Received: 26 January 2023 | Accepted: 21 August 2023

DOI: 10.1002/ajb2.16241

RESEARCH ARTICLE

Botany

Phylogenomic and morphological data reveal hidden patterns of diversity in the national tree of Brazil, *Paubrasilia echinata*

Mathew Rees[1,2] | **Linda E. Neaves**[1,3] | **Gwilym P. Lewis**[4] | **Haroldo C. de Lima**[5,6] | **Edeline Gagnon**[7,8]

[1]Tropical Diversity, Royal Botanic Garden Edinburgh, Edinburgh, EH5 3LR, UK

[2]School of GeoSciences, University of Edinburgh, Edinburgh, EH9 3FF, UK

[3]Fenner School of Environment & Society, Australian National University, Australian Capital Territory, Australia

[4]Accelerated Taxonomy Department, Royal Botanic Gardens, Kew, Richmond, TW9 3AE, UK

[5]Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Rua Pacheco Leão, 915, Rio de Janeiro, RJ 22460-030, Brazil

[6]Instituto Nacional da Mata Atlântica/INMA-MCTI, Av. José Ruschi, 4, Centro, Santa Teresa, Espírito Santo, Brazil

[7]Chair of Phytopathology, TUM School of Life Sciences, Technical University of Munich, 85352 Freising-Weihenstephan, Germany

[8]Department of Integrative BIology, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

Correspondence

Mathew Rees, Tropical Diversity, Royal Botanic Garden Edinburgh, Edinburgh, EH5 3LR, UK.
Email: mathew.rees@ed.ac.uk

Abstract

Premise: *Paubrasilia echinata* (common names, pau brasil, brazilwood) is the national tree of Brazil and an endangered species endemic to the Brazilian Atlantic Forest. Over its wide distribution of 2000 km, its leaflets morphology exhibits extensive plasticity. Three morphotypes are commonly identified based on leaf size, 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.

Methods: A morphometric study of leaf characters of herbarium specimens was coupled with genetic analyses using genotype-by-sequencing data. We used maximum-likelihood and coalescent methods to evaluate the phylogenetic and population structure of the species. We compared these with a morphological dendrogram built from hierarchical clustering.

Results: Two of the three morphotypes formed separately evolving lineages, the third morphotype formed two geographically separate lineages, and northern trees with intermediate leaf morphology formed a separate fifth lineage. 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.

Conclusions: Our genetic and morphological results provide some evidence that cultivated trees from different states in Brazil seem to have a limited genetic origin and do not reflect the broader genetic and geographical diversity of the species. As a result, more care is likely needed to preserve the overall genomic diversity of this endangered and iconic species.

KEYWORDS

Brazilian Atlantic Forest, Fabaceae, Leguminosae, Mata Atlantica, pau brasil, population genomics, RAD-seq

The Atlantic Forest of Brazil is one of the 36 biodiversity hotspots in the world (Habel et al., 2019), and 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 (Brazil Flora Group, 2022). Part of this diversity is due to the environmental heterogeneity of the Atlantic Forest, which includes a variety of marginal habitats, such as 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, semideciduous, small to medium-sized tree (c. 4–20 m tall) that grows in inland ombrophilous mesic forests and in a range of much drier habitats, including disturbed dry coastal cactus scrub,

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *American Journal of Botany* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

rocky outcrops, seasonally semideciduous 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. Although it was previously thought to be relatively common throughout the Atlantic Forest (Lima et al., 2002), current populations are highly fragmented, notably due to centuries of exploitation, first for extraction of a red dye and subsequently for manufacturing 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 and in monospecific or mixed agroforestry plantations, potentially to support future wood demand from musical instrument manufacturers, especially for violin bows (Lichtenberg et al., 2019).

Several groups have hypothesized that *P. echinata* is a species complex (Lewis, 1998; Lima et al., 2002; Juchum et al., 2008; Gagnon et al., 2016; Rodrigues et al., 2018; Macedo et al., 2019), composed of three morphotypes that are geographically and ecologically structured. The first is a common and widely distributed small-leafleted variant called *folha-pequena* or *folha de arruda* (small-leaf or rue-leaf in English; hereafter, arruda), found 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 (Figure 1; Macedo et al., 2019). The second is a medium-leafleted variant named *folha-média* or *folha de café* (medium-sized-leaf or coffee-leaf in English, hereafter café), found predominantly in the states of Espírito Santo and southern Bahia (Figure 1; Macedo et al., 2019). The third is an extremely rare and localized large-leafleted variant folha de laranja ("orange-leaf in English, hereafter laranja), found in restricted populations along the Rio Pardo Valley, in the state of Bahia (Figure 1;

