East-Timor as an unexplored yet important source of cashew (*Anacardium occidentale* L.) genetic diversity

Lara Guterres [1, 2, 3], **João Barnabé** [2, 4], **André Barros** [5], **Alberto Bento Charrua** [3, 6], **Maria Cristina Duarte** [4], **Maria M. Romeiras** [2, 4], **Filipa Monteiro** [Corresp. 2, 4]

[1] Universidade Nacional Timor Lorosa'e (UNTL), Av. Cidade de Lisboa, Díli, East-Timor., Díli, East-Timor

[2] LEAF—Linking Landscape, Environment, Agriculture and Food Research Center, Associated Laboratory TERRA, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017, Lisboa, Portugal

[3] Nova School of Business and Economics, Universidade Nova de Lisboa, Campus de Carcavelos, Rua da Holanda, n.1, Carcavelos, 2775-405, Cascais, Portugal

[4] Centre for Ecology, Evolution and Environmental Changes (cE3c)& Global Change and Sustainability Institute (CHANGE), Faculty of Sciences, University of Lisbon, 1749-017, Lisboa, Portugal

[5] Innate Immunity and Inflammation Laboratory, Instituto Gulbenkian de Ciência (IGC), Rua da Quinta Grande, 6, 2780-156, Oeiras, Portugal

[6] Department of Earth Sciences and Environment, Faculty of Science and Technology, Licungo University, P.O. Box 2025, Beira 2100, Beira, Mozambique

Corresponding Author: Filipa Monteiro
Email address: fmonteiro@isa.ulisboa.pt

Background. Cashew (*Anacardium occidentale* L.) is a crop currently grown in several tropical countries due to economic importance of cashew nuts. Despite its enormous economic worth, limited research has been conducted on the molecular diversity of cashew genetic resources. In this study, a wide comprehensive assessment of the genetic diversity of cashew in East-Timor was screened using microsatellites (SSRs) to evaluate intraspecific diversity and population structuring.

Methods. A total of 207 individuals including that from East-Timor (11), and outgroup populations from Indonesia (1) and Mozambique (2), were analyzed with 16 cashew-specific SSRs. A comprehensive sampling within East-Timor was done, thus covering cashew orchards distribution in the country. Genetic diversity indices were calculated, and population structuring determined using three different approaches: genetic distances (UPGMA and NJ), AMOVA and individual- based clustering methods through a Bayesian (STRUCTURE) and multivariate (DAPC) analyses.

Results. Population structuring revealed that the genetic diversity of cashew populations in East-Timor was higher in this study than previously reported for cashew. A higher allelic richness was found within East-Timor populations, compared with the outgroup populations (Mozambique and Indonesia), reinforced by the presence of private alleles. Moreover, our study showed that East-Timor populations are grouped into two dissimilar genetic groups, which may suggest multiple introduction events of cashew. Hence, the newly studied cashew genetic resources could be explored for future crop improvement.

Conclusions. Crop diversity underpins the productivity, resilience and adaptive capacity of agriculture. Therefore, this study provides useful information regarding genetic diversity and population structure that can be harnessed to improve cashew production in East-Timor. This data can be also important to access an in-country genetic signature and increase cashew market value.

East-Timor as an unexplored yet important source of cashew (*Anacardium occidentale* L.) genetic diversity

Lara Guterres[1,2,3], João Barnabé[1,4], André Barros [5], Alberto B. Charrua[1,2,6], Maria Cristina Duarte[4], Maria M. Romeiras[1,4], Filipa Monteiro[1,4]

[1] LEAF—Linking Landscape, Environment, Agriculture and Food Research Center, Associated Laboratory TERRA, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal
[2] Nova School of Business and Economics, Universidade Nova de Lisboa, Campus de Carcavelos, Rua da Holanda, n.1, Carcavelos, 2775-405 Cascais, Portugal.
[3] Universidade Nacional Timor Lorosa'e (UNTL), Av. Cidade de Lisboa, Díli, East-Timor.
[4] Centre for Ecology, Evolution and Environmental Changes (cE3c) & Global Change and Sustainability Institute (CHANGE), Faculty of Sciences, University of Lisbon, 1749-017 Lisbon, Portugal.
[5] Innate Immunity and Inflammation Laboratory, Instituto Gulbenkian de Ciência (IGC), Rua da Quinta Grande, 6, 2780-156 Oeiras, Portugal.
[6] Department of Earth Sciences and Environment, Faculty of Science and Technology, Licungo University, P.O. Box 2025, Beira 2100, Mozambique.

Corresponding Author:
Filipa Monteiro[1,4]
Linking Landscape, Environment, Agriculture and Food (LEAF), Instituto Superior de Agronomia (ISA), Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisbon, Portugal.
Email address: fmonteiro@isa.ulisboa.pt, fimonteiro@fc.ul.pt.

Abstract

Background. Cashew (*Anacardium occidentale* L.) is a crop currently grown in several tropical countries due to economic importance of cashew nuts. Despite its enormous economic worth, limited research has been conducted on the molecular diversity of cashew genetic resources. In this study, a wide comprehensive assessment of the genetic diversity of cashew in East-Timor was screened using microsatellites (SSRs) to evaluate intraspecific diversity and population structuring.

Methods. A total of 207 individuals including that from East-Timor (11), and outgroup populations from Indonesia (1) and Mozambique (2), were analyzed with 16 cashew-specific SSRs. A comprehensive sampling within East-Timor was done, thus covering cashew orchards distribution in the country. Genetic diversity indices were calculated, and population structuring

determined using three different approaches: genetic distances (UPGMA and NJ), AMOVA and individual- based clustering methods through a Bayesian (STRUCTURE) and multivariate (DAPC) analyses.

Results. Population structuring revealed that the genetic diversity of cashew populations in East-Timor was higher in this study than previously reported for cashew. A higher allelic richness was found within East-Timor populations, compared with the outgroup populations (Mozambique and Indonesia), reinforced by the presence of private alleles. Moreover, our study showed that East-Timor populations are grouped into two dissimilar genetic groups, which may suggest multiple introduction events of cashew. Hence, the newly studied cashew genetic resources could be explored for future crop improvement.

Conclusions. Crop diversity underpins the productivity, resilience and adaptive capacity of agriculture. Therefore, this study provides useful information regarding genetic diversity and population structure that can be harnessed to improve cashew production in East-Timor. This data can be also important to access an in-country genetic signature and increase cashew market value.

Keywords

genetic diversity; SSRs; population structuring; Southeast Asia; diversity hotspots.

Introduction

Cashew (*Anacardium occidentale* L., Anacardiaceae) is a tropical evergreen tree that can thrive in both dry and wet tropical climates. Many tropical countries use various parts of the cashew tree for consumption, medical, and industrial purposes (Salehi *et al*., 2019). Among the various parts of the plant, the cashew kernel has a high market value and has acquired a status of an export-oriented commodity in several tropical countries, also known as a cash crop (Monteiro *et al*., 2015, 2017). The highest cashew-producing countries, mainly from West African region (Guinea-Bissau and Côte d'Ivoire) and from Southeast Asia (India and Vietnam) are mainly focused on exporting cashew nuts, from which both governments and farmers have their primary income (Havik *et al*., 2018). Over the past two decades, a global market demand of cashew nuts has been increasing as a result of the rising trend in consumption. Cashews accounted for 17% of world tree nut production in 2019/20, making it the third most popular tree nut after almonds and walnuts **(**International Nut and Dried Fruit Council Foundation, 2020**)**. The popularity of cashew is mainly because of the changed food habits towards a healthier food intake. As a rich source of plant-based protein, dietary minerals and low-fat contents, cashew has become one of the most produced and valued tree nuts worldwide, together with almonds, pistachios, and walnuts **(**Pradhan *et al*., 2020**)**.

Despite its enormous economic importance, studies evaluating the molecular diversity of cashew genetic resources have been scarce. Most studies were performed with different molecular markers, namely microsatellites (SSRs, Simple Sequence Repeats) and RAPDs (Random

Amplified Polymorphic DNA), mainly done in the top cashew-producing countries. For example, genetic diversity of cashew accessions have been characterized in Ivory Coast (Kouakou *et al*., 2020), in Nigeria (Aliyu & Awopetu, 2007), in India (Archak *et al*., 2009), in Tanzanian (Mneney *et al.*, 2001), Malawi (Chipojola *et al*., 2009) and Brazil (dos Santos *et al*., 2019). Overall, these studies highlight a narrow genetic diversity of cashew accessions when compared to the cashew native origin, Brazil, where a higher diversity was observed.

Cashew was one the many tropical crops introduced by the Portuguese from Brazil into the African continent, in the sixteenth century as part of the Columbian Exchange Event (Havik *et al*., 2018). In Asia, cashew was introduced most likely through Goa (India), Portugal's main settlement in the East Indies in the sixteenth century (Massari, 1994). As a result of the good adaptation of trees to Indian soil conditions, cashew products were explored beyond the nut, mainly through the local fermented brew (the *feni*) made from ripened cashew apples. After India, cashew spread to South Asia in Moluccas (Indonesia) (Nair, 2010) and thereafter to the present day distribution across Asian countries as Vietnam, Philippines, Malaysia, Thailand, and Sri Lanka (Havik *et al*., 2018). Despite introductions at different timelines, at first, cashew was established as a rustic tree to cope deforestation at both African and Asian countries (Monteiro *et al*., 2017; Havik *et al*., 2018). However, at dissimilar paces, these two tropical regions have turned cashew amongst the most exported agriculture commodities. While African countries (except for Mozambique, Nigeria and Côte d'Ivoire) are amongst the major raw nuts' exporters with few processing, in Asia the structured value chain implemented allows the processing of imported raw nuts from less resourceful African countries (e.g. Guinea-Bissau, Senegal, Guinea, and Burkina Faso).

East-Timor is a small island country located in Asia, bordered by Indonesia. Agriculture is the main human activity providing subsistence to an estimated 80 % of the population (Harmadi & Gomes, 2013). Cashew has been developed as an industrial crop since the 1990's, with over 3,200 ha planted on 6,500 small farms. After the armed period with Indonesia, which nburnt several orchards, in 2008, about 125,000 cashew trees (about 800 ha) remained, growing in ten districts (Bobonaro, Manatuto, Oecussi, Cova Lima, Ainaro, Manufahi, Viqueque, Baucau, Lospalos and Dili), with a relatively low average yield per ha (270–300 kg/ha, MAFF, 2004; Peng *et al.*, 2009). Cashew varieties planted include the locally recognized variety from Indonesia, and more recently, varieties from Brazil and Australia (Peng *et al.*, 2009), to increase yield per ha. Nowadays, as part of the Timor-Leste Strategic Development Plan 2011-2030 (República Democrática de Timor-Leste, 2011), cashew is being engaged as an export commodity to increase the country agriculture remittance.

