Beverage Plant Research

ARTICLE

https://doi.org/10.48130/BPR-2021-0015
Beverage Plant Research **2021**, 1: 15

Selecting SNP markers reflecting population origin for cacao (*Theobroma cacao* L.) germplasm identification

Osman A. Gutiérrez[1*], Kathleen Martinez[1], Dapeng Zhang[2*], Donald S. Livingstone[3], Chris J. Turnbull[4,5], and Juan Carlos Motamayor[5]

[1] *USDA-ARS Subtropical Horticulture Research Station, 13601 Old Cutler Road, Miami, FL 33158, USA*
[2] *USDA-ARS, Beltsville Agricultural Research Center, SPCL, 10300 Baltimore Avenue, Bldg. 001, Rm. 223, BARC-W, Beltsville, MD 20705, USA*
[3] *Mars Wrigley Plant Science Laboratory, Davis, CA 95616, USA*
[4] *School of Agriculture, Policy and Development, University of Reading, RG6 6AR, UK*
[5] *Universal GeneticSolutions, LLC, Orlando, FL, USA*
* Corresponding author, E-mail: Osman.Gutierrez@usda.gov; Dapeng.Zhang@usda.gov

Abstract

Cacao is one of the most economically important agricultural commodities in the world, providing the principal ingredient for the global chocolate industry. Accurate genotype identification is essential for effective conservation and utilization of cacao germplasm. Here, we report the screening of 956 candidate SNPs, pre-selected from the 6 and 15K Theobroma cacao SNP Arrays using targeted Genotyping-by-Sequencing on 451 cacao germplasm accessions, representing ten known genetic groups from the tropical Americas. Based on call rate (No call rate < 5%), Minor Allele Frequency (MAF > 0.15) and Linkage Disequilibrium (LD ≤ 0.5), a total of 219 SNPs were selected. The efficacy of these SNP markers for population classification was compared with the previous SSR-based analysis in cacao. The population assignment results of the retained 420 cacao accessions was highly comparable with the SSR study. The matrix of genetic distance between SSR and SNP markers is highly correlated ($r = 0.718$; $P < 0.001$). These results demonstrated the consistency in using the present SNP markers for cacao germplasm identification. This is our pilot project for the development of SNP markers reflecting population origin for cacao germplasm identification. These SNP markers and the selected reference germplasm for different populations are suitable for use in cacao germplasm management and crop improvement, including genotype identification, seed gardens and nursery accreditation, and cocoa authentication. Effort is being continued with the emphasis on selecting SNP markers for the detection of sub-population structures in the primary gene pool of *T. cacao*.

Citation: Gutiérrez OA, Martinez K, Zhang D, Livingstone DS, Turnbull CJ, et al. 2021. Selecting SNP markers reflecting population origin for cacao (*Theobroma cacao* L.) germplasm identification. *Beverage Plant Research* 1: 15 https://doi.org/10.48130/BPR-2021-0015

INTRODUCTION

Cacao (*Theobroma cacao* L.) is a perennial crop cultivated by small-holder farmers in the tropical regions of the world[1]. Cacao is a worldwide commodity of great importance as its fermented dried seeds are the principal ingredient for making chocolate by the confectionary and food industries, and it is also used by cosmetic and pharmaceutical corporations. The Maya and Aztec civilizations widely cultivated cacao in Mesoamerica; however, its center of origin as well as its center of domestication is the upper Amazon area of South America[2,3]. West Africa is currently the leader in cacao production worldwide (76.0%), followed by the Americas (17.7%) and Asia (6.1%)[4].

Cacao belongs to the Malvaceae family and is a diploid organism ($2n = 2x = 20$)[5] with a genome size ranging from 411 to 470 Mbp[6,7]. Earlier classifications of cacao germplasm were conducted based mainly on morphological characteristics and it was divided into Criollo, Forastero and Trinitario (Criollo × Forastero)[8,9]. However, the development of molecular markers has facilitated a more detailed estimation of the cacao genetic diversity, which an initial study classified into ten genetic groups[10], with additional genetic groups discovered subsequently[11,12].

Mislabeling of cacao accessions has been an ongoing problem across cacao collections worldwide, and by using molecular markers (SSRs and SNPs) and selecting reference genotypes, several collections have been screened and off-types have been identified[13–15]. The consequences of the presence of off-type plants at the farmer level are that incorrect plant material usually results in unexpected and subpar economic performance. At the breeder level, segregating populations developed with the wrong parents negatively impacts the advancement of the cacao breeding program[16].

The cacao genome was initially sequenced in 2011[6,7] and since that time, the number of cacao genomes that have been sequenced, as well as the availability of sequence information, has grown substantially[17,18] and contributed to the discovery of SNPs and usage. The use of SNPs has also increased due to a reduction in sequencing costs and easy automation that allows the fingerprinting of one DNA sample with at least 5K SNPs using different next generation sequencing platforms[19,20]. Currently, there are more than 30,000 cacao accessions across all cacao collections according to the International Cocoa Germplasm Database (ICGD)[21]. Molecular authentication of cacao genotypes, therefore, is of

critical importance for downstream research and development in the cocoa industry, including germplasm identification, verification of planting materials and authentication of cacao beans and cocoa products.

Single nucleotide polymorphism (SNP) markers have been increasingly used to assist cacao germplasm management, because they are amendable to high throughput systems, have a universal data comparability and lower genotyping cost[22,23]. Several attempts have been made to develop a core set of the most informative SNPs for the identification of off-types, parental and population identification, and determination of admixture levels of the different genetic groups in different cacao collections[24–30]. These small SNP panels, ranging from 48 to 192 SNPs, have been used to generate multi-locus profiles for individual cacao trees, based on the method of 'multi-locus matching', which was used to assess the genetic integrity of genotyped germplasm[23,31].

