boab099
5. Ball MC, Cowan IR, Farquhar GD (1988) Maintenance of Leaf Temperature and the Optimisation of Carbon Gain in Relation to Water Loss in a Tropical Mangrove Forest. Funct Plant Biol 15:263–276. https://doi.org/10.1071/pp9880263
6. Borges LA, Sobrinho MS, Lopes AV (2009) Phenology, pollination, and breeding system of the threatened tree *Caesalpinia echinata* Lam. (Fabaceae), and a review of studies on the reproductive biology in the genus. Flora Morphol Distrib Funct Ecol Plants 204:111–130. https://doi.org/10.1016/j.flora.2008.01.003
7. Brown CM, Arbour JH, Jackson DA (2012) Testing of the Effect of Missing Data Estimation and Distribution in Morphometric Multivariate Data Analyses. Syst Biol 61:941–954. https://doi.org/10.1093/sysbio/sys047
8. Capblancq T, Forester BR (2021) Redundancy analysis: A Swiss Army Knife for landscape genomics. Methods Ecol Evol 12:2298–2309. https://doi.org/10.1111/2041-210x.13722
9. Cardoso MA, Provan J, Powell W et al (1998) High genetic differentiation among remnant populations of the endangered *Caesalpinia echinata* Lam. (Leguminosae-Caesalpinioideae). Mol Ecol 7:601–608. https://doi.org/10.1046/j.1365-294x.1998.00363.x
10. Cardoso SRS, Provan J, Lira C da et al (2005) High levels of genetic structuring as a result of population fragmentation in the tropical tree species *Caesalpinia echinata* Lam. Biodivers Conserv 14:1047–1057. https://doi.org/10.1007/s10531-004-8409-z
11. Carnaval AC, Moritz C (2008) Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest. J Biogeogr 35:1187–1201. https://doi.org/10.1111/j.1365-2699.2007.01870.x
12. Carnaval AC, Waltari E, Rodrigues MT et al (2014) Prediction of phylogeographic endemism in an environmentally complex biome. Proc R Soc B Biol Sci 281:20141461. https://doi.org/10.1098/RSPB.2014.1461
13. CITES (2007) *Caesalpinia echinata*. In: Cop 14 Prop. 30. The Hague, pp 1–11
14. Cunha MW, Lima CL (1992) Viagem à Terra Do Pau-brasil. Agência Brasileira DeCultura, Petrobrás

15. Cupertino FB, Melo SCO, Corrêa RX (2014) Genetic similarity among *Caesalpinia echinata* trees and their ex situ conservation. Int J Biodivers Conserv 5:655–660. https://doi.org/10.5897/ijbc11.273
16. Danecek P, Bonfield JK, Liddle J et al (2021) Twelve years of SAMtools and BCFtools. Gigascience 10:2. https://doi.org/10.1093/gigascience/giab008
17. Eaton DAR, Overcast I(2020) ipyrad: Interactive assembly and analysis of RADseq datasets. Bioinformatics
18. Eaton DAR, Spriggs EL, Park B, Donoghue MJ (2017) Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. Syst Biol 66:399–412. https://doi.org/10.1093/sysbio/syw092
19. Francis RM (2017) pophelper: an R package and web app to analyse and visualize population structure. Mol Ecol Resour 17:27–32. https://doi.org/10.1111/1755-0998.12509
20. Fonseca CR, Overton JM, Collins B, Westoby M (2000) Shifts in trait-combinations along rainfall and phosphorus gradients. J Ecol 88:964–977. https://doi.org/10.1046/j.1365-2745.2000.00506.x
21. Gagnon E, Bruneau A, Hughes CE et al (2016) A new generic system for the pantropical Caesalpinia group (Leguminosae). PhytoKeys 71:1–160. https://doi.org/10.3897/phytokeys.71.9203
22. Gagnon E, Lewis GP, Lima HC de(2020) *Paubrasilia*. In: Flora do Bras. 2020. http://reflora.jbrj.gov.br/reflora/floradobrasil/FB102193. Accessed 6 Apr 2021
23. Grazziotin FG, Monzel M, Echeverrigaray S, Bonatto SL (2006) Phylogeography of the *Bothrops jararaca* complex (Serpentes: Viperidae): past fragmentation and island colonization in the Brazilian Atlantic Forest. Mol Ecol 15:3969–3982. https://doi.org/10.1111/j.1365-294X.2006.03057.x
24. GRDC (2020) Major River Basins of the World / Global Runoff Data Centre, GRDC. 2nd, rev. ext. Federal Institute of Hydrology (BfG), ed. Koblenz, Germany
25. Habel JC, Rasche L, Schneider UA et al (2019) Final countdown for biodiversity hotspots. Conserv Lett 12. https://doi.org/10.1111/conl.12668
26. Jombart T, Bateman A (2008) Adegenet: A R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405. https://doi.org/10.1093/bioinformatics/btn129
27. Jombart T, Devillard S, Balloux F Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11: 94. https://doi.org/10.1186/1471-2156-11-94Josse Husson J(2010) F (2016) missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. J Stat Softw 70:. https://doi.org/10.18637/jss.v070.i01
28. Juchum FS, Costa MA, Amorim AM, Corrêa RX (2008) Phylogenetic relationships among morphotypes of *Caesalpinia echinata* lam. (Caesalpinioideae: Leguminosae) evidenced by trnL intron sequences. https://doi.org/10.1007/s00114-008-0424-6. Naturwissenschaften
29. Kassambara A, Mundt F(2020) Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7. https://CRAN.R-project.org/package=factoextra
30. Knaus BJ, Grünwald NJ VCFR: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources, 17:44–53, http://dx.doi.org/10.1111/1755-0998.12549. Leigh A,