Thus, our study focuses on the assessment of the genetic diversity of cashew populations cultivated in East-Timor, using highly informative molecular markers as microsatellites, which are effective in evaluating intra-specific genetic diversity along with the evaluation of the population structuring. To accomplish such objective, eleven cashew orchards from East-Timor were assessed with 16 cashew-specific microsatellites (Croxford *et al.*, 2006) applied to screen

in-country genetic diversity. Besides, an Indonesian population was included as an outgroup, to assess any East-Timor genetic signature, along with two populations from Mozambique, working as a continental outgroup. Our work is pioneer, as no studies have been conducted in East-Timor using a highly comprehensive sampling scheme (over 11 populations) in an emerging cash crop, as cashew.

Materials & Methods

East-Timor: country and agriculture profile

East- Timor is a small island country with an estimated population of 1,183,643 (Government of East-Timor, 2015), distributed within an area of approximately 15,000 square kilometers, organized in 13 districts (**Figure 1**). The capital is Dili. Since independence in 2002, East-Timor's primary activities are agriculture, fisheries, and forests, which have reduced the share of the country's GDP from 27.8% in 2003 to 19.8% in 2015. About 225,000 ha are cultivated land area, as among these 165,000 ha are arable land (with different annual crops) and 60,000 ha are permanent crops for rice, corn, cassava, coffee, coconut and other industrial crops (MAFF, 2004). Agriculture in East-Timor is the most important economic sector. Coffee, maize and rice paddy rises as the most important staple crops (MAFF, 2004; Borges, 2018). Despite its relatively small size, East-Timor is divided in six different agro-ecological zones (ARPAPET, 1996), that shape the diverse ecology and agriculture landscapes in the country. These agro-ecological zones, in terms of territory coverage (in %) and elevation (in m), are: North Coast Lowlands (10 %, <100 m), Northern Slopes (23 %, 100-500 m), Northern Uplands (20%, > 500 m), Southern Uplands (15 %, > 500 m), Southern Slopes (21 %, 100-500 m) and South Coast Lowlands (11%, <100 m) (ARPAPET, 1996; Fox, 2003). The North coast is far drier than the south, and the mountains have more rain than the coasts. The studied districts are included in four agro-ecological regions, namely: in Northern Coast Lowlands- Manatuto (Kribas); Northern Slopes- Baucau; Northern Uplands- Bobonaro; and South Coast Lowlands- Manatuto (Natarbora), Cova Lima, Manufahi and Viqueque.

Sampling

Cashew populations were sampled from different orchards in East-Timor (11), Indonesia (1), and Mozambique (2) (**Figure 1A-C**). Each population is represented by 12-18 individual plants from an orchard (**Table 1**), with a total of 207 samples analyzed. Leaves were sampled and preserved in silica gel until further processing. Within East-Timor (**Figure 1C**), about 11 cashew orchards populations were sampled, covering a comprehensive sampling on 6 districts and cashew producing regions (**Figure 1D-I**). A population from Indonesia was also included, namely in Kefamenanu (**Figure 1B**), district of Kota Kefamenanu located in the North Central Timor Regency which borders East-Timor's Oecusse enclave, one of the few Indonesian regions that have a land border with other countries. Hence, such population would be essential for

identifying possible genetic connection between a major region of Indonesia, the only with a strong land connection with East-Timor, and East-Timor cashew populations sampled. Two Mozambican populations (**Figure 1A**) were included, thus working not only as a possible population outgroup between continental and island regions, but also due to the strong historical connection with other Portuguese colonies during 16th and 17th centuries, as it was an important continental outpost for Southeast Asia islands explorations (de Carvalho & Mendes, 2016; Havik *et al*., 2018). All DNA samples are stored at the Instituto Superior de Agronomia, University of Lisboa (Portugal) and are available upon author request.

DNA extraction

Individual leaves collected in each cashew population were used for genomic DNA (gDNA), extracted with the InnuPREP Plant DNA Kit (Analytik Jena, Germany), following manufacturer's instructions with minor modifications. About one hundred milligrams of each leaf collected in the field were grinded briefly with a mortar and a pestle in liquid nitrogen, and about 25 mg of roughly grinded leaves were used for subsequent gDNA extraction protocol, by adding 400 µl of OPT lysis solution and ground the biological material with an Eppendorf-pestle in a 1.5mL tube. After, an initial incubation at 65°C for 1 h was performed, followed by adding 100 µl of Precipitation Buffer and a 5-min incubation at room temperature; the supernatant was recovered by centrifugation at maximum speed 11.000 x g for 5 mins. The supernatant was then transferred to a Pre-Filter Receiver and centrifuged at 11.000 x g for 1 min. Subsequently, 4µl of RNAse A solution (100 mg/ml) was added and samples were incubated for 30 mins at 37°C. After RNAse treatment, 200µl of SBS binding solution was added and then centrifuged at 11.000 x g for 2 mins. To the recovered supernatant, two washing steps with 650 µl of MS washing solution was performed with centrifugations at 11.000 x g for 1 min. 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).

Microsatellite genotyping

A set of 16 cashew-specific microsatellite (SSRs) markers already available (Croxford *et al.*, 2006) were selected, for screening 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, as a reference value to be considered as an informative marker, ii) markers with high allelic diversity, iii) and 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 (Schuelke, 2000) to assess both reaction

reproducibility and/or presence of PCR artifacts upon fragment analysis (Monteiro *et al*., 2016). Each SSR was PCR amplified in a 25µl volume reaction following cycling conditions previously described in Croxford *et al*. (2006), using HotStar Taq DNA Polymerase kit (QIAGEN, Germany), as per manufacturer's instructions. Next, 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 using the Multiplex Manager software v1.2 (Holleley & Geerts, 2009). 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*. (2013). To increase genotyping accuracy, a "PIG-tail" sequence was added at the 5′ end of each of the reverse primer (Brownstein *et al.*, 1996). PCR multiplex amplifications were carried out using the QIAGEN Multiplex PCR kit (QIAGEN, Germany), following the manufacturer's protocol. In each PCR reaction in a total volume of 25µL, the following components were added: 5 µl of 5x Q-Solution, 12.5 µl of 2x QIAGEN Multiplex PCR Master Mix (which includes HotStarTaq DNA Polymerase, PCR Buffer with 6 mM $MgCl_2$, and dNTP mix), with 50–100 ηg gDNA, 2.5 ρmol of each primer Forward and Reverse and 0.15 ρmol of the tailed fluorescently labeled primers (D1–D4), and variable volume of ddH_2O up to 25 µl. 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*. (2006) 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. After, multiplex PCR products were run in a ABI 3130XL sequencer for fragment analysis, as described earlier, and SSR allele sizes were aligned with the internal size standard. To improve SSR data quality, allele callings were checked manually, and ambiguous results were set as "missing data." The resulting genotypic matrix was used for genetic diversity and population structuring analyses. The genotypic data generated from this study is freely available at FigShare in Guterres *et al.* (2022, https://doi.org/10.6084/m9.figshare.19119041.v3).

Genetic diversity analysis

Genotyping errors were assessed using MICRO-CHECKER v2.2.3 (Van Oosterhout *et al*., 2004), and estimation of null alleles frequency was done with the EM algorithm of Dempster *et al.* (1977), as implemented in FreeNA (http://www.montpellier.inra.fr/URLB/). These values were computed as described by Chapuis & Estoup (2007), 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 Microsatellite Toolkit v.3.1.1 (Park, 2001) and GenALEx 6.5 (Peakall & Smouse, 2006), respectively, as previous described (Monteiro *et al.*, 2016). Briefly, 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 v.4.5 (Rousset, 2008). 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 [$p < 0.000298$, (0.05/168)] by applying a sequential Bonferroni correction (Rice, 2013).

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.

Estimating relations using genetic distances

Distances relationships among populations were done according to Monteiro *et al.* (2016). Specifically, Cavalli-Sforza and Edward's Chord genetic distances (*DC*, Cavalli-Sforza & Edwards, 1967) using the INA method computed in FreeNA (DC^{INA}), and Nei's *D* distance (Nei, 1972) were calculated in GenALEx 6.5. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbor-Joining (NJ) trees were produced using package *ape* v3.4. (Paradis *et al.*, 2004) for R v4.1.0 (R Core Team, 2021) based on 10,000 bootstraps values, assessed by aboot function from *poppr* v2.1.0. package (Kamvar *et al.*, 2014). Trees were further edited in FigTree v1.4.2 (Rambaut, 2014). Distances relationship among populations was done in two separate approaches: first using the three countries (East-Timor, Indonesia and Mozambique) populations and, second, by excluding continental populations from Mozambique.

Analysis of molecular variance (AMOVA)

An Analysis of Molecular Variance (AMOVA, Weir & Cockerham, 1984; Hill, 1996) was done with ARLEQUIN v3.5.1.3 (Excoffier & Lischer, 2010) to assess the hierarchical distribution of genetic variation on the populations analysed. Significance was assessed after 1000 permutations. Two three-levels AMOVAs were pursued: one using each of the three countries populations (MZ=2; IND=1; ET=11) as a group, and the second considering only East-Timor and Indonesia as groups. In each AMOVA, the total variance was partitioned into components to

account for pairwise differences between groups [V_a, (1) MZ vs ET vs IND; (2) ET vs IND populations], 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 & Cockerham (1984).

Individual-based clustering analyses

Identification of genetically distinct clusters were done under two different methodologies, as previously described in Monteiro *et al.* (2016), by: a Bayesian clustering analysis using STRUCTURE (Pritchard *et al.*, 2000) and a multivariate analysis method, DAPC (Jombart & Balloux, 2009). These two different analyses were done because STRUCTURE uses allele frequency and LD information from the dataset directly assuming HWE equilibrium, while DAPC does not considers a particular population genetics model outlining the genetic differentiation between and within groups (Jombart & Balloux, 2009).

Overall, individual-based clustering analysis were performed under two datasets: one including East-Timor, Indonesia, and Mozambique, and the second focused in East Timor and Indonesia. 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 East-Timor vs Indonesia vs Mozambique, 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 14 with 10 runs computed for each K. To the second approach, using East-Timor vs Indonesia populations, the same settings were followed by configuring K-values from 1 to 12 with 10 runs in each K. StructureHarvester v0.6.94 (Earl & Bridgett, 2012) was then used to calculate ΔK *ad hoc* statistics from Evanno *et al.* (2005), for estimating the most likely K-value. CLUMPP v1.1.2 (Jakobsson & Rosenberg, 2007) was used to average replicate runs for the selected K-value, for accounting problems with multimodality and label switching between iterations of STRUCTURE runs. CLUMPP results were then plotted with DISTRUCT v1.1 (Rosenberg, 2004).

DAPC was implemented in R (R Core Team, 2021) using *adegenet* v1.3.1 package (Jombart, 2008) using the dataset relative frequency of the alleles, since presence/absence data may not be fully informative, and, thus, may 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). DAPC script generated by our analysis is available at Figshare (Barros *et al.*, 2022).

