

East-Timor as unexplored yet important source of cashew (*Anacardium occidentale* L.) genetic diversity (#73008)

1

First submission

Guidance from your Editor

Please submit by **16 Jun 2022** for the benefit of the authors (and your \$200 publishing discount) .



Structure and Criteria

Please read the 'Structure and Criteria' page for general guidance.



Author notes

Have you read the author notes on the [guidance page](#)?



Raw data check

Review the raw data.



Image check

Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the [materials page](#).

14 Figure file(s)

8 Table file(s)



Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. BASIC REPORTING
2. EXPERIMENTAL DESIGN
3. VALIDITY OF THE FINDINGS
4. General comments
5. Confidential notes to the editor

 You can also annotate this PDF and upload it as part of your review

When ready [submit online](#).

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your [guidance page](#).

BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [PeerJ standards](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (see [PeerJ policy](#)).

EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  All underlying data have been provided; they are robust, statistically sound, & controlled.
-  Conclusions are well stated, linked to original research question & limited to supporting results.



The best reviewers use these techniques

Tip

Example

Support criticisms with evidence from the text or from other sources

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.

Organize by importance of the issues, and number your points

1. Your most important issue
2. The next most important item
3. ...
4. The least important points

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

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

Lara Guterres^{1, 2, 3}, João Barnabé^{2, 4}, André Barros⁵, Alberto Bento Charrua^{3, 6}, Maria Cristina Duarte⁴, 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), 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 value 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 were screened using microsatellites (SSRs) to evaluate intraspecific diversity and population structuring.

Methods. A total of 207 individuals from 11 populations from East-Timor, together with Indonesia (1) and Mozambique (2) working as outgroup populations, were analyzed with 16 cashew-specific SSRs. A comprehensive sampling within Est-Timor was done, thus covering cashew orchards distribution in the country. Genetic diversity indices were calculated, and population structuring used 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 a largely understudy country is higher than previous reported from cashew, where a narrow genetic diversity was described. A higher allelic richness was found within East-Timor populations, compared to the outgroup populations (Mozambique and Indonesia), reinforced by the presence of private alleles. Also, within East-Timor, a district- associated genetic diversity clustering into two genetic groups was depicted, which may point to multiple events of cashew introduction and to hotspots of cashew genetic diversity that should be explored for future crop improvement.

Conclusions. Considering that crop diversity underpins the productivity, resilience and adaptive capacity of agriculture, our study provides useful information regarding genetic diversity and population structure of cashew populations. This data can be also important to access an in-country genetic signature, increasing the crop market value of East-Timor cashew.

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

Lara Guterres^{1,2,3}, João Barnabé^{1,4}, André Barros⁵, Alberto B. Charrua^{1,2,6}, Maria C. Duarte⁴, 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), Faculty of Sciences, University of Lisbon, 1349-017 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, fmonteiro@fc.ul.pt.

Abstract

Background. Cashew (*Anacardium occidentale* L.) is a crop currently grown in several tropical countries due to economic value 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 were screened using microsatellites (SSRs) to evaluate intraspecific diversity and population structuring.

Methods. A total of 207 individuals from 11 populations from East-Timor, together with Indonesia (1) and Mozambique (2) working as outgroup populations, were analyzed with 16 cashew-specific SSRs. A comprehensive sampling within Est-Timor was done, thus covering cashew orchards distribution in the country. Genetic diversity indices were calculated, and population structuring used 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 a largely understudy country is higher than previous reported from cashew, where a narrow genetic diversity was described. A higher allelic richness was found within East-Timor populations, compared to the outgroup populations (Mozambique and Indonesia), reinforced by the presence of private alleles. Also, within East-Timor, a district- associated genetic diversity clustering into two genetic groups was depicted, which may point to multiple events of cashew introduction and to hotspots of cashew genetic diversity that should be explored for future crop improvement.

Conclusions. Considering that crop diversity underpins the productivity, resilience and adaptive capacity of agriculture, our study provides useful information regarding genetic diversity and population structure of cashew populations. This data can be also important to access an in-country genetic signature, increasing the crop market value of East-Timor cashew.

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 several decades, the demand for cashew has increased along with the rising of income global market demands. 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 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 using different molecular tools namely microsatellites (SSRs, Simple Sequence Repeats) and RAPDs (Random Amplified Polymorphologic DNA), associated with morphological characterization. Within cashew-producing countries, studies have been conducted in characterizing genetic diversity of cashew accessions in Ivory Coast (Kouakou *et al.*, 2020), Nigeria (Aliyu & Awopetu, 2007) using 187 Nigerian germplasm accessions, at

Indian germplasm accessions (Archak *et al.*, 2009), in Tanzanian cashew cultivars (Mneney, Mantell & Bennett, 2001), in Malawi (Chipojola *et al.*, 2009) and from its native origin in Brazil (dos Santos *et al.*, 2019). Overall, most studies highlight the existence of a narrow genetic diversity within each country cashew accessions when compared with cashew native origin, Brazil, where accessions were most distinct from the other regions of the world.

Cashew was one the many tropical crops introduced by the Portuguese in the sixteenth century as part of the Columbian Exchange Event (Havik *et al.*, 2018), in which several crops and livestock were introduced from Brazil into the African continent. Cashew was regarded as a rustic tree to fight afforestation (Monteiro *et al.*, 2015, 2017). In Asia, cashew has been introduced most likely through Goa (India), Portugal's main settlement in the East Indies at a time between 1560 and 1565 (Massari, 1994). Nowadays, cashew is still known as '*paragi andi*' at Kerala meaning foreign nut, '*lanka beeja*' at Orissa assuming its introduction from Sri Lanka, and '*mundiri*' indicating the shape of the nut in Tamil Nadu (India). The production of cashew trees that have proven to be well adapted to Indian soil conditions, upheld the indigenous populations to explore cashew products beyond its nuts, namely the creation of a local fermented brew (the *feni*) made from ripened cashew apples. After India, cashew spread to South Asia clearly in Moluccas (Indonesia), and there to northern Australia and North America, more specifically to Florida (Nair, 2010). Nowadays, cashew has dispersed to other Asian regions as Vietnam, Indonesia, Philippines, Malaysia, Thailand, and Sri Lanka (Havik *et al.*, 2018). India and Vietnam are now amongst the top world cashew producers. Despite introductions at different timelines, at first, cashew was established as a rustic tree to cope afforestation at both African and Asian countries (Monteiro *et al.*, 2017; Havik *et al.*, 2018). However, at dissimilar paces, these two tropical regions have transformed cashew as one of the major cash generator commodity, currently with different stages at technological levels: African countries (except for Mozambique, Nigeria and Côte d'Ivoire) as major raw cashew producers/exporters with few processing and/or value chain sectors, while in Asia a complete structured value chain and technological industry transforming sector is implemented, where also import raw cashew nuts at low prices from less resourceful African countries (e.g. Guinea-Bissau, Senegal, Guinea, Burkina Faso) occurs.