So far these SNP panels have not been evaluated for their efficacy in population and sub-population classification. This assessment is essential because inferring the population origin of a cacao germplasm can provide an additional dimension to support cacao germplasm identification. When cacao germplasm source or pedigree is unknown or the information is lost, SNP markers can help infer its probable origin and/or compared in better detail to other potential but under utilized germplasm. In many cases, a cacao germplasm accession (e.g., a breeding line) may not have a known reference standard. Therefore, the approach of 'multi-locus matching' cannot be used to ascertain whether this breeding line is mislabeled or not. In such circumstances, inferred parentage, or population origin, provided indirect evidence to assess the genetic integrity of this breeding line as previously reported[14,15]. In addition, population origin is important for cacao variety authentication, which is of considerable interest to the various stakeholders in the chocolate value chain. Production and marketing of differentiated (or specialty) high-value cocoa provides socioeconomic opportunities for cacao growers, the chocolate industry, and especially for consumers[25].

The original classification of cacao germplasm was based on SSR genotyping of 952 germplasm accessions, which led to the proposed classification of the primary gene pool into ten populations or genetic groups (Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Nacional, Nanay, and Purús)[10]. In this study we genotyped a fraction of the same DNA samples utilized in the initial experiment[10] using 956 SNP markers, which were pre-selected from the 6 and 15K Theobroma cacao SNP Array[19,20]. The objectives of this research were: a) to compare the SNP-based population structure and genetic distances with the ones obtained with SSRs in the previous study; b) to select a SNP genotyping panel that is effective in cacao population classification and individual genotype identification and c) to establish reference genotypes for each one of the ten populations for cacao authentication. The results obtained in this research would provide much needed cacao germplasm references as well as a suitable SNP marker panel for gene bank management, crop genetic improvement, seed garden verification and cocoa traceability and authentication.

MATERIALS AND METHODS

Plant materials

Leaf samples from 552 accessions (58%) out of the original 952 plant accessions previously used in identification of the ten genetic groups[10] were used in this research. The distribution of the samples based on genetic groups were as follows: Amelonado 65% (61/94), Contamana 51% (25/69), Criollo 3% (1/39), Curaray 75% (88/117), Guiana 63% (37/59), Iquitos 60% (70/117), Marañón 73% (104/143), Nacional 52% (27/52), Nanay 75% (114/152), and Purús 75% (83/110) (Supplemental Table S1). The cacao clones Matina 1-6, Criollo 22 and Pound 7 were used as controls.

DNA isolation and SNP genotyping

The DNeasy® Plant Mini kit (Qiagen Inc., Valencia, CA, USA) and the Fast Prep® Kit (MP Bio) were used to extract DNA from fresh and lyophilized cacao leaves. Leaf tissue was weighed; 100 mg for fresh and 20 mg for lyophilized tissue and added to the sample tube. A stainless-steel bead was added for the Qiagen method and garnet chips and a ceramic bead were added for the Fast Prep method. The tissue was ground in two 1-minute high-speed (30 Hz) shaking steps in a TissueLyser II (Qiagen) for the Qiagen method or the tissue was disrupted in a Fast Prep Homogenizer with three 30-second high-speed shaking steps for the Fast Prep method. After that, the protocol was followed from the manufacturers with the following modifications. 25 mg mL^{-1} polyvinylpolypyrrolidone was added to the DNeasy® kit buffer AP1 for the Qiagen method or to the CLS-VF for the Fast Prep® method. DNA was eluted from the silica column with two washes of 50 μL Buffer AE for the Qiagen method or with DES for the Fast Prep method, which were pooled, resulting in 100 μL DNA solution. DNA concentration was determined using a Qbit spectrophotometer (ThermoFisher) with absorbance at 260 nm.

A total of 956 Tcm SNP loci were used in this research (Supplemental Table S2). They were developed using Transcript-Based SNP identification, Genome-Based SNP Identification, the Matina 1-6 reference genome and Illumina Infinium SNP array technology[7,19,20]. Also, Tcm SNPs were selected based on their polymorphism between the cacao clones UF 273 Type 1 and Pound 7[32]. The distribution of SNPs across the cacao genome were as follows: Chromosome 1 (98), Chromosome 2 (97), Chromosome 3 (95), Chromosome 4 (99), Chromosome 5 (98), Chromosome 6 (98), Chromosome 7 (99), Chromosome 8 (76), Chromosome 9 (98), Chromosome 10 (98).

Libraries were prepared using the ThermoFisher AgriSeq technology. The DNA was normalized to 3.3 ng/μL for a total of 10 ng of DNA per 10 μL reaction before adding the Ion AgriSeq primer panel and the AgriSeq amplification master mix. The DNA targeted amplification was achieved with the following thermocycler profile: 99 °C for 2 min, then 15 cycles of 99 °C for 15 s and 60 °C for 4 min. The amplicons were then prepared for barcode addition using a pre-ligation enzyme reaction mix and the following program on a thermocycler: 50 °C for 10 min, 55 °C for 10 min, and 60 °C for 20 min. The IonCodeTM Barcode Adapters were ligated to the amplicons with the final thermocycler step: 22 °C for 30min; 72 °C for 10

min. The libraries were then purified with Agencourt® AMPure® XP magnetic beads and the DynaMag-96 side magnet rack (ThermoFisher). Libraries were pooled to create a final combined library stock by adding 3 μL of each of the prepared libraries into a single tube. Three hundred and eight-four cacao samples were represented in each tube. This tube was then put on the Ion Chef™ (ThermoFisher) which loaded them onto the Ion 540 chip. This chip was loaded onto the Ion GeneStudio S5 Plus™ (ThermoFisher) for sequencing. Amplicon sequences were aligned and scored with the Torrent Variant Caller plugin to determine the genotype call for each marker and sample.

Data analysis

Initial SNP analysis was performed using the Ion Torrent™ AgriSum Tool Kit (AST) plug-in that provided information about the coverage, call rate range of the samples and overall sequencing run metrics. Then, raw SNP loci data was exported to Microsoft Excel (Microsoft 365 applications), and samples that had more than 30% of missing data and SNPs loci with more than 10% of missing data were discarded. The final data used for further analysis was 420 DNA samples (accessions) and 865 Tcm SNP loci. GenAllex 6.5[33,34] was used to calculate allele frequencies for each locus of the study population, the number of alleles per locus (A), observed heterozygosity (HO), expected heterozygosity, (HE), polymorphic information content (PIC) as well as to perform pairwise genetic distance, Mantel test (SSRs vs SNPs) and SNP & Variation Suite 8.9.0[35] software was used to perform a linkage disequilibrium pruning analysis.

