

Genetic diversity and population structure of cacao (*Theobroma cacao* L.) germplasm from Sierra Leone and Togo based on KASP- SNP genotyping

Ranjana Bhattacharjee (✉ r.bhattacharjee@cgiar.org)

International Institute of Tropical Agriculture

Mohammed Mambu Luseni

Swedish University of Agricultural Sciences

Komivi Ametefe

Institut Togolais de recherche Agronomique, Centre de Recherche Agronomique-zone Forestière (CRA-F)

Paterne A. Agre

International Institute of Tropical Agriculture

P. Lava Kumar

International Institute of Tropical Agriculture

Laura J. Grenville-Briggs

Swedish University of Agricultural Sciences

Research Article

Keywords: cacao, Sierra Leone, Togo, KASP-SNPs, genetic diversity

Posted Date: September 15th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3345739/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Cacao (*Theobroma cacao* L.) is a tropical tree species belonging to Malvaceae family and originated in the lowland rainforests of the Amazon. It is one of the major agricultural commodities contributing towards the GDP (gross domestic product) of West African countries contributing about 70% of world's production.

Methodology

A sample of 376 cacao (*Theobroma cacao* L.) germplasm accessions (235 from Sierra Leone and 141 from Togo) collected from different seed gardens and farmers' fields across the cacao producing states/regions of these countries were assembled for genetic diversity and population structure studies using single nucleotide polymorphism (SNP) markers. The 376 cacao accessions were genotyped using 20 highly informative and reproducible KASP-SNPs markers. The genetic diversity among different accessions was assessed using three complementary clustering methods including the model-based population structure, discriminant analysis of principal components (DAPC), and phylogenetic trees. A network analysis was further carried out to understand the genetic relationships among cacao accessions from Sierra Leone and Togo. Analysis of molecular variance (AMOVA) was carried out to estimate the genetic differentiation among and within groups.

Results

The STRUCTURE and DAPC exhibited some consistency in the allocation of accessions in subpopulations or groups, although some discrepancy was identified in the grouping based on these two methods. Hierarchical clustering analysis grouped all the individuals into two major groups and represented several sub-clusters. AMOVA revealed high genetic diversity (86%) within accessions. A high rate of mislabeling/duplicate genotype names was revealed in both countries, which may be attributed to errors from the sources of introduction, labeling errors, and lost labels.

Conclusion

This study demonstrates the use of KASP-SNPs for fingerprinting that can help identify mislabeled accessions and provide strong evidence for improving the accuracy and efficiency in cacao germplasm management as well as distribution of correct materials to farmers.

Introduction

Cacao, *Theobroma cacao* L., is a tropical tree native to humid tropics of the central and northern parts of South America. It is the major ingredient used in the multi-billion-dollar chocolate and confectionaries industry as well as other intermediary products such as cacao butter, cacao powder, cacao cake, and cacao liquor. The coastal countries in West Africa, known as the West Africa cacao belt [1], from Sierra Leone, Guinea, and Liberia to Southern Cameroon, apart from Benin Republic, are responsible for production of about 70% of the world's cacao (<http://faostat.fao.org>). The chocolate and confectionary industry around the world is heavily dependent on cacao beans from West African countries, both due to the high production and the high quality of beans (bulk cacao, may not be specialty cacao) that are not produced in other cacao producing regions such as Asia or Central and Southern America. In 2020, cacao was the primary agricultural commodity that was exported by Cote d'Ivoire, Ghana, Nigeria, Cameroon and Sierra Leone, and the second most important commodity exported from Guinea, Togo, and Liberia, thus contributing towards the GDP (gross development price) of these countries (www.statista.com).

It is estimated that the Brazilian cacao (Amelonado type which is also known as Lower Amazon Forastero type) was first introduced into West Africa in the 19th and early 20th centuries by the Portuguese [2], and since then cacao is cultivated by smallholder farmers in this region. Cote d'Ivoire and Ghana remain the highest producers of cacao in the world accounting for over 60% of global world production of around 4.9 million tonnes in the 2021/2022 cacao season [3–4]. Among cacao producing countries in the world, Togo and Sierra Leone ranks 15th and 17th with a production of 22,522 and 14,670 metric tonnes, respectively [5]. In 2017, the World Bank Trade Statistics recorded an export earnings of about US\$ 14,461 million from cacao beans in Sierra Leone [6] even with low productivity. Over the years, the cacao yield has increased in both countries, which corresponds to an increase in the area under cultivation. However, the future yield is expected to be adversely affected by changing climatic conditions. Similar to other West African countries, cacao cultivation in Sierra Leone and Togo faces the challenges of old trees, aged farmers, black pod disease, mirids, poor access to improved planting materials, and other challenges (such as cacao swollen shoot virus disease) related with its cultivation, and management.

Cacao germplasm introduction in Sierra Leone and Togo followed the same trend as introductions in other West African countries, which is from a common source – Fernando Po. However, it is believed that in Sierra Leone, there may have been introductions from other sources (the West Indies) raising questions as to the origin of cacao in this country [7]. There is, therefore, an interesting probability that the present-day cacao in West Africa (apart from recent introductions through University of Reading and germplasm exchanges between West African countries such as introductions from Ghana to both Sierra Leone and Togo (at experimental stations of national institutes) is of dual origin. In Togo, cacao germplasm is conserved in a gene bank consisting of clones introduced from countries in the sub-region (Ghana, Côte d'Ivoire, Cameroon, and Nigeria) and international collections (University of Reading, United Kingdom). It is necessary to understand and assess the genetic relationships and genetic diversity among cacao germplasm within and between these two countries. Several efforts have been put forward to understand and assess the genetic diversity currently available in major cacao growing countries in the world, including several West African countries using both molecular markers and morphological traits [8–12].

However, there is no study available so far that targeted germplasm collection from different cacao growing regions in Sierra Leone and Togo to assess genetic diversity and understand population structure present in these two countries.

There is a general agreement that the cultivated cacao in West Africa, the region with highest production, has a narrow genetic base and faces issues of mislabeling. Studies that compared both modern and historical introductions did not detect significant genetic diversity and improvements in yield or pest and disease resistance during the last 20 years [13]. Any improvement in yield resulted from better management practices, although some of the recent varieties or hybrids developed in major West African countries produced significantly higher yields with better pest and disease resistance under intensive cultivation. Similar improvements in yield as well as other traits are needed in other countries so that there will be enough supply for the increasing demand of dry cacao beans in the world. Therefore, it is necessary that the cacao germplasm available in other producing countries within West Africa is characterized, documented, identified, conserved, and utilized. This will allow searches for promising unique genetic materials that can be used in developing strong breeding programs in these countries as well, which currently depend on Cote d'Ivoire, Ghana or Nigeria for improved varieties or planting materials.

The assessment of genetic diversity or identification of mislabeled accessions in germplasm collections with single nucleotide polymorphism (SNP) markers has been used extensively in several crops and is being adopted as a fast, high-throughput, and affordable tool for whole-genome genetic diversity analysis. SNPs have been successfully used to characterize crops such as maize [14], soybean [15] and several others. Similarly, SNP markers have been used for fingerprinting cacao germplasm collection in several studies [16–20]. To our knowledge, KASP-SNPs have been used in only two studies for cacao germplasm from West Africa [10, 11]. High-throughput sequence-based SNP markers such as the KASP (Kompetitive allele-specific PCR) platform emerged as an attractive method in comparison to traditional genotyping with SNP markers using electrophoresis systems because of low genotyping error rate, and flexibility to automation that resulted in drastic reduction in cost per data point [21]. In the present study, a subset of 20 KASP-SNPs, from a set of 100 KASP-SNPS used in genotyping cacao samples from Nigeria and Ghana, was used for genotypic characterization of cacao from Sierra Leone and Togo. This is the first study that aimed to use the subset of 20 selected KASP-SNPs to understand the extent of genetic diversity and population structure within and among cacao germplasm.