East-Timor is a small island country located in Asia, bordered by Indonesia, which economy is heavily dependent on oil and gas, since its independence in 2002. Oil and gas contributed about 80% to the country's GDP in 2012-2013 and about 99% exports. The Government is investing in the non-oil economy mainly in agriculture, which is the main human activity in East-Timor, providing subsistence to an estimated 80 percent of the population (Harmadi & Gomes, 2013). Most farmers practice subsistence farming, with nearly absent large-scale farming and limited access to markets. Coffee is the main export commodity in the country largely grown at high elevations along with other important commercial crops as '*chimeri*' (candlenut tree), vanilla and coconut, besides staple crops production as maize in uplands and rice paddy (MAFF, 2004). The Ministry of Agriculture, Forestry and Fisheries (MAFF) of East-Timor Strategic Plan 2014-2020, was implemented to contribute to the achievement of the Timor-Leste Strategic Development

Plan 2011-2030, which sets out the priorities identified to promote economic growth in rural areas and to advance the reduction of poverty and to provide better food security (República Democrática de Timor-Leste, 2011). The MAFF of East-Timor is also targeting its attention to increase its export trade through the intensive production of cash crops, to meet global market demand, namely through the exploration of cashew as a new commodity to increase the country agriculture remittance. In the last 10 years, cashew has been considered by small farmers and the government as an alternative export crop to coffee with a high market value.

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 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, so it could work 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 comprising the eastern half of the Island of Timor; the Atauro Island, north of Díli; the Jaco Island, on the easternmost end of the island; and Oecusse, an enclave on the northwestern side of the island, within Indonesia (**Figure 1**). As far as administrative division is concerned, Timor-Leste is split into 13 districts: Bobonaro, Liquiçá, Díli, Baucau, Manatuto and Lautém on the north coast; Cova Lima, Ainaro, Manufahi and Viqueque, on the south coast; Ermera and Aileu, the two landlocked districts; and Oecusse-Ambeno, the enclave in Indonesian territory. The territory has an estimated population of 1,183,643 (Government of Timor-Leste, 2015), distributed within an area of approximately 15,000 square kilometers. The capital is Díli. 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. After coffee, maize and rice paddy rises as the most important crops, which are the main staple crops. However, land suitable for rice production is limited and maize is more widely grown in the uplands including hillsides (MAFF, 2004).

Sampling

Cashew populations were sampled from different orchards in East-Timor, Indonesia, and Mozambique (**Figure 1A-C**). In each population, 12-18 individual plants were sampled, preserved in silica gel until further processing (**Table 1**). 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 is borders East-Timor's Oecusse enclave, one of 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 with a strong land connection 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, where 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 then 1 cm² 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, and the supernatant 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 1min. The gDNA was eluted in 40µl 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 nm and 260/230 nm 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 3 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. After, 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 (Holleley & Geerts, 2009), 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.* (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 a total volume of 25µL with 50–100 ng gDNA and 2.5 pmol of each primer Forward and Reverse and 0.15 pmol 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.* (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 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." Genotypic matrix generated (Guterres *et al.*, 2022) was used for further genetic analysis and population structuring.

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.* as implemented in FreeNA (<http://www.montpellier.inra.fr/URLB/>). These values were computed as described by Chapuis & Estoup (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. These included the total allele number and mean alleles per locus (N_a), 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 estimated with 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) 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, 2015). Distances relationship among populations was done in two separated approaches: first using the three countries (East-Timor, Indonesia and Mozambique) populations and, second, by excluding continental populations from Mozambique.

Hierarchical genetic analysis (AMOVA)

An Analysis of Molecular Variance (AMOVA, Weir & Cockerham, 1984; Excoffier, 2000; Hill, 2021) 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 all three countries populations (MZB=2; IND=1; ET=11) as groups and the second, narrowed to East-Timor and Indonesia as groups. In each AMOVA, the total variance was partitioned into components to account for differences between two defined groups [V_a , (1) MZ vs ET vs IND; (2) ET vs IND populations], differences among populations within those groups (V_b), 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: a Bayesian clustering analysis using STRUCTURE (Pritchard *et al.*, 2000) and a multivariate analysis method, Discriminant Analysis of Principal Components (DAPC, Jombart & Balloux, 2009). 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 (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 . The second approach using East-Timor and 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, which is based on the rate of change of the “estimated likelihood” between successive K -values. 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 *adeigenet* 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, the frequencies of null alleles were higher than 0.20, except for the markers mAoR33, mAoR12, mAoR29 and mAoR41, that were excluded after this analysis. Subsequent analysis were made with the remaining 12 SSRs markers (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 significance ($p < 0.05$); 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 (data not shown). 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 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). 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 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**).

By performing a population genetic diversity analysis presented at **Table 4**, East-Timor presented 10.67 alleles in average and the lowest value was from Indonesia with 3.42, the H_e was 0.67, 0.56 and 0.71 respectively (East-Timor, Indonesia and Mozambique), the H_o was 0.51, 0.47 and 0.54 following the same order, all countries populations presented a F positive varying 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, ~~are the lowest allelic diverse populations~~. 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 ETR3 are the populations that present greater allelic diversity (N_a , **Table 4**). ~~At the level of the~~ expected heterozygosity (H_e), the highest ~~value~~ was obtained for population ETK, ~~with a value of 0.67~~ and the lowest value ~~0.45~~ in the population ETV. When comparing the observed heterozygosity (H_o), the highest ~~value~~ was obtained for the population ETTR1 ~~with a value of 0.57~~, and the lowest ~~value~~ ETFA ~~with 0.43~~. 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. All countries~~ presented at least one private allele, with Indonesia (0.42) and East Timor displaying a high number of private alleles (6.17).

