Genet Resour Crop Evol (2025) 72:965–980
https://doi.org/10.1007/s10722-024-02024-0

RESEARCH ARTICLE

SNP-based analysis reveals high genetic structure and diversity in umbu tree (*Spondias tuberosa* Arruda), a native and endemic species of the Caatinga biome

Wellington Ferreira do Nascimento · Flaviane Malaquias Costa · Alessandro Alves-Pereira · Carlos Eduardo de Araújo Batista · Igor Araújo Santos de Carvalho · Caroline Bertocco Garcia · Allison Vieira da Silva · Edson Ferreira da Silva · Márcia Maria de Souza Gondim Dias · Fábio Rodrigo Araújo Pereira · Maria Imaculada Zucchi · Elizabeth Ann Veasey

Received: 11 April 2024 / Accepted: 14 May 2024 / Published online: 24 May 2024

Abstract Umbu (*Spondias tuberosa* Arruda) is an endemic fruit tree restricted to the Brazilian seasonally dry tropical forest called Caatinga. This study aimed to evaluate the structure and genomic diversity of umbu trees from seven locations in the Caatinga biome, distributed among four Brazilian states. Using genotyping-by-sequencing (GBS), a total of 5,336 SNPs were obtained, of which 250 showed outlier behavior. Therefore, 5,086 neutral SNPs were used for population structure and genetic diversity analyses. Both discriminant analysis of principal components (DAPC) and neighbor-joining cluster analyses classified the accessions into four groups, with a genetic structure observed among groups, disagreeing with our initial hypothesis of low genetic structure between locations. Isolation by distance ($r^2 = 0.974$;

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10722-024-02024-0.

W. F. do Nascimento (✉)
Science Center of Chapadinha, Federal University of Maranhão, Chapadinha, MA, CEP 65500-000, Brazil
e-mail: wellington.fn@ufma.br

F. M. Costa · C. E. de Araújo Batista · I. A. de Carvalho · C. B. Garcia · A. V. da Silva · E. A. Veasey
Genetics Department, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, SP, CEP 13418-900, Brazil
e-mail: flavianemcosta@hotmail.com

C. E. de Araújo Batista
e-mail: batistace@gmail.com

I. A. de Carvalho
e-mail: igor.arsc@usp.br

C. B. Garcia
e-mail: caroline.garcia@usp.br

A. V. da Silva
e-mail: allisonvsagro@usp.br

E. A. Veasey
e-mail: eann.veasey@gmail.com

A. Alves-Pereira
Genetics Department, Federal University of Amazonas, Manaus, AM, CEP 69067-005, Brazil
e-mail: ale.alves.pereira@gmail.com

E. F. da Silva
Department of Biology, Federal Rural University of Pernambuco, Recife, PE, CEP 52171-900, Brazil
e-mail: edson.fsilva4@ufrpe.com

M. M. de Souza Gondim Dias
Center of Sustainable Development of Semiarid, Federal University of Campina Grande, Sumé, PB, CEP 58540-000, Brazil
e-mail: msouzagondim@yahoo.com.br

F. R. A. Pereira
Federal Institute of Pernambuco, Campus Garanhuns, Garanhuns, PE, CEP 55299-390, Brazil
e-mail: fabiorodrigopereira@hotmail.com

M. I. Zucchi
Paulista Agency of Agrobusiness Technology, Piracicaba, SP, CEP 13400-970, Brazil
e-mail: mizucchi@gmail.com

p = 0.0015) was detected. Moderate to high levels of genetic diversity were found, with the average observed heterozygosity ($H_O = 0.221$) higher than the expected heterozygosity ($H_E = 0.199$) and with negative inbreeding coefficient (F_{IS}) values. Most genetic variation was found within locations, although high diversity between locations (22.1%) was observed. The results obtained are important for understanding the levels and distribution of genetic variation, suggesting that most locations are priorities for conservation actions, contributing with different alleles to the species' gene pool in Brazil.

Keywords Genetic diversity · Population structure · Caatinga biome · Native fruit · Endemic species

Introduction

Human activities have impacted all ecosystems on our planet, reducing their biodiversity and, consequently, their ability to maintain ecological functions and provide benefits to society (Haddad et al. 2015; Newbold et al. 2015; Miraldo et al. 2016). The Brazilian seasonally dry tropical forest called Caatinga is

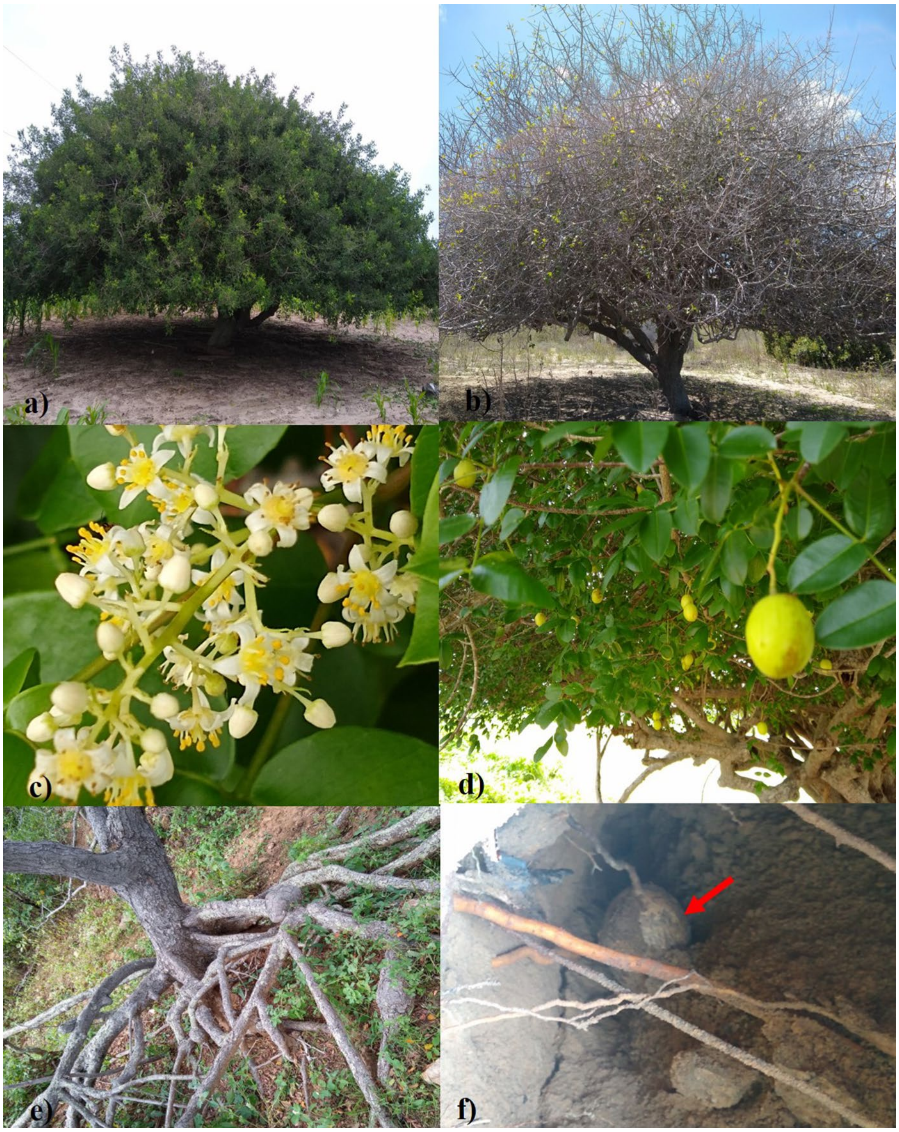

Fig. 1 Adult umbuzeiro trees during rainy (a) and dry (b) periods. Flowers (c), fruits (d) and roots (e, f), showing the xylopod (red arrow) (f). Source: Fábio Rodrigo Araújo Pereira

one of the most threatened biomes in the country due to the poorly planned use of its resources, especially concerning the removal of native vegetation (Santana and Souto 2006). Caatinga vegetation includes several endemic species, such as *Spondias tuberosa* Arruda (Anacardiaceae), popularly known as umbu tree (Fig. 1a, b) (Lins Neto et al. 2010, 2013; Mitchell and Daly 2015). Its popular name is derived from the Tupi-Guarani indigenous word "*ymb-u*", which means "the tree that gives water" (Epstein 1998). This results from a physiological adaptation of the plant forming roots with xylopods (Fig. 1e, f) capable of accumulating water, minerals, and organic solutes (Epstein 1998; Cavalcanti et al. 2010), which allows their survival in the dry season (Silva et al. 2008; Cavalcanti et al. 2010).