Results

Cacao accessions from Sierra Leone and Togo

The cacao accessions collected from both Sierra Leone and Togo represented several clones with the same name (Supplementary Table S1) even when these were collected from different trees either from the same farm or from different farms. The GPS coordinates as well as the village/farm names and the district name for each sample are provided in Supplementary Tables S2a and S2b. The majority of the

cacao samples from Sierra Leone with similar names were collected from trees located in different farms, indicating that these trees either belong to the same clone or there are issues of mislabeling. The situation is the same in Togo although samples with the same genotype names are present in the same farm as well as in different farms.

The 235 cacao samples from Sierra Leone represented 144 accessions with unique names. These 144 accessions were sourced from either Ghana (66 accessions) or from the University of Reading (77 accessions) while the source of one sample was unknown (Supplementary Table S1). Similarly, the samples from Togo represented 77 accessions with unique names. These 77 accessions were sourced from Cameroon (4 accessions), Cote d'Ivoire (11 accessions), Ghana (17 accessions), Nigeria (28 accessions), Togo (9 accessions), with a few clones representing both Cameroon and Ghana (ICS 6 x NA33; ICS16), Cote d'Ivoire and Cameroon (IFC1 x SNK13), Ghana and Cote d'Ivoire (NA2), Ghana and Nigeria (SCA 12 x NA32; C23; C26 x SCA 6), and five samples whose accession names were lost (Supplementary Table S1). For the accessions with lost names, it is assumed that these may represent PA7/A19 clones (Supplementary Table S1). There are eleven accessions/accession names common between both Sierra Leone and Togo, of which PA7, ICS60, IMC47, Na34 and Pa35 represented the University of Reading and Nigeria as the source for Sierra Leone and Togo, respectively. C20 and C70 represented Ghana as the source country for both Sierra Leone and Togo; C26 and C42 represented Ghana and Nigeria as source country for Sierra Leone and Togo, respectively; C23 represented Ghana for Sierra Leone and Nigeria/Ghana for Togo as the source country; Pound 7 represented Gene Bank of SLARI and Ghana as the source for Sierra Leone and Togo, respectively while PA7 represented Gene Bank of SLARI and Nigeria as the source (Supplementary Table S1).

Genetic diversity parameters

The 20 KASP-SNP markers used in the present study were distributed across the ten chromosomes of cacao, with two SNPs on each chromosome (Supplementary Table S3). The average PIC (polymorphic information content), H_e (expected heterozygosity), H_o (observed heterozygosity) and MAF (minor allele frequency) values for 235 cacao accessions collected from Sierra Leone were 0.22, 0.30, 0.24, and 0.21, respectively while for 141 accessions collected from Togo these values were 0.21, 0.29, 0.22, and 0.19, respectively. For the combined population of 376 cacao accessions, the average PIC, H_e , H_o , and MAF values were 0.24, 0.30, 0.26, and 0.23, respectively (Table 1). Supplementary Fig. 1 represents the distribution of H_e , H_o , MAF, and PIC values for the combined population. Of the 20 SNPs used in the study, many had a minor allele frequency (MAF) above 0.2 and showed high PIC values with a peak distribution above 0.2. Similarly, low observed and expected heterozygosity was recorded. The observed heterozygosity for the combined population was higher than individual country-level observed heterozygosity, and this is also represented by higher PIC values (Table 1).

Table 1
Descriptive statistics based on 20KASP-SNP markers across cacao accessions

Cacao Population	N	He	Ho	MAF	PIC
Sierra Leone	235	0.30	0.24	0.21	0.22
Togo	141	0.29	0.22	0.19	0.21
Combined	376	0.30	0.26	0.23	0.24
N: Number of accessions					

Population Structure and Genetic Relationships

The model-based population structure analysis combined cacao population (235 accessions from Sierra Leone and 141 accessions from Togo) showed that the delta K values from the mean log-likelihood probabilities stagnated at $K = 4$ (Fig. 1a). The 376 cacao accessions divided into four subpopulations at $K = 4$ (Fig. 1b). Based on an 80% membership probability threshold, 270 accessions (73.37%) were successfully assigned to the four sub-populations. In comparison, 106 accessions with a probability less than 80% were designated as an admixed population (Supplementary Table S1). Sub-population 1 consisted of 123 accessions (Sierra Leone: 94 accessions; and Togo: 29 accessions). Sub-population 2, 3 and 4 constituted 3.72%, 22.07% and 13.30% of the accessions, respectively with a composition of 10 accessions from Sierra Leone and 4 accessions from Togo in sub-population 2, 36 accessions from Sierra Leone and 47 accessions from Togo in sub-population 3, and 26 accessions from Sierra Leone and 24 accessions from Togo in sub-population 4 (Supplementary Table S1, Fig. 1b). The admixed group consisted of 69 accessions from Sierra Leone and 37 accessions from Togo. Few additional smaller peaks observed (Fig. 1b) implied the presence of subgroups within the four major subpopulations. Therefore, individual STRUCTURE analysis was performed for accessions representing Sierra Leone and Togo. Sub-clustering of cacao germplasm from Sierra Leone and Togo showed that delta K values stagnated at $K = 3$ and $K = 4$, respectively (Fig. 2a, b). A substantial degree of admixture was observed in cacao germplasm from Togo in comparison to the cacao germplasm from Sierra Leone. The 235 cacao accessions from Sierra Leone divided in three subpopulations with 102 accessions in subpopulation 1, 29 in subpopulation 2 and 93 in subpopulation 3, while only 11 accessions (4.7%) were in the admixed group (Supplementary Table S1; Fig. 2a). Similarly, 141 cacao accessions from Togo grouped into four subpopulations with 4 accessions in subpopulation 1, 36 in subpopulation 2, 50 in subpopulation 3 and 25 in subpopulation 4. The admixed group consisted of 26 accessions, representing 18.4% admixture among the accessions (Supplementary Table S1; Fig. 2b).

Using the Bayesian information criterion (BIC) implemented in DAPC, a maximum of $K = 4$ was obtained, which corresponded to four groups obtained for the combined population and for the germplasm collection from Togo (Fig. 3). Estimation of cluster membership revealed that cluster 3 consisted of the highest number of accessions (158) followed by cluster 2 with 116 accessions, cluster 1 with 86 accessions, and cluster 4 had the least number of accessions (16). Out of 158 accessions in cluster 3,

105 accessions (66.5%) were from Sierra Leone and 53 (33.5%) from Togo (Supplementary Table S1). The cluster 2 represented 51 accessions from Sierra Leone and 65 accessions from Togo, while cluster 1 had 68 accessions from Sierra Leone and 18 accessions from Togo. The smallest cluster 4 represented 11 (68.8%) accessions from Sierra Leone and 5 (31.3%) accessions from Togo.

Contrary to the results of STRUCTURE and DAPC, the hierarchical clustering assigned all the 376 cacao accessions to two major clusters with several sub-clusters representing a higher degree of admixture among the accessions across both countries (Fig. 4a). The hierarchical clustering performed for cacao accessions from Sierra Leone and Togo independently is represented in Fig. 5b. The cacao accessions from Sierra Leone divided into three main clusters [Fig. 4b (i)] while that from Togo divided into four main clusters [Fig. 4b (ii)] (Supplementary Table S1), which is in line with the results obtained with STRUCTURE and DAPC analyses. For Sierra Leone, cluster 1 was the largest with 119 cacao accessions while 72 accessions and 44 accessions grouped in cluster 2 and 3, respectively. For Togo, cluster 3 was the largest with 52 cacao accessions followed by cluster 1 (46 accessions), cluster 2 (29 accessions) and cluster 4 with 14 accessions (Supplementary Table S1). A network analysis between cacao accessions from Sierra Leone and Togo (Fig. 5) showed a strong genetic relationship indicating that the cacao accessions share a similar genetic background between these two countries. The central core of the QGRAPH represented a set of cacao accessions that are genetically similar to each other.