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 S2A), and, ~~after~~ an analysis narrowed to East-Timor and Indonesia (Supplementary Table S2B), 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 4**) 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 ~~are~~ depicted: one cluster (I) that includes ETTR populations (ETTR1-3) from Baucau, Viqueque (ETV), Cova Lima (ETLSU) and ETLK 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 (ETLFA), 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 populations, Indonesia, and populations from Viqueque (ETLV) and Cova Lima (ETLSU), and the third cluster include Bobonaro populations (TLMA, ETLBAT and ETLSAN) and the remaining East-Timor populations (ETLFA, ETLNA and ETLK).

When observing trees without Mozambican populations (**Figure 3**), ~~one can depict two clusters found, similarly to the clustering 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 very dissimilar NJ-generated trees (Nei's D distance, **Figure 3B**, Supplementary Fig. S2A, S2B), ~~which may be attributed to different assumptions adopted in each clustering methods, with a strict UPGMA and a relaxed (NJ) molecular clock shown previously to have implications when inferring phylogenies considering that rates of evolution may vary among microsatellite loci (Putman & Carbone, 2014).~~

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, respectively (**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%). ~~In both cases, a high molecular variation was found within individuals, as expected for a cross-pollinated species, as previously detected in other cashew genetic diversity studies (Freitas & Paxton, 1996).~~

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). 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.* method (Supplementary Figure S8), with no alternative ideal K . Considering the high ΔK -values displayed for $K = 2$, the analysis without Mozambique (**Figure 5**), two clusters were observed, the first cluster green with Baucau, Cova Lima, Manatuto and Indonesia: and the second cluster in red color with Bobonaro, Manufahi and Viqueque districts (**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 2 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, which 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. Besides, 3 cashew orchards populations were included outside East-Timor, namely one from Indonesia (nearest country) and two from Mozambique, the latter working as a continental region for detecting any bias relating to geographic isolation within East-Timor as an island country. To carry out the genetic diversity characterization, a total of 207 individuals belonging to 14 cashew populations from three tropical nations (East-Timor, Mozambique, and Indonesia) were screened using 16 cashew specific microsatellites (Croxford *et al.*, 2006).

Cashew diversity assessment

Fourteen different cashew populations were first genotyped with 16 SSRs, and after SSRs quality assessment, markers were further used for subsequent genetic diversity analysis (**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 other 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, 0.38-0.82) which indicates their high informativeness, and when compared with a recent study using 21 SSRs cashew SSR (cSSR, Savadi *et al.*, 2020) to screen 23 cashew accessions lower PIC values (0.33 average PIC-value, 0.10 to 0.38) were described. 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, though a potential influence of the lower number of SSRs loci used should not be discarded. 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 among the countries analyzed in our study is high when compared to other cashew studies; more alleles were found within East-Timor, but this can be explained by the fact that we used more individuals and populations in East-Timor. Our results show a higher allelic

richness in East-Timor populations ($N_a=10.67$, **Table 4**) than in Mozambique ($N_a=3.42$) and in Indonesia ($N_a=3.42$), which highlights the high genetic diversity in East-Timor country. Nevertheless, only with the inclusion of more populations from Mozambique and Indonesia prompt analysis could be done to determine ~~if reduced allelic diversity on cashew are effectively observed~~. Within East-Timor, ETK population from Manatuto district displays the highest allelic richness ($N_a=5.17$), followed by ETTR3 from Baucau district, with ETV from Viqueque with the lowest allelic diversity obtained ($N_a=2.25$, 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. **High allelic richness was expected in our study, as cashew is an outcrossing tree.** The populations analyzed might reflect long-term genetic diversity that can be exploited in a breeding program to improve yield and nut quality. Moreover, identifying trees with high allelic richness is necessary for conserving cashew germplasm. Allelic richness data are also useful for managing germplasm collections in terms of genetic diversity (Bataillon *et al.*, 1996). The number of private alleles found within cashew 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 cashew populations from different geographic regions, either by different allele frequencies or by unique alleles in each population. This is of importance in an agricultural crop like cashew, where the nuts are exported and processed in countries other than those where the product is originated and may lead to the valorization of a domestic product of a given geographical origin.

The ~~search~~ of 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 to a study done in Benin (West Africa), eight SSR markers were used to analyze sixty cashew morphotypes from three regions in the country (Chabi Sika *et al.*, 2013). This study revealed a low genetic diversity (Shannon index = 0.04) in the populations screened. In Brazil, in the Cerrado biome and coastal Restinga vegetation, 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 cashew populations screened (Kouakou *et al.*, 2020), which is in accordance with our results in AMOVA where most genetic diversity is depicted within populations.

Population structuring in East-Timor

AMOVA showed that most of the genetic diversity lies within individuals with little diversity present among individuals or among populations. The large proportion of diversity were 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 the heterozygous nature and high frequency of cross pollination in cashew. Cashew is primarily an allogamous species favoring cross-fertilization (Freitas & Paxton, 1996), thus allowing intraspecific hybridizations and enhancing genetic variations in cashew. 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 (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, individual-based clustering methods using a bayesian approach (STRUCTURE) and a multivariate analysis by DAPC allowed to assess the population structure, thus highlighting that genetic diversity scattering does not follows a clear geographic trend, despite a well-defined clustering observed between Mozambican populations with East-Timor/Indonesia, where few genetic diversity is shared. The population structuring using genetic distances revealed similar clustering with individual-based approaches, which supports results obtained under different analytical methods. The unique clustering attributed to Mozambican population may be related to incountry cashew varieties selection, different from the ones used by farmers in both Indonesia and East-Timor, and also due to the 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 strutting is observed which is linked to a district- associated genetic diversity. Also, surprisingly some East-Timor populations share allelic diversity with the Indonesia population specifically in Viqueque (ETV) and Manatuto (ETK and ETNA), and in less extent Baucau (ETTR3) and Manufahi (ETFA).

