


ORIGINAL RESEARCH ARTICLE

Plant Genetic Resources

American oil palm from Brazil: Genetic diversity, population structure, and core collection

Valquíria Martins Pereira^{1,*} | Jaire Alves Ferreira Filho^{1,*} | André Pereira Leão² |
 Luiz Henrique Galli Vargas¹ | Marcelo Picanço de Farias¹ | Sara de Almeida Rios³ |
 Raimundo Nonato Vieira da Cunha³ | Eduardo Fernandes Formighieri² |
 Alexandre Alonso Alves² | Manoel Teixeira Souza Júnior^{1,2} 

¹ PGBV, Univ. Federal de Lavras (UFLA), CEP 37200-000, Lavras, Minas Gerais, Brazil

² Embrapa Amazônia Ocidental, CEP 69010-970, Manaus, Amazonas, Brazil

³ Embrapa Agroenergia, CEP 70770-901, Brasília, Federal District, Brazil

Correspondence

Manoel Teixeira Souza Júnior, Embrapa Agroenergia, CEP 70770-901, Brasília, Federal District, Brazil.
 Email: manoel.souza@embrapa.br

Funding information

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The grant (01.13.0315.02 DendéPalm Project) for this study was awarded by the Brazilian Ministry of Science, Technology, and Innovation (MCTI) via the Brazilian Innovation Agency FINEP. The authors confirm that the funder had no influence over the study design, the content of article, or selection of this journal.

*These authors contributed equally to this work.

Assigned to Associate Editor Valerio Hoyos-Villegas.

Abstract

The American oil palm [*Elaeis oleifera* (Knuth) Cortés] has pronounced importance in oil palm breeding programs. Here, a germplasm bank (GB) of *E. oleifera* plants collected in the Amazon rainforest in Brazil was submitted to single nucleotide polymorphism (SNP) marker identification, selection, and use, aiming to characterize genetic diversity and population structure and to design a core collection (CC). Five hundred and fifty-three plants from 206 subsamples, collected at 19 localities spread throughout six geographic regions, were submitted to genotyping-by-sequencing analysis. A set of 1,827 high-quality SNP markers was then selected and used to run the genetic diversity and population structure analysis. The genetic diversity found is of moderate degree, and probably only a small portion of the species diversity is represented in the collection. The possible reason for that is the collecting strategy used, which collected subsamples only around the most prominent watercourses in the region. The average degree of genetic differentiation among subsamples is very high, indicating the presence of high interpopulation differentiation. The collection showed a low level of endogamy. The low average gene flow found indicates that genetic isolation caused by drift is occurring, and there is a need to review the conservation strategy. A set of 245 SNPs distributed throughout all 16 chromosomes was used to design CC based on maximizing the strategy of diversity. The optimal adjustment of the validated parameters, maintained while taking fewest subsamples, led to the choice of a model containing 20% of the entire collection as the ideal to form the CC.

Abbreviations: AMOVA, analysis of molecular variance; CC, core collection; GB, germplasm bank; GBS, genotyping-by-sequencing; PCoA, principal coordinate analysis; PIC, polymorphic information content; RAPD, random amplified polymorphic DNA; SNP, single nucleotide polymorphism; SSR, simple sequence repeat; UPGMA, unweighted pair group method with arithmetic mean; WC, whole germplasm collection.

1 | INTRODUCTION

The American oil palm [*Elaeis oleifera* (Kunth) Cortés], popularly known in Brazil as caiaué, is widely distributed in the central and northern regions of South America, which is the region from which it originated (Barcelos et al., 2015; Ooi, Barcelos, Muller, & Nascimento, 1981). It spreads widely in the Brazilian Amazon rainforest, located in the northwestern part of the country. This species has strategic importance to oil palm breeding programs in Brazil and elsewhere, as it has many desired agronomic traits to be incorporated in the genome of the African oil palm (*E. guineensis* Jacq.) (Barcelos et al., 2015). Among the most desired traits, one can list slow growth, resistance to some oil palm diseases, and better oil quality in terms of low acidity (España et al., 2018; Rios, da Cunha, Lopes, & da Silva, 2012).

In Brazil, a collection of *E. oleifera* native germplasm was established based on a series of expeditions in the Brazilian Amazon rainforest organized by Embrapa and the Center de Coopération Internationale at Rechecher Agronomique pour le Développement (CIRAD) in the early 1980s. Since then, Embrapa studies the extension and organization of the genetic diversity sampled to establish this germplasm bank (GB). This type of knowledge is necessary to define strategies for the breeding program for the development of superior American oil palm genotypes, as well as of interspecific hybrids between this species and the African oil palm (Barcelos et al., 2015).

Previous studies using molecular markers have indicated that the genetic diversity of the Brazilian GB of caiaué is more restricted than the one found in the African oil palm. However, these studies based on a limited amount of molecular markers (14 polymorphic loci out of 11 enzymatic systems, 37 restriction fragment length polymorphisms [RFLPs], 96 random amplified polymorphic DNA markers [RAPDs], and 19 simple sequence repeats [SSRs], respectively) (Arias, González, Prada, et al., 2015; Barcelos, 1998; Ghesquière, Barcelos, Santos, & Amblard, 1987; Moretzsohn et al., 2002). More recently, new studies have applied microsatellite markers (SSRs) to study genetic and phenotypic diversity and population structure, and even to establish core collection (CC) of the American and African oil palms (Arias, González, Prada, et al., 2015; Arias, González, & Romero, 2015; Bakoumé et al., 2015; Chee Chun et al., 2018; Khomphet, Eksomtramage, & Duangpan, 2018; Natawijaya et al., 2019; Zhou, Xiao, Xia, & Yang, 2015).

A CC is the result of a selection process that produces a subset of accessions from a GB, containing as much of the genetic diversity of that germplasm collection with minimal redundancy (Brown & Spillane, 1999; Odong, Jansen,

van Eeuwijk, & van Hintum, 2013). When using molecular data in the development of a CC, the strategy most preferably used is to maximize diversity while reducing the size of the subset of accessions to be maintained in the field for a long period (Odong et al., 2013; Xu et al., 2016). Producing a CC is a logistic and strategic decision to reduce the maintenance cost of germplasm and to delimit more representative subsamples of interest in the breeding program for characterization purposes. Arias, González, Prada, et al. (2015) and Arias, González, and Romero (2015) applied microsatellite markers to generate CCs for *E. oleifera* and *E. guineensis*, respectively.

Nowadays, new generation sequencing (NGS) technologies are available that allow the genotyping of an individual using thousands of molecular markers distributed throughout the entire genome at low cost (Chung, Choi, Jun, & Changsoo, 2017; Poland & Rife, 2012). Here, we report a genotyping-by-sequencing (GBS) study using 206 out of the 246 subsamples that make up the entire *E. oleifera* GB maintained by Embrapa (Rios et al., 2012); each subsample represents a half-sibling family. This study was performed based on the GBS technology of the DArTSeq platform (www.diversityarrays.com). Single nucleotide polymorphism (SNP) markers were identified, selected, and applied to characterize the genetic diversity and population structure, and to design a CC.

2 | MATERIALS AND METHODS

2.1 | Plant material and genomic DNA extraction

In this study, we used plants from the Brazilian *E. oleifera* GB at the Rio Urubu Experimental Station–Embrapa Western Amazon, located 140 km from Manaus, in the municipality of Rio Preto da Eva, Amazonas, Brazil (2°35' S, 59°28' W; 200 m asl). The climate of this municipality is tropical, and according to the classification of Köppen and Geiger, it is Af–tropical rainforest climate (<https://en.climate-data.org/>). The average temperature is 27.3 °C, with a maximum average of 28.2 °C in September and a minimum of 26.9 °C in the first quarter of the year. The annual average relative humidity is 85%, and the total insolation is 1,940 h. The average rainfall is ~2,300 mm yr⁻¹. According to the Brazilian classification of soils, the soil in this area is a yellow Latosol with a very clay texture 2:1 (dos Santos et al., 2011).

We sampled 553 plants from 206 different subsamples (half-sibling families) collected from 19 distinct localities spread throughout six distinct geographic regions in the states of Amazonas and Roraima (Figure 1, Supplemental Table S1). Each subsample had between one and