The comparison of all three methods (STRUCTURE, DAPC and hierarchical clustering) did not reveal any common clustering patterns among 376 cacao accessions except two accessions which were both from Sierra Leone and represented the same accession name (C77) (Supplementary Fig. 2; Supplementary Table S1). However, the DAPC and hierarchical clustering showed similar patterns of clustering among 135 out of 376 cacao accessions in which 127 accessions were from Sierra Leone and 8 accessions were from Togo, while the DAPC and STRUCTURE analyses showed similar patterns of grouping among 61 out of 376 cacao accessions (30 from Sierra Leone and 31 from Togo) (Supplementary Fig. 2).

Analyses of Molecular Variance and Genetic Differentiation

The AMOVA analysis revealed a variability of 86% within accessions and 14% among populations (Combined data) (Table 2). The overall F_{ST} value was 0.601. A significant level of population divergence based on pairwise F_{ST} ($p < 0.0001$) was also observed between different populations while strong genetic relationships with some divergence observed within each population (Table 3). The average F_{ST} -based population differentiation was highest for the combined population (0.096) and lowest for cacao accessions from Sierra Leone (0.049). The pairwise F_{ST} values ranged from 0.010 (combined population vs. Togo) to 0.045 (Sierra Leone vs. Togo).

Table 2
Analysis of molecular variance (AMOVA) among and within three different genetic populations (Sierra Leone, Togo and Combined)

Source	d.f.	SS	MS	Est. Var.	% Var.	<i>p</i> Value
Among populations	3	189.16	63.05	0.362	14	0.001
Among accessions	372	690.62	1.86	0.000	0	0.001
Within accessions	376	809.50	2.15	2.153	86	0.001
Total	751	1689.28		2.515	100	
Fixation index (F_{ST})	0.601					0.001
d.f.: Degree of freedom, SS: Sum of squares, MS: Mean sum of squares, Est. Var.: Estimated variance, % Var.: Percent Variance						

Table 3
Pairwise fixation index (F_{ST}) values among the three populations of cacao accessions

	Sierra Leone	Togo	Combined Population
Sierra Leone	-0.0024		
Togo	0.0451	-0.0040	
Combined Population	0.0201	0.0101	-0.0015
Average F_{ST}	0.049	0.073	0.096

Discussion

The results revealed that genotyping of cacao germplasm collection from Sierra Leone and Togo using 20 KASP-SNPs was highly efficient in assessing the genetic diversity and population structure. The 20 SNPs used were a subset of 100 KASP-SNPs carefully selected based on their polymorphic nature, reproducibility and higher efficiency from earlier studies used in fingerprinting cacao germplasm collection from Ghana [10], Nigeria [11] and other West African countries (unpublished). The biallelic nature of these selected SNPs has a lower error rate in allele calling with higher accuracy and efficiency.

The germplasm studied represent collections maintained in different seed gardens as well as farmers' fields in both countries. Currently, cacao cultivation is unstructured in Sierra Leone with cacao trees labeled as Foresterio as the main variety cultivated in the country [22]. However, the other varieties grown are called as 'Amazon cacao', 'Ghanian cacao' and 'Ivorian cacao' with significant exchange of planting materials among the farmers. The situation is similar in Togo. A thorough assessment of available genetic diversity and population structure is necessary for genetic enhancement and utilization of well-characterized and diverse germplasm in cacao improvement programs in these two countries.

In the present study, diversity indices revealed the presence of substantial genetic diversity in the cacao germplasm from both Sierra Leone and Togo indicated by average H_e (0.30 and 0.29, respectively) and H_o (0.24 and 0.22, respectively). Similar H_e was reported for cacao germplasm collections in other West African countries such as Nigeria [11] and Ghana [10] although these studies recorded higher observed heterozygosity in comparison to the current study. This difference in observed heterozygosity detected in the current study could be explained either by the lower number of KASP-SNPs used or may reflect differences in the different genetic materials. Nonetheless, the average H_e and H_o for the combined (Sierra Leone and Togo) population in the current study was comparable to that obtained for each country individually, indicating that the cacao germplasm present in both countries may share the same genetic background. The study did reveal that eleven accession names are common between both countries and the presence of mislabeled cacao accessions in seed gardens as well as farmers' fields has already been reported in most West African countries [9–11, 23]. It is also obvious that a set of common cacao germplasm is shared by major cacao producing West African countries such as Cote d'Ivoire, Ghana, Nigeria, and Cameroon including other smaller countries in the region and beyond, indicating that mislabeled and duplicated cacao germplasm may be shared with these smaller countries.

The assessment of genetic relatedness, performed by three different analyses (model-based population structure, IBS-based clustering and DAPC), revealed that the combined population (including cacao accessions from Sierra Leone and Togo) and the germplasm collection from Togo was composed of four main subpopulations while that from Sierra Leone was composed of three main subpopulations. The fact that this clustering was supported by a Bayesian approach, a genetic distance-based method, and a DAPC-based analysis, provides a strong justification for the observed population structure and genetic relationships among the cacao accessions. It is worth noticing that in the combined analysis, most cacao accessions (73.37%) were assigned to one of the four subpopulations with probabilities higher than 0.8 and only 106 out of 376 cacao accessions (28.19%) were admixtures. In a recent study of cacao germplasm from Nigeria, the rate of admixtures or off-types ranged from 10–73% among the clones in the seed gardens [11]. These kinds of admixtures and presence of off-types in cacao germplasm from both Sierra Leone (4.7%) and Togo (18.4%) probably indicates labeling errors present in introduced germplasm from other West African countries. This is also supported by the fact that lower mislabeling was observed for cacao accessions from Sierra Leone (33.33%) in comparison to Togo (45.45%). Additionally, the presence of higher level of admixed accessions or mislabeling in cacao samples from Togo probably reflects recent breeding advances involving open pollination and bi-parental crossing among the accessions coupled with strong selection pressure [24]. The labeling errors pose a higher risk of misidentification of clones which further gets multiplied when beans/pods from such clones are shared or used for the establishment of new seed gardens or when neighboring farmers share materials with each other, or farmers bring materials from neighboring countries without knowing the genetic potential, which is a common practice in both Sierra Leone and Togo. Olasupo et al [11] explained how the presence of 58% mislabeled accessions in an old seed garden translated into 100% mislabeling in a newly established seed garden in Nigeria. The presence of mislabeling in introduced germplasm from the University of Reading into Sierra Leone and the seed gardens of National Institute in Togo indicates the

loss of labels or human errors in writing the labels or misidentification, which is commonly observed in clonally propagated crops with long life cycles such as tree crops. There is an enormous implication of introduction of mislabeled or misidentified germplasm as it leads to improper prediction in the performance of such accessions and further selection in the improvement program. This may be one of the reasons for lower yields of these introduced cacao germplasm from neighboring countries, in addition to the prevalence of pests and diseases.