Population genetic structure and admixture levels were estimated utilizing the model-based Bayesian clustering methodology of Structure v2.3.4[36–38]. Since the genetic groups were previously determined by SSRs only, Criollo 22 and Matina 1-6 were additionally included as references in the analysis. The data were subjected to an admixture model. Ten independent runs were assessed for each fixed number of clusters (K value) ranging from 1 to 15, each consisting of a burn-in of 100,000 iterations and 200,000 Markov chain Monte Carlo repetitions. Results were analyzed using STRUCTURE SELECTOR[39] to identify the most likely number of clusters present based on the method of Evanno et al.[40] and Puechmaille[41]. The programs CLUMPP 1.1.2[42] and DISTRUCT 1.1[43] were used to visualize the results. Based on the result of population stratification, individuals with high assignment coefficient (Q > 0.75) were retained. Pairwise Fst, Analysis of Molecular Variance (AMOVA) and Principal Coordinate Analysis (PCoA) were performed on these populations with retained samples, using GenAllex 6.5[13,44].

Genetic relationship among the nine genetic groups was further examined using clustering analysis. Pairwise distances among populations were calculated using the Nei's[44] standard genetic distance as implemented in the program Microsatellite Analyser (MSA)[45] with 1000 boot strapping. The resulting distance matrix was then used to generate a dendrogram using the UPGMA (unweighted pair group method with arithmetic mean) algorithm[46] available in the program PHYLIP[47]. Thereafter, the dendrogram was visualized using FigTree program version 1.4.2[48].

To assess the efficacy of the selected SNP panel for population classification and individual identification, the result of STRUCTURE analysis and genetic distances generated in the present study was compared with the previous SSR-based result. The consistence of population assignment between the two marker systems was measured by Pearson's correlation. The consistency of the SNP and SSR-based distance matrix was measured using Mantel's Test, as implemented in GenAllex 6.5[33,34].

RESULTS

Data QC and data filtration

Raw data for the SNP loci and sample calls were organized in Microsoft Excel, (Microsoft 365 applications). Quality control was performed using the Quality Assurance Module from SNP Variation Suite version 8.9.0[35]. Any SNP having more than a 5% no-call rate was removed from the data set. SNPs that were in linkage disequilibrium (LD) with each other at $r^2 > 0.5$ were also removed, resulting in a data set consisting of 219 Tcm SNPs for further analysis. These loci were randomly distributed across the cacao genome and their chromosomal locations are as follows: Chromosome 1 (23), Chromosome 2 (21), Chromosome 3 (15), Chromosome 4 (27), Chromosome 5 (21), Chromosome 6 (26), Chromosome 7 (24), Chromosome 8 (18), Chromosome 9 (15), and Chromosome 10 (29) (Supplemental Table S2).

Descriptive statistics

Four hundred and twenty DNA samples and three controls produced amplification (Supplemental Table S1). Summary statistics were computed based on the 420 samples and 219 selected Tcm SNP markers and the results are presented in Supplemental Table S2. The mean value for Shannon's information index was 0.617, ranging from 0.398 to 0.693. The mean observed heterozygosity (H_{Obs}) was 0.247, ranging from 0.118 to 0.400. The mean genediversity (expected heterozygosity) was 0.428, ranging from 0.235 to 0.500. The mean fixation index (F_{IS}) was 0.419, ranging from 0.149 to 0.668. The mean minor allele frequency was 0.359, ranging from 0.150 to 0.500 (Supplemental Table S3). Mantel test showed a highly significant correlation ($r = 0.718$; $P < 0.001$) between these 219 SNPs and the 91 SSR markers reported in a previous study[10] (Fig. 1).

Inference of population structure

From the STRUCTURE analysis, the most probable number of genetically distinct groups (K) was two (Fig. 2a) based on Evanno's Delta K value[40]. However, when the result of STRUCTURE was analyzed using the method of Puechmaille[41], as implemented in STRUCTURE SELECTOR[39], all the supervised estimators (Medmedk, Medmeank, Maxmedk and Maxmeank) suggested the optimum K of nine populations (Fig. 2b).

At K = 9, seven out of the ten populations had consistent assignment results as the SSR-based study reported previously[10]. These populations include: Amelonado, Curaray, Guiana, Marañon, Nanay, and Purús (Fig. 2c; Supplemental Table S4). However, the population Nacional and Contamana were grouped together. Moreover, discrepancy was found within the Iquitos population, the germplasm from Iquitos, Peru and those from Rio Salimoes, Brazil was separated into two distinct groups. This represents 76% of the 420 samples used (Supplemental Table S1) and constitutes 44% of the

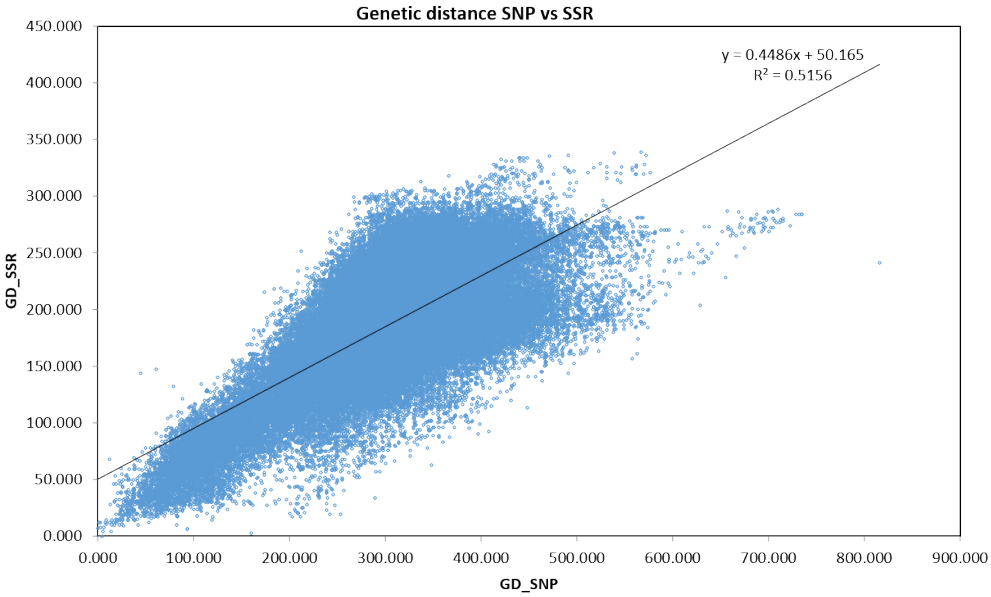

Fig. 1 Mantel test results indicating significant correlation (r = 0.718; $P < 0.001$) between SNPs and SSR markers.