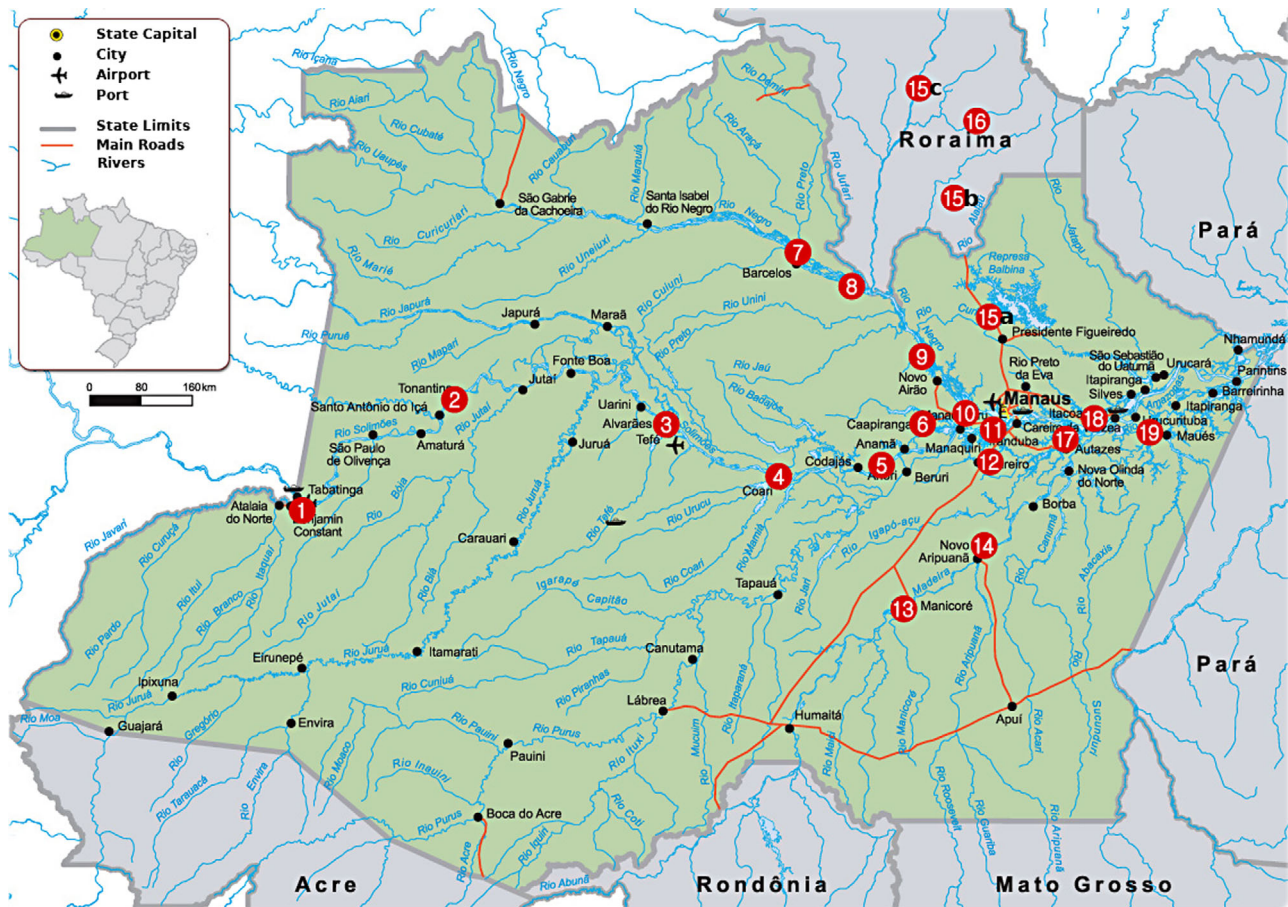


FIGURE 1 Geographic distribution of the localities where all 553 plants of *Elaeis oleifera*, from the germplasm bank (GB) at the Embrapa Western Amazon Research Institute, were collected. Plants were originally collected at six distinct geographic regions in the Brazilian Amazon forest, throughout 19 localities and 59 micro-localities. Regions (localities): Rio Solimões (1 = Benjamin Constant, 2 = Tonantins, 3 = Tefé, 4 = Coari, 5 = Anori, 6 = Manacapuru), Rio Negro (7 = Barcelos, 8 = Acajatuba, 9 = Moura), Manaus (10 = Caldeirão, 11 = Iranduba, 12 = Careiro), Rio Madeira (13 = Manicoré, 14 = Novo Aripuanã), Caracarái (15a = BR174-KM157, 15b = BR174-KM365, 15c = BR174-KM490/500, 16 = Perimetral Norte-Vila Moderna), and Rio Amazonas (17 = Autazes, 18 = Amajari, and 19 = Maués) Map Source: www.difusora24h.com.

three plants sampled. Fresh leaves collected from individual trees in the field were stored at -80°C until DNA extraction. Total DNA was extracted according to a modified cetyl-trimethyl ammonium bromide (CTAB) protocol (Doyle & Doyle, 1990). The quality and quantity of the extracted DNA were checked using NanoDrop (Thermo Scientific) and further confirmed on 0.8% agarose gel run in TBE (Tris-borate-EDTA) buffer at 80 V. Samples with a $260/280\text{ nm} \geq 1.8$ were treated with RNase, diluted to equal concentration ($100\text{ ng }\mu\text{l}^{-1}$), and submitted for genotyping by the DArTseq method to DArT.

2.2 | Genotyping by DArTseq technology

Genotyping-by-sequencing analysis of the 553 *E. oleifera* plants was performed using DArTseq Technology (Sansa-

loni, 2012). Genomic representation of the set of samples was generated by digesting the genomic DNA with a combination of two restriction enzymes, BstNI (CCWGG) and PstI (CTGCAG), and ligating PstI barcoded adapters to identify each sample. Equimolar amounts of amplification products from each sample were pooled by plate and amplified by c-Bot (Illumina) bridge polymerase chain reaction (PCR), followed by fragment sequencing on Illumina HiSeq 2000 (www.illumina.com), single-end 100 bp. The PstI adaptor includes a sequencing primer so that all generated tags were always read from the PstI restriction site.

The resulting sequences were filtered and allocated to their respective datasets, the barcodes were removed, and the sequences were trimmed at 69 bp (5 bp restriction site plus 64 bases with a minimum Q score of 10). Virtually identical reads (i.e., less than three polymorphisms) were

combined so that one or more SNPs in the read did not confuse the analysis. In parallel, a low coverage consensus sequence was generated to be used as a reference in the discovery of SNPs by aligning the 69-bp reads using the Bowtie program version 0.12 (Langmead, Trapnell, Pop, & Salzberg, 2009).

2.3 | Genetic diversity and population structure analysis

A group of 5,511 high-quality SNP markers generated by the DARt pipeline was filtered based on call rate (>0.90) and minor allele frequency (>0.05), and the number of markers reduced to 1,827. Then, this set of 1,827 SNP markers used to run the genetic diversity and population structure analysis.

DARwin software version 6.0.14 (Perrier & Jacquemoud-Collet, 2006) was used to calculate the pairwise dissimilarity coefficient matrix from allelic data, applying the simple matching index (Schlee, 1975), 80% of minimal proportion valid data for each unit pair, 1,000 bootstraps, and pairwise allele deletion. The pairwise dissimilarity coefficient matrix generated was used to perform a principal coordinate analysis (PCoA), and to construct a hierarchical clustering tree, also using DARwin 6.0.14.

Descriptive statistics analysis related to population diversity performed using GenAlEx software version 6.5 (Peakall & Smouse, 2012). Analysis of molecular variance (AMOVA) (Excoffier, Smouse, & Quattro, 1992), using the software GENES (Cruz, 2013), performed on the partition of total genetic variation between regions, and localities within or from different regions. The significance levels for the components of variance calculated using the criterion of 1,000 permutations.

Structure software version 2.3.4 (Pritchard, Stephens, & Donnelly, 2000) used to analyze population structure, based on the Bayesian model. The population structure classified in clusters (K), following its genetic similarities. The number of K tested varied from 1 to 10, with 10 interactions each (Evanno, Regnaut, & Goudet, 2005; Falush, Stephens, & Pritchard, 2003; Pritchard et al., 2000), using the Bayesian hierarchical clustering method, based on the hierarchical Bayesian clustering algorithm, thus allowing to build a hierarchy of partitions (dendrogram) based on the criterion of the maximum a posteriori probability of each partition (Iwayama & Tokunaga, 1995). The number of steps for the burning length was 10,000, and Markov chain Monte Carlo (MCMC) was 100,000. To define the number of genetic groups (most probable K), we used the criteria proposed by Evanno et al. (2005), using Structure Harvester software version 0.6.93 (Earl & vonHoldt, 2012).

2.4 | Development and evaluation of core collections

The 1,827 previously selected SNPs were submitted to analysis by the PowerCore software (Kim et al., 2007), and 500 SNP markers were selected accordingly to the highest Shannon diversity index generated. After aligning the 500 SNPs against the *E. guineensis* reference genome (Singh et al., 2013) using the BLASTN tool (standard parameters), we filtered those present only in regions unique to the genome, and of at most two gaps per sequence. A group of 245 SNPs was then obtained and used for the design of nuclear collections for *E. oleifera*.

For the development of CCs, both MSTRAT (Gouesnard et al., 2001) and PowerCore (Kim et al., 2007) software were applied. One hundred independent replicates and 100 interactions tested for each model generated when using MSTRAT. The composition of the chosen model was selected based on the highest Shannon diversity index. PowerCore used with a heuristic and a random search (Kim et al., 2007).

To calculate the genetic parameters and to assess separately the level of diversity captured in the entire germplasm collection, as well as for each one of the CC models evaluated, we applied PowerMaker software version 3.25 (Liu & Muse, 2005). The genetic parameters used to evaluate the CCs were the total number of alleles (N_A), polymorphic information content (PIC), observed heterozygosity (H_O), expected heterozygosity (H_E), and Shannon diversity index (Sh).

The genetic structure was evaluated using GenAlEx 6.5 (Peakall & Smouse, 2012), using the PCoA calculated from Nei's genetic distance matrix (Nei, 1972) of all 553 plants (representing the entire collection), in comparison with each one of the CC models generated. To compare the diversity between all CC models, and within them as well, an AMOVA analysis was performed using GenAlEx 6.5 and was calculated using a PhiPT (analogue of F_{st} , fixation index) with 999 permutations (Supplemental Table S2).

3 | RESULTS

3.1 | Genetic diversity and population structure analysis

The AMOVA results show that the genetic variability in the Brazilian GB of *E. oleifera* occurs mostly among individuals in a locality (60.7874%), followed by regions (21.2267%) and localities and region (17.9859%) (Table 1). The average dissimilarity estimated was 0.271, ranging from 0.020 to 0.367. The highest one was seen among two individuals from Rio Madeira (Manicoré-Manicoré) and Caracarái

TABLE 1 Analysis of molecular variance (AMOVA) for 553 *Elaeis oleifera* individuals, representing 19 localities grouped in six geographic regions from the Brazilian Amazon rainforest, through 1,827 single nucleotide polymorphism (SNP) markers

Source	df	SS	MS	Estimated variation	Variation %	P
Regions	5	3,103,784.0402	620,756.808	4,651.1743	21.2267	**
Localities/regions	13	845,152.8644	65,011.7588	3,941.0528	17.9859	**
Individuals/localities	534	7,112,700.7047	13,319.6642	13,319.6642	60.7874	**
Total	552	11,061,637.6094	20,039.1986	21,911.8914	100.0	–

Note. MS, medium square; SS, sum of squares.