1. Network analysis was used to unravel the genetic relationships among the cacao accessions from Sierra Leone and Togo. In the absence of pedigree records and comparison to any international reference clones, the dissection of genetic relationships among cacao accessions from these two countries through network analysis was a reliable methodology. Network analysis has been successfully used in other clonally propagated food crops such as cassava [25] and white yam [26]. In tree crops such as cacao, which is an outbreeding species with a long life cycle, open pollination is a common process wherein the source of pollen is often unknown. The extent of genetic diversity observed within cacao accessions in the present study could be attributed to this factor as well as identification of a good number of unique accessions in both the countries. Hence, these unique cacao accessions in each country can be used as parents in the cacao improvement programs after the preliminary evaluation of these accessions in multi-locations and trait profiling have been carried out. There is also a need for field conservation of unique cacao accessions in each country with provision for good management practices and irrigation facilities to mitigate against loss of accessions associated with biotic and abiotic stresses. It may also be necessary to mainstream DNA fingerprinting of introduced germplasm from neighboring countries and international collections for regular auditing of cacao accessions for their true-to-typeness and also to check for pollen contamination during hybridization through open or manual pollination.

Conclusions

In this study, we made use of KASP-SNP markers to assess the genetic diversity of cacao germplasm from Sierra Leone and Togo. The genetic relationships elucidated among the accessions in each country as well as identification of mislabeling has provided the knowledge to carefully explore future cacao improvement by identifying diverse and correct parents. This study also confirmed the reliability and accuracy of KASP-SNPs generated from next-generation sequencing-based genotyping coupled with complementary statistical analyses to generate knowledge on genetic diversity and population structure. Further studies are needed for multi-location phenotyping as well as genotyping with high-density DNA markers. In this study, we identified a high degree of mislabeling in most of the introduced materials, which has been attributed to errors from the sources of introduction, labeling errors, and lost labels. This has a consequence on the improvement programs and caution should be taken in the future, to ensure accuracy in labeling and identification of the clones/accessions before establishing seed gardens and distribution of planting materials to farmers.

Materials and Methods

Cacao sampling

A total of 235 and 141 cacao accessions respectively from Sierra Leone and Togo were sampled from different cacao growing regions (Supplementary Table S1). In Sierra Leone, samples were collected from the existing cacao farms across the three major cacao producing districts: Kenema (N: 08°02.805', W: 011°02.032'), Kailahun (N: 08°21.570', W: 010°23.477'), and Kono (N: 08°36.196', W: 010°56.458'). In addition, samples were collected from minor cacao producing districts such as Bo (N: 08°10.653', W: 011°41.720') and Pujehun Districts (N: 07°08.684', W: 011°22.652'). Samples were also collected from abandoned trees at Njala University research garden in Moyamba district (N: 08°06.493', W: 012°04.950') (Fig. 6a). Similarly, in Togo, samples were collected from 15 cacao plantations located in the areas of Agou (N: 6°49'18", E : 0°52'04"), Kloto (N: 6°82', E : 0°62'), Danyi (N: 6°49'40,2", E : 0°43'07") and Litimé (N: 7°59' ; E : 0°63') (Fig. 6b). Individual trees were tagged and geo-referenced from which fresh young cacao leaf samples were collected and dried using silica gel. The samples were then shipped to the Bioscience Center of the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. At IITA, six leaf discs of approximately 5 mm diameter from dried leaf samples of each genotype was punched into labeled 8-strip 1.1 ml propylene tubes with strip caps, up to 12 strips placed on 96-well boxes. A total of four labeled 96-well boxes were then shipped to a genotyping service provider (Intertek, Sweden) for automated DNA extraction and genotyping with 20 SNP markers using the KASP assay. Two blank controls were included in each box during genotyping.

DNA extraction, preparation and genotyping

DNA quality and quantity was checked on a 0.8% agarose gel. A total of 20 high quality SNPs were used to genotype the cacao samples. These 20 SNPs belongs to a set of 1536 SNPs identified from previous studies that used expressed sequence tags (ESTs) of a wide range of cacao tissue and organs displaying differences in the transcriptome [27, 28]. A total of 100 SNPs were then selected from this set of 1536 SNPs based on call rate, representativeness across the ten chromosomes and heterozygosity, and their use by cacao researchers for various studies [18, 29, 30]. The selected 100 SNPs were then converted using KASP™ assays at LGC Genomics (<http://www.lgcgroup.com/kasp>) in previous studies on cacao genetic diversity from West Africa [10, 11]. For the present study, a subset of 20 highly polymorphic and high quality KASP-SNPs (Supplementary Table S3) was selected from 100 KASP-SNPs, and genotyping was carried out at Intertek, Sweden. The KASP assay protocol followed the KASP manual [31] in which genotyping was carried out using high-throughput PCR SNPLine workflow using 1 µL reaction volume in 384-well PCR plate. The KASP genotyping reaction mix consisted of three components including sample DNA (10 ng), marker assay mix comprising target-specific primers, and KASP-TF™ master mix containing two universal fluorescence resonant energy transfer cassettes (FAM and HEX), passive reference dye (ROX™), Taq polymerase, free nucleotides, and MgCl₂ in an optimized buffer solution. The SNP assay mix is specific to each marker and consisted of kompetitive allele-specific forward and reverse primer. After PCR, the plates were fluorescently read, and allele calls were made using KRAKEN™ software (LGC

Biosearch Technologies, UK) and scored on a Cartesian plot, (cluster plot), in order to assign each DNA sample to a genotype class, namely homozygous for either allele 1 or 2 or heterozygous in case of biallelic SNPs.

Data analyses

Genetic diversity analyses carried out using minor allele frequency (MAF), polymorphism information content (PIC), expected heterozygosity (He) and observed heterozygosity (Ho) parameters were estimated using vcftools and plink 1.9 [32]. A genetic distance matrix-based identity by state (IBS) was generated and used for hierarchical clustering. An unrooted phylogenetic tree was constructed for visualizing how closely accessions were related within each country using the ape (analyses of phylogenetics and evolution) library package (Paradis et al., 2004) and phangon, an R package [33]. The dissimilarity matrix was then used to construct the network relationships among the cacao accessions from Sierra Leone and Togo using QGRAPH [34] implemented in R. In addition, and as a complementary analysis, Discriminant Analysis of Principal Component (DAPC) was carried out using 'genind object' and the find.clusters function in the adegenet package. In order to properly assign the accessions to groups, the Bayesian information criterion (BIC) was used to determine the optimum number of clusters to be retained. A binary file was generated from the filtered VCF file and was then subjected to cross-validation approaches for population structure analysis. A cut-off value of 50% ancestry suggested through the Admixture analysis was used to estimate membership probabilities of the accessions for the groups identified [35]. The genetic differentiation among and within groups (individual country level and combined) was estimated using the analysis of molecular variance (AMOVA), and the significance was tested with a non-parametric approach with 999 permutations using GenAlex v. 6.503 [36]. The coefficient of genetic differentiation among the populations (Sierra Leone, Togo and combined) was calculated based on pairwise F_{ST} (fixation index) to estimate the genetic distance and relationships among the three populations used in the study.

Declarations

Acknowledgments

Authors acknowledge Swedish Research Council Development research grant (2017-05621) to the Swedish University of Agricultural Sciences and the International Institute of Tropical Agriculture for funding this research. Authors acknowledge the services and support of Petra van Roggen and team at Intertek for the KASP-SNP genotyping. Authors also acknowledge all the cacao farmers in Sierra Leone and Togo for their cooperation during leaf sample collection.

Authors' contributions

Conceptualization, R.B., P.L.K. and L.G.B. Methodology, R.B., P.A.A., M.M.L. and K.A. Supervision, R.B., P.L.A. and L.G.B. Writing original draft, R.B. Manuscript review and editing, M.M.L., K.A., P.A.A., P.L.K. and L.G.B.

Funding

This study was supported through a research grant from Swedish Research Council Development to L.G.B., R.B. and P.L.K. Additional financial support was received from Partnerskap Alnarp, Sweden for the KASP-SNP genotyping.