When excluding Mozambican populations from population structuring analysis, the genetic clustering of two groups can be depicted, where Viqueque, Manufahi and Bobonaro districts display a unique allelic diversity; while Baucau, Cova Lima and Manatuto has a high genetic similarity with Indonesia population. These dissimilar results when including vs excluding Mozambican populations, highlights the complex intraspecific genetic diversity of cashew in the context of a continental (Mozambique) country where cashew orchards have been implemented at long-term with improved varieties, while in Indonesia and East-Timor surely a different historical context of 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.

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 with the remaining districts suggests a diverse genetic build-up which could be attributed to different cashew varieties being planted.

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, Covalima (south) district showed the highest inbreeding coefficient ($F=0.26$). Both Bobonaro and Viqueque are districts where a significant share of exchange of planting ~~may occurred~~ in Bobonaro from Indonesia and in Viqueque from Australia, its genetic richness might have benefited from different introduced seeds and more varied germplasm imported than the other regions.

The difference between expected heterozygosity and observed heterozygosity might be due to evolutionary factors that occurred in the studied accessions or internal genetic factors (such as gene incompatibility). In addition, cashew grower's preference when they selected seeds for establishing new orchards might be one of the reasons. Indeed, when establishing new cashew orchards, some producers used seeds from a single tree with good traits (preferentially high-yielding 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 (0.014 ± 0.004) was indicative of a common origin of Ivorian cashew trees. 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 reported in other countries (e.g. Benin (Kouami, 2020), except in Brazil. In the case of East-Timor, two different scenario of cashew genetic diversity is conceivable, namely: i) possible multiple introductions from Bobonaro and Viqueque districts, where these regions may be 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 genetically highly divergent accessions among the studied accessions could be used in the heterosis breeding of cashew in future to improve yield, quality and other traits.

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

618 The wide distribution of cashew in its primary center of diversity in Brazil has been attributed to
 619 water currents in which the mature fruit will float in addition to the role played by bats in seed
 620 dispersal (Johnson, 1973). However, outside its center of origin, namely in India, cashew is
 621 located mostly in marginal lands devoid of flowing water and there are no reports of bird- or bat-
 622 mediated seed dispersal (Archak *et al.*, 2009). Hence, cashew distribution along the entire coastal
 623 region in India in a relatively short span, of around 450 years since its introduction, has been
 624 attributed to anthropogenic efforts, rather than through natural means alone (Archak *et al.*, 2009).
 625 Considering this important premise, the present genetic diversity analysis results are discussed
 626 accordingly.

627 The extent and distribution of genetic diversity as revealed by the present study provide some
 628 clues of cashew introduction mode and expansion in East-Timor. The existence of substantial
 629 overall genetic diversity and the genetic grouping into two distinct genetic groups in East-Timor
 630 point to multiple events of introduction comprising different founder populations. If the
 631 introduction occurred as a one-time event, founder effect would have been reflected in low
 632 genetic diversity. Furthermore, none of the genetic groups was confined to a particular
 633 geographic region, yet instead by a surprisingly district-associated genetic diversity trend. The
 634 country's current agricultural landscapes have been heavily influenced by the stimulation of food
 635 production mechanisms implemented by the Portuguese Agronomic Mission in the first three-
 636 quarters of the twentieth century, by promoting rice on the coastal plains and leaving the
 637 mountains for coffee (Metzner, 2017) . Nowadays, cashew expansion has been promoted mainly
 638 through the introduction of different varieties from Brazil, Indonesia, Australia and a so-called
 639 “native” cashew. According to the information obtained from the National Directorate of
 640 Industrial Crops and Agribusiness, the Ministry of Agriculture and Fisheries (MAFF) of East-
 641 Timor, the Portuguese brought the cashew to East-Timor (ET) in the 18th century and were
 642 planted as ornamental plants in various districts. Cashew was planted by communities in Maliana
 643 during the Indonesian occupation in 1983, with the help of a Timorese governor and agronomist
 644 Mário Viegas Carrascalão, who encouraged cashew plantations for smallholder benefits (Buss &
 645 Ferreira, 2010). After, cashew was developed as an industrial crop in the 1990's in ET under
 646 Indonesian occupation, mostly in poor and dry or semi-dry regions (Odete *et al.*, 2017) in 10
 647 districts with about 3,200 ha land area planted on 6,500 small farms (60% of them in the
 648 Bobonaro district). As of 2008, a significant decrease due to fires from the armed period in 1999,
 649 with about 800 ha of cashew trees remained growing in ten districts (Bobonaro, Manatuto,
 650 Oecusse, Cova Lima, Ainaro, Manufahi, Viqueque, Baucau, Lospalos and Díli). In Cova Lima
 651 district, cashew orchards were implemented in 1990s and few new cashew varieties have been
 652 introduced. In Manatuto and Baucau districts, cashew orchards were planted with accessions that
 653 farmers exchanged within region, namely those that gave higher yield and that were adapted to
 654 local agro-ecological conditions. In Bobonaro and Viqueque districts, during fieldwork, land
 655 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 that shaped current East-Timor cashew genetic diversity panorama.

It is generally believed that the Portuguese carried the cashew nut crop to Africa and India in the sixteenth century, and the Spanish probably took it to the Central American countries and the Philippines (Singh, 2018). Thus, cashew nut was introduced into India through the western coast, most probably Goa in the 16th century CE (1560) by the Portuguese (Sauer, 1993). From there, cashew was spread throughout Southeast Asia and parts of Africa, with Portuguese navigators in the 16th century brought seeds to India and Mozambique (Brücher, 1991) more as a source of wine and brandy than for the nuts. Only later, cashew was seen as a tree used for preventing soil erosion in the coastal region, until the commercial engagement into cashew nuts as we know today. Considering that India was an important country where cashew was first introduced in Asia as a commercial crop, in future studies inclusion of Indian cashew populations could help explain the genetic diversity within East-Timor and assist on the historical introduction in the country. As 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) within Nigerian cashew germplasm (Aliyu, 2012), thus highlighting a narrow genetic diversity within germplasm collection. This is particularly important as East-Timor aims to invest in the cashew nut exports by challenging directly with competitive countries, as India and Vietnam in the surrounding region. In fact, evaluation and assessing genetic diversity will allow to detect the presence of any East-Timor genetic signature that will acknowledge the valorization of cashew nuts from the country, and prospect agrobiodiversity hotspots for future cashew germplasm programs. Traditional cashew breeding has been slow, mainly due to the heterozygous nature and high frequency of cross pollination, hence necessitating the clonal propagation for maintaining the genetic integrity of a superior genotype (Johnson, 1973). Yet recent advances in cashew genomics may result in the development of new molecular tools and breeding strategies that increase genetic gains and speed the development of superior cultivars and genetic stocks (Savadi *et al.*, 2020).