Sevanto S, Close JD, Nicotra AB(2017) (2017) The influence of leaf size and shape on leaf thermal dynamics: does theory hold up under natural conditions? Plant Cell Environ 40:237–248. https://doi.org/10.1111/pce.12857

31. Leigh A, Sevanto S, Close JD, Nicotra AB (2017) The influence of leaf size and shape on leaf thermal dynamics: does theory hold up under natural conditions? Plant Cell Environ 40:237–248. https://doi.org/10.1111/pce.12857
32. Leite YLR, Costa LP, Loss AC et al(2016) Neotropical forest expansion during the last glacial period challenges refuge hypothesis. Proceedings of the National Academy of Sciences 113:1008–1013. https://doi.org/10.1073/pnas.1513062113
33. Lewis GP (1998) *Caesalpinia*: a revision of the Poincianella-Erythrostemon group. Royal Botanic Gardens, Kew
34. Lichtenberg S, Huber-Sannwald E, Nehren U, Reyes-Agüero JA (2019) Use and Conservation of the Threatened Brazilian National Tree *Paubrasilia echinata* Lam.: A Potential for Rio de Janeiro State? In: Nehren U, Schlüter S, Raedig C et al (eds) Strategies and Tools for a Sustainable Rural Rio de Janeiro. Springer International Publishing, Cham, pp 205–219. https://doi.org/10.1007/978-3-319-89644-1_14
35. Lima HC, Lewis GP, Bueno E(2002) Pau-brasil: uma biografia. In: Bueno E (ed) Pau-brasil, Axis Mundi, Sa˜o Paulo, pp 39–76
36. Lira CF, Cardoso SRS, Ferreira PCG et al (2003) Long-term population isolation in the endangered tropical tree species *Caesalpinia echinata* Lam. revealed by chloroplast microsatellites. Mol Ecol 12:3219–3225. https://doi.org/10.1046/j.1365-294X.2003.01991.x
37. Macedo TM, Costa CG, de Lima HC, Barros CF (2020) Wood anatomy of historic French violin bows made of Pernambuco wood. IAWA J 41:320–332. https://doi.org/10.1163/22941932-bja10011
38. Macedo TM, de Lima HC, de Souza ND et al (2019) Infraspecific variation of *Paubrasilia echinata* (Fabaceae) wood along a latitudinal gradient in Brazil. Flora 258:151437. https://doi.org/10.1016/j.flora.2019.151437
39. Malhado ACM, Whittaker RJ, Malhi Y et al (2010) Are compound leaves an adaptation to seasonal drought or to rapid growth? Evidence from the Amazon rain forest. Glob Ecol Biogeogr 19:852–862. https://doi.org/10.1111/j.1466-8238.2010.00567.x
40. McDonald PG, Fonseca CR, Overton JM, Westoby M (2003) Leaf-size divergence along rainfall and soil-nutrient gradients: is the method of size reduction common among clades? Funct Ecol 17:50–57. https://doi.org/10.1046/j.1365-2435.2003.00698.x
41. Meinzer FC, Goldstein G, Jackson P et al (1995) Environmental and physiological regulation of transpiration in tropical forest gap species: the influence of boundary layer and hydraulic properties. Oecologia 101:514–522. https://doi.org/10.1007/BF00329432
42. Melo SCO, Gaiotto FA, Cupertino FB et al (2007) Microsatellite markers for *Caesalpinia echinata* Lam. (Brazilwood), a tree that named a country. Conserv Genet 8:1269–1271. https://doi.org/10.1007/s10592-006-9274-y

