UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Molecular diversity assessment of a pantropical cash crop

João Miguel Ricardo Barnabé

Mestrado em Biologia Molecular e Genética

Dissertação orientada por:
Professora Doutora Filipa Monteiro
Professora Doutora Maria Romeiras

2022

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Molecular diversity assessment of a pantropical cash crop

João Miguel Ricardo Barnabé

Mestrado em Biologia Molecular e Genética

Dissertação orientada por:
Professora Doutora Filipa Monteiro
Professora Doutora Maria Romeiras

2022

ACKNOWLEDGEMENTS

This master's thesis would not have been possible without significant assistance, for which I am eternally grateful.

I would like to thank my supervisors, Professor Dr. Filipa Monteiro and Professor Dr. Maria Romeiras, for always believing in my ability to complete this stage of my life, with their knowledge, friendship, and rigor I was learning how to overcome my limitations, managing to improve in all aspects, thank you for all the support we have walked these past years.

I also want to thank my family, especially my parents and sisters, for their unwavering support. Many thanks to my dear friend Rita Moura for her unwavering and intellectual support, as well as to my friends and colleagues Maria João, Amidu Barai, and Lara Guterres. To Miguel Brilhante, Guilherme Roxo, and all of my colleagues who helped me.

I also want to thank João Barnabé for his perseverance, resilience, and inner drive.

RESUMO

O cajueiro (*Anacardium occidentale* L.) é uma planta tropical que pertence à família Anacardiaceae, que inclui culturas agrícolas com importância económica. Para além do cajueiro, que é nativo do Brasil, inclui a manga (*Mangifera indica* L.) de origem asiática e o pistácio (*Pistacia vera* L.) de origem mediterrânica. O cajueiro adquiriu um estatuto de cultura de rendimento devido ao elevado valor da castanha de caju nos mercados internacionais. A introdução do cajueiro foi inicialmente promovida com o objetivo de combater a desertificação e erosão do solo, e mais tarde, adquiriu elevada importância económica nos países produtores. Atualmente, o cajueiro encontra-se distribuído em quase todas as regiões tropicais do globo, sendo uma importante fonte de rendimento ao nível das comunidades rurais dos países produtores. Reconhece-se que o cajueiro é uma cultura agrícola com grande valor económico nas zonas tropicais, e nos últimos anos tem-se assistido a um aumento do interesse da castanha de caju devido a uma forte procura no mercado mundial. Apesar da grande expansão da cultura a nível pantropical, poucos estudos foram desenvolvidos para avaliar e caracterizar a diversidade genética dos recursos genéticos desta cultura, contribuindo para a falta de informação orientada para a utilização racional e sustentável do cajueiro, apesar do reconhecido valor económico que esta espécie representa a nível mundial. Com exceção de Portugal, os países membros da Comunidade dos Países de Língua Portuguesa (CPLP) estão dispersos pelos três continentes do hemisfério sul: América, África e Ásia, representando uma grande diversidade agroecológica incluindo países como o Brasil, Guiné-Bissau ou Moçambique a pequenos arquipélagos como Timor-Leste na Ásia ou Cabo Verde ou São Tomé e Príncipe no continente africano.

O principal objetivo deste trabalho pretende caracterizar a diversidade genética e avaliar a estrutura populacional do cajueiro a nível pantropical, utilizando os países da CPLP como um caso de estudo, onde se incluem alguns dos maiores países produtores de caju. Para avaliar a variabilidade e diversidade intraespecífica do cajueiro, foram usados marcadores moleculares do tipo SSRs (Simple Sequence Repeats) ou microssatélites, que são frequentemente aplicados na avaliação intraespecífica da diversidade genética de várias culturas agrícolas. Neste estudo, foram avaliadas um total de 343 amostras, provenientes de 18 populações de cajueiros de 6 países da CPLP: Brazil, Guiné-Bissau, Timor-Leste, Cabo Verde, São Tomé e Príncipe e Moçambique. Cada população é constituída por 15-20 indivíduos. No Brazil, centro de origem do cajueiro, incluíram-se duas populações (BR1 e BR2) para além duas populações de espécies de Anacardium (A. giganteum e A. humile) que funcionaram como outgroups. Em Timor-Leste, foram incluídas 4 populações (ETK, ETTR, ETSU e ETV), e uma população da Indonésia (IND) foi introduzida para avaliar a relação com Timor-Leste. No continente africano, foram avaliadas duas populações de Moçambique (MZB e MZD), três de Cabo Verde provenientes de ilhas diferentes (Fogo-CVF, São Nicolau- CFSN e Santiago-CVST), cinco populações da Guiné-Bissau (as quais incluem regiões continentais de Quinara- GBQ, Biombo- GBBIO, Gabu- GBGAB e do arquipélago dos Bijagós, Ilha da Formosa-GBFORM, e Ilha de Bolama- GBBOL) e uma população de São Tomé e Príncipe (STMO). A partir de 16 SSRs previamente desenvolvidos para o cajueiro, a genotipagem foram realizadas através de um processo de otimização e avaliação dos genótipos por análise de fragmentos. Dos 16 marcadores inicialmente selecionados, 2 loci (mAoR12 e mAoR33) foram eliminados após a deteção de alelos nulos. As análises de diversidade genética r toda a determinação da estrutura populacional foi subsequentemente efetuada com os restantes 14 marcadores moleculares. Na avaliação da aplicação dos 14 marcadores nas amostras testadas, todos apresentaram valores de Índice de polimorfismo (PIC) superiores a 0,50, com uma média de 0,73, indicando que são marcadores polimórficos e altamente informativos para a análise da estrutura populacional e índices de diversidade genética.

As populações com maior valor de diversidade genética estão presentes em Timor-Leste e Guiné-Bissau, sendo (ETK) no distrito de Manatuto (Timor-Leste) e na Região de Gabu (Guiné-Bissau) 4.29, a população com menor diversidade genética foi encontrada em Viqueque 2.36 (Timor-Leste) e em São Tomé e príncipe 2.57, valor medio de alelos.

Ao nível de estrutura populacional, foram realizadas 3 abordagens: i) com base em matrizes de distâncias genéticas (DC^{INA}) através da representação por dendrogramas de UPGMA e NJ, ii) Análise da variância molecular (AMOVA) e iii) análises bayesiana (STRUCTURE) e multivariada (DAPC) para avaliar a partilha de fluxo genético a nível populacional. Quanto aos resultados das matrizes de distância DC^{INA} através de UPGMA e NJ, observaram-se 4 agrupamentos principais: um que agrupa os outgroups (A. giganteum e A. humile), um grupo que compreende as populações de Timor-Leste com a Indonésia, outro com as populações da Guiné-Bissau e São Tomé e Príncipe, e o maior grupo que inclui as populações provenientes do Brasil, Moçambique e Cabo Verde.

Na análise da estrutura populacional o primeiro agrupamento constituído por Timor-Leste e Indonésia, pode-se especular que a população de Baucau poderia ser sugerida como o local de introdução do cajueiro em Timor-Leste através da Indonésia. O segundo cluster inclui populações da Guiné-Bissau e de São Tomé e Príncipe, destacando-se assim um fluxo genético entre os dois antigos países coloniais portugueses. O terceiro agrupamento inclui populações de Cabo Verde com Moçambique e Brasil, e está também correlacionado com o fluxo genético. A região de Quinara (GBQ) na Guiné-Bissau é uma região continental que é um território de passagem entre as ilhas Formosa e Bolama do arquipélago de Bijagós e o restante território continental, e onde se observa uma grande diversidade diferente a nível da castanha e maçãs o que poderia explicar um pouco o seu complexo fluxo genético observado nas árvores UPGMA e nas análises bayesianas.

A Análise da Variância Molecular (AMOVA) foi utilizada como medida da variabilidade da diversidade genética ao nível das populações de cajueiro, e os resultados revelaram que existe uma quantidade considerável de variação entre indivíduos para ambos os tipos de agrupamentos (todas as populações, incluindo outgroups; e apenas entre populações de cajueiro). As análises bayesianas permitiram determinar a existência de uma diversidade intraespecífica complexa do cajueiro a nível pantropical, mostrando o pouco fluxo genético entre as espécies de Anacardium (usados como outgroups) e as populações de cajueiro amostradas. De acordo com os resultados do STRUCTURE e do DAPC, existe uma diversidade genética significativa para uma cultura agrícola introduzida e uma baixa partilha de fluxo genético entre as regiões tropicais estudadas. A presença do fluxo genético entre as populações do centro de origem do cajueiro, Brasil, e as populações de Cabo Verde e Moçambique é um dado a salientar, a qual pode ser explicada pelo extenso intercâmbio histórico entre os dois continentes, iniciado no século XVI durante o período colombiano, quando os portugueses transportariam várias culturas agrícolas para o continente africano.

Também na analise de $K = 5$, que é o segundo K ótimo, revelou se que em STRUCTURE com o conjunto de dados de 20 populações, juntamente com os registos históricos, sublinharam a importância das rotas comerciais presididas por Portugal na proliferação de várias plantas económicas significativas desde o século XVI, como parte do período de intercâmbio colombiano, onde se destaca o papel do arquipélago de Cabo Verde nestes comércios para a aclimatação de espécies de culturas tropicais antes da agricultura na Península Ibérica e no continente africano para fazer face às despesas coloniais. Para além do fluxo genético nestes três países, observámos que na Ilha da Formosa na Guiné-Bissau, o agrupamento genético com Cabo Verde, Moçambique e Brasil é partilhado, o que poderia ser indicativo de uma potencial localização de populações de cajueiro aquando da introdução de cultura na Guiné-Bissau. Da região do Sudeste Asiático (Timor-Leste e Indonésia), observou-se uma diversidade genética específica para esta região tropical, que poderia indicar uma potencial rota diferente de introdução do cajueiro.

Os resultados deste trabalho sobre a distribuição da diversidade genética do cajueiro proporcionam uma breve visão do mecanismo de introdução da cultura na região de Timor-Leste e no continente africano,

onde a presença de uma diversidade significativa e complexa sugere numerosos eventos de introdução nos países ao longo da costa do continente africano. A fim de compreender a complexa variação intra-específica presente em Timor-Leste e no continente africano, um estudo futuro deverá incluir populações da Índia e de Angola. Este estudo é a primeira abordagem abrangente da diversidade intraespecífica do cajueiro utilizando uma técnica molecular, e enfatiza a necessidade de iniciativas nacionais para a conservação do germoplasma. Os resultados obtidos contribuem para o desenvolvimento de estratégias de gestão e conservação dos recursos genéticos do cajueiro a nível pantropical, utilizando os países da CPLP como estudo de caso, dado o papel preponderante dos portugueses na expansão das culturas agrícolas no século XVI e a elevada importância do cajueiro como cultura de rendimento na maioria dos países estudados.

Palavras-chave: cajueiro, microssatélites, diversidade genética, estrutura populacional, recursos genéticos.

ABSTRACT

The cashew tree (*Anacardium occidentale* L.) is a tropical plant from Anacardiaceae family, that has acquired a high economic importance as a cash crop in several tropical countries. It is assumed that in the middle of the sixteenth century the Portuguese brought the cashew tree from Brazil (center of origin) to India, and later expanded to Southeast Asia, and introduced in Africa at the same time in Mozambique. The socio-economic importance of the cashew tree in tropical regions is recognized, and in recent decades, the demand for cashew has increased at the global market. Despite the recognized value that this specie represents worldwide, few studies have addressed molecular markers to assess the distribution and structuring of cashew genetic diversity. In this work, microsatellites were used to study the intraspecific diversity of cashew across different tropical regions, from Brazil (South America) to Guinea-Bissau and Mozambique (Africa) and East Timor, Indonesia (Asia), using CPLP countries as a case study, which also cover cashew pantropical distribution. Despite its huge commercial value, few studies have been pushed forward towards the molecular diversity assessment of cashew genetic resources, which highlights the importance of the study conducted. As a result of extensive fieldwork in Guinea-Bissau and East Timor, complemented with samples collected in Brazil, Cabo Verde, Mozambique, and São Tomé and Principe, a set of cashew specific microsatellites or simple sequence repeats (SSRs) were selected for the characterization of the intra-specific diversity and population structuring of cashew.

Overall, about 343 individuals (i.e. 309 of *A. occidentale*; 16 of *A. giganteum* and 18 *A. humile*) corresponding to 20 different populations were screened and genotyped with 16 SSRs. mAoR12 and mAoR33 loci were discarded due to presence of null alleles in almost all populations. The remaining 14 markers were found to be polymorphic and further genetic diversity and population structuring analysis were conducted. The Analysis of Molecular Variance (AMOVA) was performed to quantify the genetic variability between and within populations, revealing that most of the genetic diversity lies within individuals, reflecting the heterozygous nature of cashew. In terms of diversity by population, the populations with the greatest diversity were found in Guinea-Bissau and Timor-Leste. The population structure can be observed using the DC^{INA} distance matrices in which the presence of 4 main clusters was observed. Individual-based clustering methods using a Bayesian approach (STRUCTURE) and a multivariate analysis by DAPC allowed to assess the population structuring, thus highlighting that genetic diversity scattering does follows a geographical trend under a continental distribution. Overall, our data reports the first comprehensive study on cashew intraspecific diversity using a continental approach thus highlighting the need to perform conservation programs focused on a country standpoint. This is especially important considering that the result of a limited gene flow across countries were observed, which may be due to local selection of accessions within each country.

This work demonstrates a national and continental signature, highlighting that cashew introduction was not a single event in history, and that there were several points of entry into Africa. The findings obtained in this work could be a baseline for the assessment of the diversity of cashew genetic resources at a pantropical scale, as one of the most economically important cash crops in tropical regions, still largely understudied.

Keywords: cashew, microsatellites, genetic diversity, populational structure; genetic resources.

APPENDIX: PUBLICATIONS & CONFERENCES

De acordo com o disposto no artigo n.º 19 do Regulamento de Estudos de Pós-Graduação da Universidade de Lisboa, Despacho n.º 2950/2015, publicado no Diário da República, 2.ª série — N.º 57 — 23 de março de 2015, foram incluídos nesta dissertação os resultados apresentados em:

PUBLICATIONS:

Guterres, L.; **Barnabé**, **J.**; Barros, A.; Charrua, A.B; Duarte, M.C.; Romeiras, M.M.; Monteiro, F. (**2022**). Population structure and genetic diversity of cashew (*Anacardium occidentale* L.) in East Timor. Under review at *PeerJ* [IF: 2.98, Q1 Agriculture and Biological Sciences].

Guterres L; Duarte MC; Catarino S; Roxo G; **Barnabé J**; Sebastiana M; Monteiro F, Romeiras MM. Diversity of legumes in the cashew agroforestry system in East Timor and their role for sustainable agriculture. Under review at *Frontiers in Sustainable Food Systems* [Q1 Agronomy and Crop Science].

PRESENTATIONS AT CONFERENCES:

Oral communications:

- Barai AS, **Barnabé J**, Correia Z, Ferreira MR, Diniz I, Costa G, Duarte MC, Romeiras MM, Batista D, Catarino L, Monteiro F. (**2022**). "Desafios atuais e perspetivas da agroecossistema do cajueiro na Guiné-Bissau." 2ª Encontro de Jovens Investigadores da CPLP sobre África, 25-27 May 2022, ISEG, Lisboa, Portugal.

- Monteiro F, Vidigal P, **Barnabé J**, Bahia MJ, Barai AS, Duarte MC, Catarino L, Romeiras MM (**2021**). "Avaliação das paisagens agrícolas na África Subsaariana no contexto da resiliência às alterações climáticas: o caso do cajueiro na Guiné-Bissau." XXII Encontro da Rede de Estudos Ambientais de Países de Língua Portuguesa (REALP), 29 November-2 December 2021, Universidade de Cabo Verde (Uni-CV), Cidade da Praia, Ilha de Santiago, Cabo Verde.

- Barai AS, **Barnabé J**, Correia Z, Ferreira MR, Diniz I, Costa G, Duarte MC, Romeiras MM, Batista D, Catarino L, Monteiro F. (**2021**). "Fighting cashew tree enemies in Guinea-Bissau: a collaborative project to support the control of diseases, pests and weeds." 7th Annual Meeting Frontiers in cE3c, 30 September-1 October 2021. cE3c, Faculty of Sciences, University of Lisbon, Lisbon, Portugal.

- **Barnabé J**, Guterres L, Charrua AB, Pena AR, Baldé A, Catarino L, Batista D, Romeiras MM, Monteiro F. (**2021**). "Molecular diversity assessment of a pantropical cash crop." 17th May 2021, Centre for Ecology, Evolution and Environmental Changes - cE3c, Lisbon, Portugal.

- Monteiro F, **Barnabé J**, Guterres L, Charrua AB, Pena AR, Baldé A, Catarino L, Batista D, Romeiras MM. (**2020**). "Molecular assessment of cashew diversity unravels distinctive differentiation routes in CPLP countries." 6th Annual Meeting Frontiers in cE3c, 1-2- October 2020. cE3c, Faculty of Sciences, University of Lisbon (Lisbon, Portugal).

Posters:

- Guterres L, **Barnabé J**, Barros AB, Charrua AB, Duarte MC, Romeiras MM, Monteiro F. (**2022**). "First look into the genetic diversity of cashew (*Anacardium occidentale* L.) in East Timor." Encontro Ciência 2021. 1st July 2021, Lisbon, Portugal.

- **Barnabé J**, Guterres L, Barros A, Charrua A, Duarte MC, Romeiras MM, Monteiro F. (**2022**). "Molecular diversity assessment of a pantropical cash crop (*Anacardium occidentale* L.)." Encontro Ciência 2022, 16-18th May 2022, Lisbon, Portugal.

- Barai A, Guterres L, **Barnabé J**, Correia Z, Ferreira MR, Diniz I, Costa G, Duarte MC, Romeiras MM, Batista D, Catarino L, Monteiro F. (**2022**). "Phytosanitary assessment of cashew (*Anacardium occidentale* L.): learning from Guinea-Bissau (West-Africa) and East Timor (Southern Asia)." Encontro Ciência 2022. 16-18th May 2022, Lisbon, Portugal.

- Bahia MJ, **Barnabé J**, Barai J, Diniz I, Batistas D, Sebastiana M, Monteiro F. (**2022**). **"**Molecular characterization of *Fusarium* sp. in cashew (*Anacardium occidentale* L.) from Guinea-Bissau." Encontro Ciência 2022. 16-18th May 2022, Lisbon, Portugal.

ONLINE RESOURCES:

- Guterres, L; **Barnabé, J**; Monteiro, F. Genotypic Matrix for Studying the Cashew Population Structure and Genetic Diversity in East Timor. *Figshare Dataset* 2022, doi:https://doi.org/10.6084/m9.figshare.19119041.v3.

- Barros, A; **Barnabé, J**; Guterres, Lara; Monteiro, F. Script for Performing DAPC Analysis Applied to the Study of Population Structure and Genetic Diversity in Cashew from East Timor. *Figshare Software*. 2022, https://doi.org/10.6084/m9.figshare.19117889.v3.

OTHER RELEVANT ACTIVITIES:
Participação nas atividades desenvolvidas no âmbito da Noite Europeia dos Investigadores 2021 (24 de setembro de 2021).

This research was funded by Fundação para a Ciência e Tecnologia (FCT) through the project "**GenoCash-** ***Phenotype to genotype assessment of cashew genetic resources for sustainable production in Guinea-Bissau (West Africa)***" PTDC/ASP-AGR/0760/2020.