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), which 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 and exploration 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, and to increase the genetic diversity of on-farm orchards at the short

run and to the identification of genetic resources for the development of cashew genetic management.

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 with 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 regarding to become a more important crop, 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 poorly 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 Fundação para a Ciência e Tecnologia (FCT) under the GenoCash project (PTDC/ASP-AGR/0760/2020).

Data Availability Statement

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

Conflicts of Interest

The authors declare that they have no competing interests.

References

- Ahmed OAMS, Saddi A. 2012. Am-Isometric operators in semi-Hilbertian spaces. *Linear Algebra and its applications*, 436(10), 3930-3942.
- Aliyu OM, Awopetu JA. 2007. Multivariate analysis of cashew (*Anacardium occidentale* L.) germplasm in Nigeria. *Silvae Genetica* 56:170–179. DOI: 10.1515/sg-2007-0026.
- Aliyu OM. 2012. Chapter 9: Genetic Diversity of Nigerian Cashew Germplasm. *In* Genetic Diversity in Plants, Çalışkan M. (Ed.). IntechOpen. Pp 163-184. <https://doi.org/10.5772/32892>.
- Archak S, Gaikwad AB, Swamy KRM, Karihaloo JL. 2009. Genetic analysis and historical perspective of cashew (*Anacardium occidentale* L.) introduction into India. *Genome* 52:222–230. DOI: 10.1139/G08-119.
- Barros A, Barnabé J, Guterres L, Monteiro F. 2022. Script for performing DAPC analysis applied to the study of population structure and genetic diversity in cashew from East-Timor. *Figshare Software*. DOI: <https://doi.org/10.6084/m9.figshare.19117889.v3>.
- Bataillon M, David J L, & Schoen D J. 1996. Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics*, 144 (1), 409-417.
- Borges D. 2018. Cultures du Timor-Oriental: processus d’objectification. *Plural Pluriel - revue des cultures de langue portugaise*. (Avalaible on <https://www.pluralpluriel.org/index.php/revue/issue/view/16>).
- Brownstein MJ, Carpten JD, Smith JR. 1996. Modulation of non-templated nucleotide addition by Taq DNA polymerase: Primer modifications that facilitate genotyping. *BioTechniques* 20:1004–1010. DOI: 10.2144/96206st01.
- Brücher H. 1991. *Useful plants of neotropical origin and their wild relatives*. DOI: 10.1016/0308-521x(91)90152-z.
- Buss PM, Ferreira JR. 2010. 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* 4:106–118. DOI: 10.3395/reciis.v4i1.351pt.
- Cavalli-Sforza LL, Edwards AW. 1967. Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics* 19:233–257. DOI: 10.2307/2406616.
- Chabi Sika K, Adoukonou-Sagbadja H, Ahoton LE, Adebo I, Adigoun FA, Saidou, A & Baba-Moussa, L. 2013. Indigenous knowledge and traditional management of cashew (*Anacardium*

- occidentale L.) genetic resources in Benin. *J. Exp. Biol. Agric. Sci*, 1(5), 375-382.
- 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.
- Excoffier L, Smouse PE, & Quattro J. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2), 479-491.
- 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.
- Government of East-Timor. 2015. Population and Housing Census 2015: Preliminary Results. Direção 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>.
- Guterres MO. (2017). Improvements in Growing Organic Cashews in Timor-Leste (Doctoral dissertation, Charles Darwin University (Australia)).

- 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 and Dried Fruits Statistical Yearbook 2019/20. Reus, Spain. Pp 1-80. Available online 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, Minhobo 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.
- Metzner JK. 2017. Man and Environment in Eastern Timor. Canberra, ACT: Development Studies Centre, The Australian National University.
- 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.