**Significant at the .01 probability level.

(BR174-KM490), whereas the lowest was between two individuals from Rio Amazonas—both from Autazes-Quirimiri A (data not shown). The Rio Madeira region, with the highest number of individuals (159), was the one with the highest average diversity (2,179,577.8679). The Caracarái region, on the other hand, had the lowest diversity (625,635.5849), probably because it has the smallest number of individuals (53). The Rio Negro, Manaus, Rio Amazonas, and Rio Solimões regions had average diversity of 765,831.5079, 1,392,968.0421, 1,432,145.0989, and 1,561,695.4674, respectively; and 63, 95, 91, and 92 individuals, respectively.

Barcelos, with only two individuals, is the location with the lower average diversity (398), followed by Manacapuru (10,841.3333) and Benjamin Constant (13,867.6667), both with three individuals. The average diversity for the remaining localities was 58,759.6667 for Transcal (six individuals), 99,968.5556 for Anori (nine), 160.180 for Perimetral Norte (18), 171,345.75 for Tonantins (12), 177,258.9444 for Caldeirão (18), 205,353.4286 for Tefé (14), 261,127.1053 for Nova Aripuanã (19), 321,509.7241 for Acajatuba (29), 356,315.375 for Moura (32), 414,190.1071 for Autazes (288), 427,082.0968 for Amatari (31), 439,854.6286 for BR174 (35), 472,611.5938 for Maués (32), 762,932.9259 for Coari (54), 943,159.1029 for Careiro (68), and 1,808,944.7 for Manicoré (140).

In the graphical representation of the PCoA, shown in Figure 2, one can see that there is a higher consistency for grouping in the Rio Solimões, Rio Madeira, Rio Negro, and Caracarái regions, whereas the other two remaining geographic regions show a bigger dispersion pattern. The graphic dispersion of 206 subsamples (data not shown) and of 553 plants, obtained by PCoA method using the Jaccard's dissimilarity coefficient, presented the same grouping pattern, but with more consistency for the latter due to the bigger number of samples in the analysis. These results show that the groups have a strong relationship with the region from where their constituents were collected. The two first principal components account for

34.61% of the total variance for the 553 plants, being 25.17% for Axis 1 and 9.44 for Axis 2.

The hierarchical clustering tree was generated using the DARwin 6.0.14 software (unweighted pair group method with arithmetic mean [UPGMA], threshold equality of 0%), when using the set of 1,827 high-quality SNP markers, and divided these 553 plants into two major branches (Figure 3). The neighbor-joining method was applied to construct a radial unrooted tree, which is presented as Supplemental Figure S1. The Madeira/Solimões branch encompasses 137 individuals from Rio Madeira (86.16% of the total number of individuals from this region) and 90 from Rio Solimões (94.74%). The Rio Madeira region was represented by a total of 159 plants in this study, with 140 from the Manicoré and 19 from the Novo Aripuanã families. All but one of the 19 Novo Aripuanã plants did not assemble to the Madeira/Solimões branch; instead, it assembled in the Rio Amazonas sub-branch. On the other hand, the main branch encompasses all but six individuals from the remaining four geographic regions, together with almost all plants from Novo Aripuanã (Figure 3). All three plants from Manacapuru in the Rio Solimões region assembled in the main branch, more specifically in the Rio Negro sub-branch. All 53 plants from Caracarái assembled in the main branch; there was a sub-branch specific for this region composed by 47 out of the 53 plants (Figure 3). All but two of the 63 plants from the Rio Negro region assembled in the main branch. A sub-branch specific for the Rio Negro region can also be seen in Figure 3. Most of the 91 plants from Rio Amazonas assembled in a specific sub-branch for this region; however, all plants from Amarati-São Sebastião and Maués-Bom Jardim assembled in a different sub-branch of the main branch, together with most of the Manaus plants.

Results from the descriptive statistics analysis are shown in Table 2. The average number of alleles per loci (*A*) was 1.311. The lowest *A* was found in Benjamin Constant (1.130), in the Rio Solimões region, and Barcelos (1.155), in the Rio Negro region. The highest *A* was found

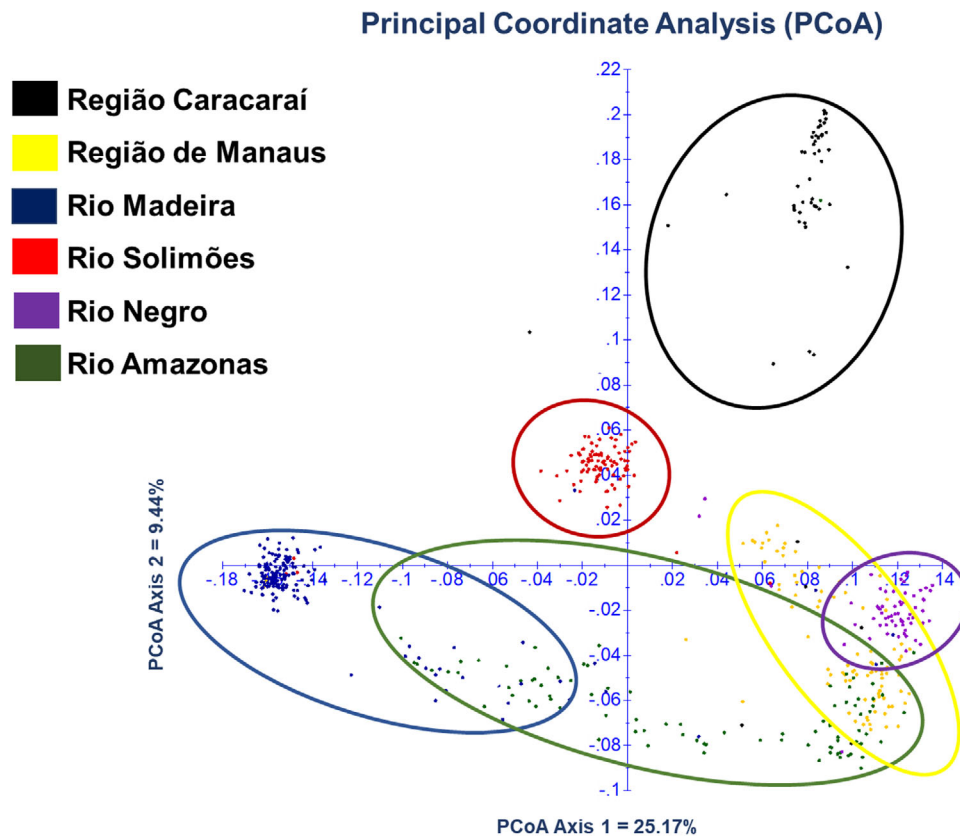


FIGURE 2 Principal components analysis (PCoA), based upon geographic regions, generated by means of the DARwin 6.0.14 software when using a set of 1,827 high-quality single nucleotide polymorphism (SNP) markers for 553 individuals from the Brazilian germplasm bank (GB) of *Elaeis oleifera*

in Nova Aripuanã (1.421), in the Rio Madeira region, and Maués (1.410), in the Rio Amazonas region. The observed heterozygosity (H_o), which shows the proportion of individuals that are heterozygous at a given locus, showed an average of 0.177, and localities Novo Aripuanã (0.237) and Anori (0.225) presented the populations with the highest H_o values. Caldeirão, Manacapuru, Anori, Benjamin Constant, and Barcelos show negative fixation indices (f), which can indicate excess of heterozygous individuals in the populations found in these localities (Table 2). The highest f was found in BR174 (0.270), from the Caracarái region. The average f for all 19 localities evaluated was 0.070, which shows an overall relatively low level of endogamy.

The average degree of genetic differentiation (F_{ST}) observed among the 19 subsamples was 0.326 (Table 3), which is considered very high and an indication of the presence of high interpopulation differentiation (Wright, 1978; Yeh, 2000). The average correlation between gametes that unite to produce the individuals (F_{IT}), relative to the gametes of the whole germplasm collection, was 0.372; although the average inbreeding coefficient (F_{IS}) estimated was 0.088 (Table 3). The F_{IS} and the fixation index (f), seen

in Table 2, showed similar behavior, showing once more that the GB of *E. oleifera* from the Amazon rainforest in Brazil has a generally low level of endogamy.

The results obtained using the Structure Harvester software show that the $K = 2$ and $K = 3$ models were the best ones to explain the population structure of the *E. oleifera* whole germplasm collection (Figure 4). The number of groups has varied between $K = 2$ and $K = 4$ when analyzing all 19 populations (Figure 5). The $K = 3$ model was accepted as the best one to explain the genetic structure of the *E. oleifera* populations. These results suggest strong genetic interpopulation differentiation.