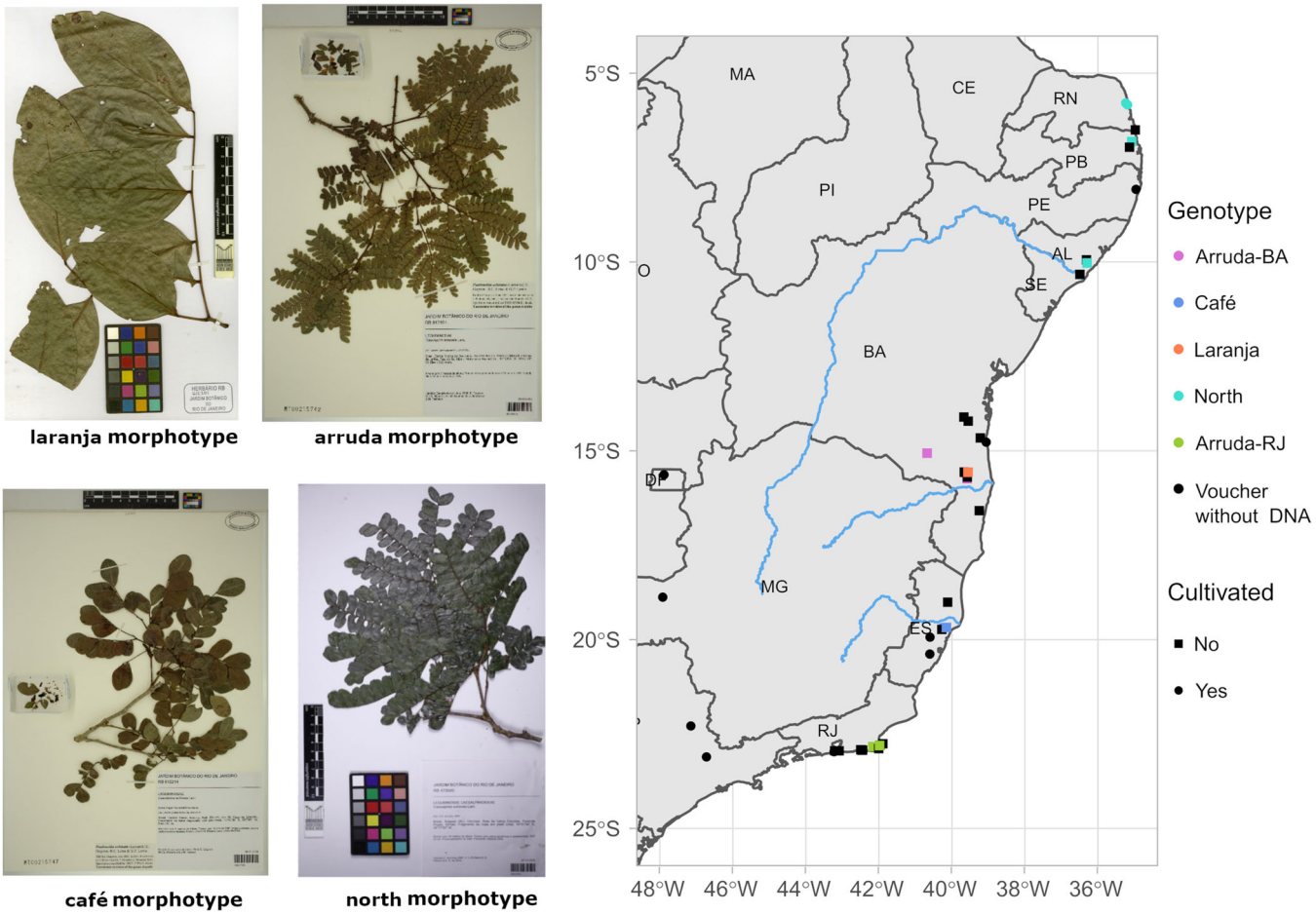


FIGURE 1 Herbarium specimens and distribution of *Paubrasilia echinata* samples used for the morphometric and genetic studies, excluding cultivated specimens of unknown provenance. Inset on the left shows the morphological plasticity in *P. echinata* displayed by herbarium specimens. From top left to bottom right, the laranja morphotype from Bahia (*Lima 7905*), the arruda morphotype from Bahia (*Lima 7896*), the café morphotype from Espírito Santo (*Lima 7918*) and the north morphotype from Alagoas (*Lima 6961*). The laranja morphotype has a bipinnate leaf with two pinnae and each pinna has a few large leaflets. Other specimens of the arruda, café and north morphotypes have bipinnate leaves with multiple pinnae and leaflets. On the right, the distribution map shows circles as cultivated specimens and squares are wild-collected specimens. Colored markers represent genetic clusters of field-collected material, as shown on Figure 2B. Black markers represent herbarium specimens used only in the morphometric analyses. Rivers on the map, from north to south: Rio Sao Francisco, Rio Jequitinhonha, and Rio Doce (from GRDC, 2020).

Lewis, 1998; Juchum et al., 2008), predominantly in ombrophilous mesic forests. However, there are other populations of pau brasil in the northern Brazilian states of Rio Grande do Norte, Paraiba, Pernambuco and Alagoas, which display a wider variation in leaflet size (Figure 1); the latter have been hypothesized to belong to the arruda morphotype (Lewis, 1998; Lima et al., 2002).

There is some evidence that leaflet size in pau brasil is not due to phenotypic plasticity; researchers from the Centro de Pesquisas do Cacau (CEPEC) have been growing seedlings from these three morphotypes side by side and observed that the differences in leaflet size did not appear to be related to environmental factors (Robelio de Santana, CEPEC, personal communication). Furthermore, bow makers often use the names sartori, peccati, and pernambuco, which reflect different characteristics of the wood used to make violin bows (Macedo et al., 2020). These sometimes correspond to specific populations; for example, peccati wood refers specifically to wood from trees with the café morphotype in Espírito Santo. Significant differences in heartwood color have also been used to help distinguish the three varieties (Lewis, 1998; Lima et al., 2002). While studies based on wood anatomy and density (Macedo et al., 2019) have confirmed these differences, they did not find sufficient evidence to justify the description of new taxa. Studies looking at karyotype diversity and genome size have also not found any support for segregation into different species or subspecies (Van-Lume et al., 2017; Rodrigues et al., 2018). Furthermore, the geographical distribution and morphological characters that help diagnose and delimit the three morphotypes based on leaflet size have never been fully described. There have been suggestions that the arruda phenotype not only has smaller leaflets, but also more pinnae (5–10) per leaf and more leaflets (12–21) per pinna; the pinnae are approximately 4 cm long (Macedo et al., 2019). The café morphotype has larger leaves, but fewer pinnae per leaf (3–5) and fewer leaflets per pinna (3–8); the pinnae are approximately 7 cm in length (Macedo et al., 2019). For the laranja morphotype, descriptions in the literature only mention that leaflets can reach 12 cm long; no other details are given (Lewis, 1998; Lima et al., 2002).

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 taxa (either varieties, subspecies, or species in their own right). For example, the widespread arruda morphotype might actually represent several distinct lineages, given the high amount of genetic structuring that has been observed in other taxa of the Atlantic Forest (reviewed by Peres et al., 2020). Cardoso et al. (1998), Lira et al. (2003), and Cardoso et al. (2005), investigating the structure and genetic diversity in remnant populations of pau brasil, found that they were genetically highly differentiated and formed three groups, corresponding to a clade containing all the southern populations located in the state of Rio de Janeiro, a clade with populations from the northeast of Brazil, and a third group containing populations from southern Bahia and Espírito Santo. However, relationships between these groups were poorly supported, and none of these studies included information about which morphotype(s) were sampled, nor included citation of specimen vouchers deposited in herbaria. Thus, their results cannot be used to determine whether they support the monophyly of the three morphotypes described in *P. echinata*.

To investigate whether the three known morphotypes of *P. echinata* correspond to distinct evolutionary lineages, we combined population genomics and morphometric analyses of field-collected and herbarium samples from across the entire range of *P. echinata*, including for the first-time populations of the rare laranja morphotype. Specifically, we addressed 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 used these results to evaluate whether infraspecific taxa or separate species should be formally described and discuss how they bring new insight into the conservation of these morphotypes.

MATERIALS AND METHODS

Taxon sampling for molecular analyses

Twenty-eight specimens were collected in November 2014 from wild populations of *P. echinata* growing in remnant patches of the Brazilian Atlantic Forest. These included the arruda morphotype from southern Bahia (hereafter arruda-BA, five individuals/two localities) and Rio de Janeiro (hereafter arruda-RJ, four individuals/two localities), the café morphotype from Espírito Santo (five individuals/one locality), and the laranja morphotype from a forest patch in Bahia (13 individuals/one locality) (Figure 1; Appendix S1, Table S1). 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*). From all localities, we collected individuals along a transect from the center of the forest fragment to its margin. In addition to field-collected specimens, we sampled DNA from 13 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, 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. 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. 8123*) was collected in the Rio de Janeiro Botanical Garden arboretum, where several morphotypes of pau brasil are cultivated.

In summary, our full genomic data set included 48 samples of herbarium and field-collected specimens. Thirteen were assigned to the laranja morphotype, seven to the café morphotype, and 14 to the arruda morphotype (nine arruda-BA and five arruda-RJ), plus an additional four samples from northern Brazilian populations of *P. echinata* (north group), and three cultivated specimens of unknown provenance. The outgroup comprised seven taxa of the genus *Cenostigma* Tul. that were sampled mostly from herbarium material (Appendix S1, Table S1).

DNA extraction and sequencing

DNA was extracted from samples using 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 with high levels of secondary metabolites. Library preparation and sequencing were carried out at the Institute of Integrative Biology and Systems (IBIS; University of Laval, Canada) using a modified genotyping-by-sequencing (GBS) approach with single end reads and two restriction enzymes: the rare enzyme PstI (CTGCAG) and common 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 noncoding regions. The samples were sequenced on one lane of an Illumina HiSeq 4000 (Illumina, Foster City, CA, USA) 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 a second sample (Lima 7905_15) was contaminated. Both were removed from further analyses (Appendix S2). All the generated genomic data are 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, at 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 restriction-site-associated DNA (RAD) tags, by specifying the data type as "ddrad". Because the three morphotypes of *P. echinata* were confirmed 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 data set with the outgroup (46 samples), and on a second data set 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; Appendix S2). For the data set containing the outgroup, 7165 loci were retained after filtering, with 41,496 SNPs (26.14% missing sites); for the data set with the outgroup pruned, 7123 loci were retained, with 31,439 SNPs (18.45% missing sites) (Appendix S2).