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

GenoCash

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

LIST OF SUPPLEMENTARY DATA

ABBREVIATIONS AND ACRONYMS

AFLP- Amplified fragment length polymorphism
AG- *Anacardium giganteum*
AH- *Anacardium humile*
AMOVA- Analysis of Molecular Variance
BIC- Bayesian information criterion
BR- Brazil
CLUMPP- Cluster Matching and Permutation Program
CPLP- Comunidade dos Países de Língua Portuguesa
cSSR- cashew simple sequence repeat
CVF- Cabo Verde, Ilha do Fogo
CVSN- Cabo Verde, Ilha de São Nicolau
CVST- Cabo Verde, Ilha de Santiago
DAPC- Discriminant analysis of principal components
DC- Cavalli–-Sforza and Edwards genetic distance
DC^{INA}- Cavalli–-Sforza and Edwards genetic distance with INA (i.e., including null alleles) method
DNA- DeoxyriboNucleic Acid
EM- Expectation- maximization
ENA- Excluding Null Alleles
FAO- Food and Agriculture Organization of the United Nations
FCT- Fundação para a Ciência e a Tecnologia
GBBIO- Guinea-Bissau, Biombo
GBBOL- Guinea-Bissau, Bolama
GBFORM- Guinea-Bissau, Formosa
GBGAB- Guinea-Bissau, Gabú
GBQ- Guinea-Bissau Quinara
H_E- Expected Heterozigoty
H_O- Expected Heterozigoty
HWE- Hardy-Weinberg Equilibrium
IND- Indonesia
ISBN- International Standard Book Number
ISSR- Inter Simple Sequence Repeats
MCMC- Markov chain Monte Carlo
MSE- Mean squared error
MZB- Momzabique, Beira
MZD- Mozambique, Dondo
NA- not applicable
NGS- Next generation sequencing
NJ- Neighbor-joining
PCA- Principal Component Analysis
PGR- Plant genetic resource
PIC- Polymorphism Information Content
RAPD- Random Amplified Polymorphic
RFLP- Restriction fragment length polymorphism
RNA RiboNucleic Acid
SNP- Single-nucleotide Polymorphism
SSRs- Simple Sequence Repeats

STOM- São Tomé and Príncipe
ETK- East Timor, Kribas
ETSU- East Timor, Suai
ETTR- East Timor, Triloca
ETV East Timor, Viqueque
UPGMA- Unweighted Pair Group Method with Arithmetic Mean
USA- United States of America
USDA- United States Department of Agriculture

1. INTRODUCTION

1.1. Cashew as a cash crop commodity

1.1.1. Anacardiaceae family

The Anacardiaceae family holds about 600 described species and is natively distributed across Latin America. Cashew (*Anacardium occidentale* L.), mango (*Mangifera indica* L.) and pistachio (*Pistacia vera* L.) are the 3 main agricultural products of the family Anacardiaceae; pistachio is a Mediterranean crop [1], while mango and cashew are from tropical origins.

The *Anacardium* genus includes ca. 44 species from trees to shrubs and subshrubs, being the most significant genus of the Anacardiaceae family at economical level [2]. From this genus, *Anacardium occidentale* L. (**Figure 1**), commonly designated as cashew, is an evergreen tree with its center of origin in the northeastern part of Brazil [3]. Though, it is currently distributed across the globe due to its introduction for commercial purposes in several tropical areas, including Mozambique, Tanzania, Kenya, and Guinea-Bissau [4], where the cashew nut is a highly marketable product. Besides the cashew nut, the apple (i.e., pseudo fruit/swollen hypocarp) (**Figure 1**) is both consumed locally and used to produce "*Feni*," an alcoholic beverage. In South America, especially Brazil, and African cashew-producing countries, the juice from the apple is marketed widely as a popular drink. Moreover, the cashew nutshell liquid is used for industrial purposes and has medicinal value.

Figure 1. Illustration of the cashew plant representing the apple (fleshy false fruit) with nut (fruit), as well as leaves and inflorescences (Source: Flore d'Amérique).

The cashew tree can grow up to 20 meters but normally attains between 10-15 m, depending on the soil and environmental conditions. It usually starts to flourish by the third year, and reaches maximum output by the 8th year [2,4]. Cashew is a very rustic tree, thus developing very well without specific cares if the environmental conditions are favorable, namely average yearly temperatures between 17 and 38 ºC are optimal. Cashew processing has been a key economic activity for many tropical nations and

provides a range of foods and industrial products, cashew nut (**Figure 2A**) being the most valued and marketable one [5].

Some of the other *Anacardium* species have economic potential but are currently underutilized. While several species are mainly used in construction (*A. excelsum* (Bertero ex Kunth) Skeels) and ornamental purposes (*A. spruceanum* Benth. ex Engl), only two other species, *A. giganteum* W.Hancock ex Engl. (**Figure 2B**) and *A. humile* A. St.-Hil. (**Figure 2C**), have edible hypocarps and seeds, which are very appreciated by local people [2]. However, these species have not been cultivated and/or explored, *A. occidentale* being the only species of the genus that has been exploited crop, cultivated widely at the pantropical scale; the remaining species are only found in South America [6–8].

Figure 2. Apples and nuts from *Anacardium* trees. The nut (fruit) is attached to the bottom of the red/orange hypocarp, the apple (pseudofruit) from cashew – *Anacardium occidentale* (**A**), *A. giganteum* (**B**) and *A. humile* (**C**).

Nowadays, *Anacardium occidentale* is grown in the tropical regions of the Old and New Worlds (**Figure 3**). The *Anacardium* genus is endemic to tropical America and has two diversification hotspots, the Central Amazon and Brazil's Planalto. Four species occur near Manaus (*A. occidentale*, *A. humile*, *A. nanum* A.St.-Hil., and *A. corymbosum* Barb.Rodr.), three of which occupy the same habitat. The cashew distribution follows five distribution patterns [2] and the geographical limits of its cultivation are latitudes and longitudes of 27°N and 28°S, respectively ([9]; **Figure 3**).

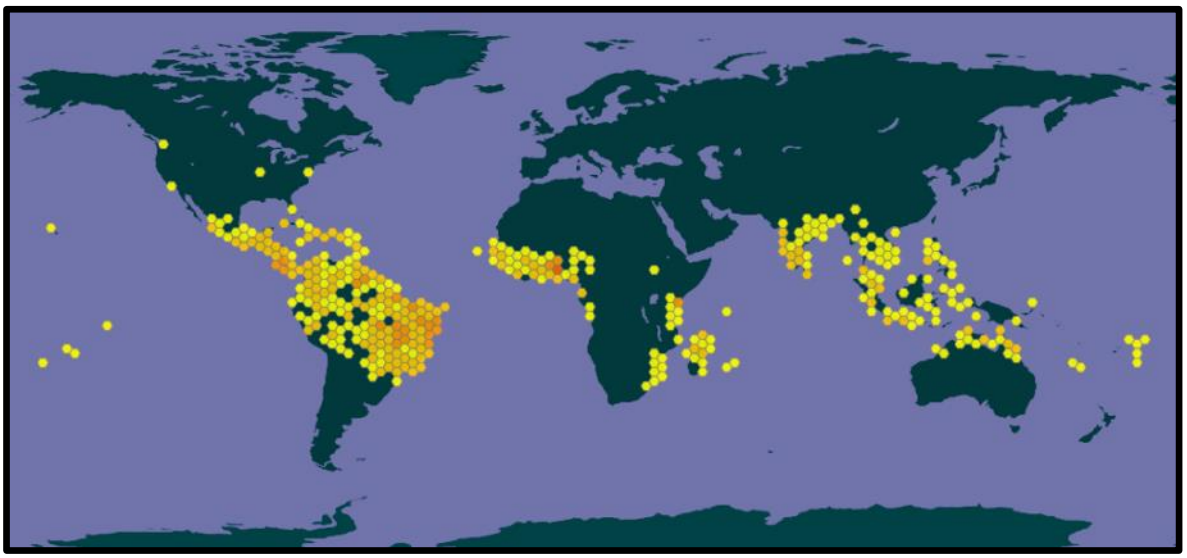

Figure 3. Worldwide distribution and occurrence trends of cashew (*Anacardium occidentale* L.). The highest predominance is observed in Southeast Asia, South America, and Western and Eastern Africa. Source: GBIF. Accessed on 23 November 2021.

Anacardium genus has the following five distribution patterns:

1. The Andes separate *A. excelsum* from its congeners, both taxonomically and geographically. The Andes' uplift was most likely the driving force for *A. excelsum*'s early separation from the rest of the genus.
2. *A. giganteum* and *A. spruceanum* are found in Amazonia and Guyana, respectively.
3. *A. occidentale*, the most common species in the genus, presents disjunct populations in Brazil's Planalto, eastern Brazil's Restingas, the Amazon basin's savannas, and Colombia and Venezuela's Ilanos. It should be noted that this species' native range is masked by its extensive cultivation in both the Old and New Worlds.
4. *A. humile*, *A. nanum*, and *A. corymbosum* are three closely related species found only in central Brazil's Planalto.
5. *A. corymbosum*, found only in south-central Mato Grosso (Brazil), is an allospecies of *A. nanum* and *A. fruticosum* J.D.Mitch. & S.A.Mori., the latter with a restricted distribution to Guyana's upper Mazaruni River basin. It is linked to *A. parvifolium* Ducke, which grows in the Amazon.

Johnson [10] studied the issues regarding the genesis and range of the genus, suggesting it originated in the Restinga (low vegetation found in the sandy soil along the coast of eastern and northeastern Brazil). Cultivated and wild populations of cashew species from eastern Brazil share chartaceous leaf blades and long petioles. *Anacardium* is native and occasionally a dominant feature of the Cerrados (savannah-like vegetation) on Central and Amazonian Brazil [2,5,11,12]. The Cerrado populations of *A. occidentale* differ from the Restinga populations by having undulated, thickly coriaceous leaves with short and stout petioles. The hypocarps (cashew apples) of Cerrado trees are usually smaller and sometimes have a more acidic flavor than those of the Restinga. The natural distribution of *A. occidentale* extends from northern South America to south of São Paulo, Brazil. It is probably not native to Central America, the West Indies, or South America west of the Andes. It is believed that *A. occidentale* originally evolved in the Cerrados of Central Brazil and later colonized the more recent Restinga zones of the coast [2,13]. Very likely, most cashew cultivated widely was brought from Restinga to other tropical regions, where similar morphological features are observed, at both leaf and apple/nut levels.

1.1.2. Cashew as a cash crop commodity across tropical regions

A cash crop is an agricultural crop that is grown to obtain profit. The word is used to differentiate between marketed and subsistence crops. Cashew is considered a cash crop because the main goal of its production is to obtain the cashew nut for commercial purposes, with few by-products beyond its fruit. During the Columbian Exchange in the 16th century, a strong transfer of crops between South America and Africa allowed the establishment of cultivation fields with exotic crops, which became the economic foundations of several tropical countries [11]. In the 16th century, the Portuguese imported cashew from Brazil onto the African continent and to India as a rustic tree to fight afforestation and alleviate soil erosion, and only in the 1950s did it start to be seen as an agriculture commodity [4,11]. In India, cashew was spread primarily to control soil erosion in coastal areas [2], and only after the nuts were seen as food and its apple could be used to make a good wine.

According to FAO [14], the total annual production of cashew nut is close to 1.0 million tons, with Vietnam (30%), Nigeria (21%) and India (17%) as the three main producers. On a global scale, cashew exports account for major monetary revenues in several tropical countries (**Figure 4A**). In 2018, West Africa region alone accounted for half of the global cashew supply [15], with Guinea-Bissau (GB) ranked as 3rd in West Africa and as 6th worldwide in the list of top cashew-producing countries. The

demand for cashew has risen over the past several decades along with the growth of global market demands, which have been complemented with an increase in its harvest area (**Figure 4B**).

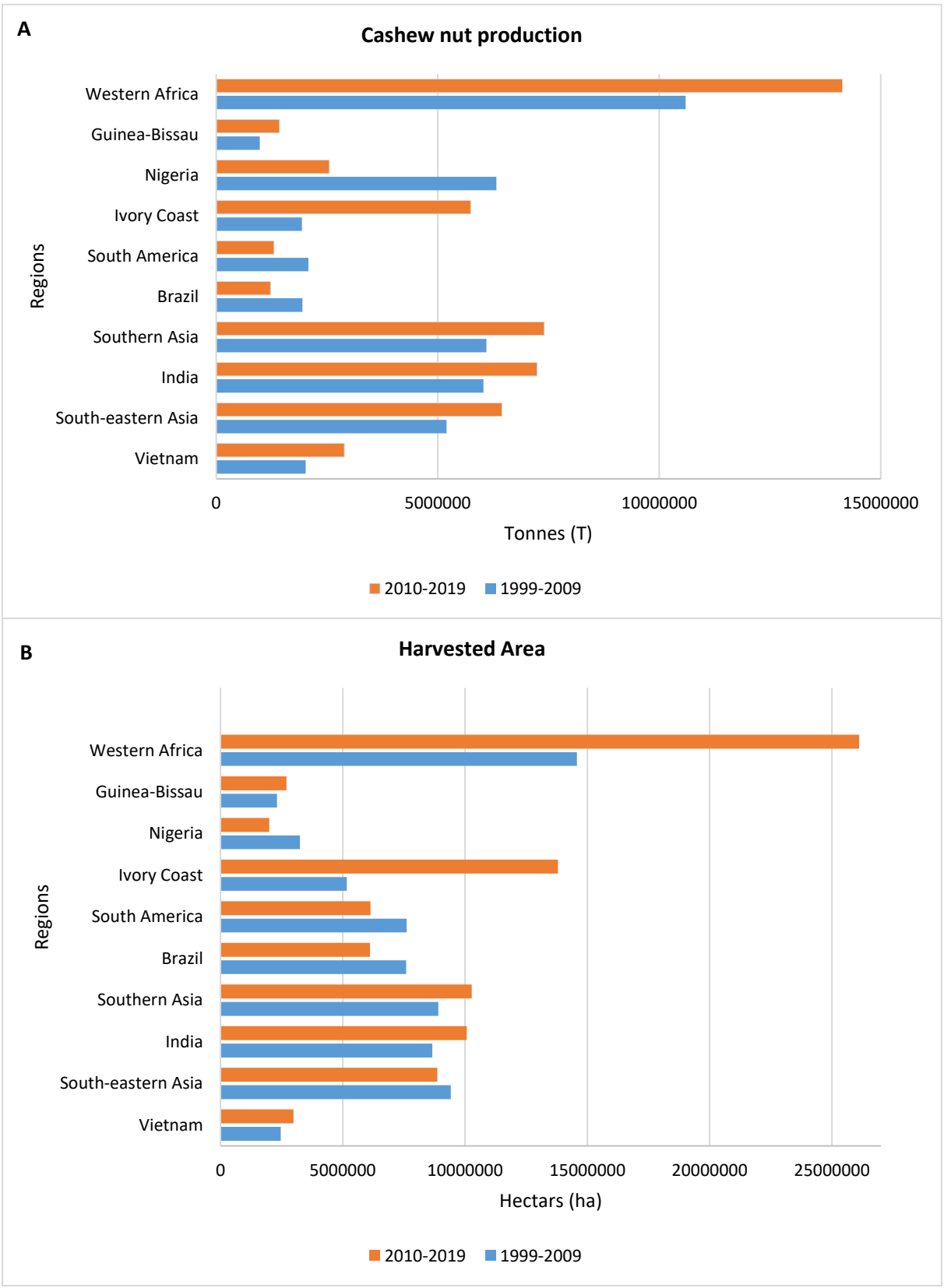

Figure 4. World cashew production in tons (A) and harvest area (B) over the periods 1999-2009 and 2010-2019, from key exporting countries in Asia, South America, and Africa. Source: FAOSTAT 2022 (accessed on January 2022).

As stated above, over the past several decades, the demand for cashew has increased with the rising of global market demands. The consumption trend was particularly high in recent years, when the global market consumed 87% more cashews than 10 years ago [16]. The attractiveness of cashew is mainly due to changing food habits towards healthy diets. As a rich source of plant-based protein, minerals and low-fat contents, cashew is one of the top three most produced and valued tree nuts worldwide, together with almonds, and walnuts (**Figure 5**; [17]).

The cashew nut is a highly valued product composed of 21% protein and 22% carbohydrate, with the right combination of amino acids, minerals, and vitamins, and 47% fat, from which 82% are unsaturated fatty acids, not contributing to cholesterol enhancement and also involved in balancing or decreasing overall blood cholesterol [8,18,19]. Therefore, nutritionally, it can be considered a healthy snack, turning cashew nuts into a valuable agriculture product [20].

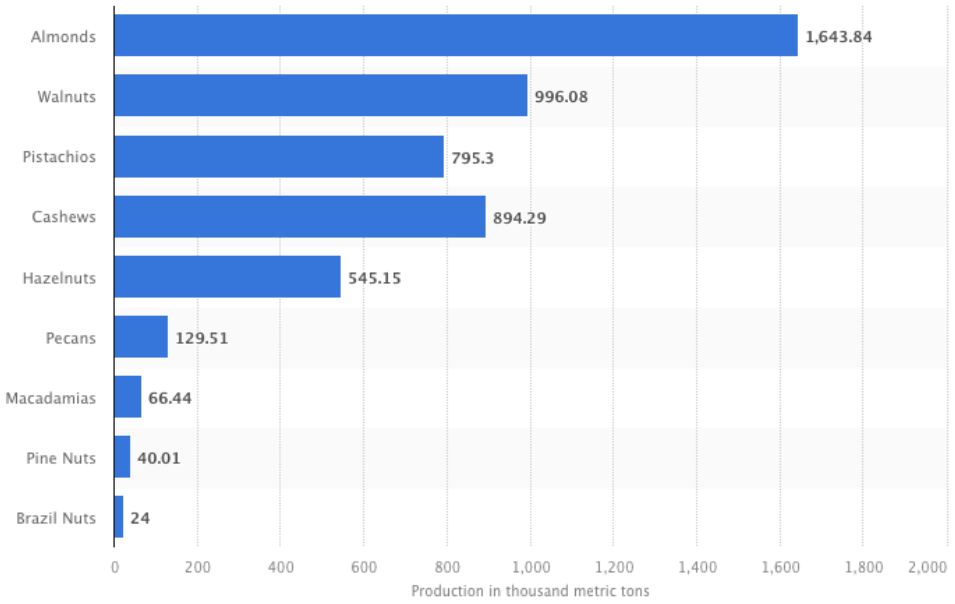

Figure 5. Tree nuts produced globally, by kind, in 2020/2021 (in 1,000 metric tons). Source: Statista 2022.

1.1.3. Agriculture development in CPLP (Community of Portuguese Language Countries)

In the last decades, demand for cashew has risen in worldwide markets (USDA, 2015). Several Portuguese Language countries from the CPLP (*Comunidade dos Países de Língua Portuguesa*), such as Guinea-Bissau, Mozambique and Brazil are top cashew producers. CPLP is an international body and political union of nations spanning four continents, where Portuguese is an official language [13]: Portugal, Brazil, Angola, Mozambique, Cape Verde, Guinea-Bissau, São Tomé and Principe, East Timor, and Equatorial Guinea [2] (**Figure 6**).

Considering the distribution of CPLP across the three continents (South America, Africa and Asia), and the cashew production and economic importance in many of its countries, it can be considered an intesresting case study to assess cashew genetic diversity. Namely, from its center of origin (Brazil, South America), to African regions where it is a top agriculture commodity (Guinea-Bissau in West Africa and Mozambique in East Africa) and to Southeast Asia (East Timor) where this crop was established as a cash crop. Besides all these considerations, the countries' distribution covers the main tropical regions where cashew is cultivated, thus allowing to prospect a pantropical scale of cashew diversity.

Figure 6. Member countries of CPLP. The countries where the samples were taken for the present study are indicated by a circle. 1- Brazil, 2- Cabo Verde, 3-Portugal, 4- Guinea-Bissau, 5- São Tomé and Principe, 6- Angola, 7- Equatorial-Guinea, 8- Mozambique, 9- East Timor. Source: CPLP (accessed on January 2022).