Availability of data and materials

Data can be obtained upon request from the corresponding author.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare the absence of any potential conflict of interest while conducting the research

References

1. International Trade Centre, “Cacao: A Guide to Trade Practices”, International Trade Centre, 2001, UNCTAD/WTO, Geneva, Cocoa: A guide to trade practices (intracen.org)
2. Bartley BGD. “The genetic diversity of cacao and its utilization”, 2005, CABI Publishing, Wallingford, UK.
3. ICCO, “ICCO Quarterly bulletin of cocoa statistics, vol XLVII, Cocoa year 2020/2021”, <https://www.icco.org/icco-documentation/quarterly-bulletin-of-cocoa-statistics/>
4. ICCO, “ICCO quarterly bulletin of cocoa statistics, vol XLVIII, Cocoa year 2021/2022”, <https://www.icco.org/icco-documentation/quarterly-bulletin-of-cocoa-statistics/>
5. FAOSTAT, “Food and Agriculture Organization of the United Nations Database of Agricultural Production”. FAO Statistical Databases, 2022, Accessed on 25 March 2023. <http://www.fao.org/faostat/>.
6. EU, “Ex-post evaluation of agriculture for development (A4D)”. 2019, Online, <https://www.eeas.europa.eu/sites/default/files/a4devalfinalreport.pdf>, Accessed on 14th September 2022.
7. Howes FN, “The early introduction of Cacao to West Africa”. African Affairs. vol. 45, no. 180, pp. 152, 1969.
8. N’Goran JAK, Laurent V, Risterucci AM, and Lanaud C, “The genetic structure of cacao populations (*Theobroma cacao* L.) revealed by RFLP analysis”, Euphyt., vol. 115, pp. 83–90, 2000.

9. Aikpokpodion PO, Kolesnikova-Allen M, Adetimirin VO, Gultinan MJ, Eskes A, Motamayor JC, and Schnell RJ, "Population structure and molecular characterization of Nigerian field genebank collections of cacao, *Theobroma cacao* L.", *Silv. Genet.*, vol. 59, no., pp. 273–285, 2010.
10. Padi FK, Ofori A, Takrama J, Djan E, Opoku SY, Dadzie AM, Bhattacharjee R, Motamayor JC, and Zhang D, "The impact of SNP fingerprinting and parentage analysis on the effectiveness of variety recommendations in cacao". *Tree Genet. Genomes*, vol. 11, pp. 1-14, 2015.
11. Olasupo FO, Adewale DB, Aikpokpodion PO, Muyiwa AA, Bhattacharjee R, Gutierrez OA, Motamayor JC, Schnell RJ, Ebai S, and Zhang D, "Genetic identity and diversity of Nigerian cacao genebank collections verified by single nucleotide polymorphisms (SNPs): a guide to field genebank management and utilization", *Tree Genet. Genomes*, vol. 14, no. 2, pp. 32, 2018, DOI: 10.1007/s11295-018-1244-2
12. Adenuga OO and Ariyo OJ, "Diversity analysis of cacao (*Theobroma cacao*) genotypes in Nigeria based on juvenile phenotypic plant traits", *Internat. J. Fruit Sci*, vol. 20, pp. S1348-S1359, 2020, DOI: 10.1080/15538362.2020.1786487
13. Bekele F and Phillips-Mora W, "Cocoa Breeding", In: *Advances in Plant Breeding: Industrial and Food Crops*. Al-Khayri J, Jain S, Johnson D (eds.), Springer-Verlag, New York, pp. 409-487, 2019.
14. Van Inghelandt D, Melchinger AE, Lebreton C, and Stich B, "Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers", *Theoret. Appl. Genet.*, vol. 120, pp. 1289-1299, 2010.
15. Liu Z, Li H, Wen Z, Fan X, Li Y, Guan R, Guo Y, Wang S, Wang D, and Qiu L, "Comparison of genetic diversity between Chinese and American soybean (*Glycine max* (L.)) accessions revealed by high-density SNPs", *Front. Pl. Sci.*, vol. 8, pp. 2014, 2017, doi: 10.3389/fpls.2017.02014
16. Motamayor J, Schnell R, and Kuhn D, "Applying SNP marker technology in the cacao breeding programme in Ghana", *Afr. Crop Sci. J.*, vol. 20, pp. 67–75, 2012.
17. Livingstone D, Royaert S, Stack C, Mockaitis K, May G, Farmer A, Saski C, Schnell R, Kuhn D, and Motamayor JC, "Making a chocolate chip: development and evaluation of a 6K SNP array for *Theobroma cacao*", *DNA Res.*, vol. 22, no. 4, pp. 279-291, 2015.
18. Lukman ZD, Susilo A, Dinarti D, Bailey BA, Mischke BS, and Meinhardt LW, "Genetic identity, ancestry and parentage in farmer selections of cacao from Aceh, Indonesia revealed by single nucleotide polymorphism (SNP) markers", *Trop. Pl. Biol.*, vol. 7, pp. 133–143, 2014.
19. Motilal LA, Sankar A, Gopaulchan D, and Umaharan P, "Cocoa". In: *Biotechnology of plantation crops*, Chowdappa P, Karun A, Rajesh MK, Ramesh SV (eds.), Daya Publishing House, New Delhi, pp. 313–354, 2017.
20. Mahabir A, Motilal LA, Gopaulchan D, Ramkissoo S, Sankar A, and Umaharan P, "Development of a core SNP panel for cacao (*Theobroma cacao* L.) identity analysis", *Genome*, vol. 63, pp. 103–114, 2020.
21. Semagn FK, Babu R, Hearne S, and Olsen M, "Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop

- improvement”, Mol. Breed., vol. 33, pp. 1–14, 2014.
22. Moinina A, Lahlali R, and Boulf M, “Management practices to improve the cocoa bean value chain in Sierra Leone: Cocoa bean production in Sierra Leone”, Moroccan J. Agric. Sci., vol. 4, no. 2, pp. 45-52, 2023.
 23. Takrama J, Kun J, Meinhardt L, Mischke S, Opuku S, Padi FK, and Zhang D, “Verification of genetic identity of introduced cacao germplasm in Ghana using single nucleotide polymorphism (SNP) markers”, Afr. J. Biotech., vol. 13, pp. 2127–2136, 2014.
 24. Darkwa K, Olanmi B, Asiedu R, and Asfaw A, “Review of empirical and emerging breeding methods and tools for yam (*Dioscorea* spp.) improvement: Status and prospects”, Pl. Breed., vol. 139, no. 3, pp. 474-497, 2019.
 25. Rabbi IY, Kulakow PA, Manu-Aduening JA, Dankyi AA, Asibou JY, Parkes EY, Abdoulaye T, Girma G, Gedil MA, Ramu, P, Reyes B, and Maredia M, “Tracking crop varieties using genotyping-by-sequencing markers: A case study using cassava (*Manihot esculenta* Crantz)”, BMC Genet., vol. 16, pp. 115, 2015, <https://doi.org/10.1186/s12863-015-0273-1>
 26. Bhattacharjee R, Agre P, Bauchet G, De Koeber D, Lopez-Montes A, Lava Kumar P, Abberton M, Adebola P, Asfaw A, and Asiedu R, “Genotyping-by-Sequencing to unlock genetic diversity and population structure in white yam (*Dioscorea rotundata* Poir.)”, Agron., vol. 10, no. 9, pp. 1437, 2020, <https://doi.org/10.3390/agronomy10091437>
 27. Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci A, da Silva C, Cascardo J, Allegre M, Kuhn D, Verica J, Courtois B, Loo G, Babin R, Sounigo O, Ducamp M, Guiltinan MJ, Ruiz M, Alemanno L, Machado R, Phillips W, Schnell R, Gilmour M, Rosenquist E, Butler D, Maximova S, and Lanaud C, “Towards the understanding of the cocoa transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions”, BMC Genomics, vol. 9, pp. 512, 2008, doi: 10.1186/1471-2164-9-512
 28. Allegre M, Argout X, Boccara M, Fouet O, Roguet Y, Berard A, Thevenin JM, Chauveau A, Rivallan R, Clement D, Courtois B, Gramacho K, Boland-Auge A, Tahi M, Umaharan P, Brunel D, and Lanaud C, “Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of *Theobroma cacao* L.”, DNA Res., vol. 19, no. 1, pp. 23–35, 2012.
 29. Fang W, Meinhardt LW, Mischke BS, Bellato C, Motilal L, and Zhang D, “Accurate determination of genetic identity for a single cacao bean, using molecular markers with a nanofluidic system, ensures cocoa authenticity and traceability”, J. Agric. Food Chem., vol. 62, no. 2, pp. 481-487, 2014.
 30. Cosme-Reyes SM, Cuevas HE, Zhang D, Oleksyk TK, and Irish BM, “Genetic diversity of naturalized cacao (*Theobroma cacao* L.) in Puerto Rico”, Tree Genet. Genomes, vol. 12, pp. 88, 2016, <https://doi.org/10.1007/s11295-016-1045-4>
 31. LGC Genomics, “KASPTm genotyping chemistry user guide and manual”, 2013, <http://www.lgcgenomics.com/genotyping/kasp-genotyping>