43. Menini Neto L, Furtado SG, Zappi DC et al (2016) Biogeography of epiphytic angiosperms in the Brazilian Atlantic forest, a world biodiversity hotspot. Rev Bras Bot 39:261–273. https://doi.org/10.1007/s40415-015-0238-7
44. Morellato LPC, Haddad CFB (2000) Introduction: The Brazilian Atlantic Forest. Biotropica 32:786. https://doi.org/10.1646/0006-3606(2000)032[0786:itbaf]2.0.co;2
45. Neves DM, Dexter KG, Pennington RT et al (2017) Dissecting a biodiversity hotspot: The importance of environmentally marginal habitats in the Atlantic Forest Domain of South America. Divers Distrib 23:898–909. https://doi.org/10.1111/ddi.12581
46. Nicotra AB, Leigh A, Boyce CK et al (2011) The evolution and functional significance of leaf shape in the angiosperms. Funct Plant Biol 38:535–552. https://doi.org/10.1071/FP11057
47. Oksanen J, Blanchet FG, Kindt R et al(2013) vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan
48. Peres EA, Pinto-da-Rocha R, Lohmann LG(2020) Patterns of species and lineage diversity in the Atlantic Rainforest of Brazil. In: Rull and Carnaval (eds) Neotropical diversification: patterns and processes. Springer, Cham, pp. 415–447
49. Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253. https://doi.org/10.1371/journal.pone.0032253
50. Prance GT (1982) Forest refuges: evidence from woody angiosperms. In: Prance GT (ed) Biological Diversification in the Tropics. Columbia Univ. Press, New York, pp 137–158
51. R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
52. Raposo do Amaral F, Albers PK, Edwards SV, Miyaki CY (2013) Multilocus tests of Pleistocene refugia and ancient divergence in a pair of Atlantic Forest antbirds (*Myrmeciza*). Mol Ecol 22:3996–4013. https://doi.org/10.1111/mec.12361
53. Ribeiro MC, Metzger JP, Martensen AC et al (2009) The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. Biol Conserv 142:1141–1153. https://doi.org/10.1016/j.biocon.2009.02.021
54. Rodrigues PS, Souza MM, Melo CAF et al (2018) Karyotype diversity and 2C DNA content in species of the Caesalpinia group. BMC Genet. https://doi.org/10.1186/s12863-018-0610-2
55. Scarano FR (2002) Structure, function and floristic relationships of plant communities in stressful habitats marginal to the Brazilian Atlantic rainforest. Ann Bot 90:517–524. https://doi.org/10.1093/aob/mcf189
56. Silva JMC, Sousa MC, Castelletti CHM (2004) Areas of endemism for passerine birds in the Atlantic forest, South America. Glob Ecol Biogeogr 13:85–92. https://doi.org/10.1111/j.1466-882X.2004.00077.x
57. Snir S, Rao S (2012) Quartet MaxCut: A fast algorithm for amalgamating quartet trees. Mol Phylogenet Evol 62:1–8. https://doi.org/10.1016/j.ympev.2011.06.021

58. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033
59. Tang Y, Horikoshi M, Li W (2016) Ggfortify: Unified interface to visualize statistical results of popular r packages. R J 8:478–489. https://doi.org/10.32614/rj-2016-060
60. The Brazil Flora Group (2022) Brazilian Flora 2020: Leveraging the power of a collaborative scientific network. Taxon 71:178–198. https://onlinelibrary.wiley.com/doi/full/10.1002/tax.12640
61. Thomé MTC, Zamudio KR, Giovanelli JGR et al (2010) Phylogeography of endemic toads and post-Pliocene persistence of the Brazilian Atlantic Forest. Mol Phylogenet Evol 55:1018–1031. https://doi.org/10.1016/j.ympev.2010.02.003
62. Trevisan CC, Batalha-Filho H, Garda AA et al (2020) Cryptic diversity and ancient diversification in the northern Atlantic Forest Pristimantis (Amphibia, Anura, Craugastoridae). Mol Phylogenet Evol 148:106811. https://doi.org/10.1016/j.ympev.2020.106811
63. Valero-Mora PM(2010) ggplot2: Elegant Graphics for Data Analysis. J Stat Softw 35. https://doi.org/10.18637/jss.v035.b01
64. Van-Lume B, Esposito T, Diniz-Filho JAF et al (2017) Heterochromatic and cytomolecular diversification in the Caesalpinia group (Leguminosae): Relationships between phylogenetic and cytogeographical data. Perspect Plant Ecol Evol Syst 29:51–63. https://doi.org/10.1016/j.ppees.2017.11.004
65. Varty(1998) *Caesalpinia echinata*. The IUCN Red List of Threatened Species 1998: e.T33974A9818224. https://www.iucnredlist.org/species/33974/9818224
66. Weeks JP (2010) Plink: An R package for linking mixed-format tests using IRT-based methods. J Stat Softw 35:1–33. https://doi.org/10.18637/jss.v035.i12
67. Yu G, Smith DK, Zhu H et al (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. https://doi.org/10.1111/2041-210X.12628