At **Figure 7**, cashew nuts production (tons) and harvest area (ha) in CPLP countries are shown (FAOSTAT, 2022). No data are available for Cabo Verde, São Tomé and Principe and East Timor, although cashew is produced in these countries, highlighting the lack of documented cashew exports to international markets, which may be due to local selling to processing countries such as India and Vietnam. In its center of origin (Brazil), the production of cashew nuts is one of the highest among the CPLP countries, with a mean value of 150,922 tons (**Figure 7A**), decreasing since 2011 to values of 138,754 tons in 2019, which may be attributed to a reduction of the harvested area, as shown in **Figure 7B**. When compared with other CPLP's countries, Guinea-Bissau was the second highest producer until 2011, when it became the main producer (**Figure 7A**) which was accompanied by the increase harvest area (**Figure 7B**). Mozambique is the third CPLP highest cashew producer, and in both Guinea-Bissau and Mozambique this crop is an important commodity for government and people revenues. Angola looks to be a significant cashew producer but has not yet established itself as a top contributor.

However, when comparing the ratio of production to harvested area, we do not observe the expected rise in output. Therefore, it is necessary to make improvements through selection of high-yield cashew varieties.

As an illustration of cashew's importance, Guinea-Bissau alone accounts for almost 90% of all exports being a key revenue for both government and rural communities. It was estimated that about 40% of the typical food consumption depended on cashew sales, which highlights the importance of cashew cultivation for family economies [4]. The development of a cashew's chain value will deeply influence the producing-countries, by promoting an added-value of cashew products (nuts and apple) besides providing a more stable income for both smallholder farmers and government.

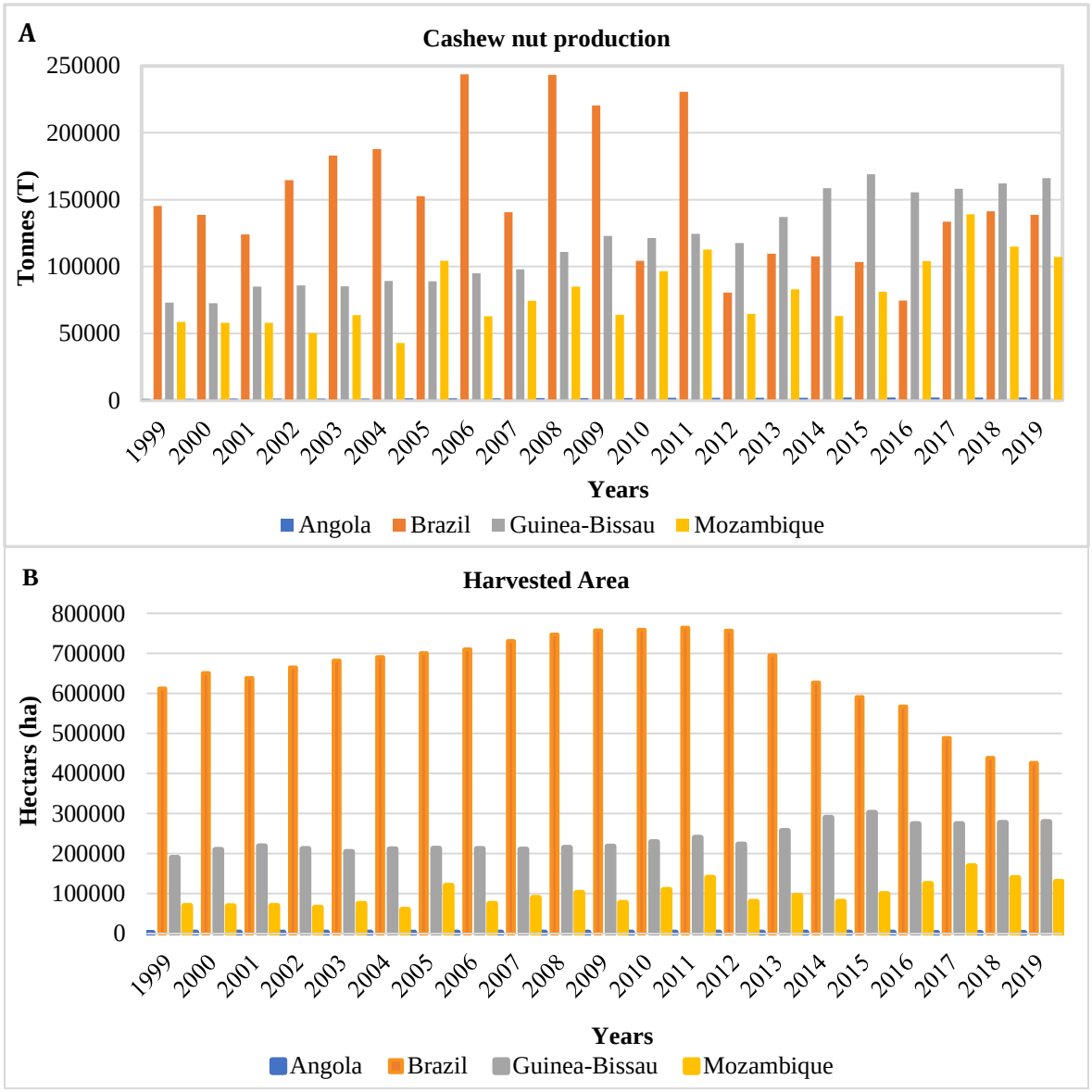

Figure 7. Cashew nut production from 1999 to 2019 in four CPLP countries, in tons (T, **A**) and harvest area in hectares (ha, **B**). Source: FAOSTAT (accessed on 16th February 2022).

1.2. Microsatellites, a tool to study cashew genetic diversity

Molecular markers are useful tools to measure genetic diversity among agricultural species (**Table 1**). Examples of those are RAPDs, ISSRs, AFLP, microsatellites and more recent markers such as SCAR, which are now more widely genotyped, often in combination with morphological descriptors [21,22]. Unlike standard analyses of PCR, the **RAPD**s (Rando Amplified Polymorphism DNAs) does not require knowledge of the target organism's DNA sequence; Restriction Fragment Length Polymorphism (**RFLP**) is a form of polymorphism resulting from the DNA sequence variation detected by restrictive enzymes. RFLPs are employed as markers in genetic maps; **SCAR** (Sequence Characterized Amplified Regions) markers are developed with a pair of longer primers (usually the extended sequence of a **RAPD**). Amplified Fragment Length Polymorphism (**AFLP**) is a PCR method that employs selective amplification to produce and analyze single fingerprints for genomes of relevance to a sub-ensemble of digested DNA fragments. The benefit is that the target genome and its high repeatability and sensitiveness to identify polymorphism at a DNA sequence level does not require any previous knowledge. **AFLP**, widely used for plant and microbial trials, has been utilized to determine genetic diversity in

several applications, for example, to estimate population-level phylogenies within species or between closely related species; but it has a high cost, and often clusters at the centromeres and telomeres. The main advantage of **ISSRs** (Inter Simple Sequence Repeats) is that no sequence data for primer construction are needed; however, ISSRs are randomly distributed throughout the genome like **RAPDs**, and thus can have reproducibility problems. Molecular markers advantages and disadvantages are described in **Table 1.**

Table 1. Types of molecular markers and their advantages and disadvantages.

Molecular Markers	Advantages	Disadvantages
RFLP	Co-dominants, no prior sequence information needed.	Require high quality and quantity of DNA.
RAPD	Less quantity of DNA needed. Low cost.	Contamination of DNA. Need a highly standard purification protocol.
SCAR	Simpler patterns than RAPD.	Sequence information needed.
AFLP	Large number of amplicons, no prior sequence information or probe.	Dominant markers. Hight cost. Often cluster at centromeres and telomeres.
SSR	Require small quantity of genomic DNA. Highly polymorphic. Easy interpretation.	High cost. Quite complex discovery procedure.
ISSR	No prior sequence information needed. Variation within unique regions of the genome may be found at several loci simultaneously.	Dominant markers. Complex detection system.

SSRs (Simple Sequence Repeats), also known as microsatellites or short tandem repeats, are short 2–8 nucleotide motifs, repeated in tandem for a few to hundreds of times at many independent loci in eukaryotic genomes [23]. Microsatellites or simple sequence repeats (SSRs) are among the most effective markers among the many molecular markers used, primarily because they are co-dominant single locus markers [24]. SSRs are extremely informative molecular markers, inheritable and co-dominant, which are frequently used as markers in the characterization of genetic diversity, conservation, and traceability of germplasm in plants.

SSRs are polymorphic loci that derive one to six base pairs in length from the repetition of short sequence motifs. Microsatellites have several features which make them ideal for the analysis of plant genomes:

1. Co-dominance, which makes it possible to study plant hybrids with commercial varieties.
2. In general, the amplified fragments are small (between 100 and 500 base pairs), resulting in positive amplifications of PCR even in highly degraded DNA.
3. Due to the polyploid nature of the genome of several significant crop species, as stated in the section on plant diversity, a limited number of selected SSRs can provide a high capacity for intraspecific discrimination.

Next-Generation Sequencing (NGS) offers a new opportunity to examine uncharacterized systems for sequencing. Targeted Genotyping by Sequencing (Hi-SNPseq) combines multifunctional PCR with high performance sequences to execute multiplex PCR enhancements in a single tube using site-specific primers [25,26]. This methodology retrieves thousands of SNPs (Single Nucleotide Polymorphisms), based on a high-throughput sequencing, and by comparing several samples one can determine a SNP associated to an agronomical trait and/or to a geographical region. It is frequently applied to genetic research in population and disease-related genes [3,26]; however, its bioinformatic analysis is still laborious, and to assess a prompt genetic diversity analysis in a germplasm collection is still far from applicable at country level.

1.3. Genetic diversity in cashew

Despite its enormous economic worth, few research has been conducted to evaluate the molecular diversity of cashew genetic resources. No worldwide study has been done so far, with only country-specific approaches, namely in Ivory Coast [27], Brazil [21] and India [6] germplasm accessions. Mneney *et al.* [28] studied cultivars from different regions of the world and 20 elite Tanzanian varieties using RAPDs, and found a high degree of genetic similarity among accessions from Tanzania, India, and Mozambique, whereas accessions from Brazil were most distinct from the other regions of the world. The work made by Chipojola *et al.* [29] using 4 Malawian populations and morphological markers revealed a high similarity among populations, with a similar result reported in a genetic diversity study conducted in 187 germplasm accessions from Nigeria [30], which indicate the existence of a narrow genetic base of cashew germplasm.

Given the importance of cashew as an agriculture commodity in several tropical regions, a concerted research effort is necessary to determine agrobiodiversity hotspots associated to a geographical region. Considering the lack of a global genetic diversity studies including its native origin (Brazil) and other cashew producing regions, the present study is pioneer and is based on a comprehensive sampling that covers cashew main distribution trends in 3 different continents, from South America to African Continent and Southeast Asia. Rather than using all cashew producing countries in each continent, this study focused on the cashew producing CPLP countries, which will allow to hint on the introduction history of cashew in the African and Southeast Asia continents, based on the role of the Portuguese traders from the 16th century as part of the Columbia Exchange period.

The main objective of this work is to characterize the genetic diversity of cashew from diverse tropical regions, using CPLP countries as a case study (**Figure 8**). By including the native origin of cashew, Brazil, together with the closest *Anacardium* species from Brazil, *A. giganteum* and *A. humile*, it will be possible to determine the presence of any genetic flow between cultivated cashews and the wild sister species. Besides, by including several populations from each CPLP country, it will be possible to determine hotspots of agrobiodiversity in cashew, namely at continental and/or at country levels for future germplasm efforts. Moreover, in this work it is tentatively performed the association between genetic diversity patterns with the historical cultivation of cashew.

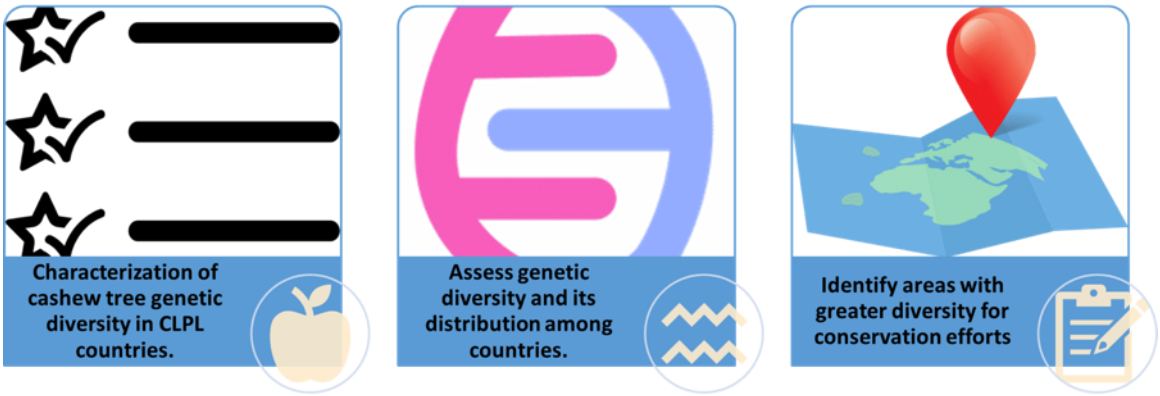

Figure 8. Main objectives of the molecular diversity assessment of cashew.

2. MATERIAL AND METHODS

2.1. Sampling

Cashew populations were sampled from seven different CPLP countries: Brazil (N=2), Cabo Verde (N=3), Guinea-Bissau (N=5), São Tomé and Principe (N=1), East Timor (N=4), Indonesia (N=1), and Mozambique (N=2) (**Figure 9, Table S1**). Thus, 18 cashew populations were sampled from Portuguese-speaking countries where cashew has acquired significant economic importance. In each population, 16-20 individuals were sampled and preserved in silica gel until further processing (Table 2). Also, two populations were sampled from the closest *Anacardium* species originated in Brazil, *A. giganteum* and *A. humile*, which will be used to level the intraspecific diversity of cashew populations, considering its high phylogenetic relationship as wild relatives of the crop (*A. occidentale*). All leaf samples are stored at the Instituto Superior de Agronomia/University of Lisbon and are available upon author request.

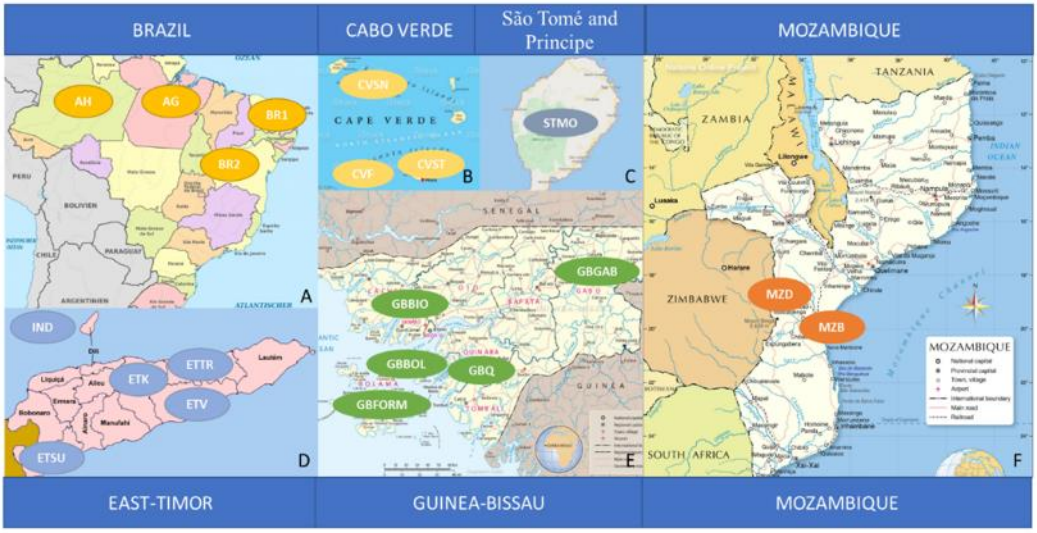

Figure 9. Sampled populations: acronyms by country, district, and location.

2.2. DNA extraction

Individual leaves collected from each cashew population were used to obtain genomic DNA (gDNA), extracted with the InnuPREP Plant DNA Kit (Analytik Jena, Germany), following the manufacturer's instructions with minor modifications. About 100 mg of each leaf collected in the field were grinded briefly with a mortar and a pestle in liquid nitrogen, and then 1 cm^2 of roughly grinded leaves were used for further gDNA extraction, by adding 400 µL of OPT lysis solution and grinding the biological material with an Eppendorf-pestle in a 1.5 mL tube. Afterwards, an initial incubation at 65 °C for 1 h was performed, followed by the addition of 100 µL of Precipitation Buffer and a 5-min incubation at room temperature; the supernatant was recovered by centrifugation at maximum speed 13,000 x g for 5 mins. The supernatant was then transferred to a Pre-Filter Receiver and centrifuged at the same speed for 1 min. Subsequently, 4 µL of RNAse A solution (100 mg/mL) was added and samples were incubated for 30 mins at 37 °C. After RNAse treatment, 200 µL of SBS binding solution was added and then centrifuged at 11,000 x g for 2 mins. The recovered supernatant then underwent two washing steps with 650 µL of MS washing solution and centrifugations at 11,000 x g for 1min. The gDNA was eluted in 40

μL of AE buffer, left to incubate at room temperature for 15 mins and recovered by centrifugation at 11,000 x g for 1 min. DNA purity and concentration were measured at 260/280 ηm and 260/230 ηm using a spectrophotometer (NanoDrop-1000, Thermo Scientific), while DNA integrity was verified by agarose gel electrophoresis at 0.8% in 1x TAE running Buffer (Merck) for 30 mins at 90 Volts and then visualized in a GelDoc XR image system (BioRad, USA).

2.3. Microsatellite genotyping

A set of 16 cashew-specific microsatellite (SSRs, Simple Sequence Repeats) markers already available [22] were selected to for screen the genetic diversity of the populations under study, following three major criteria: (i) markers with a Polymorphism Information Content (PIC) value higher than 0.5, a threshold reference value to be considered an informative marker, (ii) markers with high allelic diversity, and (iii) dinucleotide repeats markers, to enable a clearer interpretation upon microsatellite genotyping, and thus avoiding genotyping errors.

