NRC Research Press

ARTICLE

Development of a core SNP panel for cacao (*Theobroma cacao* L.) identity analysis

Amrita Mahabir, Lambert A. Motilal, David Gopaulchan, Saila Ramkissoon, Antoinette Sankar, and Pathmanathan Umaharan

Abstract: Single nucleotide polymorphisms (SNPs) are preferred markers for DNA fingerprinting and diversity studies in cacao (*Theobroma cacao* L.). Yet, a consensus SNP panel with a minimum number of SNPs for optimal identity analysis is unavailable for cacao. An initial set of 146 SNP panels of varying sizes were assembled based on heterozygosity, linkage disequilibrium (LD), linkage group (LG) distribution, major allele frequency, minor allele frequency (M_iAF), polymorphism information content (PIC), and random distribution. These panels were assessed to determine their ability to distinguish among a training set of 155 accessions. The panels with the best separation ability were supplemented with additional SNPs to create 16 designer panels, which separated all 155 accessions. The 16 designer SNP panels were then assessed on a dataset of 1220 accessions coming from 10 ancestral groups. Increasing the number of SNPs generally yielded improved resolution of genetic identities with concomitant reduction of synonymous groups. The number and choice of SNPs were critical factors with LD, M_iAF, and PIC being important selection attributes but an even LG distribution was unnecessary. A robust set of 96 SNPs is recommended as a minimal core SNP panel for cacao DNA fingerprinting to the international cacao community.

Key words: DNA fingerprinting, SNP panel, identity analysis, *Theobroma cacao* accessions, genetic diversity.

Résumé : Les polymorphismes mononucléotidiques (SNP) sont les marqueurs de choix pour établir des empreintes génétiques et réaliser des études de diversité chez le cacaoyer (*Theobroma cacao* L.). Et pourtant, il n'existe aucun jeu de SNP comptant le nombre minimal de SNP requis pour une identification optimale. Initialement, 146 jeux de SNP de tailles diverses ont été assemblés sur la base de l'hétérozygotie, du déséquilibre de liaison (LD), de leur distribution sur les groupes de liaison (LG), de la fréquence de l'allèle majeur, la fréquence de l'allèle mineur (M_iAF), le contenu en information (PIC) et leur distribution. Ces jeux ont été examinés pour évaluer leur capacité à distinguer les individus au sein d'une collection de 155 accessions. Les jeux offrant la meilleure capacité de séparation ont été bonifiés par l'ajout de SNP additionnels pour générer 16 jeux faits sur mesure, lesquels permettaient tous de distinguer les 155 accessions. Ces 16 jeux sur mesure ont ensuite été évalués sur une collection de 1220 accessions appartenant aux 10 groupes ancestraux. L'accroissement du nombre de SNP augmentait généralement la résolution des identités génétiques et permettait de réduire le nombre de groupes synonymes. Le nombre et le choix des SNP étaient des facteurs critiques. Le LD, le M_iAF et le PIC étaient également importants, tandis que leur distribution sur les LG n'était pas nécessaire. Un jeu robuste de 96 SNP est recommandé comme étant un jeu minimal essentiel pour la caractérisation génétique du cacaoyer et pour la communauté internationale travaillant sur le cacaoyer. [Traduit par la Rédaction]

Mots-clés : empreintes génétiques, jeu de SNP, analyse d'identité, accessions du *Theobroma cacao*, diversité génétique.

Introduction

Theobroma cacao L. ($2n = 2x = 20$) belongs to the family Malvaceae (Alverson et al. 1999; Bayer et al. 1999). The centre of origin and diversity is in Amazonian South America (Cuatrecasas 1964; Motamayor et al. 2008) with the greatest diversity and earliest use occurring in the upper Amazon region of northwest South America (Zarrillo et al. 2018). Cacao is a commercially important industrial tree crop within the top 10 global agricultural commodities (Utro et al. 2012). The fermented and dried cotyledons of the seeds are raw ingredients in the multi-billion dollar confectionery industry. Cacao is an important cash crop in over 50 countries, mainly on small-holder farms, particularly in West Africa where over 70% of the world's cacao

Received 11 April 2019. Accepted 19 October 2019.

A. Mahabir, L.A. Motilal, D. Gopaulchan, S. Ramkissoon, A. Sankar, and P. Umaharan. Cocoa Research Centre, Sir Frank Stockdale Bldg., The University of the West Indies, St. Augustine, 330912, Trinidad, Trinidad and Tobago.

Corresponding author: Lambert A. Motilal (email: lamotilal@yahoo.com).

Published at www.nrcresearchpress.com/gen on 4 November 2019.

is produced (ICCO 2017). Genetic resources exist in over 60 germplasm collections in various countries (Motilal 2018), in farmers' fields, and endemically in Amazonian South America (Zhang and Motilal 2016).

Cacao has been traditionally classified into three agro-morphological groups: Criollo, Forastero, and Trinitario (Cheesman 1944; Cuatrecasas 1964; Toxopeus 1985). Diversity in cacao classically relied on morphological traits primarily based on flower, fruit, and seed descriptors (Engels et al. 1980; Bekele and Butler 2000) until the development of molecular markers. A variety of molecular markers have been employed to fingerprint and assess genetic diversity in cacao (Motilal et al. 2017 and references therein). Molecular marker information from microsatellites has been used to sort cacao into 10 ancestral groups (Motamayor et al. 2008). New collections from the wild in Bolivia enabled the identification of an additional population (Zhang et al. 2012) and a subsequent reclassification into 13 genetic clusters (Motamayor et al. 2010). The ancestral groups are distributed across a variety of accession groups. Accession groups are named according to the collection expedition (Turnbull and Hadley 2019) with the result that some accession groups contain individuals belonging to more than one ancestral group. For instance, the accession PA 120 [PER] belongs to the Parinari accession group, which fits into the Marañon population group of Motamayor et al. (2008), whereas POUND 26 fits into the Nanay ancestral group and POUND 31 fits into the Contamana ancestral group (Zhang et al. 2009). Currently, there are 29 500 accession names in the International Cocoa Germplasm Database (ICGD; Turnbull and Hadley 2019) with over 24 000 accessions being distributed over 40 cacao collections (CacaoNet 2012). The majority of these accessions need to be fingerprinted and (or) have multi-locus profiles deposited in the ICGD.

The use of single nucleotide polymorphism (SNP) molecular markers is increasingly used in identity and diversity studies in plants. In comparison to other molecular markers, SNPs have gained popularity owing to their high abundance in genomes, amenability for automation, and high-throughput in processing many samples with many markers, and increasing cost-effectiveness. Collard and Mackill (2008) reported that SNPs can increase the efficiency and precision of breeding. Also, SNPs have been used to characterize crops such as maize (Van Inghelandt et al. 2010) and soybean (Liu et al. 2017). In a genotyping by sequencing approach, Singh et al. (2019) mapped SNPs in *Aegilops tauschii* and found evidence that putatively different accessions of the wild wheat relative were duplicates of the same cultivar within and among genebanks. Duplication and mislabelling of accessions are common to genebanks (van Hintum 2000; Hurka et al. 2004) including cacao (Motilal et al. 2017; Motilal 2018). Currently, cacao is maintained as living trees in field genebanks. An accession refers to a tree that is clonally propagated. Due to the historic and continued repeated movement of germplasm material within and among cocoa countries, errors in collection of budwood, propagation, greenhouse inventory, transportation for transplanting, planting out in the field, and rootstock dominance of plants derived from grafting, have led to the loss of, and confusion of, tree identities. This is compounded by the practise in the past of renaming accessions when they were translocated to different countries. Additionally, seedlings established from a mother tree may have been given the name of the mother tree despite being sexual recombinants and hence genotypically different. Putative clonal copies of an accession in a cocoa genebank may therefore be the same or dissimilar. In the latter case, the putative copies may be sorted into groups of variable numbers with each group being assigned a different accession name. It is possible for putative copies to belong to the same genetic group but be dissimilar because they represent different accessions of that group. Errors in nomenclature can therefore only be resolved provided that each individual accession can be distinguished.