Population genomics and phylogenetic analyses

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 $r^2 < 0.2$.

Phylogenetic analyses of the entire data set, including the outgroup taxa, were carried out using maximum likelihood (ML) and quartet inference methods. The ML 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 were 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 Quartet MaxCut (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 ggtree v. 3.2.0 (Yu et al., 2017) using R v. 4.1.0 (R Core Team, 2021).

We used the maximum likelihood algorithm in ADMIXTURE v.1.3.0 (Alexander et al., 2009) to determine the number of genetic clusters present in the data, using the data set pruned of the outgroup. 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 visualized using custom R scripts based on the R 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 data sets 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 adegent v.2.1.8 (Jombart and Bateman, 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 Bayesian information criterion (BIC) value. To select how many principal components should be retained to maintain a high discriminant power without overfitting the data, we used a cross-validation method and the xvalDapc() function, with 1000 repetitions, and the method-based calculation of the a.score using the optim.a.scores() function.

Morphometric analyses

The morphological analysis used a total of 79 herbarium specimens from five herbaria (K, MT, MO, NY, RB; Appendix S3). When accounting for duplicate specimens, they represented 56 individual collections, distributed across 42 localities of the entire extant range of *P. echinata*, from the state 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. We assigned these to four categories: the laranja group (six specimens), the café group (Espírito Santo, three specimens), the arruda group (individuals with small leaflets from Bahia and Rio de Janeiro, 22 specimens), and the north group, with specimens growing in states of northern Brazil (Alagoas, Pernambuco, Paraíba, and Rio Grande do Norte, eight specimens). Another nine specimens came from cultivated trees for which we were able to trace their provenance and assign them to the groups described above. Finally, eight specimens came from cultivated trees for which we were unable to trace their provenance, and hence did not assign them any category (Appendix S3).

Morphological characters of herbarium specimens were measured using an architect ruler and a dissecting microscope. Duplicate specimens of *Lima et al. 7905* were only available as high-resolution digital images and were measured using the software imageJ2 (Rueden et al., 2017). A total of 17 quantitative and qualitative variables were initially selected for the morphometric analysis (described in Appendix S3). For traits related to leaflet size, 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. Duplicate specimens representing the same field collection were merged for each morphological variable (Appendix S3).

As the total amount of missing data was low (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 for bias in the data set associated with collection date by calculating correlation values between the month of collection and a subset of quantitative leaf variables. Because 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 (Appendix S3, Description of measurements). The distribution of the data was visualized and checked for normality using qqplots and the Shapiro–Wilk test. As a result, values for rachis length were square-root-transformed, and values for petiole length, pinna length, and leaflet length were log-transformed. To avoid infinite values with log transformation, zero values were replaced by half the measurement precision (0.05 mm). Variables were subsequently standardized using the function scale().

To examine the distribution of specimens in the morphometric space, based on their assigned morphotypes, we ran a principal component analysis (PCA) with the five traits, using ggplot v. 3.4.1 (Wickham, 2016) and ggfortify v. 0.4.16 packages in R (Tang et al., 2016). Subsequently, a Ward hierarchical clustering analysis was performed on a Euclidean distance matrix calculated from the morphometric data set, 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() in the R package factoextra version 1.0.7; Kassambara and Mundt, 2020) and the Calinski criterion (function cascadeKM() in the R package vegan version 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 nontransformed values. We assigned individuals in our data set to one of the genetic lineages detected in our population structure analyses. To reduce potential bias or error in our data set, we removed genetically admixed samples and cultivated samples of unknown provenance, leaving 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: arruda-RJ, arruda-BA, café, Laranja, and north. A nonparametric Kruskal–Wallis test was used to determine whether there were significant differences amongst the medians of each group, followed by pairwise comparisons using Dunn's test, to determine which groups differed significantly (α = 0.05).

RESULTS

Population genomic and phylogenetic analyses

In our ML phylogenetic analyses, *Lima et al. 7910* was found to be sister to the arruda-BA clade, with bootstrap support

above 95%, whereas *Santana et al. 5* was found to be sister to this clade + *Lima et al. 7910*, albeit with poor bootstrap support, below 60% (Figure 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. 5* was found to be sister to the laranja clade, with 74% bootstrap support (Appendix S1, Figure S1). For all other samples, both the ML and quartet phylogenies recovered five groups, with over 95% bootstrap support for all clades (Figure 2A; Appendix S1). Relationships among the five clades 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 (Appendix S1, Figure 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 (results not shown).

For the ADMIXTURE analysis, five genetic clusters were recovered: (1) a cluster containing individuals of the arruda-BA morphotype from Bahia; (2) a cluster containing individuals of the arruda-RJ morphotype from Rio de Janeiro; (3) a cluster from Espírito Santo, containing all the café morphotypes; (4) a cluster from southern Bahia containing all but one of the specimens assigned to the laranja morphotype; (5) a cluster composed of individuals from populations in the northern provinces of Brazil (Rio Grande do Norte,