Results

SSRs genotyping and statistics

All 16 SSRs were tested in singleplex reactions at the estimated optimal annealing temperature, and only after this initial quality assessment, SSRs markers were grouped into 4-plex reactions (**Table 2**). For the 16 SSRs loci, 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 SSRs screened. In 60 of the 224 locus-comparisons, null alleles frequencies were higher than 0.20 in markers mAoR33, mAoR12, mAoR29 and mAoR41, thus subsequent analyses were made with the remaining 12 SSRs (**Table 3**). Deviations from Hardy–Weinberg Equilibrium (HWE) were observed in most loci except for mAoR48, mAoR42, mAoR3, mAoR17, mAoR35, with 84 locus-population combinations statistically significant ($p \geq 0.05$); while after sequential Bonferroni correction only two loci (mAoR3 and mAoR35,) displayed significant deviations, matching 33 of the 168 locus population combinations (Supplementary Table S1). All 12 loci were in linkage equilibrium after Bonferroni correction, thus being non-correlated, and alleles independently segregated and inherited (Supplementary Table S2). Negative fixation index (*F*) estimates were observed in two loci, mAoR17 (-0.03) and mAoR7 (-0.04) (**Table 3**), which can reflect more heterozygotes than expected or other population structure complexities.

Genetic diversity estimates

Overall, a total of 157 alleles were detected in the 207 individuals analyzed (**Table 3**). All loci screened were polymorphic. The total number of alleles per locus ranged from 4 (mAoR2) to 25 (mAoR3) with an average of 13.08 alleles per locus (**Table 3**). Overall, Polymorphic Information Content (PIC) values ranged from 0.40 (mAoR16) to 0.83 (mAoR17) with a mean value of 0.67 (**Table 3**). In our 12-loci dataset, observed heterozygosity (H_o) varied from 0.25 (mAoR35) to 0.73 (mAoR17) with a mean of 0.40 (**Table 3**); and expected heterozygosity (H_e) varied between 0.45 (mAoR16) and 0.84 (mAoR17).

By performing a population genetic diversity analysis presented (**Table 4**), East-Timor presented 10.67 alleles in average and the lowest value was from Indonesia with 3.42. Expected heterozigosity (H_e) was 0.67, 0.56 and 0.71, respectively from East-Timor, Indonesia and Mozambique, while H_o was 0.51, 0.47 and 0.54, accordingly. All populations presented a *F* positive variating from 0.12-0.25. The populations with the highest number of alleles were observed in East-Timor, most precisely in the population of ETK which presents in average 5.17 alleles, followed by ETR3 with 5.08 alleles, while ETBAT and ETV with 2.42 and 2.25 alleles, respectively. The allele numbers of the SSRs by comparing each of the populations screened, enable to depict which population harbors more allelic diversity. When analyzing this parameter, we can see that ETK and ETTR3 are the populations that present greater allelic diversity (Na, **Table 4**). Concerning the expected heterozygosity (H_e), the highest was obtained in ETK population, and the lowest in ETV population. When comparing the observed heterozygosity (H_o), the highest was obtained for the population ETTR1, and the lowest value in ETFA. Fixation index (*F*) was positive in all populations except for ETSAN and ETV, the positive

values in the remaining populations may indicate genetic stability and a higher rate of inbreeding. The absence or existence of private alleles is important to account for, since it can allow to identify a specific genetic signature, as private alleles are alleles that are only detected in this population. East-Timor displayed a high number of private alleles (6), with Mozambique harboring 2, while in Indonesia populations no private alleles were detected.

Population structuring analyses

Estimating relations among populations through genetic distances

UPGMA and NJ trees were built using Nei's *D* and DC^{INA} (*FreeNA*) genetic distances across accessions screened for East-Timor, Indonesia, and Mozambique (Supplementary Table S3), and an analysis narrowed to East-Timor and Indonesia (Supplementary Table S4) was done. Under the two analysis approaches, similar tree topology's structure was observed with both Nei's *D* (Supplementary Figure S1, S2) and DC^{INA} (**Figures 2 and 3**) matrices, thus indicating a reliable topology regardless of the different genetic distance's algorithms used. As such, only DC^{INA} distances matrices-derived trees are presented in Figure 2 for East-Timor, Indonesia, and Mozambique analyses and in Figure 3 for analyses narrowed to East-Timor and Indonesia. In the UPGMA (**Figure 2A**) and NJ (**Figure 2B**) derived trees, three clusters depicted: one cluster (I) that includes ETTR populations (ETTR1-3) from Baucau, Viqueque (ETV), Cova Lima (ETSU) and ETK population from Manatuto, grouped with Indonesia (IND); a second cluster (II) comprising the remaining East-Timor populations from Manatuto (ETNA), all populations from Bobonaro district (ETMA, ETBAT, ETSAN) and Manufahi population (ETFA), and the third (III) representing the Mozambican populations (MZB and MZD). This result support two different genetic clusters within East-Timor: one including only East-Timor populations (Cluster III) and the other with a high genetic membership with the Indonesian population (Cluster II). With the NJ tree (**Figure 2B**), the dendrogram derived from DC^{INA} distance matrix also presented three different clusters: cluster I shows the same grouping with Mozambican populations as observed in the UPGMA (**Figure 2A**), the second cluster is configured by Bacau, Viqueque (ETV) and Cova Lima (ETSU) populations with Indonesia, and the third cluster include Bobonaro populations (ETMA, ETBAT and ETSAN) and the remaining East-Timor populations (ETFA, ETNA and ETK).

When observing trees without Mozambican populations (**Figure 3**), the two clusters are similar to the observed for clusters I and II in Figure 2.

Conversely to UPGMA derived dendrograms using the whole populations dataset (**Figure 2A**, Suppl. Fig. S1A) and the East-Timor/Indonesia dataset (**Figure 3A**, Suppl. Fig. S2B), the two genetic distance algorithms produced dissimilar NJ-generated trees (Nei's *D* distance, **Figure 3B,** Supplementary Fig. S2A, S2B), indicative of a complex population structuring.

Analysis of molecular variance

When grouping countries dataset (MZ, IND, ET), AMOVA results showed that molecular variation was mainly found within individuals (76%), whereas variation among populations and among individuals within population explained 13% of the total genetic differentiation, in both cases (**Table 5**). Regarding East-Timor vs. Indonesia dataset, a similar scenario was depicted, with genetic differentiation within individuals (73%) also higher than among individuals (14%) or among populations (13%).

Individual-based clustering using Bayesian and a multivariate discriminant analysis to uncover population structure

Two different approaches were done: 1) covering all populations from East-Timor, Indonesia, and Mozambique, and 2) excluding Mozambique, to uncover more in-depth individual clustering within East-Timor populations, using Indonesia population as outgroup.

In the first approach, STRUCTURE was run considering the highest range of clusters conceivable (K = 1–15). This analysis assigned K = 5 as the optimal number of groups based on Evanno *et al*. (2005) ΔK method (Supplementary Figure S3). According to the results obtained from this first approach with all populations from East-Timor, Indonesia and Mozambique, in K = 5, Mozambique was grouped in a single cluster (orange cluster) (**Figure 4A**), Manatuto populations are grouped in 2 principal clusters (brown and red), Baucau populations mostly in one cluster (blue) except for ETTR3 that have a mix of all clusters except with Mozambican cluster; Cova Lima is only grouped in a single cluster (green cluster) as well as Bobonaro populations grouped in the red cluster. Manufahi population is mostly in a red cluster sharing genetic flow with Viqueque and Indonesia, which are grouped in two clusters (brown) (**Figure 4A**).

DAPC analysis was made without any *a priori* group assignment. For the first dataset, the clustering analysis determined that K = 10 was the one with the best combination of mean and 95% CI of BIC (Supplementary Figure S4). However, we do not see an "elbow" effect, rather a considerable plateau in the number of clusters with a significant overlap among confidence intervals for the nearby number of clusters. To make the results comparable with the ones produced by STRUCTURE, we decided to use the optimal number of clusters for the later analysis, K = 5 (**Figure 4B**). Using the function *xvalDAPC*, 40 PCs (highest successful assignment – 88.2 %, with the lowest mean squared error – 0.178) were retained and 4 Discriminant Functions, thus conserving 87.5% of variance (Supplementary Figure S5).

Cross validation using the *xvalDapc* function outcome the number of PCA axes retained against the proportion of successful outcome prediction, which allowed retaining 40 PCA axes (considering the highest successful assignment- 88%, with the lowest mean squared error, MSE-17%) and 2 Discriminant Functions (explaining 87.5% of cumulative variance), for inferring the 5 genetic clusters (Supplementary Figure S6). When displaying loading plots from both Discriminant Functions, one can determine which variables (i.e., alleles/loci) contributed the most for the five-clustering assemblage. Considering both DAPC results (K = 9 and K = 10), the

same variables are responsible for cluster assemblage (i.e., 172 (mAoR6), 328 (mAoR17), 174 (mAoR35), 170 (mAoR47), which highlight the importance of these alleles for cluster discrimination (Supplementary Figure S6 and S9).

For the Bayesian analysis of the second approach, which included East-Timor and Indonesia populations, STRUCTURE was run considering the highest range of clusters conceivable (K = 1–13). This analysis assigned K= 2 as the optimal number of groups based on Evanno *et al*. (2005) method (Supplementary Figure S8), with no alternative ideal K. Considering the high ΔK-values displayed for K= 2 (**Figure 5),** two clusters were observed, the first cluster, in green, with Baucau, Cova Lima, Manatuto and Indonesia; and the second cluster, in red color, with Bobonaro, Manufahi and Viqueque populations **(Figure 5A**). For the DAPC analysis and based on the best number of clusters from STRUCTURE, K = 2 (**Figure 5B**), the cross-validation allowed retaining 20 PCA axes (considering the highest successful assignment- 93.5%, with the lowest mean squared error, MSE - 10%) and 2 Discriminant Functions (explaining 96.4% of cumulative variance), for inferring the two genetic clusters. For K = 2, there are differences between the STRUCTURE and DAPC analyses, a different admixture scenario is depicted, namely in DAPC genetic diversity in Baucau which is not shared with Indonesia, while Bobonaro, Manufahi and Viqueque populations has a common allelic diversity to Indonesia (**Figure 5B**), contrarywise to STRUCTURE optimal clustering (**Figure 5A**). These analytical inconsistencies may be related to different pre-requisites associated to both methods: STRUCTURE assumes that markers are not linked and that populations are panmictic (Pritchard *et al.*, 2000), while DAPC are more convenient approaches for populations that are clonal or partially clonal. In this case, STRUCTURE provides a more realist observation of the genetic diversity of cashew populations, given the panmictic nature and absence of clonal populations even in cashew varieties.

Discussion

A comprehensive sampling of cashew orchards in East-Timor was conducted by collecting 11 populations from major cashew producing districts in the country, to assess and characterize the genetic diversity and population structuring. A total of 207 individuals belonging to 14 cashew populations from three tropical countries (East-Timor, Mozambique, and Indonesia), using Indonesia and Mozambique populations (n=3) as outgroups, were screened using 16 cashew specific microsatellites.