The determination of whether different accessions are duplicates (same multi-locus fingerprint profile) or mislabelled (different from expected genetic background or multi-locus fingerprint profile) requires having sufficient discriminatory multi-locus profiles for comparisons. Until recently, microsatellite markers were used in cacao to resolve such issues but these markers are being increasingly supplanted by SNPs (Motilal et al. 2017; Motilal 2018). Argout et al. (2008, 2011) first identified SNP loci from expressed genes in a wide range of cacao tissues. The first two cacao genome maps (Argout et al. 2011; Saski et al. 2011) were subsequently followed by the sequencing of 200 genomes (Cornejo et al. 2018). Recently, over 6000 SNPs were identified by Livingstone et al. (2015), 13 000 SNPs by Livingstone et al. (2017), and 7 million SNPs by Cornejo et al. (2018) in cacao using transcript and genome-based approaches.

Nevertheless, there is yet no firm consensus on a minimal SNP panel for cacao accession identification. Previous approaches using microsatellites (SSRs) had recommended a set of 15 SSR loci for identity analysis to the cacao community (Saunders et al. 2004). Motilal et al. (2009) showed that the composition of the SSR primer panel (quantity and choice of marker) was critical to obtaining full resolution among unique accessions. Cacao SNP genotyping has typically used 96 SNPs originating from Michel Boccara who screened over 1500 SNPs developed for Illumina's GoldenGate Assay based on Argout et al. (2008) as reported in Ji et al. (2013), Fang et al. (2014), Osorio-Guarin et al. (2017), and Arevalo-Gardini et al. (2019). The selection was reportedly based on the level of polymorphism and their distribution across the 10 chromosomes (Ji et al. 2013; Fang et al. 2014).

Ji et al. (2013) in a study of 115 cacao on-farm varieties and 70 SNPs, found that the 26 most informative SNPs

could be used to differentiate among 115 varieties with 99.999% certainty. However, these authors did not indicate the chromosomal locations or surrounding sequences of the 26 SNP markers. Fang et al. (2014) used 48 SNPs to demonstrate the feasibility of SNPs in cacao authentication and traceability. Takrama et al. (2014) used 53 SNPs to reliably separate 39 accessions. Livingstone et al. (2015), using a 6K SNP chip array, found that 30 SNP loci were adequate to differentiate between three pairwise combinations of closely related individuals but did not indicate whether the 30 loci could discriminate amongst all the 1152 accessions or provide details of the 30 loci that were used. Separation of closely related accessions however, does not guarantee that these same SNPs will work well in separating accessions of diverse origins. Padi et al. (2015) reported 64 SNPs that discriminated amongst 2424 individuals, although their panel could not resolve a set of three Amelonado accessions and a set of two Marañon accessions. In an empirical study involving 81 accessions and 546 SNPs, Motilal et al. (2017) recommended two panels of 96 SNPs for identity analysis of cacao. These 192 SNPs were present on all 10 chromosomes, some had high information content and some also resolved closely related accessions. However, a systematic assessment of the number of SNPs and the method used to compile SNP panels was lacking in previous work. Furthermore, notwithstanding the SNP studies already conducted, the cacao community is yet to decide as to which SNPs should be included in a consensus identity panel.

Thus, although SNP panels are well developed in cacao, an efficient common panel of SNPs is needed to uniquely identify cacao accessions, resolve nomenclature errors in germplasm collections, identify progenies from breeding trials, and to correctly position newly collected germplasm relative to established genebank collections. To enable wide-spread adoption of a consensus panel, ideally as few SNPs as possible should be included in the panel to minimize the cost of genotyping samples and to allow designer panels to be created from the consensus panel by adding in discriminatory SNPs relevant to the user-defined material under study. If the cacao community adopted a minimal consensus panel of SNPs to use for genotyping, it would simplify the process of depositing data into the ICGD and allow comparison of globally collected samples.

Here, we reassessed the SNPs proposed by Motilal et al. (2017) to identify a panel with minimal SNPs that could be used as a core panel for identity resolution in cacao. We present results of 146 SNP panels based on informativeness, linkage disequilibrium (LD), linkage group (LG) distribution, and allele frequency on a training set of 155 cacao accessions. The best panels were then validated on 1220 accessions to identify a minimal core SNP panel for cacao DNA fingerprinting.

Materials and methods

Healthy leaves were collected from 155 unique cacao trees from the International Cocoa Genebank Trinidad (ICGT; 149 sampled trees), Jamaica (three sampled trees), and Haiti (three sampled trees). Four to five leaf discs (6 mm diameter) were punched from each cleaned leaf and placed in collection plates from LGC Genomics, UK. This training set of 155 samples represented accessions with background from 9 of the 10 ancestral groups of Motamayor et al. (2008) comprising of 15 Amelonado, 12 Contamana, 4 Criollo, 10 Guiana, 19 Iquitos, 13 Marañon, 27 Nanay, and mixed individuals drawn from Curaray (11), Nacional (7), and various (38) accessions with mixed lineages (Refractario, Amelonado, Contamana, Curaray, and Nacional). The samples were submitted to LGC Genomics for SNP genotyping using 192 SNPs developed by Centre de coopération internationale en recherche agronomique pour le développement (Motilal et al. 2017). These SNPs were reduced to 182 SNPs (less than 5.85% missing data). The major allele frequency (M_jAF), minor allele frequency (M_iAF; $0 < M_iAF < 0.5$), and expected heterozygosity (H_e) in the genetic data of 182 SNPs from 155 accessions were obtained using GenAlEx v6.503 (Peakall and Smouse 2006, 2012). Three SNPs were monomorphic in the dataset of 155 accessions, and the M_jAF and H_e ranged between 0.503–0.997 and 0.006–0.502, respectively, for the polymorphic SNPs. The polymorphism information content (PIC) as found using Cervus v3.03 (Marshall et al. 1998) ranged between 0.006–0.375 for the 179 polymorphic SNP markers.

A total of 146 initial SNP panels were constructed as follows:

- Random panels in stepwise increments of five SNPs from 5 to 105 inclusive were created. Three random panels at each of the 21 incremental steps were generated for a subtotal of 63 panels.
- Twenty-four panels were designed using the H_e, M_jAF, or PIC statistic. Eight panels based on each statistic were created in stepwise increments of 12 SNPs from 12 to 96 inclusive. The SNPs were preferentially included based on the highest values for the relevant statistic and were not restricted by linkage group (chromosome; LG). In these panels, the relevant statistic for H_e, M_jAF, and PIC ranged between 0.407–0.502, 0.693–0.997, and 0.323–0.375, respectively.
- Nine panels were created based only on the presence and location of markers with the goal of achieving as even a distribution as possible across the 10 chromosomes. Panels were created in incremental steps of 10 SNPs from 20 to 100 random SNPs by adding in one SNP per LG at each step.
- Twenty-one panels based on linkage disequilibrium (LD) values in cacao (average LD = 8.63 cM; Motilal et al. 2016) were created. Panels contained 5–35 SNPs in incremental steps of five SNPs. Strings of SNPs that were not in LD were created by adding in a random SNP that

was not in LD with the previous ones. Three random panels were created at each incremental size. At set sizes of 20, 25, 30, and 35 SNPs there were 1, 3, 5, and 26 common SNPs across the three random panels in each set size. The maximum number of SNPs that had no two SNPs in LD was 35 in this dataset.

- Twenty-nine panels based on M_iAF were created. Sets of 24, 48, 72, and 96 SNPs were identified in each cluster, except for the Contamana, Nanay, and Nacional/Curaray clusters in which panels of 96 SNPs could not be identified because of the incidence of monomorphic SNP markers. Only 46 SNPs were identified under M_iAF selection from the Guiana cluster that had 136 monomorphic SNPs, so selection based on this group was not considered. General panels based on common SNPs across genetic clusters were compiled containing 24, 48, 72, and 96 SNP markers. Panels were increasingly nested within each other.