Before multiplexing, each SSR marker was validated in single-plex polymerase chain reactions (PCR) using a three-primer PCR approach [31], to assess reaction reproducibility and/or presence of PCR artifacts upon fragment analysis [12]. Each SSR was PCR amplified in a 25 μL volume reaction following cycling conditions previously described in Croxford *et al*. [22], using the HotStar Taq DNA Polymerase kit (QIAGEN, Germany), as per manufacturer's instructions. Afterwards, SSRs amplified fragments were run in an ABI 3130XL sequencer (Applied Biosystems) with the internal size standard GS500 LIZ (Applied Biosystems, USA) at STAB VIDA company (Costa da Caparica, Portugal), while allele calling was performed in GeneMapper v 3.7 (Applied Biosystems, USA). A thorough markers selection to ensure the success of co-amplification loci was assessed by using the Multiplex Manager software v1.2 [32], which allowed building four SSRs panels assembled in 4-plex PCR reactions (Multiplex A, B, C, and D; **Table 2**), using four universal forward fluorescently labelled primers following Culley *et al*. [33]. To increase genotyping accuracy, a "PIG-tail" sequence was added at the 5'end of each of the reverse primer [34]. PCR multiplex amplifications were carried out using the QIAGEN Multiplex PCR kit (QIAGEN, Germany), following the manufacturer' s protocol, in a total volume of 25 μL with 50–100 ηg gDNA and 2.5 ρmol of each primer Forward and Reverse and 0.15 ρmol of the tailed fluorescently labeled primers (D1–D4). Reactions were done in 96 well-plates and on each plate one sample was repeated per run, thus working as positive control for allele scoring. Negative PCR controls were included. Initially, a hot-start step at 95 °C for 15 min was performed, followed by a touchdown cycling protocol adapted from Croxford *et al*. [22] as follows: 5 cycles of denaturation at 95 °C for 45 s, primer annealing at 68 °C for 5 min with −2 °C/cycle; a sequence extension at 72 °C for 1 min; 5 cycles of denaturation at 95 °C for 45 s, primers annealing (58 °C for Multiplexes A, C and D and 60 °C for Multiplex B) for 2 min with −2 °C/cycle and an extension step for 1 min at 72 °C; 27 cycles at 95 °C for 45 s, 47 °C for 75 s, and 72 °C for 1 min; followed by a final extension step at 72 °C for 10 min. Then, multiplex PCR products were run in an ABI 3130XL sequencer for fragment analysis as described earlier, and SSR allele sizes were aligned with the internal size standard, further scored using the binning function in GeneMapper v3.7 (Applied Biosystems, USA). To improve the SSR marker data quality, allele assignments were checked manually, and ambiguous results were set as "missing data."

Table 2. Loci used to screen the 20 *Anacardium* sp. populations sampled. Primers sequences, multiplexing scheme, and amplicon size range/ expected size (bp) are provided. Multiplex arrangement is identified.

Locus	Repeat motif	Primers (5'-3')	Tailed Primer	Size range (Expected Size)	Multiplex
mAoR6	$(AT)_5(GT)_{12}$	F: CAAAACTAGCCGGAATCTAGC	D2	118–186 (143)	**A**
		R: GTTTCTTCCCCATCAAACCCTTATGAC			
mAoR17	$(GA)_{24}$	F: GCAATGTGCAGACATGGTTC	D1	122-184 (124)	
		R: GTTTCTTGGTTTCGCATGGAAGAAGAG			
mAoR7	$(AT)_2(GT)_5AT(GT)_5$	F: AACCTTCACTCCTCTGAAGC	D4	158-198 (178)	
		R: GTTTCTTGTGAATCCAAAGCGTGTG			
mAoR48	$(GAA)_6(GA)_3$	F: CAGCGAGTGGCTTACGAAAT	D3	130-186 (177)	
		R: GTTTCTTGACCATGGGCTTGATACGTC			
mAoR3	$(AC)_{12}(AAAAT)_2$	F: CAGAACCGTCACTCCACTCC	D4	140-282 (231)	**B**
		R: GTTTCTTATCCAGACGAAGAAGCGATG			
mAoR42	$(CAT)_9TAT(CTT)_7$	F: ACTGTCACGTCAATGGCATC	D2	160-232 (204)	
		R: GTTTCTTGCGAAGGTCAAAGAGCAGTC			
mAoR52	$(GT)_{16}(TA)_2$	F: GCTATGACCCTTGGGAACTC	D1	142-244 (202)	
		R: GTTTCTTGTGACACAACCAAAACCACA			
mAoR11	$(AT)_3(AC)_{16}$	F: ATCCAACAGCCACAATCCTC	D3	142-248 (234)	
		R: GTTTCTTCTTACAGCCCCAAACTCTCG			
mAoR2	$(CA)_{10}(TA)_6$	F: GGCCATGGGAAACAACAA	D3	172-322 (366)	**C**
		R: GTTTCTTGGAAGGGCATTATGGGTAAG			
mAoR33	$(CT)_{18}(AT)_{19}$	F: CATCCTTTTGCCAATTAAAAACA	D4	322-404 (354)	
		R: GTTTCTTCACGTGTATTGTGCTCACTCG			
mAoR35	$(AG)_{14}$	F: TCTTTCGTTCCAATGCTCCTC	D2	142-180 (165)	
		R: GTTTCTTCATGTGACAGTTCGGCTGTT			
mAoR47	$(TAAA)_2(TA)_7(AAT)$	F: AAGAGCTGCGACCAATGTTT	D1	166-272 (161)	
		R: GTTTCTTCTTCTTGAACTTGACAC-TTCATCCA			
mAoR12	$(AC)_{12}ATAC(AT)_4$	F: CACCAAGATTGTGCTCCTG	D2	322-362 (324)	**D**
		R: GTTTCTTAAACTACGTCCGGTCACACA			
mAoR16	$(GT)_8(TA))_{17}(GT)_3$	F: GGAGAAAGCAGTGGAGTTGC	D1	245-335 (256)	
		R: GTTTCTTCAAGTGAGTCCTCTCACTCTCA			
mAoR29	$(TG)_{10}$	F: GGAGAAGAAAAGTTAGGTTTGAC	D3	164-364 (316)	
		R: GTTTCTTCGTCTTCTTCCACATGCTTC			
mAoR41	$(ACC)_7(AC)_3$	F: GCTTAGCCGGCACGATATTA	D4	162-177 (151)	
		R: GTTTCTTAGCTCACCTCGTTTCGTTTC			

Following Culley *et al*,. [33], D1 (6-FAM): M13 (-21), 5'-TGTAAAACGACGGCCAGT-3'; D2 (NED,): T7term, 5'-CTAGTTATTGCTCAGCGGT-3'; D3 (VIC): M13modA, 5'-TAGGAGTGCAGCAAGCAT-3'; D4 (PET): M13modB, 5'-CACTGCTTAGAGCGATGC-3'. Underlined sequence at each reverse primer (GTTTCTT) identifies the "PIG-tail" sequence.

2.4. Genetic diversity analyses

Genotyping errors were assessed using MICRO-CHECKER v2.2.3 [35], and estimation of null alleles frequency was done with the EM algorithm of Dempster *et al*. [36] as implemented in *FreeNA*

(http://www.montpellier.inra.fr/URLB/). These values were computed as described by Chapuis and Estoup [37], with 10,000 bootstrap iterations, alternatively using and not using the Excluding Null Alleles (ENA) method, after assessment of null allele frequencies. Polymorphic Information Content (PIC) and genetic diversity indices were calculated with the Microsatellite Toolkit v.3.1.1 [38] and GenALEx 6.5 [39], respectively. These included the total allele number and mean alleles per locus (Na), private alleles, inbreeding coefficient (fixation index, *F*), observed (H_O) and expected (H_E) heterozygosity. Deviations from Hardy–Weinberg equilibrium (HWE) were assessed for each locus-population combination and linkage disequilibrium (LD) to determine the extent of distortion from independent segregation of loci using GenePop v4.5 [40]. Statistical significance for both HWE and LD was tested by running a Monte Carlo Markov Chain (MCMC) consisting of 10,000 iterations each, and *p*-values were corrected for multiple comparisons [p <0.00018, (0.05/280] by applying a sequential Bonferroni correction [41].

2.5. Population structuring

Population structure was addressed using three approaches: (i) estimating relations among populations using genetic distances; (ii) hierarchical genetic analysis by AMOVA; and (iii) individual-based clustering with a Bayesian (STRUCTURE) and a multivariate (DAPC, Discriminant Analysis of Principal Components) analyses.

2.5.1. Estimating relations among populations using genetic distances

Distances relationships among populations were estimated with Cavalli-Sforza and Edward's Chord genetic distances (*DC*, [42]) using the INA method computed in *FreeNA* (DC^{INA}), and Nei's *D* distance [43] calculated in GenALEx 6.5 [39]. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbor-Joining (NJ) trees were produced using the package *ape* v3.4. [44] of R v4.1.0 [45], based on 10,000 bootstraps values, assessed by *aboot* function from *poppr* v2.1.0. Package [46]. Trees were further edited in FigTree v1.4.2 [47]. Distances relationship among populations were determined by two separated approaches: first, using all the populations and, second, by excluding the two *Anacardium* species used as outgroups (*A. giganteum* and *A. humile*).

2.5.2. Hierarchical genetic analysis (AMOVA)

An Analysis of Molecular Variance (AMOVA, [48–50]) was done with ARLEQUIN v3.5.1.3 [51] to assess the hierarchical distribution of genetic variation of the analysed populations. Significance was assessed after 1000 permutations. Two three-levels AMOVAs were pursued: one using all the twenty populations as groups, and the second narrowed to the cashew populations by excluding outgroups (AG and AH populations). In each AMOVA, the total variance was partitioned into components to account for differences between defined groups (V_a), differences among populations within those groups (V_b), and differences among individuals within populations (V_c). Variance components (V_a, V_b, and V_c) were used to calculate the fixation indices (*F*-statistics; F_{CT}, F_{SC}, F_{ST}) according to Weir and Cockerham (1984) [48].

2.5.3. Individual based clustering

Identification of genetically distinct clusters was done with two different methodologies: a Bayesian clustering analysis using STRUCTURE [52] and a multivariate analysis method, the Discri-

minant Analysis of Principal Components (DAPC, [53]). These two different individual cluster assignment approaches were followed since, while STRUCTURE uses allele frequency and LD information from the dataset directly, DAPC is a multivariate method which attempts to summarize the genetic differentiation between groups, while overlooking within-group variation and not relying on a particular population genetics model free of HWE assumptions [53].

Overall, individual-based clustering analyses were performed of two datasets: one including all populations (N=20), and another focused-on cashew populations (N=18) to evaluate intraspecific genetic variability. The Bayesian model-based clustering algorithm implemented in STRUCTURE v.2.3.4 was used to identify genetic clusters under a model assuming admixture and correlated allele frequencies without using population information. In the first approach, including all twenty populations, analyses were set for a burn-in period length to 100,000 followed by 1,000,000 MCMC iterations with *K*-values set from 1 to 21 with 10 runs computed for each *K*. For the second approach, restricted to cashew populations (N=18), the same settings were followed by configuring *K*-values from 1 to 19 with 10 runs in each *K*. StructureHarvester v0.6.94 [54] was then used to calculate Δ*K ad hoc* statistics from Evanno *et al*. [55] to estimate the most likely *K*-value, which is based on the rate of change of the estimated likelihood between successive *K*-values. CLUMPP v1.1.2 [56] was used to average replicate runs for the selected *K*-value, and to account for problems with multimodality and label switching between iterations of STRUCTURE runs. CLUMPP results were then plotted with DISTRUCT v1.1 [56].

DAPC was implemented in R [45] using *adegenet* v1.3.1 package [53] and the dataset relative frequency of the alleles, since presence/absence data may not be fully informative and thus overlook relevant patterns in allele frequency. The function *find.clusters* was used to find the ideal *K*-value, based on the computation of Bayesian Information Criterion (BIC) scores, maintaining default parameters and retaining all principal components (PCs). Cross validation using the *xvalDapc* function was pursued to determine the optimal number of PCs to retain in the Discriminant Analysis (DA).

3. RESULTS

3.1. SSRs genotyping and statistics

All 16 SSRs were tested in single-plex reactions at the estimated optimal annealing temperature, and only after this initial quality assessment, were SSRs markers grouped into 4-plex reactions (**Table 2**). For the SSRs loci screened, allele profiles were clear and easy to score. No errors in the genotypic data matrix were detected, indicating the absence of potential errors associated with stuttering bands or large allele dropout in the SSRs used. In 76 of the 320 locus-comparisons, the frequencies of null alleles were higher than 0.20 (data not shown) in the markers mAoR12 and mAoR33, in almost all populations. Thus, these two loci were removed for subsequent analyses, and the overall data were analyzed with 14 SSRs (**Table 3**). Deviations from the Hardy–Weinberg Equilibrium (HWE) were observed in most loci except mAoR3, with 191 locus-population combinations statistically significant ($p < 0.05$), while after sequential Bonferroni correction only three loci (mAoR3, mAoR17 and mAoR35) displayed significant deviations, matching 108 of the 280 locus-population combinations (**Table S2**). All 14 loci were in linkage equilibrium after Bonferroni correction, thus being non-correlated, and alleles independently segregated and inherited (data not shown). Negative fixation index (*F*) estimates were observed in two loci, mAoR17 (-0.155), mAoR47 (-0.038) and mAoR16 (-0.024) (**Table 3**), which can reflect more heterozygotes than expected or other population structure complexities.

Table 3. Marker's diversity measurements. The level of genetic diversity of each SSR marker was described with the parameters number of alleles, Polymorphic Information Content (PIC), gene diversity (expected heterozygosity, H_e), observed heterozygosity (H_o) and inbreeding/fixation coefficient (*F*). A total of 343 individual samples were analyzed.

Locus	Allele number	PIC	H_e	H_o	*F*
mAoR48	17	0.66	0.71	0.48	0.037
mAoR6	24	0.80	0.82	0.53	0.142
mAoR17	26	0.86	0.87	0.77	-0.155
mAoR7	19	0.75	0.78	0.54	0.012
mAoR11	22	0.81	0.83	0.46	0.145
mAoR3	36	0.89	0.90	0.57	0.186
mAoR42	19	0.69	0.73	0.54	0.021
mAoR52	21	0.81	0.82	0.53	0.084
mAoR2	9	0.55	0.60	0.27	0.328
mAoR35	12	0.71	0.73	0.40	0.347
mAoR47	13	0.69	0.72	0.61	-0.038
mAoR16	21	0.72	0.74	0.51	-0.024
mAoR29	15	0.84	0.86	0.32	0.385
mAoR41	11	0.67	0.69	0.22	0.517
Total	309	1	1	1	1
Mean	19	0.73	0.76	0.52	0.113

3.2. Genetic diversity estimates

Overall, a total of 309 alleles were detected in the 343 individuals analyzed (**Table 4**). All the screened loci were polymorphic. The total number of alleles per locus ranged from 9 (mAoR2) to 36 (mAoR3), with an average of 19.3 alleles per locus (**Table 4**). Overall, Polymorphic Information Content (PIC) values ranged from 0.55 (mAoR2) to 0.89 (mAoR3), with a mean value of 0.73 (**Table 4**). In our 14-loci dataset, the observed heterozygosity (H_o) varied from 0.22 (mAoR41) to 0.77 (mAoR17) with a mean of 0.52 (**Table 3**), and the expected heterozygosity (H_e) varied between 0.60 (mAoR2) and 0.90 (mAoR3) with a mean of 0.76. The Fixation Index *F* (also called the Inbreeding Coefficient) exhibits values from −1 to +1. Values close to zero are expected under random mating, while substantial positive values indicate inbreeding or undetected null alleles. Negative values denote excess of heterozygosity, due to negative assortative mating, or selection for heterozygotes. Overall, Positive *F*-values were observed across all loci except mAoR16, mAoR17, and mAoR47 (**Table 4**), thus revealing that they are at or near Hardy–Weinberg equilibrium, further supported by the lower observed heterozygosity values against the expected under HWE (**Table 4**). A population genetic diversity analysis (**Table 4**) showed that Guinea-Bissau presented 9.43 alleles on average, and that the lowest value was from São Tomé (2.57). H_e was 0.73 for Guinea-Bissau and 0.66 for the population BR1.

Table 4. Genetic diversity analysis by countries and populations. Data are provided by 14 SSRs, followed by a country analysis and, afterwards, a population-based genetic diversity assessment. Sample size (N). Countries: East Timor (4 populations), Indonesia (1 population), Brazil (2 populations), Mozambique (2 populations), Cabo Verde (3 populations), Guinea-Bissau (5 populations), São Tomé and Principe (1 population) and the outgroups (AG, AH). Populations: 20 different populations. Genetic diversity indices for each group are presented: expected heterozygosity (H_e) and observed heterozygosity (H_o), inbreeding/fixation coefficient (*F*), Private alleles, and mean alleles per locus (Na).

	Sample size	**Na**	**H_e**	**H_o**	**Private Alleles**	***F***
Countries						
East Timor	67	7.79	0.66	0.45	2.57	0.32
Indonesia	18	3.29	0.56	0.41	0.50	0.23
Brazil	33	4.57	0.66	0.58	0.57	0.10
Mozambique	34	5.36	0.71	0.49	0.93	0.29
Cabo Verde	47	4.71	0.65	0.57	0.86	0.11
Guinea-Bissau	90	9.43	0.73	0.43	3.79	0.40
São Tomé and Principe	20	2.57	0.34	0.29	0.29	0.16
AG	16	3.43	0.58	0.75	1.07	-0.31
AH	18	2.36	0.39	0.45	0.14	-0.19
Populations						
ETK	17	4.86	0.65	0.517	0.86	0.20
ETTR	16	4.71	0.57	0.463	0.57	0.17
ETSU	16	3.50	0.63	0.390	0.43	0.36
ETV	18	2.36	0.45	0.404	0.07	0.05
IND	18	3.29	0.54	0.411	0.43	0.23
BR1	16	4.07	0.66	0.613	0.21	0.06
BR2	17	3.57	0.60	0.553	0.21	0.09
MZB	17	4.14	0.65	0.445	0.43	0.31
MZD	17	3.50	0.63	0.540	0.21	0.13
CVSN	15	3.14	0.57	0.596	0.14	-0.08
CVF	16	3.07	0.53	0.612	0.00	-0.18
CVST	16	3.29	0.58	0.512	0.14	0.10
GBBOL	20	3.29	0.57	0.338	0.64	0.38
GBGAB	19	4.29	0.62	0.364	0.43	0.39
GBBIO	16	3.50	0.55	0.470	0.07	0.12
GBQ	17	3.07	0.54	0.297	1.36	0.47
GBFORM	18	3.79	0.61	0.704	0.50	-0.16
STOM	20	2.57	0.33	0.290	0.21	0.16

The lowest value of H_e was for the country São Tomé and Principe, with 0.34 and respectively population; the H_o was 0.74 for AG (*A. giganteum* population), 0.58 for Brazil and 0.70 for GBFORM (Guinea-Bissau), the lowest value of H_o being found for São Tomé and Principe (0.29). All the analyzed populations within each country presented a *F* positive varying from 0.10 to 0.40, except for the outgroup populations AG and AH (-0.31 and -0.19, respectively). The populations with the highest number of alleles were observed in Guinea-Bissau (9.43 Na) and East Timor (7.79), most precisely in the population of GBGAB, with an average of 4.29 alleles, followed by ETK with 4.86 alleles, respectively; São Tomé and Príncipe (STOM) presented the population with the smallest Na, 2.57 alleles, representing the lowest allelic diverse populations. By comparing the number of alleles of the SSRs in the screened populations, it can be determined which population harbors the highest allelic diversity. The fixation

index (*F*) was positive in almost every population indicating genetic stability and a high rate of inbreeding. The absence or existence of private alleles is important to determine, as it may allow to discriminate specific genetic populations. These alleles are very important because they can be used to create a unique genetic signature for each cashew population in each sampled geographic region. The countries that exhibited at least one private allele were East Timor and Guinea-Bissau, and the population of *A. giganteum* (2.57, 3.79 and 1.07, respectively).

3.3. Population structuring analyses

3.3.1. Estimating relations among populations using genetic distances

UPGMA and NJ trees were built using DC^{INA} genetic distance across accessions screened for all populations, including outgroups and only for the *A. occidentale* populations. Regarding the distance's method, a similar structure was observed with both Nei's *D* (data not shown) and DC^{INA} matrices. This indicates a reliable topology, regardless of the different genetic distance's algorithms used, but DC^{INA} was chosen because it has the advantage that it corrects for null alleles, if there are any. As such, only UPGMA trees derived from DC^{INA} distances matrices are presented in **Figure 10**. In the UPGMA tree for the 20- populations dataset analysis, four clusters are depicted (**Figure 10A**): **(I)** with all the populations of East Timor and Indonesia; **(II)** with populations of Guinea-Bissau and São Tomé and Principe; **(III)** with Brazil, Mozambique and Cabo Verde populations, and the GBQ population from Guinea-Bissau, in a mixed cluster; and **(IV)** with the outgroup populations, AG and AH (*A. giganteum* and *A. humile*, respectively), completely separated from cashew populations.