Umbu is an incipiently domesticated deciduous fruit tree (Lins Neto et al. 2013, 2014). Although it is native to the Caatinga, it occurs frequently in areas near the Atlantic Forest (Balbino et al. 2018), from the north of Minas Gerais State in the Southeast region to the most northern point of the Northeastern region in Brazil (Santos 1997). In this semi-arid region, this fruit tree represents an important food and medicinal resource for local residents, in addition to having high market potential (Albuquerque et al. 2007; Siqueira et al. 2016; Mertens et al. 2017; Cordeiro et al. 2018). The exploitation of its fruits (Fig. 1d) is mainly based on extractivism (Mertens et al. 2017), directly proportional to the flavor (if bittersweet), size, and quantity of pulp (Lins Neto et al. 2010), as they are commercially exploited for "*in natura*" consumption and preparation of juices, jellies, ice creams, sweets, and frozen pulp (Mertens et al. 2017). The fruits and leaves are also used as fodder for small domestic mammals such as sheep and goats (Cavalcanti et al. 2004). In traditional medicine, different parts of the plant have been used to treat venereal diseases, digestive disorders, diarrhea, diabetes, menstrual disturbances, and placental delivery (Albuquerque et al. 2007; Siqueira et al. 2016; Cordeiro et al. 2018). Siqueira et al. (2016) reported evidence of an anti-inflammatory action using the leaves, suggesting potential therapeutic benefits for inflammatory conditions. The pharmacological potential of the leaf extract as an antioxidant and antifungal agent was also demonstrated by Cordeiro et al. (2018).

S. tuberosa is an andromonoecious species with gametophytic self-incompatibility (Leite and Machado 2010), pollinated mainly by bees and wasps (Nadia et al. 2007; Almeida et al. 2011). It is a predominantly allogamous species, with an estimated outcrossing rate of 71,9% (Santos et al. 2011), and variation between 80.4% (multi-locus) and 84.1% (single locus) (Santos and Gama 2013). However, fruit production is low, considering the high number of flowers produced (Fig. 1c). According to Stephenson (1981), this may be related to extrinsic factors, such as limiting environmental resources, and it may also be related to intrinsic factors, such as zoochoric fruits of high energy value aborted at a young age (Nadia et al. 2007).

In the Caatinga, the exploration of extensive pastures is the predominant anthropic disturbance (Alves et al. 2009) and combined with the destruction of their habitat (Mertens et al. 2017; Balbino et al. 2018), induces a reduction of the umbu tree populations, which may compromise the genetic diversity of the species (Mitchell and Daly 2015). It is supposed that much of the existing genetic variability of *S. tuberosa* has been lost due to indiscriminate deforestation, rapid advance of agricultural frontiers for the plantation of exotic crops, and urban expansion in their respective areas of occurrence, which may have been enhanced by natural threats, such as climate change (Mertens et al. 2017). Therefore, understanding the species' genetic diversity is essential to rationalize the use of its genetic resources, elaborate efficient conservation strategies, and develop genetic improvement programs (Souza et al. 2016).

Several molecular markers have been used to assess the genetic diversity of this species, such as RAPD (Random Amplified Polymorphic DNA) (Moreira et al. 2007), AFLP (Amplified Fragment Length Polymorphism) (Santos et al. 2008, 2011; Santos and Oliveira 2008), ISSR (Inter-simple Sequence Repeat) (Lins Neto et al. 2013), and SSR (Simple Sequence Repeat) (Balbino et al. 2018, 2019; Santos et al. 2021a,b). Only one study was found using SNP (Single Nucleotide Polymorphisms) markers in *S. tuberosa* (Nobre et al. 2018) addressing the hybrid origin of *S. bahiensis* P. Carvalho, van Den Berg & M. Machado.

The genotyping-by-sequencing (GBS) technique is based on the complexity reduction of genomic DNA by restriction enzymes and on the use of barcode DNA adapters to produce multiplexed libraries of samples that are submitted to next-generation sequencing (NGS) (Poland and Rife 2012). With this combination, the technique has demonstrated the ability to produce thousands of SNPs in several species, including fruit trees (Goonetilleke et al. 2018). SNP markers are the most abundant genetic polymorphisms in the genome. In addition to the evaluation of neutral variation, they enable the study and identification of regions of the genome that might be under natural selection in the population (outlier loci), that is, regions possibly associated with adaptation (Luikart et al. 2003; Cortinovis et al. 2020; Alves-Pereira et al. 2020, 2022). The present study aimed to assess the genomic diversity and structure of umbu trees (*S. tuberosa*) from seven locations in the Caatinga biome with SNP markers obtained through the GBS technique. Two complimentary hypotheses were tested: a) because *S. tuberosa* is an endemic species of the Caatinga, predominantly allogamous, presenting gametophytic self-incompatibility, we expected to find low genetic structure between locations; b) the species is in a state of genetic vulnerability, highly threatened by anthropogenic activities, showing a reduction in its populations and, consequently, in its genetic diversity.

Table 1 Sampling locations of umbu trees (*Spondias tuberosa*), from seven locations distributed in the states of Minas Gerais (MG), Bahia (BA), Paraíba (PB), and Pernambuco (PE), in the Caatinga biome, including number of individuals (N) sampled, geographic coordinates and climate data (https://koppenbrasil.github.io/)

Locations	N	Latitude	Longitude	Mean Annual Temp. (ºC)	Annual Rain (mm)	Altitude (m)	Koppen[a]
Espinosa-MG	3	15°05′27.5″S	42°47′49.0″W	22.2	798.3	675.5	As
Senhor do Bonfim-BA	19	10°18′23.0″S	40°09′44.0″W	23.8	679.7	498.1	As
São Vicente do Seridó-PB	3	06°53′40.2″S	36°24′12.3″W	22.9	455.7	573.8	BSh
Queimadas-PB	7	07°26′08.4″S	35°53′12.4″W	23.7	639.0	407.1	As
Boqueirão-PB	6	07°27′19.2″S	36°06′1.5″W	23.7	466.9	424.5	BSh
Cabaceiras-PB	13	07°13′22.3″S	35°53′41.3″W	23.6	429.3	444.1	BSh
Lagoa Grande-PE	20	08°57′27.0″S	40°11′40.0″W	24.5	525.3	429.9	BSh

[a] Koppen classification: As (Tropical savannah: warm, with winter and autumn rains); BSh (Semi-arid: hot and dry, with winter rains)

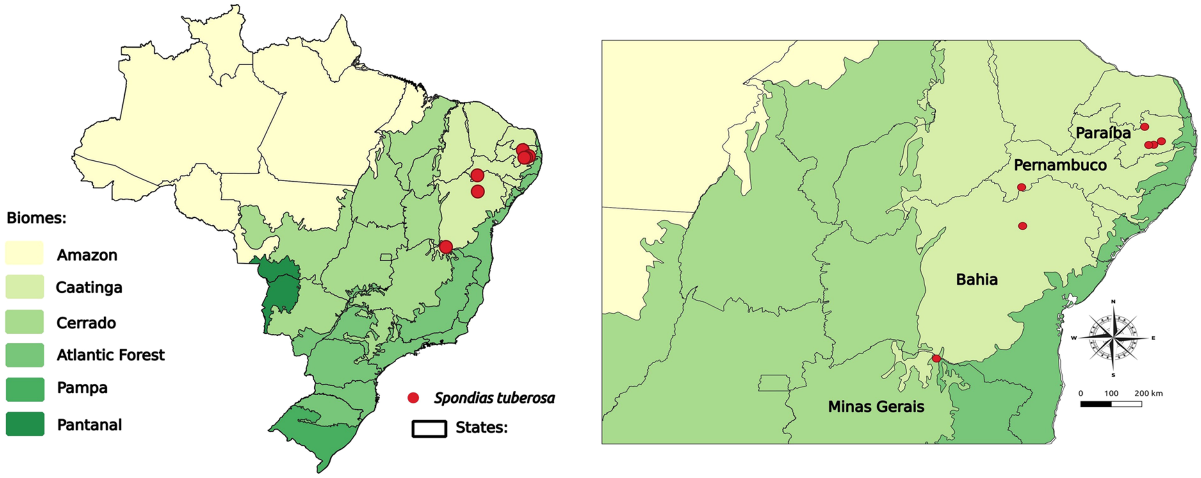

Fig. 2 Map of Brazil with the sampling location of umbu trees (*Spondias tuberosa*), in the states of Minas Gerais (MG), Bahia (BA), Pernambuco (PE), and Paraíba (PB)

Material and methods

Sampling, DNA extraction, and quantification

In the Caatinga biome, the umbu tree occurs in areas of native vegetation but mainly in anthropized ones such as those currently cultivated, pasturelands, homegardens and areas of regeneration of native vegetation after being abandoned for agricultural use. In these areas, young leaves, stored in a plastic bag with silica, were sampled from 71 umbu trees in seven locations (Table 1; Fig. 2), in Minas Gerais State (MG), between the municipalities of Espinosa and Monte Azul, named Espinosa in this study, in cultivated areas and areas of regeneration of native vegetation; in Bahia State (BA), between the municipalities of Jaguarari and Senhor do Bonfim, named as Senhor do Bonfim in this study, in anthropized remnant fragments; in Pernambuco State (PE), between the municipalities of Lagoa Grande and Santa Maria da Boa Vista, named as Lagoa Grande in this study, in cultivated areas and in areas of regeneration of native vegetation; and four areas in Paraiba State (PB), in the municipalities of São Vicente do Seridó and Queimadas, in homegardens and agriculture cultivated areas; Boqueirão, only in agriculture cultivated areas; and Cabaceiras, in areas of natural vegetation and homegardens, with cattle and goat farming. It was established that the minimum distance among sampled individuals was 500 m to minimize the probability of collecting related individuals and a maximum of 3,000 m to represent the sampled area.