- 865 Año 1987.
- 866 Mneney EE, Mantell SH, Bennett M. 2001. Use of random amplified polymorphic DNA (RAPD)
- 867 markers to reveal genetic diversity within and between populations of cashew (*Anacardium*
- 868 *occidentale* L.). *Journal of Horticultural Science and Biotechnology* 76:375–383. DOI:
- 869 10.1080/14620316.2001.11511380.
- 870 Monteiro F, Catarino L, Batista D, Indjai B, Duarte MC, Romeiras MM. 2017. Cashew as a high
- 871 agricultural commodity in West Africa: Insights towards sustainable production in Guinea-
- 872 Bissau. *Sustainability* 9:1–14. DOI: 10.3390/su9091666.
- 873 Monteiro F, Romeiras MM, Figueiredo A, Sebastiana M, Baldé A, Catarino L, Batista D. 2015.
- 874 Tracking cashew economically important diseases in the West African region using
- 875 metagenomics. *Frontiers in Plant Science* 6:1–6. DOI: 10.3389/fpls.2015.00482.
- 876 Monteiro F, Vidigal P, Barros AB, Monteiro A, Oliveira HR. 2016. Genetic Distinctiveness of
- 877 Rye In situ Accessions from Portugal Unveils a New Hotspot of Unexplored Genetic
- 878 Resources. 7:1–17. DOI: 10.3389/fpls.2016.01334.
- 879 Nair KP. 2010. The Agronomy and Economy of Important Tree Crops of the Developing World.
- 880 *Elsevier Wordmark*. DOI: ISBN: 9780123846785.
- 881 Nei M. 1972. Genetic Distance between Populations. *The American Naturalist* 106:283–292. DOI:
- 882 10.1086/282771.
- 883 Nybom H. 2004. Comparison of different nuclear DNA markers for estimating intraspecific
- 884 genetic diversity in plants. *Molecular Ecology* 13:1143–1155. DOI: 10.1111/j.1365-
- 885 294X.2004.02141.x.
- 886 Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R
- 887 language. 20:289–290. DOI: 10.1093/bioinformatics/btg412.
- 888 Park SDE. 2001. Trypanotolerance in West African cattle and the population genetic effects of
- 889 selection [PhD thesis]. [Dublin (Ireland)]: University of Dublin.
- 890 Peakall R, Smouse PE. 2006. GENALEX 6: Genetic analysis in Excel. Population genetic software
- 891 for teaching and research. *Molecular Ecology Notes* 6:288–295. DOI: 10.1111/j.1471-
- 892 8286.2005.01155.x.
- 893 Pradhan C, Peter N, & Dileep N. 2020. Nuts as Dietary Source of Fatty Acids and Micro Nutrients
- 894 in Human Health. In V. Rao, L. Rao, M. Ahiduzzaman, & A. K. M. A. Islam (Eds.), Nuts and
- 895 Nut Products in Human Health and Nutrition. IntechOpen.
- 896 <https://doi.org/10.5772/intechopen.94327>
- 897 Pritchard J K, Stephens M, & Donnelly P. 2000. Inference of population structure using multilocus
- 898 genotype data. *Genetics*, 155(2), 945-959.
- 899 Putman AI, Carbone I. 2014. Challenges in analysis and interpretation of microsatellite data for
- 900 population genetic studies. *Ecology and Evolution* 4:4399–4428. DOI: 10.1002/ece3.1305.
- 901 R Core Team. 2021 R: A Language and Environment for Statistical Computing. R Found. Stat.
- 902 Comput. Vienna, Austria.
- 903 Rambaut A. 2014. FigTree v1. 4.2, a graphical viewer of phylogenetic trees. Available from left
- 904 angle bracket <http://tree.bio.ed.ac.uk/software/figtree/> right angle bracket.
- 905 República Democrática de Timor-Leste. 2011 East-Timor Strategic Development Plan 2011 –
- 906 2030. Pp 1- 228. [https://www.adb.org/sites/default/files/linked-documents/cobp-tim-2014-](https://www.adb.org/sites/default/files/linked-documents/cobp-tim-2014-2016-sd-02.pdf)
- 907 [2016-sd-02.pdf](https://www.adb.org/sites/default/files/linked-documents/cobp-tim-2014-2016-sd-02.pdf).
- 908 Rice WR. 2013. Analyzing Tables of Statistical Tests. *Evolution*, 43(1): 223-225.
- 909 Rosenberg NA. 2004. DISTRUCT: a program for the graphical display of population structure.
- 910 *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.
- Sauer JD. 1993. Historical Geography of Crop Plants. *CRC Press*. DOI: <https://doi.org/10.1201/9780203751909>.
- 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.
- Savadi S, Muralidhara BM, Preethi P. 2020. Advances in genomics of cashew tree: molecular tools and strategies for accelerated breeding. *Tree Genetics and Genomes* 16. DOI: 10.1007/s11295-020-01453-z.
- 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 the total number of individuals sampled by population (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). Total samples analyzed= 207.

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 (N_a), expected heterozygosity (H_e) and observed heterozygosity (H_o) with corresponding standard de-viation (SD) values, private alleles and inbreeding/fixation coefficient (F).

Table 5. AMOVA results including fixation indices F_{CT} , F_{SC} , and F_{ST} .

Figures

Figure 1. Geographical location of cashew populations sampled in the three tropical countries: Mozambique (A), Indonesia, in Kefamenanu (B) and East-Timor (C). Cashew orchards some of the plantations sampled: Natarbora-Manatuto district (D), Mailiana in Bobonaro district (E), Triloca in Baucau district (F), Fatucahi in Manufahi (G), Kefamenanu in Indonesia (H), Viqueque (I).

Figure 2. UPGMA (A) and NJ (B) trees generated from *FreeNA* using matrix DC^{INA} respectively, representing the population from East-Timor, Indonesia, and Mozambique. 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} respectively, representing the population from East-Timor and Indonesia. 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**).

Supplementary Materials

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

Table S2. Pairwise F_{ST} (lower-left matrix) and F_{ST}^{ENA} (upper-right matrix) between all populations of East-Timor, Indonesia and Mozambique.

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

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

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

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

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

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

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

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

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

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

Figure 1

~~Geographical location of cashew populations sampled~~ in the three tropical countries: Mozambique (A), Indonesia, in Kefamenanu (B) and East-Timor (C-I).

Cashew orchards of some the plantations sampled: Natarbora-Manatuto district (D), Mailiana in Bobonaro district (E), Triloca in Baucau district (F), Fatucahi in Manufahi (G), Kefamenanu in Indonesia (H), Viqueque (I).



Figure 1. Geographical location of cashew populations sampled in the three tropical countries: Mozambique, province of Sofala (A), Indonesia, in Kefamenanu (B) and East-Timor (C). Cashew orchards some of the plantations sampled: Natarbora-Manatuto district (D), Mailiana in Bobonaro district (E), Triloca in Baucau district (F), Fatucahi in Manufahi (G), Kefamenanu in Indonesia (H), Viqueque (I).

Figure 2

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

Legend: (□ Manatuto, □ Bobonaro, ■ Baucau, ■ Cova Lima, ■ Manufahi, ■ Viqueque, ♦ Indonesia, ▲ Mozambique).

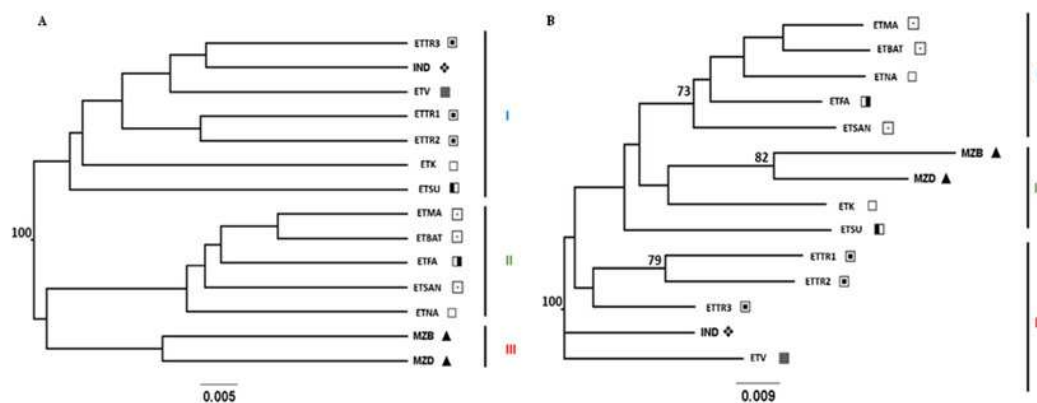


Figure 2. UPGMA (A) and NJ (B) trees generated from *FreeNA* using matrix DC^{INA} respectively, representing the population from East-Timor, Indonesia, and Mozambique. Legend: (□ Manatuto, □ Bobonaro, ■ Baucau, ■ Cova Lima, ■ Manufahi, ■ Viqueque, ♦ Indonesia, ▲ Mozambique).