samples used in the initial classification of the ten genetic groups[10]. Their distribution based on the genetics groups is as follows: Amelonado 46% (43/94), Contamana 51% (18/69), Criollo 3% (1/39), Curaray 56% (66/117), Guiana 47% (28/59), Iquitos 34% (40/117), Marañón 45% (65/143), Nacional 21% (11/52), Nanay 64% (98/152), and Purús 45% (50/110). The highest DNA amplification was obtained in samples from the Nanay group and the lowest in the Criollo group. Due to this reason, the Criollo sample was not included in the PCA and Structure Analysis. The samples with Q-value ≥ 0.75 were selected as reference clones for each of the corresponding populations (Supplemental Table S3).

Relationship among different populations

Principal coordinate analysis based on the results of the STRUCTURE analysis is presented in Fig. 3a and 3b, which provides a complementary illustration of the relationship among the nine genetic groups. The plane of the first three main axes accounted for 23.1%, 7.6%, and 3.8% of total variation, respectively. The distinctiveness of the nine clusters was clearly revealed. The results of the analysis of molecular variance (AMOVA) provide additional evidence supporting the significant population differentiation (Table 1). The within-population molecular variance accounted for 47.0%, whereas among populations, molecular variance was 53.0%. The inter population differentiation was highly significant as shown by Phi-statistics[49] ($P < 0.001$) (Table 2). The Fst value ranged from 0.038 (Nacional vs Contamana) to 0.194 (Amelonado vs Nanay), with an average of 0.109 among all the populations (Table 3).

The UPGMA tree (Fig. 4) provided complementary information regarding the inter-population relationships. The cluster pattern is largely consistent with the previous SSR-based result[10]. Same as the result of STRUCTURE stratification, Population Nacional and Contamana were grouped together, which is also compatible with the results of PCoA. In addition, population Guiana and Amelonado were grouped together in the UPGMA tree, whereas population Iquitos and Nanay fell in the same main group. Purús I and Purús II were grouped together. All the branches were supported by the bootstrapping value above 50%, ranging from 516 to 1000 in the consensus tree (Fig. 4).

DISCUSSION

SNP genotyping using the Thermo Fisher Agri Seq technology

Despite great progress in genomics research on cacao, availability of cost-effective molecular tools to support routine germplasm management has been scarce. Developing SNP markers using available sequences could fill the gap between genomic research and downstream applications by cacao breeders and germplasm collection curators. In the present study, we genotyped 956 Tcm SNPs selected from the previously published arrays[7,19,20] and used them to genotype a diverse panel of 451 cacao accessions. These cacao accessions are all wild and were used in a previously reported SSR analysis of genetic diversity[10] in wild cacao populations, based on which the classification of cacao germplasm into ten populations (or genetic groups) were proposed. The repeated genotyping on the same genetic materials using SNP markers enabled direct comparison between the results obtained by both marker systems. It also allowed us to identify cacao germplasm that can serve as a reference standard in population stratification.

We obtained a high success rate (> 95%) for marker validation, which demonstrated that using the ThermoFisherAgri-Seq technology targeted sequencing is an effective method for cacao genotyping. This technology is a targeted Genotype By Sequencing (GBS) that utilizes a multiplexed PCR chemistry where large numbers of markers can be targeted and uniformly amplified in a single reaction. The genotyping result showed that it is a suitable technology for large scale genotyping, which can serve as a complementary approach to the currently used methods (e.g., KASP, TaqMan-based