3.2 | Designing core collection for *Elaeis oleifera* from the Brazilian rainforest

Out of the 500 SNP markers selected accordingly to the highest Shannon diversity index, only 245 (or 49%) mapped against the genome of *E. guineensis* and were consequently used for designing the CC. They were distributed throughout all 16 chromosomes, ranging from 5 to 31 SNPs per chromosome. The highest number of SNPs, 31, was

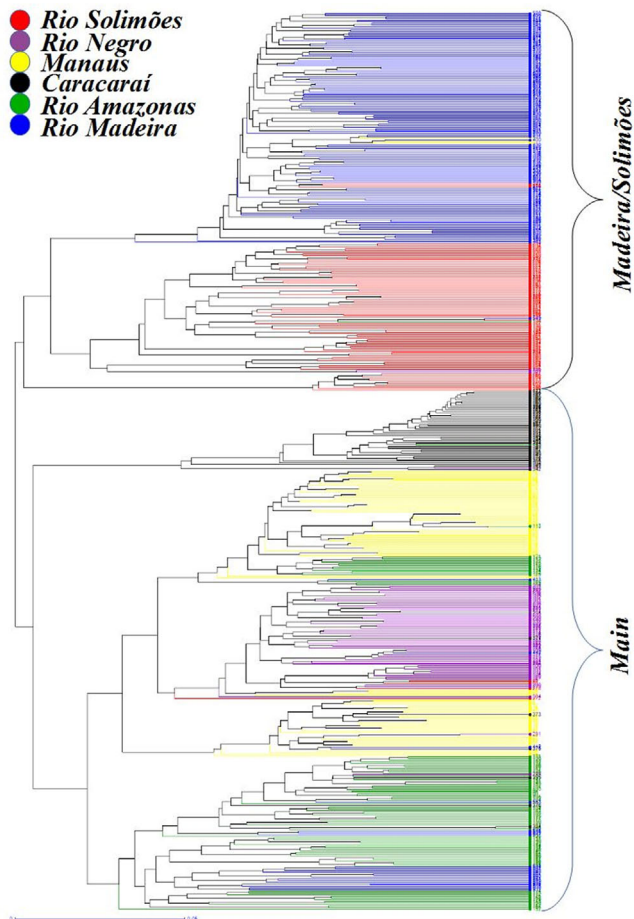


FIGURE 3 The hierarchical clustering tree generated by means of the DARwin 6.0.14 software (unweighted pair group method with arithmetic mean [UPGMA] method, threshold equality of 0%), when using a set of 1,827 high quality single nucleotide polymorphism (SNP) markers for 553 individuals from the Brazilian germplasm bank (GB) of *Elaeis oleifera*

mapped to chromosome 1, whereas the smallest number was mapped to chromosomes 15 and 16, five each (data not shown).

Six models of the CC were evaluated in this study (MS1, MS2, MS3, MS4, MS5, and MS6), containing 6.7, 10, 19.7, 23, 25, and 50% of the entire collection (WC) size, or 37, 55, 109, 127, 138, and 276 individuals, respectively. These models were generated by the MSTRAT software (Gouesnard et al., 2001), whereas the CCs established by the PowerCore software (Kim et al., 2007) were composed of 26 (PC1) and 16 (PC2) individuals (random and heuristic searches, respectively).

All models tested in this present study presented 100% of the alleles found in the WC (Table 4). The MS6 model was the only one to have a higher PIC value (0.356), whereas MS4 and MS5 had similar PIC values (0.355), and

the remaining models had lower values (between 0.351 and 0.354) than the one from the WC (0.355). All models presented expected heterozygosity (H_e) larger than its respective observed one (H_o). The PC1 model had the highest H_o value, 0.314, and MS5 had the lowest one, 0.273. The observed heterozygosity in all models was bigger than the one in the WC. On the contrary, only MS6 had expected heterozygosity larger than the one in the WC, 0.464 (Table 4). The Shannon's diversity index (H) for the entire collection was 0.693, whereas it ranged from 0.677 to 0.692 for the different models tested, increasing in value as the number of individuals per model increased (Table 4).

Regarding the six geographic regions where the *E. oleifera* germplasm was collected, all CC models tested did maintain at least one individual per region (data not shown). The MS3 model presented the following distribution of individuals by geographic region: Manaus (19%), Rio Amazonas (13%), Rio Solimões (20%), Rio Negro (13%), Caracarái (10%), and Rio Madeira (25%). Rio Madeira, the region with the most individuals in the WC, and Caracarái, the region with fewest, maintained this proportion in PC2, MS3, and MS6 models.

None of the individuals was present in all eight models tested, whereas four individuals were present in seven out of the eight models. However, 209 individuals were present only in one of the models tested. Only the PC1 model presented one individual per subsample; for the others, the number of subsamples with more than one individual ranged from 1 to 79 (data not shown).

The AMOVA results showed that all molecular variation (100%) is within models, and none was between models (Table 5). When considering each model, it ranged from 1.96 to 35.33%. The higher genetic variation belongs to the MS5 and MS6 models, with 17.63 and 35.33%, respectively, whereas PC1 presented the smallest one, 1.96%.

The PCoA results showed that all CC models tested were representative of what concerns the distribution of the entire collection (Figure 6). For the different clusters, there was an average of 46.17% for the genetic variation. Only the PC1 model did not present more than one individual per subsample, possibly due to the reduced size of this model (16 subsamples). For the other models, the variation in the number of subsamples with more than one individual was from 1 to 79. This result may indicate that there is considerable genetic variability within subsamples of the same population. Models with the smallest number of individuals, such as PC1, PC2, and MS1, were able to satisfactorily represent the distribution of the entire collection, with subsamples spread in all quadrants of the PCoA (Figure 6).

TABLE 2 Descriptive statistics of the regions and localities where the *Elaeis oleifera* plants were collected in the Brazilian Amazon rainforest: population size (N), number of alleles (N_a), number of alleles per loci (A), information index (I), observed heterozygosity (H_o), expected heterozygosity (H_e), and fixation index (f), and percentage of polymorphic loci ($\%P$) for the 19 localities

Geographic region	Locality	N	N_a	A	I	H_o	H_e	f	$\%P$
Manaus	Caldeirão	18	1.663 (0.011)	1.313 (0.009)	0.278 (0.006)	0.199 (0.005)	0.181 (0.005)	-0.074 (0.005)	66.28
	Careiro	68	1.962 (0.004)	1.357 (0.008)	0.351 (0.005)	0.181 (0.004)	0.211 (0.004)	0.162 (0.006)	96.22
	Irاندuba	6	1.536 (0.012)	1.276 (0.008)	0.258 (0.006)	0.167 (0.005)	0.168 (0.004)	0.006 (0.10)	53.59
Rio Amazonas	Amarati	31	1.851 (0.008)	1.367 (0.008)	0.348 (0.006)	0.186 (0.004)	0.223 (0.004)	0.107 (0.008)	85.06
	Autazes	68	1.896 (0.012)	1.363 (0.008)	0.357 (0.006)	0.173 (0.005)	0.226 (0.004)	0.205 (0.010)	89.60
	Maués	32	1.870 (0.008)	1.410 (0.008)	0.378 (0.006)	0.212 (0.004)	0.245 (0.004)	0.109 (0.006)	87.03
Rio Solimões	Anori	9	1.672 (0.011)	1.357 (0.008)	0.326 (0.006)	0.225 (0.005)	0.214 (0.004)	-0.056 (0.008)	67.21
	Benjamin Constant	3	1.245 (0.012)	1.130 (0.009)	0.153 (0.006)	0.129 (0.006)	0.101 (0.004)	-0.257 (0.008)	28.30
	Coari	54	1.863 (0.008)	1.386 (0.008)	0.356 (0.006)	0.199 (0.004)	0.231 (0.004)	0.105 (0.006)	86.26
Tonantins	Tefé	14	1.716 (0.011)	1.371 (0.008)	0.337 (0.006)	0.167 (0.004)	0.221 (0.004)	0.185 (0.009)	71.65
	Tonantins	12	1.776 (0.010)	1.366 (0.008)	0.347 (0.006)	0.175 (0.004)	0.224 (0.004)	0.158 (0.009)	77.56
	Manacaputu	3	1.313 (0.011)	1.210 (0.008)	0.179 (0.006)	0.185 (0.007)	0.121 (0.004)	-0.454 (0.009)	31.42
Rio Negro	Acajatuba	29	1.818 (0.009)	1.310 (0.008)	0.301 (0.006)	0.185 (0.005)	0.190 (0.004)	0.042 (0.006)	81.83
	Barcelos	2	1.221 (0.011)	1.155 (0.008)	0.144 (0.006)	0.141 (0.007)	0.099 (0.004)	-0.410 (0.011)	23.81
	Moura	32	1.741 (0.010)	1.292 (0.008)	0.280 (0.006)	0.162 (0.004)	0.178 (0.004)	0.045 (0.005)	74.06
Caracará	BRI74	35	1.84 (0.008)	1.261 (0.007)	0.275 (0.005)	0.110 (0.003)	0.168 (0.004)	0.270 (0.009)	84.62
	Vila Moderna	18	1.799 (0.009)	1.215 (0.006)	0.249 (0.005)	0.140 (0.004)	0.147 (0.004)	0.052 (0.005)	79.91
Rio Madeira	Manicoré	140	1.975 (0.004)	1.345 (0.008)	0.335 (0.006)	0.194 (0.004)	0.212 (0.004)	0.109 (0.005)	97.54
	Novo Aripuanã	19	1.848 (0.008)	1.421 (0.009)	0.379 (0.006)	0.237 (0.005)	0.248 (0.004)	0.040 (0.007)	84.84
Total of individuals	533	1.716 (0.002)	1.311 (0.002)	0.296 (0.001)	0.177 (0.001)	0.190 (0.001)	0.070 (0.002)	-	

Note. Values in parentheses are estimates of standard deviation.