The extraction of genomic DNA was performed using the protocol described by Inglis et al. (2018) with modifications, including three to four prewashes with sorbitol buffer [100 mM Tris–HCl pH 8.0, 0.35 M Sorbitol, 5 mM EDTA pH 8.0, 1% (w/v) Polyvinylpyrrolidone (average molecular weight 40,000; PVP-40)]. The quantification and analysis of DNA quality were performed through electrophoresis in a 1% agarose gel (w/v) stained with Gel Red (Biotium). DNA was quantified based on the phage λ molecular size standards (Invitrogen) at different concentrations (20, 50, and 100 ng μL^{-1}) and validated with a Qubit4 fluorometer (Invitrogen). After quantification, DNA samples were normalized to a 20 ng/uL concentration for GBS library preparation.

Genomic library and SNPs identification

The genomic library was prepared following the protocol described by Poland et al. (2012). Briefly describing, high-quality genomic DNA (140 ng per sample) was digested at 37 °C for 12 h using a combination of a *PstI* rare-cutting enzyme (NEB-New England Biolabs) with a *MseI* frequent-cutting enzyme (NEB-New England Biolabs). The fragments generated from the digestion for each sample were ligated to adapters containing specific barcode sequences using the enzyme DNA T4 ligase and grouped together in a 96-plex. The multiplex was column purified and amplified for PCR enrichment, and submitted to a new purification step. The library was then qualitatively evaluated using the BioAnalyzer system (Agilent Technologies) and quantified using the NEBNext® Library Quant Kit for Illumina (New England Biolabs) on the CFX 384 Touch Real Time PCR Detection System (Bio-Rad Laboratories). Subsequently, the library was sequenced in a flowcell using a sequencer on the HiSeq2500 Illumina platform, with the company EcoMol, at the Genomics Center of ESALQ/USP.

The overall sequencing quality was assessed with FastQC (Andrews 2010), and the removal of low-quality sequences containing adapters and trimming of sequences to 80 bases was performed with Trimmomatic 0.39 (Bolger et al. 2014). De novo identification of SNP markers was performed with Stacks-1.42 (Catchen et al. 2011). Sequence demultiplexing for each sample and checking the integrity of restriction sites was performed with the process_radtags module. The initial assembly of loci for each sample was performed with the ustacks module with the parameters of minimum sequencing depth (-m) of 3×and maximum distance allowed between sequences of the same locus (-M) of 2 bases. A catalogue of loci was obtained with the cstacks program allowing a maximum distance between loci of different samples (-n) of 2 bases. The sstacks module was used to compare the loci of each sample with the catalogue loci, and the rxstacks module was used to remove the loci with a lower probability (–lnl_lim -10). The populations module was used for final data filtering, considering SNP markers as the loci with a minimum sequencing depth of 5x, frequency of the rarest allele (MAF) ≥ 0.01, presence of the SNP in at least 90% of the samples of each one of the sampled locations, and retaining only one SNP per sequenced tag. Sequencing quality metrics of SNP markers were obtained with VCFtools 0.1.17 (Danecek et al. 2011).

Statistical analyses

Identification of possible outlier loci

The identification of possible outlier loci was performed considering the sampled locations. Three complementary tests were performed: Pcadapt (Luu et al. 2017) in which the outlier loci are associated with the genetic groups observed in a principal component analysis (PCA); FstHet (Flanagan and Jones 2017) for identifying loci with excessive high or low F_{ST} values in relation to a neutral distribution, and BayeScan (Foll and Gaggiotti 2008), a Bayesian analysis for estimating posterior probabilities to verify whether or not each locus reflects selection. The pcadapt analysis was performed considering the first four principal components, which suggested great agreement between genetic groups and sample locations. In this analysis, SNP markers with q-values <0.1 were considered as outliers. The fstHet analysis was performed based on the beta hat estimate (Cockerham and Weir 1993) (analogous to Wrigth's F_{ST}), considering as outliers the SNP markers above or below a 95% confidence interval constructed based on 1000 bootstraps. The above analyses were performed with the R packages (R Core Team 2018), pcadapt (Luu et al. 2017) and fsthet (Flanagan and Jones 2017). BayeScan 2.1 (Foll and Gaggiotti 2008) was used to perform 20 pilot runs, with 100,000 iterations each, followed by 250,000 burn-in steps and 25,000 steps with intervals of 50 (total of 1,500,000 iterations). It was considered in the model that the probability of including selection was 3 × lower than that of not including selection. In this analysis, SNP markers with $FDR < 0.05$ were considered outliers.

False positives are frequent in the detection of outlier loci (Luikart et al. 2003). For this reason, the final set of outlier markers consisted of the loci identified in at least two of the three applied tests, as suggested by Luikart et al. (2003) and considered by Alves-Pereira et al. (2020, 2022).

Statistical analyses carried out with neutral SNP loci

The genetic structure and diversity analyses were performed with neutral SNPs, excluding the outlier loci according to the criterion described above. Discriminant analysis of principal components (DAPC) was performed with the adegenet package (Jombart 2008) in the R program (R Development Core Team 2018). The number of clusters from the DAPC was calculated by the K-means method, which runs different probabilities of cluster numbers, and by using the locations as a *priori* groupings. The analyses were continued based on the locations, as they summarized a greater percentage of the genetic variation.

The genetic relationship between the samples was performed by cluster analysis with the neighbor-joining method, using Nei's genetic distances (Nei 1978) performed with the ape package (Paradis and Schliep 2019) in the R program (R Development Core Team 2018). The dendrogram was edited with FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/softwere/figtree/). Pairwise F_{ST} matrices among locations and among the groups delimited by DAPC were calculated with the poppr package (Kamvar et al. 2014) in the R program (R Development Core Team 2018).

The genetic diversity parameters for the sampled locations of the total number of alleles (A), observed heterozygosity (H_O), and expected heterozygosity (H_E), in addition to the inbreeding coefficient (f), were estimated using the hierfstat (Goudet and Jombart 2020) and poppr (Kamvar et al. 2014) packages in the R program (R Development Core Team 2018). The distribution of genetic variability between and within locations was detected using the analysis of molecular variance (AMOVA) with the hierfstat (Goudet and Jombart 2020) and poppr (Kamvar et al. 2014) packages in the R program (R Development Core Team 2018). To verify the existence of isolation by distance, the Mantel test was performed with the ade4 package (Dray and Dufour 2007; Dray et al. 2007; Bougeard and Dray 2018; Thioulouse et al. 2018), aiming to evaluate the correlation between the genetic divergence from the F_{ST} values of the pairwise matrix between locations and the geographic distances (km), generated from the geographic coordinates, obtained with the geodist package (Padgham and Sumner 2020). A second Mantel test was performed with only the four populations of Paraíba, geographically located closer to each other. The significance level was considered based on 20,000 permutations.

Results

SNP detection and outlier SNPs loci

A total of 5,336 SNP markers were identified for 71 samples (mean sequencing depth = 55.4x; standard deviation = 33.3; mean of 0.54% missing data). Of these, 1,624 SNPs were identified as outlier markers (1,029 by the pcadapt program, 468 by the fsthet program, and 127 by the BayeScan program) (Fig. S3). The final set of outlier markers consisted of 250 SNPs identified by at least two of the three tests performed (Fig. S3). Therefore, the analyses of genetic diversity and population structure were performed with 5,086 SNPs considered neutral, excluding from the total the 250 SNPs identified as outlier loci.

Genomic structure among umbu tree locations

The number of clusters was obtained from the DAPC calculated by the K-means method, and by using the locations as a *priori* groupings. By using the K-means method, two groups were detected whereas group I included the locations from the States of Bahia, Pernambuco, and Minas Gerais, while group II included the four locations of Paraíba State (Fig. S1 and S2). The K-means method retained only one principal component that explained 12.4% of the total variation. In the DAPC analysis based on locations, six principal components were retained, of which the first three explained 26.7% of the variation (Fig. 3a, b). Therefore, the analyses were continued based on the locations, as they summarized a greater percentage of the genetic variation.