While looking to the UPGMA generated for cashew populations alone (**Figure 10B**), four main clusters are observed: **(I)** with all populations from East Timor grouped with Indonesia, and BR1 from Ceará Region in Brazil; **(II)** all populations from Guinea-Bissau clustered with São Tomé and Principe (STOM); **(III)** with Mozambique populations, the Cabo Verde populations from Fogo (CVF) and Santiago (CVST) islands; and **(IV)** with a population from Brazil (BR2, from Brasília) and a population from Cabo Verde (São Nicolau island).

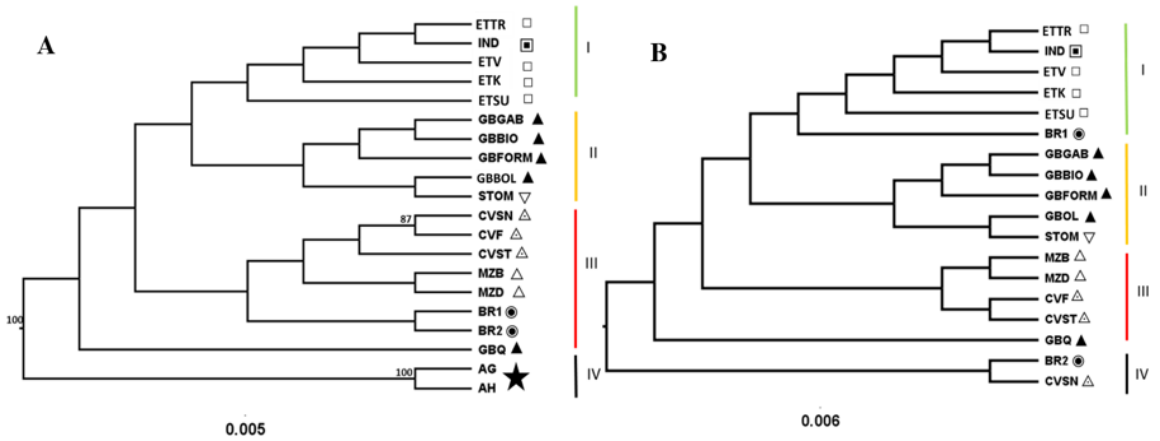

Figure 10. UPGMA trees generated from *FreeNA* using DC^{INA} matrix representing the populations from all 7 countries sampled in this study, with outgroups (**A**) and without outgroups (**B**). Clusters are represented by I, II, III and IV, with green, yellow, red and black colors, respectively. Legend: □ East Timor, ▣ Indonesia, ◉ Brazil, ▽ São Tomé and Principe, △ Cabo Verde, ▲ Guinea-Bissau, △ Mozambique, and the outgroups ★.

The NJ trees for both dataset analyses (**Figure S3**) showed a similar result, when looking at the outgroups we also have four clusters: the first (i) with some populations of Guinea-Bissau (GBGAB, GBBIO, and GBBOL) associated with the population of São Tomé and Principe (STOM), the second (ii) cluster with all populations of Cabo Verde (CVSN, CVF, and CVST) with Brazil populations (BR1 and BR2), Mozambique and one population from Guinea-Bissau (GBFORM); the third (iii) cluster with East Timor populations (ETTR, ETV, ETK, and ETSU) and the Indonesia population (IND), and lastly the fourth (iv) cluster with outgroup populations (AG and AH) and the GBQ population from Guinea-Bissau. The NJ tree analysis without any outgroup populations was grouped into four clusters, the first of which has Brazil populations (BR1 and BR2) with all Cabo Verde populations (CVSN, CVF, and CVST) joining Mozambique populations (MZB and MZD) and one Guinea-Bissau population (GBFORM), the second includes East Timor populations (ETTR, ETV, ETK, and ETSU) and the Indonesia population (IND); the third cluster (iii) comprises some Guinea-Bissau populations (GBGAB, GBBIO, and GBQ), while the fourth cluster includes GBBOL population from Bolama island (GB) coupled with the São Tomé and Principe population (STOM).

3.3.2. Hierarchical genetic analysis (AMOVA)

When grouping all populations, including the outgroups, AMOVA results showed that molecular variation was mainly found within individuals (64%), whereas variation among populations and among individuals explained 11% and 25 % of the total genetic variability, respectively (**Table 5**). Without outgroups, a similar scenario resulted for countries, with genetic variation being higher within individuals (64%), rather than among individuals (14%) or populations (22%).

Table 5. AMOVA results including fixation indices F_{CT}, F_{SC} and F_{ST}. The genetic differentiation among all populations (including outgroups) and countries (excluding outgroups) is defined as F_{CT}, among accessions within groups as F_{SC}, and within accessions as F_{ST}. $^{*}p < 0.001$.

Source of variation	df	Sum of Squares	Variance components	Variation (%)	Fixation indices
All populations					
among populations	19	851	V_a= 1.19	25.02	F_{CT}=0.144*
among individuals	323	1315	V_b= 0.51	10.80	F_{SC}=0.250*
within individuals	343	1045	V_c =3.05	64.18	F_{ST}=0.358*
Countries (excluding outgroups)					
among populations	17	660	V_a= 1.01	21.68	F_{CT}=0.179*
among individuals	291	1247	V_b= 0.65	14.03	F_{SC}=0.217*
within individuals	343	922	V_c= 2.98	64.29	F_{ST}=0.357*

3.3.3. Clustering analysis using bayesian and multivariate approaches

Two different Bayesian analysis approaches were done: 1) covering all populations including the outgroups (*A. giganteum* and *A. humile*) and 2) excluding outgroups, to uncover more in-depth individual clustering within *A. occidentale* populations.

In the first approach, all the 20 sampled populations were considered. STRUCTURE was run considering the highest range of clusters conceivable (K= 1–21). This analysis assigned K= 17 as the optimal number of groups based on the Evanno *et al*. [55] method, with K= 5 also displaying high ΔK-values (**Figure S2**). Analyzing the optimum K (K=17, **Figure 11A**), East Timor has a genetic admixture

between ETK (Manatuto) and ETTR (Baucau) – blue cluster –, followed by ETSU from Covalima with a unique cluster (purple). ETV has its own cluster (white), and so does Indonesia (dark green cluster); the population BR1 from Brazil clusters with CVSN and CVF from Cabo Verde, while Mozambique populations (MZB and MZD) share a cluster (pink), and CVST has its own cluster (light red). In Guinea-Bissau, GBBOL and GBQ populations have a unique cluster, GBGAB shares a cluster with GBBIO, while the island population GBFORM has its own cluster with some individuals shared with the continental populations of GBGAB and GBBIO; STOM from São Tomé and Principe has its own cluster (light pink), and the outgroup populations do not share any genetic diversity with the cashew populations. The second ideal K= 5 was also analyzed due to high ΔK-values (**Figure S2A**) obtained from the Bayesian analysis. In K= 5 (**Figure 11B**), all cashew populations are differentiated from the outgroups *A. giganteum* (AG) and *A. humile* (AH). Regarding cashew populations clustering, all East Timor and Indonesia populations are represented by a unique cluster (pink); Brazil populations (BR1 and BR2) share genetic flow with GBQ (Guinea-Bissau) and STOM (São Tomé and Principe) populations. Also, a shared clustered including Brazil, Mozambique and Cabo Verde populations is observed (blue cluster). A third cluster (in orange) mostly includes Guinea-Bissau and São Tomé and Principe (STOM) populations. The fourth cluster (grey color) is found for Mozambique and the GBGAB population from Guinea-Bissau (**Figure 11B**).

DAPC analysis was made without any *a priori* group assignment. Also considering the same approach as in the STRUCTURE analysis, analyses with and without outgroup populations were performed, to derive the appropriate number of genetic clusters. For the 20-populations dataset and DAPC analysis of K = 17, the cross validation used the *xvaldapc* function outcome to determine the number of PCA (i.e., alleles/loci) that contributed the most for the three-clustering assembly. For this K clustering, cross-validation analysis following the Occam's razor principle determined the retention of 20 PCA axes (97.50% of successful assignment with 4.72% of MSE), capturing 60% of cumulative variance, and with only 1 Discriminant Function to describe the 16 genetic clusters (data not shown). As such, alleles 184 (mAoR7), 203 (mAoR42), and 364 (mAoR29) are responsible for most of the genetic variation explaining the three genetic cluster assignments (**Figure S4**). A scatterplot allows an the overview of the 3 genetic groups clustering in both analyses; it should be noted that the outgroup populations are in a separate group (**Figure S5**), and when performing a DAPC membership probability plot as in STRUCTURE (**Figure 11C**), one can depict a similar clustering assignment as determined with K= 17 in STRUCTURE for the approach considering the outgroup populations: East Timor and Indonesia are part of a same cluster (dark), and so do Brazil populations (BR1 and BR2) (dark orange cluster), but some BR1 individuals share the cluster with East Timor and Indonesia. Mozambique populations display two clusters (light yellow, for MZB, and light purple for MZD), while all Cabo Verde populations (CVSN, CVF, and CVST) are grouped in a single and unique cluster (dark green). For Guinea-Bissau populations, the dominant cluster is represented by a light beige color, separated from GBBOL, which shares a unique cluster (blue) with GBQ (closest continental region) and GBFORM (island population), which have a mixture of green and grey colored clusters. São Tomé and Principe population (STOM) appears in its own cluster (dark brown). As expected, the outgroup populations (AG and AH) do not share any genetic diversity with cashew populations, being grouped in a separate cluster (dark purple). The second ideal K determined by STRUCTURE (K= 5) was also chosen in DAPC analysis (**Figure 11D**). The results show a cluster with East Timor, Brazil, Mozambique, and Cabo Verde populations grey), three Guinea-Bissau populations (GBBOL, GBGAB, GBBIO) are represented in one cluster (red), and the fourth (GBQ) is represented in another (blue cluster) and in another cluster (grey). São Tomé and Principe population shares the same cluster of Guinea-Bissau populations (red), and outgroup populations (AG and AH) represent a separate cluster (orange) (**Figure 11D**).

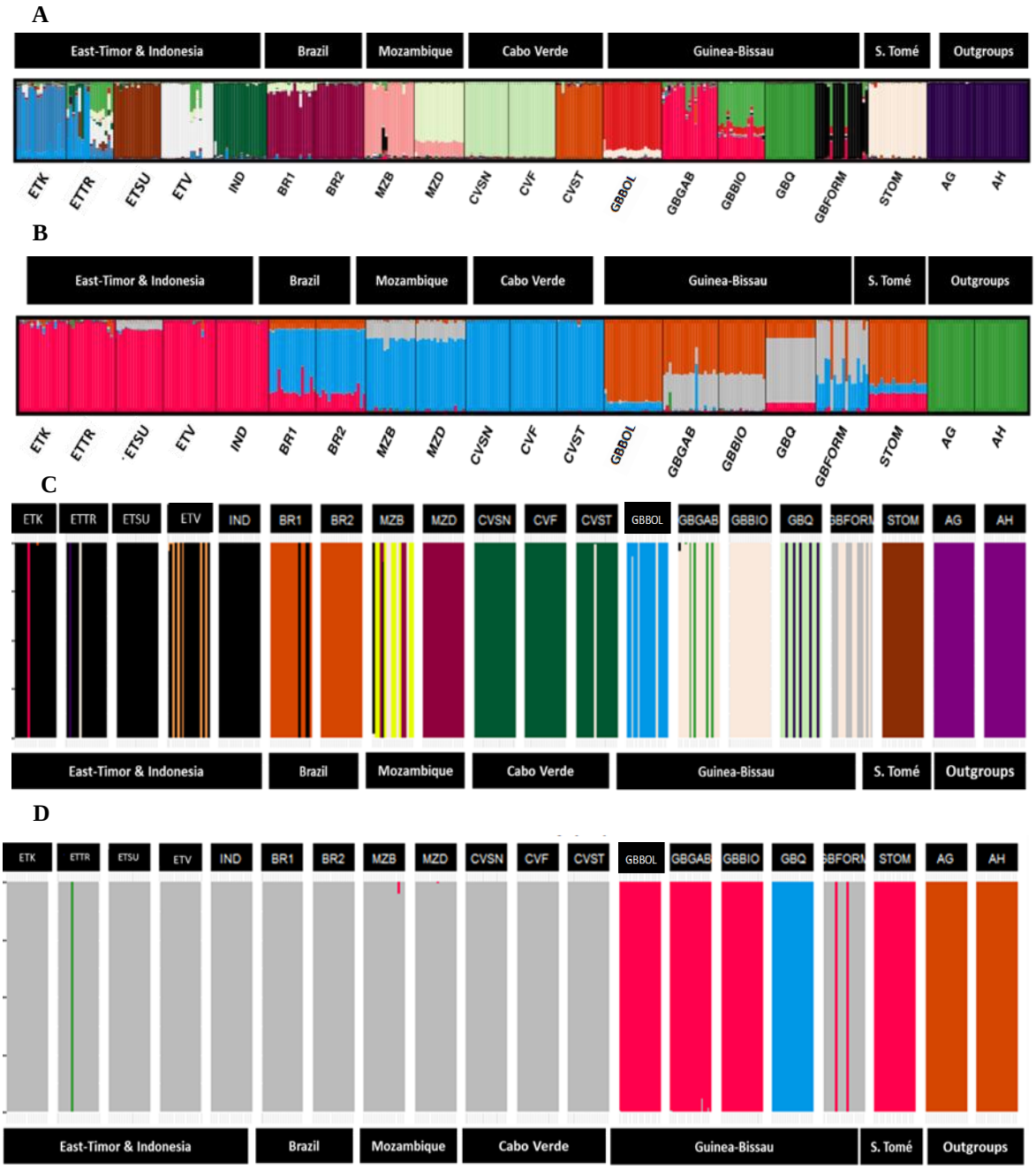

Figure 11. Clustering based on SSR data using STRUCTURE (*K* = 17, A; *K*= 5, B) and DAPC (*K* = 17, C; *K* = 5, D) analyses. The length of each section is proportional to the estimated ancestry value of the individual accession to each one of the K clusters for STRUCTURE and memberships probabilities for DAPC analysis. Each individual is represented by a vertical bar according to each K section. Labels on the x-axis indicate populations IDs. This analysis included all twenty populations, including the two outgroups.

For the second Bayesian analysis approach, only cashew populations were analyzed. STRUCTURE was run considering the highest range of clusters conceivable (*K*= 1–19) and, based on Evanno *et al*. [55] method, the optimal *K* was 16 (**Figure S2B**). The results (**Figure 12A**) revealed that East Timor populations are grouped into 4 different clusters with admixture among them, with Indonesia being in a single grey cluster. The populations from Brazil (BR1 and BR2) are grouped in a unique cluster, as for Mozambique populations In Cabo Verde, there are 2 major clusters, one including the

CVSN and CVF populations (in orange), and the other formed by CVST. In Guinea Bissau there are at least 5 major clusters. São Tomé and Principe population (STOM) appears in its own cluster.

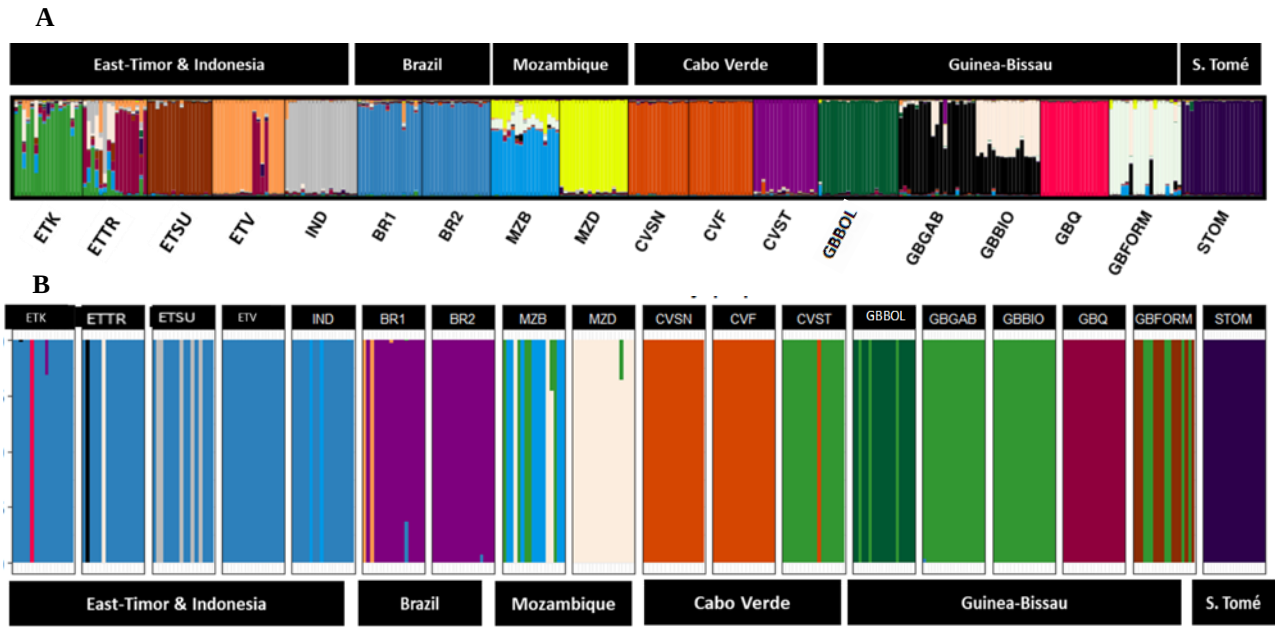

Figure 12. Clustering for *K*=16 based on SSR data from cashew populations (N=18) using STRUCTURE (**A**) and DAPC (**B**) analyses. The length of each section is proportional to the estimated ancestry value of the individual accession to each one of the *K* clusters for STRUCTURE and memberships probabilities for DAPC analysis. Each individual is represented by a vertical bar according to each *K* section. Labels on the x-axis indicate populations IDs. This analysis was carried out excluding outgroup populations.

Multivariate analysis through DAPC using the optimal *K* number from STRUCTURE (*K*= 16) was also performed (**Figure 12B**). For this *K* clustering, cross-validation analysis following the Occam's razor principle determined the retention of 20 PCA axes (97.50% of successful assignment with 5.72% of MSE), capturing 60% of cumulative variance, and with only 1 Discriminant Function for describing the 16 genetic clusters. (Data not shown). Following this cluster assignment, 16 different genetic backgrounds of 18 cashew populations were obtained, thus reinforcing a complex intraspecific variability, namely: one cluster groups mostly East Timor populations (blue); a specific Brazilian clustering (purple); Mozambique shows two clusters: MZB shares a flux with East Timor and Indonesia (blue cluster) while MZD has its own cluster (light beige); Cabo Verde populations are distributed into 2 different clusters, one with CVSN and CVF (orange) and another with CVST (light green); Guinea-Bissau populations GBGAB and GBBIO are grouped (light green cluster), GBBOL and GBQ have their own clusters (darker green and red-purple, respectively), and GBFORM is a combination of the red-purple and light green clusters; São Tomé and Principe corresponds to a single cluster (dark blue). Considering the overall pattern of genetic clustering and observed intraspecific variation, STRUCTURE and DAPC produced similar results, showing optimal clustering of populations, and highlighting a complex intraspecific variability of the screened cashew populations.