Sixteen designer panels were then constructed based on the performances of the aforementioned panels to maximally and unambiguously separate the 155 samples. Several of the initial 146 panels were chosen that achieved near resolution. Variable numbers of selected SNPs were then added in as needed to obtain the lowest number of SNPs in a panel to resolve all of the 155 reference accession samples. A total of 162 panels were therefore assessed on the training set of the 155 samples.

The ability of each panel to distinguish amongst the 155 reference samples was assessed using Cervus v3.03 (Marshall et al. 1998) with fuzzy matching being set at five loci. Samples were declared as exact matches if the samples had the same genotypic data for all the SNP markers. Samples with exact matching to each other constituted an unresolved group such that the smallest unresolved group will contain a sample pair (two accessions). A set of SNPs will separate accessions into variable groups with variable numbers of unresolved sample pairs in each group. An optimal panel will separate unique accessions into the largest number of groups in which each group will contain only one accession (zero unresolved sample pairs). Mismatches were identified as SNPs at which the genotypic data of the samples differed. Fuzzy matching indicated samples that could become similar if genetic differences were ignored and an upper tolerance limit of five mismatches was used. A sample pair could then be matched at all but one, two, three, four, or five SNP markers and could be considered as having the same identity under fuzzy matching. An optimal panel will resolve all non-identical samples by having mismatch events at many SNP markers.

The resolution abilities of the 16 designer panels were tested on a data set of 1220 accessions from the ICGT and for which data on 170 SNPs were available. The 1220 accessions contained 138 accessions from the training set and had ancestral contributions from the 10 genetic groups of Motamayor et al. (2008). Fuzzy matching was

Fig. 1. Identity resolution from single nucleotide polymorphism (SNP) panels based on random sets and linkage disequilibrium (LD) assessed as (A) unresolved sample pairs and mismatch events in (B) random and (C) linkage disequilibrium sets. Values are means ± standard error of the mean from three independent sets at each SNP sample size. Mismatch values are log transformed. Samples with exact matching to each other constitute an unresolved group, and the smallest group will contain an unresolved sample pair. Mismatches were identified as number of SNPs at which the genotypic data of the samples differed for samples that were otherwise equivalent for all other SNP markers. The maximum number of SNPs under LD and random selections was 35 and 105, respectively.

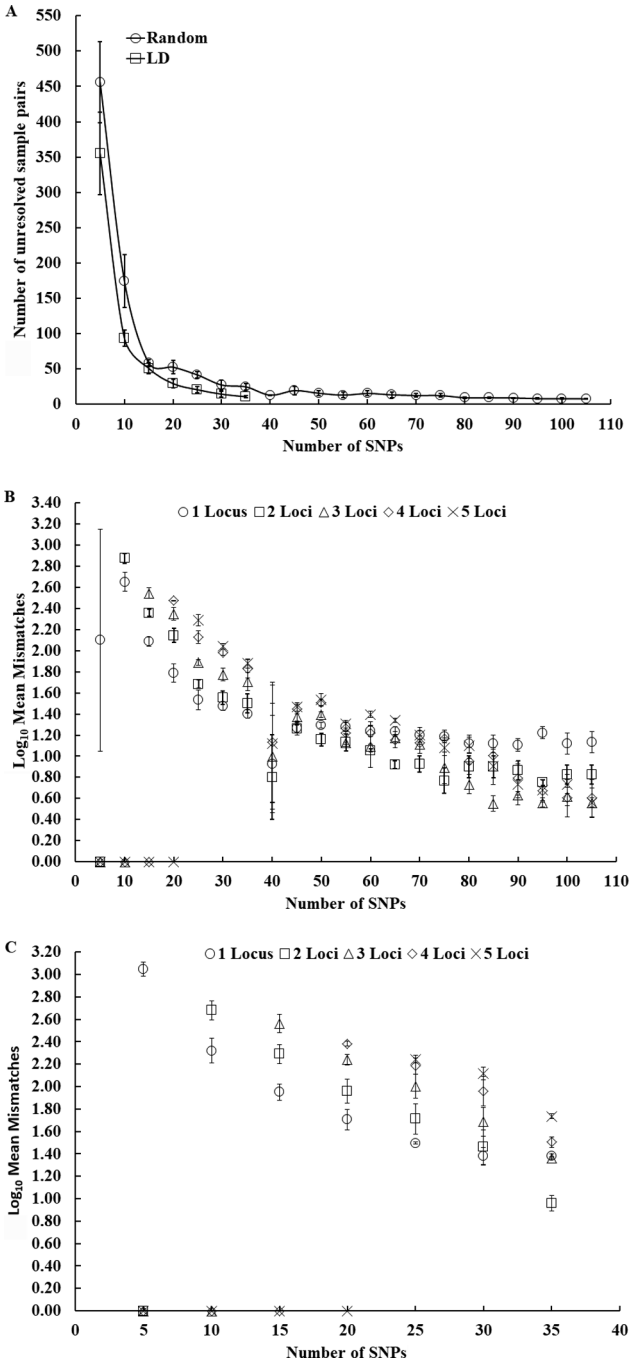


Table 1. Identity resolution of 155 cacao (*Theobroma cacao* L.) samples with 62 single nucleotide polymorphism (SNP) panels.

Panel[a]	No. of SNPs	No. of unresolved groups[b]	No. of unresolved sample pairs	No. of mismatches at				
				1 SNP	2 SNPs	3 SNPs	4 SNPs	5 SNPs
H_e	12	41	78	135	232	413	493	590
	24	16	26	38	46	60	97	113
	36	12	16	16	26	21	27	22
	48	9	10	13	24	13	8	12
	60	8	9	8	5	20	12	5
	72	7	8	9	5	13	12	3
	84	7	8	8	6	11	8	4
	96	5	5	11	5	11	4	1
M_jAF	12	133	7897	1394	703	455	794	298
	24	121	5194	2266	619	334	667	354
	36	87	1173	2125	2027	1420	693	244
	48	46	121	398	610	814	806	551
	60	30	45	94	182	179	189	181
	72	22	27	45	52	110	65	50
	84	16	17	18	33	46	53	49
	96	12	13	15	6	5	22	31
PIC	12	22	78	135	232	413	493	590
	24	15	26	38	46	60	97	113
	36	9	16	16	26	21	27	22
	48	9	10	12	24	14	6	15
	60	7	9	8	5	20	12	5
	72	7	8	9	5	13	12	3
	84	6	8	8	6	11	8	4
	96	5	5	11	5	11	4	1
LG	20	19	35	25	80	135	266	450
	30	16	33	9	35	41	62	86
	40	8	10	24	18	20	23	28
	50	8	15	23	1	13	11	16
	60	7	14	19	13	7	14	24
	70	6	8	16	8	7	8	6
	80	7	9	4	5	13	19	7
	90	13	15	15	3	4	10	11
	100	5	7	15	10	5	1	3
M_iAF (Amel)	24	19	39	33	67	190	276	355
	48	18	23	18	15	14	14	30
	72	7	9	14	5	16	12	6
	96	5	6	8	9	3	6	1
M_iAF (IMC)	24	18	36	20	47	105	198	207
	48	11	13	8	17	9	15	11
	72	7	7	7	8	16	5	6
	96	4	4	9	9	5	8	8
M_iAF (PA)	24	13	31	51	70	101	106	107
	48	11	13	8	16	20	31	49
	72	8	10	4	7	9	17	6
	96	5	6	8	6	4	8	2
M_iAF (Refro)	24	16	32	26	51	56	98	141
	48	8	10	18	15	10	19	12
	72	6	7	16	5	8	9	9
	96	4	5	15	6	7	4	4
M_iAF (NaCur)	24	12	19	43	38	55	126	176
	48	12	16	33	18	17	15	23
	72	7	8	6	30	17	11	15

Table 1 (*concluded*).