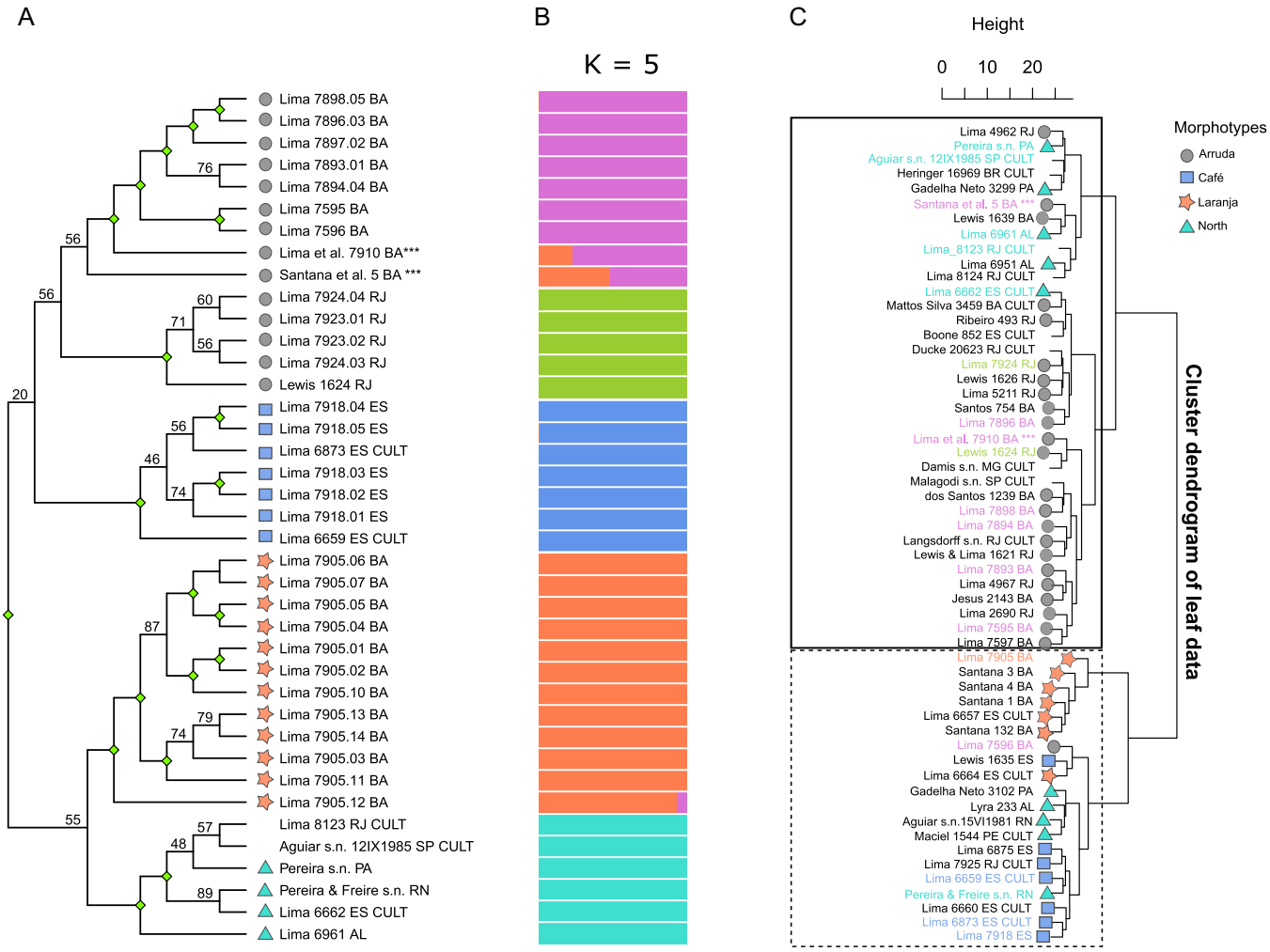


FIGURE 2 Phylogenetic tree, ADMIXTURE plots, and hierarchical clustering dendrogram of the morphometric data set for specimens of *Paubrasilia echinata*. (A) Maximum likelihood phylogenetic tree, with outgroup pruned out. Bootstrap support values are given at each node, except for values ≥95%, which are represented as green diamonds. Morphotypes assigned before the genetic analyses are given at the tips of the tree; see key to colored symbols in panel C. Tip labels indicate collector, collector number, state collected in Brazil, and whether specimens are cultivated or not. *** Asterisks indicate putative hybrids between laranja and arruda-BA clades. (B) ADMIXTURE plots showing posterior probability of assignment to each genetic cluster for $K = 5$. Each bar is plotted opposite the corresponding sample in panel A. (C) Hierarchical clustering of morphological data using Ward's criterion on transformed variables. The height scale indicates the distance between each point in the distance matrix. Solid and dashed black boxes indicate the clusters found by the Calinski and Silhouette criteria for $K = 2$. Samples included in the genetic analyses are indicated by colored tip labels according to the five genetic clusters suggested by the ADMIXTURE analysis. Tip labels indicate collector, collector number, state collected in Brazil, and whether specimens are cultivated or not. Specimens with morphotypes assigned before the clustering analysis are indicated by colored symbols at the branch tips.

Paraíba, Alagoas) and cultivated individuals in the south (Espírito Santo, Rio de Janeiro, and São Paulo) (Figure 2B). In addition, two individuals collected in southern Bahia, *Lima et al. 7910* and *Santana et al. 5*, which were identified as belonging to the arruda morphotype, displayed admixtures with the laranja and arruda-BA genotypes from southern Bahia (Figure 2B). The lowest cross-validation error was found for $K = 5$ (average 0.43), followed by $K = 4$ (0.45) and $K = 6$ (0.46) (Figure 2B; Appendix S1, Figure S2). For $K = 4$, the same clusters were found, except that the clusters from northern Brazil and Rio de Janeiro were merged.

The DAPC recovered the same five genetic clusters as did the ADMIXTURE analysis, with the *k*-means clustering algorithm in the R package adegent retrieving $K = 5$ clusters as the best score (BIC = 283.7) (Figure 3A). 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.

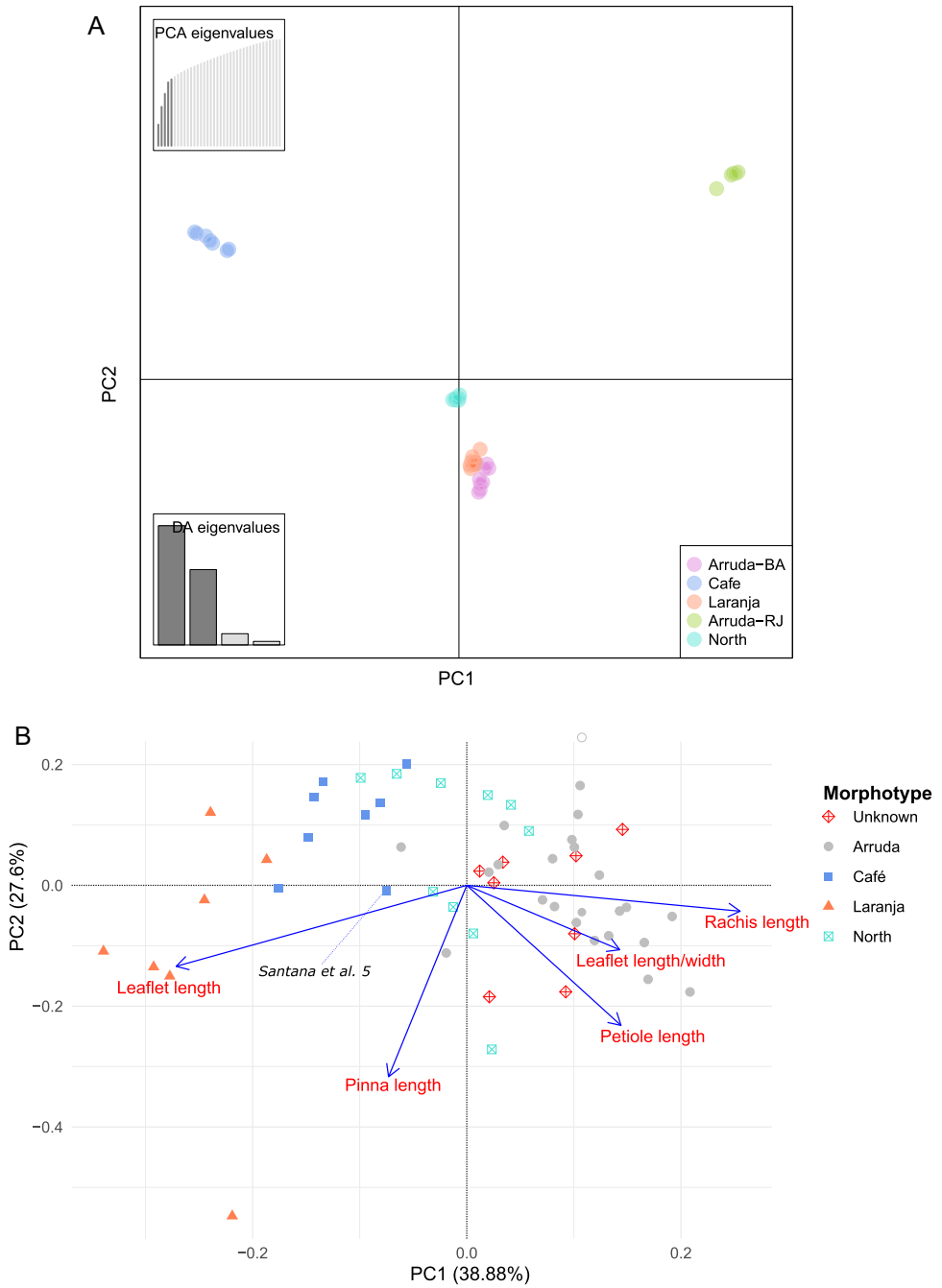


FIGURE 3 Results of the discriminant analysis of principal components (DAPC) and PCA of morphometric traits. (A) DAPC of genotype-by-sequencing data, with outgroup removed (39 individuals). Top left inset shows results of the cumulative PCA eigenvalues; bottom left inset shows eigenvalues for four axes. (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.

