


RESEARCH

Open Access



Genetic diversity and population structure of superior shea trees (*Vitellaria paradoxa* subsp. *paradoxa*) using SNP markers for the establishment of a core collection in Côte d'Ivoire

Affi Jean Paul Attikora^{1*} , Saraka Didier Martial Yao^{2,3}, Dougba Noel Dago², Souleymane Silué², Caroline De Clerck⁴, Yves Kwibuka⁵, Nafan Diarrassouba^{2,3}, Taofic Alabi^{2,6}, Enoch G. Achigan-Dako⁷ and Ludivine Lassois¹

Abstract

Background The shea tree is a well-known carbon sink in Africa that requires a sustainable conservation of its gene pool. However, the genetic structure of its population is not well studied, especially in Côte d'Ivoire. In this study, 333 superior shea tree genotypes conserved in situ in Côte d'Ivoire were collected and genotyped with the aim of investigating its genetic diversity and population structure to facilitate suitable conservation and support future breeding efforts to adapt to climate change effects.

Results A total of 7,559 filtered high-quality single nucleotide polymorphisms (SNPs) were identified using the genotyping by sequencing technology. The gene diversity (HE) ranged between 0.1 to 0.5 with an average of 0.26, while the polymorphism information content (PIC) value ranged between 0.1 to 0.5 with an average of 0.24, indicating a moderate genetic diversity among the studied genotypes. The population structure model classified the 333 genotypes into three genetic groups (GP1, GP2, and GP3). GP1 contained shea trees that mainly originated from the Poro, Tchologo, and Hambol districts, while GP2 and GP3 contained shea trees collected from the Bagoué district. Analysis of molecular variance (AMOVA) identified 55% variance within populations and 45% variance within individuals, indicating a very low genetic differentiation (or very high gene exchange) between these three groups ($F_{ST}=0.004$, gene flow $Nm=59.02$). Morphologically, GP1 displayed spreading tree growth habit, oval nut shape, higher mean nut weight (10.62 g), wide leaf (limb width = 4.63 cm), and small trunk size (trunk circumference = 133.4 cm). Meanwhile, GP2 and GP3 showed similar morphological characteristics: erect and spreading tree growth habit, ovoid nut shape, lower mean nut weight (GP2: 8.89 g; GP3: 8.36 g), thin leaf (limb width = 4.45 cm), and large trunk size (GP2: 160.5 cm, GP3: 149.1 cm). A core set of 100 superior shea trees, representing 30% of the original population size and including individuals from all four study districts, was proposed using the "maximum length sub-tree function" in DARwin v. 6.0.21.

*Correspondence:

Affi Jean Paul Attikora

affi jean0121@gmail.com; ajpattikora@uliege.be

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Conclusion These findings provide new knowledge of the genetic diversity and population structure of Ivorian shea tree genetic resources for the design of effective collection and conservation strategies for the efficient use of inbreeding.

Keywords *Vitellaria paradoxa*, Single nucleotide polymorphisms (SNPs), Population structure, Genetic diversity, DArTseq, Analysis of molecular variance (AMOVA), Core collection

Background

The shea tree (*Vitellaria paradoxa*, C. F. Gaertn) is an African native plant species. It has been recorded in 21 countries across semi-arid Sub-Saharan Africa within a wide belt of more than 3.4 million km [1]. It supports an estimated 16.2 million of shea nut collectors [1]. *V. paradoxa* is distributed from Senegal in the West to Uganda in the East [2]. The genus *Vitellaria* comprises a single species belonging to the Sapotaceae family, and it includes two subspecies: subsp. *nilotica*, primarily found in East Africa, and subsp. *paradoxa*, which occurs in West Africa from Senegal to the Central African Republic [3, 4].

In Côte d'Ivoire, shea trees are predominantly found in the northern region. This species exhibits monoecious characteristics and possesses a bisexual reproduction pattern that makes it rely primarily on insect-mediated cross-pollination. However, the plant has also hermaphroditic flowers capable of self-pollination [2]. The dissemination of the shea fruits is primarily achieved through barochory, with secondary dispersal by various animals such as birds, monkeys, rodents, and even humans [5].

Shea tree fruits have been considered as a significant source of economic benefit in a number of countries within the semi-arid savanna regions of West Africa. In fact, dating back to medieval times, their kernels have been used in the production of an important primary derivative: shea butter [6, 7]. Additionally, due to the presence of edible fatty acids that are used in the food, cosmetic, and pharmaceutical industries, shea butter holds a prominent position as a multimillion-dollar export commodity [8–10].

Despite its importance, the International Union for Conservation of Nature [11] has classified the shea tree as one of the plant species facing significant threats and experiencing vulnerability. It is confronted with extensive degradation, which is primarily driven by activities such as charcoal production and frequent uncontrolled bushfires. Furthermore, factors like population growth, which reduces the duration of fallow periods, and the systematic collection of fruit from beneath shea trees by local communities, pose significant obstacles to the natural regeneration of the species [3]. In addition, the overexploitation of shea trees and the expansion of new, more profitable agricultural crops such as cashew

plantations in northern Côte d'Ivoire are contributing to the decline in shea tree densities.

To safeguard the genetic resources of *Vitellaria paradoxa* in West Africa, national and regional strategies [12] for the identification and preservation of local shea tree varieties have been promoted [5]. Several additional initiatives aimed at identifying and conserving shea tree resources have also been launched to address these concerns [13, 14].

In Côte d'Ivoire, superior shea trees have been identified and rereferred in several districts based on a participatory survey. These trees constitute the in situ collection of shea trees in the country [13]. The genetic diversity and structure of this population have been not studied. In addition, a recent study on the morphological traits and sustainability of part of these trees demonstrated that the superior shea trees conserved on farmers' lands are threatened because of biotic and abiotic pressures [15]. Consequently, establishing a core collection based on the genetic diversity of superior shea trees will be helpful for suitable conservation and management [15].

Studying the genetic diversity and population structure is important for designing effective conservation and breeding programs [16], as well as for characterizing the natural selection history and genetic relationships of *V. paradoxa* [17].

Several authors have mentioned molecular markers such as random amplified polymorphic DNA (RAPDs) [18–20] and single sequence repeats (SSRs) [5, 20–27] in studies of shea tree species. Most of these studies have used these molecular markers to access the genetic diversity and population structure of shea tree species. Recently, single nucleotide polymorphism (SNP) markers were applied to study the genetic diversity and population structure of the Ugandan *nilotica* subspecies of the shea tree [28].

The approach of establishing core collections has emerged to increase the efficiency of the conservation and use of plant genetic resources while preserving the genetic diversity of the entire collection as much as possible [29, 30]. For plant species with recalcitrant seeds, the establishment of a core collection is the most suitable and cost-effective alternative method for in situ conservation of their genetic resources [31].

In the present study, the genotyping by sequency technology was used to genotype a panel of 333 superior *V. paradoxa* trees from the four key shea production districts in Côte d'Ivoire. The objectives were: (1) to characterize the genetic diversity and population structure of the shea tree using SNP data; (2) to characterize the genetic differentiation among and within Ivoirian shea populations; and (3) to establish a core collection for suitable conservation and management. This study is the first to use SNP markers to assess the genetic diversity and population structure in a *V. paradoxa* subspecies of Côte d'Ivoire. It lays a foundation for the effective conservation of shea trees and future genome wide association studies (GWAS) in shea tree breeding programs.

Methods

Plant material

From an initial collection of 1,200 superior shea trees previously identified by a shea breeding program from Côte d'Ivoire now known as the *Centre Africain de Recherches et d'Applications sur le Karité* (CRAK, the African Center for Shea Research and Application), 333 were randomly selected based on geographical distribution and population density. Superior shea trees were identified as described in a previous study [15]. A participatory survey with farmers allowed for the selection of superior trees

based on criteria such as high fruit yield, sweet taste of the fruit pulp, large fruit size, early flowering every year, and periodicity of fruit production [15]. These genotypes are being conserved in situ in four northern districts of Côte d'Ivoire: Bagoué, Hambol, Poro, and Tchologo (Fig. 1). To obtain samples, two mature leaves from the lower part of the tree were collected from each individual shea tree between May and July 2020 and 2021. The collected leaves were immediately dried using silica gel and then stored at 4 °C, awaiting DNA extraction.

The savannas of northern Côte d'Ivoire, where shea trees grow, are divided into two main zones: the Sudanese and sub-Sudanese savannas, based on climatic factors and differences in vegetation [15]. The Sudanese savanna (Bagoué, Poro, and Tchologo districts) corresponds to the main production zone of the shea tree with a monomodal rainfall pattern (1,200 mm/year) [32, 33]; the sub-Sudanese savanna (Hambol district) corresponds to a transitional production zone with a bimodal rainfall pattern (1,050 mm/year) [32, 34]. The average annual temperature is around 27 °C. The vegetation is Sudano-Guinean and consists of wooded savanna and grassy savanna with scattered gallery forests, particularly along waterways [35]. The pedology of this zone is characterized by three subclasses of ferralitic soils: soils on basic rocks, tropical ferruginous soils and hydromorphic soils