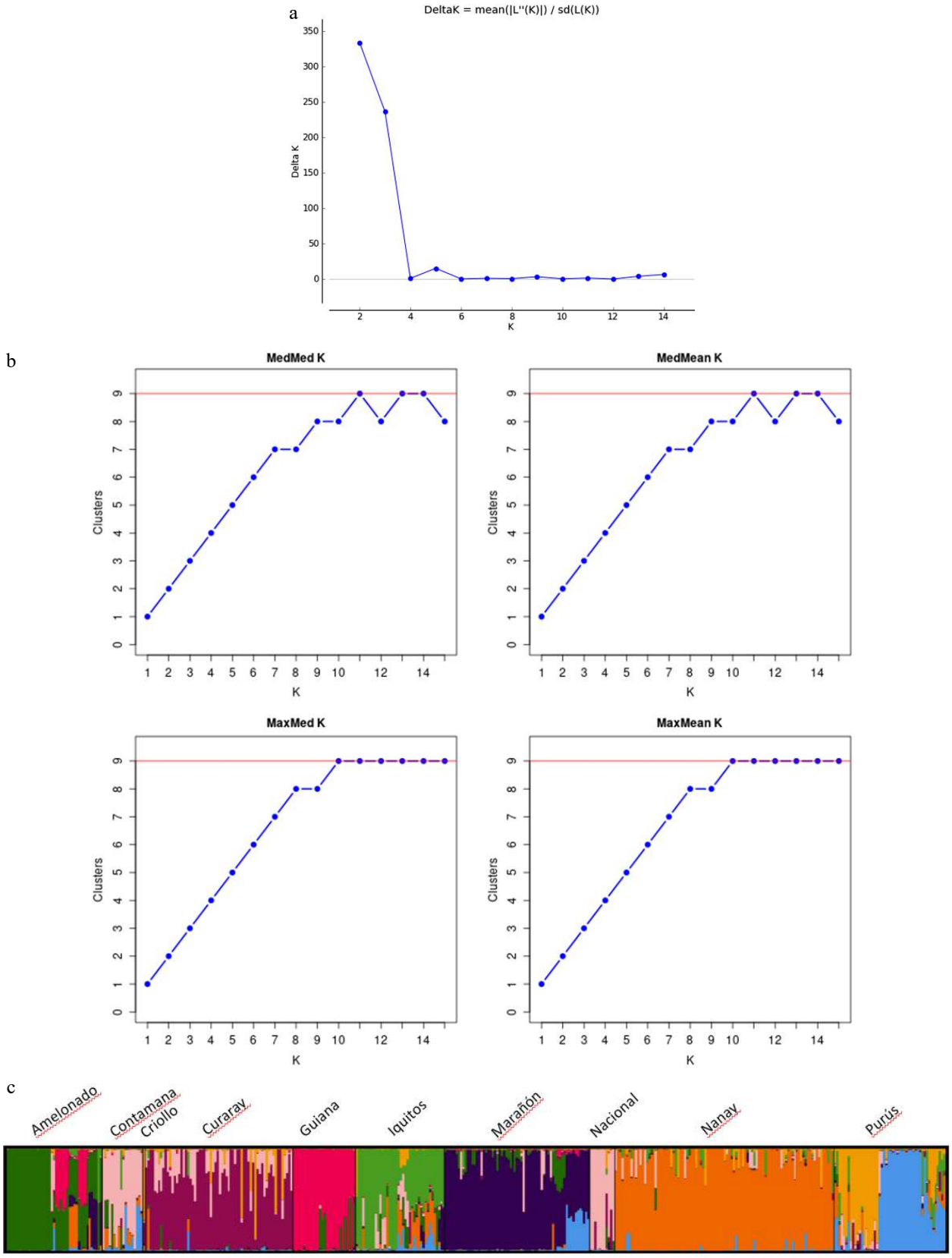

Fig. 2 (a) Number of clusters based on the Evanno's Delta K value[10]. (b) Inferred clusters obtained using the method of Puechmaille. (c) Population structure of the 420 cacao accessions (*Theobroma cacao* L.) germplasm collections containing representative genotypes of the nine cacao genetic groups obtained using Structure v2.3.3. Black vertical lines indicate the separation of the genetic groups. Multiple colors within the genetic group imply admixed individuals.

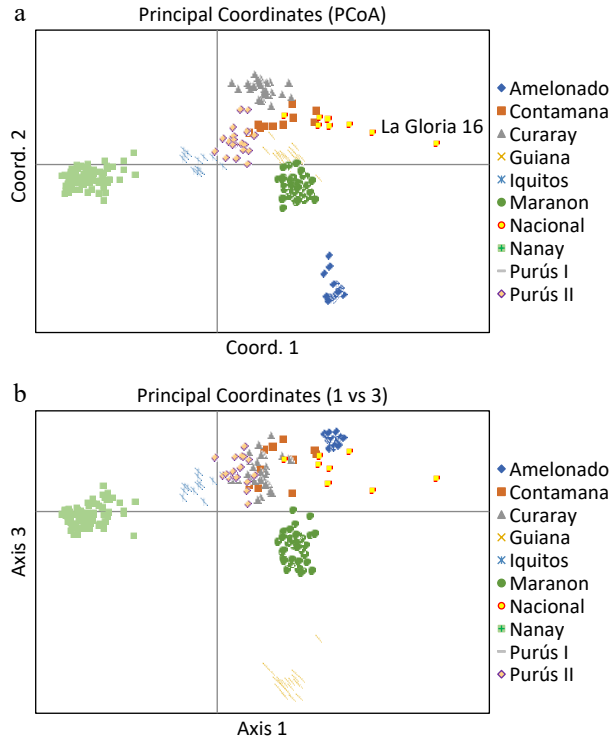

Fig. 3 Principal Coordinates Analysis plots of 420 cacao accessions belonging to nine genetic groups. The plane of the first three main axes accounted for: first axis = 23.1%, the second = 7.6% and the third = 3.8% of the total variation.

quantitative PCR, DArT markers and Maldi-TOF mass spectrometry (MS) for cacao germplasm identification.

Population structure and inter-population relationships

The delta K calculated by Evanno's method[40] indicated K = 2 was the most likely genetic clusters in the 420 samples retained in data analysis. This discrepancy to the known genetic groups could be explained by the uneven sampling of the ten known populations, where some populations were not proportionally represented by enough samples, thus, were not classified as an independent genetic cluster by STRUCTURE[50]. Distinct subpopulations with reduced sampling tended to be merged, while at the same time, individuals from extensively sampled subpopulations were generally split, despite belonging to the same panmictic population[41]. Moreover, because the Delta K method detects the uppermost hierarchical level of genetic structure, this can also lead to underestimating the number of genetic clusters in this collection.

To correct the uneven sample size from different populations, we used the method of Puechmaille[41], as implemented in the program STRUCTURE SELECTOR. The four new supervised methods, 'MedMeaK' (median of means), 'MaxMeaK' (maximum of means), 'MedMedK' (median of medians) and 'MaxMedK' (maximum of medians) were applied when individual samples can be grouped based on prior knowledge (e.g., sampling location/region). A subpopulation was considered as belonging to a cluster if its arithmetic mean (for MedMeaK and MaxMeaK) or its median (for MedMedK and MaxMedK) membership coefficient to that cluster was greater than a threshold value (set to 0.5), thus ensuring that a subpopulation cannot belong to more than one cluster[41].

Using Puechmaille's method[41], a result of nine genetic clusters were obtained (with the threshold value set to 0.5), which differentiated population of Amelonado, Curaray, Guiana, Iquitos, Marañon, Nanay, and Purús. However, the populations of Nacional and Contamana were grouped together. This lack of differentiation was likely due to the sampling bias in the National population. The National population is native to the rainforest of Southern Ecuador. Still, the core member of this population was represented by the landraces from the cocoa producing regions in the Pacific coast, including La Gloria and Las Brisas. In the present study, there was only one sample ('La Gloria 16') which was included in the National population. As shown in the PCoA (Fig. 3a), La Gloria 16 was distanced away from the rest of the samples in the Nacional population which shared higher similarity with

Table 1. Analysis of molecular variance (AMOVA) for the nine cacao genetics groups.