A strong structure was found among the umbu tree samples from different locations, especially among the sampled states (Fig. 3c), establishing

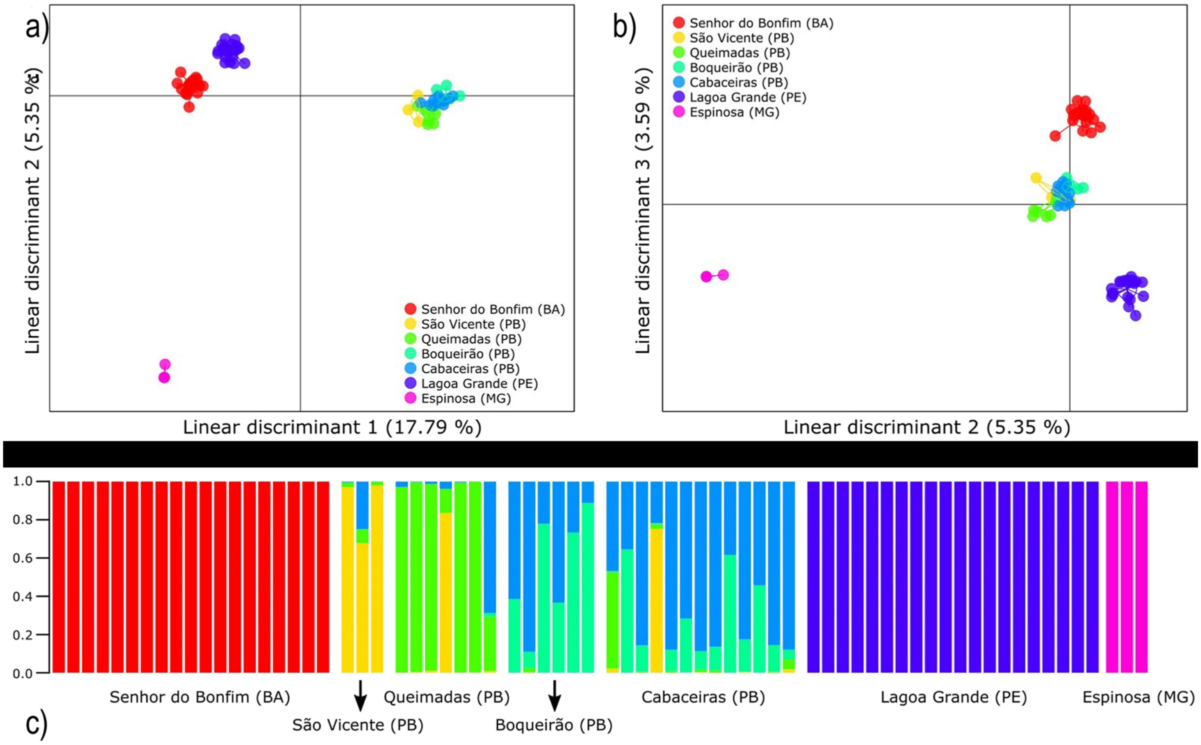

Fig. 3 Discriminant Analysis of Principal Components (DAPC) performed based on 5,086 single nucleotide polymorphism (SNP) markers for 71 accessions of umbu tree (*Spondias tuberosa*): a) Scatter plot of locations (collection sites) considering components 1 and 2; b) Scatter plot of locations considering components 2 and 3; c) Clustering probability analysis according to results generated in the DAPC. Each bar represents an individual, and the white lines separate the locations originating in the states of Bahia (BA), Paraíba (PB), Pernambuco (PE), and Minas Gerais (MG)

an optimal number of groups corresponding to four groups (Fig. 3a, b). The locations of Espinosa-MG, Senhor do Bonfim-BA, and Lagoa Grande-PE formed three isolated groups, all genetically different from each other. As for the four populations collected in Paraíba, all geographically close to each other (Fig. 2), there was an overlapping of genotypes suggesting greater genetic similarity among them.

The neighbor-joining dendrogram based on Nei's genetic distances (1978) clustered the samples into four groups: group I, consisting of individuals from Lagoa Grande-PE; group II, consisting of individuals from Senhor do Bonfim-BA; group III, consisting of individuals from Espinosa-MG; and group IV grouping the individuals from the four locations sampled in Paraíba, with the individuals from São Vicente do Seridó-PB being the most divergent in relation to the individuals from the other locations (Fig. 4).

The Mantel test showed a high and significant correlation between genetic distances (F_{ST}) and geographic distances (km) ($r^2 = 0.974$; $p = 0.0015$) between pairs of locations (Fig. 5a). However, when

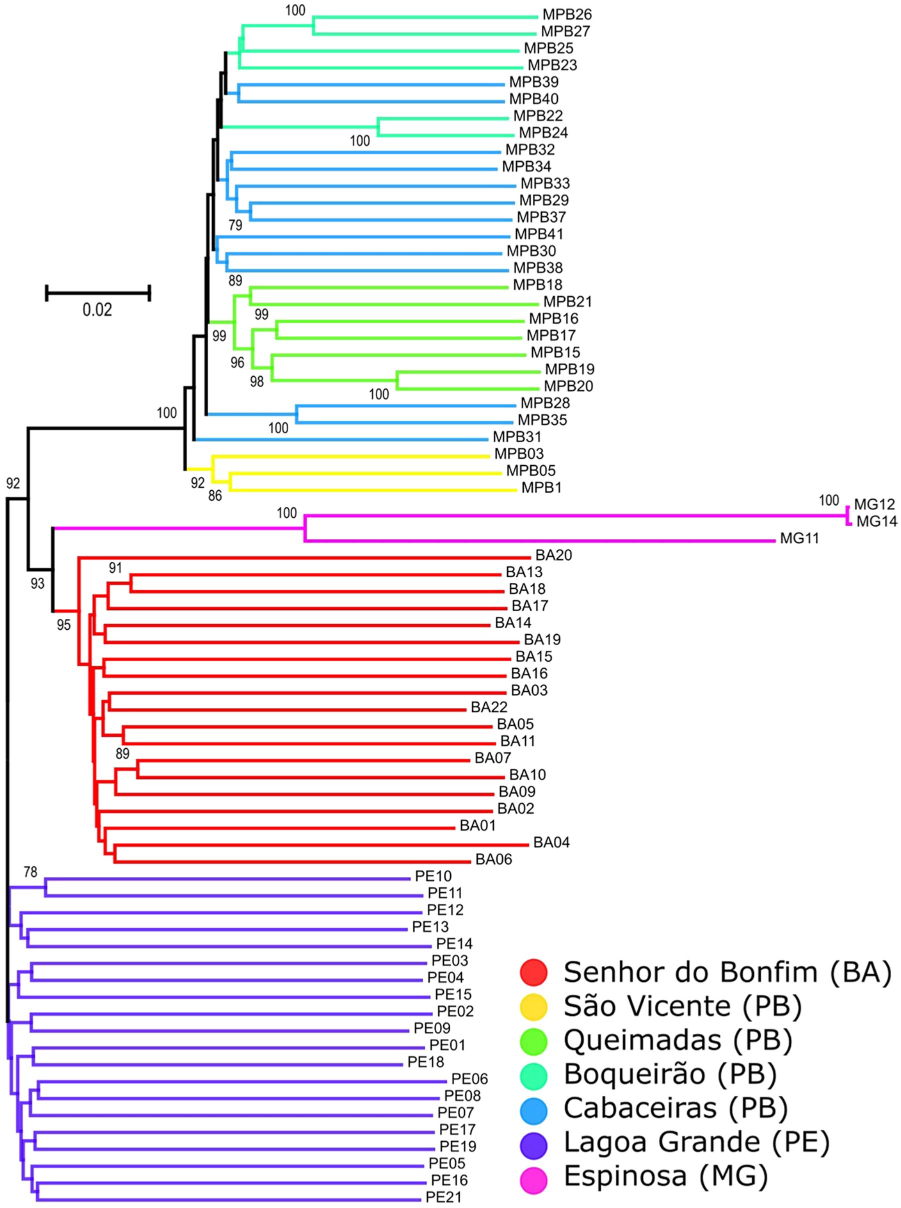

Fig. 4 Neighbor-joining dendrogram built with Nei's (1978) genetic distances among 71 individuals of umbu tree (*Spondias tuberosa*), based on 5,086 single nucleotide polymorphisms (SNPs)

the Mantel test was conducted with only the four locations from Paraíba (Fig. 5b) the result was non-significant ($r^2=0.575$; $p=0.212$), which was already expected due to their greater geographical and genetic proximity (Figs. 2, 3 and 4).