Figure 3

UPGMA (A) and NJ (B) trees generated from *FreeNA* using matrix DC^{INA} respectively, representing the population from East-Timor and Indonesia.

Legend: (□ Manatuto, □ Bobonaro, ■ Baucau, ■ Cova Lima, ■ Manufahi, ■ Viqueque, ♦ Indonesia).

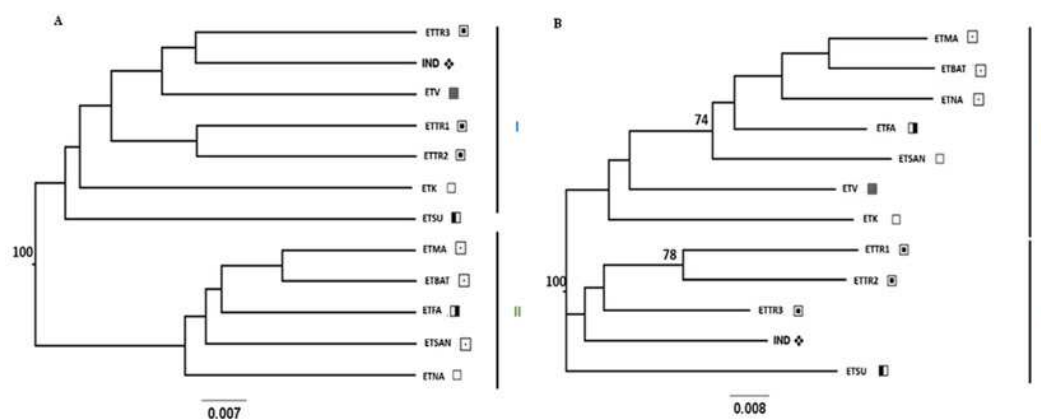


Figure 3. UPGMA (A) and NJ (B) trees generated from *FreeNA* using matrix DC^{INA} respectively, representing the population from East-Timor and Indonesia. Legend: (□ Manatuto, □ Bobonaro, ■ Baucau, ■ Cova Lima, ■ Manufahi, ■ Viqueque, ♦ Indonesia).

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).

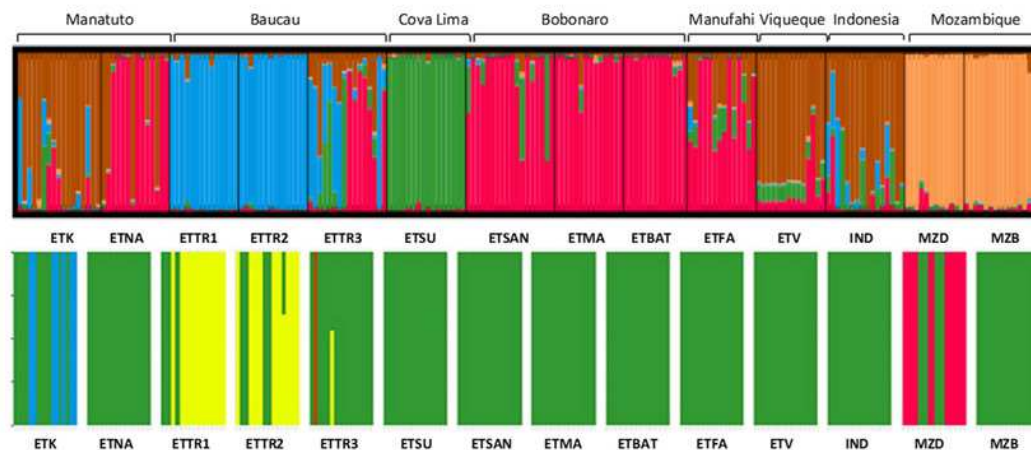


Figure 4. Clustering based on SSR data of the optimal K -means using STRUCTURE (A) for populations of Mozambique, East-Timor, and Indonesia; and DAPC analyses representation of $K = 5$ (B).

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).

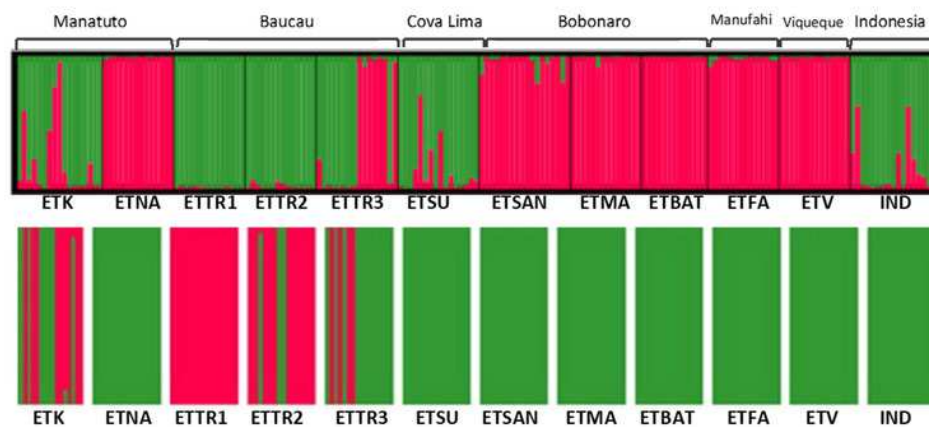


Figure 5. Optimal K - means individual-based clustering using STRUCTURE ($K = 2$, A) for populations of East-Timor and Indonesia and DAPC analyses of $K=2$ (B).