Figures

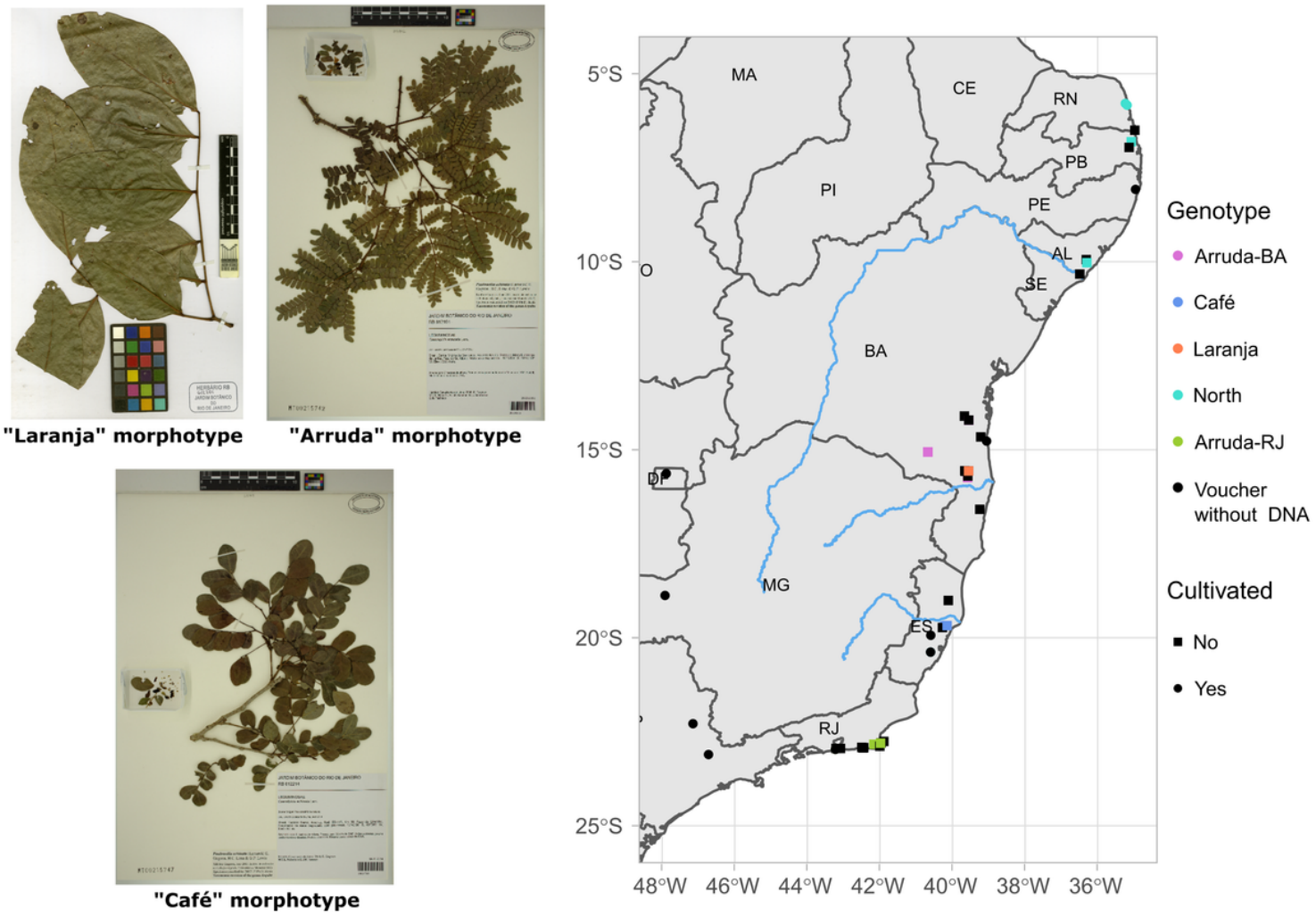


Figure 1

Herbarium specimens and distribution of *Paubrasilia echinata* samples used for the morphometric and genetic studies, excluding cultivated specimens of unknown provenance. Coloured points represent genetic clusters of field-collected material, as shown on Fig. 2b. Black points represent herbarium specimens used only in the morphometric analyses. Rivers shown on the map, from North to South, are the Rio Sao Francisco, Rio Jequitinhonha, and Rio Doce (taken from GRDC, 2020). Inset on the left shows the morphological plasticity in *P. echinata* displayed by herbarium specimens of the three commonly identified morphotypes. From top to bottom, "Laranja" (*Lima et al. 7905*, Bahia), "Arruda" (*Lima et al. 7896*, Bahia) and "Café" morphotypes (*Lima et al. 7918*, Espirito Santo)