Source of variation	d.f.	Sum of squares	Means squares	Est. Var	%
Among populations	9	12,047.62	1,338.62	53.84	53
Within populations	215	11,874.38	47.69	47.69	47
Total	435	23,922.00	16.57380	101.53	100

Table 2. Pairwise Population PhiPT Values among nine cacao germplasm groups.

Populations	Amelonado	Contamana	Curaray	Guiana	Iquitos	Marañon	Nacional	Nanay	Purús I	Purús II
Amelonado	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Contamana	0.633	0.000	0.001	0.001	0.001	0.001	0.011	0.001	0.001	0.001
Curaray	0.645	0.407	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Guiana	0.702	0.561	0.540	0.000	0.001	0.001	0.001	0.001	0.001	0.001
Iquitos	0.695	0.462	0.500	0.634	0.000	0.001	0.001	0.001	0.001	0.001
Maranon	0.539	0.406	0.443	0.477	0.444	0.000	0.001	0.001	0.001	0.001
Nacional	0.622	0.095	0.436	0.561	0.499	0.395	0.000	0.001	0.001	0.001
Nanay	0.712	0.593	0.558	0.644	0.489	0.547	0.673	0.000	0.001	0.001
Purús I	0.637	0.319	0.373	0.568	0.437	0.404	0.317	0.555	0.000	0.001
Purús II	0.577	0.300	0.382	0.508	0.360	0.367	0.336	0.483	0.268	0.000

Note: PhiPT Values below diagonal. Probability, P (r and ≥ data) based on 999 permutations is shown above the diagonal.

Table 3. Pairwise Population Fst Values based on the result of population stratification. Within each population, samples with the assignment coefficient > 0.75 were retained for analysis.

	Amelonado	Contamana	Curaray	Guiana	Iquitos	Marañon	Nacional	Nanay	Purús I	Purús II
Amelonado	0.000									
Contamana	0.181	0.000								
Curaray	0.179	0.092	0.000							
Guiana	0.180	0.143	0.124	0.000						
Iquitos	0.180	0.117	0.113	0.150	0.000					
Maranon	0.129	0.094	0.095	0.104	0.098	0.000				
Nacional	0.166	0.038	0.101	0.138	0.129	0.091	0.000			
Nanay	0.194	0.135	0.121	0.150	0.087	0.115	0.176	0.000		
Purús I	0.168	0.085	0.078	0.133	0.100	0.087	0.083	0.114	0.000	
Purús II	0.157	0.083	0.086	0.125	0.082	0.077	0.094	0.092	0.065	0.000

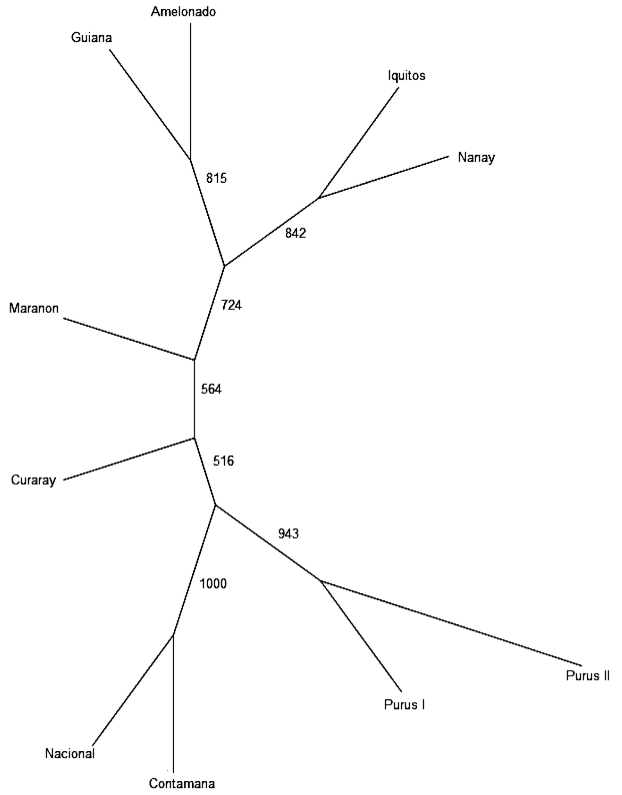

Fig. 4 UPGMA tree indicating the relationships among the genetic groups.

the Contamana population. The biased sampling led to a small estimation of Fst (0.038) between Nacional and Contamana in the present study.

At K = 9, the STRUCTURE analysis also split the Purús population into two clusters (Fig. 2c; Supplemental Table S3). The first cluster included most of the wild cacao collected from Napo river in the Ecuadorian Amazon. These samples were classified as members of Purús population, but they all had a low assignment coefficient (Q = 0.39–0.61) in the previous SSR-based analysis. The second Purús cluster comprised exclusively of the samples from Purús river, Brazil. Collecting sites of the two cacao groups are more than 1,000 km apart. Fst between the two groups is 0.065, showing that the two groups are substantially differentiated. Therefore, the separation of these two clusters is well justified and they probably can be considered as different populations.

The population relationship revealed by the UPGMA dendrogram (Fig. 4), together with the result of STRUCTURE stratification (Fig. 2c) and PCoA plot (Fig. 3a & b)), is compatible with the previous SSR-based result. These results demonstrated that these selected SNPs can be used to assess population origin for cacao germplasm. Such information is highly useful for cacao germplasm identification because it can also be used to infer ancestry/parentage/pedigree for cacao germplasm that may not have known identity or passport data. This approach can play a complementary role to the currently used method for cacao germplasm identification, which directly compare the candidate tree with the known cacao accession, based on the reference SNP profiles. Whereas this approach is highly effective for cacao germplasm identification, it lacks capacity to deal with germplasm that do not have reference standard. SNP marker-based information on population origin, ancestry, parentage, and pedigree, therefore, will be appreciated by cacao breeders, genebank curators and cacao research community in general.