Morphometric analyses

For the hierarchical clustering analysis, both the Calinski and Silhouette methods determined the ideal number of clusters in our morphometric data set to be two ($K = 2$). The first of these clusters contained all the specimens from southern Bahia attributed to the laranja morphotype (Figure 2C). This cluster also contained all the café morphotypes, and five of the nine specimens representing collections of non-cultivated individuals from the states of northern Brazil (Alagoas, Pernambuco, Paraíba, and Rio Grande do Norte). Only a single specimen of the arruda morphotype was found to be nested within this cluster (*Lima et al. 7596*; Figure 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 of nine specimens from northeastern Brazil (Alagoas and Paraíba) were found within this cluster. 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.

The first and second axis of the PCA explained 38.9% and 27.6%, respectively, of the variance in the data. Leaflet length and rachis length provided the strongest contribution to explaining the variance in PC1, whereas average pinna length contributed most to explaining the variance in PC2 (Figure 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 with most cultivated specimens for which we could not determine geographical origin. Specimens from the northern Brazilian states overlapped with the café and arruda morphotypes and with cultivated specimens of unknown origin.

Finally, both the boxplots and Kruskall–Wallis analysis (Figure 4, Table 1) revealed significant differences amongst the mean values of these five groups for four of our analyzed traits, but subsequent pairwise analyses did not find unequivocal evidence for clear morphometric distinction among these groups. The results show that the laranja morphotype had significantly larger leaflets than the north, arruda-RJ, and arruda-BA morphotypes, but not the café morphotype (Table 1). We found that specimens assigned to the laranja morphotype had average leaflet lengths between 5–12 cm, with a reduced number of pinnae per leaf (2–3) and leaflets per pinna (4–6), whereas specimens assigned to the café morphotype had leaflets 2.3–4.4 cm long, and a higher average number of pinnae per leaf (2–4) and leaflets per pinna (5–6). (Figure 4, Table 1; Appendix S3). Specimens assigned to the arruda morphotype, from Rio de Janeiro and Bahia had average leaflet lengths between c. 1–2.84 cm and had a much higher number of pinnae per leaf (4–9) and leaflets per pinnae (9–20) (Figure 4, Table 1; Appendix S1, Table S2).

DISCUSSION

Our study is the first to use genomic sequencing to examine the overall population structure of *P. echinata* throughout the Atlantic Coast of Brazil, using representative specimens from all three putative morphotypes, 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 phylogenetic reconstruction, ADMIXTURE analysis and DAPC (Figures 2A, B, 3A; Appendices S1, S2). Of these five groups, only the laranja and café groups supported our initial hypothesis. Furthermore, the arruda group split into two geographical clades, one containing specimens from Rio de Janeiro and the other containing specimens from coastal Atlantic forests of the state of Bahia; two individuals also had indications of admixture (*Lima et al. 7910* and *Santana et al. 5*; Figure 2B). The fifth clade contained all specimens from northeastern Brazil and three cultivated specimens from various geographic locations, including São Paulo, Espírito Santo, and the Rio de Janeiro Botanical Garden.

The morphometric analyses did not support the existence of three distinct morphotypes (Figures 2C, 3B, 4). The laranja, café, and arruda morphotypes grouped together in most of these analyses. However, samples from the northern states of Brazil disrupted this pattern, clustering with the arruda and café individuals in the clustering dendrogram (Figure 2C) and overlapping with these groups in the PCA (Figure 3B). Samples that clustered with the café and arruda morphotypes were found to belong to the same genetic lineage (Figure 2). Furthermore, one individual admixed between the laranja and arruda-BA groups (*Santana et al. 5*) appeared to have a leaf morphology intermediate between these two groups (Figure 3B).

In summary, while our genetic and morphometric analyses recovered clusters that correspond to the café and laranja morphotypes, our analyses also showed that the small-leafleted arruda group separates into two genetic clusters that are sister taxa. Furthermore, a fifth genetic cluster whose leaflet size and morphology overlaps with the arruda and café groups helps to explain why three infraspecific taxa of *P. echinata* have not been formally recognized to date. Equally, our results do not support the description of five distinct species, as it would be difficult to distinguish them based on leaflet morphology alone. Nevertheless, our results do provide important 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 among different genetic clusters shows commonalities with

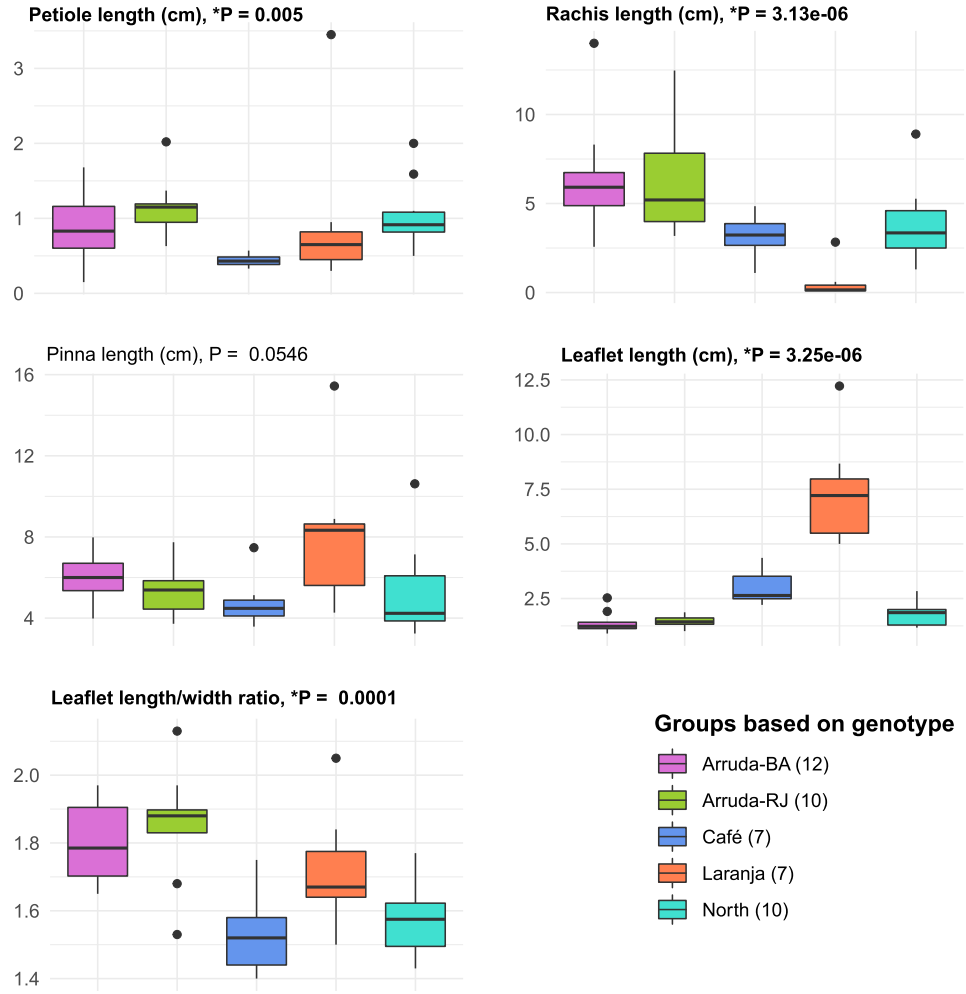


FIGURE 4 Box plots of five leaf traits based on genetic clusters. *P*-values from the Kruskall–Wallis test are given next to each variable (Bold values with an asterisk indicate significance at $P < 0.01$.). Cultivated specimens of unknown provenance and two admixed individuals (*Lima et al. 7910* and *Santana et al. 5*) were not included in the analysis. The sample size for each group is indicated in parentheses in the key. Whiskers extend to ±1.5 × interquartile range (IQR).