TABLE 3 Wright's F statistics and the number of migrants per generation (N_m) for 19 localities in the Brazilian Amazon rainforest where *Elaeis oleifera* were collected

Parameter ^a	Avg.	SD
F_{ST}	0.326	0.004
F_{IS}	0.088	0.004
F_{IT}	0.372	0.005
N_m	0.801	0.030

^a F_{ST} , Degree of genetic differentiation; F_{IS} , inbreeding coefficient; F_{IT} , the correlation between gametes that unite to produce the individuals, relative to the gametes of the total population; N_m , gene flow.

4 | DISCUSSION

4.1 | Genetic diversity and population structure of the Brazilian *Elaeis oleifera* germplasm bank

Although it is not produced commercially, the American oil palm has pronounced importance to oil palm breeding programs in Brazil and elsewhere (Rios et al., 2012). Accordingly to Barcelos, Amblard, Berthaud, and Seguin (2002), there are four distinct genetic groups of *E. oleifera*: Brazil, French Guyana/Surinam, Peru, and the north of Colombia/Central America. There are already a few interspecific hybrids between the commercially available African and American oil palm and the Brazilian genetic group of *E. oleifera* is parental to most of them—Manicoré (BRS Manicoré from Embrapa, and [Mangenot × Manicoré] × La Mé from PalmElit SAS), Manaus (Amazon from ASD Costa Rica), and Coari (Coari × La Mé, Coari × Yangambi) (Rios et al., 2012).

The average dissimilarity found in this study (0.271) is slightly lower than that previously found by Moretzsohn

et al. (2002) (0.325). There are a few differences between these two studies that we must consider. First, the amount and type of molecular marker used (96 RAPD vs. 1,827 SNP markers), the number of plants (individuals, 175 vs. 553), and the number of localities (populations, 13 vs. 19). In Moretzsohn et al. (2002), the caiaué accessions grouped into three main branches, based on the UPGMA method; the third (and biggest) branch (Group III) had the majority of the accessions of Rio Madeira (84%) and Rio Solimões (70%). In the present study, the Madeira/Solimões branch encompasses 86.16% of the total number of individuals from Rio Madeira and 94.74% of the total number of individuals from Rio Solimões. These differences in percentage between Moretzsohn et al. (2002) and our study, specifically regarding the Madeira/Solimões branch, are probably due to the difference in the number of individuals used. While they used 57 and 13 accessions from Rio Madeira and Rio Solimões, respectively, we used 159 and 95. Besides, our study had three localities from the Rio Solimões that were absent from theirs (Benjamin Constant, Coari, and Manacapuru), which together added more 58 plants to the Rio Solimões population.

When comparing the two other branches (Groups I and II) from Moretzsohn et al. (2002) and the main one from the present study, it is clear that Group II is equivalent to the Caracarái population. The main differences appeared when comparing the remaining sub-branches of the main branch with Group I (Moretzsohn et al., 2002). In the present study, the populations making up these sub-branches (Rio Negro, Manaus, and Rio Amazonas) were better separated from each other than in Group I (Moretzsohn et al., 2002). Besides the differences in the number of individuals and populations used in this study, we have to consider that the number of markers may have favored the identification of additional similarities between subsamples imperceptible when using a limited amount of them.

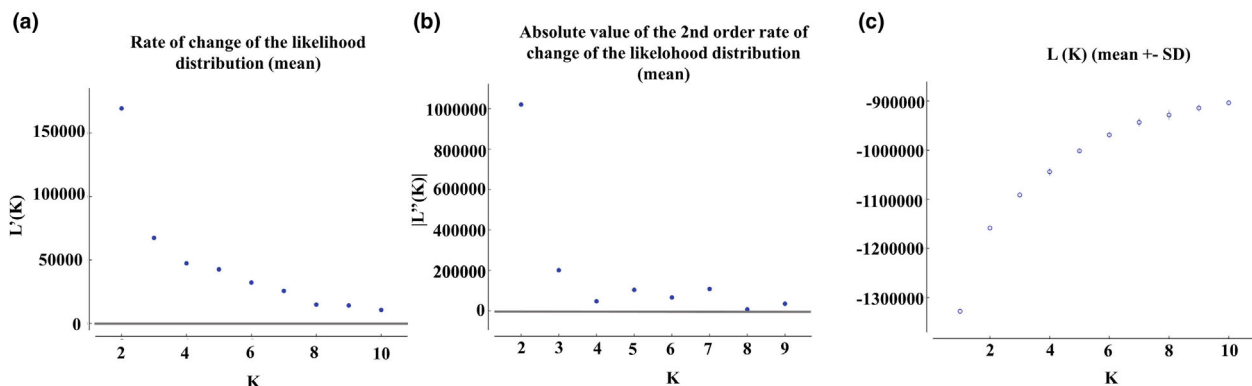


FIGURE 4 Average values for (a) rate of change of the likelihood distribution (mean \pm SD) calculated as $L'(K) = L(K) - L(K - 1)$, (b) absolute value of the second-order rate of change of the likelihood distribution (mean \pm SD) calculated as $|L''(K)| = |L'(K + 1) - L'(K)|$, and (c) likelihood distribution (mean \pm SD) with the most likely number of populations (K) for 10 repetitions of simulation using the Structure software, with the $K = 1$ –10 for the whole germplasm collection (WC)

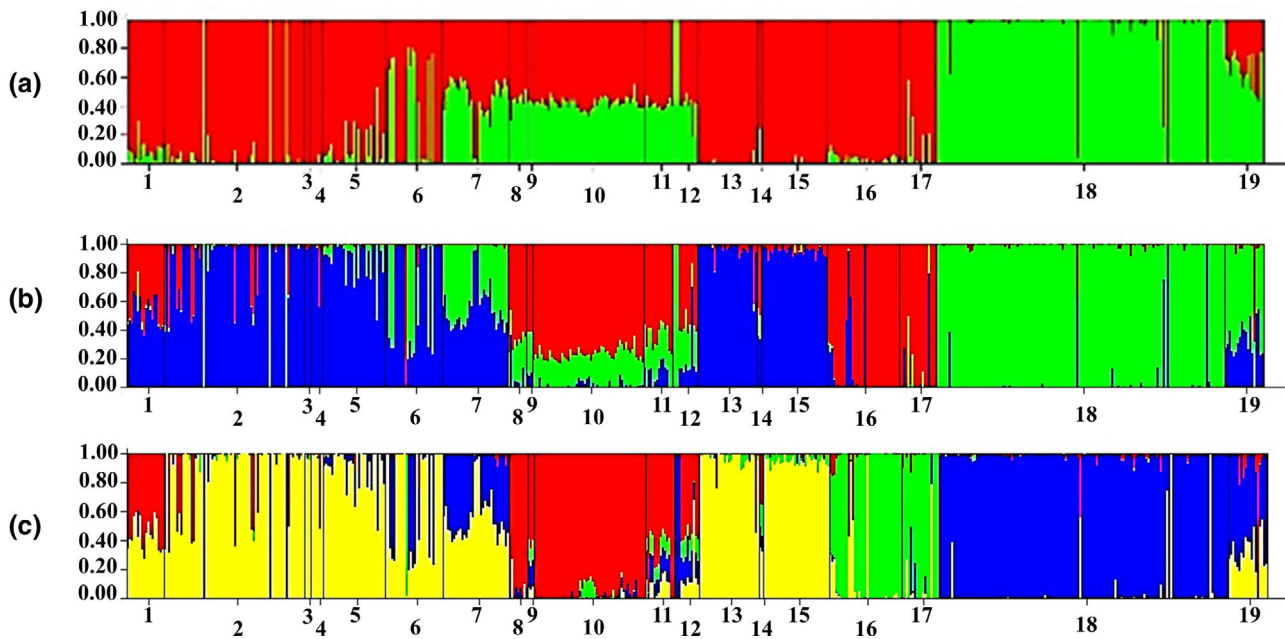


FIGURE 5 Genetic structure of *Elaeis oleifera* populations from 19 localities, spread throughout six regions in the Brazilian Amazon forest, using the Structure Harvester software, with (a) $K = 2$, (b) $K = 3$, and (c) $K = 4$ for the whole germplasm collection (WC)

Our results show that the genetic diversity in the *E. oleifera* is of moderate degree, and it corroborates with previous results found by Barcelos et al. (2002) who, when evaluating genetic diversity in populations of the species, identified four major groups also related according to geographic region: Brazil, French Guyana/Surinam, Peru, and the north of Colombia/Central America. Similar results were found by Araya, Alvarado, and Escobar (2011) when characterizing the genetic diversity of the *E. oleifera* GB from the Agricultural Services & Development (ASD) in Costa Rica using eight microsatellite markers. Rios et al. (2012), based on the study of phenotypic characteristics, also reported the existence of variability among the subsamples of *E. oleifera* in the GB from Brazil. The present study focused on the discovery of the extent of the genetic diversity of this group of individuals from Brazil.

Independent of the higher consistency for grouping only to four out of six regions (Rio Solimões, Rio Madeira, Rio Negro, and Caracará), it was possible to separate all six regions (Figure 2). These results contrast with those obtained by Moretzsohn et al. (2002) and may be related to two factors: (a) the sampling amplitude (comprising 206 subsamples belonging to six distinct regions and thus a sampling of 84% of the entire GB) and (b) the use of a higher marker density allows us to sample previously untapped regions of the genome, contributing to different estimates of dissimilarity.

The plants used in this study, which make up the *E. oleifera* GB from Brazil, come originally from a few seeds collected from a small number of matrices. The found-

ing effect may explain the grouping pattern seen in Figure 3, and associated with it the phenomenon of gene flow through four main watercourses (Rio Solimões, Rio Negro, Rio Madeira, and Rio Amazonas), one long highway (Caracará region), and a region where three of the watercourses join resulting in a new water course (Manaus). These results also indicate that probably only a small portion of the species diversity is represented in the GB as a result of the collecting strategy used, which collected subsamples only in the regions of the most prominent watercourses in the region. To increase the diversity of the collection and making use of these data, one strategy would be to collect new subsamples in distant regions of the rivers (less subject to anthropic and founder effect), prioritizing the sampling of a large number of subsamples and individuals from these subsamples (Escobar, 1981).