Cashew diversity assessment

Fourteen different cashew populations were first genotyped with 16 SSRs, and SSRs quality assessment was done (**Table 3**). Four loci (mAoR12, mAoR33, mAoR41, mAoR29) were

discarded due an excess of null alleles in almost all populations (null allele frequency > 0.20), despite being polymorphic across the cashew populations analyzed and thus being an informative marker for future diversity analysis. Thus, subsequent genetic diversity indices and population structuring analyses were performed using 12 loci. PIC-values obtained were high (average PIC = 0.65) which indicates high informativeness, and compared with a recent study using 21 cashew SSR (cSSR, Savadi *et al.*, 2020) values were higher (0.33 average PIC-value). Differences on PIC-values from our study might be due to distinct plant material sources compared to former reports, which may influence the number of alleles detected at each SSR locus, though a potential influence of the lower number of SSRs loci used should not be eliminated. When analyzing null alleles presence and their effect on population structure, only 60 of the 224 locus-comparisons harbored null alleles with a frequency higher than 0.20. Overall, based on the results of the preceding analysis, it is possible to predict that the SSRs loci used are suitable for downstream genetic diversity analysis.

Genetic diversity observed in our study is high when compared to other cashew studies. Within East-Timor, more alleles were found, which can be explained partly by a higher sampling effort but also may be indicative of a reservoir of alleles associated to traits conferring tolerance to various biotic and abiotic stresses or to local adaptation conditions. Our results show a higher allelic richness in East-Timor populations (Na=10.67, **Table 4**) than in Mozambique (Na=3.42) and in Indonesia (Na=3.42), which highlights the high genetic diversity in East-Timor. Nevertheless, only with the inclusion of more populations from Mozambique and Indonesia could corroborate if Eat-Timor higher allelic richness. Within East-Timor, ETK population from Manatuto district display the highest allelic richness (Na=5.17), followed by ETTR3 from Baucau district, with ETV from Viqueque having the lowest allelic diversity obtained (Na=2.25, Table 4). Since in Viqueque district cashew orchards are less frequent in comparison with Baucau and Manatuto districts, where implementation of several orchards is occurring over the last 20 years, less diversity is observed. The populations analyzed might reflect long-term genetic diversity that can be exploited in a breeding program to improve yield and nut quality (Kouakou *et al.*, 2020). Moreover, identifying trees with high allelic richness is necessary for conserving cashew germplasm (Bataillon *et al.*, 1996). The number of private alleles found within accessions is an important diversity measurement since these alleles represent genotypic-specific allelic build-up. Contrasting with previous SSRs studies in cashew (Savadi *et al.*, 2020), private alleles were detected in our study, which may allow for the configuration of a unique genetic signature of our populations, either by different allele frequencies or by unique alleles in each population. This is of major importance in an agricultural crop like cashew, where nuts are, frequently, exported and processed outside the producing countries, as it may lead to the valorization of a domestic product with geographical origin.

The information regarding genetic diversity and population structuring in cashew is the primary impetus for this work; from 2006 to 2020, only a few papers were published especially in Southeast Asia and West Africa region. In comparison with a study done in Benin (West Africa),

our study show a much higher genetic diversity, where a lower genetic diversity (Shannon index = 0.04) was observed in 60 cashew morphotypes using eight SSR markers (Sika *et al*., 2013). In Brazil, wild Brazilian populations of cashew were studied (dos Santos *et al*., 2019), 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. In Côte d'Ivoire, genetic diversity of cashew was studied using SSRs and revealed an overall heterozygosity deficit and a high intra-population genetic diversity among the screened cashew populations (Kouakou *et al*., 2020), which is in accordance with our results in AMOVA where most genetic diversity is depicted within populations. Our results also agree with a recent study from Burkina Faso, where a substantial genetic diversity was observed across the 18 cashew accessions screened with 4 SSRs (Moumouni *et al.*, 2022).

Population structuring in East-Timor

AMOVA showed that most of the genetic diversity lies within populations with little diversity present among populations or between populations in each country. The large proportion of diversity was found within accessions for the two types of groupings (countries MZ vs IND vs ET and ET vs IND) suggesting a high gene flow between populations, which is in accordance with previous studies (Freitas & Paxton, 1996). Cashew is an allogamous species favoring cross-fertilization (Freitas & Paxton, 1996), thus allowing intraspecific hybridizations and heightening genetic variation. Overall, outcrossing species tend to have higher genetic variation within-populations, whereas selfing species or species with a mixed mating system are often genetically less variable (Nybom, 2004). Since cashew is an outcrossing, negative to low inbreeding coefficients (*F*) were expected, which agrees with former studies (Freitas & Paxton, 1996; Layek *et al*., 2021).

Moreover, population structuring revealead that genetic diversity scattering does not follows a clear geographic trend, despite a well-defined clustering observed between Mozambican populations with the remaining East-Timorese/Indonesian populations. The unique clustering attributed to Mozambican populations may be related to in-country cashew varieties selection, different from the ones used by farmers in both Indonesia and East-Timor, and also due to being a continental region in comparison to the geographical isolation of the island countries, Indonesia and East-Timor. When narrowing to East-Timor, a complex population structuring is observed which is linked to a district- associated genetic diversity. Within the country, Viqueque, Manufahi and Bobonaro districts display a unique allelic diversity; while Baucau, Cova Lima and Manatuto has a high genetic similarity with the Indonesian population. These dissimilar results when including vs excluding Mozambican populations, highlights the complex genetic diversity of cashew in the context of a continental (Mozambique) country where orchards have been implemented at long-term with improved varieties, while in Indonesia and East-Timor surely a different historical context with few progression into improved varieties, is being

applied. These results are in accordance with previous results in India (Archak *et al*., 2009), where cashew genetic diversity lies within geographical populations and also the sharing of allele frequencies among populations does not translates into an in-country population structuring.

The clustering of cashew populations in this study had an uncommon, yet existing, relationship with geographical region under a district-wise distribution, which is contrary to previous reports in India (Archak *et al*., 2009), where no relationship with the geographic region was observed. One of the reasons for a genetic diversity pattern according to district sampled may be related to current cashew orchards in each district been engaged into different cashew varieties according to seeds availability and farmers' preferences, many of which associated to better trees local performance and yield. Considering that few cashew varieties have been introduced in the country, this genetic diversity distribution may be associated to inter-exchange of seed material adapted to similar ecological conditions. Despite a high genetic diversity attributed to the high heterozygosity, allogamous nature and high gene flow found in cashew (Mitchell & Mori, 1987; Borges, 2018), also obtained in our study, in certain East-Timor districts as Bobonaro, Viqueque and Manufahi, the dissimilar genetic clustering from the remaining districts suggests a diverse genetic build-up, which could be attributed to different cashew varieties being planted.

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 (Monteiro *et al.*, 2016). Negative values denote excess of heterozygosity, due to negative assortative mating, or selection for heterozygotes. Overall, positive *F*-values were observed across all populations (**Table 4**), thus revealing that populations are at or near Hardy–Weinberg equilibrium, further supported by the lower observed heterozygosity values against the expected under HWE (**Table 4**). Among the cashew populations collected (associated with different geographic regions), inbreeding coefficient was lowest in the northern Bobonaro district (ETSAN, *F*=-0.04; ETMA, *F*=0.14; ETBAT, *F*=0.01, **Table 4**) and southern Viqueque district (ETV, *F*= -0.09), possibly because growers from both regions easily exchange seeds with the other regions. In contrast, Cova Lima district (south) showed the highest inbreeding coefficient (*F*=0.26). Both Bobonaro and Viqueque districts display a significant share of exchange of planting material: in Bobonaro from Indonesia and in Viqueque from Australia. Thus, its genetic richness might benefited from different introduced seeds and more varied germplasm imported than in other regions. In the studied populations, positive *F* values was observed which may indicate some level of inbreeding. The difference between expected heterozygosity and observed heterozygosity might be due to evolutionary factors (as inbreeding) or internal genetic factors (such as gene incompatibility). In addition, cashew grower's seeds preference for establishing new orchards might be another explanation. Indeed, when establishing new cashew orchards, some producers used seeds from a single tree with good traits (preferentially high-yield and large-nuts). Moreover, according to Ahmed & Saddi (2012), the genetic makeup of a given population can vary over time in response to

evolutionary forces that in turn affect the heterozygosity of the population relative to the Hardy-Weinberg equilibrium.

Considering the 11 populations studied from East-Timor, population differentiation (F_{ST}= 0.129, Table 5) was relatively moderate in comparison to other studies in Côte d'Ivoire (Kouakou *et al*., 2020) where a low differentiation was indicative of a common origin of Ivorian cashew trees. However, it cannot be discarded that moderate differentiation could be associated to a lesser number of molecular markers used in our study (12 SSRs vs 18 SSRs in Kouakou *et al*. (2020)). In India, a similarly low genetic diversity among cashew trees was associated to a relatively recent introduction in the country (Archak *et al*., 2009), which is also reported in other countries (e.g. Benin, Kouami *et al.*, 2020), except in Brazil. In East-Timor case, two different scenarios of cashew genetic diversity are conceivable, namely: i) possible multiple introductions from Bobonaro and Viqueque districts, these regions being regarded as cashew introduction hotspots in East-Timor; and ii) introduction of new cashew varieties on the remaining districts according to farmers preference.

Also, the moderate genetic variability among the populations screened could be due to the relatively recent introductions into East-Timor on evolutionary timescale and the allogamous nature of cashew resulting in the high gene flow and exchange of genetic material. The unexplored yet high genetically diverse accessions from East-Timor could be used in future cashew breeding programs to improve yield, quality, and other traits.

East-Timor as an explored yet important source of cashew genetic diversity

The wide distribution of cashew in its primary center of diversity, in Brazil, has been attributed to water currents in which the mature fruit will float in addition to the role played by bats in seed dispersal (Johnson, 1973). However, outside its center of origin, cashew distribution, since its introduction in India, has been attributed to anthropogenic efforts, rather than through natural means alone, (Archak *et al*., 2009). Considering this important premise, the present genetic diversity analysis is discussed accordingly. The extent and distribution of genetic diversity as revealed by the present study provide some clues of cashew introduction mode and expansion in East-Timor. The existence of substantial overall genetic diversity and populations' grouping into two distinct genetic groups in East-Timor point to multiple events of introduction comprising different founder populations. These results are in accordance with a recent report that suggests several introductions in Burkina Faso, and then, well performed cashew cultivars were disseminated through the same route across producing areas, without a single cultivar selection to a particular producing region (Moumouni *et al.*, 2022). If the introduction occurred as a one-time event, founder effect would have been reflected in low genetic diversity. Nowadays, cashew expansion in East-Timor has been promoted mainly through the introduction of different varieties from Brazil, Indonesia, Australia and a so-called "native" cashew. According to the information obtained from the National Directorate of Industrial Crops and Agribusiness

(Ministry of Agriculture, Forestry and Fisheries (MAFF) of East-Timor), the Portuguese brought the cashew to East-Timor (ET) in the 18th century, where it was planted as ornamental plants in various districts. Since the establishment of cashew as an industrial crop in the 1990's (Buss & Ferreira, 2010; Odete *et al*., 2017), no proper orchards management have been applied in terms of varieties introduced. In Cova Lima district, cashew orchards were implemented in 1990s and few new cashew varieties have been introduced. At Manatuto and Baucau districts, cashew orchards were planted with accessions retrieving higher yield and better-adapted to local agro-ecological conditions. During fieldwork at Bobonaro and Viqueque districts, land farmers reported that several cashew accessions were being introduced, namely from Indonesia and Australia, hence such a different genetic clustering may be associated to farmers preferences thus shaping current East-Timor cashew genetic diversity panorama.