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, and Daly M, "PLINK: A tool set for whole-genome association and population-based linkage analyses", *Amer. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, 2007.
33. Schliep K, Potts AJ, Morrison DA, and Grimm GW, "Intertwining phylogenetic trees and networks", *Meth. Ecol. Evol.*, vol. 8, no. 10, 2017, <https://doi.org/10.1111/2041-210X.12760>
34. Epskamp S, Cramer OJC, Waldorp LJ, Schmittmann VD, and Borsboom D, "qgraph: Network visualizations of relationships in psychometric data", *J. Stat. Software*, vol., 48, pp. 1–18, 2012.
35. Paradis E, Claude J, and Strimmer K, "APE: Analyses of phylogenetics and evolution in R language", *Bioinfor. App. Note*, vol. 20, no. 2, pp. 289-290, 2004.
36. Peakall R, and Smouse PE, "GenALEX 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update:", *Bioinfor.*, vol. 28, pp. 2537–2539, 2012.

Figures

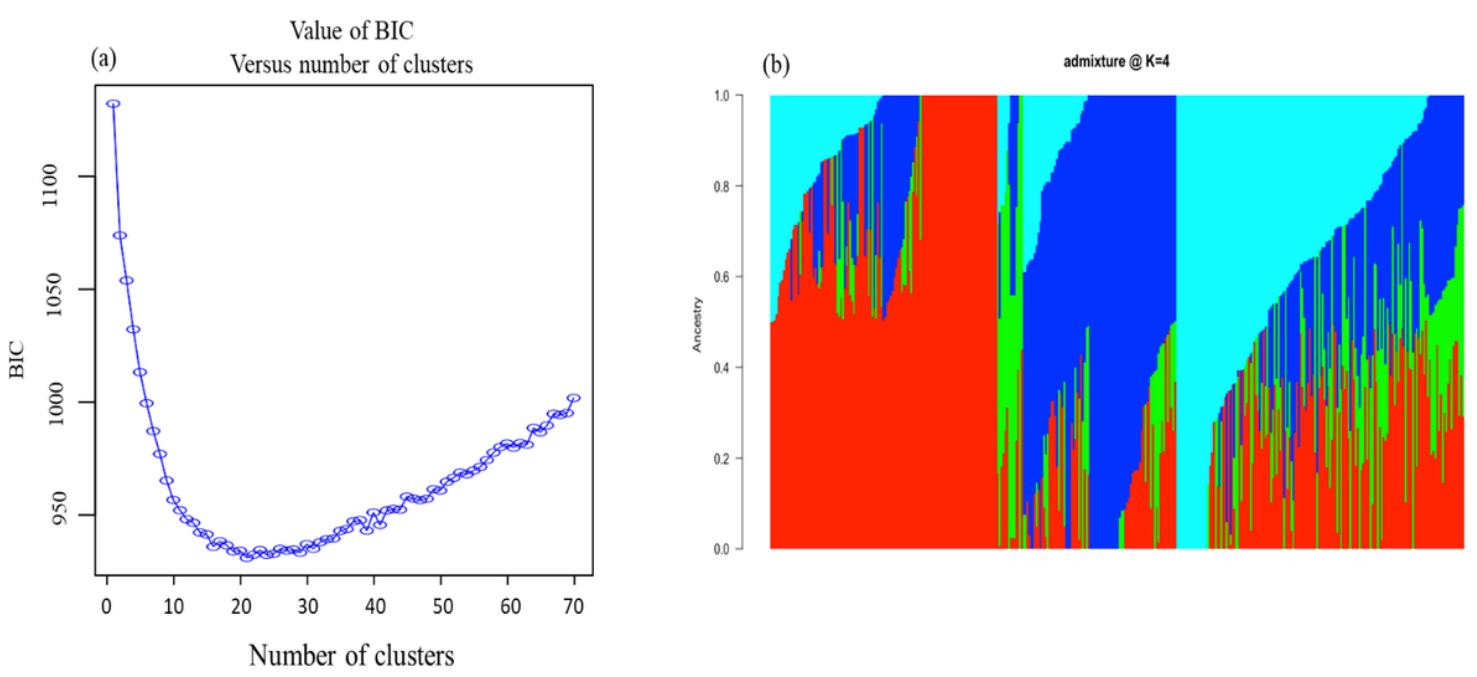


Figure 1

Graphical representation of the population structure of 376 cacao accessions. (a) Plot of mean likelihood of delta K against the number of K groups. (b) Subpopulations at K = 4. The colours represent four subpopulations of 376 accessions: Red: Subpopulation 1, Blue: Subpopulation 2, Light Blue: Subpopulation 3, and Green: Subpopulation 4