4. DISCUSSION

In this study, 343 individuals from 18 cashew (*Anacardium occidentale*) populations and two outgroups, *Anacardium* species (*A. giganteum* and *A. humile*) were selected from six Portuguese-speaking countries (Brazil, Cabo Verde, São Tomé and Principe, Guinea-Bissau, Mozambique, East Timor) and Indonesia (Southeast Asia). The outgroup species included in this study are native to the cashew center of origin (Brazil), thus allowing to determine the presence of genetic flow between wild and cultivated *Anacardium* species. By performing a comprehensive sampling of six cashew producing countries that belong to the Portuguese-speaking countries community (CPLP), a characterization of cashew genetic diversity and population structuring can be performed at a pantropical scale, thus covering crops center of origin (Brazil), several African countries where cashew has a significant economic importance and, East Timor rises as the most recent CPLP country that has engaged into cashew as a cash crop.

4.1. The significance of SSRs markers as genetic diversity indicators

Overall, the 16 microsatellite markers used in this study revealed to be applicable in all samples screened, with easy amplification and genotyping. Upon SSRs data analysis, mAoR12 and mAoR33 loci were excluded for subsequent analysis due to the detection of null alleles in almost all populations, despite being polymorphic and displaying high reproducibility in a multiplex PCR amplification. After SSRs quality assessment, 343 individuals from 20 different populations were genotyped with the remaining 14 loci (**Table 3**) and genetic diversity and population structuring analyses were conducted. All markers revealed to be polymorphic across the cashew populations analyzed and thus being informative markers for posterior diversity analysis. After excluding mAoR12 and mAoR33 loci, only 8 of the 280 locus-comparisons harbored null alleles with a frequency higher than 0.20. Overall, considering the above analysis, one can predict that the SSRs loci selected to screen the cashew populations under study are suitable for downstream genetic diversity analysis. PIC-values obtained in our study were high (average PIC= 0.73) which indicates their high informativeness. Our PIC values were higher when compared with a recent study using 21 cashew SSRs (cSSRs) to screen 23 cashew germplasm accessions which displayed a 0.33 average PIC-value for all loci used [25,57]. Deviations on PIC values depicted in our study might be due to differences in plant material sources compared to former reports, which may influence the number of alleles detected at each SSR locus. Private alleles were detected, contrasting with other SSRs studies in cashew [25] where no results were reported. The presence of private alleles allows to build a distinctive genetic profile associated to different geographic locations using frequencies and single alleles of each locus. This feature is extremely relevant to agrobiodiversity and crop studies such as in cashew, where nuts are exported and processed in countries other than the product's origin and the identification of a genetic profile may contribute to the valorization of a product with geographical provenance. In this context, one can depict the presence of private alleles in East Timor and Guinea-Bissau, which can be further explored to provide a genetic signature profile. For the remaining countries screened few private alleles were observed, and one can speculate that alleles are present in one or more cashew populations regardless of geographical origin thus representing significant shared allelic diversity. Our outcomes show a significant genetic diversity among the countries.

Population structuring revealed that genetic diversity appears to follow a geographic trend, with a well-defined cluster observed in populations from continental regions. This study provides useful information on genetic diversity hotspots, which can be used in future breeding efforts to improve genetics and characterize new cashew varieties. Moreover, the genetic diversity build-up obtained in our study

points out to cashew agrobiodiversity hotspots such as Gabú in Guinea-Bissau and Kribas (Manatuto) in East Timor.

4.2. Cashew diversity and population structuring in CPLP countries

The wide distribution of cashew in its primary center of diversity in Brazil has been attributed to water currents in which the mature fruit float in addition to the role played by fruit bats in seed dispersal [10]. In the current cashew distribution, its expansion has been attributed to anthropogenic efforts, rather than through natural means alone [6]. Considering this important premise, the present genetic diversity analysis results are discussed accordingly. Molecular studies in cashew diversity have been scarce, namely using informative markers as SSRs. In West Africa, most precisely in Nigeria previous reports [58] examined 10 microsatellite loci, but only mAoR2 and mAoR47 were polymorphic, thus revealing a considerable amount of redundancy within the Nigerian cashew germplasm. In Ivory Coast, a recent study on cashew genetic diversity with 18 SSR markers highlighted a high intra-population diversity thus denoting a noteworthy gene flow within populations [24]. Previous studies using RAPD and ISSR markers were used to examine the diversity and genetic connection of 100 cashew germplasm accessions, in which a significant amount of genetic diversity among accessions was reported [59]. Since 2020, studies on cashew genetic diversity have been pushed forward due to its high economic importance in tropical regions [25,27,57], yet have been done at a regional level, thus not capturing a global view of the genetic diversity of cashew current tropical distribution.

The current study's findings on the amount and distribution of genetic variation shows some insight into the mechanism of introduction and expansion of cashew in African continent and the Southeast Asia in East Timor. The presence of a significant overall genetic variation hints to numerous events of introduction involving different founder populations across CPLP countries. The founder effect would have been reflected in low genetic diversity if the introduction had been a one-time event. Furthermore, none of the genetic groups were restricted to a specific geographic location, but rather by an unexpected continental-associated genetic diversity trend.

Overall, no genetic flow was detected between *Anacardium* wild species from Brazil, included as outgroups, and cashew populations from several geographical provenances. Even populations sampled from the center of origin of cashew, in Brazil, no genetic flow was observed denoting a low hybridization between wild *Anacardium* species and cultivated/wild crop populations. Previous studies in the Cerrado biome and coastal Restinga vegetation, wild Brazilian populations of cashew were studied [21], and the genetic diversity in wild populations was higher than in domesticated ones, despite a weak distinction between wild and domesticated groups and with no correlation between genetic and geographical interpopulation distance. Our results suggests that populations screened from native cashew origin (Brazil) should be considered for *in situ* conservation programs from both wild (Ceará) and cultivated (Brasília) populations to assess gene flow and allelic diversity transfer regarding agronomically desirable traits.

In Ivory Coast, genetic diversity of cashew was studied using SSRs and revealed an overall heterozygosity deficit and a high intra-population genetic diversity among cashew populations screened [27], which is in accordance with our results in AMOVA where a high share of diversity lies within populations. Regarding population structuring, AMOVA showed that most of the genetic diversity lies within individuals. The large proportion of diversity found within individuals and within populations for the two groupings analyzed (all populations including outgroup species and excluding outgroups) suggests a high gene flow between populations, which may be attributed to the wind-pollinated reproduction of cashew allied with its outcrossing habit [60]. Since cashew is an outcrossing plant, negative to low inbreeding coefficients (F) were expected, which agrees with previous studies [61,62]. Cashew

is primarily an allogamous species favoring cross-fertilization [61], thus allowing intraspecific hybridizations and enhancing genetic variation. Outcrossing plant species tend to have higher genetic variation within-populations, whereas selfing species or species with a mixed mating system are often genetically less variable [63].

The population structuring using genetic distances revealed similar clustering with individual-based approaches, which supports results obtained under different analytical methods. The presence of four different clusters in the UPGMA tree with outgroup species (**Figure 10**) indicate that there is a clear separation between the wild *Anacardium* species (*A. giganteum* and *A. humile*) and the cashew populations screened. In the first cluster, East Timor with Indonesia are clustered, and one can speculate that Triloca population could be suggested as the place of entry of cashew in East Timor through Indonesia. The second cluster includes Guinea-Bissau populations and São Tomé and Principe, thus highlighting a strong gene flow between the two former Portuguese colonial countries. The third cluster includes Cabo Verde populations with Mozambique and Brazil, and it is also correlated with the gene flow in **Figure 11B** by the blue cluster. Quinara (GBQ) in Guinea-Bissau is a continental region which is a passing region between Formosa and Bolama Islands from Bijagós archipelago and the remaining continental territory, and where a different cashew accessions diversity at nut and apple is observed (FMonteiro, pers. comm.), which could explain somewhat its specific genetic cluster observed in the UPGMA trees (**Figure 10**) and at STRUCTURE analyses (**Figure 12**).

The clustering of cashew populations in this study had an uncommon, yet existing relationship with geographical region under a continental-wise distribution, which is contrary to previous reports in India [6], where no relationship with the geographic region was observed. Despite a high genetic diversity attributed to the high heterozygosity, allogamous nature and high gene flow found in cashew [64,65], also obtained in our study, at continental level dissimilar genetic clustering suggests a diverse genetic build-up when outgroup species are included; however, excluding outgroup species, a different and complex intra-specific diversity is observed.

Individual-based clustering methods using a Bayesian approach (STRUCTURE) and a multivariate analysis by DAPC allowed to assess the population structuring, thus highlighting that genetic diversity scattering does follows a geographical trend under a continental distribution. STRUCTURE and DAPC analyses suggest a high genetic diversity and genetic flow among the various tropical regions screened in our study. When comparing the analysis of K = 17 with outgroups **(Figure 11**) and the analysis without outgroups K =16 (**Figure 12**), we can see that both outgroup populations are considered just one cluster, which confirms the absence of genetic flow between the two *Anacardium* species and cashew populations. Excluding outgroups makes the optimal K decrease to 16, highlighting a complex intra-specific genetic diversity and population structuring. Another noteworthy remark is in the Bayesian analysis namely at K = 5 (**Figure 11B**), which is the second optimal K at STRUCURUE with the 20-populations dataset, where a gene flow between the Brazilian populations at the cashew origin, with Cabo Verde and Mozambican populations are depicted, which could be explained by the fact that in these countries cashews may not been improved into distinct varieties but rather been maintained by farmer's preferences and crop performance, thus maintaining a common genetic diversity. Also, historical records underlined the significance of Portuguese-led trade routes in the proliferation of several significant economic plants since the 16th century, as part of Columbian exchange period [11], where it is highlighted the role Cabo Verde archipelago in these trades for acclimating tropical crop species prior to agriculture purposes at the Iberian Peninsula and African continent to meet colonial expenses [11]. In addition to the gene flow in these three countries, we observed that in GBFORM from Formosa Island at Guinea-Bissau, genetic clustering with Cabo Verde, Mozambique and Brazil are shared (in blue cluster, **Figure 11B**), which could be indicative of a potential location of cashew remnant populations upon crop introduction at Guinea-Bissau. Also, in K= 5, the center of origin of cashew has clearly a gene flow with the remaining populations. From the Southeast Asia region (East Timor and Indonesia), a specific

genetic diversity (red cluster, **Figure 11B**) was observed, which could indicate a potential different route of cashew introduction/expansion, which remains to be addressed.

Overall, our data reports the first comprehensive study on cashew intraspecific diversity using a continental approach thus highlighting the need to perform conservation programs focused on a country standpoint. This is especially important considering that the result of a limited gene flow across countries were observed, which may be due to local selection of accessions within each country.

4.3. Historical notes on cashew at pantropical scale

Cashew introduction in both African and Asian continents has been subject of debate throughout the years, with the general belief that the Portuguese brought the cashew to the East Indies, Africa, and India, and that the Spanish took it to Central America and the Philippines [66]. According to this assumption, the Portuguese introduced the cashew to India via the western coast, most likely Goa, in the 16th century (1560) [67]; while in the African continent, probably through Cabo Verde [11] and to other coastal areas. Cashew was then spread throughout Southeast Asia and Africa, with Portuguese navigators bringing seeds to India and Mozambique in the 16th century, more for the wine and brandy than for the nuts [68]. Cashew was only later recognized as a tree used to prevent soil erosion in the coastal region, until its expansion has occurred as we know it today.

The genetic flow among Brazilian populations that share genetic diversity with the Cabo Verde, Mozambique and São Tomé and Príncipe populations, along the two islands populations from Formosa and Bolama at Guinea-Bissau is of major importance, since it links the genetic diversity of cashew from its center of origin to the African Continent, at the coastal areas sampled (**Figure 13**). This shared genetic diversity and gene flow may be linked to historical accounts that point out to several staple (e.g. rice, etc) and commercial (e.g. potato, maize, etc) crops have been brought from the New World (South America) to Old World (Africa) as part of the Columbian exchange event in 16th century [11], in which cashew could be included. Clearly, at the African continent multiple introduction events from diverse routing countries have occurred. After, by the 1950s, cashew was seen as a viable tree for coping with soil erosion and afforestation by the Portuguese colonial empire, not only in Africa but also in India, where it was seen as an optimal plant to cope with monsoons [11].

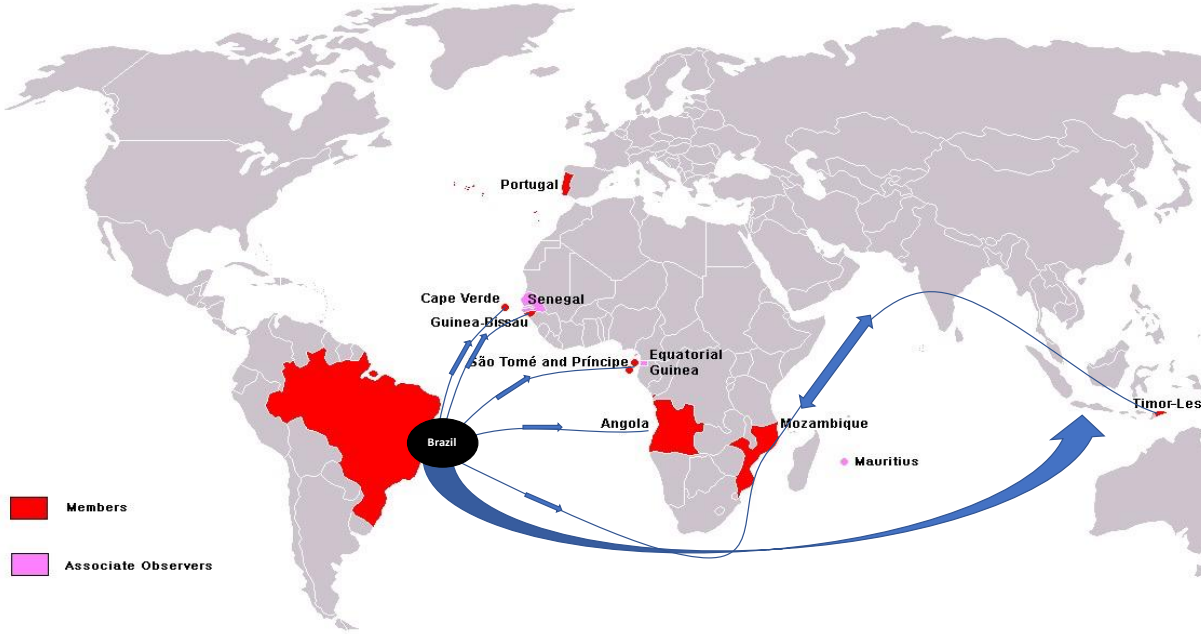

Figure 13. Schematic proposal of *A. occidentale* expansion map from Brazil to the studied tropical regions, namely CPLP countries. Adapted from Ferrão (1993).

The remaining CPLP countries, all former Portuguese colonies, were also engaged in cashew at first as a rustic tree and only after as a commercial crop, with different time advances in establishing cashew nuts as an export-commodity. Despite the multiple-event introduction on the African continent, in Guinea-Bissau at Formosa (GBFORM) and Bolama (GBBOL) islands, a genetic diversity linked to other African countries are observed, whereas at continental regions a high and different genetic build up are represented. Both Bolama and Formosa Islands were among the first regions in GB where cashew was explored as a potential commercial crop and, only then, expanded throughout the continental territory [4,11]. The results suggest that continental Quinara region from Guinea-Bissau display the least genetic diversity when compared to other continental populations which a different genetic diversity is depicted and may be related to an increased number of different cashew plantations according to farmer's preferences, as opposite to Bolama and Formosa Islands, where new plantations are scarce due to the abandonment of new orchards as part of human migration from rural to urban areas in the search for better life opportunities.

Regarding East Timor populations, there is a surprising genetic diversity clustering with some populations of Guinea-Bissau and Mozambique, not present in Brazil populations. This could indicate two different scenarios for East Timor: one, there is a route of cashew introduction from Brazil to East Timor through India, and second, directly via African continent. For testing this hypothesis, cashew populations from India should be included in future analyses. Mozambique also shares a genetic flow with Cabo Verde, which may imply that there is a separate point of entry in Mozambique, which one cannot be discarded through India or also from Angola. Thus, considering India's importance as a former commercial trading route in Asia and the CPLP country Angola that was not possible to sample for this work, the inclusion of Indian and Angolan cashew populations in future studies could help explain the complex intra-specific diversity within East Timor and within African continent.

The findings obtained in this work could be a baseline for the assessment of the diversity of cashew genetic resources at a pantropical scale, as one of the most economically important cash crops in tropical regions, still largely understudied.

5. Final remarks

Overall, our study successfully illustrates the significance of comparing the genetic diversity and structure of eighteen population of *A. occidentale* in CPLP countries. The SSR markers revealed prominent levels of polymorphisms and PIC values, which goes in accordance with the premise that they are reliable indicators of the existent genetic diversity at an inter-population level. In fact, the SSR markers revealed the genetic relationships and diversity of cashew cultivated in Guinea-Bissau. Identification of alleles underlying such distinctive diversity would be of utmost importance for determining their usefulness for incorporating future cashew selection programs, and to additionally propose on farm conservation policies at country level. Finally, we can tell a story of how these species occurred over time and the speed at which it acquired diversity. A particularly interesting fact is since these species are a perennial tree with a large life cycle, exhibit a high degree of polymorphism in less than 90 years. This study also provides useful information that can be used in a future breeding program for genetic improvement and, the characterization of new varieties. Plant genetic diversity (PGR) allows plant growers to develop new and improved cultivars with appealing characteristics that combine farmer (yield potential and large seeds) and breeders (pest and disease resistance) preferences. Since the beginning of agriculture, natural genetic variability within crop species has been used to meet the needs of subsistence agriculture. Cashew is an example of an agricultural crop in desperate need of sustainable production standards, owing to a lack of improved varieties and germplasm banks. To our knowledge, this is the first study to perform genetic analysis on larger numbers of individuals from the 18 cashew

populations present in CPLP countries. In addition, this data is relevant for developing management and conservation strategies for cashew genetic resources.