Considering that India was an important country, where cashew was first introduced as a commercial crop in Asia (Singh, 2018), future studies should include Indian cashew populations aiming to explain the genetic diversity within East-Timor and assist on the historical introduction in the country. As also reported in India (Archak *et al*., 2009), in East-Timor a relatively significant genetic diversity for an introduced species was observed, thus supporting the possibility of cashew being introduced repeatedly over time. Contrary to our data, a significant level of redundancy (homogenous group) was reported within Nigerian cashew germplasm (Aliyu, 2012), thus highlighting a narrow genetic diversity. Our results in East-Timor are of major importance, since the country aims to invest in cashew nut exports market, and thus assessing genetic diversity will allow detecting the existence of an East-Timor genetic signature that will acknowledge the valorization of cashew nuts. Thus, this should be taken in consideration by the local Government, since East-Timor would compete directly with countries such as India and Vietnam, which are among the major world cashew nuts exporters.

This study provides useful information on genetic diversity assessment on cashew populations in a largely understudy country, where cashew nuts is becoming an important export-oriented crop, and thus the genetic diversity build-up obtained in our study, point out to a cashew genetic diversity hotspot. Cashew has been implemented under a monoculture system and land cropping area has been increasing to meet global market needs (Monteiro *et al.*, 2017). This, together with a scenario of rising potential of pest and diseases (Monteiro *et al.*, 2015; 2017) and the current narrow genetic diversity of cashew orchards is a major concern to its future sustainable production. Thus, the incorporation of genetic resources with new genetic diversity should be foreseen towards the development of varieties with improved agronomic traits, such as higher yield, biotic and abiotic stress tolerance. Also the increase of the genetic diversity of on-farm orchards at the short run and the identification of genetic resources for the development of cashew genetic management should not be neglected.

Conclusions

Our results show a higher allelic richness in East-Timor populations than in Mozambique and Indonesia, reinforced by the presence of private alleles. Genetic diversity was observed within populations, in accordance with former studies. Population structuring revealed that the genetic diversity seems to follow a geographic trend, with a well-defined cluster observed in Mozambican populations and other in East-Timor/Indonesia. Within East-Timor, a district-associated genetic diversity clustering into two genetic groups was seen, which may point to multiple events of cashew introduction. This study provides useful information on genetic diversity hotspots, which can be used to improve genetics and characterize new types in a future breeding effort. East-Timor is one of the countries where cashew nuts are becoming an important tradeable agriculture product, and the genetic diversity build-up obtained in our study point out to cashew agrobiodiversity hotspots. The findings of this study are also applicable to the development or preservation of genetic resources for cashew in a understudied country as East-Timor, towards the development of a management and conservation plan of cashew genetic resources. Considering that East-Timor is engaging into cashew as an important crop, the genetic diversity build-up obtained would be important for assessing an in-country genetic signature to increase the crop market value, and thus competitiveness.

Acknowledgements

The authors would like to acknowledge all farmers from East-Timor for their significant contribution during fieldwork, the Ministry of Agriculture, Forestry and Fisheries (MAFF) of East-Timor for logistics support, and Sílvia Catarino for map representation in Figure 1.

Author Contributions

Conceptualization, F.M. and M.M.R.; methodology, F.M, L.G., J.B. and A.B.B.; formal analysis, J.B., A.B.B., A.C. and F.M.; investigation, L.G. and J.B.; writing—original draft preparation, L.G., J.B. and F.M.; writing—review and editing, A.B.B., A.C., M.C.D. and M.M.R.; supervision, M.M.R. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by FCT – Fundação para a Ciência e a Tecnologia, I.P. under the project GenoCash (PTDC/ASP-AGR/0760/2020). Fellowships were funded by Portuguese National Funds through FCT, Portugal: SFRH/BD/135358/2017 to L.G. and SFRH/BD/135360/2017 to A.C., and research units: UID/AGR/04129/2020 (LEAF); UID/BIA/00329/2020 (cE3c). Fellowship to J.B. was funded by FAO/UN (TCP/GBS/3801). The APC was funded by Fundacão para a Ciência e Tecnologia (FCT) under the GenoCash project (PTDC/ASP-AGR/0760/2020).

691

Data Availability Statement

693 Data is contained within the article or supplementary material. Genotypic matrix
694 (https://doi.org/10.6084/m9.figshare.19119041.v3) and DAPC script
695 (https://doi.org/10.6084/m9.figshare.19117889.v3) are available at the online figshare repository.

Conflicts of Interest

697 The authors declare that they have no competing interests.

698

References

700

701 Aliyu OM, Awopetu JA. 2007. Multivariate analysis of cashew (*Anacardium occidentale* L.)
702 germplasm in Nigeria. *Silvae Genetica* 56:170–179. DOI: 10.1515/sg-2007-0026.
703 Aliyu OM. 2012. Chapter 9: Genetic Diversity of Nigerian Cashew Germplasm. *In* Genetic
704 Diversity in Plants, Çalişkan M. (Ed.). IntechOpen. Pp 163-184.
705 https://doi.org/10.5772/32892.
706 ARPAPET. 1996. Agro-climatic Zones of East Timor. (Lindsay Evans, April, 1996). Indonesia-
707 Australia Development Cooperation, Agricultural and Regional Planning Assistance
708 Program East Timor, Kantor Wilayah Departmen Pertanian Propinsi Timor Timur, Jalan
709 Estrada de Balide, Dili, Timor Timur 88112, Indonesia.
710 Archak S, Gaikwad AB, Swamy KRM, Karihaloo JL. 2009. Genetic analysis and historical
711 perspective of cashew (*Anacardium occidentale* L.) introduction into India. *Genome*
712 52:222–230. DOI: 10.1139/G08-119.
713 Barros A, Barnabé J, Guterres L, Monteiro F. 2022. Script for performing DAPC analysis
714 applied to the study of population structure and genetic diversity in cashew from East-
715 Timor. *Figshare Software.* DOI: https://doi.org/10.6084/m9.figshare.19117889.v3.
716 Bataillon M, David J L, & Schoen D J. 1996. Neutral genetic markers and conservation genetics:
717 simulated germplasm collections. *Genetics*, 144 (1), 409-417.
718 Borges D. 2018. Cultures du Timor-Oriental: processus d'objectification. *Plural Pluriel - revue*
719 *des cultures de langue portugaise*. (Avalaible on
720 https://www.pluralpluriel.org/index.php/revue/issue/view/16).
721 Brownstein MJ, Carpten JD, Smith JR. 1996. Modulation of non-templated nucleotide addition
722 by Taq DNA polymerase: Primer modifications that facilitate genotyping. *BioTechniques*
723 20:1004–1010. DOI: 10.2144/96206st01.
724 Buss PM, Ferreira JR. 2010. Diplomacia da saúde e cooperação Sul-Sul: as experiências da
725 Unasul saúde e do Plano Estratégico de Cooperação em Saúde da Comunidade de Países de
726 Língua Portuguesa (CPLP). *Reciis* 4:106–118. DOI: 10.3395/reciis.v4i1.351pt.
727 Cavalli-Sforza LL, Edwards AW. 1967. Phylogenetic analysis. Models and estimation
728 procedures. *American Journal of Human Genetics* 19:233–257. DOI: 10.2307/2406616.
729 Sika KC, Adoukonou-Sagbadja H, Ahoton LE, Adebo I, Adigoun FA, Saidou, A & Baba-
730 Moussa, L. 2013. Indigenous knowledge and traditional management of cashew
731 (*Anacardium occidentale* L.) genetic resources in Benin. J. Exp. Biol. Agric. Sci, 1(5), 375-
732 382.
733 Chapuis MP, Estoup A. 2007. Microsatellite null alleles and estimation of population

differentiation. *Molecular Biology and Evolution* 24:621–631. DOI: 10.1093/molbev/msl191.

Chipojola FM, Mwase WF, Kwapata MB, Bokosi JM, Joyce P, Maliro MF. 2009. Morphological characterization of cashew (*Anacardium occidentale* L.) in four populations in Malawi. *African Journal of Biotechnology* 8:5173–5181.

Croxford AE, Robson M, Wilkinson MJ. 2006. Characterization and PCR multiplexing of polymorphic microsatellite loci in cashew (*Anacardium occidentale* L.) and their cross-species utilization. *Molecular Ecology Notes* 6:249–251. DOI: 10.1111/j.1471-8286.2005.01208.x.

Culley TM, Stamper TI, Stokes RL, Brzyski JR, Hardiman NA, Klooster MR, Merritt BJ. 2013. An Efficient Technique for Primer Development and Application that Integrates Fluorescent Labeling and Multiplex PCR. *Applications in Plant Sciences* 1:1300027. DOI: 10.3732/apps.1300027.

de Carvalho BRP, Mendes H. 2016. Cashew chain value in Guiné-Bissau: Challenges and contributions for food security: A case study for Guiné-Bissau. *International Journal on Food System Dynamics* 7:1–13. DOI: 10.18461/ijfsd.v7i1.711.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39, 1–38.

dos Santos JO, Mayo SJ, Bittencourt CB, de Andrade IM. 2019. Genetic diversity in wild populations of the restinga ecotype of the cashew (*Anacardium occidentale* L.) in coastal Piauí, Brazil. *Plant Systematics and Evolution* 305:913–924. DOI: 10.1007/s00606-019-01611-4.

Earl DA, Bridgett M. 2012. STRUCTURE HARVESTER : a website and program for visualizing STRUCTURE output and implementing the Evanno method. :359–361. DOI: 10.1007/s12686-011-9548-7.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology* 14:2611–2620. DOI: 10.1111/j.1365-294X.2005.02553.x.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10:564–567. DOI: 10.1111/j.1755-0998.2010.02847.x.

Freitas BM, Paxton RJ. 1996. The role of wind and insects in cashew (*Anacardium occidentale* L.) pollination in NE Brazil. *Journal of Agricultural Science* 126:319–326. DOI: 10.1017/s0021859600074876.