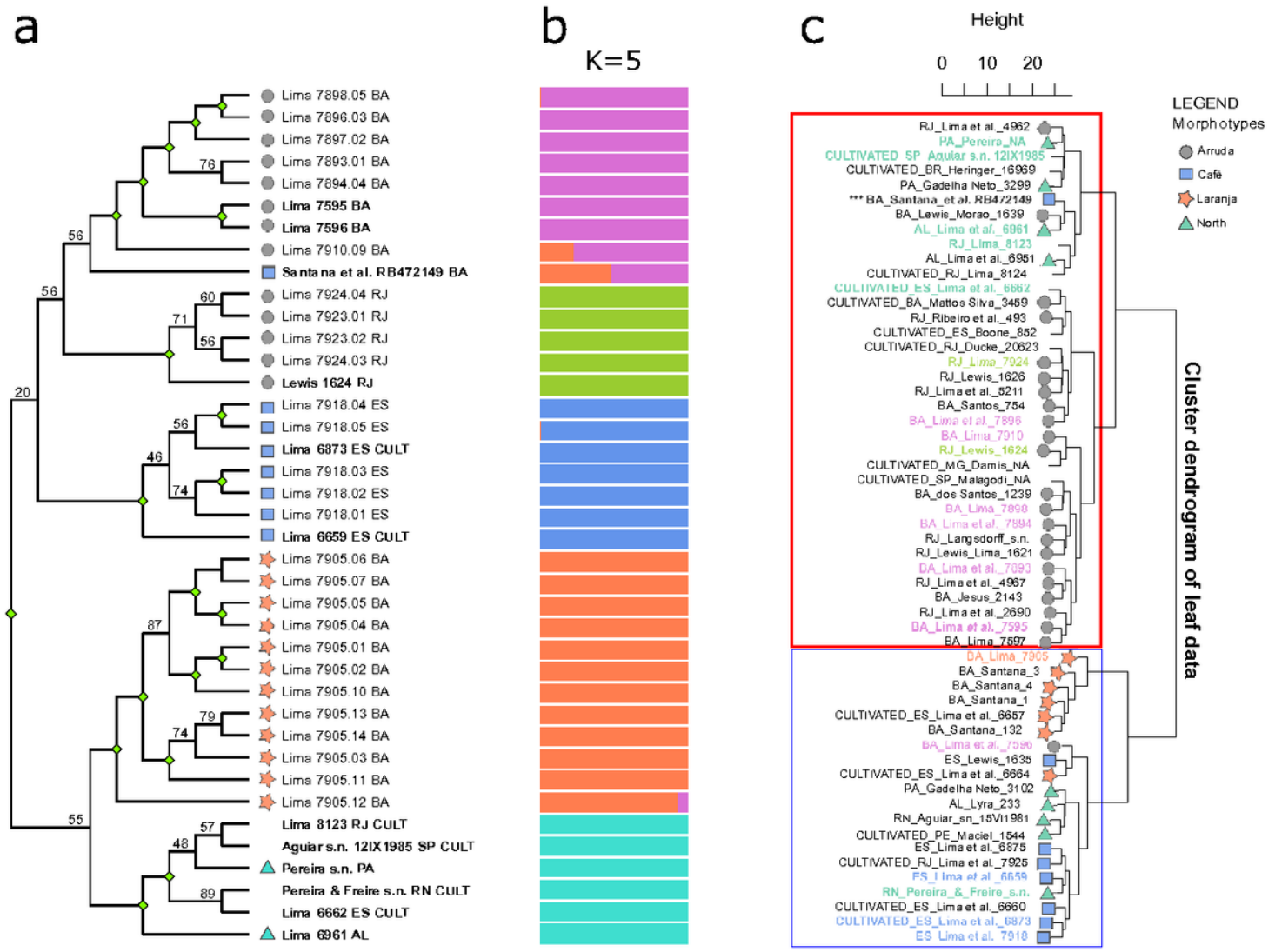


Figure 2

Comparison of phylogenies and clustering of genetic dataset with results of the hierarchical clustering dendrogram of the morphometric dataset. (a) Maximum likelihood phylogenetic tree, with outgroup pruned out. Bootstrap support values are displayed at each node, except for values of 95% and above, which are represented by green dots. Samples in bold represent herbarium specimens. Assigned morphotypes are indicated at the tip of the trees, as listed in the legend of Fig. 2c. Tips indicate collector, collector number, state, as well as whether specimens are cultivated or not. (b) ADMIXTURE plots showing posterior probability of assignment to each clade, for K=5. (c) Hierarchical clustering of morphological data using Ward distance on transformed variables. Boxes indicate the clusters found by the Calinski and Silhouette criterion, for K=2. Samples included in the genetic analyses are indicated by bold tips, coloured according to the five clusters suggested by the ADMIXTURE analysis. Assigned morphotypes are indicated at the tip, as listed in the legend of the figure. Tips indicate whether specimens are cultivated or not, state of collection, collector, and collector number.