Panel[a]	No. of SNPs	No. of unresolved groups[b]	No. of unresolved sample pairs	No. of mismatches at 1 SNP	2 SNPs	3 SNPs	4 SNPs	5 SNPs
M_iAF	24	13	15	18	43	36	44	51
(NA)	48	9	10	6	13	18	15	15
	72	6	7	7	7	14	9	8
M_iAF	24	9	45	21	26	52	88	134
(SCA)	48	13	20	24	28	10	20	31
	72	8	10	5	18	16	19	16
M_iAF	24	9	16	30	25	49	80	140
(Gen)	48	10	11	22	9	6	15	17
	72	6	6	15	11	3	6	4
	96	5	5	8	7	9	3	1

[a]Panels based on selections from expected heterozygosity (H_e), major allele frequency (M_jAF), polymorphism information content (PIC), distribution on linkage groups (LG), and minor allele frequency (M_iAF) from Amelonado (Amel), Contamana (SCA), general (Gen), Iquitos (IMC), Marañon (PA), Nacional/Curaray (NaCur), Nanay (NA), and Refractario (Refro) clusters.

[b]Two samples with exact matching to each other at all SNPs constitute an unresolved sample pair and all sample pairs that are equivalent to each other form an unresolved group.

set at five loci with minimal matching set at 24 SNPs for CRC48 panel and 50 SNPs for the other 15 panels in Cervus v3.03 (Marshall et al. 1998). Accumulated probability of identity among siblings (total PID_{sib}; Waits et al. 2001) was obtained from Cervus v3.03 (Marshall et al. 1998) for the panels on 1220 accessions. The PID_{sib} is defined as the probability that two sibling individuals drawn at random from a population have the same multi-locus genotype (Waits et al. 2001). The total PID_{sib} estimates the likelihood of individuals being the same when genotyped over all the SNPs given the samples in the dataset and will indicate the most conservative number of loci required to resolve all individuals, including relatives (Waits et al. 2001).

Results

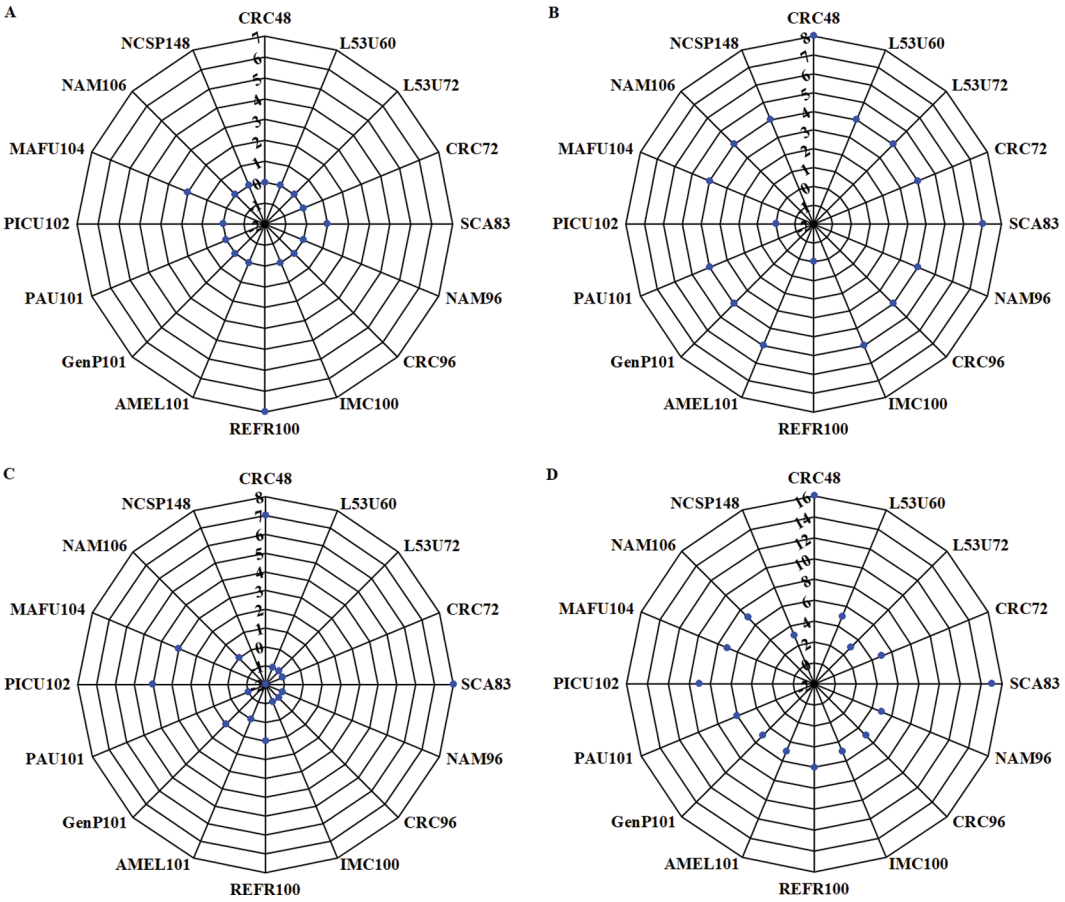
Generally, the resolution ability of the SNP panels increased with increasing number of SNPs tending towards full resolution among the training set of 155 accessions and a reduced number of closely matched fuzzy equivalents (Fig. 1; Table 1). None of the 146 initial panels, across all tested composition strategies (H_e, M_iAF, M_jAF, LD, LG, PIC, and random), unambiguously separated the 155 cacao samples. At 100 random SNPs, an average of eight sample pairs were genotypically equivalent across all loci and on average 13 sample pairs were different at one locus, 6 sample pairs differed at two loci, 3.3 sample pairs differed at three loci, 3.7 sample pairs differed at four loci, and 4.7 sample pairs differed at five loci. In general, panels based on the M_jAF performed the worst, having at least 100 times more unresolved sample pairs at the lowest panel sizes and was the last ranked panel over all panel sizes (Table 1).

The number of unresolved sample pairs decreased in an exponential manner tapering off around 30 SNPs for both random and LD sets (Fig. 1). Panels based on LD generally performed better than those based on random selection. The LD panel of 25 SNPs performed better on average than the LG panels (20 and 30 SNPs) and the H_e, M_iAF, and PIC panels (24 SNPs), with the exception of the M_iAF panels, based on Nacional/Curaray, Nanay, and General groups (Table 1; Fig. 1). However, one of the replicates of the LD panels of 25 SNPs outperformed all others having only 11 unresolved sample pairs and 31 fuzzy matches at one locus. When a panel size of 35 SNPs was used, the LD panels (on average and for each individual panel) had fewer unresolved sample pairs but more fuzzy matches at one locus than H_e and PIC panels with 36 SNPs (Table 1; Fig. 1). On average, the LD panel of 35 SNPs was slightly worse than the LG panel of 40 SNPs with 11 instead of 10 unresolved sample pairs. However, one of the replicates of the LD panel of 35 SNPs had the same resolution and one replicate achieved better resolution with nine unresolved sample pairs than the LG panel of 40 SNPs.

Comparison of panels based on LG and random selection for panel sizes of 20–100 inclusive showed better resolution across all panel sizes, except for 90 SNPs, for selection based on LG. However, some of the replicates of the random panels were comparable with, or better than, the LG panel for 30, 50, 60, 80, 90, and 100 panel sizes. When LG panels (30, 50, 80, and 100) were compared with H_e, M_iAF, and PIC panels with comparable but lower number of SNPs, only the 50 SNP panel had the LG set as a top performer but was second to the M_iAF Refractario 50 SNP panel.