Nonetheless, additional effort remains needed to assess the efficacy of these markers regarding the differentiation power at the level of subpopulations. More germplasm from each wild population, with reliable GPS data, need to be analyzed. Especially, in the SSR-based study, a significant fraction of the individual accessions did not have a high value of assignment coefficient (Q-value). A total of 217 samples (of the 952; 22.8%) had a Q value below 0.70 and 56 samples (5.6%) had a Q value below 0.50[10]. This low assignment coefficients indicates that structure of sub-populations in these wild populations needs further investigation. SNP markers that can efficiently detect these variations among sub-populations need to be selected and used for cacao germplasm identification. Now the major river systems in Peru have been sampled for wild cacao populations[51,52]. Further analysis that includes all the wild populations in these regions will likely provide more insight about the structure of subpopulations in the center of origin of this species.

CONCLUSIONS

Various SNP genotyping sets have been used for cacao germplasm identification. However, these panels have not been systematically evaluated for optimum genotyping efficiency, as well as for population and sub-population classification. The ideal genotyping panel should comprise a minimum number of SNP markers but have a maximum

discriminating power. Moreover, the capacity to infer the population origin of a given cacao accession is essential to support cacao germplasm identification when the reference SNP profile is not available. For an efficient germplasm identification, Linkage Disequilibrium is one of the critical factors because each SNP marker is expected to be independently informative. In the present study, we evaluated 956 SNPs on 451 wild cacao samples with known population origin. Based on the criteria of LD ≤ 0.5, call rate > 95% and Minor Allele Frequency (MAF > 0.15), we selected a total of 219 SNPs. Population stratification demonstrated their efficacy in high compatibility with previously reported SSR markers. Mantel Test of distance matrix between SSR and SNP markers showed a high correlation (r = 0.718; $P < 0.001$). In addition, the present study generated complementary insight regarding the classification of wild cacao populations and subpopulations in the Amazon region. These newly selected SNPs can also be combined with the previously identified SNP markers, e.g., the TcSNPs that have been commonly used in cacao germplasm identification, to form different genotyping panels. The generated SNP profiles can be converted into a simple bar code and be used in many other downstream applications, such as nursery accreditation, clone registration and the authentication of geographically referenced cocoa beans. This is our pilot project for the development of SNP markers reflecting population origin for cacao (*Theobroma cacao* L.) germplasm identification. Marker evaluation is being continued with the emphasis on selecting SNP markers to detect sub-population structures in the primary gene pool of *T. cacao*.

ACKNOWLEDGMENTS

The authors are grateful to Mr. Wilber Quintanilla and Ms. Ashley Johnson at USDA-ARS for their excellent technical assistance. This work has been supported by the USDA-ARS-SHRS Development and Application of Genomic-assisted Breeding Strategies to Produce Disease-resistant Cacao Genetic Resources (Project No. 6038-21000-025-000-D). Also funding for this projectwas provided by MARS, Inc.; Trust Agreement No. 6038-21000-025-12-T: Genomic Enhancement of Theobroma cacao.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Information accompanies this paper at (http://www.maxapress.com/article/doi/10.48130/BPR-2021-0015)

Dates

Received 30 September 2021; Accepted 26 November 2021; Published online 27 December 2021

REFERENCES

1. Cope FW. 1984. Cacao Theobroma cacao (Sterculiaceae). In *Evolution of Crop Plants*, ed. SimmondsNW. London, UK: Longman. pp. 285–89
2. Bartley BGD. 2005. *The genetic diversity of cacao and its utilization*. Wallingford: CABI https://doi.org/10.1079/9780851996196.0000
3. Zarrillo S, Gaikwad N, Lanaud C, Powis T, Viot C, et al. 2018. The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. *Nature Ecology & Evolution* 2:1879–88
4. ICCO. 2021. *ICCO Quarterly Bulletin of Cocoa Statistics, Vol. XLVII, No. 2, Cocoa year 2020/21*. https://www.icco.org/statistics/#tab-id-1
5. Whitlock BA, Baum DA. 1999. Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene *Vicilin*. *Systematic Botany* 24:128
6. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, et al. 2011. The genome of *Theobroma cacao*. *Nature Genetics* 43:101–8
7. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, et al. 2013. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology* 14:r53
8. Cheesman EE. 1932. *The economic botany of cacao*. A critical survey of the literature to the end of 1930. Tropical. Agriculture. 9:16 pp
9. Cuatrecasas J. 1964. Cacao and its allies. A taxonomic revision of the genus Theobroma. *Contributions from the United States National Herbarium* 35:379–605
10. Motamayor JC, Lachenaud P, da Silva E Mota JW, Loor R, Kuhn DN, et al. 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). *PLoS ONE* 3:e3311
11. Motamayor JC, Lachenaud P, Da Silva E Mota JW, Loor RG, Martinez WJ, et al. 2010. No mas forastero: a new protocol for meaningful cacao germplasm classification. *Proc. 16th International Cocoa Research Conference, Bali, Indonesia*. 2010: 179–85. Indonesia: Cocoa Producers' Alliance
12. Zhang D, Martínez WJ, Johnson ES, Somarriba E, Phillips-Mora W, et al. 2012. Genetic diversity and spatial structure in a new distinct *Theobroma cacao* L. population in Bolivia. *Genetic Resources and Crop Evolution* 59:239–52
13. Motilal L, Butler D. 2003. Verification of identities in global cacao germplasm collections. *Genetic Resources and Crop Evolution* 50:799–807
14. Olasupo FO, Adewale DB, Aikpokpodion PO, Muyiwa AA, Bhattacharjee R, et al. 2018. Genetic identity and diversity of Nigerian cacao genebank collections verified by single nucleotide polymorphisms (SNPs): a guide to field genebank management and utilization. *Tree Genetics & Genomes* 14:32
15. Padi FK, Ofori A, Takrama J, Djan E, Opoku SY, et al. 2015. The impact of SNP fingerprinting and parentage analysis on the effectiveness of variety recommendations in cacao. *Tree Genetics & Genomes* 11:44
16. DuVal A, Gezan SA, Mustiga G, Stack C, Marelli JP, et al. 2017. Genetic parameters and the impact of off-types for *Theobroma cacao* L. in a breeding program in Brazil. *Frontiers in Plant Science* 8:2059
17. Cornejo OE, Yee MC, Dominguez V, Andrews M, Sockell A, et al. 2018. Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Communications Biology* 1:167
18. Hämälä T, Wafula EK, Guiltinan MJ, Ralph PE, dePamphilis CW, et al. 2021. Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *PNAS* 118:e2102914118
19. Livingstone D, Royaert S, Stack C, Mockaitis K, May G, et al. 2015. Making a chocolate chip: development and evaluation of a 6K SNP array for *Theobroma cacao*. *DNA Research* 22:279–91
20. Livingstone D, Stack C, Mustiga GM, Rodezno DC, Suarez C, et al. 2017. A larger chocolate chip – development of a 15K *Theobroma cacao* L. SNP array to create high-density linkage maps. *Frontiers in Plant Science* 8:2008