According to Mitton (2013), population structure is the pattern of genetic variation among populations and is produced by the joint action of gene flow, genetic drift, and natural selection. The population genetic structure analysis has been based on principles underlying Wright's F statistics (Wright, 1978). The fixation index (f) is one of the most important parameters in population genetics, as it measures the balance between homozygous and heterozygous in populations. The average f found in this study to all 19 localities evaluated was 0.070, which shows an overall low level of endogamy (Table 2). The explanation for populations that contain more loci in homozygosis and fewer in heterozygosis must be associated with the reproductive system and/or genetic drift. The fact that entomophilous

TABLE 4 Genetic parameters evaluated for the whole germplasm collection (WC) and for the different core collection models of *Elaeis oleifera* from the Brazilian Amazon rainforest: PC1 (2.9% of the WC), PC2 (4.7%), MS1 (6.7%), MS2 (10%), MS3 (19.7%), MS4 (23%), MS5 (25%), and MS6 (50%)

Parameter ^a	WC	PC1	PC2	MS1	MS2	MS3	MS4	MS5	MS6
N_i	533	16	26	37	55	109	127	138	276
N_s	206	16	25	35	52	86	105	108	197
N_A	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
PIC	0.355	0.351	0.351	0.352	0.352	0.354	0.355	0.355	0.356
H_o	0.271	0.314	0.289	0.275	0.277	0.274	0.280	0.273	0.282
H_e	0.463	0.457	0.456	0.458	0.459	0.461	0.463	0.462	0.464
H	0.693	0.677	0.683	0.687	0.689	0.691	0.691	0.691	0.692

^a N_i , number of individuals; N_s , number of subsamples; N_A , total number of alleles; PIC, polymorphic informative content; H_o , observed heterozygosity; H_e , expected heterozygosity; H , Shannon's diversity index.

TABLE 5 Analysis of molecular variance (AMOVA) for different models of core collection of *Elaeis oleifera* from the Brazilian Amazon rainforest

Source	df	SS	MS	Variation %
Between models	7	1,703,687	234.384	0.0
Within models	776	194,324,048	250.418	100
PC1	15	3,812,438	254.163	1.96
PC2	25	6,286,346	251.454	3.23
MS1	36	9,004,324	250.120	4.63
MS2	54	13,519,527	250.362	6.96
MS3	108	27,052,679	250.488	13.92
MS4	126	31,729,339	251.820	16.33
MS5	137	34,262,884	250.094	17.63
MS6	275	68,656,511	249.660	35.33
TOTAL	783	196,027,735	–	100.00

Note. MS, medium square; SS, sum of squares.

pollination is the most important method of pollination in *Elaeis* spp. (Meléndez & Ponce, 2016) could explain, in part, the overall relatively low level of endogamy found in this study, whereas population size and/or genetic drift could explain the negative fixation index found in some subsamples (Table 2).

A high gene flow between populations will make them indistinguishable, meaning that if gene flow were unopposed by other forces, the populations connected by gene flow would ultimately share the same alleles at the same frequencies (Mitton, 2013). According to Wright (1978), when the average gene flow value is less than one, it shows genetic isolation; the average gene flow estimated for the *E. oleifera* GB from Brazil was 0.801, which is considered low (Table 3). This value may be an indication that genetic isolation caused by drift is occurring, and there is a need to review the conservation strategy for this species. The gene flow value calculated from the genetic divergence reflects

the gene flow that occurred over a long period (Smouse & Sork, 2004). The estimate does not indicate whether gene flow is occurring in a given reproductive event but calculates the levels of gene flow that must have occurred to produce the observed patterns of genetic structure. Natural selection can oppose the homogenizing effect of gene flow, sustaining genetic differences among populations linked by gene flow (Mitton, 2013). Genetic drift and gene flow, besides selection, can change allele frequencies in a population; it is an error to assume selection will always drive populations toward the most well-adapted state (Andrews, 2010; Sork, 2015).

The results obtained using Structure Harvester software suggest strong genetic interpopulation differentiation among the 19 populations from the *E. oleifera* GB from Brazil (Figure 5). It is possible to notice that the native populations of *E. oleifera* from Brazil used in this study have a moderate diversity, a fact confirmed by the statistical

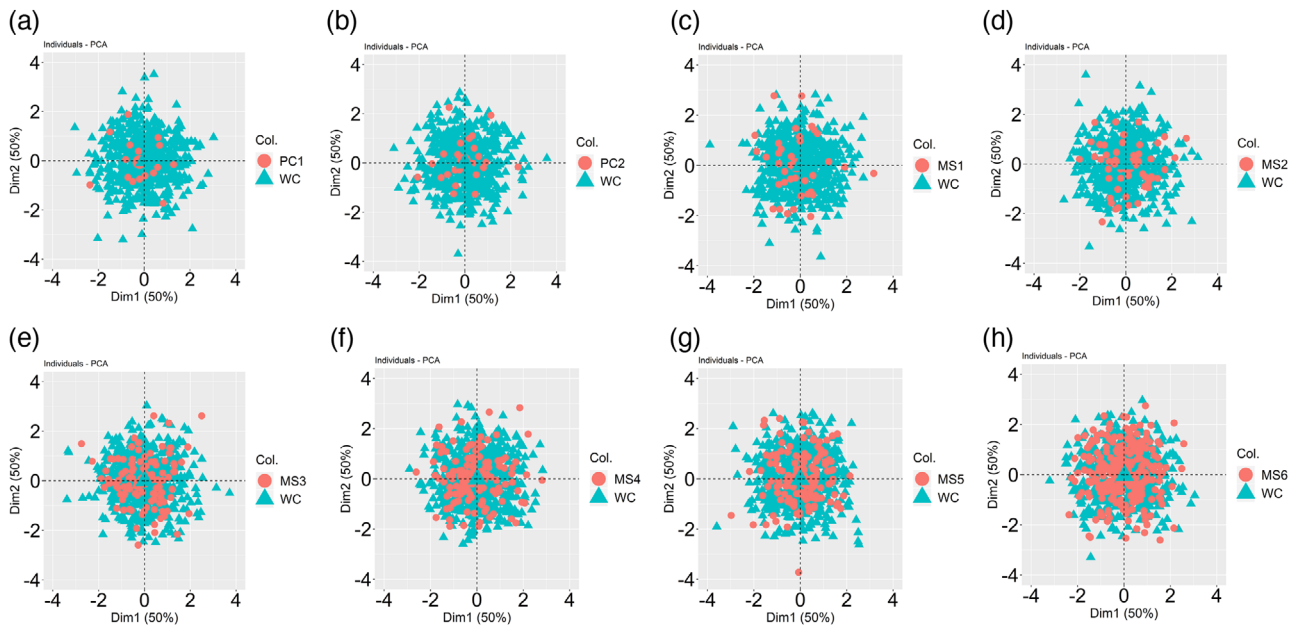


FIGURE 6 Principal coordinate analysis (PCoA) for the whole germplasm collection (WC) of *Elaeis oleifera* with each one of the eight core collection models tested (PC1 [4.7% of the WC], PC2 [2.9%], MS1 [6.7%], MS2 [10%], MS3 [19.7%], MS4 [23%], MS5 [25%] and MS6 [50%])

analyses of genetic diversity, and that these populations have a clear genetic structure as shown by Wright's F statistic and structured by Bayesian analysis, as well as show the results of effective population size (N_m). There is genetic variability between and within populations, which can be explored for conservation, commercial exploitation, and selection for application in genetic improvement of the species and provide a broad view of the structure as a whole to facilitate future applicability of breeders (Ithnin, Teh, & Ratnam, 2017; Natawijaya et al., 2019; Osorio-Guarín et al., 2019).

4.2 | Core collection for *Elaeis oleifera* from the Brazilian rainforest

For the genetic parameters evaluated in this study, there was little change from those presented by the eight models of nuclear collections tested to that obtained for the entire collection. Besides, all models maintained subsamples representing the six regions of the state of Amazonas from where the GB accessions were collected.

The genetic parameters evaluated did not vary drastically for the models generated by the methods of MSTRAT (Gouesnard et al., 2001) and PowerCore software (Kim et al., 2007) (Table 4). The M strategy (Shoen & Brown, 1993) can efficiently select subsamples and accessions that maximize the genetic diversity presented by the entire collection, and evidence suggests that it is capable of maintaining diversity genetics and allelic for the different CCs generated, when used with molecular markers (Marita,

Rodriguez, & Nienhuis, 2000). The implementation of this strategy by MSTRAT software allows the alleles to be selected interactively, maintaining the diversity by allelic richness criteria in the analysis. In contrast, PowerCore software uses a search using a heuristic algorithm that can represent all alleles identified by molecular markers and all classes of phenotypic observations in the development of CCs.

All eight CC models tested in this study maintained 100% of the SNPs alleles from the entire collection (Table 4). The M strategy (Shoen & Brown, 1993) is capable of maintaining diversity genetics and allelic for the different CCs generated, when used with molecular markers (Marita et al., 2000). However, maintaining the total number of alleles in a CC is not always possible depending on the species or the dataset. In wheat (*Triticum aestivum* L.), only 98% of the alleles remained in the nuclear collection (Balfourier et al., 2007). Belaj et al. (2012), using microsatellite markers for the construction of a CC in *Olea europaea* L., obtained an average reduction of 10% in the total amount of alleles in the entire collection for the five CC models generated.