Fox JJ. 2003. Drawing from the past to prepare for the future: responding to the challenges of food security in East Timor. *In* Agriculture: New Directions for a New Nation — East Timor (Timor-Leste), da Costa H, Piggin C, Fox J, da Cruz CJ (eds.). *Proceedings of workshop 1–3 October 2002, Dili, East Timor ACIAR Proceedings* 113: 105-114.

Government of East-Timor. 2015. Population and Housing Census 2015: Preliminary Results. Direcção Geral de Estatística de East-Timor. Pp 1- 39. http://www.statistics.gov.tl/wp-content/uploads/2015/10/1-Preliminary-Results-4-Printing-Company-19102015.pdf.

Guterres L, Barnabé J, Monteiro F. 2022. Genotypic matrix for studying the cashew population structure and genetic diversity in East-Timor. *Figshare Dataset*. DOI: https://doi.org/10.6084/m9.figshare.19119041.v3.

Harmadi SHB & Gomes RA. 2013. Developing Timor-Leste's Non-Oil Economy: Challenges and Prospects. Journal of Southeast Asian Economies, 30(3), 309–321.

http://www.jstor.org/stable/43264687
Havik PJ, Monteiro F, Catarino S, Correia AM, Catarino L, Romeiras MM. 2018. Agro-economic transitions in Guinea-Bissau (West Africa): Historical trends and current insights. *Sustainability* 10:1–19. DOI: 10.3390/su10103408.
Hill W G. 1996. Genetic Data Analysis II. By Bruce S. Weir, Sunderland, Massachusetts. Sinauer Associates, Inc. 445 pages. ISBN 0-87893-902-4. Genetics Research, 68(2), 187-187.
Holleley CE, Geerts PG. 2009. Multiplex Manager 1.0: A cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques* 46:511–517. DOI: 10.2144/000113156.
(International Nut and Dried Fruit Council Foundation. 2020. Nuts & Dried Fruits Statistical Yearbook 2019 / 2020. Pp 1-80. Available at https://www.nutfruit.org/files/tech/1587539172_INC_Statistical_Yearbook_2019-2020.pdf
Jakobsson M, Rosenberg NA. 2007. CLUMPP : a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. 23:1801–1806. DOI: 10.1093/bioinformatics/btm233.
Johnson JW. 1973. The botany, origin, and spread of the cashew *Anacardium occidentale* L. *Journal of Endodontics* 1:1–7. DOI: 10.1016/S0099-2399(06)81513-X.
Jombart T, Balloux F. 2009. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *PLoS Computational Biology* 5. DOI: 10.1371/journal.pcbi.1000455.
Jombart T. 2008. *Adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403–1405. doi: 10.1093/bioinformatics/btn129
Kamvar ZN, Tabima JF, Grünwald NJ. 2014. *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.
Kouakou CK, Adopo AN, Djaha AJB, N'da DP, N'da HA, Bi IAZ, Koffi KK, Djidji H, Minhibo MY, Dosso M, N'guessan AÉ. 2020. Genetic characterization of promising high-yielding cashew (*Anacardium occidentale* L.) cultivars from Côte d'Ivoire. *Biotechnology, Agronomy and Society and Environment* 24:46–58. DOI: 10.25518/1780-4507.18464.
Kouami N'D, Adolphe A, Hubert A-S, Barnabas W, Raphiou M, SaliouB, Vinou Yémalin Alfred VY. 2020. Yield and Nut Quality of 29 Cashew Mother Trees (*Anacardium occidentale* L) Established At the Germplasm of Ouoghi in Central Region of Benin. *International Journal of Advanced Research* 8:1144–1152. DOI: 10.21474/ijar01/11946.
Layek U, Bera K, Bera B, Bisui S, Pattanayek SK, Karmakar P. 2021. Assessment of yield enhancement in cashew (*Anacardium occidentale* L.) by the pollinator sharing effect of magnetic bee-friendly plants in India. *Acta Ecologica Sinica* 41:243–252. DOI: 10.1016/j.chnaes.2021.05.003.
MAFF. 2004. MAFF (Ministry of Agriculture, Forestry and Fisheries of East-Timor). :MAFF website. Available at: http://www.gov.east-ti.
Massari F. 1994 Introduction. In ''The World Cashew Economy.' NOMISMA, L'Inchiostroblu, Bol. Italy. (A. M. Del, pp. 3–4.) 443831467999735473/102933.
Mitchell JD, Mori SA. 1987. The cashew and its relatives (*Anacardium*: Anacardiaceae). El marañón y sus parientes (*Anacardium*: Anacardiaceae). *Biblioteca OET: M* 42:v. 42, 1-76. Año 1987.
Mneney EE, Mantell SH, Bennett M. 2001. Use of random amplified polymorphic DNA (RAPD) markers to reveal genetic diversity within and between populations of cashew

(*Anacardium occidentale* L.). *Journal of Horticultural Science and Biotechnology* 76:375–383. DOI: 10.1080/14620316.2001.11511380.
Monteiro F, Catarino L, Batista D, Indjai B, Duarte MC, Romeiras MM. 2017. Cashew as a high agricultural commodity in West Africa: Insights towards sustainable production in Guinea-Bissau. *Sustainability* 9:1–14. DOI: 10.3390/su9091666.
Monteiro F, Romeiras MM, Figueiredo A, Sebastiana M, Baldé A, Catarino L, Batista D. 2015. Tracking cashew economically important diseases in the West African region using metagenomics. *Frontiers in Plant Science* 6:1–6. DOI: 10.3389/fpls.2015.00482.
Monteiro F, Vidigal P, Barros AB, Monteiro A, Oliveira HR. 2016. Genetic Distinctiveness of Rye In situ Accessions from Portugal Unveils a New Hotspot of Unexplored Genetic Resources. *Frontiers in Plant Science* 7:1–17. DOI: 10.3389/fpls.2016.01334.
Moumouni K, Vianney TW, Larbouga B, Issa W. 2022. Genetic diversity assessment among 18 elite cashew tree genotypes (*Anacardium occidentale* L.) selected in Western Burkina Faso. *Journal of Plant Breeding and Crop Science* 14(1): 1-11. https://doi.org/10.5897/JPBCS2021.0986
Nei M. 1972. Genetic Distance between Populations. *The American Naturalist* 106:283–292. DOI: 10.1086/282771.
Nybom H. 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology* 13:1143–1155. DOI: 10.1111/j.1365-294X.2004.02141.x.
Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. 20:289–290. DOI: 10.1093/bioinformatics/btg412.
Park SDE. 2001. Trypanotolerance in West African cattle and the population genetic effects of selection [PhD thesis]. [Dublin (Ireland)]: University of Dublin.
Peakall R, Smouse PE. 2006. GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288–295. DOI: 10.1111/j.1471-8286.2005.01155.x.
Peng RK, Christian K, Gibb, K. 2009. The Improvement of Cashew Crop in East Timor. Report to the Ministry of Agriculture and Fisheries of East Timor.
Pradhan C, Peter N, & Dileep N. 2020. Nuts as Dietary Source of Fatty Acids and Micro Nutrients in Human Health. *In* V. Rao, L. Rao, M. Ahiduzzaman, & A. K. M. A. Islam (Eds.), Nuts and Nut Products in Human Health and Nutrition. IntechOpen. https://doi.org/10.5772/intechopen.94327
Pritchard J K, Stephens M, & Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.
R Core Team. 2021 R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. Vienna, Austria.
Rambaut A. 2014. FigTree v1. 4.2, a graphical viewer of phylogenetic trees. Available from left angle bracket http://tree. bio. ed. ac. uk/software/figtree/right angle bracket.
República Democrática de Timor-Leste. 2011 East-Timor Strategic Development Plan 2011 – 2030. Pp 1- 228. https://www.adb.org/sites/default/files/linked-documents/cobp-tim-2014-2016-sd-02.pdf.
Rice WR. 2013. Analyzing Tables of Statistical Tests. *Evolution*, 43(1): 223-225.
Rosenberg NA. 2004. DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology* 4:137–138. doi: 10.1046/j.1471-8286.2003.00566.x
Rousset F. 2008. GENEPOP'007: A complete re-implementation of the GENEPOP software for

Windows and Linux. *Molecular Ecology Resources* 8:103–106. DOI: 10.1111/j.1471-8286.2007.01931.x.

Salehi B, Gültekin-Özgüven M, Kirkin C, Özçelik B, Morais-Braga MFB, Carneiro JNP, Bezerra CF, Da Silva TG, Coutinho HDM, Amina B, Armstrong L, Selamoglu Z, Sevindik M, Yousaf Z, Sharifi-Rad J, Muddathir AM, Devkota HP, Martorell M, Jugran AK, Martins N, Cho WC. 2019. *Anacardium* plants: Chemical, nutritional composition and biotechnological applications. *Biomolecules* 9:1–34. DOI: 10.3390/biom9090465.

Savadi S, Megha KSVS, Mohana BMMGS. 2020. Genetic diversity and identification of interspecific hybrids of Anacardium species using microsatellites. *Brazilian Journal of Botany* 2026. DOI: 10.1007/s40415-020-00678-5.

Schuelke M. 2000. An economic method for the fluorescent labeling of PCR fragments A poor man's approach to genotyping for research and high-throughput diagnostics. *PRism* 18:1–2.

Singh AK. 2018. Early History of Crop Presence/Introduction in India: III. *Anacardium occidentale* L. *Cashew Nut. Asian Agri-History* 22. DOI: 10.18311/aah/2018/21389.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P. 2004. MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* 4:535–538. DOI: 10.1111/j.1471-8286.2004.00684.x.

Weir BS, Cockerham CC. 1984. Estimating *F*-Statistics for the analysis of population structure. *Evolution* 38:1358–1370. doi: 10.2307/2408641.

Tables

Table 1. Populations by country, district and location, geographical coordinates and total individuals sampled (N).

Table 2. Loci used to screen 14 cashew populations. Primers sequences, multiplexing scheme, amplicon size range (bp) and amplicon expected size (bp) are provided.

Table 3. Marker's diversity measurements. The level of genetic diversity of each SSR marker was described with the parameters of total number of alleles, Polymorphism Information Content (PIC), gene diversity (expected heterozygosity, H_e), observed heterozygosity (H_o), inbreeding/fixation coefficient (F). About 207 individuals from 14 populations were analyzed for each locus.

Table 4. Genetic diversity indices by a country analysis and a population approach scheme. Countries: Mozambique (MZ, 2 populations), Indonesia (IND, 1 population) and East-Timor (ET, 11 populations); Number of populations: 14. Sample size (N). Genetic diversity indices for each group were assessed by mean alleles per locus (Na), expected heterozygosity (He) and observed heterozygosity (HO) with corresponding standard deviation (SD) values, private alleles (P_A) and inbreeding/fixation coefficient (F).

Table 5. AMOVA results including fixation indices F_{CT}, F_{SC}, and F_{ST}.
Legend: V_a, variance among populations; V_b, variance within populations; V_c, variance within individuals.