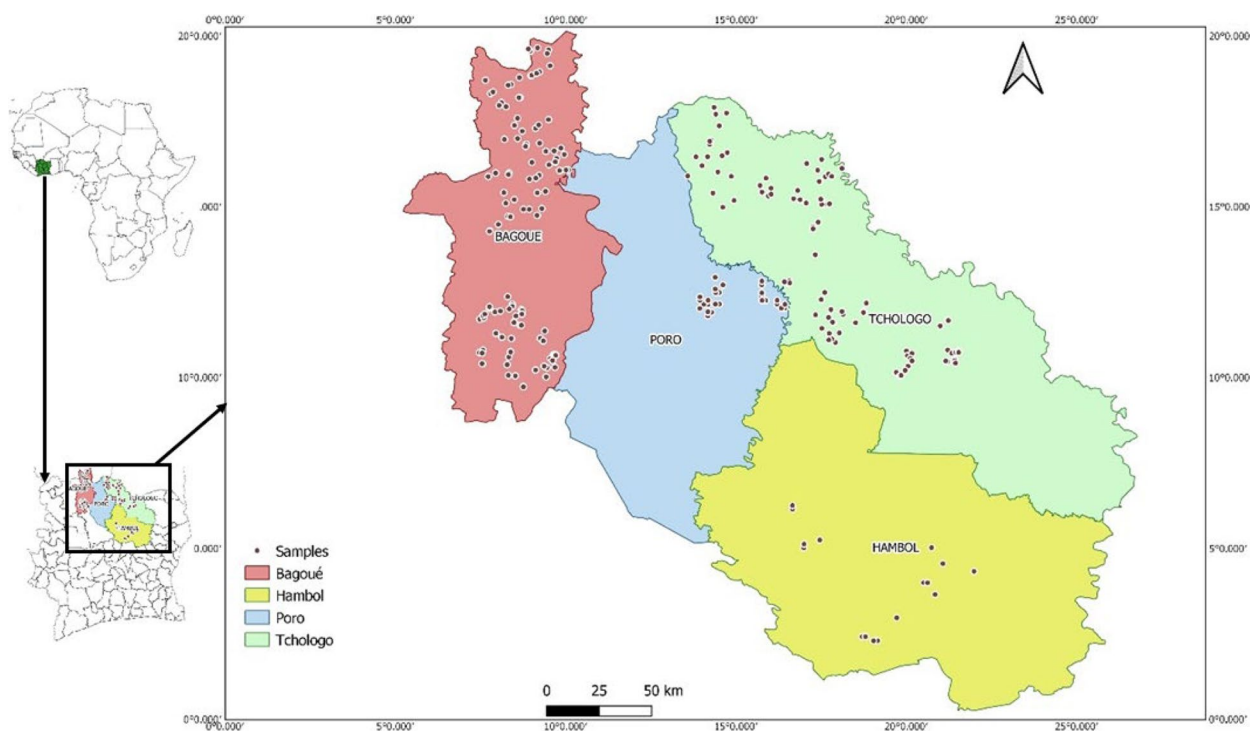


Fig. 1 Spatial distribution of the 333 superior shea trees (163 in Bagoué, 53 in Poro, 97 in Tchologo and 20 in Hambol) from northern Côte d'Ivoire; the colors represent the different districts (red for Bagoué, blue for Poro, green for Tchologo, and yellow for Hambol); the dots represent the samples

[32]. The main crops grown in these districts are cotton, cashew nuts and mangoes. The tree and shrub species commonly found in the study area are *Vitellaria paradoxa*, *Paria biglobosa*, *Pilliosigma thonningii*, *Parinari curratellifolia*, *Terminalia avicennioides* and *Ficus sciaphylla* [36].

DNA extraction and sequencing

The collected samples were sent to SEQART AFRICA, located at the International Livestock Research Institute (ILRI) in Nairobi, for genotyping. DNA extraction was conducted using the NucleoMag plant DNA extraction kit (Takara Bio USA), following the manufacturer recommendations. The concentration of extracted genomic DNA varied between 50 and 100 ng/ μ l. DNA integrity was checked on 0.8% agarose gel loaded. Libraries were constructed according to the DArTseq complexity reduction method through digestion of the genomic DNA using a combination of *Pst*I and *Mse*I restriction enzymes and ligation of barcoded adapters and the common adapter, followed by PCR amplification of adapter-ligated fragments [37]. Libraries were sequenced using single read sequencing runs for 77 bases. The sequencing was carried out using the Illumina HiSeq 2500 system.

SEQART AFRICA uses genotyping by sequencing DArTseq technology, which provides rapid, high-quality, and affordable genome profiling, even from the most complex polyploid genomes. DArTseq markers scoring was achieved using DArTsoft14, an in-house marker scoring pipeline based on algorithms. DArTseq SNP markers were scored as binary for the presence or absence (1 and 0, respectively) of the restriction fragment with the marker sequence in the genomic region of the corresponding sample.

SNP markers were aligned to the Vitpa_HiCPO_Assembly reference genome freely accessible online at <https://bioinformatics.psb.ugent.be/orcae/overview/Vitpa> to locate their corresponding chromosome positions.

SNP marker filtering

Two criteria were used to discard low-quality SNP markers and ensure data integrity. First, regarding the proportion of missing data (>20%), SNP markers with missing data exceeding 20% were excluded from the dataset. Second, regarding the minor allele frequency (MAF), SNPs with a minor allele frequency of less than 5% were considered rare and were therefore discarded.

The next step consisted of selecting the SNPs with substantial information content for further analysis. Specifically, a threshold based on the polymorphism information content (PIC) value equal to or greater than 0.1 was established. In addition, only biallelic SNP markers were kept for this study.

After this rigorous quality control process, a dataset consisting of 7,559 SNP markers and the 333 superior shea tree genotypes were considered for further analyses. This stringent filtering ensured that our dataset was of high quality and suitable for robust genetic analysis.

Genetic properties of markers

A custom Perl script was used to compute the allele counts and allele frequencies from the selected SNPs. Furthermore, to estimate markers-associated statistics such as observed heterozygosity (HO), expected heterozygosity (HE) and minor allele frequency (MAF), the GenAEx version 6.503 software was used [38]. The PIC values were computed using the formula proposed by Botstein et al. [39] using the Excel software [39]. The web-based SNIPlay software was used to determine transversion and transition mutations [40]. The Plink software “recordeA” function [41] was used for the generation of the SNP dosage format 0, 1, and 2, respectively representing the homozygote, the homozygote alternative and the heterozygote.

Population structure analysis

To assess the population structure of superior shea trees, three complementary methods were employed: Bayesian model-based clustering using Structure version 2.3.4 software [42], principal coordinates analysis (PCoA) was performed using GenAEx version 6.503 software, and discriminant analysis of principal components (DAPC) was performed using the R software version 4.3.0.

For the Bayesian model-based clustering analysis, the Markov chain Monte Carlo (MCMC) method with the admixture model excluding the LOCPRIOR option was used. This analysis was iteratively run 10 times for each K value ranging from 1 to 10. A burn-in period of 50,000 iterations followed by 100,000 MCMC iterations was used. Additionally, we assumed an admixture model with correlated allele frequencies. The most probable K value for each test was determined using the delta K (ΔK) [43] method based on the rate of change in $[\text{LnP}(D)]$ between successive K-values. Genotypes with membership probabilities greater than 0.7 were considered as belonging to the same group.

DAPC was used to complement the model-based population structure results obtained from Structure. DAPC is a multivariate method designed to identify and describe clusters of genetically related individuals [44]. It was performed in R software version 4.3.0 using the package “adegenet” with the function “find.clusters”. In the absence of a predefined grouping pattern, DAPC employs sequential K-means and model selection to establish genetic clusters based on genetic data. The bayesian information criterion (BIC) guided the

determination of the optimal number of genetic clusters (K) to best describe the data. The calculation of the α -score was instrumental in retaining the optimal number of principal components. DAPC also furnished membership probabilities for each individual with respect to each identified group, which is comparable to the admixture proportions obtained from Structure.

A neighbor-joining (NJ) phylogenetic tree was reconstructed using R software (version 4.3.0) based on Nei's genetic distance with 1,000 bootstrap replicates. The tree was customized using the online tree annotation platform iTOL (Interactive Tree of Life) [45].

The number of clusters determined by Bayesian model-based clustering was subsequently used in the analysis of molecular variance (AMOVA) to assess the genetic differentiation of the genotypes. This comprehensive AMOVA was performed using the GenAlEx software version 6.503, and it allowed for the estimation of the fixation index (F_{ST}) and the gene flow per haploid number of migrants (Nm). F_{ST} values range from 0 (no differentiation between groups) to 1 (complete differentiation). Moreover, genetic diversity indices such as the number of different alleles (N_a), the number of effective alleles (N_e), the number of loci with private alleles, Shannon's information index (I), observed heterozygosity (HO), and expected heterozygosity (HE) were also computed using the above-mentioned software [38].

Morphological characteristics of the groups obtained with bayesian model-based clustering from structure

To evaluate the morphological characteristics of 160 genotyped shea trees (availability of their morphological traits), 11 quantitative and qualitative morphological traits (Table S1), described in previous studies [15, 46] were used. A comparison of the morphological traits between the groups obtained using SNP markers with Structure was also accessed. Principal component analysis (PCA) and a heat map of the correlation matrix were performed to structure the quantitative traits. Finally, a Mantel test with 10,000 permutations and Spearman correlation method was performed for matrices comparison between morphological traits and SNP markers to access the relationship between both markers.

Design of the superior shea tree core collection

The design of the core collection followed the methodology proposed in a previous study [47]. DARwin software version 6.0.21 was used for the reconstruction of the diversity of trees using SNP dataset [48]. Dissimilarities were computed and transformed into Euclidean distances. The un-weighted neighbor-joining method was applied to the Euclidean distances to build a tree with all genotypes. The "maximum length sub tree" function was

then used to identify the individuals of the core collection: The "maximum length sub-tree" is a step-by-step process that successively eliminates redundant individuals. We then selected the last 100 individuals that retain the largest diversity. The size of the core collection was fixed a priori [49], and the efficiency of the strategy was assessed by comparing and keeping the total number of alleles captured for each run using the same software. The size of the core collection was expressed as the ratio of individuals kept in the core collection to the number of individuals in the entire collection. Principal component analysis (PCA) was plotted using R software (version 4.3.0) to see the distribution of the core sample relative to the entire sample.

Results

Distribution of SNPs, genetic diversity, and polymorphism information content in the *Vitellaria paradoxa* genome

After the filtering process, 7,559 SNPs, representing 17.7% of the 42,705 SNPs initially yielded, were retained. These 7,559 SNPs were unequally distributed across the 12 chromosomes with, an average marker density of 1 marker per 87.38 kb.

A genome-wide SNP marker analysis revealed that chromosome 2 had the highest number of SNPs, with 12.9% (978 SNPs) of the filtered SNPs, while chromosome 12 had the lowest number, with 6.2% (469 SNPs). In terms of marker density, chromosome 6 had the highest marker density with 1 marker per 74.67 kb, while chromosome 4 had the lowest density with 1 marker per 105.77 kb (Table 1).

Table 1 Genomic distributions of the 7,559 filtered SNPs physically mapped on the 12 chromosomes of *Vitellaria paradoxa* and the corresponding SNPs densities

Chromosomes	No. of SNPs	% SNPs	Length (Mpb)	Density (kb)
Chr01	791	10.46	80,731,948	102.06
Chr02	978	12.94	74,439,616	76.11
Chr03	662	8.76	57,704,473	86.17
Chr04	564	7.46	59,651,551	105.77
Chr05	654	8.65	59,580,608	91.10
Chr06	659	8.72	49,210,429	74.67
Chr07	606	8.02	55,380,075	91.39
Chr08	664	8.78	52,298,408	78.76
Chr09	488	6.46	47,443,901	97.22
Chr010	522	6.91	46,597,090	89.27
Chr011	502	6.64	38,413,107	76.52
Chr012	469	6.2	37,276,254	79.48
Mean	630	8.33	-	87.38
Total	7559	100	-	-

Within the collected shea genome, transition-type SNPs (4,647, or 61.48% of SNP markers) were more common than transversion-type SNPs (2,992, or 38.52% of SNP markers). This resulted in a ratio of transitions to conversion SNPs of 1.6 (4,647/2,912). Specifically, transition A/G and T/C types were more common than G/A and C/T types. For transversion, the types T/G, A/T, A/C, and G/C were more common than G/T, T/A, C/A, and C/G (Table 2).