Based on F_{ST} estimates (Table 2), Espinosa-MG is genetically the most distinct from the other locations, with F_{ST} values ranging from 0.202 (Espinosa-MG and Lagoa Grande-PE) to 0.330 (Espinosa-MG and Boqueirão-PB). The locations of Senhor do

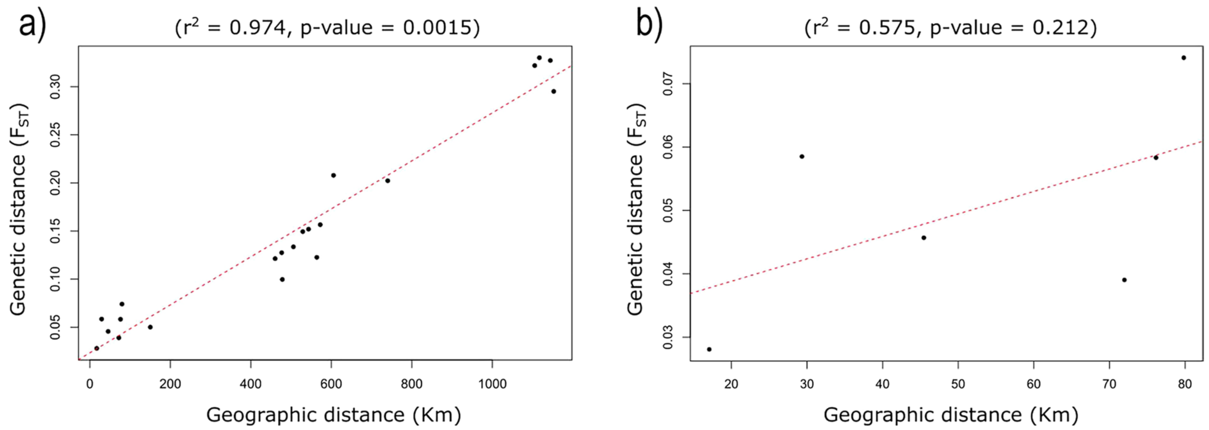

Fig. 5 Correlation between genetic distances (F_{ST}) and geographic distances (km) between pairs of locations from the four states (a), and between pairs of locations from Paraíba (b) performed for umbu tree (*Spondias tuberosa*)

Table 2 F_{ST} estimates between pairs of locations (lower diagonal) and respective 95% confidence intervals (upper diagonal) for the umbu trees (*Spondias tuberosa*) locations

	Senhor do Bonfim-BA	São Vicente-PB	Queimadas-PB	Boqueirão-PB	Cabaceiras-PB	Lagoa Grande-PE	Espinosa-MG
Senhor do Bonfim-BA		(0.113:0.131)	(0.149:0.164)	(0.144:0.160)	(0.143:0.156)	(0.047:0.053)	(0.198:0.218)
São Vicente-PB	0.123		(0.065:0.083)	(0.049:0.067)	(0.032:0.046)	(0.091:0.108)	(0.283:0.308)
Queimadas-PB	0.157	0.074		(0.051:0.066)	(0.041:0.051)	(0.127:0.140)	(0.315:0.340)
Boqueirão-PB	0.152	0.058	0.059		(0.024:0.033)	(0.121:0.135)	(0.318:0.342)
Cabaceiras-PB	0.149	0.039	0.046	0.028		(0.115:0.127)	(0.310:0.334)
Lagoa Grande-PE	0.050	0.100	0.134	0.128	0.121		(0.193:0.211)
Espinosa-MG	0.208	0.295	0.327	0.330	0.322	0.202	

Table 3 Analysis of molecular variance (AMOVA) based on 5,086 single nucleotide polymorphisms (SNPs) used to identify sources of genetic variability among and within sampled locations of umbu trees (*Spondias tuberosa*)

Source of variation	DF[a]	SS	MS	% variation	PhiST	p-value
Among locations	6	12,822.9	2137.1	22.1	0.221	0.00005
Within locations	64	37,242.1	581.9	77.9		
Total	70	50,065.0	715.2			

[a] DF = degrees of freedom, SS = sum of squares, MS = mean squares, PhiST = estimate analogous to F_{ST}

Bonfim-BA and Lagoa Grande-PE are genetically closer to each other (Fst = 0.050) but genetically distant from the others. The locations in Paraíba are genetically closer to each other, with the F_{ST} values ranging from 0.028 (Boqueirão and Cabaceiras) to 0.074 (São Vicente and Queimadas).

The AMOVA showed that most of the observed genetic variation was found within locations (77.9%) (Table 3). However, the variation among locations (F_{ST} = 0.221) is high (Hartl and Clark 2007) and significant, corroborating the results observed in the DAPC and in the dendrogram (Figs. 3 and 4).

Genetic diversity for the umbu tree locations

In the analysis of genetic diversity for each location (Table 4), the total number of alleles ranged from 7,231 (São Vicente-PB) to 9,402 (Lagoa Grande-PE), with an average of 8,136.7 alleles. All locations presented polymorphism greater than 71%, and allelic richness ranged from 1.352 (São Vicente-PB) to 1.538 (Lagoa Grande-PE). Results show moderate to high levels of genetic diversity for the species, with an excess of heterozygotes in all locations except for Lagoa Grande-PE. This can be evidenced by the negative and close to zero inbreeding coefficients (mean F_{IS} = -0.117). The locations of Espinosa-MG, Lagoa Grande-PE and Senhor do Bonfim-BA stand out as having the highest heterozygosities, while the four locations in Paraíba (PB) State showed the lowest heterozygosities.

Discussion

For the first time, SNP markers obtained by the GBS method have been used to assess the genetic diversity and population structure of *S. tuberosa*. This study found variable levels of genetic structure among the sampled locations of umbu tree. When comparing the location of Espinosa-MG with the other locations, the F_{ST} values were all above 0.20, varying from 0.202 to 0.330, which are considered high to very high according to Hartl and Clark (2007). And except for the four geographically closer locations from Paraíba all the other locations showed moderate genetic structure among each other, with F_{ST} varying from 0.05 to 0.157. The same patterns of genetic structure were observed in the DAPC, where the locations from Paraíba formed a single admixed group, which was genetically divergent in relation to the groups of individuals from the other states. This result refutes the first hypothesis of this study, which expected low genetic structure among locations, considering that *S. tuberosa* presents allogamy and gametophytic

Table 4 Genetic diversity parameters and inbreeding coefficient for umbu trees (*Spondias tuberosa*) locations evaluated with 5,086 single nucleotide polymorphisms (SNPs)

Locations	N[a]	A	P(%)	Ar	H_O	H_E	F_{IS}	F_{IS} CI95%
Espinosa-MG	3	7,787	76.6	1.453	0.294	0.215	–0.364	–0.787:-0.120
Senhor do Bonfim-BA	19	9,257	91.0	1.528	0.244	0.244	–0.002	–0.029: 0.019
São Vicente-PB	3	7,231	71.1	1.352	0.198	0.162	–0.225	–0.801:-0.027
Queimadas-PB	7	7,641	75.1	1.371	0.185	0.171	–0.080	–0.203:-0.021
Boqueirão-PB	6	7,581	74.5	1.369	0.189	0.171	–0.107	–0.287:-0.030
Cabaceiras-PB	13	8,058	79.2	1.403	0.193	0.185	–0.043	–0.084:-0.018
Lagoa Grande-PE	20	9,402	92.4	1.538	0.246	0.247	0.004	–0.017: 0.019
Mean	–	8,136.7	80.0	1.431	0.221	0.199	–0.117	

[a] N = number of samples, A = total number of alleles, P% = percentage of polymorphic loci, Ar = mean allelic richness per locus, H_O = observed heterozygosity, H_E = expected heterozygosity, F_{IS} = inbreeding coefficient, CI95% = confidence interval of 95%

self-incompatibility. The variable levels of genetic structure observed in this study may be related to the edaphoclimatic differences among sampled locations due to the vast extension of the Caatinga biome. According to Balbino et al. (2018), the Caatinga area has approximately 850,000 km^2, is characterized by discontinuous ecoregions, which consist of different types of vegetation, average annual temperatures that vary between 27 °C and 29 °C and precipitation ranging from 300 to 800 mm; it also comprises large plateaus up to 1,000 m and lowland peneplains.

The lower levels of genetic structure among the four locations in Paraíba, suggested by the small F_{ST} values and the overlap of individuals from different locations in the DAPC, may be due to their smaller geographic distances than the other locations. Such geographic proximity always results in higher edaphoclimatic uniformity and provides a larger exchange of fruits through trade and by relations among the community of local people, favoring gene flow. Furthermore, the umbu fruits serve as food for many Caatinga animals, both domestic and wild animals. Therefore, it may have provided more gene flow among the closer locations such as Queimadas, Boqueirão, and Cabaceiras in the Paraíba State.