Figures

Figure 1. Geographical location of cashew populations studied at Mozambique (**A**), Indonesia, (**B**) and East-Timor (**C**), and illustrative orchards from East-Timor (**D**-**H**) and Indonesia (**I**).
Legend: Illustrative orchards from some plantations sampled in East-Timor (**D-H**): Natarbora-Manatuto district (**D**), Mailiana in Bobonaro district (**E**), Triloca in Baucau district (**F**), Fatucahi in Manufahi (**G**), Viqueque (**H**); and in Kefamenanu in Indonesia (**I**).

Figure 2. UPGMA (**A**) and NJ (**B**) trees generated from *FreeNA* using matrix DC^{INA} of East-Timor, Indonesia, and Mozambique dataset. Legend: (□ Manatuto, ⊡ Bobonaro, ▣ Baucau, ◧ Cova Lima, ◨ Manufahi, ▦ Viqueque, ❖ Indonesia, ▲Mozambique).

Figure 3. UPGMA (**A**) and NJ (**B**) trees generated from *FreeNA* using matrix DC^{INA} presenting East-Timor and Indonesia populations. Legend: (□ Manatuto, ⊡ Bobonaro, ▣ Baucau, ◧ Cova Lima, ◨ Manufahi, ▦ Viqueque, ❖ Indonesia).

Figure 4. Clustering based on SSR data of the optimal *K*-means using STRUCTURE (**A**) for populations of Mozambique, East-Timor, and Indonesia; and DAPC analyses representation of *K* = 5 (**B**).

Figure 5. Optimal *K*- means individual-based clustering using STRUCTURE (*K* = 2, **A**) for populations of East-Timor and Indonesia and DAPC analyses of *K*=2 (**B**).

988
989

Supplementary Materials

991 **Table S1**. Hardy–Weinberg equilibrium (HWE) test for each locus-population combination
992 using GenePop v4.5.

993 **Table S2.** Linkage Disequilibrium (LD) test for each locus-population combination using
994 GenePop v4.5.

995 **Table S3**. Pairwise F_{ST} (lower-left matrix) and F_{ST} ^{ENA} (upper-right matrix) between all
996 populations of East-Timor, Indonesia and Mozambique.

997 **Table S4**. Pairwise F_{ST} (lower-left matrix) and F_{ST} ^{ENA} (upper-right matrix) between all
998 populations of East-Timor and Indonesia.

999 **Figure S1**. UPGMA (A) and NJ (B) trees generated using matrix Nei's *D* distance, respectively,
1000 representing the population from East-Timor, Indonesia, and Mozambique.;

1001 **Figure S2**. UPGMA (A) and NJ (B) trees generated using matrix Nei's *D* distance, respectively,
1002 representing the population from East-Timor and Indonesia.

1003 **Figure S3**. STRUCTURE ad hoc statistics retrieved by StructureHarvester using 1 to 15 possible
1004 clusters (*K*). Variation of Δ*K* values according to Evanno *et al.* (2005) method for populations
1005 from East-Timor, Indonesia and Mozambique.

1006 **Figure S4**. DAPC results inference of the number of clusters using *find.clusters* function with a
1007 $K = 10$ (left) and $K = 5$ (right).

1008 **Figure S5**. Scatterplot of DAPC for $K = 5$ assignment.

1009 **Figure S6**. Loading plots of the two Discriminant Functions following DAPC analysis with a K
1010 $= 5$.

1011 **Figure S7**. STRUCTURE ad hoc statistics retrieved by StructureHarvester using 1 to 12 possible
1012 clusters (*K*). Variation of Δ*K* values according to Evanno *et al.* (2005) method for populations
1013 from East-Timor and Indonesia.

1014 **Figure S8**. Number of clusters inferred by DAPC *find.clusters* function with a $K = 5$ and $K = 2$.

1015 **Figure S9**. Loading plot of the DF1 following DAPC analysis with a K=2, after assigning a
1016 0.045 as threshold.

1017
1018
1019
1020
1021

Figure 1

Figure 1. Geographical location of cashew populations studied at Mozambique (A), Indonesia, in Kefamenanu (B) and East-Timor (C), and illustrative orchards from East-Timor (D-H) and Indonesia (I).

Illustrative orchards from some plantations sampled in East-Timor (D-H): Natarbora-Manatuto district (D), Mailiana in Bobonaro district (E), Triloca in Baucau district (F), Fatucahi in Manufahi (G), Viqueque (H); and in Kefamenanu in Indonesia (I).

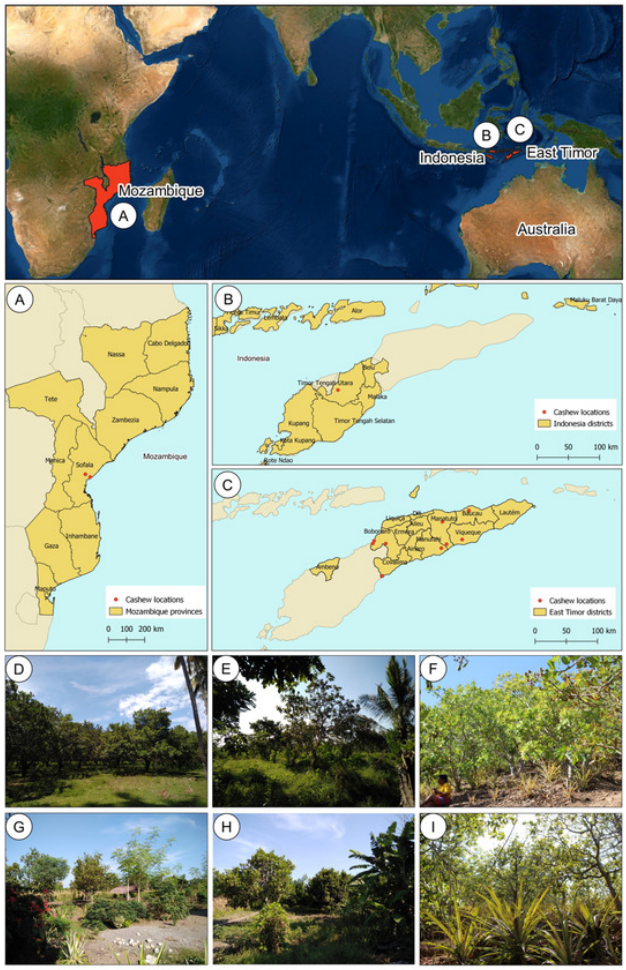

Figure 1. Geographical location of cashew populations studied at Mozambique (**A**), Indonesia, in Kefamenanu (**B**) and East-Timor (**C**), and illustrative orchards from East-Timor (**D-H**) and Indonesia (**I**).

Legend: Illustrative orchards from some plantations sampled in East-Timor (**D-H**): Natarbora-Manatuto district (**D**), Mailiana in Bobonaro district (**E**), Triloca in Baucau district (**F**), Fatucahi in Manufahi (**G**), Viqueque (**H**); and in Kefamenanu in Indonesia (**I**).

Eliminou: sampled in the three tropical countries:

Eliminou: . Cashew

Eliminou: of the

Eliminou: Kefamenanu in Indonesia

Eliminou: ,

Eliminou: Viqueque

Figure 2

UPGMA (A) and NJ (B) trees generated from *FreeNA* using matrix DC^{INA} of East-Timor, Indonesia, and Mozambique dataset.

Legend: (□ Manatuto, ⊡ Bobonaro, ▣ Baucau, ◧ Cova Lima, ◨ Manufahi, ▦ Viqueque, ❖ Indonesia, ▲Mozambique).

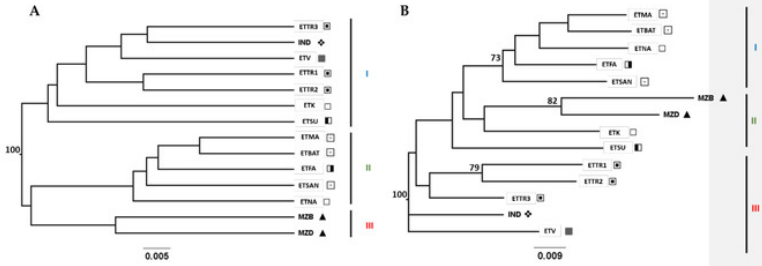

Figure 2. UPGMA (**A**) and NJ (**B**) trees generated from *FreeNA* using matrix DC^{INA} of East-Timor, Indonesia, and Mozambique dataset. Legend: (□ Manatuto, ⊡ Bobonaro, ▣ Baucau, ◧ Cova Lima, ◨ Manufahi, ▦ Viqueque, ❖ Indonesia, ▲Mozambique).

Eliminou: respectively representing the population from

Eliminou: dataset

Figure 3

UPGMA (A) and NJ (B) trees generated from *FreeNA* using matrix DC^{INA} presenting East-Timor and Indonesia populations.

Legend: (□ Manatuto, ⊡ Bobonaro, ▣ Baucau, ◧ Cova Lima, ◨ Manufahi, ▦ Viqueque, ❖ Indonesia).

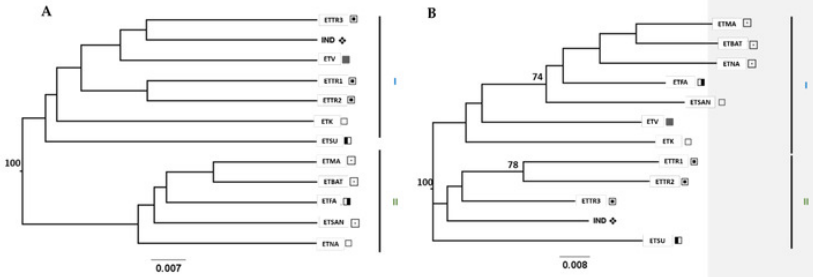

Figure 3. UPGMA (**A**) and NJ (**B**) trees generated from *FreeNA* using matrix DC^{INA} presenting East-Timor and Indonesia populations. Legend: (□ Manatuto, ⊡ Bobonaro, ▣ Baucau, ◧ Cova Lima, ◨ Manufahi, ▦ Viqueque, ❖ Indonesia).

Eliminou: respectively re

Eliminou: the population from

Figure 4

Clustering based on SSR data of the optimal *K*-means using STRUCTURE (A) for populations of Mozambique, East-Timor, and Indonesia; and DAPC analyses representation of $K = 5$ (B).