The average observed heterozygosity (*HO*) in this study was 0.17, while the expected heterozygosity (*HE*) values varied from 0.1 (for 152 SNPs) to 0.5 (for 1,438 SNPs), with an average of 0.26. In parallel, the PIC values ranged from 0.1 (for 237 SNPs) to 0.5 (for 879 SNPs), with an

average of 0.24 (Fig S1.a and Fig. S1.b). A consistent number of SNPs (5,104 or 67.52% of the filtered SNPs) had a MAF value less than 0.2 (Fig. S1.c).

Population structure and genetic relationships

The population structure and the genetic relationships among the studied population were investigated using the Structure version 2.3.4 software. The *K* value was used to estimate the number of clusters in the shea tree population based on genotypic data across the whole genome. The optimal *K* value was determined by plotting the number of clusters (*K*) against ΔK . It showed a sharp peak at *K*=3 (Fig. 2a), suggesting that the studied population can be clustered into three groups with different

Table 2 Percentage of transition and transversion SNPs across the *Vitellaria paradoxa* genome

SNP type	Transition SNPs		Transversion SNPs			
	A/G	T/C	A/T	A/C	G/T	G/C
Number of SNPs	2316	2331	851	685	673	703
Frequency %	30.6%	30.8%	11.3%	9.1%	8.9%	9.3%
Total (Percent of total)	4647 (61.5%)		2912 (38.5%)			

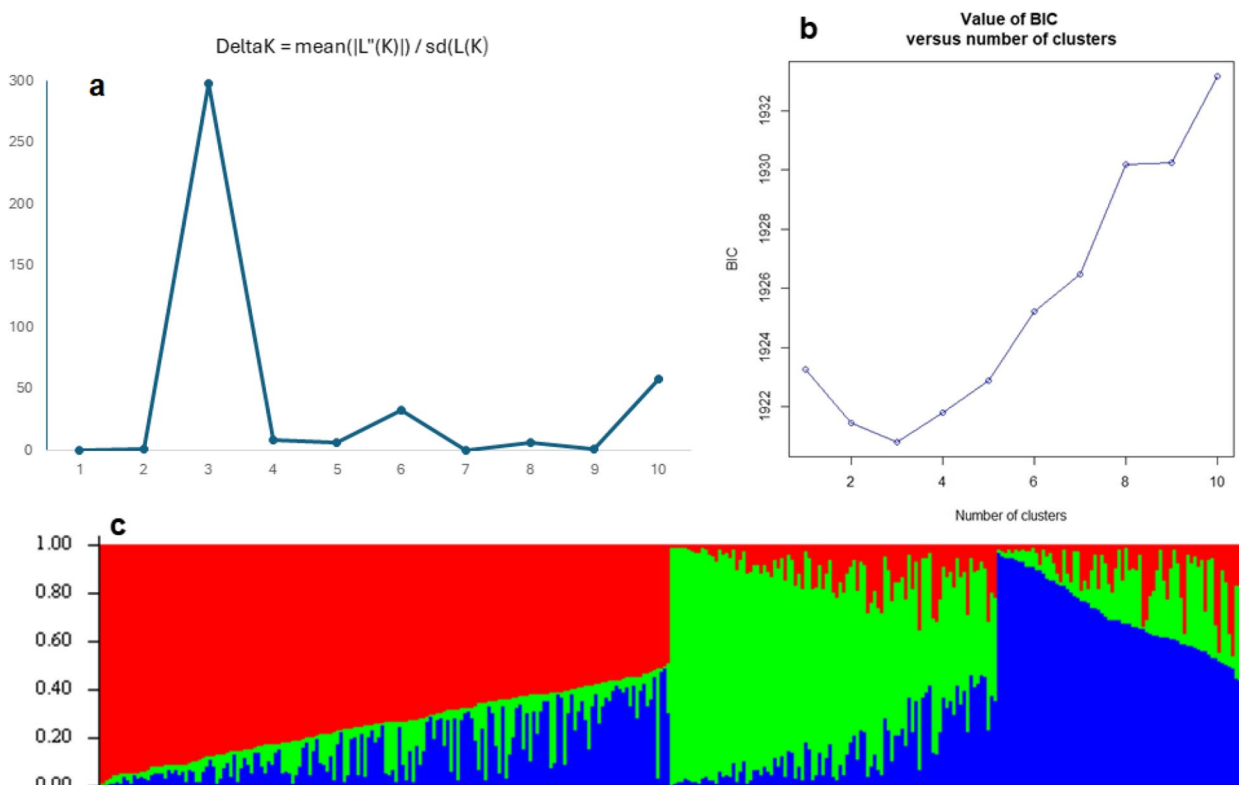


Fig. 2 **a** Delta K for various number of clusters (*K*); **b** Values of the BIC (1,920.82) versus the number of clusters; **c** Bayesian model-based analysis (*K*=3) of 333 *Vitellaria paradoxa* individuals; accessions in red are clustered into GP1, accessions in green are clustered in GP2, and accessions in blue are clustered into GP3

genetic backgrounds (GP1, GP2, GP3). The number of genotypes in each group was 168 in GP1, 93 in GP2, and 72 in GP3. Among the population, 154 genotypes were considered as admixed (Table S2). Consequently, these results were considered in the subsequent population genetics analyses.

The net nucleotide distance between the groups is shown in Table S2, and the maximum distance was recorded between GP1 and GP2 (0.003). The genetic distance between GP1 and GP3 (0.0026) was closely related to the distance between GP2 and GP3 (0.0025). The results from Structure estimated the fixation index (F_{ST}) for each group and suggested a significant divergence within the three groups (Table S2). GP3 displayed the highest F_{ST} value (0.0182), while GP1 had the lowest F_{ST} value (0.0177). In addition, the heterozygosity values were 0.259, 0.261 and 0.262 for GP1, GP2, and GP3, respectively (Table S2).

Consistent with the findings from Bayesian model-based clustering in Structure, the discriminant analysis of principal components also suggested three distinct clusters based on the value of BIC (1,920.82) (Fig. S2.a). In the DAPC results, group1, group2, and group3 have, respectively, 93, 72, and 168 individuals. The probability of each individual belonging to a single cluster was 100%. Therefore, no individual was considered mixed (Fig. S2.a and Fig. S3). The principal coordinates analysis (PCoA) showed that GP2 is an intermediate group between GP1 and GP3 (Fig. S4). In fact, the Structure software triangular plot showed that most of the shea trees from the Hambol, Poro, and Tchologo districts were clustered in GP1, whereas genotypes included in GP2 and GP3 are mainly from the Bagoué district (Fig. S2.b). The same results were observed with the PCoA plot (Fig. S5).

In addition to the three methods above, a neighbor-joining phylogenetic tree also clustered the shea trees into three groups (Fig. 3). Based on the true clades confirmed with the bootstrap values, the samples should be clustered by origin (Fig. 3). This confirms the true structure of our shea tree samples using Structure.

Genetic differentiation of populations

The three groups from the Bayesian model-based clustering in Structure were then used to calculate various genetic parameters such as AMOVA, Nei's genetic distance, and genetic diversity indices using GenAlEx 6.503 software. The results of these analyses are shown in Table 3. Overall, the results showed a low level of genetic differentiation between groups but a high level of genetic differentiation within groups. In addition, Nei's genetic distance analysis revealed a very low fixation index value (F_{ST} : 0.004) and a significant number of migrants (Nm : 59.02), confirming the low level of genetic differentiation

between the three genetic groups. The pairwise F_{ST} values were found to be 0.005, 0.004, and 0.003 for GP1-GP2, GP1-GP3, and GP2-GP3, respectively (Table 3).

Allelic pattern across populations

The three groups showed a grand mean of 2 for the number of different alleles (N_a) and 1.395 for the number of effective alleles (N_e) (Table 4). Within the whole population, the means for Shannon's index (I), gene diversity (HE), and unbiased gene diversity (uHE) were 0.414, 0.258, and 0.260, respectively. The three genetic groups were closely related in terms of diversity metrics (Table 4). The percentage of polymorphic loci per population (PPL) was 100% for GP1 and GP2, while GP3 had a PPL of 99.93%.

Comparison of our results with previous studies

Several studies assessed the genetic diversity in shea tree during the last two decades (Table S3). SSR markers were widely used to assess the genetic diversity of shea tree, representing 54.55% of the studies. SNP markers were the most recent markers used, representing 27.27% including our study. Only two studies mentioned RAPD markers in shea genetic diversity study (18.18%). Moderate genetic diversity was observed in our study and a study conducted in Uganda using SNP markers. However, the HE obtained in the present study (HE: 0.26) was higher than the HE obtained in Uganda (HE: 0.21). Using the same type of marker, a study conducted in Ghana showed low genetic diversity, with an HE value of 0.041. However, the population genetic differentiation parameter such as the fixation index (F_{ST}) was higher in the Ugandan population than in our study and in the population of Ghana (Table S3).

For other studies using SSR markers, low HE value (0.32) to moderate HE value (0.73) was observed. The HE values of SSR markers were higher than HE in our study because SSRs are co-dominant and the HE value varies from 0 to 1, whereas for SNP and RAPD markers, HE is calculated as dominant and the value ranges from 0 to 0.5. Overall, studies conducted in a single country or subspecies showed low genetic differentiation. In contrast, studies conducted in the natural range of *V. paradoxa* showed high genetic differentiation, and the authors attributed this differentiation to the subspecies *nilotica* and *paradoxa* (Table S3).