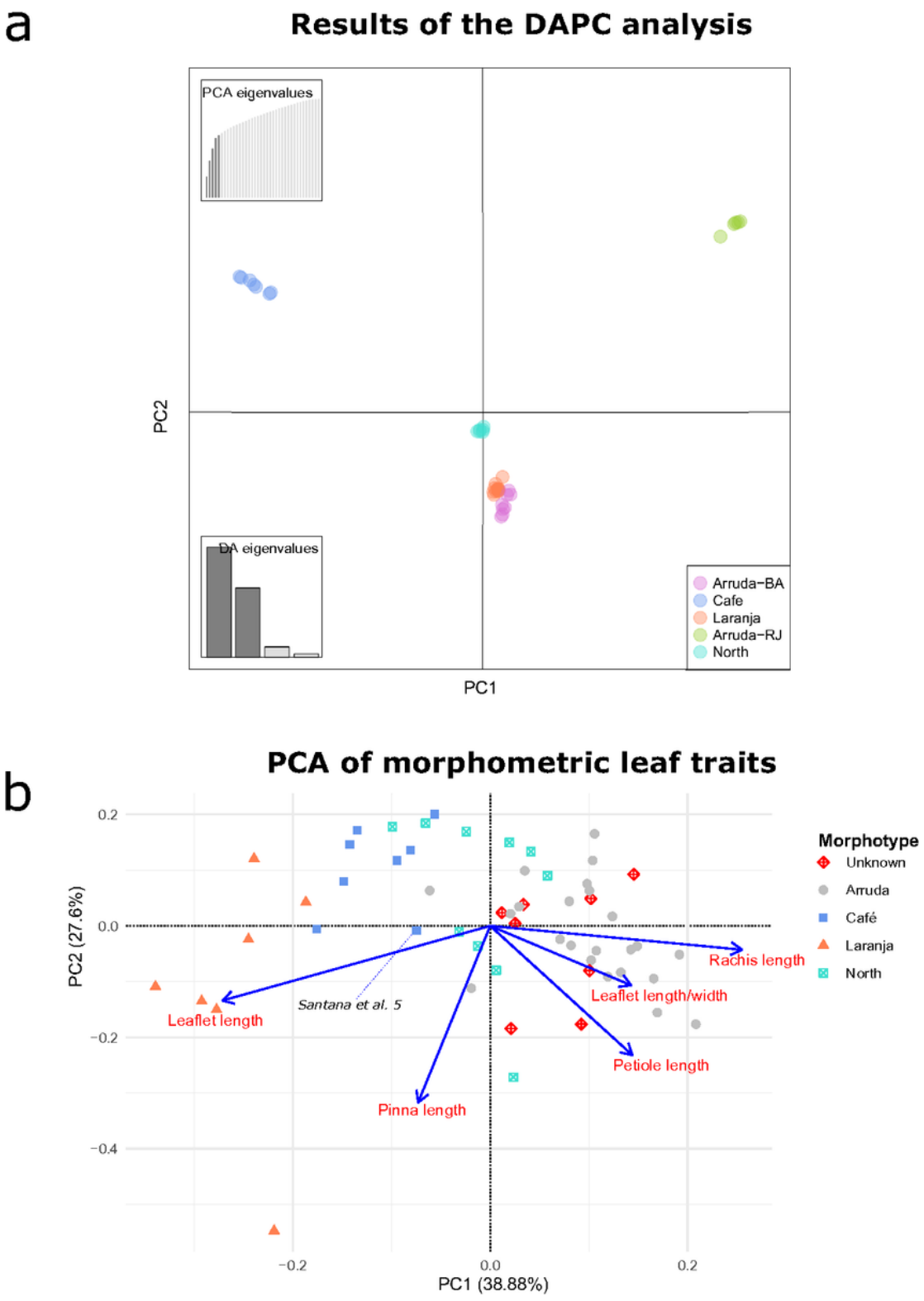


Figure 3

Results of the DAPC analysis and the PCA analysis of morphometric traits. (a) DAPC analysis of Genotype-by-sequencing data, with outgroup removed (39 individuals). Top left inset shows results of the cumulative PCA eigenvalues; bottom left inset shows discriminant analysis eigenvalues for four axes. (b) PCA of five morphometric leaf traits. Arrows indicate projection of variables in PCA space. One admixed individual, *Santana et al. 5*, is indicated on the PCA.