The H_e values were higher than those obtained for the H_o index in all models tested (Table 4). This can be related to a high number of homozygotes as a result of the inbreeding effect in some populations. It can also be the result of a sampling effect in the entire collection.

The AMOVA results showing that all molecular variation is within and not between models was expected due to the close genetic composition between the models (Table 5). de Oliveira et al. (2014) obtained similar results in

cassava (*Manihot esculenta* Cranz) also using SNP markers, with 0.30 and 99.70% of the variation between and within the models, respectively.

The PCoA calculated from Nei's genetic distance matrix (Nei, 1972) for the WC, in comparison with each one of the models generated, showed that all of them were representative of the WC, showing an average explanation of 46.17% for the genetic variation; however, this value decreased as the number of subsamples increased in the tested model (Figure 6). The PCoA analysis is an exploratory method that assists in the elaboration of more concrete hypotheses based on the data collected, and it shows evidence that is related to the level of explanation for the existing genetic relationships (Mingoti, 2005). Our results showed that for all models plotted against the WC, most individuals are dispersed in the panel, representing subsamples in the four quadrants of the graphic dispersion (Figure 6).

In general, a CC can be 10% of the original size, which represents ~70% of the diversity of the original collection (Brown & Spillane, 1999). However, different percentages can be accepted due to fact that the stabilization of the CC depends on several factors, such as the size of the original collection, the quality of the data collected for the characterization, the assessment of the stratification of the original collection, and the sampling strategy used (Cochran, 1977). A good CC incorporates the maximum of the diversity of the species with a minimum of redundancy in the smallest possible size to facilitate the manipulation of it (Brown & Spillane, 1999; Odong et al., 2013).

Recently, Arias, González, Prada, et al. (2015) analyzed the phenotypic and genetic diversity of accessions of caiaué from Brazil, Peru, Ecuador, and Colombia to select those individuals who would form a CC representing the maximum allelic diversity. These authors used 37 plants from two regions in Brazil—Coari, and Manaus, which are represented in our—and demonstrated that 100% of the alleles and 100% of the genetic diversity found in that material could be included in a sample of roughly 34% of the initial collection. By using a much higher number of plants (553) and a much bigger group of markers (245 SNPs distributed throughout all 16 chromosomes), it is possible to see in the present study that an even smaller percentage of plants can generate the same results.

The *E. oleifera* GB has ~4,000 plants conserved in 29 ha at the Rio Urubu Experimental Station (Rios et al., 2012). The maintenance of such a GB in a field in the heart of the Amazon rainforest has several economical and logistic constraints, keeping the collection at constant risk. Despite the genetic diversity present in the caiaué (Barcelos, 1998; Ghesquière et al., 1987; Moretzsohn et al., 2002; and this study), its exploitation by the breeding program is still inef-

fective. Only 20% of the subsamples in the BAG underwent characterization, and most of the additional characterization and evaluation necessary are no longer possible considering the age of the plants.

Creating a caiaué CC in Brazil is relevant for both the scientific and technological development of the oil palm industry in Brazil. Designing a CC for this species would also reduce the cost of the necessary studies to be carried out by Embrapa and partners for this important and strategic native species. The MS3 model had a Shannon diversity index of 0.691, remaining very close to the one evaluated for the entire collection (0.693). Besides, the other genetic parameters evaluated showed little variation (Table 4). Thus, the MS3 model that retains 19.7% of the total individuals of the WC is the most appropriate choice to generate an *E. oleifera* CC at Embrapa. By maximizing genetic diversity and reducing the number of genotypes, the MS3 collection can facilitate studies of variability and correlation with morphological traits of agronomic interest to the breeding program developed at Embrapa. The MS3 model presented the following distribution of individuals by geographic region: Manaus (19%), Rio Amazonas (13%), Rio Solimões (20%), Rio Negro (13%), Caracarai (10%), and Rio Madeira (25%). However, it is known from such studies with different species that the genetic diversity of a CC can be better achieved when molecular data are used in conjunction with phenotypic data (Balfourier et al., 2007; Belaj et al., 2012).

There is a high similarity between the genomes of *E. oleifera* and *E. guineensis*, so it might be possible to transfer the SNPs developed in this study to assess the genetic diversity and establishment of core collection in *E. guineensis*. The presence of common alleles could occur due to the genetic proximity between these two species.

ACKNOWLEDGMENTS

The authors acknowledge funding to V.M.P., J.A.F.F., L.H.G.V., and M.P.F. from the Coordination for the Improvement of Higher Education Personnel (CAPES), a Foundation within the Ministry of Education in Brazil, via the Graduate Program in Plant Biotechnology, Federal University of Lavras (UFLA).

AUTHOR CONTRIBUTIONS

M.T.S. Jr., A.A.A., E.F.F., S.A.R., and R.N.V.C. conceived the experiment(s). V.M.P., J.A.F.F., A.P.L., L.H.G.V., and M.P.F. conducted the experiment(s). M.T.S. Jr., A.A.A., E.F.F., V.M.P., J.A.F.F., and A.P.L. analyzed the results. M.T.S. Jr., V.M.P., and J.A.F.F. wrote the manuscript. All authors reviewed the manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ORCID

Manoel Teixeira Souza Júnior  <https://orcid.org/0000-0002-6590-9333>

REFERENCES

- Andrews, C. A. (2010). Natural selection, genetic drift, and gene flow do not act in isolation in natural populations. *Nature Education Knowledge*, 3(10).
- Araya, E., Alvarado, A., & Escobar, R. (2011). *Use of DNA markers for fingerprinting compact clones and determining the genetic relationship between E. oleifera germoplasm origins*. San José: Agricultural Services & Development.
- Arias, D., González, M., Prada, F., Ayala-Díaz, I., Montoya, C., Daza, E., & Romero, H. M. (2015a). Genetic and phenotypic diversity of natural American oil palm (*Elaeis oleifera* (H.B.K.) Cortés) accessions. *Tree Genetics & Genomes*, 11, 1–13. <https://doi.org/10.1007/s11295-015-0946-y>
- Arias, D., González, M., & Romero, H. M. (2015). Genetic diversity and establishment of a core collection of oil palm (*Elaeis guineensis* Jacq.) based on molecular data. *Plant Genetic Resources: Characterization and Utilization*, 13, 256–265. <https://doi.org/10.1017/S1479262114001026>
- Bakoumé, C., Wickneswari, R., Siju, S., Rajanaidu, N., Kushairi, A., & Billotte, N. (2015). Genetic diversity of the world's largest oil palm (*Elaeis guineensis* Jacq.) field genebank accessions using microsatellite markers. *Genetic Resources and Crop Evolution*, 62, 349–360. <https://doi.org/10.1007/s10722-014-0156-8>
- Balfourier, F., Roussel, V., Strelchenko, P., Exbrayat-Vinson, F., Sourdille, P., Boutet, G., ... Charmet, G. (2007). A worldwide bread wheat core collection arrayed in a 384-well plate. *Theoretical and Applied Genetics*, 114, 1265–1275. <https://doi.org/10.1007/s00122-007-0517-1>
- Barcelos, E. (1998). *Étude de la diversité génétique du genre Elaeis (E. oleifera (Kunth) et E. guineensis Jacq.), par marqueurs moléculaires (RFLP et AFLP)*. (Ph.D. dissertation, École Nationale Supérieure Agronomique de Montpellier).
- Barcelos, E., Amblard, P., Berthaud, J., & Seguin, M. (2002). Genetic diversity and relationship in American and African oil palm as revealed by RFLP and AFLP molecular markers. *Pesquisa Agropecuária Brasileira*, 37, 1105–1114.
- Barcelos, E., de Almeida Rios, S., Cunha, R. N. V., Lopes, R., Motoike, S. Y., Babiychuk, E., ... Kushnir, S. (2015). Oil palm natural diversity and the potential for yield improvement. *Frontiers in Plant Science*, 6. <https://doi.org/10.3389/fpls.2015.00190>
- Belaj, A., Dominguez-García, M.d.C., Atienza, S. G., Urdiroz, N. M., De la Rosa, R., Satovic, Z., ... Del Río, C. (2012). Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genetics & Genomes*, 8, 365–378. <https://doi.org/10.1007/s11295-011-0447-6>
- Brown, A. H. D. & Spillane, C. (1999). Implementing core collections: Principles, procedures, progress, problems and promise. In R. C. Johnson & T. Hodgkin (Eds.), *Core collections for today and tomorrow* (pp. 1–10). Madison, WI: CSSA.
- Chee Chun, T., Chua Kia, L., Lee Chong, H., Cheah Suan, C., Jakim, B., Au Wai, F., ... Tan Soon, G. (2018). Genetic diversity and inbreeding level in deli dura and avros advanced breeding materials in oil palm (*Elaeis guineensis* Jacq.) using microsatellite markers. *Journal of Oil Palm Research*, 30, 366–379.
- Chung, Y. S., Choi, S. C., Jun, T., & Changsoo, K. (2017). Genotyping-by-sequencing: A promising tool for plant genetics research and breeding. *Horticulture, Environment and Biotechnology*, 58, 425–431. <https://doi.org/10.1007/s13580-017-0297-8>
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley and Sons.
- Cruz, C. D. (2013). Genes: A software package for analysis in experimental statistics and quantitative genetics. *Acta Scientiarum Agronomy*, 35, 271–276. <https://doi.org/10.4025/actasciagron.v35i3.21251>
- de Oliveira, E. J., Ferreira, C. F., da Silva Santos, V., de Jesus, O. N., Oliveira, G. A., & da Silva, M. S. (2014). Potential of SNP markers for the characterization of Brazilian cassava germplasm. *Theoretical and Applied Genetics*, 127, 1423–1440. <https://doi.org/10.1007/s00122-014-2309-8>
- dos Santos, H. G., de Carvalho Junior, W., Dart, R. de O., Aglio, M. L. D., de Sousa, J. S., Pares, J. G., ... de Oliveira, A. P. (2011). *O novo mapa de solos do Brasil: Legenda atualizada*. Rio de Janeiro, Brazil: Embrapa Solos.
- Doyle, J. J., & Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. *Focus*, 12, 13–15.
- Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4, 359–361. <https://doi.org/10.1007/s12686-011-9548-7>
- Escobar, R. (1981). Preliminary results of the collection and evaluation of the American oil palm (*Elaeis oleifera* H.B.K. Cortes) in Costa Rica, Panama and Colombia. In E. Pushparajah & P. Chew (Eds.), *The oil palm in agriculture in the eighties* (pp. 79–97). Kuala Lumpur: Incorporated Society of Planters.
- España, M. D., Mendonça, S., Carmona, P. A. O., Guimarães, M. B., da Cunha, R. N. V., & Souza, Jr., M. T. (2018). Chemical characterization of the American oil palm from the Brazilian Amazon forest. *Crop Science*, 58, 1982–1990. <https://doi.org/10.2135/cropsci2018.04.0231>
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, 14, 2611–2620. <https://doi.org/10.1111/j.1365-294x.2005.02553.x>
- Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, 131, 479–491.
- Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164, 1567–1587.
- Ghesquière, M., Barcelos, E., Santos, M. M., & Amblard, P. (1987). Polymorphisme enzymatique che *Elaeis oleifera* H. B. K. (*Elaeis melanococca*) analyse des populations du Bassin amazonien. *Oléagineux*, 42, 143–153.
- Gouesnard, B., Bataillon, T. M., Decoux, G., Rozale, C., Schoen, D. J., & David, J. L. (2001). MSTRAT: An algorithm for building