6. REFERENCES

1. CPLP *Estatísticas Da CPLP*; 2012, vol 1; pag 303 Instituto Nacional de Estatística, I.P; ISBN 9789892501109.
2. Gates, D.M. *Tree Crops*; 2020; Vol. 29; ISBN 9783030621391.
3. Monteiro, F.; Romeiras, M.M.; Figueiredo, A.; Sebastiana, M.; Baldé, A.; Catarino, L.; Batista, D. Tracking Cashew Economically Important Diseases in the West African Region Using Metagenomics. *Front. Plant Sci.* **2015**, *6*, 1–6, doi:10.3389/fpls.2015.00482.
4. Monteiro, F.; Catarino, L.; Batista, D.; Indjai, B.; Duarte, M.C.; Romeiras, M.M. Cashew as a High Agricultural Commodity in West Africa: Insights towards Sustainable Production in Guinea-Bissau. *Sustain.* **2017**, *9*, 1–14, doi:10.3390/su9091666.
5. Oliveira, N.N.; Mothé, C.G.; Mothé, M.G.; de Oliveira, L.G. Cashew Nut and Cashew Apple: A Scientific and Technological Monitoring Worldwide Review. *J. Food Sci. Technol.* **2020**, *57*, 12–21, doi:10.1007/s13197-019-04051-7.
6. Archak, S.; Gaikwad, A.B.; Swamy, K.R.M.; Karihaloo, J.L. Genetic Analysis and Historical Perspective of Cashew (*Anacardium occidentale* L.) Introduction into India. *Genome* **2009**, *52*, 222–230, doi:10.1139/G08-119.
7. Royo, V.D.A.; Mercadante-Simões, M.O.; Ribeiro, L.M.; De Oliveira, D.A.; Aguiar, M.M.R.; Costa, E.R.; Ferreira, P.R.B. Anatomy, Histochemistry, and Antifungal Activity of *Anacardium humile* (Anacardiaceae) Leaf. *Microsc. Microanal.* **2015**, *21*, 1549–1561, doi:10.1017/S1431927615015457.
8. Salehi, B.; Gültekin-Özgüven, M.; Kirkin, C.; Özçelik, B.; Morais-Braga, M.F.B.; Carneiro, J.N.P.; Bezerra, C.F.; Da Silva, T.G.; Coutinho, H.D.M.; Amina, B.; *et al*. *Anacardium* Plants: Chemical,Nutritional Composition and Biotechnological Applications. *Biomolecules* **2019**, *9*, 1–34, doi:10.3390/biom9090465.
9. Nambiar, M.C. Cashew [*Anacardium occidentale L.*, Ecophysiology]. **1977**.
10. Johnson, J.W. The Botany, Origin, and Spread of the Cashew *Anacardium occidenta*le L. *J. Endod.* **1973**, *1*, 1–7, doi:10.1016/S0099-2399(06)81513-X.
11. Havik, P.J.; Monteiro, F.; Catarino, S.; Correia, A.M.; Catarino, L.; Romeiras, M.M. Agro-Economic Transitions in Guinea-Bissau (West Africa): Historical Trends and Current Insights. *Sustain.* **2018**, *10*, 1–19, doi:10.3390/su10103408.
12. Monteiro, F.; Vidigal, P.; Barros, A.B.; Monteiro, A.; Oliveira, H.R.; Viegas, W. Genetic Distinctiveness of Rye *In Situ* Accessions from Portugal Unveils a New Hotspot of Unexplored Genetic Resources. **2016**, *7*, 1–17, doi:10.3389/fpls.2016.01334.
13. Buss, P.M.; Ferreira, J.R. Diplomacia Da Saúde e Cooperação Sul-Sul: As Experiências Da Unasul Saúde e Do Plano Estratégico de Cooperação Em Saúde Da Comunidade de Países de Língua Portuguesa (CPLP). *Reciis* **2010**, *4*, 106–118, doi:10.3395/reciis.v4i1.351pt.
14. FAO/INFOODS density database, version 2.0. In Food and agriculture organization; Source: FAOSTAT 2022 (accessed on January 2022).
15. Ricau, P. The West African Cashew Sector in 2018: General Trends and Country Profiles. *Anal. cashew Prod. trade West Africa., 30.* **2019**, *30*.
16. Nuts and Dried Fruits Statistical Yearbook 2019/20. Reus, S.A. online at https://www. nutfruit. org/files/tech/1587539172_INC_Statistical_Yearbook_201.-2020. pd. *International Nut and*

Dried Fruit Council Foundation; 2020;

17. USDA. *Nutr. Data Lab. Agric. Res. Serv.* **2015**.
18. Tedong, L.; Madiraju, P.; Martineau, L.C.; Vallerand, D.; Arnason, J.T.; Desire, D.D.P.; Lavoie, L.; Kamtchouing, P.; Haddad, P.S. Hydro-Ethanolic Extract of Cashew Tree (*Anacardium occidentale* L.) Nut and Its Principal Compound, Anacardic Acid, Stimulate Glucose Uptake in C2C12 Muscle Cells. *Mol. Nutr. Food Res.* **2010**, *54*, 1753–1762, doi:10.1002/mnfr.201000045.
19. Sharma, P.; Gaur, V.K.; Sirohi, R.; Larroche, C.; Kim, S.H.; Pandey, A. Valorization of Cashew Nut Processing Residues for Industrial Applications. *Ind. Crops Prod.* **2020**, *152*, 112550, doi:10.1016/j.indcrop.2020.112550.
20. Adeigbe, O.O.; Olasupo, F.O.; Adewale, B.D.; Muyiwa, A.A. A Review on Cashew Research and Production in Nigeria in the Last Four Decades. *Sci. Res. Essays* **2015**, *10*, 196–209, doi:10.5897/sre2014.5953.
21. dos Santos, J.O.; Mayo, S.J.; Bittencourt, C.B.; de Andrade, I.M. Genetic Diversity in Wild Populations of the Restinga Ecotype of the Cashew (*Anacardium occidentale* L.) in Coastal Piauí, Brazil. *Plant Syst. Evol.* **2019**, *305*, 913–924, doi:10.1007/s00606-019-01611-4.
22. Croxford, A.E.; Robson, M.; Wilkinson, M.J. Characterization and PCR Multiplexing of Polymorphic Microsatellite Loci in Cashew (*Anacardium occidental*e L.) and Their Cross-Species Utilization. *Mol. Ecol. Notes* **2006**, *6*, 249–251, doi:10.1111/j.1471-8286.2005.01208.x.
23. Rahman, M.-; Engineering, G. *Genetics and Genomics of Cotton*; 2009; ISBN 9780387708102.
24. Cavalcanti, J.J.V.; Wilkinson, M.J. The First Genetic Maps of Cashew (*Anacardium occidentale* L.). *Euphytica* **2007**, *157*, 131–143, doi:10.1007/s10681-007-9403-9.
25. Savadi, S.; Muralidhara, B.M.; Preethi, P. Advances in Genomics of Cashew Tree: Molecular Tools and Strategies for Accelerated Breeding. *Tree Genet. Genomes* **2020**, *16*, doi:10.1007/s11295-020-01453-z.
26. Perea, C.; De La Hoz, J.F.; Cruz, D.F.; Lobaton, J.D.; Izquierdo, P.; Quintero, J.C.; Raatz, B.; Duitama, J. Bioinformatic Analysis of Genotype by Sequencing (GBS) Data with NGSEP. *BMC Genomics* **2016**, *17*, doi:10.1186/s12864-016-2827-7.
27. Kouakou, C.K.; Adopo, A.N.; Djaha, A.J.B.; N'da, D.P.; N'da, H.A.; Bi, I.A.Z.; Koffi, K.K.; Djidji, H.; Minhibo, M.Y.; Dosso, M.; et al. Genetic Characterization of Promising High-Yielding Cashew (*Anacardium occidentale* l.) Cultivars from Côte d'Ivoire. *Biotechnol. Agron. Soc. Environ.* **2020**, *24*, 46–58, doi:10.25518/1780-4507.18464.
28. Mneney, E.E.; Mantell, S.H.; Bennett, M. Use of Random Amplified Polymorphic DNA (RAPD) Markers to Reveal Genetic Diversity within and between Populations of Cashew (*Anacardium occidentale* L.). *J. Hortic. Sci. Biotechnol.* **2001**, *76*, 375–383, doi:10.1080/14620316.2001.11511380.
29. Chipojola, F.M.; Mwase, W.F.; Kwapata, M.B.; Bokosi, J.M.; Joyce, P.; Maliro, M.F. Morphological Characterization of Cashew (*Anacardium occidentale* L.) in Four Populations in Malawi. *African J. Biotechnol.* **2009**, *8*, 5173–5181.
30. Aliyu, O.M.; Awopetu, J.A. Multivariate Analysis of Cashew (*Anacardium occidentale* L.) Germplasm in Nigeria. *Silvae Genet.* **2007**, *56*, 170–179, doi:10.1515/sg-2007-0026.
31. Schuelke, M. An Economic Method for the Fluorescent Labeling of PCR Fragments A Poor Man ' s Approach to Genotyping for Research and High-Throughput Diagnostics . *Prism* **2000**, *18*, 1–2.
32. Holleley, C.E.; Geerts, P.G. Multiplex Manager 1.0: A Cross-Platform Computer Program That Plans and Optimizes Multiplex PCR. *Biotechniques* **2009**, *46*, 511–517, doi:10.2144/000113156.
33. Culley, T.M.; Stamper, T.I.; Stokes, R.L.; Brzyski, J.R.; Hardiman, N.A.; Klooster, M.R.; Merritt, B.J. An Efficient Technique for Primer Development and Application That Integrates Fluorescent Labeling and Multiplex PCR. *Appl. Plant Sci.* **2013**, *1*, 1300027,

doi:10.3732/apps.1300027.

34. Brownstein, M.J.; Carpten, J.D.; Smith, J.R. Modulation of Non-Templated Nucleotide Addition by Taq DNA Polymerase: Primer Modifications That Facilitate Genotyping. *Biotechniques* **1996**, *20*, 1004–1010, doi:10.2144/96206st01.

35. Van Oosterhout, C.; Hutchinson, W.F.; Wills, D.P.M.; Shipley, P. MICRO-CHECKER: Software for Identifying and Correcting Genotyping Errors in Microsatellite Data. *Mol. Ecol. Notes* **2004**, *4*, 535–538, doi:10.1111/j.1471-8286.2004.00684.x.

36. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data Via the EM Algorithm . *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22, doi:10.1111/j.2517-6161.1977.tb01600.x.

37. Chapuis, M.P.; Estoup, A. Microsatellite Null Alleles and Estimation of Population Differentiation. *Mol. Biol. Evol.* **2007**, *24*, 621–631, doi:10.1093/molbev/msl191.

38. Shaibi, T.; Lattorff, H.M.G.; Moritz, R.F.A. A Microsatellite DNA Toolkit for Studying Population Structure in *Apis mellifera*. *Mol. Ecol. Resour.* **2008**, *8*, 1034–1036, doi:10.1111/j.1755-0998.2008.02146.x.

39. Peakall, R.; Smouse, P.E. GENALEX 6: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research. *Mol. Ecol. Notes* **2006**, *6*, 288–295, doi:10.1111/j.1471-8286.2005.01155.x.

40. Rousset, F. GENEPOP'007: A Complete Re-Implementation of the GENEPOP Software for Windows and Linux. *Mol. Ecol. Resour.* **2008**, *8*, 103–106, doi:10.1111/j.1471-8286.2007.01931.x.

41. Rice, W.R. Analyzing Tables of Statistical Tests. *Evol. Vol. 43, No. 1, pp. 223-225* **2013**, *52*, 647–674.

42. Cavalli-Sforza, L.L.; Edwards, A.W. Phylogenetic Analysis. Models and Estimation Procedures. *Am. J. Hum. Genet.* **1967**, *19*, 233–257, doi:10.2307/2406616.

43. Nei, M. Genetic Distance between Populations. *Am. Nat.* **1972**, *106*, 283–292, doi:10.1086/282771.

44. Paradis, E.; Claude, J.; Strimmer, K. APE : Analyses of Phylogenetics and Evolution in R Language. **2004**, *20*, 289–290, doi:10.1093/bioinformatics/btg412.

45. R Core Team 2021 R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput. Vienna, Austria* 2021.

46. Kamvar, Z.N.; Tabima, J.F.; Gr̈unwald, N.J. Poppr: An R Package for Genetic Analysis of Populations with Clonal, Partially Clonal, and/or Sexual Reproduction. *PeerJ* **2014**, 1–14, doi:10.7717/peerj.281.

47. Rambaut, Andrew. FigTree. Tree figure drawing tool. http://tree. bio. ed. ac. uk/software/figtree/, **2009**.

48. Weir, B.; Clark Cockerham, C. Estimating F-Statistics for the Analysis of Population Structure Author (s): Published by : Society for the Study of Evolution Stable URL : Http://Www.Jstor.Org/Stable/2408641 . *Evolution (N. Y).* **2011**, *38*, 1358–1370.

49. Laurent Excoffier, * Peter E. Smouse* and Joseph M. Quattro*' Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genet. Soc. Am.* **2000**, *198*, 283–291, doi:10.3354/meps198283.

50. Hill, W.G. Book Reviews Genetic Data Analysis II. By. **2021**, 87893.

51. Excoffier, L.; Lischer, H.E.L. Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol. Ecol. Resour.* **2010**, *10*, 564–567, doi:10.1111/j.1755-0998.2010.02847.x.

52. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. **2000**, 155(2), 945-959.

53. Thibaut Jombart, S.D.; Balloux, F. Discriminant Analysis of Principal Components: A New Method for the Analysis of Genetically Structured Populations. *PLoS Comput. Biol.* **2009**, *5*, doi:10.1371/journal.pcbi.1000455.

54. Earl, D.A.; Bridgett, M. STRUCTURE HARVESTER : A Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. **2012**, 359–361, doi:10.1007/s12686-011-9548-7.

55. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the Number of Clusters of Individuals Using the Software STRUCTURE: A Simulation Study. *Mol. Ecol.* **2005**, *14*, 2611–2620, doi:10.1111/j.1365-294X.2005.02553.x.

56. Jakobsson, M.; Rosenberg, N.A. CLUMPP : A Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure. **2007**, *23*, 1801–1806, doi:10.1093/bioinformatics/btm233.

57. Savadi, S.; Megha, K.S.V.S.; Mohana, B.M.M.G.S. Genetic Diversity and Identification of Interspecific Hybrids of Anacardium Species Using Microsatellites. *Brazilian J. Bot.* **2020**, *2026*, doi:10.1007/s40415-020-00678-5.

58. Mashood, O. Genetic Diversity of Nigerian Cashew Germplasm. *Genet. Divers. Plants* **2012**, doi:10.5772/32892.

59. Thimmappaiah, S. M. G., Shobha, D., & Anil, S.R. Assessment of Cashew Species for Molecular Diversity. *J. Plant. Crop.* **2009**, *2*, 146–151.

60. Bennett, R.A.; Thiagarajah, M.R.; King, J.R.; Rahman, M.H. Interspecific Cross of *Brassica oleracea* Var. *alboglabra* and *B. napus*: Effects of Growth Condition and Silique Age on the Efficiency of Hybrid Production, and Inheritance of Erucic Acid in the Self-Pollinated Backcross Generation. *Euphytica* **2008**, *164*, 593–601, doi:10.1007/s10681-008-9788-0.

61. Freitas, B.M.; Paxton, R.J. The Role of Wind and Insects in Cashew (Anacardium occidentale L.) Pollination in NE Brazil. *J. Agric. Sci.* **1996**, *126*, 319–326, doi:10.1017/s0021859600074876.

62. Layek, U.; Bera, K.; Bera, B.; Bisui, S.; Pattanayek, S.K.; Karmakar, P. Assessment of Yield Enhancement in Cashew (*Anacardium occidentale* L.) by the Pollinator Sharing Effect of Magnetic Bee-Friendly Plants in India. *Acta Ecol. Sin.* **2021**, *41*, 243–252, doi:10.1016/j.chnaes.2021.05.003.

63. Nybom, H. Comparison of Different Nuclear DNA Markers for Estimating Intraspecific Genetic Diversity in Plants. *Mol. Ecol.* **2004**, *13*, 1143–1155, doi:10.1111/j.1365-294X.2004.02141.x.

64. Mitchell, J.D; Mori, S.A. The Cashew and Its Relatives (*Anacardium*: Anacardiaceae). El Marañón y Sus Parientes (*Anacardium*: Anacardiaceae). *Bibl. OET M* **1987**, *42*, v. 42, 1-76. Año 1987.

65. Daniel Borges Cultures Du Timor-Oriental : Processus d'objectification. *Plur. Plur. - Rev. des Cult. Lang. Port.* **2018**.

66. Singh, A.K. Early History of Crop Presence/Introduction in India: III. *Anacardium occidentale* L. *Cashew Nut. Asian Agri-History* **2018**, *22*, doi:10.18311/aah/2018/21389.

67. Jonathan D. Sauer Historical Geography of Crop Plants. *CRC Press* **1993**, doi:https://doi.org/10.1201/9780203751909.

68. Brücher, H. *Useful Plants of Neotropical Origin and Their Wild Relatives*; **1991**; Vol. 35; ISBN 9783642733154.

LIST OF SUPPLEMENTARY DATA

Table S1. Populations by country, district and location, geographical coordinates and the total number of individuals sampled by population (N).

Country	District	Location	Population	Latitude	Longitude	N
East Timor	Manatuto	Kribas	ETK	-8.650	125.983	17
	Baucau	Triloca L3	ETTR	-8.490	126.370	16
	Cova Lima	Suai	ETSU	-9.432	125.108	16
	Viqueque	Viqueque	ETV	-8.907	126.273	18
Indonesia	Flores Island	Flores	IND	-8.681	121.580	18
Brazil	Ceára	Fortaleza	BR1	-3.878	-38.562	16
	Brasília	Brasília	BR2	-15.856	-47.760	17
	Pará	*A. giganteum*	AG	-5.697	-52.207	16
	Amazonia	*A. humile*	AH	-3.468	-62.220	18
Mozambique	Sofala	Beira	MZB	-19.804	34.887	17
		Dondo	MZD	-19.611	34.735	17
Cabo Verde	São Nicolau	São Nicolau	CVSN	16.592	-24.276	15
	Fogo island	Fogo island	CVF	14.925	-24.384	16
	Santiago island	Santiago island	CVST	15.063	-23.625	16
Guinea-Bissau	Bolama	Praia de Ofir	GBBOL	11.555	-15.470	20
	Gabú	Granja do Estado	GBGAB	12.248	-14.224	19
	Biombo	Ponta Romana	GBBIO	11.800	-15.705	16
	Quinara	Buba	GBQ	11.612	-14.964	17
	Bolama	Formosa island	GBFORM	11.471	-15.683	18
São Tomé and Principe	São Tomé	Roça	STOM	0.296	6.530	20

Table S2. Hardy–Weinberg equilibrium (HWE) test for each locus-population combination using GenePop v4.5. Statistical significance was assessed by running 10.000 iterations Monte Carlo Markov Chain (MCMC) test. *p*-values were corrected by multiple comparisons applying a sequential Bonferroni correction ($p<0.00017857$. [0.05/280). $p<0.05$. light blue; $p<0.00017857$. dark blue.)

	mAoR48	mAoR6	mAoR17	mAoR7	mAoR11	mAoR3	mAoR42	mAoR52	mAoR2	mAoR35	mAoR47	mAoR16	mAoR29	mAoR41
TLK	0.0005	0.1857	0.0000	0.1253	0.0808	0.0000	0.0401	0.4497	0.1174	0.0000	0.0083	0.0005	0.0022	0.0001
TLTR	0.0051	0.0037	0.1475	0.0000	0.0021	0.0000	0.0043	1.0000	1.0000	1.0000	0.1815	1.0000	0.0000	0.0001
TLSU	0.0011	0.0000	0.0006	0.0003	0.0000	0.0000	0.0570	0.0006	0.1258	0.0000	0.2556	0.6222	0.0000	0.0001
TLV	0.0116	0.0942	0.0007	1.0000	0.3484	0.3274	1.0000	0.3575	0.2381	0.0481	1.0000	1	0.0000	0.0000
IND	0.4357	0.0000	0.0000	0.0006	0.0336	0.0004	0.0004	0.1205	1.0000	0.0000	0.0002	1	0.0000	0.0000
BR1	1.0000	0.0081	0.0002	0.1882	0.0106	0.0001	0.1916	0.0001	0.0002	0.0001	0.0000	0.3300	0.0000	0.0000
BR2	1.0000	1.0000	0.0374	0.0322	0.3375	0.0000	0.0002	0.0000	0.0000	0.0000	0.0188	0.0044	0.0384	0.0463
AG	0.0043	0.0000	0.0000	0.0002	0.4503	0.0001	0.0001	0.0001	1	0.0009	0.0003	0.0003	0.0000	0.0061
AH	1	0.0000	0.0000	0.0000	0.5076	0.0392	0.1005	1	1	0.0000	0.0000	0.2399	0.0402	1
MZB	0.0256	0.0001	0.0004	0.8045	0.0000	0.0000	0.0028	0.9465	0.0001	0.0000	0.0000	0.0003	0.0000	0.0000
MZD	0.4729	0.2951	0.6625	0.6089	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0095	0.0014	0.0005
CVSN	0.1427	0.0218	0.0009	0.2213	0.2411	0.0000	0.6265	0.0088	1.0000	0.0000	0.6803	0.0001	0.0017	0.5055
CVF	0.1512	0.0045	0.0004	0.0955	0.0924	0.0000	1.0000	0.0048	1.0000	0.0022	0.7727	0.0000	0.0002	1.0000
CVST	0.0001	0.2372	0.0012	0.0965	0.0001	0.0000	0.0060	0.2607	0.0000	0.0110	0.1172	0.0001	0.0002	0.0001
GBBOL	0.2745	0.0000	0.0016	0.0000	0.3994	0.0000	0.0009	0.0000	0.0000	0.0000	0.0000	1	0.0000	0.0639
GBGAB	0.0000	0.0000	0.1358	0.0000	0.0000	0.0000	0.0000	0.0000	0.3000	1.0000	0.0000	0.0000	0.0267	0.0000
GBBIO	0.3448	0.0002	0.7676	0.0000	0.0003	0.0000	0.5426	0.0029	0.2096	0.0009	0.0046	0.0000	0.5094	0.0002
GBQ	0.0000	0.0022	0.0000	0.0000	0.0000	0.0000	0.7641	0.0000	1	0.0003	0.0917	0.0000	0.0000	0.0038
GBFORM	0.0025	0.0031	0.0002	0.0001	0.0000	0.0001	0.0000	0.0003	0.0105	0.3918	0.0241	0.0000	0.0000	0.0001
STOM	1	0.0008	0.0020	1.0000	1	0.0000	0.0000	0.0005	1	0.0000	0.0000	0.5238	1	1