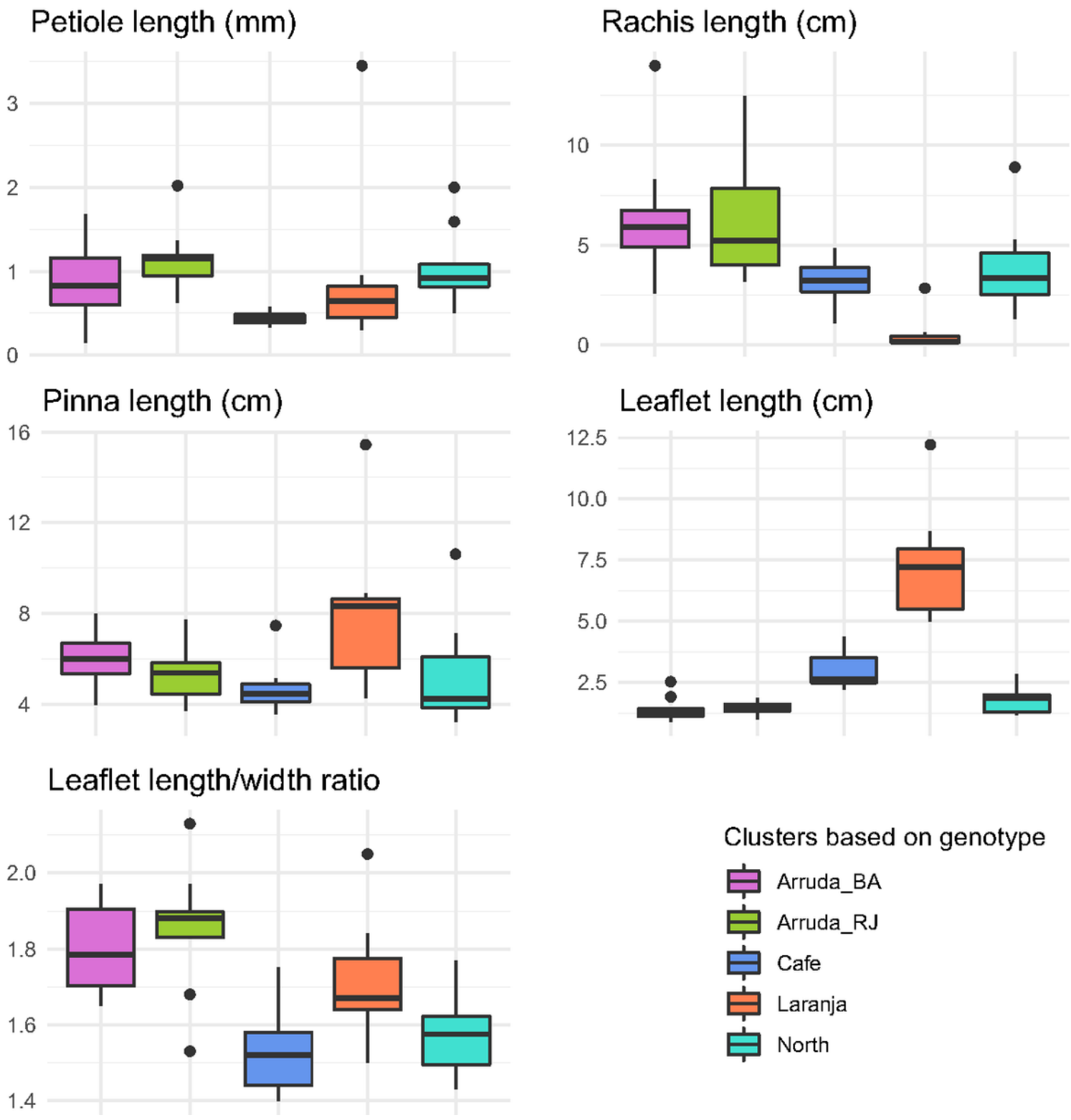


Figure 4

Box plots of six leaf traits based on genetic clusters. Cultivated specimens of unknown provenance removed, as well as two admixed individuals (*Lima et al. 7910* and *Santana et al. 5*)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation1FINAL.pdf
- SupplementaryInformation2FINAL.xlsx
- SupplementaryInformation3Final.xlsx