21. Turnbull CJ, Hadley P. 2021. International Cocoa Germplasm Database (ICGD). CRA Ltd./ICE Futures Europe/University of Reading, UK.
22. Livingstone DS, Motamayor JC, Schnell RJ, Cariaga K, Freeman B, et al. 2011. Development of single nucleotide polymorphism markers in *Theobroma cacao* and comparison to simple sequence repeat markers for genotyping of Cameroon clones. *Molecular Breeding* 27:93–106
23. Takrama J, Kun J, Meinhardt L, Mischke S, Opoku SY, et al. 2014. Verification of genetic identity of introduced cacao germplasm in Ghana using single nucleotide polymorphism (SNP) markers. *African Journal of Biotechnology* 13:2127–36
24. Dadzie AM, Livingstone DS, Opoku SY, Takrama J, Padi F, et al. 2013. Conversion of microsatellite markers to single nucleotide polymorphism (SNP) markers for genetic fingerprinting of *Theobroma cacao* L. *Journal of Crop Improvement* 27:215–41
25. Fang W, Meinhardt LW, Mischke S, Bellato CM, Motilal L, et al. 2014. Accurate determination of genetic identity for a single cacao bean, using molecular markers with a nanofluidic system, ensures cocoa authentication. *Journal of Agricultural and Food Chemistry* 62:481–87
26. Ji K, Zhang D, Motilal LA, Boccara M, Lachenaud P, et al. 2013. Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. *Genetic Resources and Crop Evolution* 60:441–53
27. Li Y, Zhang D, Motilal LA, Lachenaud P, Mischke S, et al. 2021. Traditional varieties of cacao (*Theobroma cacao*) in Madagascar: their origin and dispersal revealed by SNP markers. *Beverage Plant Research* 1:4
28. Mahabir A, Motilal LA, Gopaulchan D, Ramkissoon S, Sankar A, et al. 2019. Development of a core SNP panel for cacao (*Theobroma cacao* L.) identity analysis. *Genome* 63:103–14
29. Takrama J, Dadzie AM, Opoku SY, Padi FK, Adomako B, et al. 2012. Applying SNP marker technology in the cacao breeding programme in Ghana. *African Crop Science Journal* 20:67–75
30. Wang B, Motilal LA, Meinhardt LW, Yin J, Zhang D. 2020. Molecular characterization of a cacao germplasm collection maintained in Yunnan, China using single nucleotide polymorphism (SNP) markers. *Tropical Plant Biology* 13:359–70
31. Mata-Quirós A, Arciniegas-Leal A, Phillips-Mora W, Meinhardt L, Zhang D. 2017. Understanding the genetic structure and parentage of the clonal series of cacao UF, CC, PMCT and ARF preserved in the International Cacao Collection at CATIE (IC3). *Proc. International Symposium on Cocoa Research* (*ISCR*), Lima, Peru. pp. 13–17
32. Gutiérrez OA, Puig AS, Phillips-Mora W, Bailey BA, Ali SS, et al. 2021. SNP markers associated with resistance to frosty pod and black pod rot diseases in an F1 population of *Theobroma cacao* L. *Tree Genetics & Genomes* 17:28
33. Peakall R, Smouse PE. 2006. GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288–95
34. Peakall R, Smouse PE. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics* 28:2537–39
35. Golden Helix, Inc. 2021. SNP & Variation Suite™. Bozeman, MT, USA.
36. Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–87
37. Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–59
38. Pritchard JK, Wen X, Falush D. 2010. Documentation for structure software: Version 2. 3
39. Li Y, Liu J. 2018. StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. *Molecular Ecology Resources* 18:176–77
40. Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* 14:2611–20
41. Puechmaille SJ. 2016. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Molecular Ecology Resources* 16:608–27
42. Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–6
43. Rosenberg NA. 2004. DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* 4:137–38
44. Nei M. 1972. Genetic distance between populations. *The American Naturalist* 106:283–92
45. Dieringer D, Schlötterer C. 2003. Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes* 3:167–69
46. Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38:1409–38
47. Felsenstein J. 1989. PHYLIP - phylogeny inference package, (version 3.2). *Cladistics* 5:164–66
48. Rambaut A. 2014. FigTree-v1.4.2. A graphical viewer of phylogenetic trees. http://tree.bio.ed.ac.uk/software/figtree2014
49. Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479–91
50. Kalinowski ST. 2011. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity* 106:625–32
51. Arevalo-Gardini E, Meinhardt LW, Zuñiga LC, Arévalo-Gardni J, Motilal L, et al. 2019. Genetic identity and origin of "Piura Porcelana" – a fine-flavored traditional variety of cacao (*Theoborma cacao*) from the Peruvian Amazon. *Tree Genetics & Genomes* 15:11
52. Zhang D, Motilal L. 2016. Origin, Dispersal, and Current Global Distribution of Cacao Genetic Diversity. In *Cacao Diseases: A History of Old Enemies and New Encounters*, ed. Bailey BA, Meinhardt LW. Switzerland: Springer, Cham. pp. 3–31 https://doi.org/10.1007/978-3-319-24789-2_1