The M_iAF panels had the least number of unresolved sample pairs when panels were composed from the Iquitos, Marañon, Nanay, and Refractario clusters at lower panel sizes, but the differences among panels was reduced as the number of SNPs increased. The M_iAF panels

Fig. 2. Incidence of mismatches at (A) one, (B) two, (C) three, and (D) four loci with 16 single nucleotide polymorphism (SNP) designer panels on 155 cacao (*Theobroma cacao* L.) accessions relative to the full panel of 182 SNP markers. CRC48, set of 38 SNPs that are not in linkage disequilibrium with TcSNP0013 with 10 additional SNPs; L53U60, SNPs that separate samples with 1–10 loci differences with seven additional SNPs; L53U72, L53U60 panel with 12 additional SNPs; CRC72, CRC48 with 24 additional SNPs; SCA83, 72 SNPs from minor allele frequency (M_iAF) in Contamana cluster with 11 additional SNPs; NAM96, random set of 90 with six additional SNPs; CRC96, CRC72 with 24 additional SNPs; IMC100, 96 SNPs from M_iAF in Iquitos cluster with four additional SNPs; REFR100, 96 SNPs from M_iAF in Refractario cluster with four additional SNPs; AMEL101, 96 SNPs from M_iAF in Amelonado cluster with five additional SNPs; GenP101, common set of 96 SNPs from M_iAF in genetic clusters with five additional SNPs; PAU101, 96 SNPs from M_iAF in Marañon cluster with five additional SNPs; PICU102, panel of 96 SNPs with highest polymorphism information content (PIC) with six additional SNPs; MAFU104, panel of 96 SNPs with highest major allele frequency with eight additional SNPs; NAM106, 10 SNPs on each of the 10 cacao chromosomes and six additional SNPs; NCSP148, union of NAM96 and NAM106 panels. Mismatches were identified as number of SNPs at which the genotypic data of the samples differed for samples that were otherwise equivalent for all other SNP markers.



of 96 SNPs based on the Amelonado, Iquitos, Marañon, and General clusters performed the best with only 8–9 sample pairs that differed at only one locus. In contrast, the LG panel, composed as 10 SNPs per chromosome (100 SNPs), had 15 sample pairs that differed at one locus. When the total number of sample pairs over fuzzy matching at one locus or two loci was considered, panels of 96 SNPs based on H_e, PIC, and M_iAF (Amelonado, Iquitos, Marañon, and General clusters) had 14–18 closely matched sample pairs with M_iAF panels based on Marañon and General clusters performing the best with 14 and 15 matched pairs, respectively (Table 1).

All 16 designer panels separated the 155 accessions from each other without having any exact matches. Fuzzy matches at 1–4 loci were observed relative to the full complement of 182 SNPs (Fig. 2). The worst performing panels were SCA83, CRC48, MAFU104, and REFR100, with 36, 35, 29, and 28 total fuzzy matches for one to three loci, whereas the full panel of 182 SNPs had 20 fuzzy matches. Seven panels (L53U60, L53U72, CRC72, CRC96, NAM96, PAU101, NCSP148), three panels (IMC100, PICU102, NAM106), and two panels (AMEL101, GenP101) had 22–23, 24, and 25 total fuzzy matches, respectively, over three loci, but the relative performance of these panels to the full complement of 182 SNPs were similar (Fig. 2).

The set of 170 SNPs separated the 1220 varieties with 12, 39, and 23 pairs of fuzzy matches at one, two, and three SNP loci, respectively. Comparative performance of the 16 designer panels to the resolution ability of

Table 2. Exact matching and probability of identities of designer panels on 1220 cacao varieties.

Panel[a]	No. of SNPs[b]	No. of unresolved groups[c]	No. of unresolved sample pairs	Maximum no. of unresolved pairs in a group	Total PID_{sib}[d]
CRC48	48	6	6	1	4.640×10^{-9}
L53U60	60	5	5	1	3.230×10^{-12}
L53U72	72	4	4	1	4.621×10^{-14}
CRC72	72	3	3	1	7.305×10^{-14}
SCA83	83	15	21	3	2.481×10^{-14}
NAM96	96	4	4	1	9.899×10^{-19}
CRC96	96	2	2	1	3.864×10^{-18}
MAFU104	97	8	18	10	1.306×10^{-14}
PICU102	98	5	7	3	1.209×10^{-20}
IMC100	100	8	10	3	9.140×10^{-20}
REFR100	100	7	9	3	6.650×10^{-20}
AMEL101	101	3	3	1	2.782×10^{-19}
GenP101	101	4	4	1	1.503×10^{-19}
PAU101	101	5	5	1	4.579×10^{-20}
NAM106	106	2	2	1	3.548×10^{-19}
NCSP148	148	1	1	1	8.165×10^{-27}
Full170	170	0	0	0	6.838×10^{-30}

[a]CRC48, set of 38 SNPs in linkage disequilibrium with TcSNP0013 with 10 additional SNPs; L53U60, SNPs that separate samples with 1–10 loci differences with seven additional SNPs; L53U72, L53U60 panel with 12 additional SNPs; CRC72, CRC48 with 24 additional SNPs; SCA83, 72 SNPs from minor allele frequency (M_iAF) in Contamana cluster with 11 additional SNPs; NAM96, random set of 90 with six additional SNPs; CRC96, CRC72 with 24 additional SNPs; IMC100, 96 SNPs from M_iAF in Iquitos cluster with four additional loci; REFR100, 96 SNPs from M_iAF in Refractario cluster with four additional SNPs; AMEL101, 96 SNPs from M_iAF in Amelonado cluster with five additional SNPs; GenP101, common set of 96 SNPs from M_iAF in genetic clusters with five additional SNPs; PAU101, 96 SNPs from M_iAF in Marañon cluster with five additional SNPs; PICU102, panel of 96 SNPs with highest polymorphism information content (PIC) with six additional SNPs, four SNPs removed because of missing data; MAFU104, panel of 96 SNPs with highest major allele frequency with eight additional SNPs, seven SNPs removed because of missing data; NAM106, 10 SNPs on each of the 10 cacao chromosomes and six additional SNPs; NCSP148, union of NAM96 and NAM106 panels; Full170, maximal set of 170 SNPs.

[b]Single nucleotide polymorphism markers.

[c]Two samples with exact matching to each other at all SNPs constitute an unresolved sample pair and all sample pairs that are equivalent to each other form an unresolved group.

[d]Accumulated probability of identity among siblings (Waits et al. 2001).

170 SNPs on 1220 varieties showed that the worst performers were SCA83, MAFU104, IMC100, and REFR100 panels and best performers were CRC96, AMEL101, NAM106, and NCSP148 when exact matches (Table 2) and fuzzy matches (Fig. 3) were considered. The designer panel with the lowest number of SNPs (CRC48) had less exact matches than panels with more SNPs like SCA83, MAFU104, IMC100, and REFR100 but had much more fuzzy matches (Fig. 3).