Design of the core germplasm collection

A core germplasm collection is a critical step in creating a manageable and representative sample that can reflect the diversity within the larger germplasm collection. This process is also relevant to modern plant breeding efforts. Maintaining genetic diversity within

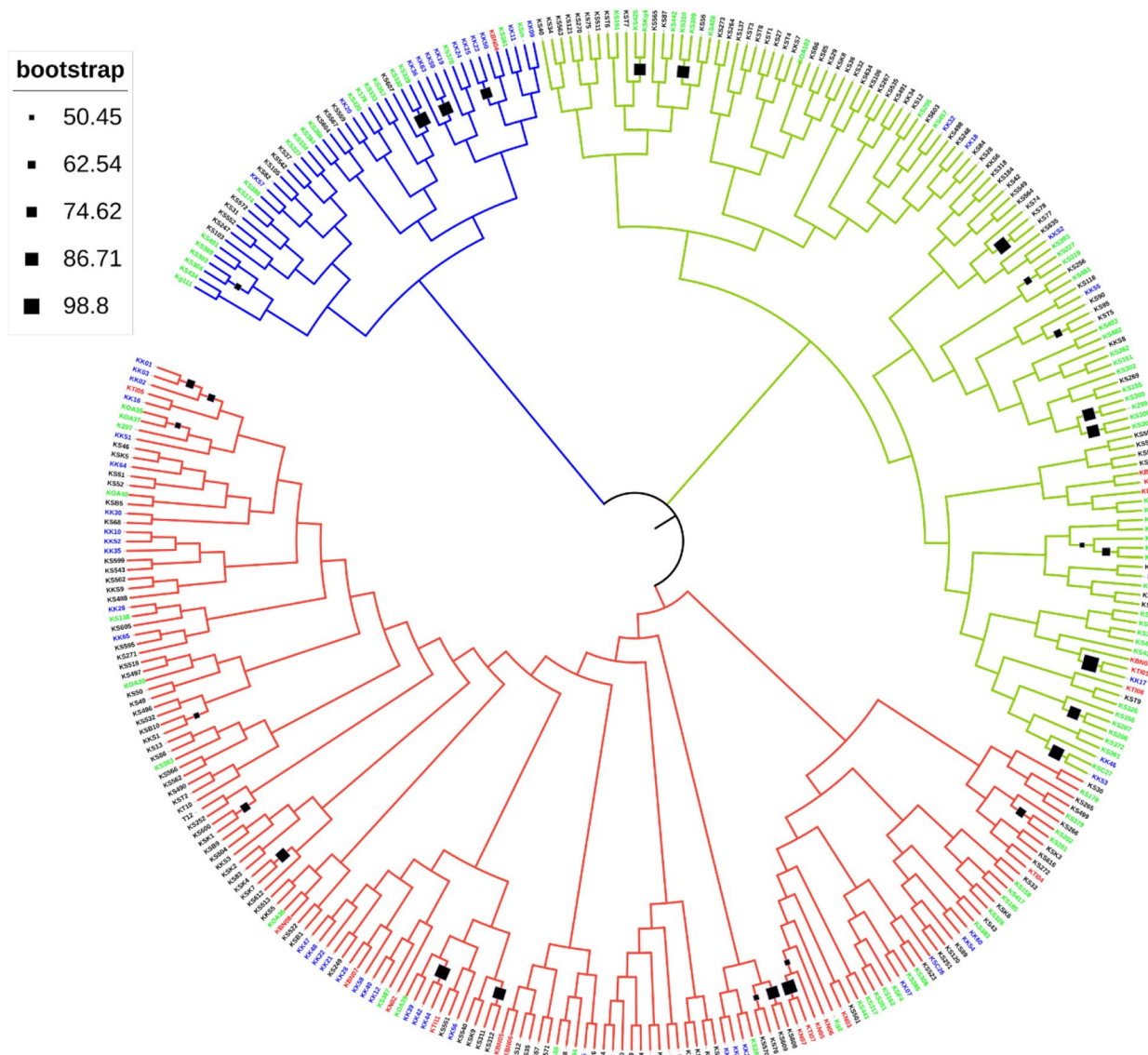


Fig. 3 Neighbor-joining phylogenetic tree with bootstrap values (black squares represent the bootstrap values; 1,000 replicates) showing relationships between the 333 superior shea trees based on Nei’s genetic distance matrix using 7,559 SNP markers; branch colors represent genetic groups (red for GP1, light green for GP2, and blue for GP3; label colors represent the origin of the genotypes (red for Hambol, black for Bagoué, green for Tchologo, and blue for Poro)

the core collection is essential, and the primary criterion for selecting its members is the average genetic distance in the population. Using the “maximum length sub-tree” function in DARwin 6.0.21, we successfully designed a core germplasm set of 100 individuals, representing 30% of the entire population (Fig. S6). Of the core germplasm, 49% of the shea trees are from the district of Bagoué. These individuals belong to GP2 (30 individuals) and GP3 (19 individuals). The remaining 51% of the core germplasm, are from the other three districts (Hambol, Poro, and Tchologo) and belong

mainly to GP1 (Table 5). The grand mean diversity metrics of this core collection set were similar to those of the entire population (Table S4).

A phylogenetic tree of the 100 individuals in the core set was reconstructed with the neighbor joining method based on Nei’s genetic distance. The obtained dendrogram is similar to that obtained with the whole population (Fig. S6).

A good spatial representation of the core set within the entire sample was observed in the PCA plot of the entire panel of 333 shea trees (Fig. 4).

Table 3 Analysis of molecular variance (AMOVA) among the 333 superior shea trees based on genetic variation among and within the identified groups

Source	Df	SS	MS	Estimated Variance	%
Among Pops	2	6607.628	3303.814	5.763	0%
Within Pops	330	696,593.908	2110.891	750.481	55%
Within Indiv	333	203,106.500	609.929	609.929	45%
Total	665	906,308.036		1366.173	100%
Fixation index (F_{ST})	0.004				
Nm (Haploid)	59.02				
Pairwise F_{ST} values	GP1-GP2	GP1-GP3	GP2-GP3		
	0.005	0.004	0.003		

Table 4 The means of different genetic parameters in each of the three groups

Pop	N	Na	Ne	I	HO	HE	uHE	F	PPL
GP1	154.418	2	1.395	0.415	0.173	0.259	0.259	0.327	100%
GP2	87.645	2	1.395	0.414	0.172	0.258	0.260	0.328	100%
GP3	65.345	1.999	1.396	0.413	0.173	0.258	0.260	0.320	99.93
Mean	102.469	2	1.395	0.414	0.173	0.258	0.260	0.325	99.98%

N number of samples, Na number of different alleles, Ne number of effective alleles, I Shannon’s index, HO observed heterozygosity, HE diversity index, uHE unbiased diversity index, PPL percentage of polymorphic loci

Table 5 Percentage of individuals from each district and genetic group in the core set

District	N population				N core set			
	GP1 (%)	GP2 (%)	GP3 (%)	Total	GP1 (%)	GP2 (%)	GP3 (%)	Total
Bagoué	-	95 (28.53)	68 (20.42)	163	-	30 (30)	19 (19)	49
Hambol	20(6.01)	-	-	20	8 (8)	-	-	8
Poro	51 (15.31)	-	2 (0.6)	53	28 (30)	-	2 (2)	30
Tchologo	96 (28.83)	-	1 (0.3)	97	12 (12)	-	1 (1)	13
Total	167 (50.15)	95 (28.53)	71 (21.32)	333	48 (48)	30 (30)	22 (22)	100

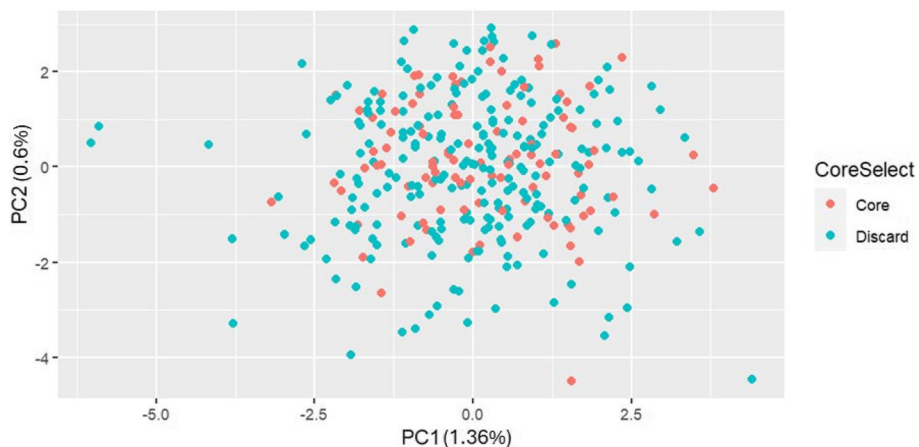


Fig. 4 Principal component analysis plot showing the distribution of the core sample within the full panel of 333 shea trees. Blue dot represents the discarded genotypes and red dot represents the core sample

Morphological characteristics of the genotyped shea trees

The analysis of the morphological characteristics of 160 genotyped shea trees revealed significant variations (Table S5). Principal component analysis (PCA) identified three principal components (PCs) that explained 73.03% of the total variance observed among shea trees population (Table S6). The PC1 (x-axis) explained 34.55% of the total variance. The leaf traits (PL, LL, and LW) and nut traits (NL, NWD and NWG) are correlated positively towards PC1 (Fig. 5). However, trunk circumference (TC) correlated negatively towards PC1. Similarly, PC2 (y-axis) accounted for 26.03% of the total variation. The nut traits (NL, NWD, and NWG) and TC positively correlated towards PC2 while the leaf traits (PL, LL, and LW) were negatively correlated (Fig. 5). PC3 captured 12.45% of the total variance, and TC and NWG were positively correlated. In contrast, NWD and NGW were negatively correlated towards PC3 (Table S6).

The correlation heatmap analysis revealed three classes of studied traits: (i) trunk circumference, (ii) nut size (nut length, nut width, and nut weight) and (iii) leaf size (limb width, limb length, and petiole length).

Positive correlations were observed within the descriptors of each identified class (see color key in Fig. S7).

Using clustering from Structure as a priori groups, significant differences were found between the three groups in nut weight (F: 14.11; $p < 0.001$), trunk circumference (F: 5.04; $p < 0.01$), and limb width (F: 4.16; $p < 0.05$). GP1 showed a weak mean of trunk circumference (133.4 cm), large leaf size (limb width = 4.63 cm), and a high mean of nut weight (10.62 g), whereas GP2 showed a high mean of trunk circumference (160.5 cm), a thin leaf (limb width = 4.46 cm), and a weak mean of nut weight (8.89 g). GP3 had the lowest value for significant traits such as nut weight (8.36 g), and limb width (4.44 cm) and, it was intermediate for trunk circumference (149.1 cm). A detailed examination of the results shows that GP1 differs from GP2 and GP3, while no significant differences are observed between GP2 and GP3 (Table S7). There were no statistical differences between the groups for the other quantitative traits. The means of these quantitative traits are presented in Table S7.

Meanwhile, all the qualitative traits used to characterize the three groups showed statistically significant differences (Table S7).