Mantel's test results also suggested isolation by distance when considering all the sampled locations but not when considering only the closest locations from the Paraíba State, in accordance with the observed patterns of genetic structure revealed by F_{ST} values and DAPC. These differences might also be explained by climatic differences in the area covered. Table 1 shows similar average annual temperatures among all collection sites, with greater variation in relation to precipitation and altitude. Espinosa-MG and Senhor do Bonfim-BA occur in As climate (tropical savannah: warm, with winter and autumn rains), with higher annual precipitation values (798.3 mm and 679.7, respectively) and altitude above 490 m, while Lagoa Grande-PE has a BSh climate (semi-arid: hot and dry, with winter rains), with precipitation of 525.3 mm and altitude of 429.9 m. It is important to note that although the three samples from Espinosa-MG certainly do not represent the genetic diversity of umbu populations from the state of Minas Gerais, these were more isolated and genetically distant. An interesting observation is that Espinosa-MG is located at a transition among the Caatinga, Atlantic Forest and Cerrado biomes (Fig. 2), which could also explain its high differentiation from the other locations, implying an adaptation to different ecological conditions, such as soil type, temperatures, etc. In relation to Paraíba, the Queimadas location, with an As climate, presents much higher precipitation (639 mm) than the other three locations, all with a BSh climate, including São Vicente do Seridó. This latter location is situated at a higher altitude (573.8 m) when compared to the PB locations, which might explain its slight differentiation.

Other research (Santos et al. 2008, 2021b; Lins Neto et al. 2013; Balbino et al. 2018) also observed high structuring between umbu locations, using different markers with lower genomic coverage. Santos et al. (2008) studied the genetic variation in 15 ecoregions of the Brazilian semi-arid region using AFLP markers. They observed that the genetic diversity of umbu trees was not uniformly dispersed and was highly structured between ecoregions (31.38%), suggesting restricted gene flow between populations. Balbino et al. (2018) studied the phylogeographic pattern of umbu trees using chloroplast sequences and six nuclear SSR markers in individuals from 20 locations in the states of Alagoas and Minas Gerais. They observed moderate genetic structure (13% of variation among populations) with SSR markers and described two genetic groups: a larger one containing most of the Caatinga populations and a small group closer to the Atlantic Forest, identifying the Caatinga as a large and continuous refuge and the region close to the interface between the Caatinga and the Atlantic Forest as a second refuge. Analyzing populations from Minas Gerais, Bahia, and Pernambuco, like our study, Santos et al. (2021b), based on nuclear SSR markers, observed that accessions from Bahia and Pernambuco formed a separate group from accessions from Minas Gerais, with a genetic structure equivalent to 12% among groups. In our study, Minas Gerais, Bahia, and Pernambuco locations were separated into three distinct groups, showing high genetic structure (22.1% among locations in AMOVA); they could be considered as three separate populations. This result was expected since SNP markers are more efficient in separating genetic groups than other markers (Huq et al. 2016; Leitwein et al. 2020). High genomic structure (38.6%) among locations from three Brazilian biomes based on SNP markers was reported for a fruit tree of the same genus (*S. monbim*, known as "cajá") (Silva

2021). SNP markers were used by Garcia et al. (2024) in another fruit species (*Platonia insignis*, known as "bacuri"), for which even higher levels of diversity between locations were found (68.3%).

The results of the present study regarding the genetic diversity of umbu locations show that the Caatinga populations of this species present moderate to high levels of diversity ($H_O = 0.221$ and $H_E = 0.199$, on average), with most of the variability (77.9%) occurring within locations. This result partially refutes our second hypothesis, as despite the species being in a state of genetic vulnerability, with a reduction in its populations (Mertens et al. 2017), *S. tuberosa* still maintains moderate to high levels of diversity. It should be noted that these locations occupy a region of the Brazilian semi-arid with intense anthropic action; their xylopods are used to extract water during the dry period, and the plant has a reduced capacity for regeneration (Mertens et al. 2017). Still, *S. tuberosa* is not currently at imminent risk of extinction (Mitchell and Daly 2015; Mertens et al. 2017), although its genetic diversity may be compromised due to the destruction of its habitat. The Caatinga biome lost around 150,000 km^2 of primary vegetation between 1985 and 2020, a reduction of 26.4%, with 112,000 km^2 replaced by agriculture, and some other areas are compromised by accelerated desertification (Marques 2022).

Even in this context of vulnerability, the umbu trees maintain levels of heterozygosity and diversity higher than those found for species of the same genus, such as *S. mombin*, which presented $H_O = 0.17$ and $H_E = 0.19$ on average (Silva 2021); and phylogenetically distant species, such as the fruit tree *Platonia insignis* ($H_O = 0.081$; $H_E = 0.092$ on average) (Garcia et al. 2024). Similar genetic diversity estimates were found for other tropical trees, such as cacao (*Theobroma cacao* L.) cultivars in Honduras and Nicaragua ($H_O = 0.206$; $H_E = 0.367$, on average) (Ji et al. 2013), *Parkia platycephala* Benth. located outside and inside the Sete Cidades National Park, in the state of Piauí, in a transition zone between Caatinga and Cerrado ($H_O = 0.29$; $H_E = 0.29$, on average) (Morais et al. 2023).

Negative inbreeding coefficients indicate an excess of heterozygotes in the studied umbu locations. These results, in addition to the high genetic diversity observed within locations, are in line with the predominantly allogamous reproductive system for the species, which also exhibits gametophytic self-incompatibility (Leite and Machado 2010; Santos and Gama 2013; Santos et al. 2021a, b). *S. tuberosa* is an andromonoecious species and, therefore, has equal numbers of hermaphrodite and male flowers on the same individual. Thus, the large quantity of pollen grains produced increases the fertilization viability of hermaphrodite flowers, also increasing male sexual expression (Nadia et al. 2007). Increased male sexual expression can favor cross-pollination through increased pollen flow (Symon 1979; Medan and D'Ambrogio 1998), which is enhanced by the action of pollinators. These are essential to generate new genotypic combinations and to maintain high levels of genetic variation in umbu populations. Umbu flowers have a slight sweet odor, which attracts visits from various pollinators. Nadia et al. (2007) reported 17 species of insects, including seven wasps, six bees, and four flies, as pollinators of umbu plants. Bees, *Scaptotrigona postica flavisetis* and *Trigona fuscipennis*, were the main pollinators, with emphasis also on wasps, mainly *Polybia ignobilis*. It is noteworthy that umbu tree has zoochoric fruits (Nadia et al. 2007) and thus has a series of characteristics, such as the presence of an edible portion involving the seed and attractive colors, which stimulate and facilitate its consumption by animals and, consequently, the dispersal of its seeds. Thus, andromonoecy, self-incompatibility, and zoochory are advantageous characteristics for maintaining the variability of umbu populations. Another important aspect to note is that local people along the Caatinga biome have an old habit as part of their culture of protecting umbu trees against fire and deforestation, especially plants that produce sweet fruits (Silva E.F., personal communication). This practice certainly provides important support for the conservation of the species; however, it results in inadvertent selection and favors recombination between plants with sweet fruits to the detriment of plants that produce acidic fruits, which are often eliminated.

In conclusion, for the in situ conservation of umbu genetic resources, we can suggest that all sampled sites should be considered as priority, as they have high genetic diversity and different alleles in relation to the species' gene pool. The four locations in Paraíba can be considered a single population, as they are genetically closer, although divergent individuals were also observed within them. Additionally, Minas

Gerais, Pernambuco, and Bahia locations can be considered genetically different populations. Likewise, if the interest is ex situ conservation, it is recommended to collect seeds in all locations, in each state covered by the Caatinga biome. Finally, very low levels of inbreeding were detected within the umbu locations, which seems to contradict our hypothesis of genetic vulnerability, which is a promising result for the conservation of the genetic resources of umbu tree in the Brazilian Caatinga.

Author contributions Planning and design of research: WFN, MIZ and EAV; Funding: EAV and MIZ; Material preparation, field collection expedition, and laboratory analysis: WFN, MMSGD, EFS, IASC, DPR, FRAP, CEB and CBG; Statistical analysis: AAP and FMC; The first draft of the manuscript was written by EAV and WFN; All authors commented on previous versions of the manuscript and contributed to the final manuscript.

Funding This study was financially supported by *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP, 2019/04100–6). Scholarships were provided by FAPESP (2019/15544–2 to IASC and 2021/04698–9 to FMC), by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), (309445/2020–5 to EAV).

Data availability All data generated or analyzed during this study are included in this article and available in the Mendeley repository: https://data.mendeley.com/datasets/xpxs3fkdzz/1

Declarations

Competing interests The authors declare no competing interests.