major biogeographical patterns and species community turnovers found in other plant and animal taxa of the Brazilian Atlantic Forest, recently reviewed by 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 taxa, including tree species (Prance, 1982), passerine birds (Silva et al., 2004), and amphibians (Trevisan et al., 2020), and 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 hypothesized 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 that have suggested that climatic oscillation in the recent Pleistocene period led to differences in the stability of different areas of the Atlantic Forest, with the Central Bahian and northern regions far more paleoclimatically stable than the forest South of the Rio Doce (Carnaval and Moritz, 2008; Carnaval et al., 2014). 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) or have suggested that earlier events

TABLE 1 Results of the pairwise Dunn comparison test for the four traits which differed significantly in the Kruskall–Wallis test (see Figure 4). Threshold for *P*-values: 0.01. (**$P < 0.01$; NS: not significant, $P > 0.01$). Rows shaded in gray indicate comparisons between the laranja group and other genetic lineages detected in this study.

Group 1	Group2	Petiole length (cm)	Rachis length (cm)	Leaflet length (cm)	Leaflet length/ width
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	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

before the late Pleistocene could be responsible for these biogeographic patterns (Grazziotin et al., 2006; Thomé et al., 2010; Raposo do Amaral et al., 2013).

Genetic and morphological differentiation in *P. echinata* could also result from local adaptation to different soil types and climatic conditions across the Atlantic Forest Biome, including the differing environments in the restinga and tabuleiro forests. 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). In environments with higher levels of humidity, such as the ombrophilous forests where the laranja and café groups are found, larger leaflets can potentially reduce transpiration and minimize latent heat loss (Wright et al., 2017). 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 needed before we can conclude that leaflet size in *P. echinata* is an adaptation to local environments.

Finally, if we hypothesize that natural populations of pau brasil are locally adapted, the expectation is that other morphological traits adapted 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 has wood traits that suggest adaption to higher drought and higher temperatures. These traits included a higher percentage of axial parenchyma and wider vessels in samples from Rio de Janeiro, which would guarantee greater conductance, whereas populations from the states of Rio Grande do Norte 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 have shaped the diversity of *P. echinata* in the Brazilian Atlantic Forest, future studies should focus on expanding population-level sampling to detect any 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, 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 (Lira et al., 2003; Cardoso et al., 2005). Our results, based on greater 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 be given to the southern populations in Bahia and Espírito Santo, particularly with regards to populations belonging to the café and laranja morphotypes, which are hyper-localized in distribution and currently are not included in any protected area, contrary to the other three lineages. It is even more important to focus on these populations given that our analyses showed evidence in two individuals for genetic introgression between the laranja and arruda-BA genotypes. One of these putative hybrids (*Lima et al. 7910*) was collected from the margin of the laranja forest patch, in a region where arruda-BA has been recorded, but it is not clear whether the hybridization occurred between trees from native populations of these two lineages or resulted from the planting of different genotypes next to each other (Figure 2).

In addition, our morphometric analyses showed that most specimens from cultivated trees fall in the cluster composed of specimens from the arruda and north groups. One explanation is that cultivated trees in Brazil may be primarily derived from a limited number of genetic pools. Many trees were planted in urban areas between 1970 and 1995 on the initiative of Prof. Roldão S. Fontes, of the

Federal University of Pernambuco (Fontes, 1995). 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 if further genetic sampling shows that cultivated trees in urban environments, natural plantations, and potentially also in nature reserves and protected areas are possibly propagated from a similar genetic stock. Northern Brazil was once the historical center for the exploitation and export of pau brasil and pernambuco wood is an alternative common name for the species *Paubrasilia echinata*, recognizing the state of Pernambuco in northern Brazil (Macedo et al., 2020).

To verify these hypotheses, we need to include more specimens from cultivated urban trees and cacao plantations in future genetic analyses. The full extent to which hybridization occurs between different cultivated specimens of *P. echinata* is unknown and should be further investigated. While conservation of *P. echinata* as an urban tree or in cacao plantations is one solution to the long-term survival of the species, it would be preferable to plant selected genotypes that are representative of local populations.

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

Challenges using herbarium specimens for morphometric analyses

Since we did not use all the same specimens in the morphometric and genetic analyses, the comparison of the groups between the two data sets 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, we collected 37 of the 56 specimens used in our study, so we are confident of their status as wild or cultivated.

Factors other than genotype can lead to phenotypic differences in leaflet size. In addition to natural variation among 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. 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 color of the heartwood (Cunha and Lima, 1992; Lewis, 1998), but this trait is not recorded on most herbarium specimens.

CONCLUSIONS

By combining morphological and genetic data from across the extant distribution of *Paubrasilia echinata*, we found 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 appear to correspond to the previously defined laranja, café, and arruda leaflet morphotypes, individuals from northern Brazil were found to have leaf traits that overlap with both the café and arruda morphotypes. Further investigation into the morphological differences among subpopulations 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 rarer 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 morphotypes should be formally recognized as distinct infraspecific taxa is an ongoing question. More work is needed on the laranja morphotype to understand how widespread this 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.

AUTHOR CONTRIBUTIONS

M.R.: data curation; formal analysis; investigation; project administration; validation; visualization; writing original draft, reviewing, and editing. L.N.: supervision; reviewing and editing. G.L.: supervision; reviewing and editing. H.d.L.: conceptualization; reviewing and editing. E.G.: conceptualization; funding acquisition; project administration; supervision; reviewing and editing.

ACKNOWLEDGMENTS

We thank Dr. Royce Steeves for advice on DNA sequencing, Dr. Anik Dutta for advice on bioinformatic analyses, and undergraduate research assistant Carlo Cormier for extracting DNA. We also thank collaborators and partners at the Centro de Pesquisas do Cacau (CEPEC) in Ilhéus, Bahia, Brazil, including Robelio Duarte de Santana. This work was made possible thanks to a grant from the Department of Biology of the University of Moncton (New Brunswick, Canada) and a postdoctoral fellowship to E.G. from the

Fonds de recherche du Québec en Nature et technologies (FQRNT). Fieldwork was made possible by a travel grant from the Biodiversity Research Centre of Quebec. The Research/Scientific Computing teams at The James Hutton Institute and NIAB provided computational resources and technical support for the "UK's *Crop Diversity* Bioinformatics HPC" (BBSRC grant BB/S019669/1), which contributed to the results reported here. We also thank the three reviewers, including Andrew Schnabel, who provided constructive reviews and feedback that greatly improved the manuscript.

COMPETING INTERESTS

The authors have no relevant financial or nonfinancial interests to disclose.

DATA AVAILABILITY STATEMENT

The scripts used for the analyses are available in the Github repository: https://github.com/MyosotisMatt/Paubrasil. All the generated genomic data are available at NCBI's Sequence Read Archive (SRA) under reference PRJNA689870 at https://www.ncbi.nlm.nih.gov/sra/PRJNA689870.