Table 1(on next page)

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

1 **Table 1.** Populations by country, district and location, geographical coordinates and the total
2 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
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

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'-TGTAACGACGGCCAGT-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: CAAACTAGCCGGAATCTAGC R: <u>GTTTCTTCCCCATCAAACCTTATGAC</u>	D2	118-186 (143)	A
mAoR17	(GA) ₂₄	F: GCAATGTGCAGACATGGTTC R: <u>GTTTCTTGGTTTCGCATGGAAGAAGAG</u>	D1	122-184 (124)	
mAoR7	(AT) ₂ (GT) ₅ AT(GT) ₅	F: AACCTTCACTCCTCTGAAGC R: <u>GTTTCTTGTGAATCCAAAGCGTGTG</u>	D4	158-198 (178)	
mAoR48	(GAA) ₆ (GA) ₃	F: CAGCGAGTGGCTTACGAAAT R: <u>GTTTCTTGACCATGGGCTTGATACGTC</u>	D3	130-186 (177)	
mAoR3	(AC) ₁₂ (AAAAT) ₂	F: CAGAACCCTCACTCCACTCC R: <u>GTTTCTTATCCAGACGAAGAAGCGATG</u>	D4	140-282 (231)	B
mAoR42	(CAT) ₉ TAT(CTT) ₇	F: ACTGTCACGTCAATGGCATC R: <u>GTTTCTTGCGAAGGTCAAAGAGCAGTC</u>	D2	160-232(204)	
mAoR52	(GT) ₁₆ (TA) ₂	F: GCTATGACCTTGGGAACTC R: <u>GTTTCTTGTGACACAACCAAAACCACA</u>	D1	142-244 (202)	
mAoR11	(AT) ₃ (AC) ₁₆	F: ATCCAACAGCCACAATCCTC R: <u>GTTTCTTCTTACAGCCCCAACTCTCG</u>	D3	142-248 (234)	
mAoR2	(CA) ₁₀ (TA) ₆	F: GGCCATGGGAAACAACAA R: <u>GTTTCTTGGAAGGGCATTATGGGTAAG</u>	D3	172-322 (366)	C
mAoR33	(CT) ₁₈ (AT) ₁₉	F: CATCCTTTTGCCAATTAACAA R: <u>GTTTCTTCACGTGTATTGTGCTCACTCG</u>	D4	322-404 (354)	
mAoR35	(AG) ₁₄	F: <u>TCTTTCGTTCCAATGCTCCTC</u> R: <u>GTTTCTTCATGTGACAGTTCGGCTGTT</u>	D2	142-180 (165)	
mAoR47	(TAAA) ₂ (TA) ₇ (AAT)	F: AAGAGCTGCGACCAATGTTT R: <u>GTTTCTTCTTCTTGAACCTGACACTTCATCCA</u>	D1	166-272 (161)	
mAoR12	(AC) ₁₂ ATAC(AT) ₄	F: CACCAAGATTGTGCTCCTG R: <u>GTTTCTTAAACTACGTCCGGTCACACA</u>	D2	322-362 (324)	D
mAoR16	(GT) ₈ (TA) ₁₇ (GT) ₃	F: GGAGAAAGCAGTGGAGTTGC R: <u>GTTTCTTCAAGTGAGTCTCTCACTCTCA</u>	D1	245-335 (256)	
mAoR29	(TG) ₁₀	F: GGAGAAGAAAAGTTAGGTTTGAC R: <u>GTTTCTTCGTCTTCTTCCACATGCTTC</u>	D3	164-364 (316)	
mAoR41	(ACC) ₇ (AC) ₃	F: GCTTAGCCGGCAGCATATTA R: <u>GTTTCTTAGCTCACCTCGTTTCGTTTC</u>	D4	162-177 (151)	

3

4

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). Total samples analyzed= 207.

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). Total samples analyzed= 207.

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
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 (N_a), expected heterozygosity (H_e) and observed heterozygosity (H_o) with corresponding standard deviation (SD) values, private alleles 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 and inbreeding/fixation coefficient (F).

	N	Na	Na SD	H_e	H_e SD	H_o	H_o SD	Private Alleles	F
Countries									
East Timor (ET)	164	10.67	4.96	0.670	0.033	0.506	0.012	6.17	0.24
Indonesia (IND)	16	3.42	1.93	0.561	0.063	0.465	0.037	0.42	0.12
Mozambique (MZ)	27	5.50	2.28	0.711	0.028	0.535	0.028	1.92	0.25
Mean	69	6.530	3.057	0.647	0.0413	0.502	0.0257	2.837	0.203
Populations									
ETK	17	5.17	1.75	0.67	0.04	0.56	0.04	1.08	0.14
ETNA	14	3.75	1.14	0.65	0.03	0.56	0.04	0.17	0.10
ETTR1	14	3.83	1.53	0.65	0.04	0.57	0.04	0.33	0.06
ETTR2	14	3.67	1.72	0.58	0.06	0.49	0.04	0.50	0.11
ETTR3	16	5.08	2.54	0.59	0.06	0.51	0.04	0.67	0.07
ETSU	16	3.58	1.44	0.65	0.04	0.46	0.04	0.58	0.26
ETSAN	18	3.58	1.83	0.57	0.05	0.57	0.03	0.33	-0.04
ETMA	14	3.00	0.85	0.57	0.03	0.46	0.04	0.00	0.14
ETBAT	13	2.42	0.79	0.48	0.05	0.45	0.04	0.00	0.01
ETFA	14	2.75	0.75	0.51	0.05	0.43	0.04	0.08	0.10
ETV	14	2.25	0.62	0.45	0.06	0.45	0.04	0.08	-0.09
IND	16	3.42	1.93	0.56	0.06	0.47	0.04	0.42	0.12
MZB	12	4.33	1.50	0.69	0.03	0.51	0.04	1.00	0.23
MZD	15	3.50	1.17	0.65	0.03	0.56	0.04	0.50	0.12
Mean	14.79	3.60	1.40	0.59	0.05	0.50	0.04	0.41	0.09

Table 5(on next page)

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

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}

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} . $*p < 0.001$.