Ethical approval This research was registered in the National System for the Management of Genetic Heritage and Associated Traditional Knowledge (SisGen) (registration nº A3AF200). It does not contain any research with human participants or animals performed by any authors.

References

Albuquerque UP, Medeiros PM, Almeida ALS, Monteiro JM, Freitas Lins Neto EM, Melo JG, Santos JP (2007) Medicinal plants of the caatinga (semi-arid) vegetation of NE Brazil: a quantitative approach. J Ethnopharmacol 3:325–354. https://doi.org/10.1016/j.jep.2007.08.017

Almeida ALS, Albuquerque UP, Castro CC (2011) Reproductive biology of *Spondias tuberosa* Arruda (Anacardiaceae), an endemic fructiferous species of the caatinga (dry forest), under different management conditions in northeastern Brazil. J Arid Environ 75:330–337. https://doi.org/10.1016/j.jaridenv.2010.11.003

Alves JJ, Araujo MP, Nascimento SS (2009) Degradação da caatinga: uma investigação ecogeográfica. Rev Caatinga 22:126–135. https://doi.org/10.14393/RCG92715740

Alves-Pereira A, Clement CR, Picanço-Rodrigues D, Veasey EA, Dequigiovanni G, Ramos SLF, Pinheiro JB, Souza AP, Zucchi MI (2020) A population genomics appraisal suggests independent dispersals for bitter and sweet manioc in Brazilian Amazonia. Evol Appl 13:342–361. https://doi.org/10.1111/eva.12873

Alves-Pereira A, Zucchi MI, Clement CR, Viana JPG, Pinheiro JB, Veasey EA, de Souza AP (2022) Selective signatures and high genome-wide diversity in traditional Brazilian manioc (*Manihot esculenta* Crantz) varieties. Sci Rep 12(1):1268. https://doi.org/10.1038/s41598-022-05160-8

Andrews S (2010) FastQC: A quality control tool for high throughput sequence data [Online]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 28 February 2024

Balbino E, Caetano B, Almeida C (2018) Phylogeographic structure of *Spondias tuberosa* Arruda Câmara (Anacardiaceae): seasonally dry tropical forest as a large and continuous Refuge. Tree Genet Genomes 14:67. https://doi.org/10.1007/s11295-018-1279-4

Balbino E, Martins G, Morais S, Almeida C (2019) Genome survey and development of 18 microsatellite markers to assess genetic diversity in *Spondias tuberosa* Arruda Câmara (Anacardiaceae) and cross-amplifcation in congeneric species. Mol Biol Rep 46:3511–3517. https://doi.org/10.1007/s11033-019-04768-w

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

Bougeard S, Dray S (2018) Supervised multiblock analysis in R with the ade4 package. J Stat Softw 86:1–17. https://doi.org/10.18637/jss.v086.i01

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes Genom Genet 1(3):171–182. https://doi.org/10.1534/g3.111.000240

Cavalcanti NB, Drumond MA, Resende GM (2004) Uso das folhas do umbuzeiro (*Spondias tuberosa* Arruda) na alimentação de caprinos e ovinos no Semi-Árido Nordestino. Agrossilvicultura 1(2):131–134. https://ainfo.cnptia.embrapa.br/digital/bitstream/item/176401/1/Agrossilvicultura-v.1-n.2-p.-131 134–2004.pdf

Cavalcanti NB, Resende GM, Brito LTL (2010) O crescimento de plantas de imbuzeiro (*Spondias tuberosa* ARRUDA) no semiárido de Pernambuco. Eng Ambiental 7:21–31. https://ainfo.cnptia.embrapa.br/digital/bitstream/item/37112/1/Nilton-2010.pdf

Cockerham CC, Weir BS (1993) Estimation of gene flow from F-statistics. Evol 47(3):855–863. https://doi.org/10.2307/2410189

Cordeiro BMPC, Santos NDL, Ferreira MRA, Araújo LCC, Carvalho Junior AR, Santos ADC, Oliveira AP, Silva AG, Falcão EPS, Correia MTS, Almeida JGS, Silva LCN, Soares LAL, Napoleão TH, Silva MV, Paiva PMG (2018) Hexane extract from *Spondias tuberosa* (Anacardiaceae) leaves has antioxidant activity and is an anti-*Candida* agent by causing mitochondrial and lysosomal damages.

BMC Complement Altern Med 18:284. https://doi.org/10.1186/s12906-018-2350-2

Cortinovis G, Frascarelli G, Di Vittori V, Papa R (2020) Current state and perspectives in population genomics of the common bean. Plants 9(3):330. https://doi.org/10.3390/plants9030330

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinform 27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dray S, Dufour A (2007) The ade4 Package: Implementing the duality diagram for ecologists. J Stat Softw 22:1–20. https://doi.org/10.18637/jss.v022.i04.

Dray S, Dufour A, Chessel D (2007) The ade4 Package – II: Two-Table and K-Table Methods. R News 7(2):47–52. https://cran.r-project.org/doc/Rnews/. Accessed 28 February 2024

Epstein L (1998) A riqueza do umbuzeiro. Revista Bahia Agrícola 2:31–34. http://www.seagri.ba.gov.br/content/riqueza-do-umbuzeiro

Flanagan SP, Jones AG (2017) Constraints on the FST-heterozygosity outlier approach. J Hered 108(5):561–573. https://doi.org/10.1093/jhered/esx048

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. Genetics 180(2):977–993. https://doi.org/10.1534/genetics.108.092221

Garcia CB, Silva AV, Carvalho IAS, Nascimento WF, Ramos SLF, Rodrigues DP, Zucchi MI, Costa FM, Alves-Pereira A, Batista CEA, Amaral DD, Veasey EA (2024) Low diversity and high genetic structure for *Platonia insignis* Mart., an endangered fruit tree species. Plants 13:1033. https://doi.org/10.3390/plants13071033

Goonetilleke SN, March TJ, Wirthensohn MG, Arús P, Walker AR, Mather DE (2018) Genotyping by Sequencing in Almond: SNP discovery, linkage mapping, and marker design. G3: Genes Genom Genet 8:161–172. https://doi.org/10.1534/g3.117.300376

Goudet J, Jombart T (2020) hierfstat: Estimation and Tests of Hierarchical F-Statistics. R package version 0.5–7. https://CRAN.R-project.org/package=hierfstat. Accessed 28 February 2024

Haddad NM, Brudvig LA, Clobert J, Davies KF, Gonzalez A, Holt RD, Lovejoy TE, Sexton JO, Austin MP, Collins CD, Cook WM, Damschen EI, Ewers RM, Foster BL, Jenkins CN, King AJ, Laurance WF, Levey DJ, Margules CR, Melbourne BA, Nicholls AO, Orrock JL, Song D-X, Townshend JR (2015) Habitat fragmentation and its lasting impact on Earth's ecosystems. Sci Adv 1:1–10. https://doi.org/10.1126/sciadv.1500052

Hartl DL, Clark AG (2007) Principles of population genetics. Oxford University Press, Oxford

https://doi.org/10.1371/journal.pone.0206085

Huq MA, Akter S, Nou IS, Kim HT, Jung YJ, Kang KK (2016) Identification of functional SNPs in genes and their effects on plant phenotypes. J Plant Biotechnol 43:1–11. https://doi.org/10.5010/JPB.2016.43.1.1

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

Ji K, Zhang D, Motilal LA, Boccara M, Lachenaud P, Meinhardt LW (2013) Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. Genet Resour Crop Evol 60:441–453. https://doi.org/10.1007/s10722-012-9847-1

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinform 24:1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Kamvar ZN, Tabima JF, Grünwald NJ (2014) Pppr: a R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281. https://doi.org/10.7717/peerj.281

Leite AVL, Machado IC (2010) Reproductive biology of woody species in Caatinga, a dry forest of northeastern Brazil. J Arid Environ 74:1374–1380. https://doi.org/10.1016/j.jaridenv.2010.05.029

Leitwein M, Duranton M, Rougemont Q, Gagnaire P-A, Bernatchez L (2020) Using haplotype information for conservation genomics. Trends Ecol Evol 35(3):245–258. https://doi.org/10.1016/j.tree.2019.10.012

Lins Neto EMF, Peroni N, Albuquerque UP (2010) Traditional knowledge and management of Umbu (*Spondias tuberosa*, Anacardiaceae): an endemic species from the semi-arid region of Northeastern Brazil. Econ Bot 64(1):11–21. https://doi.org/10.1007/s12231-009-9106-3

Lins Neto EMF, Oliveira LF, Britto FB, Albuquerque UP (2013) Traditional knowledge, genetic and morphological diversity in populations of *Spondias tuberosa* Arruda (Anacardiaceae). Genet Resour Crop Evol 60:1389–1406. https://doi.org/10.1007/s10722-012-9928-1