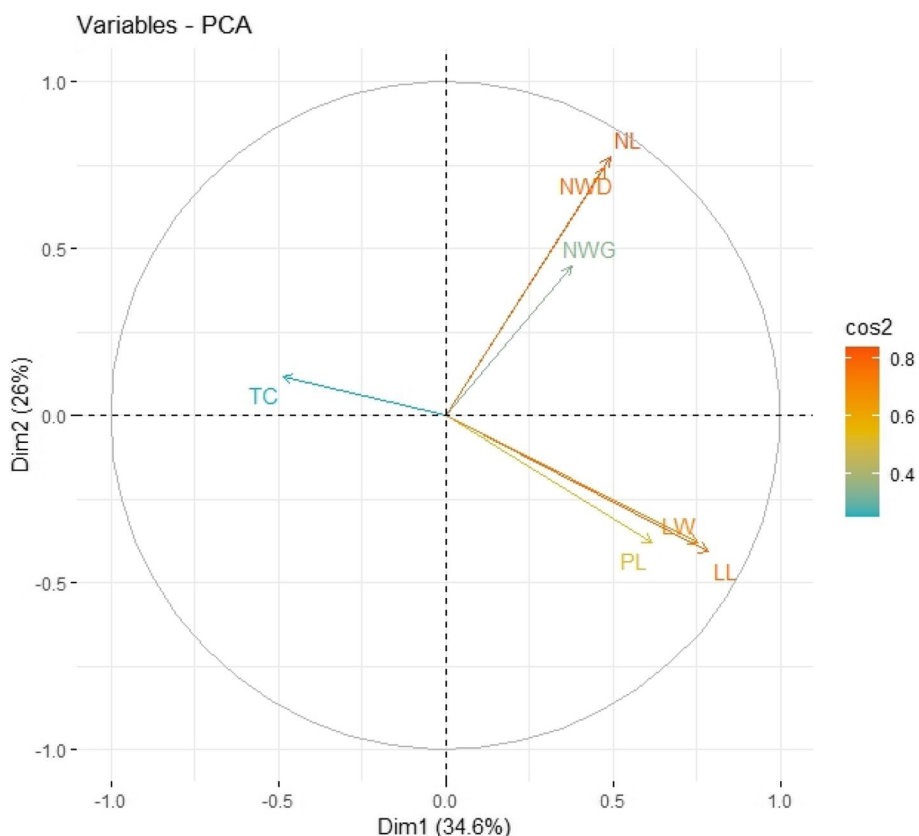


Fig. 5 Estimated PC1 and PC2 for quantitative morphological traits of 160 genotyped shea trees. TC: trunk circumference, PL: petiole length, LL: limb length, LW: limb width, NL: nut length, NWD: nut width, NWG: nut weight

While the seed coat colors of “creamy” and “dull brown” were present only in GP1, the spheroid seed shape was found only in GP2 and GP3 (Table S7). In terms of group differentiation, the same trend observed for quantitative traits was also observed for qualitative traits. Therefore, the three groups obtained with SNP data can only be clustered into two groups based on morphological characteristics: (I): GP1 and (II): GP2 and GP3 as a single group.

The Mantel test results revealed that there is no relationship between morphological traits and SNP markers (Mantel $r = 0.0176$, $p = 0.3324$).

Morphological characteristics of the core collection

Of the 160 genotyped shea trees, 52 were captured in the established core collection. The analysis of the morphological characteristics of these core individuals showed important variations for the quantitative traits (Table S8). For example, trunk circumference varied from 65.5 to

242 cm with a mean of 19.7 cm. Concerning qualitative traits, observations demonstrated that almost all the modalities of traits were found in the core collection. Only “spheroid shape”, a modality of seed shape (SEES) was not captured in the core collection (Table 6). However, that modality had a very low proportion (2.5%) in the entire collection. In addition, multivariate analyses showed that there is no significant difference between the core collection morphological characteristics compared to those of the entire collection (Table 6).

Discussion

In order to assess the genetic diversity within *Vitellaria paradoxa*, a collection of 333 genotypes was gathered from the in situ superior shea collection in Côte d’Ivoire. The genotypic information from these individuals was used to investigate genetic diversity and population genetics, which could provide valuable insights for future breeding efforts, such as genome-wide association

Table 6 Quantitative and qualitative traits associated with the core and entire collection of *Vitellaria paradoxa*

Quantitative Traits		Means ± standard deviation		F-value	p-value
		Initial collection (N= 160)	Core set (N=52)		
Trunk Circumference (cm)		144.6 ± 45.93	149.7 ± 44.34	0.48	0.49
Petiole length (cm)		8.29 ± 1.49	8.12 ± 1.61	0.51	0.475
Limb length (cm)		14.77 ± 2.31	14.33 ± 2.13	1.44	0.232
Limb width (cm)		4.61 ± 0.79	4.44 ± 0.78	1.82	0.18
Nut length (cm)		3.11 ± 0.38	3.66 ± 0.42	0.19	0.67
Nut width (cm)		2.9 ± 0.3	2.26 ± 0.41	0.78	0.38
Nut weight (g)		9.55 ± 2.61	10.09 ± 2.73	1.6	0.21
Qualitative traits	Modalities	Proportions N(%)		χ ²	p-value
		Initial collection (N= 160)	Core set (N=52)		
Tree Growth Habit (TGH)	Erect	55 (34.38)	13 (25)	2.73	0.26
	Semi-erect	60 (37.5)	26 (50)		
	Spreading	45 (28.12)	13 (25)		
Leaf Apex Shape (LAS)	Acute	47 (29.38)	18 (34.62)	4.4	0.22
	Acuminate	24 (15)	13 (25)		
	Retuse	49 (30.63)	11 (21.15)		
	Obtuse	40 (25)	10 (19.23)		
Adult Leaf Color (ALC)	Light green	13 (8.13)	3 (5.77)	0.86	0.65
	Green	133 (83.13)	46 (88.46)		
	Dark green	14 (8.75)	3 (5.77)		
Seed Coat Color (SCC)	Creamish	13 (8.13)	8 (17.33)	3.08	0.54
	Dull brown	3 (1.88)	1 (1.92)		
	Brown	78 (48.75)	20 (38.46)		
	Pale brown	15 (9.38)	5 (9.62)		
	Dark brown	51 (31.88)	18 (34.62)		
Seed shape (SEES)	Spheroid	4 (2.5)	0 (0)	2.81	0.42
	Ellipsoid	37 (23.13)	16 (30.77)		
	Oval	65 (40.63)	24 (46.15)		
	Ovoid	54 (33.75)	14 (26.92)		

studies (GWAS). To the best of our knowledge, this study is the first one that uses single nucleotide polymorphisms (SNPs) to investigate the genetic diversity and population structure of *Vitellaria paradoxa* subspecies in Côte d'Ivoire.

Single nucleotide polymorphism markers and mutation types

In this study, 7,559 high-quality SNP markers were retained after data filtering. The analysis of SNPs distribution and mutation types in the genome was conducted. A higher frequency of transitions than transversions was observed (transitions/transversions=1.6). A ratio of 1.3 in transition SNPs and transversion SNPs was obtained in shea trees in Uganda [28]. Similar transition/transversion results have been reported in other plant species, such as *Hevea brasiliensis* [50], *Camellia sativa* [16], *Vigna unguiculata* [47], *Colocasia esculenta* [51], and *Oryza sativa* [52], and in living organisms in general [53]. This suggests that during the natural selection of *V. paradoxa*, transition mutations tend to be more tolerated than transversion mutations.

This may be due to the presence of higher frequencies of synonymous mutations in the protein-coding sequences [53]. However, the non-synonymous SNPs are of interest for this study because they generate new variants that are important in breeding programs [28]. In natural selection, non-synonymous SNPs are important for the conservation of shea tree species because the variants become more adapted to environmental changes [28]. Transition mutation types are also important in evaluating the distribution, extent, and amount of genetic variation among and within shea tree populations [28]. The frequencies of A/G and C/T transitions were similar (A/G: 30.6% and T/C: 30.8%). This result is consistent with those obtained for shea tree and rice species [28, 52].

Gene diversity

The genetic diversity and the population structure of a species are good indicators of its management and conservation status [28].

The findings in this study revealed moderate level of genetic diversity (HE: 0.26) and moderate polymorphism information content (PIC: 0.24) in *V. paradoxa*. This is consistent with what has been reported for shea trees in East Africa using SNP [28] and SSR markers [25], in West Africa with SSR markers [24, 26] and in the natural range of the species using RAPD markers [18, 19]. This suggests that the SNP markers used in our study were reasonably informative markers. In contrast, a *V. paradoxa* diversity study in Ghana using SNP markers revealed very low genetic diversity [54]. Other studies showed low genetic

diversity in *V. paradoxa* using SSR markers [21, 55]. These results suggest that the genetic diversity in shea tree species is low to moderate in its natural range.

The moderate genetic diversity may be the result of selection based on farmer-preferred traits such as large fruit size, tasty pulp, and high oil content [56, 57]. Furthermore, this moderate genetic diversity provides an opportunity to generate shea varieties that can be successfully grown in shea belts in different geographical districts [28]. Studies involving other plant species such as *Camelina sativa* [16], *Ziziphus jujuba* [58], and winter wheat [59] observed the same trend and attributed these findings to the bi-allelic nature and low mutation rates of SNPs. Our marker density is sufficient to perform genome-wide association studies, as a genotype panel with minor allele frequency (MAF) > 0.1 is desirable for genome-wide association mapping [17, 52].

Population structure and relationships

Population structure analyses are essential for understanding genetic diversity and facilitating subsequent association mapping studies [59]. In our study, the cluster pattern in Structure grouped the shea trees into three groups, which concurs with the DAPC, PCoA, and the neighbor-joining tree reports. These results were also consistent with a previous study using SNP markers in *V. paradoxa* [54] in Ghana. However, our results were different from those obtained in Uganda using SNP markers. The authors clustered their accessions into two groups. These results can confirm that West African shea trees exhibit higher genetic diversity than East African shea trees using SSR markers [5]. The structure of the Ivorian shea population collection suggests a geographical effect. In fact, the genotypes clustered in GP1 are from the districts of Poro, Tchologo, and Hambol, while the genotypes collected in GP2 and GP3 come from the district of Bagoué. There are thus two geographical groups of shea trees in our collection: one group in the western part of the Poro district, which can also be divided into two subgroups representing GP2 and GP3, and the second geographical group located in the eastern part of the Poro district (see Fig. 1). Similar results were observed in the Ugandan shea tree population structure study [28]. These findings suggest that these geographical groups can serve as genetic resources for hybridization programs to create improved varieties and align with the objectives of the annotated genome [17].

However, a certain number of individuals showed a mixed genetic profile across the three groups (Fig. 2c). This observation is consistent with findings on *V. paradoxa* using DArTseq SNP markers [28, 54]. Gene flow between individuals in the neighborhood or individuals overlapping the study areas may be responsible for the

presence of these admixed individuals. As a result, a limited number of accessions may show clear membership to one group, while the majority may show some degree of membership to the three groups. This indicates that common allelic/gene combinations continue among the collection of shea trees.