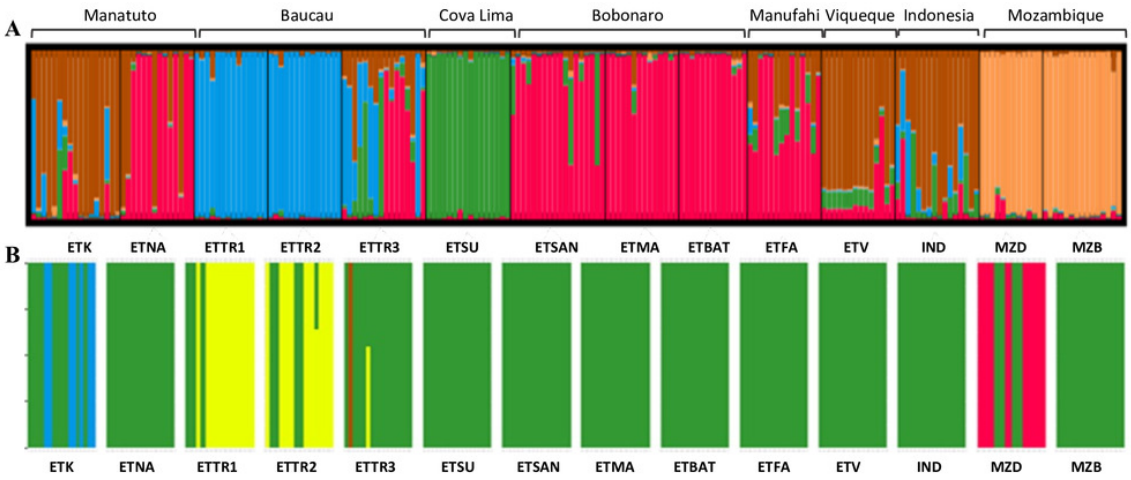

Figure 4. Clustering based on SSR data of the optimal *K*-means using STRUCTURE (**A**) for populations of Mozambique, East-Timor, and Indonesia; and DAPC analyses representation of $K = 5$ (**B**).

Figure 5

Optimal *K*- means individual-based clustering using STRUCTURE (*K* = 2, A) for populations of East-Timor and Indonesia and DAPC analyses of *K*=2 (B).

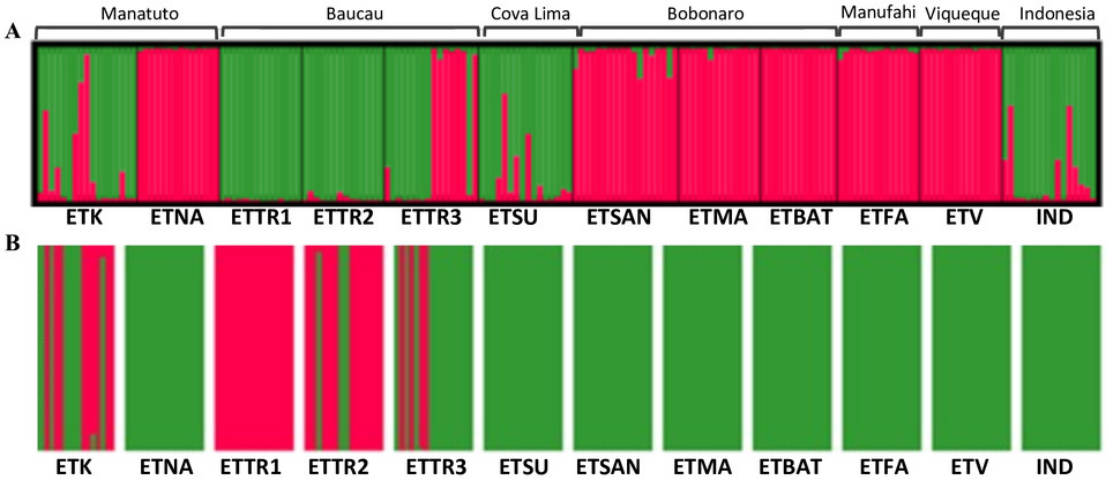

Figure 5. Optimal *K*- means individual-based clustering using STRUCTURE (*K* = 2, **A**) for populations of East-Timor and Indonesia and DAPC analyses of *K*=2 (**B**).

Table 1(on next page)

Populations by country, district and location, geographical coordinates and total individuals sampled (N).

Table 1. 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 (ET)	Manatuto	Kribas	ETK	-8.650	125.983	17
		Natarbora	ETNA	-8.974	126.045	14
	Baucau	Triloca	ETTR1	-8.490	126.370	14
			ETTR2	-8.491	126.370	14
			ETTR3	-8.487	126.370	16
	Cova Lima	Suai	ETSU	-9.432	125.108	16
	Bobonaro	Sanirin	ETSAN	-8.925	124.986	18
		Maliana	ETMA	-8.966	125.156	14
		Batugade	ETBAT	-8.965	124.966	13
	Manufahi	Fatucahi	ETFA	-9.032	125.968	14
	Viqueque	Viqueque	ETV	-8.907	126.273	14
Indonesia (IND)	Kota Kefamenanu	Kefamenanu	IND	-9.437	124.485	16
Mozambique (MZ)	Sofala	Beira	MZB	-19.723	34.982	12
		Dondo	MZD	-19.571	34.715	15
Total						**207**

Table 2(on next page)

Loci used to screen 14 cashew populations. Primers sequences, multiplexing scheme, amplicon size range (bp) and amplicon expected size (bp) are provided.

Following Culley *et al*. [19], 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”.

Table 2. Loci used to screen 14 cashew populations. Primers sequences, multiplexing scheme, amplicon size range (bp) and amplicon expected size (bp) are provided.

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: GTTTCTTCTTCTTGAACTTGACACTTCATCCA			
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			

Table 3(on next page)

Marker’s diversity measurements.

The level of genetic diversity of each SSR marker was described with the parameters of total number of alleles, Polymorphism Information Content (PIC), gene diversity (expected heterozygosity, H_e), observed heterozygosity (H_o), inbreeding/fixation coefficient (*F*). About 207 individuals from 14 populations were analyzed for each locus.

Table 3. Marker's diversity measurements. The level of genetic diversity of each SSR marker was described with the parameters of total number of alleles, Polymorphism Information Content (PIC), gene diversity (expected heterozygosity, H_e), observed heterozygosity (H_o), inbreeding/fixation coefficient (*F*). About ~~Total samples analyzed=~~ 207 individuals from 14 populations were analyzed for each locus.~~.~~

Locus	Allele number	PIC	H_e	H_o	*F*
mAoR48	11	0.64	0.69	0.49	0.2
mAoR6	17	0.68	0.71	0.55	0.11
mAoR17	19	0.83	0.84	0.73	-0.03
mAoR7	15	0.77	0.8	0.64	-0.04
mAoR11	16	0.66	0.69	0.47	0.19
mAoR3	25	0.74	0.76	0.48	0.24
mAoR42	14	0.65	0.69	0.59	0.03
mAoR52	14	0.67	0.7	0.57	0.05
mAoR2	4	0.48	0.57	0.44	0.13
mAoR35	9	0.62	0.65	0.28	0.48
mAoR47	7	0.63	0.69	0.49	0.06
mAoR16	6	0.4	0.45	0.34	0.03
Total	157				
Mean	13.08	0.65	0.69	0.51	0.12

Table 4(on next page)

Genetic diversity indices by a country analysis and a population approach scheme.

Countries: Mozambique (MZ, 2 populations), Indonesia (IND, 1 population) and East-Timor (ET, 11 populations); Number of populations: 14. Sample size (N). Genetic diversity indices for each group were assessed by mean alleles per locus (Na), expected heterozygosity (H_e) and observed heterozygosity (H_O) with corresponding standard deviation (SD) values, private alleles (P_A) and inbreeding/fixation coefficient (F).

Table 4. Genetic diversity indices by a country analysis and a population approach scheme. Countries: Mozambique (MZ, 2 populations), Indonesia (IND, 1 population) and East-Timor (ET, 11 populations); Number of populations: 14. Sample size (N). Genetic diversity indices for each group were assessed by mean alleles per locus (Na), expected heterozygosity (H_e) and observed heterozygosity (H_O) with corresponding standard deviation (SD) values, private alleles (P_A) and inbreeding/fixation coefficient (F).

	N	Na	Na SD	H_e	H_e SD	H_o	H_o SD	P_A	F
Countries									
East Timor (ET)	164	10.67	4.96	0.670	0.033	0.506	0.012	6	0.24
Indonesia (IND)	16	3.42	1.93	0.561	0.063	0.465	0.037	0	0.12
Mozambique (MZ)	27	5.50	2.28	0.711	0.028	0.535	0.028	2	0.25
Mean	69	6.53	3.06	0.647	0.0413	0.502	0.0257	-	0.20
Populations									
ETK	17	5.17	1.75	0.67	0.04	0.56	0.04	1	0.14
ETNA	14	3.75	1.14	0.65	0.03	0.56	0.04	0	0.10
ETTR1	14	3.83	1.53	0.65	0.04	0.57	0.04	0	0.06
ETTR2	14	3.67	1.72	0.58	0.06	0.49	0.04	0	0.11
ETTR3	16	5.08	2.54	0.59	0.06	0.51	0.04	0	0.07
ETSU	16	3.58	1.44	0.65	0.04	0.46	0.04	0	0.26
ETSAN	18	3.58	1.83	0.57	0.05	0.57	0.03	0	-0.04
ETMA	14	3.00	0.85	0.57	0.03	0.46	0.04	0	0.14
ETBAT	13	2.42	0.79	0.48	0.05	0.45	0.04	0	0.01
ETFA	14	2.75	0.75	0.51	0.05	0.43	0.04	0	0.10
ETV	14	2.25	0.62	0.45	0.06	0.45	0.04	0	-0.09
IND	16	3.42	1.93	0.56	0.06	0.47	0.04	0	0.12
MZB	12	4.33	1.50	0.69	0.03	0.51	0.04	1	0.23
MZD	15	3.50	1.17	0.65	0.03	0.56	0.04	0	0.12
Mean	14.79	3.60	1.40	0.59	0.05	0.50	0.04	-	0.09

Table 5(on next page)

AMOVA results including fixation indices F_{CT}, F_{SC}, and F_{ST}.

Legend: V_a, variance among populations; V_b, variance within populations; V_c, variance within individuals. The genetic differentiation between countries and East-Timor *vs* Indonesia populations is denoted as F_{CT}, among individuals within populations as F_{SC} and within individuals as F_{ST} . *$p < 0.001$.

Table 5. AMOVA results including fixation indices F_{CT}, F_{SC}, and F_{ST}. Legend: V_a, variance among populations; V_b, variance within populations; V_c, variance within individuals.

Source of variation	df	Sum of Squares	Variance components	Variation (%)	Fixation indices
Countries (MZ, IND, ET)					
Among pops	13	168.537	V_a= 0.34692	12.6	F_{CT}=0.128*
Among individuals	193	523.90	V_b= 0.30894	11.22	F_{SC}=0.126*
Within individuals	207	434.00	V_c = 2.09662	76.17	F_{ST}=0.238*
Countries (ET, IND)					
Among pops	11	91.172	V_a= 0.16546	12.97	F_{CT}= 0.273*
Among individuals	168	288.058	V_b= 0.12973	14.4	F_{SC}= 0.165*
Within individuals	180	221.00	V_c =1.22778	72.63	F_{ST}= 0.129*

The genetic differentiation between countries and East-Timor *vs* Indonesia populations is denoted as F_{CT}, among individuals within populations as F_{SC} and within individuals as F_{ST} . *$p < 0.001$.