ORCID

Mathew Rees http://orcid.org/0000-0002-6410-1260
Linda E. Neaves http://orcid.org/0000-0002-5626-1029
Gwilym P. Lewis http://orcid.org/0000-0003-2599-4577
Edeline Gagnon http://orcid.org/0000-0003-3212-9688

REFERENCES

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19: 1655–1664.

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data. Website: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arseneau, J. R., R. Steeves, and M. Laflamme. 2017. Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues. *Molecular Ecology Resources* 17: 686–693.

Ball, M. C., I. R. Cowan, and G. D. Farquhar. 1988. Maintenance of leaf temperature and the optimisation of carbon gain in relation to water loss in a tropical mangrove forest. *Functional Plant Biology* 15: 263–276.

Borges, L. A., M. S. Sobrinho, and A. V. Lopes. 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 - Morphology, Distribution, Functional Ecology of Plants* 204: 111–130.

Brazil Flora Group. 2022. Brazilian Flora 2020: Leveraging the power of a collaborative scientific network. *Taxon* 71: 178–198.

Brown, C. M., J. H. Arbour, and D. A. Jackson. 2012. Testing of the effect of missing data estimation and distribution in morphometric multivariate data analyses. *Systematic Biology* 61: 941–954.

Capblancq, T., and B. R. Forester. 2021. Redundancy analysis: a Swiss Army Knife for landscape genomics. *Methods in Ecology and Evolution* 12: 2298–2309.

Cardoso, M. A., J. Provan, W. Powell, P. C. G. Ferreira, and D. E. de Oliveira. 1998. High genetic differentiation among remnant populations of the endangered *Caesalpinia echinata* Lam. (Leguminosae-Caesalpinioideae). *Molecular Ecology* 7: 601–608.

Cardoso, S. R. S., J. Provan, C. D. F. Lira, L. D. O. R. Pereira, P. C. G. F. Ferreira, and M. A. Cardoso. 2005. High levels of genetic structuring as a result of population fragmentation in the tropical tree species *Caesalpinia echinata* Lam. *Biodiversity & Conservation* 14: 1047–1057.

Carnaval, A. C., and C. Moritz. 2008. Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic Forest. *Journal of Biogeography* 35: 1187–1201.

Carnaval, A. C., E. Waltari, M. T. Rodrigues, D. Rosauer, J. VanDerWal, R. Damasceno, I. Prates, et al. 2014. Prediction of phylogeographic endemism in an environmentally complex biome. *Proceedings of the Royal Society, B, Biological Sciences* 281: 20141461.

CITES [Convention on International Trade of Endangered Species of Wild Fauna and Flora]. 2007. CoP14 Prop. 30, Consideration of proposals for amendment of Appendices I and II. Fourteenth meeting of the Conference of the Parties, The Hague, Netherlands. CITES, Geneva, Switzerland. Website: https://cites.org/sites/default/files/eng/cop/14/prop/E14-P30.pdf [accessed 2 October 2023].

Cunha, M. W., and C. L. Lima. 1992. Viagem à terra do pau-brasil. Agência Brasileira de Cultura, Rio de Janeiro, Brasil.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10: giab008.

Eaton, D. A. R., and I. Overcast. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* 36: 2592–2594.

Eaton, D. A. R., E. L. Spriggs, B. Park, and M. J. Donoghue. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology* 66: 399–412.

Fonseca, C. R., J. M. Overton, B. Collins, and M. Westoby. 2000. Shifts in trait-combinations along rainfall and phosphorus gradients. *Journal of Ecology* 88: 964–977.

Fontes, R. S. 1995. Pau-Brasil, um sonho de resgate. FUNBRASIL, Recife, Brasil.

Francis, R. M. 2017. pophelper: an R package and web app to analyse and visualize population structure. *Molecular Ecology Resources* 17: 27–32.

Gagnon, E., A. Bruneau, C. E. Hughes, L. P. de Queiroz, and G. P. Lewis. 2016. A new generic system for the pantropical *Caesalpinia* group (Leguminosae). *PhytoKeys* 71: 1–160.

Gagnon, E., G. P. Lewis, and H. C. de Lima. 2020. *Paubrasilia*. *In* Flora do Brasil. 2020. Website: http://reflora.jbrj.gov.br/reflora/floradobrasil/FB102193 [accessed 6 April 2021].

Grazziotin, F. G., M. Monzel, S. Echeverrigaray, and S. L. Bonatto. 2006. Phylogeography of the *Bothrops jararaca* complex (Serpentes: Viperidae): past fragmentation and island colonization in the Brazilian Atlantic Forest. *Molecular Ecology* 15: 3969–3982.

Global Runoff Data Centre [GRDC]. 2020. Major river basins of the world. Global Runoff Data Centre, 2nd, revised, extended edition. Federal Institute of Hydrology [BfG, Bundesanstalt für Gewässerkunde], Koblenz, Germany.

Habel, J. C., L. Rasche, U. A. Schneider, J. O. Engler, E. Schmid, D. Rodder, S. T. Meyer, et al. 2019. Final countdown for biodiversity hotspots. *Conservation Letters* 12: e12668.

Jombart, T., and A. Bateman. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403–1405.

Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11: 94.

Josse, J., and F. Husson. 2016. missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software* 70: 1–31.

Juchum, F. S., M. A. Costa, A. M. Amorim, and R. X. Corrêa. 2008. Phylogenetic relationships among morphotypes of *Caesalpinia echinata* lam. (Caesalpinioideae: Leguminosae) evidenced by *trnL* intron sequences. *Naturwissenschaften* 95: 1085–1091.

Kassambara, A., and F. Mundt. 2020. factoextra: extract and visualize the results of multivariate data analyses. R package version 1.0.7. Website: https://CRAN.R-project.org/package=factoextra

Knaus, B. J., and N. J. Grünwald. 2017. vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources* 17: 44–53.

Leigh, A., S. Sevanto, J. D. Close, and A. B. Nicotra. 2017 The influence of leaf size and shape on leaf thermal dynamics: does theory hold up under natural conditions? *Plant Cell & Environment* 40: 237–248.

Leite, Y. L. R., L. P. Costa, A. C. Loss, and R. Pardini. 2016. Neotropical forest expansion during the last glacial period challenges refuge hypothesis. *Proceedings of the National Academy of Sciences, USA* 113: 1008–1013.

Lewis, G. P. 1998. Caesalpinia: a revision of the Poincianella-Erythrostemon group. Royal Botanic Gardens, Kew, UK.

Lichtenberg, S., E. Huber-Sannwald, U. Nehren, and J. A. Reyes-Agüero. 2019. Use and conservation of the threatened Brazilian national tree *Paubrasilia echinata* Lam.: a potential for Rio de Janeiro State? *In* U. Nehren, S. Schlüter, C. Raedig, D. Sattler, and H. Hissa [eds.], Strategies and tools for a sustainable rural Rio de Janeiro, 205–219. Springer International, Cham, Switzerland.

Lima, H. C., G. P. Lewis, and E. Bueno. 2002. Pau-brasil: uma biografia. *In* E. Bueno [ed.], Pau-brasil, 39–76. Axis Mundi, Sao Paulo, Brazil.

Lira, C. F., S. R. S. Cardoso, P. C. G. Ferreira, M. A. Cardoso, and J. Provan. 2003. Long-term population isolation in the endangered tropical tree species *Caesalpinia echinata* Lam. revealed by chloroplast microsatellites. *Molecular Ecology* 12: 3219–3225.