Lins Neto EMF, Peroni N, Casas A, Parra F, Aguirre X, Guillén S, Albuquerque UP (2014) Brazilian and Mexican experiences in the study of incipient domestication. J Ethnobiol Ethnomed 10:33. https://doi.org/10.1186/1746-4269-10-33

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nature Reviews Genet 4(12):981–994. https://doi.org/10.1038/nrg1226. (**PMID: 14631358**)

Luu K, Bazin E, Blum MG (2017) pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour 17(1):67–77. https://doi.org/10.1111/1755-0998.12592

Marques L (2022) Brasil, 200 anos de devastação: O que restará do país após 2022? Estud Av 36(105):169–185. https://doi.org/10.1590/s0103-4014.2022.36105.011

Medan D, D'Ambrogio AC (1998) Reproductive biology of the andromonoecious shrub *Trevoa quinquenervia* (Rhamnaceae). Bot J Linn 126:191–206. https://doi.org/10.1111/j.1095-8339.1998.tb02526.x

Mertens J, Germer J, Siqueira Filho JA, Sauerborn J (2017) *Spondias tuberosa* Arruda (Anacardiaceae), a threatened tree of the Brazilian Caatinga? Braz J Biol 77:542–552. https://doi.org/10.1590/1519-6984.18715

Miraldo A, Li S, Borregaard MK, Flórez-rodríguez A, Gopalakrishnan S, Rizvanovic M, Wang Z, Rahbeck C, Marske KA, Nogués-Bravo D (2016) An anthropocene map of genetic diversity. Science 353(6307):1532–1535. https://doi.org/10.1126/science.aaf4381

Mitchell JD, Daly DC (2015) A revision of *Spondias* L. (Anacardiaceae) in the Neotropics. PhytoKeys 55:1–92. https://doi.org/10.3897/phytokeys.55.8489

Morais JGS, Costa MF, Alves-Pereira A, Zucchi MI, Baldin PJ, Araujo ASF, Silva VB, Ferreira-Gomes RL, Lopes ACA (2023) Genomic population structure of *Parkia platycephala* Benth. (Leguminosae) from Northeastern Brazil. Genet Resour Crop Evol 70:251–261. https://doi.org/10.1007/s10722-022-01431-5

Moreira PA, Pimenta MAS, Saturnino HM, Gonçalves NP, Oliveira DA (2007) Variabilidade genética de umbuzeiro na Região Norte do Estado de Minas Gerais. Revista Brasil Biol 5:279–281. https://seer.ufrgs.br/index.php/rbrasbioci/article/view/115915/63196

Nadia TL, Machado IC, Lopes AV (2007) Polinização de *Spondias tuberosa* Arruda (Anacardiaceae) e análise da partilha de polinizadores com *Ziziphus joazeiro* Mart. (Rhamnaceae), espécies frutíferas e endêmicas da caatinga. Revista Brasil Bot 30:89–100. https://doi.org/10.1590/S0100-84042007000100009

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genet 89:583–590. https://doi.org/10.1093/genetics/89.3.583

Newbold T, Hudson LN, Hill SLL, Contu S, Lysenko I, Senior RA, Börger L, Bennett DJ, Choimes A, Collen B, Day J, Palma A, Díaz S, Echeverria-Londoño S, Edgar MJ, Feldman A, Garon M, Harrison MLK, Alhusseini T, Ingram DJ, Itescu Y, Kattge J, Kemp V, Kirkpatrick L, Kleyer M, Correia DLP, Martin CD, Meiri S, Novosolov M, Pan Y, Philips HRP, Purves DW, Robinson A, Simpson J, Tuck SL, Weiher E, White HJ, Ewers RM, Mace GM, Scharlemann JPW, Purvis A (2015) Global effects of land use on local terrestrial biodiversity. Nature 520:45–50. https://doi.org/10.1038/nature14324

Nobre LLM, Santos JDO, Leite R, Almeida C (2018) Phylogenomic and single nucleotide polymorphism analyses revealed the hybrid origin of *Spondias bahiensis* (family Anacardiaceae): de novo genome sequencing and comparative genomics. Genet Mol Biol 41(4):878–883. https://doi.org/10.1590/1678-4685-GMB-2017-0256

Padgham M, Sumner MD (2020) Geodist: Fast, Dependency-Free Geodesic Distance Calculations. R package version 0.0.6. https://CRAN.R-project.org/package=geodist. Accessed 29 February 2024

Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinform 35:526–528. https://doi.org/10.1093/bioinformatics/bty633

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5(3):92. https://doi.org/10.3835/plantgenome2012.05.0005

Poland JA, Brow 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

R Development Core Team (2018) R: A language and environment for statistical computing. R. Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org. Accessed 15 December 2018

Santana JAS, Souto JS (2006) Diversidade e estrutura fitossociológica da caatinga na Estação Ecológica do Seridó - RN. Rev Biol Ciênc Terra 6(2):232–242. https://www.redalyc.org/articulo.oa?id=50060215

Santos CAF, Oliveira VR (2008) Inter-relações genéticas entre espécies do gênero *Spondias* com base em marcadores AFLP. Rev Bras Frutic 30(3):731–735. https://doi.org/10.1590/S0100-29452008000300028

Santos CAF, Rodrigues MA, Zucchi MI (2008) Variabilidade genética do umbuzeiro no Semi-Árido brasileiro, por meio de marcadores AFLP. Pesqui Agropecu Bras 43:1037–1043. https://doi.org/10.1590/S0100-204X2008000800013

Santos CAF, Gama RNCS (2013) An AFLP estimation of the outcrossing rate of *Spondias tuberosa* (Anacardiaceae), an endemic species to the Brazilian semiarid region. Ver Biol Trop 61(2):577–582. https://doi.org/10.15517/rbt.v61i2.11150.

Santos CAF, Oliveira VR, Rodrigues MA, Ribeiro HLC, Drumond MA (2011) Estimativas de polinização cruzada em população de *Spondias tuberosa* Arruda (Anacardiaceae) usando marcador AFLP. Rev. Árvore 35(3):691–697. https://www.scielo.br/j/hb/a/LNPNdFrq3RTywVPdvxP986j/?format=pdf

Santos VN, Gama RNCS, Santos CAF (2021a) Development and transferability of microsatellite loci for *Spondias tuberosa* (Anacardiaceae: Sapindales), a species endemic to the Brazilian semi-arid region. *Genet Mol Res* 20(2):gmr18778. https://doi.org/10.4238/gmr18778

Santos VN, Santos CAF, Oliveira VR, Costa AES, Silva FFS (2021b) Diversity and genetic structure of *Spondias tuberosa* (Anacardiaceae) accessions based on microsatellite loci. Rev Biol Trop 69(2):640–648. https://doi.org/10.15517/rbt.v69i2.44194

Santos CAF (1997) Dispersão da variabilidade fenotípica do umbuzeiro no semi-árido brasileiro. Pesqui Agropecu Bras 32:923–930. https://seer.sct.embrapa.br/index.php/pab/article/view/4729

Silva EC, Nogueira RJMC, Araújo FP, Melo NF, Azevedo Neto ADA (2008) Physiological responses to salt stress in young umbu plants. Environ Exp Bot 63:147–157. https://doi.org/10.1016/j.envexpbot.2007.11.010

Silva AV (2021) Genômica populacional da cajazeira (*Spondias mombin* L.). Dissertation, Luiz de Queiroz College of Agriculture, University of São Paulo, São Paulo, Brazil

Siqueira EMS, Félix-Silva J, Araújo LML, Fernandes JM, Cabral B, Gomes JAS, Roque AA, Tomaz JC, Lopes NP, Fernandes-Pedrosa MF, Giordani RB, Zucolloto SM (2016) *Spondias tuberosa* (Anacardiaceae) leaves: profiling phenolic compounds by HPLC-DAD and LC–MS/MS and in vivo anti-inflammatory activity. Biomed Chromatogr 30:1656–1665. https://doi.org/10.1002/bmc.3738

Souza IGB, Souza VAB, Silva KJD, Lima PSC (2016) Multivariate analysis of ‘bacuri’ reproductive and vegetative morphology. *Comun Sci* 7:232–240. https://doi.org/10.14295/CS.v7i2.779

Stephenson AG (1981) Flower and fruit abortion: proximate causes and ultimate functions. Ann Rev Ecol Syst 12:253–279. https://www.jstor.org/stable/2097112

Symon DE (1979) Sex forms in *Solanum* (Solanaceae) and the role of pollen collecting insects. In: Hawkes JG, Lester RN, Skelding AD (eds) The biology and taxonomy of the Solanaceae, 1st. Academic Press, London, pp 385–398

Thioulouse J, Dray S, Dufour A, Siberchicot A, Jombart T, Pavoine S (2018) Multivariate analysis of ecological data with ade4. Springer, New York

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.