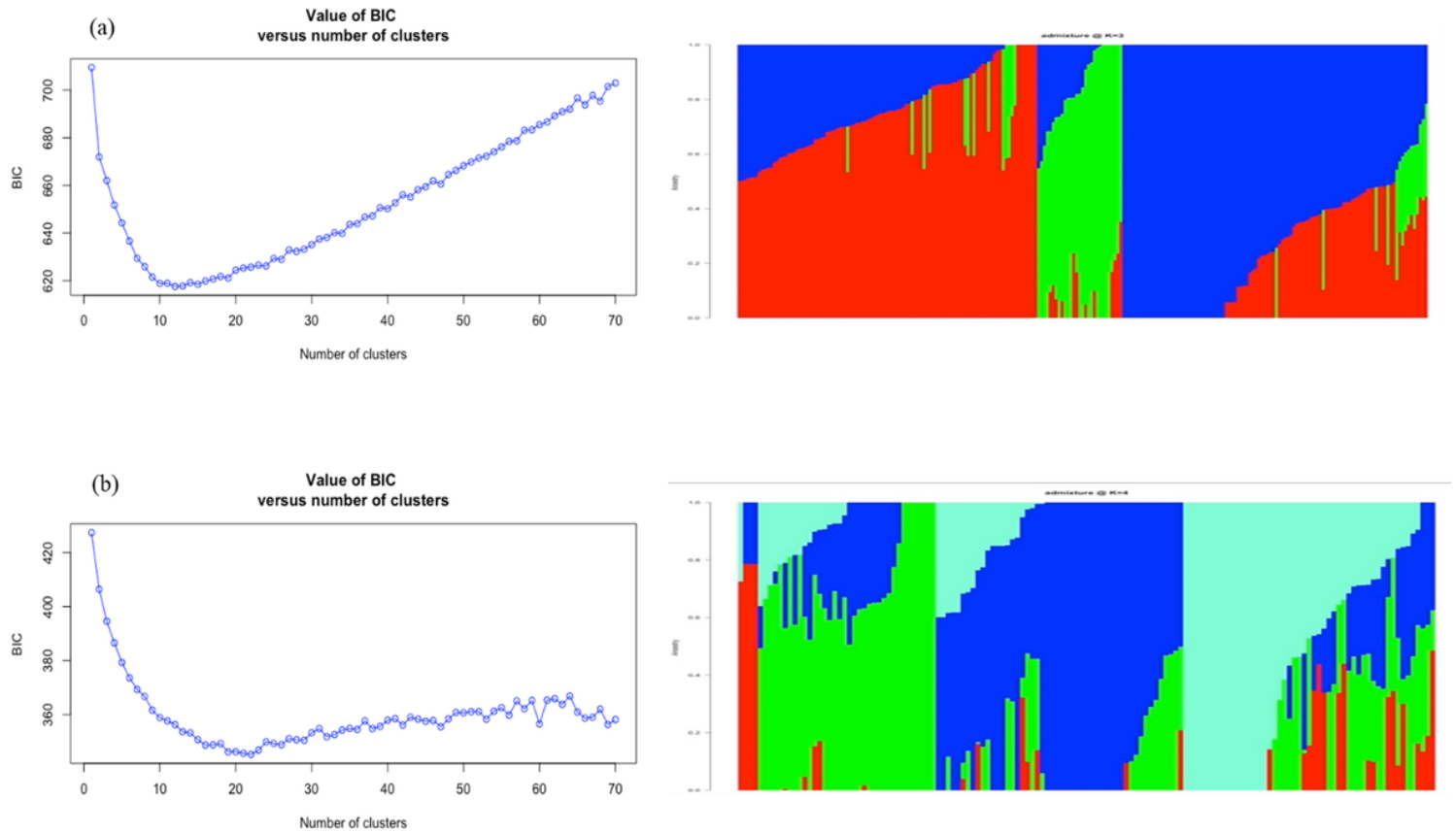


Figure 2

Population structure(a) Sierra Leone: Plot of mean likelihood of delta K against the number of K groups; Subpopulations at K = 3. The colours represent three subpopulations of 235 cacao accessions: Red – Subpopulation 1, Green – Subpopulation 2, Blue – Subpopulation 3; (b) Togo: Plot of mean likelihood of delta K against the number of K groups; Subpopulations at K = 4. The colours represent four subpopulations of 141 cacao accessions: Red – Subpopulation 1, Green – Subpopulation 2, Blue – Subpopulation 3, Turquoise – Subpopulation 4

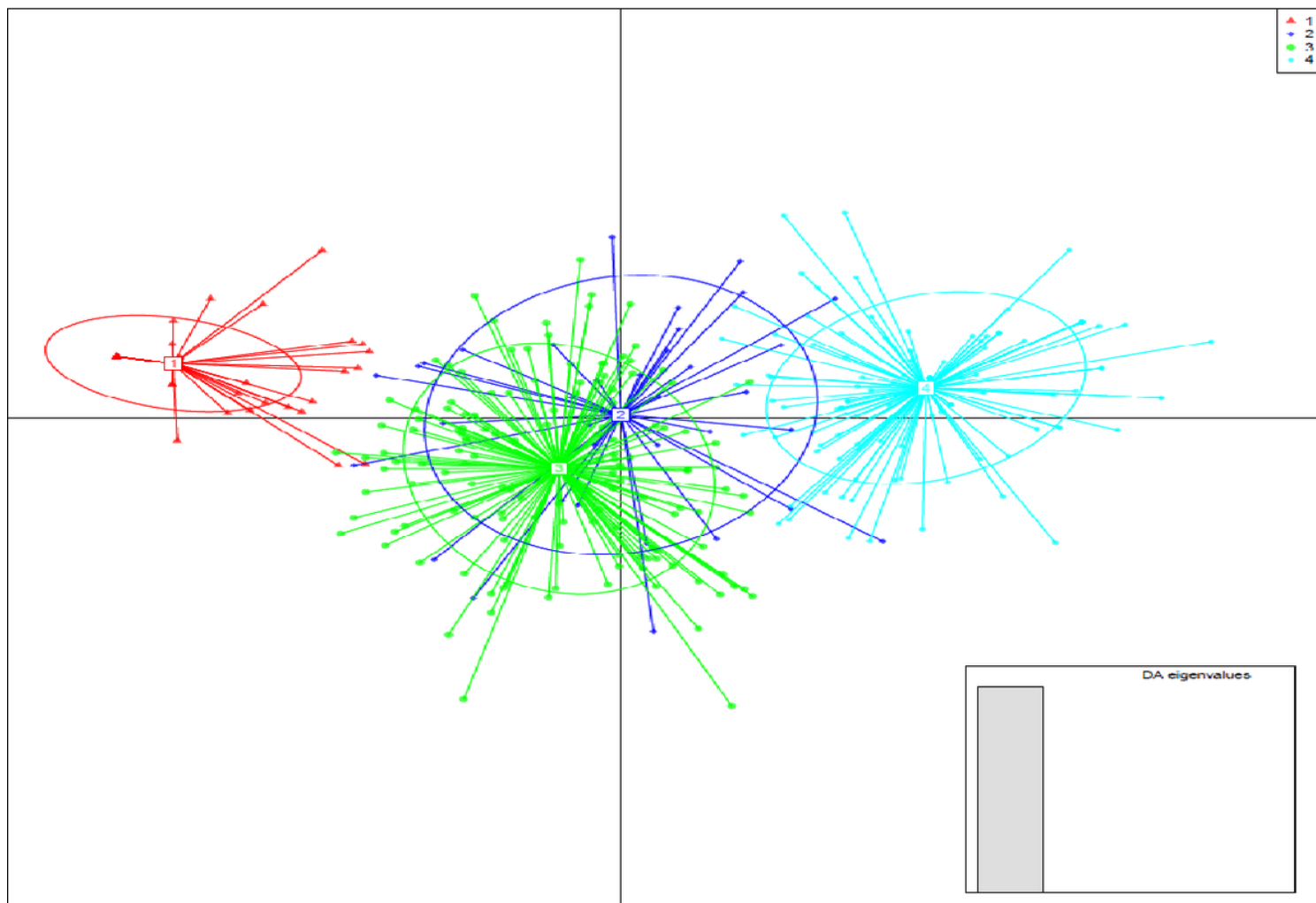
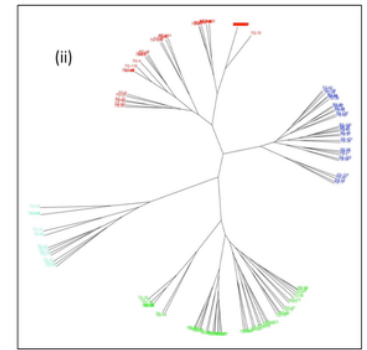
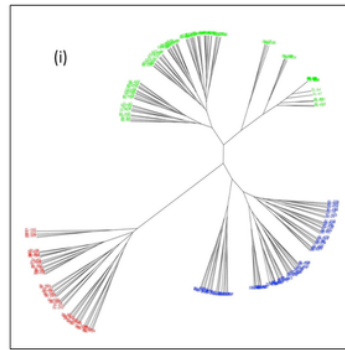
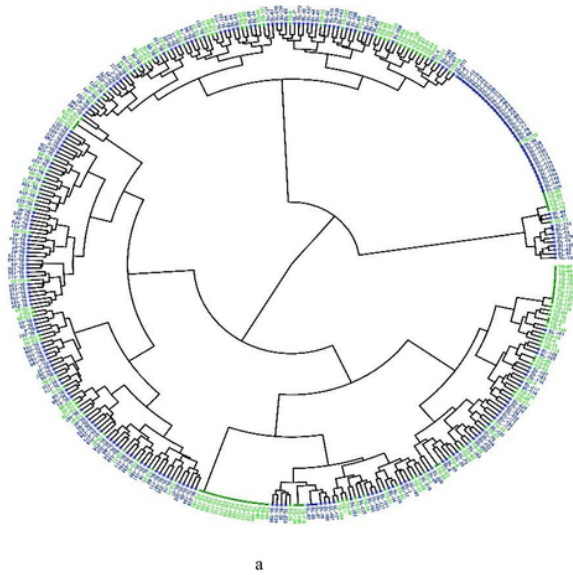