Genetic differentiation of populations

The F_{ST} value is the most relevant F-statistic used to study the degree of genetic differentiation between and within populations [52]. In this study, a fixation index (F_{ST}) value of 0.004 was found for the whole population and the low pairwise F_{ST} values found between the three groups (Table 3) indicate low genetic differentiation between these groups [52, 60]. Similar results were found in the genetic differentiation in *V. paradoxa* populations implying SNP markers [28, 54] and SSR markers [21]. The neighbor-joining tree based on genetic distances confirmed the low differentiation between groups, with few bootstrap values higher than 50.45, with the main branches having bootstrap values lower than 50.45 (Fig. 3). This trend has been reported in shea tree populations using nuclear SSR markers [24]. These results suggest the presence of extensive and anthropogenic gene flow, outcrossing and admixture in the study area. Comparison of our findings with previous studies revealed that low genetic differentiation is observed in shea tree populations when the study is carried out in a small area, limited to one or two countries [21, 25, 54, 55, 61] or within a subspecies [24]. This trend is expected in the shea tree due to its status as a long-lived woody perennial, insect-pollinated outcross, and widespread in a continuous range [24]. However, high genetic differentiation is observed when the study is conducted in the natural range of the subspecies [5, 18, 19]. This high differentiation can be explained by the fact that these studies included both subspecies of *V. paradoxa*, and differentiation is observed between the *nilotica* and *paradoxa* subspecies [5, 18, 19]. These findings suggest that many individual shea trees should be considered for efficient sampling of genetic diversity within a population. This is consistent with the proposal to develop a breeding population of *Vitellaria paradoxa* [18].

The findings from Structure are consistent with the outcomes from the AMOVA, where the total variation was primarily attributed to variations within the groups. Furthermore, it has been suggested that a high N_m value, indicative of substantial gene flow, can lead to low differentiation between populations [16]. Our study aligned with this observation, as a very high N_m value of 59.02 was obtained. This suggests that in terms of gene flow, the N_m value obtained in our study was higher than that reported in shea trees using SSR markers [55]. Therefore,

our results suggest that population genetic differentiation using SNP markers is more informative than other markers.

Allelic pattern and genetic diversity indices

The allelic patterns and genetic diversity indices offered valuable insights into the genetic diversity present within each of the three genetic groups. While the three groups exhibited relatively close levels of expected heterozygosity (HE), GP1 showed a slightly higher HE compared to GP2 and GP3. This indicates that GP1 has a slightly higher level of diversity compared to the other groups, as HE considers both the number of alleles (referred to as richness) and the distribution (or evenness) of those alleles within a population. This could be explained by the origin of the genotypes clustered in GP1. In fact, the genotypes in GP1 are from three different districts (Hambol, Poro, and Tchologo), while the genotypes in GP2 and GP3 are mainly from the Bagoué district.

Genetic diversity ranged from low to high. This result is consistent with recent studies using SNP markers [28] and previous population genetic studies using SSR markers [5, 21, 24, 26, 55]. Understanding the genetic diversity within *V. paradoxa* populations is essential, this is fundamental element for robust shea tree improvement programs and for future studies using genomic screening methods such as marker-assisted screening (MAS) and genome-wide association studies (GWAS) [17].

Core collection

Sustaining living collections, a common practice for perennial tree crops, can be a costly and labor-intensive endeavor. Creating core collections is an efficient strategy for managing germplasm, effectively reducing costs while retaining the highest possible genetic diversity within the germplasm pool while minimizing redundancy [29, 30]. Developing core collections has been the focus of various approaches [62–64], and the choice of the most appropriate evaluation methods depends on the specific objectives associated with these core collections [65].

In this study, we used the “maximum length sub-tree function” of DARwin version 6.0.21 to carefully curate a core set of 100 superior shea trees (30% of the initial population). The analysis indicates that the core collection successfully captures the full genetic diversity of the entire population. This is evidenced by the PCA plot, which showed good coverage of the core sample across the whole panel (Fig. 4), and the representation of each genetic group within the core, which includes 48% from GP1, 30% from GP2 and 22% from GP3. These proportions reflect the overall genetic structure of the entire population, ensuring that no significant genetic diversity is lost in the core. In addition, several genetic parameters

in the core closely mirror those of the total population: heterozygosity, allele frequency, and polymorphism rates. This similarity underscores the effectiveness of the core collection in maintaining the genetic variation present in the larger population, making it a robust subset for further genetic studies and conservation efforts. Thus, the core collection can safely be used for genetic research, breeding programs, and conservation strategies without compromising the genetic integrity of the original population [66].

The proportion of this core set is similar to the proportion established for other crops, such as palm oil (31.2%) [67]. However, it differs from the core set proportions observed for *E. oleifera* (6.4%) [68], *S. superba* (19.87%) [69], and *P. massoniana* (19.46%) [66]. A core collection sample size of 5–10% of the original germplasm resources can be sufficient to represent over 70% of the genetic variation present in the entire germplasm [29]. For effective conservation of the genetic diversity of the entire population, a size of 20–30% of the population is required for the core set [70].

However, there is no universal approach to selecting a core size. It depends on factors such as the extent of variability and redundancy within the collection, the resources available for core set management, and the frequency of species regeneration [70, 71].

Field genebanks provide convenient and immediate access to germplasm resources. However, relying solely on in situ conservation is not the most reliable long-term conservation strategy. Shea trees can be uprooted by adverse weather conditions, such as tropical cyclones [15]. To address this, a shift in focus for germplasm conservation of shea trees can be directed towards the identified core set. A viable approach to germplasm conservation is to clone the identified core set through grafting for ex situ conservation. This method ensures the precise conservation of maximum genetic diversity while reducing redundancy and avoiding genetic erosion. Furthermore, the number of accessions to be planted in the field can be reduced, resulting in a more manageable and cost-effective approach compared to progeny trials.

The use of molecular markers to establish a core collection offers distinct advantages because they can accurately capture genetic diversity regardless of plant growth status, developmental stage, and environmental conditions [72]. Importantly, genetic diversity is often positively associated with population persistence and resilience to environmental changes [73].

Morphological characteristics of the genotyped shea trees

Variability in superior shea trees has been widely demonstrated using morphological traits [15, 46]. A similar trend was observed in this study. Using genotype

clustering from the Structure software showed significant differences between the groups for trunk circumference, nut weight, and leaf width. These differences are essentially between GP1 and GP2 or GP3. No significant differences were observed between GP2 and GP3. These results suggest that the three groups obtained with molecular markers reflected two morphological groups. This can be explained by the significant influence of the environment on the expression of morphological traits. A savanna gradient has been reported in the study area [46]. The effect of climate on the expression of shea tree morphological traits has been reported in several countries [46]. In addition, we suggest that the SNP markers that structured Bagoué superior shea trees into two groups are synonymous SNPs, while those that separated GP1 and GP2/GP3 could be designed as nonsynonymous SNPs. SNPs may change the encoded amino acids (nonsynonymous SNP) and change the amino acid sequence or be silent (synonymous SNP), thereby maintaining the amino acid sequence or simply occurring in the noncoding regions [28]. A Mantel test realized between morphological traits and SNP markers did not show significant differences (Mantel $r=0.0176$; $p=0.3324$). This suggests that the morphological characteristics observed in the groups are not the direct effects of the SNP markers. In addition to environmental conditions, the varying mineral composition of the soil from one ecological zone to another could also have an effect on the morphological trait expression. Studies have reported soil mineral composition effects on shea tree morphology in Mali [74], Eastern Ghana [75], and West Africa [76].

The morphological characteristics of the core collection showed a similar trend to those of the entire collection. A similar trend was found in a core establishment in *Synsepalum dulcificum* [31]. These results suggest that the core collection captured the full morphological characteristics of the entire collection. It is also important to ensure that the core collection is well representative of the whole collection in all genetic aspects, confirming the quality of the established core based on morphological traits as well as molecular markers (SNPs). Therefore, morphological traits could be confidently used in the establishment of core collections of *V. paradoxa* species without loss of genetic diversity.

Conclusion

In this study, we used high-throughput DArTseq technology to investigate the genetic diversity and population structure of *Vitellaria paradoxa* in Côte d'Ivoire. The primary objective was to explore the potential utility of SNP markers for various genomic analyses in

the context of genetic improvement efforts. Our data revealed that the collection under investigation exhibited a moderate degree of genetic diversity.

This rich genetic diversity serves as a promising foundation upon which to develop novel *Vitellaria* cultivars boasting desirable traits, including high yield potential, high oil production, and resilience to biotic and abiotic stresses, all while being well-suited for adaptation to diverse environmental conditions.

Furthermore, our research unveiled the presence of three genetic groups within the study population. The differentiation between these groups can be attributed to a combination of factors and natural selection pressures. Notably, the three groups demonstrated very close diversity across multiple parameters, including Shannon's information index (*I*), expected heterozygosity (*HE*), and unbiased expected heterozygosity (*uHE*).

A core collection of 100 superior shea trees, representing 30% of the entire population was captured. This study marks the inaugural endeavor to molecularly characterize and validate the creation of a core set for shea tree germplasm resources. The core collection successfully captured all the variation in morphological traits and the alleles present within the accessions, while preserving the genetic diversity and structure of the original population.

These findings provide important information for suitable conservation and future allelic/gene identification using genome-wide association studies (GWAS) and marker-assisted selection (MAS) to enhance genetic gain in *V. paradoxa* breeding programs.

Abbreviations

AMOVA	Analysis of molecular variance
DAPC	Discriminant Analysis of Principal Components
<i>F_{ST}</i>	Genetic differentiation
PCoA	Principal Coordinate Analysis
PCA	Principal Component Analysis
<i>HE</i>	Expected heterozygosity
<i>HO</i>	Observed heterozygosity
<i>I</i>	Shannon's information index of diversity
GBS	Genotyping by-sequencing
GWAS	Genome-wide association study
PIC	Polymorphism Information Content
SNP	Single nucleotide polymorphism
MAS	Marker-assisted selection
TC	Trunk Circumference
PL	Petiole length
LL	Limb length
LW	Limb width
NL	Nut length
NWD	Nut width
NWG	Nut weight
TGH	Tree Growth Habit
LAS	Leaf Apex Shape
ALC	Adult Leaf Color
SEES	Seed Coat Color

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12870-024-05617-0>.

Supplementary files: A.pdf file containing tables and figures that support findings in the manuscript.