- germ plasm core collections by maximizing allelic or phenotypic richness. *Journal of Heredity*, 92, 93–94. <https://doi.org/10.1093/jhered/92.1.93>
- Ithnin, M., Teh, C., & Ratnam, W. (2017). Genetic diversity of *Elaeis oleifera* (HBK) Cortés populations using cross species SSRs: Implications for germplasm utilization and conservation. *BMC Genetics*, 18. <https://doi.org/10.1186/s12863-017-0505-7>
- Iwayama, M., & Tokunaga, T. (1995). Hierarchical Bayesian clustering for automatic text classification. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1322–1327). International Joint Conference on Artificial Intelligence.
- Khomphet, T., Eksomtramage, T., & Duangpan, S. (2018). Genetic variation of improved oil palm tenera hybrid populations using morphological and SSR markers. *Songklanakarinn Journal of Science and Technology*, 40, 1329–1335.
- Kim, K. W., Chung, H., Cho, G., Ma, K., Chandrabalan, D., Gwag, J., ... Park, Y. (2007). PowerCore: A program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics*, 23, 2155–2162. <https://doi.org/10.1093/bioinformatics/btm313>
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10. <https://doi.org/10.1186/gb-2009-10-3-r25>
- Liu, K., & Muse, S. V. (2005). PowerMarker: An integrated analysis environment for genetic marker analysis. *Bioinformatics*, 21, 2128–2129. <https://doi.org/10.1093/bioinformatics/bti282>
- Marita, J. M., Rodriguez, J. M., & Nienhuis, J. (2000). Development of an algorithm identifying maximally diverse core collections. *Genetic Resources and Crop Evolution*, 47, 515–526. <https://doi.org/10.1023/A:1008784610962>
- Meléndez, M. R., & Ponce, W. P. (2016). Pollination in the oil palms *Elaeis guineensis*, *E. oleifera* and their hybrids (OxG), in tropical America. *Pesquisa Agropecuária Tropical*, 3, 46–48. <http://doi.org/10.1590/1983-40632016v4638196>
- Mingoti, S. A. (2005). *Análise de Dados Através de Métodos de Estatística Multivariada: Uma abordagem Aplicada*. Belo Horizonte, Brazil: Editora UFMG.
- Mitton, J. B. (2013). Gene flow. In S. Maloy & K. Hughes (Eds.), *Brenner's encyclopedia of genetics* (2nd ed.). Cambridge, MA: Academic Press.
- Moretzsohn, M. C., Ferreira, M. A., Amaral, Z. P. S., Coelho, P. J. A., Grattapaglia, D., & Ferreira, M. E. (2002). Genetic diversity of Brazilian oil palm (*Elaeis oleifera*) germplasm collected in the Amazon forest. *Euphytica*, 124, 35–45.
- Natawijaya, A., Ardie, S. W., Syukur, M., Maskromo, I., Hartana, A., & Sudarsono, S. (2019). Genetic structure and diversity between and within African and American oil palm species based on microsatellite markers. *Biodiversitas*, 20, 1233–1240. <https://doi.org/10.13057/biodiv/d200501>
- Nei, M. (1972). Genetic distance between populations. *American Naturalist*, 106, 283–292.
- Odong, T. L., Jansen, J., van Eeuwijk, F. A., & van Hintum, T. J. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theoretical and Applied Genetics*, 126, 289–305. <https://doi.org/10.1007/s00122-012-1971-y>
- Ooi, S., Barcelos, E., Muller, A., & Nascimento, J. (1981). Oil palm genetic resources: Native *E. oleifera* populations in Brazil offer promising sources. *Pesquisa Agropecuária Brasileira*, 16, 385–395.
- Osorio-Guarín, J. A., Garzón-Martínez, G. A., Delgadillo-Duran, P., Bastidas, S., Moreno, L. P., Enciso-Rodríguez, F. E. ... Barrero, L. S. (2019). Genome-wide association study (GWAS) for morphological and yield-related traits in an oil palm hybrid (*Elaeis oleifera* × *Elaeis guineensis*) population. *BMC Plant Biology*, 19. <https://doi.org/10.1186/s12870-019-2153-8>
- Peakall, R., & Smouse, P. E. (2012). GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research: An update. *Bioinformatics*, 28, 2537–2539. <https://doi.org/10.1093/bioinformatics/bts460>
- Perrier, X., & Jacquemoud-Collet, J. P. (2006). DARwin software. CIRAD. Retrieved from <http://darwin.cirad.fr/>
- Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome*, 5, 92–102. <https://doi.org/10.3835/plantgenome2012.05.0005>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.
- Rios, S. A., da Cunha, R. N. V., Lopes, R., & da Silva, E. B. (2012). *Recursos genéticos de palma de óleo (Elaeis guineensis Jacq.) e caiaué (Elaeis oleifera (HBK) Cortés)*. Manaus, Brazil: Embrapa Amazônia Ocidental. Retrieved from <https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/949588/1/Doc96A5.pdf>
- Sansaloni, C. P. (2012). *Desenvolvimento e aplicações de DArT (Diversity Arrays Technology) e genotipagem por sequenciamento (Genotyping-by-Sequencing) para análise genética em Eucalyptus* (Ph.D. dissertation, Universidade de Brasília). Retrieved from <https://repositorio.unb.br/handle/10482/13400>
- Schlee, D. (1975). Numerical taxonomy. The principles and practice of numerical classification by P. H. A. Sneath, R. R. Sokal, W. H. Freeman. *Systematic Zoology*, 24, 263–268. <https://doi.org/10.2307/2412767>
- Schoen, D. J., & Brown, A. H. (1993). Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. *Proceedings of the National Academy of Sciences of the United States of America*, 90, 10623–10627. <https://doi.org/10.1073/pnas.90.22.10623>
- Singh, R., Ong-Abdullah, M., Low, E., Manaf, M. A. A., Rosli, R., Nookiah, R., ... Sambanthamurthi, R. (2013). Oil palm genome sequence reveals divergence of interfertile species in Old and New Worlds. *Nature*, 500, 335–339. <https://doi.org/10.1038/nature12309>
- Smouse, P. E., & Sork, V. L. (2004). Measuring pollen flow in forest trees: A comparison of alternative approaches. *Forest Ecology and Management*, 197, 21–38. <https://doi.org/10.1016/j.foreco.2004.05.049>
- Sork, V. L. (2015). Gene flow and natural selection shape spatial patterns of genes in tree populations: Implications for evolutionary processes and applications. *Evolutionary applications*, 9, 291–310. <https://doi.org/10.1111/eva.12316>
- Wright, S. (1978). *Evolution and the genetics of population, variability within and among natural populations*. Chicago, IL: The University of Chicago Press.

- Yeh, F. C. (2000). Population genetics. In A. Young & D. Boshier & T. Boyle (Eds.), *Forest conservation genetics: Principles and practice* (pp. 21–37). Collingwood, VIC, Australia: CSIRO.
- Xu, C., Gao, J., Du, Z., Li, D., Wang, Z., Li, Y., & Pang, X. (2016). Identifying the genetic diversity, genetic structure and a core collection of *Ziziphus jujuba* Mill. var. *jujuba* accessions using microsatellite markers. *Scientific Reports*, 6. <https://doi.org/10.1038/srep31503>
- Zhou, L. X., Xiao, Y., Xia, W., & Yang, Y. D. (2015). Analysis of genetic diversity and population structure of oil palm (*Elaeis guineensis*) from China and Malaysia based on species-specific simple sequence repeat markers. *Genetics and Molecular Research*, 14, 16247–16254. <https://doi.org/10.4238/2015.December.8.15>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Pereira VM, Filho JAF, Leão AP, et al. American oil palm from Brazil: Genetic diversity, population structure, and core collection. *Crop Science*. 2020;60:3212–3227. <https://doi.org/10.1002/csc2.20276>