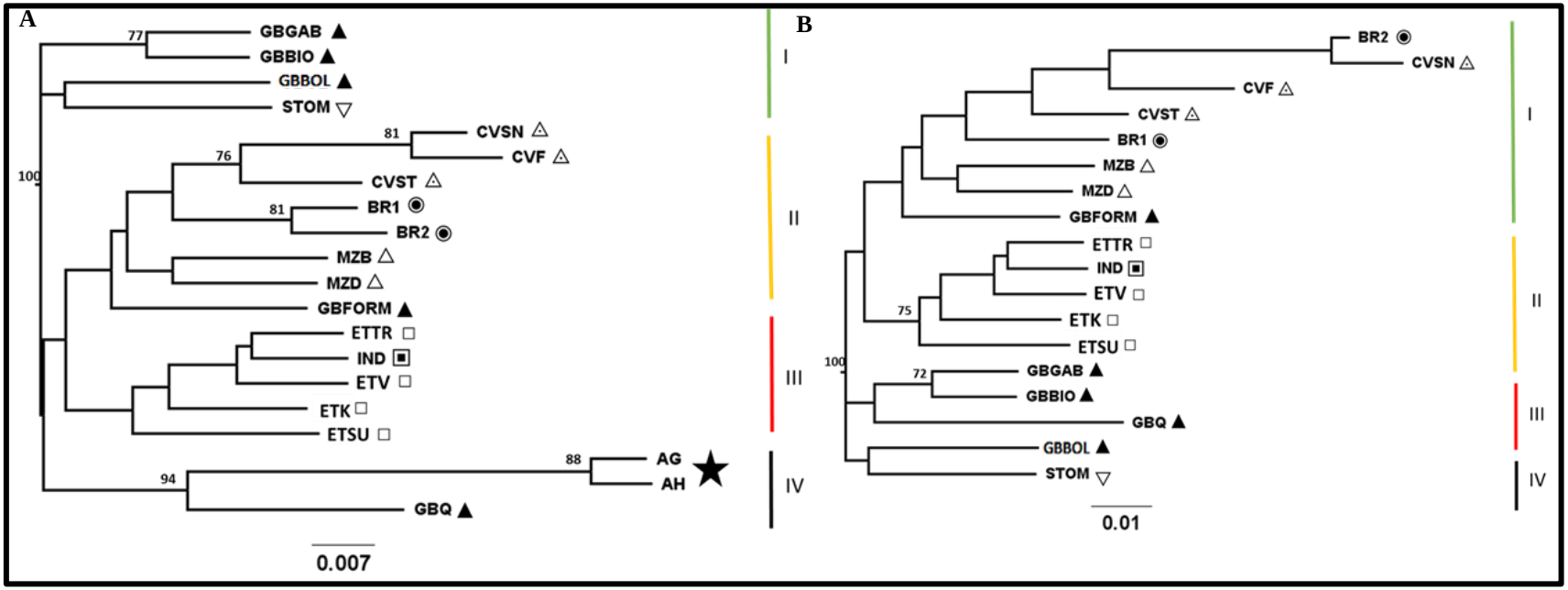

Figure S1. NJ trees generated from *FreeNA* using matrix DC^{INA} respectively, representing all 7 countries used in this study including (**A**) and excluding (**B**) outgroups. Clusters are represented by (I, II, III and IV) with coloring (Green, blue, yellow, and red). Legend: (□ East Timor, ▣ Indonesia, ◉ Brazil, ▽ São Tomé and Principe, △ Cabo Verde, ▲ Guinea-Bissau, △ Mozambique and the outgroups ★.

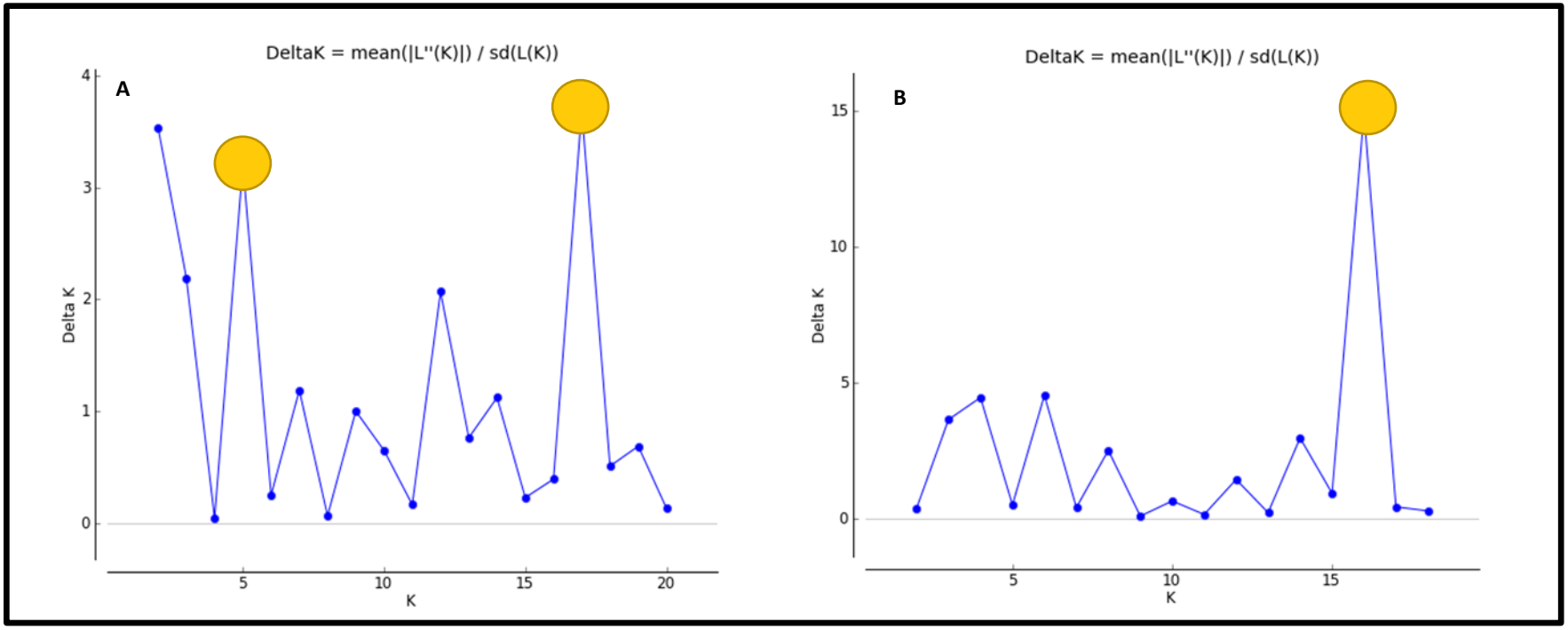

Figure S2. STRUCTURE *ad hoc* statistics retrieved by StructureHarvester using 1 to 10 possible clusters (*K*). Variation of Δ*K* values according to Evanno *et al*. [55] method for 20-populations dataset including outgroups (**A**) and excluding outgroups with only cashew populations (n=18) (**B**). Delta *K* is the most used method for determining the number of clusters to select. The parameter *K*, which describes the number of subpopulations that make up the total population, is an important value to quantity in the analysis of the structured populations. The model evidence, which is equivalent to the model's likelihood, is ideally used to infer *K*, being the highest values identified by an orange circle.

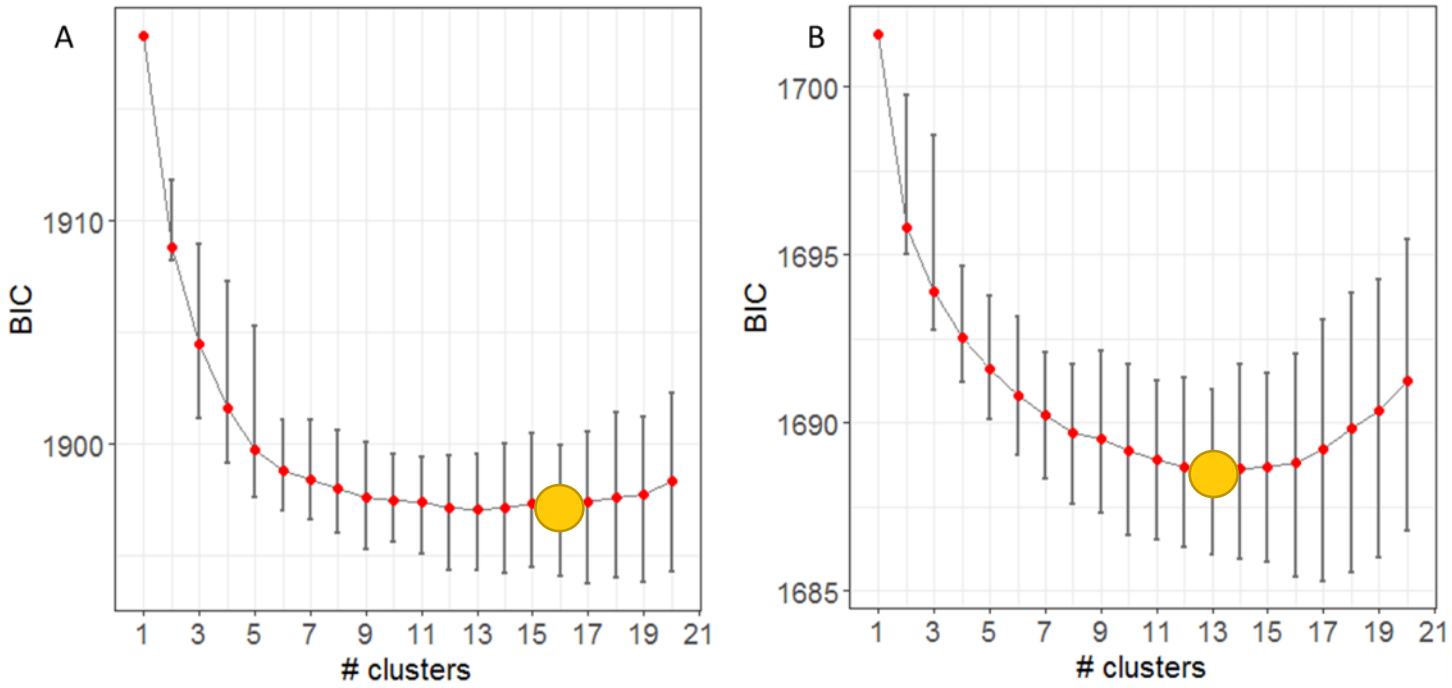

Figure S3. DAPC results inference of the number of clusters using *find.clusters* function with a *K*=16 (**A**) including the outgroup populations and *K*=13 with only the *A. occidentale* populations (**B**). Identification of the clusters is achieved by *find.clusters*, The function *find.clusters* first transforms the data using PCA, prompting the user to specify the number of retained PCs interactively. Then, unless the argument *n.clust* is provided, it runs the *K*-means algorithm (function *kmeans* from the *stats* package) with increasing values of *K* and computes associated summary statistics (by default, BIC). To determine the optimal number of clusters, *K*-means is run sequentially with increasing *K* values, and various clustering solutions are compared using the Bayesian Information Criterion (BIC). Ideally, the best clustering solution should have the lowest BIC. In practice, an elbow in the curve of BIC values as a function of *K* often indicates the 'best' BIC. The lowest values “optimal” are identified with an orange circle.

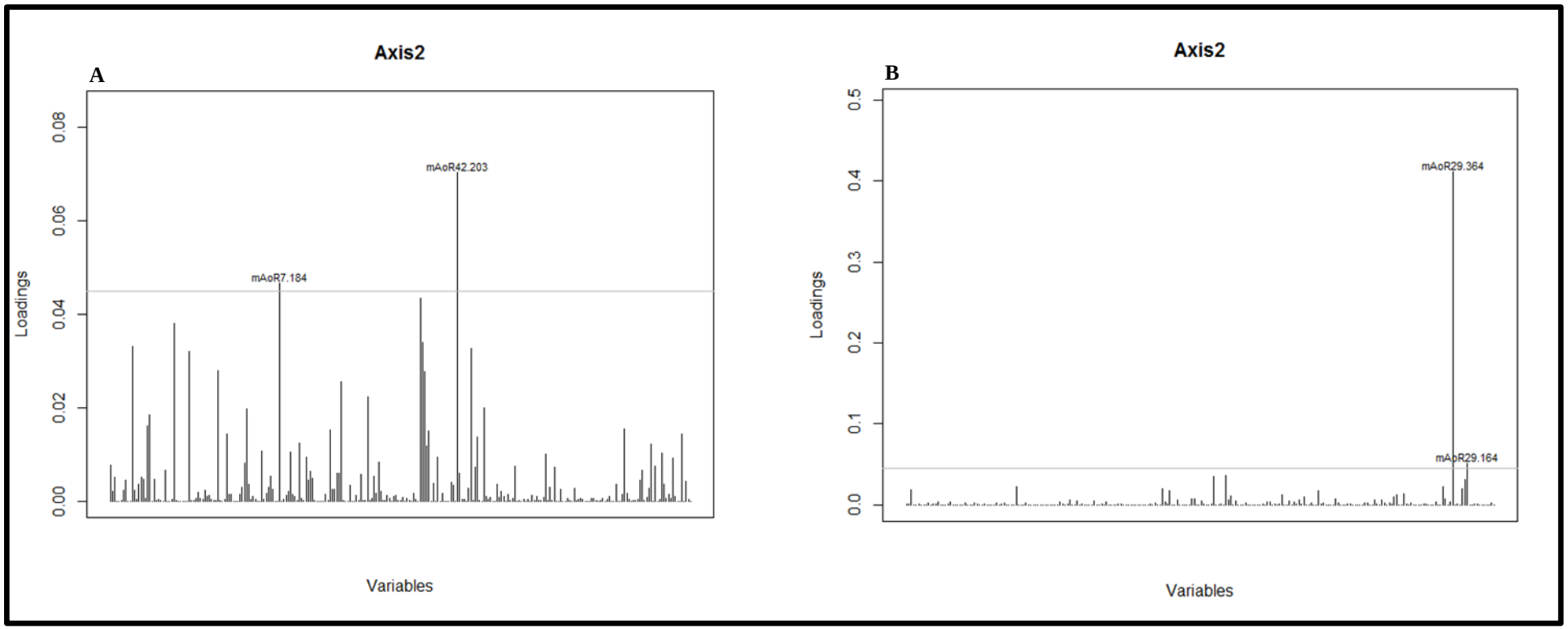

Figure S4. Loading plots of the two Discriminant Functions following DAPC analysis with a *K* = 5 (**A**) and *K* = 17 (**B**) for the 20- populations dataset, including outgroups.

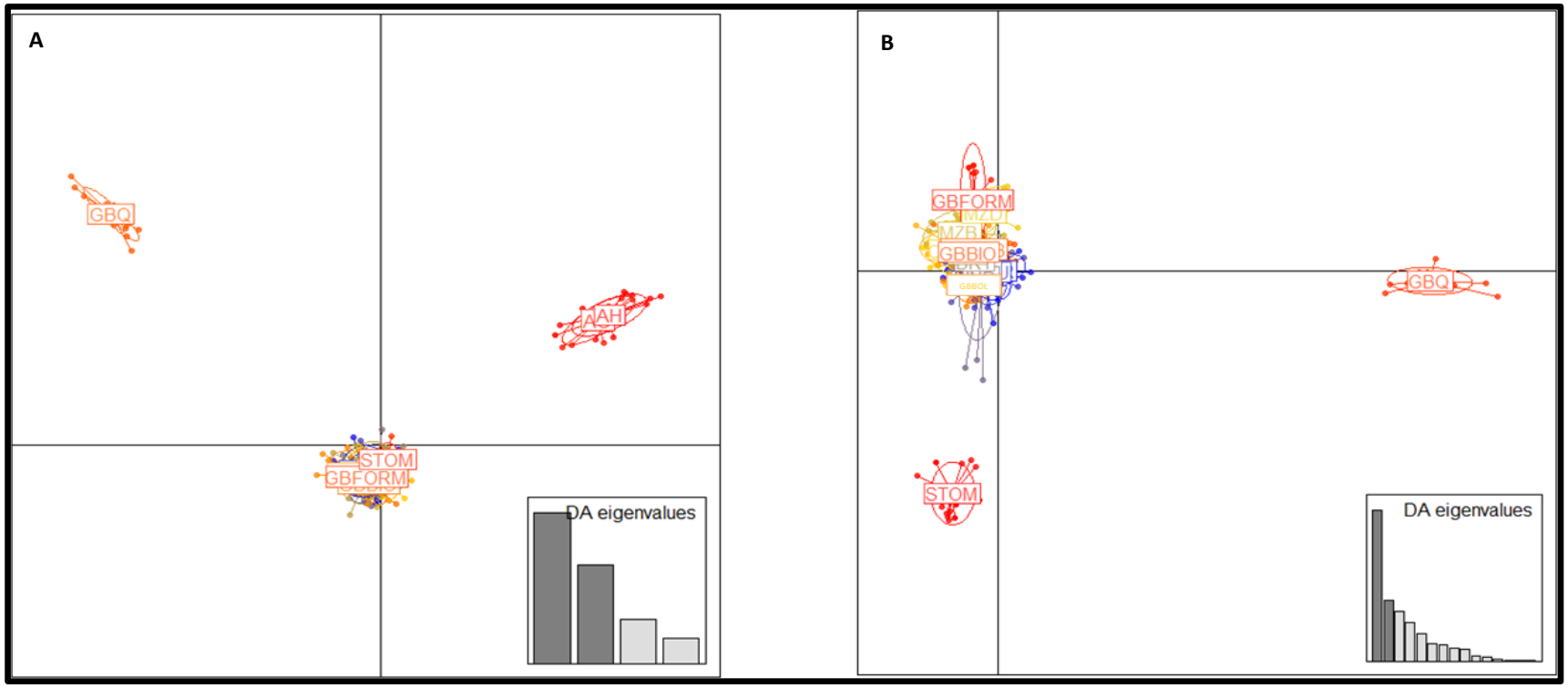

Figure S5. Scatterplot shows the two principal components of the DAPC, and clusters are numbered and displayed by different colors, while dots represent individuals. The 2 Discriminant Functions hereby represented explain 93% of cumulative variance of the dataset for all twenty populations for *K* = 5 (**A**) and *K* =16 (**B**) considering the analysis without outgroups.