Macedo, T. M., C. G. Costa, H. C. de Lima, and C. F. Barros. 2020. Wood anatomy of historic French violin bows made of Pernambuco wood. *IAWA Journal* 41: 320–332.

Macedo, T. M., H. C. de Lima, N. D. de Souza, C. G. Costa, and C. F. Barros. 2019. Infraspecific variation of *Paubrasilia echinata* (Fabaceae) wood along a latitudinal gradient in Brazil. *Flora* 258: 151437.

McDonald, P. G., C. R. Fonseca, J. M. Overton, and M. Westoby. 2003. Leaf-size divergence along rainfall and soil-nutrient gradients: Is the method of size reduction common among clades? *Functional Ecology* 17: 50–57.

Menini Neto, L., S. G. Furtado, D. C. Zappi, A. T. Oliveira-Filho, and R. C. Forzza. 2016. Biogeography of epiphytic angiosperms in the Brazilian Atlantic Forest, a world biodiversity hotspot. *Brazilian Journal of Botany* 39: 261–273.

Morellato, L. P. C., and C. F. B. Haddad. 2000. Introduction: The Brazilian Atlantic Forest. *Biotropica* 32: 786–792.

Neves, D. M., K. G. Dexter, R. T. Pennington, A. S. M. Valente, M. L. Bueno, P. V. Eisenlohr, M. A. L. Fontes, et al. 2017. Dissecting a biodiversity hotspot: the importance of environmentally marginal habitats in the Atlantic Forest Domain of South America. *Diversity & Distribution* 23: 898–909.

Nicotra, A. B., A. Leigh, C. K. Boyce, C. S. Jones, K. J. Niklas, D. L. Royer, and H. Tsukaya. 2011. The evolution and functional significance of leaf shape in the angiosperms. *Functional Plant Biology* 38: 535–552.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, et al. 2013. vegan: community ecology package. Website: https://CRAN.R-project.org/package=vegan

Peres, E. A., R. Pinto-da-Rocha, L. G. Lohmann, F. A. Michelangeli, C. Y. Miyaki, and A. C. Carnaval. 2020. Patterns of species and lineage diversity in the Atlantic Rainforest of Brazil. *In* V. Rull and A. C. Carnaval [eds.], Neotropical diversification: patterns and processes, 415–447. Springer, Cham, Switzerland.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J. L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7: p. e32253.

Prance, G. T. 1982. Forest refuges: evidence from woody angiosperms. *In* G. T. Prance [ed.], Biological diversification in the tropics, 137–158. Columbia University Press, NY, NY, USA.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website: https://www.R-project.org/

Raposo do Amaral, F., P. K. Albers, S. V. Edwards, and C. Y. Miyaki. 2013. Multilocus tests of Pleistocene refugia and ancient divergence in a pair of Atlantic Forest antbirds (*Myrmeciza*). *Molecular Ecology* 22: 3996–4013.

Ribeiro, M. C., J. P. Metzger, A. C. Martensen, F. J. Ponzoni, and M. M. Hirota. 2009. The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. *Biological Conservation* 142: 1141–1153.

Rodrigues, P. S., M. M. Souza, C. A. F. Melo, N. S. Pereira, and R. X. Correa. 2018. Karyotype diversity and 2C DNA content in species of the *Caesalpinia* group. *BMC Genetics* 19: 25.

Rueden, C. T., J. Schindelin, M. C. Hiner, B. E. DeZonia, A. E. Walter, E. T. Arena, and K. W. Eliceiri. 2017. ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* 18: 529.

Scarano, F. R. 2002. Structure, function and floristic relationships of plant communities in stressful habitats marginal to the Brazilian Atlantic rainforest. *Annals of Botany* 90: 517–524.

Silva, J. M. C., M. C. Sousa, and C. H. M. Castelletti. 2004. Areas of endemism for passerine birds in the Atlantic forest, South America. *Global Ecology and Biogeography* 13: 85–92.

Snir, S., and S. Rao. 2012. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. *Molecular Phylogenetics and Evolution* 62: 1–8.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Tang, Y., M. Horikoshi, and W. Li. 2016. ggfortify: Unified interface to visualize statistical results of popular r packages. *R Journal* 8: 478–489.

Thomé, M. T. C., K. R. Zamudio, J. G. R. Giovanelli, C. F. B. Haddad, F. A. Baldissera, and J. Alexandrino. 2010. Phylogeography of endemic toads and post-Pliocene persistence of the Brazilian Atlantic Forest. *Molecular Phylogenetics and Evolution* 55: 1018–1031.

Trevisan, C. C., H. Batalha-Filho, A. A. Garda, L. Menezes, U. R. Dias, M. Solé, C. Canedo, et al. 2020. Cryptic diversity and ancient diversification in the northern Atlantic Forest *Pristimantis* (Amphibia, Anura, Craugastoridae). *Molecular Phylogenetics and Evolution* 148: 106811.

Van-Lume, B., T. Esposito, J. A. F. Diniz-Filho, E. Gagnon, G. P. Lewis, and G. Souza. 2017. Heterochromatic and cytomolecular diversification in the *Caesalpinia* group (Leguminosae): relationships between phylogenetic and cytogeographical data. *Perspective in Plant Ecology, Evolution and Systematics*. 29: 51–63.

Varty, N. 1998. *Caesalpinia echinata*. The IUCN red list of threatened species 1998: e.T33974A9818224. Website: https://www.iucnredlist.org/species/33974/9818224 [accessed 01 May 2020].

Weeks, J. P. 2010. Plink: an R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software* 35: 1–33.

Wickham, H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag, NY, NY, USA.

Wright, I. J., N. Dong, V. Maire, I. C. Prentice, M. Westoby, S. Díaz, R. V. Gallagher, et al. 2017. Global climatic drivers of leaf size. *Science* 357: 917–921.

Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8: 28–36.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

APPENDIX S1.

Table S1. Samples used in the genotype-by-sequencing analysis.

Table S2. Morphometric measures for each of the five genetic clusters, showing the mean, median, and standard deviation of the five leaf traits.

Figure S1. Quartet phylogenetic tree, constructed with Tetrad, with outgroup.

Figure S2. ADMIXTURE results for K = 1–10. (A) Bar plots of the cross-validation results for K = 1–10, repeated 10 times. (B) ADMIXTURE plots showing posterior probability of assignment to each cluster for K = 2–10.

APPENDIX S2. Statistics from Ipyrad. (1) Demultiplexing_results: log output after demultiplexing files using Ipyrad. (2) Paubrasil+Outgroup: log output after running steps 2–7 of Ipyrad, with both the samples from *Paubrasilia* and the outgroup. (3) Paubrasil_only: log output after running steps 3–7 of Ipyrad, with only the samples from *Paubrasilia echinata*.

APPENDIX S3. Morphometric data set. (1) Paubrasilia_morphometrics_Raw_D: raw measurements for 17 morphometric variables for all herbarium samples included in this study, including duplicates. (2) *Paubrasilia*_merged_measures: merged measurements of 17 morphometric variables for herbarium samples included in this study. (3) Description of measurements: Description of morphometric measurements of all 17 variables, how they were taken on the sample or measured.

How to cite this article: Rees, M., L. E. Neaves, G. P. Lewis, H. C. de Lima, and E. Gagnon. 2023. Phylogenomic and morphological data reveal hidden patterns of diversity in the national tree of Brazil, *Paubrasilia echinata*. *American Journal of Botany* 110(11): e16241. https://doi.org/10.1002/ajb2.16241