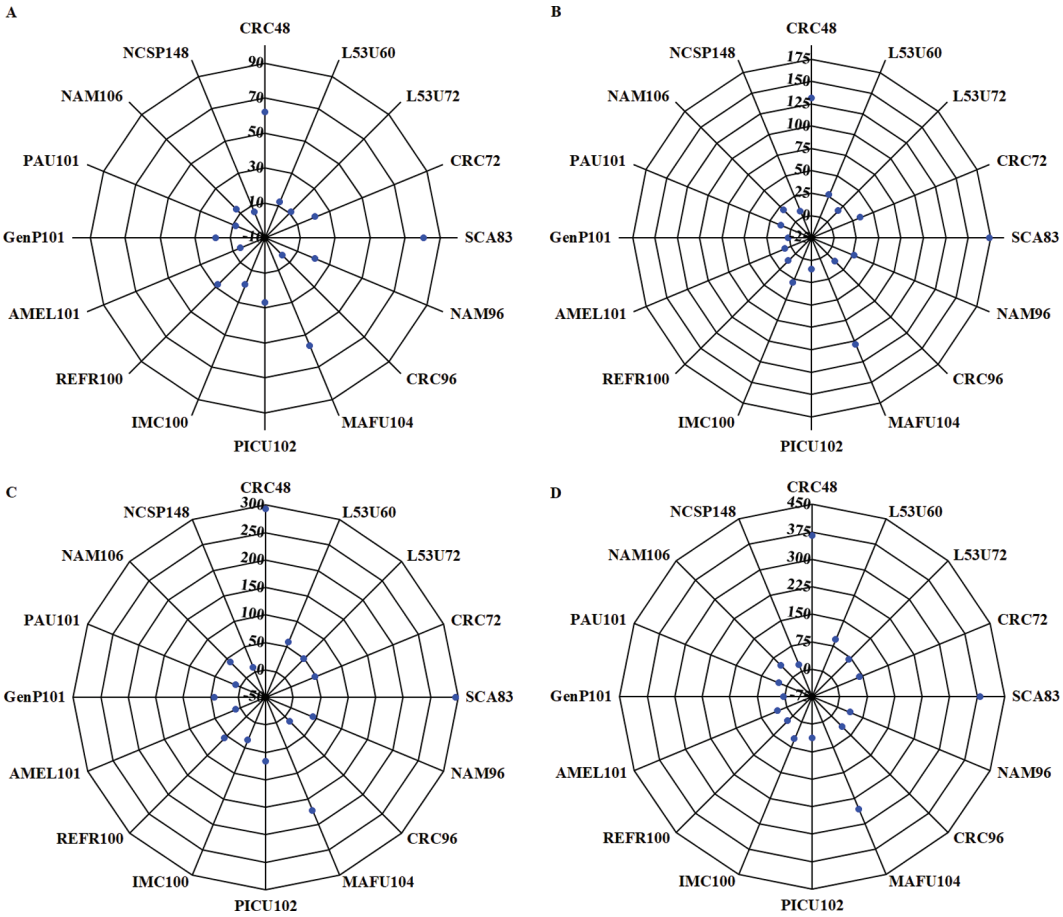
The accumulated PID_{sib} values on the 1220 accessions ranged from 4.640×10^{-9} (CRC48) to 6.838×10^{-30} (Full170) and generally decreased as the number of SNPs increased (Table 2). Accumulated PID_{sib} values in MAFU104 were four, five, and six orders of magnitude greater than the CRC96, NAM96, and PICU102 panels although their number of SNPs was similar. Although NAM96 had a lower accumulated PID_{sib} value than CRC96, the latter had less exact matches (Table 2) and less fuzzy matches at one to three loci (Fig. 3) than the NAM96 panel.

Designer panels with 100–106 SNPs yielded an accumulated PID_{sib} in the order of 10^{-19} to 10^{-20}, which was 10 times greater than that of 170 SNPs in the maximal possible set.

Discussion

In this study, we examined different combinations of 179 SNPs drawn from a recommended panel of 192 SNPs (Motilal et al. 2017) to identify the minimum number of SNPs that could perform similarly to the full panel for identity analysis in cacao. Panels created from any method (random, H_e, LD, LG, M_iAF, M_jAF, PIC) did not fully resolve all 155 reference samples in the training set. However, the judicious use of additional SNPs to maximise mismatches amongst samples could be used to obtain panels which resulted in zero exact matches in the training set. The resolving ability of the panels was not solely determined by the method of selection as the performance ranking of the panels was inconsistent. Nevertheless, panels based on M_iAF or LD were usually the best. The best genetic group for an identity panel based on M_iAF varied with the number of SNPs involved but panels based on the Nanay and General set usually per-

Fig. 3. Incidence of mismatches at (A) one, (B) two, (C) three, and (D) four loci with 16 designer panels on 1220 cacao (*Theobroma cacao* L.) accessions relative to a maximal set of 170 single nucleotide polymorphism (SNP) markers. CRC48, set of 38 SNPs that are not in linkage disequilibrium with TcSNP0013 with 10 additional SNPs; L53U60, SNPs that separate samples with 1–10 loci differences with seven additional SNPs; L53U72, L53U60 panel with 12 additional SNPs; CRC72, CRC48 with 24 additional SNPs; SCA83, 72 SNPs from minor allele frequency (M_iAF) in Contamana cluster with 11 additional SNPs; NAM96, random set of 90 with six additional SNPs; CRC96, CRC72 with 24 additional SNPs; IMC100, 96 SNPs from M_iAF in Iquitos cluster with four additional SNPs; REFR100, 96 SNPs from M_iAF in Refractario cluster with four additional SNPs; AMEL101, 96 SNPs from M_iAF in Amelonado cluster with five additional SNPs; GenP101, common set of 96 SNPs from M_iAF in genetic clusters with five additional SNPs; PAU101, 96 SNPs from M_iAF in Marañon cluster with five additional SNPs; PICU102, panel of 96 SNPs with highest polymorphism information content (PIC) with six additional SNPs, four SNPs removed because of missing data; MAFU104, panel of 96 SNPs with highest major allele frequency with eight additional SNPs, seven SNPs removed because of missing data; NAM106, 10 SNPs on each of the 10 cacao chromosomes and six additional SNPs; NCSP148, union of NAM96 and NAM106 panels. Mismatches were identified as number of SNPs at which the genotypic data of the samples differed for samples that were otherwise equivalent for all other SNP markers.



formed optimally yielding the least number of unresolved groups, least number of unresolved sample pairs and least number of fuzzy matches combined for one to two SNP markers. A set of 96 SNPs (CRC96) starting with 28 SNPs in LD was the best panel with the least number of SNPs to resolve the real world dataset of 1220 cacao accessions. The panels PICU102 and MAFU104 were missing four and seven SNPs respectively in the set of 1220 varieties and their resolving power may have been better if there was no missing data. Still, the PICU102 panel (98 SNPs) was clearly better than MAFU104 (97 SNPs) and IMC100 and REFR100 (each with 100 SNPs) panels. The panel PICU102 could have had similar resolution to the CRC96 and AMEL101 panels if there were no missing data.

Identity issues in cacao genebanks were previously resolved with microsatellites (Saunders et al. 2004; Motilal 2018; Motilal et al. 2009, 2017; Zhang et al. 2009) but SNPs are becoming the marker of choice (Takrama et al. 2014; Livingstone et al. 2015; Motilal 2018). Singh et al. (2019) employed about 20 000 SNPs to detect substantial mislabelling of a wild wheat relative within and among genebanks. Identity resolution in cacao may not be dependent on such large numbers of SNP markers. Our results show that for identity analysis in cacao, the number of SNP markers must be complemented by choosing

SNPs that can resolve closely related samples even at the expense of having low discriminatory power. A similar result for SSR markers in cacao was previously reported (Motilal et al. 2009). Our results also corroborated that of Yoon et al. (2007) who reported that the efficacy of the SNPs, as well as, the size and diversity of the population being investigated would influence the composition of the SNP panel. These results indicate that there is an ascertainment bias in selecting SNP panels for identity resolution. Earlier studies with low numbers of SNPs and accessions (Ji et al. 2013; Fang et al. 2014; Lukman et al. 2014; Takrama et al. 2014) were therefore fortunate in achieving reliable separation and showed that the polymorphic SNPs used in these studies had good separation ability. Thirty-seven of the SNPs provided in cacao identity studies that listed the SNPs used (Ji et al. 2013; Fang et al. 2014; Takrama et al. 2014; Padi et al. 2015) were included in the recommended CRC96 panel (Table S1[1]). Ascertainment bias may occur when studies use widely divergent samples making it easier to identify fewer SNPs that can nevertheless discriminate amongst all accessions. Ascertainment bias can also occur when markers are based only on closely related individuals, as the selected SNPs may be unable to discriminate amongst more diverse accessions.