Acknowledgements

We acknowledge the support of the plant Genetics laboratory at Gembloux Agro Bio-Tech and the Department of Biochemistry-Genetics, Educational and Research Unit of Genetics at the Péléforo Gon Coulibaly University in Korhogo, which made this study possible. As a person, we would like to thank Mr Blé Kpagni Antoine (Senior Technician at University Peleforo Gon Coulibaly of Korhogo) is acknowledged for putting us in contact with the farmers. We also thank Ekra Jean-Yves (Ph.D student at Sokoine University of Agriculture, Tanzania), Soro Nangalourou Adama and the ANADER agents who accompanied us in the field.

Authors' contributions

A.J.P.A., E.A.D and L.L. conceptualized the topic. A.J.P.A., D.N.D, and E.A.D analyzed the data. A.J.P.A., Y.K., D.N.D, L.S., E.A.D and L.L. set up the methodology. A.J.P.A., S.S., S.D.M.Y., D.N.D., E.A.D., and L.L. provided resources. N.D., T.A. and L.L. supervised the study. A.J.P.A., S.D.M.Y., D.N.D., T.A., E.A.D., C.D.C., N.D., and L.L. validated the work. A.J.P.A., T.A., and L.L. wrote the main manuscript text. All authors reviewed approved the final manuscript.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the University of Liege Scientific Research mobility (2019/MOB/02924 and 2021/MOB/00089) and the ULiege-PACODEL "Valorization / Reinforcement" Grant.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the Science Data Bank (ScienceDB) repository, <https://doi.org/10.57760/sciencedb.17559>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Plant Genetics and Rhizosphere Processes Lab, University of Liege, Gembloux Agro Bio-Tech, Terra Research Center, Passage Des Déportés 2, Gembloux 5030, Belgium. ²Faculty of Biological Sciences, Department of Biochemistry-Genetics, Educational and Research Unit of Genetic, University of Peleforo Gon Coulibaly (UPGC), Korhogo BP 1328, Côte d'Ivoire. ³African Center for Shea Research and Application (CRAK), Korhogo, Côte d'Ivoire. ⁴AgricultureLife, University of Liege, Gembloux Agro Bio-Tech, Passage Des Déportés 2, Gembloux 5030, Belgium. ⁵Faculté Des Sciences Agronomiques, Département de Production Végétale, Université Catholique de Bukavu (UCB), Bukavu, Democratic Republic of the Congo. ⁶Functional and Evolutionary Entomology, University of Liege, Gembloux Agro Bio-Tech, Passage Des Déportés 2, 5030 Gembloux, Belgium. ⁷Genetics, Biotechnology, and Seed Science Unit (GBioS), Department of Plant Sciences, Faculty of Agronomic Sciences, University of Abomey-Calavi, 01 BP 526, Abomey-Calavi, Benin.

Received: 25 March 2024 Accepted: 23 September 2024

Published online: 01 October 2024

References

- Naughton CC, Lovett PN, Mihelcic JR. Land suitability modeling of shea (*Vitellaria paradoxa*) distribution across sub-Saharan Africa. *Appl Geogr*. 2015;58:217–27.
- Sanou H, Lamien N. *Vitellaria paradoxa*, shea butter tree. Conservation and Sustainable Use of Genetic Resources of Priority Food Tree Species in sub-Saharan Africa. Rome: Biovers Internat; 2011.
- Boffa JM, Knudson DM, Yameogo G, Nikiema P. Shea nut (*Vitellaria paradoxa*) production and collection in agroforestry parklands of Burkina Faso. Rome: Non-Wood For Prod FAO; 1996.
- Maranz S, Kpikpi W, Wiesman Z, De Saint Sauveur A, Chapagain B. Nutritional values and indigenous preferences for Shea Fruits (*Vitellaria paradoxa* C.F. Gaertn. F.) in African Agroforestry Parklands. *Econ Bot*. 2004;58(4):588–600.
- Allal F, Sanou H, Millet L, Vaillant A, Camus-Kulandaivelu L, Logossa ZA, et al. Past climate changes explain the phylogeography of *Vitellaria paradoxa* over Africa. *Heredity* août. 2011;107(2):174–86.
- Neumann K, Kahlheber S, Uebel D. Remains of woody plants from Saouga, a medieval west African village. *Veg Hist Archaeobotany*. 1998;7(2):57–77.
- Masters ET, Yidana JA, Lovett PN. Reinforcing sound management through trade: Shea tree products in Africa. *Unasylva*. 2004;55(219):46–52.
- Carette C, Malotaux M, Leeuwen MV, Tolkamp M, Kassaw A, Dordaa S, et al. Shea nut and butter in Ghana Opportunities and constraints for local processing. In 2009. Disponible sur: <https://www.semanticscholar.org/paper/Shea-nut-and-butter-in-Ghana-Opportunities-and-for-Carette-Malotaux/bd5a492a393875a3bd8701c61a0f47b9e2baa1bf?sort=relevance&citationIntent=methodology>. Cité 19 Oct 2023.
- Glew D, Lovett PN. Life cycle analysis of shea butter use in cosmetics: from parklands to product, low carbon opportunities. *J Clean Prod*. 2014;68:73–80.
- Rousseau K, Gautier D, Wardell DA. Coping with the upheavals of globalization in the shea value chain: the maintenance and relevance of upstream shea nut supply chain organization in western Burkina Faso. *World Dev*. 2015;66:413–27.
- IUCN. IUCN Red List of Threatened Species: *Vitellaria paradoxa*. IUCN Red List Threat Species. 1 janv 1998; Disponible sur: <https://www.iucnredlist.org/en>. Cité 27 Sep 2023.
- Boffa JM. Opportunities and challenges in the improvement of the shea (*Vitellaria paradoxa*) resource and its management. Occasional paper. 2015;24:54.
- Diarrassouba N, Yao SDM, Traoré B. Identification participative et caractérisation des arbres élités de karité dans la zone de production en Côte d'Ivoire. 2017 p. 15 pages. University Peleforo Gon Coulibaly, Côte d'Ivoire (projet FIRCA/Karité). Report No.: N° 069/2016.
- Sandwidi A, Diallo BO, Lamien N, Vinceti B, Sanon K, Coulibaly P, et al. Participatory identification and characterisation of shea butter tree (*Vitellaria paradoxa* C.F. Gaertn.) ethnovarieties in Burkina Faso. *Fruits Int J Trop Subtrop Hortif*. 2018;73(3):141–52.
- Attikora AJP, Diarrassouba N, Yao SDM, Clerck CD, Silue S, Alabi T, et al. Morphological traits and sustainability of plus shea trees (*Vitellaria paradoxa* C.F. Gaertn.) in Côte d'Ivoire. *Biotechnol Agron Société Environ*. 25 sept 2023; Disponible sur: <https://orbi.uliege.be/handle/2268/307173>. Cité 19 Oct 2023.
- Luo Z, Brock J, Dyer JM, Kutchan T, Schachtman D, Augustin M, et al. Genetic Diversity and Population Structure of a *Camelina sativa* Spring Panel. *Front Plant Sci*. 2019;10. Disponible sur: <https://www.frontiersin.org/articles/10.3389/fpls.2019.00184>. Cité 19 Oct 2023.
- Hale I, Ma X, Melo ATO, Padi FK, Hendre PS, Kingan SB, et al. Genomic Resources to Guide Improvement of the Shea Tree. *Front Plant Sci*. 2021;12. Disponible sur: <https://www.frontiersin.org/articles/10.3389/fpls.2021.720670>.
- Bouvet JM, Fontaine C, Sanou H, Cardé C. An analysis of the pattern of genetic variation in *Vitellaria paradoxa* using RAPD markers. *Agrofor Syst*. 2004;60(1):61–9.
- Fontaine C, Lovett PN, Sanou H, Maley J, Bouvet JM. Genetic diversity of the shea tree (*Vitellaria paradoxa* C.F. Gaertn.), detected by RAPD and chloroplast microsatellite markers. *Heredity*. 2004;93(6):639–48.
- Attikora AJP, Silué S, Yao SDM, De Clerck C, Shumbe L, Diarrassouba N, et al. An innovative optimized protocol for high-quality genomic DNA extraction from recalcitrant Shea tree (*Vitellaria paradoxa*, C.F. Gaertn.) plant and its suitability for downstream applications. *Mol Biol Rep*. 2024;51(1):171.
- Kelly B, Hardy O, Bouvet JM. Temporal and spatial genetic structure in *Vitellaria paradoxa* (shea tree) in an agroforestry system in southern Mali. *Mol Ecol*. 2004;13(5):1231–40.
- Cardé C, Vaillant A, Sanou H, Kelly BA, Bouvet JM. Characterization of microsatellite markers in the shea tree (*Vitellaria paradoxa* C. F. Gaertn) in Mali. *Mol Ecol Notes*. 2005;5(3):524–6.
- Allal F, Vaillant A, Sanou H, Kelly B, Bouvet JM. Isolation and characterization of new microsatellite markers in shea tree (*Vitellaria paradoxa* C. F. Gaertn). *Mol Ecol Resour*. 2008;8(4):822–4.
- Logossa ZA, Camus-Kulandaivelu L, Allal F, Vaillant A, Sanou H, Kokou K, et al. Molecular data reveal isolation by distance and past population expansion for the shea tree (*Vitellaria paradoxa* C.F. Gaertn) in West Africa. *Mol Ecol*. 2011;20(19):4009–27.
- Gwali S, Vaillant A, Nakabonge G, Okullo JBL, Eilu G, Muchugi A, et al. Genetic diversity in shea tree (*Vitellaria paradoxa* subspecies nilotica) ethno-varieties in Uganda assessed with microsatellite markers. *For Trees Livelihoods*. 2015;24(3):163–75.
- Abdulai I, Krutovsky KV, Finkeldey R. Morphological and genetic diversity of shea tree (*Vitellaria paradoxa*) in the savannah regions of Ghana. *Genet Resour Crop Evol*. 2017;64(6):1253–68.
- Mohammed HI, Mohammed Z, Warra A, Abdulrahman Y, Sabo I, Ibrahim G, et al. Physicochemical and Genetic Diversity Studies of *Vitellaria paradoxa* in Northern Nigeria. *J Curr Biomed Res*. 2022;2(1):19–37.
- Odoi JB, Adjei EA, Hendre P, Nantongo JS, Ozimati AA, Badji A, et al. Genetic diversity and population structure among Ugandan shea tree (*Vitellaria paradoxa* subsp. nilotica) accessions based on DArTSeq markers. *Crop Sci*. 2023;63(4):2297–309.
- Brown AHD. Core collections: a practical approach to genetic resources management. *Genome*. 1989;31(2):818–24.
- Frankel OH, Brown A. Plant genetic resources today: a critical appraisal. In: Holden J, Williams J (eds) *Crop genetic resources: conservation and evaluation*. London: George Allen and Unwin; 1984. p. 249–57.
- Tchokponhoué DA, Achigan-Dako EG, N'Danikou S, Nyadanu D, Kahane R, Houéto J, et al. Phenotypic variation, functional traits repeatability and core collection inference in *Synsepalum dulcificum* (Schumacher & Thonn.) Daniell reveals the Dahomey Gap as a centre of diversity. *Sci Rep*. 2020;10(1):19538.
- Dekoula CS, Kouame B, N'goran EK, Yao FG, Ehounou JN, Soro N. Impact De La Variabilité Pluviométrique Sur La Saison Culturelle Dans La Zone De Production Cotonnière En Côte d'Ivoire. *Eur Sci J ESJ*. 2018;14(12):143–143.
- Brou YT. Climat, mutations socio-économiques et paysages en Côte d'Ivoire. [Université des Sciences et Technologies de LILLE, (2005) 212.]: Université Lille1 - Sciences et Technologies (Lille, France). 2005.
- Kouamé B, Ehounou JN, Kassim KE, Dekoula CS, Yao GF, N'goran EK, et al. Caractérisation Des Paramètres Agroclimatiques Clés De La Saison Culturelle En Zone De Contact Forêt/Savane De Côte-d'Ivoire. *Eur Sci J ESJ*. 2018;14(36):243–243.
- N'Guessan KA, Diarrassouba N, Alui KA, Nangha KY, Fofana JJ, Yao-Kouame A. Indicateurs de dégradation physique des sols dans le Nord de la Côte d'Ivoire: Cas de Boundiali et Ferkessedougou. *Afr Sci Rev Int Sci Technol*. 2015;11(3):115–28.
- Koffi AA, Kouassi FA, N'Goran SBK, Soro D. Les Loranthaceae, parasites des arbres et arbustes : cas du département de Katiola, au nord de la Côte d'Ivoire. *Int J Biol Chem Sci*. 2014;8(6):2552–9.
- Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, et al. Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods Mol Biol Clifton NJ*. 2012;888:67–89.
- Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinforma Oxf Engl*. 2012;28(19):2537–9.
- Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 1980;32(3):314–31.
- Dereeper A, Nicolas S, Le Cunff L, Bacilieri R, Doligez A, Peros JP, et al. SNIPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics*. 2011;12(1):134.

41. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.* 2007;81(3):559–75.
42. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics.* 2000;155(2):945–59.
43. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol.* 2005;14(8):2611–20.
44. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 2010;11(1):94.
45. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 2021;49(W1):W293–6.
46. Yao SDM, Diarassouba N, Attikora A, Fofana IJ, Dago DN, Silue S. Morphological diversity patterns among selected elite Shea trees (*Vitellaria paradoxa* C.F. Gaertn.) from Tchologo and Bagoué districts in Northern Côte d'Ivoire. *Int J Genet Mol Biol.* 2020;12:1–10.
47. Ketema S, Tesfaye B, Kenehi G, Fenta BA, Assefa E, Greliche N, et al. DAR-Seq SNP-based markers revealed high genetic diversity and structured population in Ethiopian cowpea [*Vigna unguiculata* (L.) germplasm]. *PLOS ONE.* 2020;15(10):e0239122.
48. Perrier X, Jacquemoud-Collet JP. DARwin software: dissimilarity analysis and representation for Windows (version 6.0.021). 2006. Available at <http://darwin.cirad.fr/>. Accessed 29 Sep 2024.
49. Lassois L, Denancé C, Ravon E, Guyader A, Guisnel R, Hibrand-Saint-Oyant L, et al. Genetic Diversity, Population Structure, Parentage Analysis, and Construction of Core Collections in the French Apple Germplasm Based on SSR Markers. *Plant Mol Biol Rep.* 2016;34(4):827–44.
50. Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Junior EJS, de Souza Gonçalves P, et al. De Novo Assembly and Transcriptome Analysis of Rubber Tree (*Hevea brasiliensis*) and SNP Markers Development for Rubber Biosynthesis Pathways. *PLOS ONE.* 2014;9(7):665.
51. Fufa TW, Abtey WG, Amadi CO, Oselebe HO. DARSeq SNP-based genetic diversity and population structure studies among taro [*Colocasia esculenta* (L.) Schott] accessions sourced from Nigeria and Vanuatu. *PLOS ONE.* 2022;17(11):302.
52. Peringottillam M, Kunhiraman Vasumathy S, Selvakumar HKK, Alagu M. Genetic diversity and population structure of rice (*Oryza sativa* L.) landraces from Kerala, India analyzed through genotyping-by-sequencing. *Mol Genet Genomics.* 2022;297(1):169–82.
53. Guo C, McDowell IC, Nodzinski M, Scholtens DM, Allen AS, Lowe WL, et al. Transversions have larger regulatory effects than transitions. *BMC Genomics.* 2017;18(1):394.
54. Anyomi WE, Barnor MT, Eleblu JSY, Danquah A, Avicor SW, Ofori K, et al. Elucidation of the Genetic Diversity within Some In Situ Shea Germplasm in Ghana. *Agronomy.* 2023;13(9):2256.
55. Sanou H, Lovett PN, Bouvet JM. Comparison of quantitative and molecular variation in agroforestry populations of the shea tree in (*Vitellaria paradoxa* C.F. Gaertn.) Mali. *Mol Ecol.* 2005;14(8):2601–10.
56. Odoi JB, Muchugi A, Okia CA, Gwali S, Odong TL. Local knowledge, identification and selection of shea tree (*Vitellaria paradoxa*) ethnovarieties for pre-breeding in Uganda. *J Agric Nat Resour Sci.* 2020;7(1):22–33.
57. Lovett PN, Haq N. Diversity of the Sheanut tree (*Vitellaria paradoxa* C.F. Gaertn.) in Ghana. *Genet Resour Crop Evol.* 2000;47(3):293–304.
58. Chen W, Hou L, Zhang Z, Pang X, Li Y. Genetic Diversity, Population Structure, and Linkage Disequilibrium of a Core Collection of *Ziziphus jujuba* Assessed with Genome-wide SNPs Developed by Genotyping-by-sequencing and SSR Markers. *Front Plant Sci.* 2017;8:575.
59. Eltaher S, Sallam A, Belamkar V, Emara HA, Nower AA, Salem KFM, et al. Genetic Diversity and Population Structure of F3:6 Nebraska Winter Wheat Genotypes Using Genotyping-By-Sequencing. *Front Genet.* 2018;9. Disponible sur: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00076>. Cited 19 Oct 2023.
60. Frankham R, Ballou JD, Briscoe DA. Higher Education from Cambridge University Press. Cambridge University Press; 2010. Introduction to Conservation Genetics. Disponible sur: <https://www.cambridge.org/highereducation/books/introduction-to-conservation-genetics/696B4E558C93F7FBF9C33D6358EA7425>. Cited 19 Oct 2023.
61. Odoi JB, Adjei EA, Hendre P, Nantongo JS, Ozimati AA, Badji A, et al. Genetic diversity and population structure among Ugandan shea tree (*Vitellaria paradoxa* subsp. nilotica) accessions based on DARSeq markers. *Crop Sci.* 2023;63(4):63(4):2297–309.
62. Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, et al. PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics.* 2007;23(16):2155–62.
63. De Beukelaer H, Davenport GF, Fack V. Core Hunter 3: flexible core subset selection. *BMC Bioinformatics.* 2018;19(1):203.
64. Odong TL, van Heerwaarden J, Jansen J, van Hintum TJL, van Eeuwijk FA. Statistical Techniques for Defining Reference Sets of Accessions and Microsatellite Markers. *Crop Sci.* 2011;51(6):2401–11.
65. Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJL. Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theor Appl Genet.* 2013;126(2):289–305.
66. Bai Q, Cai Y, He B, Liu W, Pan Q, Zhang Q. Core set construction and association analysis of *Pinus massoniana* from Guangdong province in southern China using SLAF-seq. *Sci Rep.* 2019;9(1):13157.
67. Gan ST, Teo CJ, Manirasa S, Wong WC, Wong CK. Assessment of genetic diversity and population structure of oil palm (*Elaeis guineensis* Jacq.) field genebank: A step towards molecular-assisted germplasm conservation. *PLOS ONE.* 2021;16(7):e0255418.
68. Ithnin M, Teh CK, Ratnam W. Genetic diversity of *Elaeis oleifera* (HBK) Cortes populations using cross species SSRs: implication's for germplasm utilization and conservation. *BMC Genet.* 2017;18(1):37.
69. Bai Q, He B, Cai Y, Lian H, Zhang Q, Liang D, et al. Genetic Diversity and Population Structure of *Schima superba* From Southern China. *Front Ecol Evol.* 2022;10. Disponible sur: <https://www.frontiersin.org/articles/10.3389/fevo.2022.879512>. Cited 19 Oct 2023.
70. Yonezawa K, Nomura T, Morishima H, Nonura T. Sampling strategies for use in stratified germplasm collections. In 1995. Disponible sur: <https://www.semanticscholar.org/paper/Sampling-strategies-for-use-in-stratified-germplasm-Yonezawa-Nomura/d24d31793c09f5c9044085d6ebab1bc924a36d1a>. Cited 19 Oct 2023.
71. Hintum TJL van, Brown AHD, Spillane C. Core Collections of Plant Genetic Resources. Rome: Bioversity International; 2000. p. 51.
72. Agarwal M, Shrivastava N, Padh H. Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep.* 2008;27(4):617–31.
73. Schlottfeldt S, Walter MEMT, de Carvalho ACPLF, Soares TN, Telles MPC, Loyola RD, et al. Multi-objective optimization for plant germplasm collection conservation of genetic resources based on molecular variability. *Tree Genet Genomes.* 2015;11(2):16.
74. Sanou H, Picard N, Lovett PN, Dembélé M, Korbo A, Diarisso D, et al. Phenotypic Variation of Agromorphological Traits of the Shea Tree, *Vitellaria paradoxa* C.F. Gaertn., in Mali. *Genet Resour Crop Evol.* 2006;53(1):145–61.
75. Moore S. The role of *Vitellaria paradoxa* in poverty reduction and food security in the Upper East region of Ghana. *Earth Environ.* 2008;3:209–45.
76. Bondé L, Ouédraogo O, Ouédraogo I, Thiombiano A, Boussim JI. Variability and estimating in fruiting of shea tree (*Vitellaria paradoxa* C.F. Gaertn.) associated to climatic conditions in West Africa: implications for sustainable management and development. *Plant Prod Sci.* 2019;22(2):143–58.

Publisher's Note

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