Figure 3

Discriminant analysis of principal components (DAPC) using 20 KASP-SNP markers. The axes represent the first two linear discriminants (LD). Each colour represents a cluster, while each dot represents an individual. Numbers represent different subpopulations identified by DAPC analysis



b

Figure 4

a. Phylogenetic tree for 376 cacao accessions from Sierra Leone and Togo. The blue and green color represent accessions from Sierra Leone and Togo, respectively

b. Phylogenetic tree depicting genetic relationships among cacao accessions from (i) Sierra Leone and (ii) Togo. Colors of each tree represent genetic groups within each country

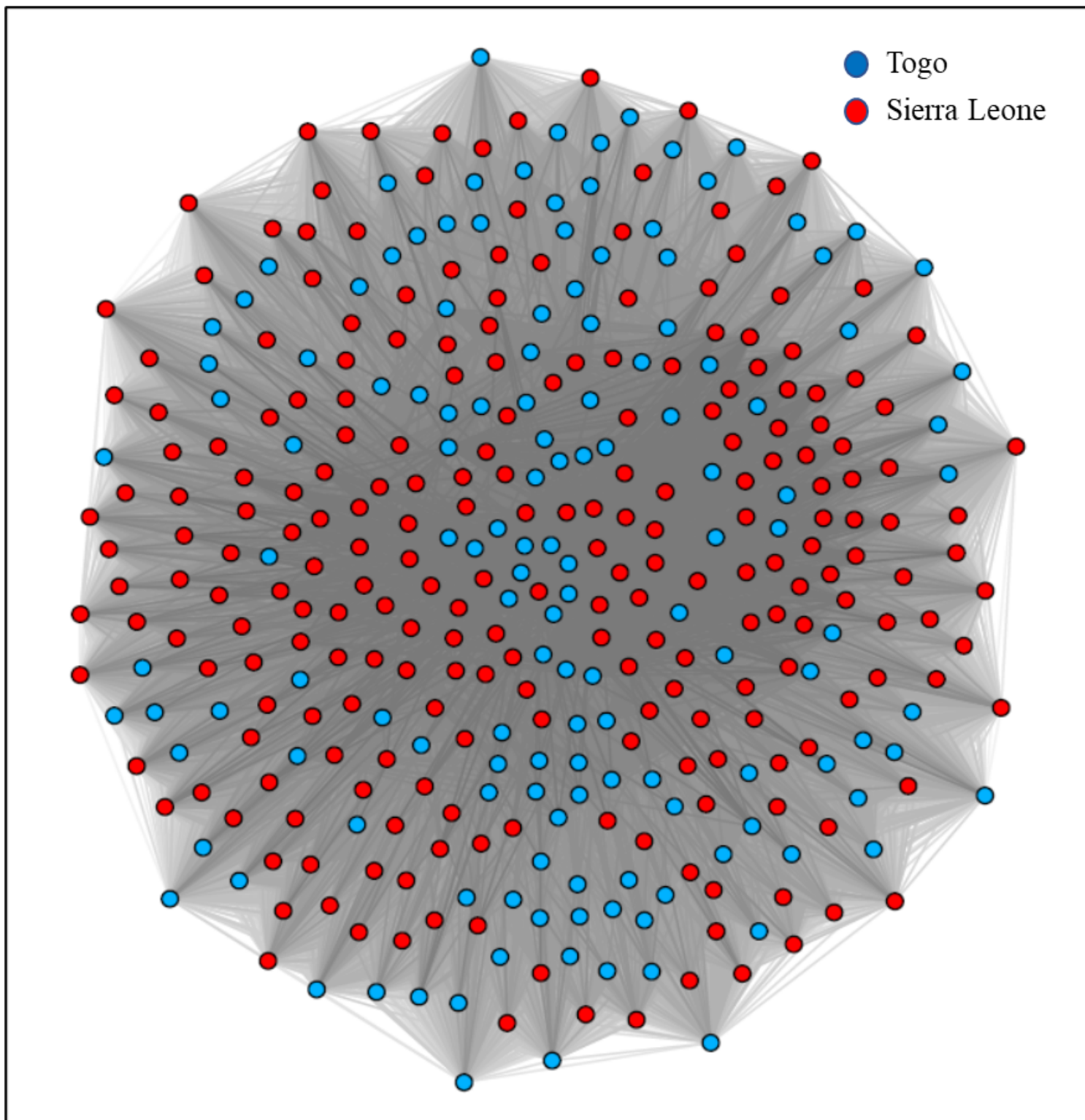
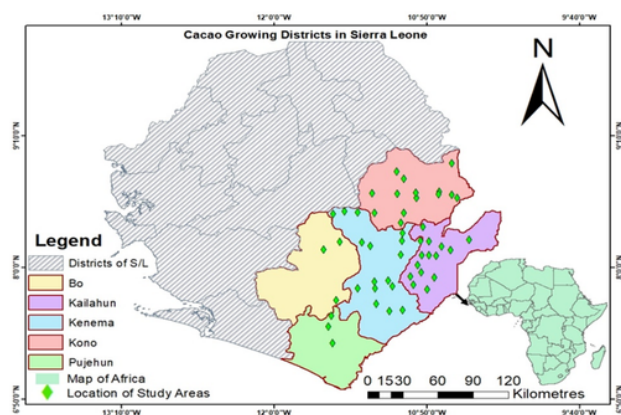
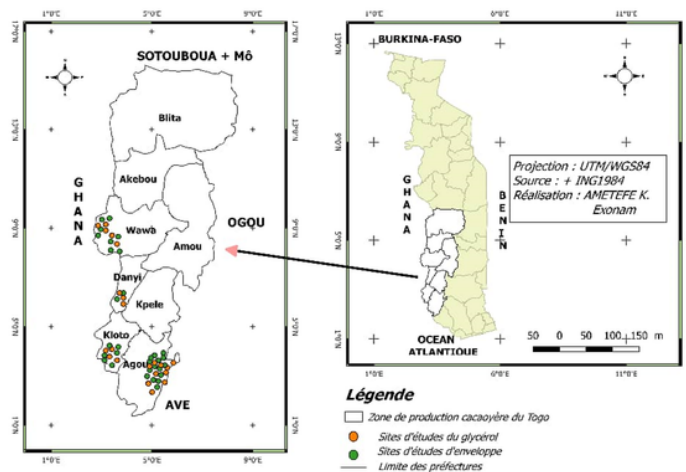


Figure 5

The genetic networks obtained using QGRAPH. The figure represents networks between cacao accessions from Sierra Leone and Togo, with the node size depicting genetic relationships among different accessions based on the observed heterozygosity and allelic richness



a



b

Figure 6

a. Map of Sierra Leone showing the districts and the locations from where samples were collected

b. Map of Togo showing the districts from where samples were collected

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigures.docx](#)
- [SupplementaryTable1.xlsx](#)
- [SupplementaryTable2aand2b.xlsx](#)
- [SupplementaryTableS3.xlsx](#)