Increasing the number of samples highlighted the importance of choosing both the correct number of SNPs and the choice of SNP in creating a SNP panel. Similar findings were reported for soybean (Yoon et al. 2007; Liu et al. 2017). Although a minimum set of 48 SNPs, used in the CRC48 panel (designed in part on LD considerations), could completely resolve the identities of 155 accessions, the CRC48 panel performed relatively poorly in the larger set of 1220 accessions. Ten designer panels were better than the CRC48 panel at identity resolution yielding fewer unresolved sample pairs. However, this was probably a function of the number of SNPs. The CRC96 designer panel (Table S1[1]), gave better resolution than the CRC48 panel from which it was developed as evidenced by the presence of fewer mismatches on the training set of 155 accessions. The CRC48 panel was unable to resolve six sample pairs in the data set of 1220 accessions whereas the CRC96 panel performed better and was only unable to resolve two duplicate groups each with two samples (Table 2). Furthermore, even when the number of SNPs was nearly doubled, the set of 170 SNPs separated these accession pairs by only one SNP each. This indicated that with this set of 170 SNPs, the matched accessions are possible duplicates and are more related to each other than the other accessions. A PID_{sib} of about 10^{-18} was obtained for the CRC96 panel which was five orders of magnitude less than a suggested threshold for PID (Motilal et al. 2009). Nevertheless, as the number of samples to be resolved increases and as more closely related varieties are encountered, additional SNPs can be included to improve the CRC96 panel to obtain more stringent PID_{sib} values and 100% resolution of identities without close mismatching.

Closely related samples may have many SNPs with the same information and differ at only a few SNPs. An ideal panel will contain SNPs that maximise mismatches in any dataset without *a priori* knowledge of the samples to be evaluated. The amount of mismatches tolerated would need to be decided upon for each organism/marker/platform system. A common panel of SNPs that can reliably discriminate amongst accessions is a valuable resource to the cacao community and is the first step towards resolving mislabelling among and within germplasm collections, as well as, comparing and sharing diversity data. The designer panel CRC96 (Table S1[1]) is recommended as the base panel to which SNPs can be added as needed. This panel or a subset of it can be used as a panel to deposit SNP fingerprint data in the ICGD or other public domain cacao databases.

Acknowledgements

We thank Mitra Benny and Vickeisha Lall for helping with the identity analysis. Rena Kalloo, Mitra Benny, and Michel Boccara are thanked for assisting with sample collection. This study was financially supported by funding for the project “Leveraging the International Cocoa Gene Bank to Improve Competitiveness of the Cocoa Sector in the Caribbean, Using Modern Genomics” obtained from The UWI-Trinidad and Tobago Research and Development Impact Fund. Additional support was obtained from the International Fine Cocoa Innovation Centre (IFCIC), of the Cocoa Research Centre, which is funded by the European Union under the ACP Science & Technology Programme II and The University of the West Indies.

References

Alverson, W.S., Whitlock, B.A., Nyffeler, R., Bayer, C., and Baum, D.A. 1999. Phylogeny of the core Malvales: evidence from ndhF sequence data. Am. J. Bot. **86**(10): 1474–1486. doi:10.2307/2656928. PMID:10523287.

Arevalo-Gardini, E., Meinhardt, L.W., Zuñiga, L.C., Arévalo-Gardni, J., Motilal, L., and Zhang, D. 2019. Genetic identity and origin of “Piura Porcelana”—a fine-flavored traditional variety of cacao (*Theoborma cacao*) from the Peruvian Amazon. Tree Genet. Genomes, **15**: 11. doi:10.1007/s11295-019-1316-y.

Argout, X., Fouet, O., Wincker, P., Gramacho, K., Legavre, T., Sabau, X., et al. 2008. Towards the understanding of the cocoa transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions. BMC Genomics, **9**: 512. doi:10.1186/1471-2164-9-512. PMID:18973681.

Argout, X., Salse, J., Aury, J.M., Guiltinan, M.J., Droc, G., Gouzy, J., et al. 2011. The genome of *Theobroma cacao*. Nat. Genet. **43**(2): 101–109. doi:10.1038/ng.736. PMID:21186351.

[1]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2019-0071.

Bayer, C., Fay, M.F., De Bruijn, P.Y., Savolainen, V., Morton, C.M., Kubitzki, K., et al. 1999. Support for an expanded family concept of Malvaceae within a recircumscribed order Malvales: a combined analysis of plastid atpB and rbcL DNA sequences. Bot. J. Linn. Soc. **129**(4): 267–303. doi:10.1111/j.1095-8339.1999.tb00505.x.

Bekele, F., and Butler, D.R. 2000. Proposed short list of cocoa descriptors for characterization. *In* Working Procedures for Cocoa Germplasm Evaluation and Selection. *Edited by* A.B. Eskes, J.M.M. Engels, and R.A. Lass. Proceedings of the CFC/ICCO/IPGRI Project Workshop, Montpellier, France, 1998. International Plant Genetic Resources Institute, Rome, Italy, pp. 41–48.

CacaoNet. 2012. A global strategy for the conservation and use of cacao genetic resources, as the foundation for a sustainable cocoa economy. *In* Bioversity International, Montpellier, France. *Compiled by* B. Laliberté. Available from http://cacaonet.org/.

Cheesman, E.E. 1944. Notes on the nomenclature, classification and possible relationships of cacao populations. Trop. Agric. **21**: 144–159.

Collard, B.C.Y., and Mackill, D.J. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. B Biol. Sci. **363**: 557–572. doi:10.1098/rstb.2007.2170. PMID:17715053.

Cornejo, O.E., Yee, M.C., Dominguez, V., Andrews, M., Sockell, A., Strandberg, E., et al. 2018. Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. Commun. Biol. **1**: 167. doi:10.1038/s42003-018-0168-6. PMID:30345393.

Cuatrecasas, J. 1964. Cacao and its allies. A taxonomic revision of the genus *Theobroma*. *In* Systematic plant studies. Contributions from the United States National Herbarium. Smithsonian Institution, Washington, D.C., pp. 375–614.

Engels, J.M.M., Bartley, B.G.D., and Enriquez, G.A. 1980. Cacao descriptors, their states and modus operandi. Turrialba, **30**(2): 209–218.

Fang, W., Meinhardt, L.W., Mischke, S., Bellato, C.M., Motilal, L., and Zhang, D. 2014. Accurate determination of genetic identity for a single cacao bean, using molecular markers with a nanofluidic system, ensures cocoa authentication. J. Agric. Food Chem. **62**(2): 481–487. doi:10.1021/jf404402v. PMID:24354624.

Hurka, H., Neuffer, B., and Friesen, N. 2004. Plant genetic resources in botanical gardens. *In* Proceedings of the 21st International Symposium on Breeding Ornamentals, Part II. *Edited by* G. Forkmann and S. Michaelis. International Society for Horticultural Science, Leuven, Belgium, Acta Hortic. **651**: 35–44.

ICCO. 2017. Quarterly Bulletin of Cocoa Statistics, Vol. XLIII, No. 3, Cocoa year 2016/17. Available from https://www.icco.org/about-us/international-cocoa-agreements/cat_view/30-related-documents/46-statistics-production.html?limit=35&limitstart=0&order=date&dir=ASC.

Ji, K., Zhang, D., Motilal, L.A., Boccara, M., Lachenaud, P., and Meinhardt, L.W. 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. Genet. Resour. Crop Evol. **60**(2): 441–453. doi:10.1007/s10722-012-9847-1.

Liu, Z., Li, J., Fan, X., Htwe, N.M.P.S., Wang, S., Huang, W., et al. 2017. Assessing the numbers of SNPs needed to establish molecular IDs and characterize the genetic diversity of soybean cultivars derived from *Tokachi nagaha*. Crop J. **5**(4): 326–336. doi:10.1016/j.cj.2016.11.001.

Livingstone, D., Royaert, S., Stack, C., Mockaitis, K., May, G., Farmer, A., et al. 2015. Making a chocolate chip: development and evaluation of a 6K SNP array for *Theobroma cacao*. DNA Res. **22**(4): 279–291. doi:10.1093/dnares/dsv009. PMID:26070980.

Livingstone, D., III, Stack, C., Mustiga, G.M., Rodezno, D.C., Suarez, C., Amores, F., et al. 2017. A larger chocolate chip — development of a 15K *Theobroma cacao* L. SNP array to create high-density linkage maps. Front. Plant Sci. **8**: 2008. doi:10.3389/fpls.2017.02008. PMID:29259608.

Lukman, Zhang, D., Susilo, A.W., Dinarti, D., Bailey, B., Mischke, S., and Meinhardt, L.W. 2014. Genetic identity, ancestry and parentage in farmer selections of cacao from Aceh, Indonesia revealed by single nucleotide polymorphism (SNP) markers. Trop. Plant Biol. **7**(3–4): 133–143. doi:10.1007/s12042-014-9144-6.

Marshall, T.C., Slate, J., Kruuk, L.E.B., and Pemberton, J.M. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7**(5): 639–655. doi:10.1046/j.1365-294x.1998.00374.x. PMID:9633105.

Motamayor, J.C., Lachneaud, P., da Silva e Mota, J.W., Loor, R., Kuhn, D.N., Brown, J.S., and Schnell, R.J. 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). PLoS ONE, **3**(10): e3311. doi:10.1371/journal.pone.0003311. PMID:18827930.

Motamayor, J.C., Lachenaud, P., da Silva e Mota, J.W., Loor, R., Martinez, W.J., Graham, J., et al. 2010. 'No mas 'Forastero': a new protocol for meaningful cacao germplasm classification. *In* Proceedings of the 16th International Cacao Research Conference, Bali, Indonesia, 16–21 November 2009. COPAL, Nigeria, pp. 179–185.

Motilal, L.A. 2018. The role of gene banks in preserving the genetic diversity of cacao. *In* Achieving sustainable cultivation of cocoa. Vol. 1. Genetics, breeding, cultivation and quality. *Edited by* P. Umaharan. Burleigh Dodds Science Publishing Limited, Cambridge, U.K. pp. 47–99.

Motilal, L.A., Zhang, D., Umaharan, P., Mischke, S., Boccara, M., and Pinney, S. 2009. Increasing accuracy and throughput in large-scale microsatellite fingerprinting of cacao field germplasm collections. Trop. Plant Biol. **2**(1): 23–37. doi:10.1007/s12042-008-9016-z.

Motilal, L.A., Zhang, D., Mischke, S., Meinhardt, L.W., Boccara, M., Fouet, O., et al. 2016. Association mapping of seed and disease resistance traits in *Theobroma cacao* L. Planta, **244**(6): 1265–1276. doi:10.1007/s00425-016-7. PMID:27534964.

Motilal, L.A., Sankar, A., Gopaulchan, D., and Umaharan, P. 2017. Cocoa. *In* Biotechnology of plantation crops. *Edited by* P. Chowdappa, A. Karun, M.K. Rajesh, and S.V. Ramesh. Daya Publishing House, New Delhi, India. pp. 313–354.

Osorio-Guarín, J.A., Berdugo-Cely, J., Coronado, R.A., Zapata, Y.P., Quintero, C., Gallego-Sánchez, G., and Yockteng, R. 2017. Colombia a source of cacao genetic diversity as revealed by the population structure analysis of germplasm bank of *Theobroma cacao* L. Front. Plant Sci. **8**: 1994. doi:10.3389/fpls.2017.01994. PMID:29209353.

Padi, F.K., Ofori, A., Takrama, J., Djan, E., Opoku, S.Y., Dadzie, A.M., et al. 2015. The impact of SNP fingerprinting and parentage analysis on the effectiveness of variety recommendations in cacao. Tree Genet. Genomes, **11**(3): 1–14. doi:10.1007/s11295-015-0875-9.

Peakall, R., and Smouse, P.E. 2006. GenAlEx 6: genetic analysis in excel. Population genetic software for teaching and research. Mol. Ecol. Notes, **6**(1): 288–295. doi:10.1111/j.1471-8286.2005.01155.x.

Peakall, R., and Smouse, P.E. 2012. GenAlEx 6.5: genetic analysis in excel. Population genetic software for teaching and research — an update. Bioinformatics, **28**(19): 2537–2539. doi:10.1093/bioinformatics/bts460. PMID:22820204.

Saski, C.A., Feltus, F.A., Staton, M.E., Blackmon, B.P., Ficklin, S.P., Kuhn, D.N., et al. 2011. A genetically anchored physical framework for *Theobroma cacao* cv. Matina 1–6.

BMC Genomics, **12**: 413. doi:10.1186/1471-2164-12-413. PMID: 21846342.

Saunders, J.A., Mischke, S., Leamy, E.A., and Hemeida, A.A. 2004. Selection of international molecular standards for DNA fingerprinting of *Theobroma cacao*. Theor. Appl. Genet. **110**(1): 41–47. doi:10.1007/s00122-004-1762-1. PMID:15551041.

Singh, N., Wu, S., Raupp, W.J., Sehgal, S., Arora, S., Tiwari, V., et al. 2019. Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. Sci. Rep. **9**(1): 650. doi:10.1038/s41598-018-37269-0. PMID:30679756.

Takrama, J., Kun, J., Meinhardt, L., Mischke, S., Opoku, S.Y., Padi, S.K., and Zhang, D. 2014. Verification of genetic identity of introduced cacao germplasm in Ghana using single nucleotide polymorphism (SNP) markers. Afr. J. Biotechnol. **13**(21): 2127–2136. doi:10.5897/AJB2013.13331.

Toxopeus, H. 1985. Botany, types and populations. *In* Cocoa. 4th ed. *Edited by* G.A.R. Wood and R.A. Lass. Longman, London, UK. pp. 11–37.

Turnbull, C.J., and Hadley, P. 2019. International Cocoa Germplasm Database (ICGD). [Online Database.] CRA Ltd./ICE Futures Europe/University of Reading, UK. Available from http://www.icgd.reading.ac.uk (accessed 4 March 2019).

Utro, F., Cornejo, O.E., Livingstone, D., Motamayor, J.C., and Parida, L. 2012. ARG-based genome-wide analysis of cacao cultivars. BMC Bioinf. **13**(Suppl 19): S17. doi:10.1186/1471-2105-13-S19-S17. PMID:23281769.

Van Hintum, T.J.L. 2000. Duplication within and between germplasm collections. III. A quantitative model. Genet. Resour. Crop Evol. **47**(5): 507–513. doi:10.1023/A:1008703031415.

Van Inghelandt, D., Melchinger, A.E., Lebreton, C., and Stich, B. 2010. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor. Appl. Genet. **120**(7): 1289–1299. doi:10.1007/s00122-009-1256-2. PMID:20063144.

Waits, L.P., Luikart, G., and Taberlet, P. 2001. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. Mol. Ecol. **10**(1): 249–256. doi:10.1046/j.1365-294X.2001.01185.x. PMID:11251803.

Yoon, M.S., Song, Q.J., Choi, I.Y., Specht, J.E., Hyten, D.L., and Cregan, P.B. 2007. BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification. Theor. Appl. Genet. **114**(5): 885–899. doi:10.1007/s00122-006-0487-8. PMID:17219205.

Zarrillo, S., Gaikwad, N., Lanaud, C., Powis, T., Viot, C., Lesur, I., et al. 2018. The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. Nat. Ecol. Evol. **2**(12): 1879–1888. doi:10.1038/s41559-018-0697-x. PMID: 30374172.

Zhang, D., and Motilal, L. 2016. Origin, dispersal and current global distribution of cacao genetic diversity. *In* Cacao diseases: a history of old enemies and new encounters. *Edited by* B. Bailey and L. Meinhardt. Springer, New York. pp. 3–32.

Zhang, D., Boccara, M., Motilal, L., Mischke, S., Johnson, E.S., Butler, D.R., et al. 2009. Molecular characterization of an earliest cacao (*Theobroma cacao* L.) collection from upper Amazon using microsatellite DNA markers. Tree Genet. Genomes, **5**(4): 595–607. doi:10.1007/s11295-009-0212-2.

Zhang, D., Martínez, W.J., Johnson, E.S., Somarriba, E., Phillips-Mora, W., Astorga, C., et al. 2012. Genetic diversity and spatial structure in a new distinct *Theobroma cacao* L. population in Bolivia. Genet. Resour. Crop Evol. **59**(2): 239–252. doi:10.1007/s10722-011-9680-y.