ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2015) **13**(3); 256–265
doi:10.1017/S1479262114001026

Genetic diversity and establishment of a core collection of oil palm (*Elaeis guineensis* Jacq.) based on molecular data

Diana Arias[1], Maria González[1] and Hernán Romero[1,2]*

[1]*Oil Palm Biology and Breeding Program, Oil Palm Research Center, Calle 21 No. 42-55, Bogotá, Colombia and* [2]*Department of Biology, Universidad Nacional de Colombia, Bogotá, Colombia*

Received 25 July 2014; Accepted 30 October 2014 – First published online 8 December 2014

Abstract

Understanding of genetic diversity and its distribution is essential for promoting the use of genetic resources. The development of core collections using molecular tools has been proposed as a strategy for increasing the economical use and conservation of genetic resources. In this study, we investigated the genetic variation among different geographical origins and potential entries that constituted a core collection of oil palm, using 29 microsatellite markers and by evaluating 788 oil palm accessions. Our results revealed important genetic diversity ($H_T = 0.759$) between oil palm accessions from Angola and Cameroon, which exhibited a low coefficient of genetic differentiation between populations ($G_{ST} = 0.022$). However, the inclusion of oil palm accessions from Indonesia in the analysis resulted in a high coefficient of genetic differentiation between populations ($G_{ST} = 0.251$). We found that the combination of stratified sampling based on a sorting method and a heuristic algorithm was the most effective method for the development of an oil palm core collection set. Using this method, two core collections were identified. The first core collection, comprising 289 entries, contained 271 retained alleles in a sample representing 37% of the entire collection. The second one is a mini core collection, comprising 91 entries, that contained 271 retained alleles with a total H_e value of 0.72 in a sample representing 11% of the entire collection. The information reported in this study will be of great interest to oil palm researchers because new strategies for breeding programmes can be developed based on these advances.

Keywords: genetic resources; microsatellite markers; mini core collection; oil palm

Introduction

Oil palm represents the single largest source of vegetable oils worldwide, producing 45.8 million tons of oil in 2010 or constituting 26.7% of the world's production of edible oils (FAOSTAT, 2010). This crop has rapidly expanded in the tropical countries of Asia, Africa and the Americas. Although the oil palm *Elaeis guineensis* Jacq. is native to Africa, the world's largest palm oil producers are Malaysia and Indonesia (Corley and Tinker, 2003).

Breeding is regarded as a practice that results in the reduction of genetic diversity. This reduction may change the future adaptability and breeding progress of the crop. Therefore, it is necessary to preserve the genetic variation that underlies the traits required for the sustainable development of oil palm cultivation. Crop wild

* Corresponding author. E-mail: hmromeroa@unal.edu.co

relatives are particularly useful as a source of genes that confer resistance to various types of stress (Kuroda *et al*., 2009). This has encouraged oil palm breeding programmes in Colombia to conduct exploratory studies on oil palm-derived materials in Angola and Cameroon for the establishment of *ex situ* genetic collections (Rey *et al*., 2004; Arias *et al*., 2013a,b). The large number of oil palm accesions contained within these collections presented challenges for their conservation, evaluation, identification and usage. To promote the use of genetic resources, several authors have recommended the development of core collections. A core collection consists of a limited number of accessions that are selected to encompass the maximum genetic diversity and represent the genetic spectrum contained within the entire collection (Brown, 1989). A core collection gives researchers more favourable access to the gene pool and can serve as a reference point for research studies, which, in turn, can provide an overview of the characteristics of the entire collection.

Molecular marker techniques have been used in applications such as the development of collection sampling and gap identification strategies, the planning of future collections, the management and conservation of germplasm banks, the identification of duplicates and genetic contamination, and the determination of genetic change associated with breeding (Hodgkin *et al*., 2001). Molecular markers have been used for the development of core collections of both wild and cultivated plant species (Sangiri *et al*., 2007; Hao *et al*., 2008; Le Cunff *et al*., 2008; Kuroda *et al*., 2009; Zhang *et al*., 2011). These markers represent a powerful tool for germplasm characterization (Kalia *et al*., 2011). Simple sequence repeats (SSRs) have been widely used to determine the genetic diversity of core collections of *Elaeis* from germplasm banks worldwide (Billotte *et al*., 2001; Montoya *et al*., 2005; Bakoumé, 2007; Singh *et al*., 2008; Cochard *et al*., 2009; Arias *et al*., 2010; Billotte *et al*., 2010). Understanding of the genetic diversity and distribution of plants is essential for their preservation and usage. The objectives of this study were to use the molecular information generated by 29 SSR markers to determine the genetic diversity of oil palm accessions from Indonesia, Angola and Cameroon, and to select the entries that would constitute a core collection that ensured maximum allelic diversity.

Materials and methods

Plant material

The plant material used in this study is part of the *ex situ* collection of Cenipalma that is located in the Palmar de la Vizcaína experimental field (Barrancabermeja, Santander, Colombia). A total of 766 oil palm samples were analysed, of which 455 were collected from five geographical regions of the Republic of Angola, and 311 were collected from six geographical regions of the Cameroons (Rey *et al*., 2004; Arias *et al*., 2013a,b). In addition, 22 samples of oil palm Deli *dura* collected from Indonesia were analysed. The seeds of these materials were acquired from the Bogor Botanical Gardens.

DNA extraction and SSR amplification

For the analysis, young palm leaflets were used. Each leaflet sample was macerated in liquid nitrogen, and DNA was isolated using a Qiagen extraction kit (reference no. 69106), following the manufacturer's protocol. The quality of the extracted DNA was evaluated by spectrophotometry and gel electrophoresis on a 0.8% agarose. DNA extracted from each sample was diluted to a concentration of 5 ng/µl for use in subsequent amplification reactions. A total of 29 microsatellite markers were used, and according to the linkage disequilibrium test, they were located at independent loci and mapped to 14 linkage clusters (Billotte *et al*., 2005). The amplification conditions used for each marker have been described previously by Billotte *et al*. (2001) and Singh *et al*. (2008). The amplification products were resolved by electrophoresis on a 6% denaturing polyacrylamide gel containing 5 M urea and stained with silver nitrate, according to the protocols described by Qu *et al*. (2005). Alleles were identified using a 10 base-pair (bp) molecular marker with fragment sizes ranging from 10 to 330 bp.

Data analysis

Genetic differentiation and diversity

Allele frequencies, total number of alleles, average polymorphic information content and allelic richness (R_a) were calculated for each SSR locus, and evaluated based on the algorithms included in the (FSTAT software, http://www2.unil.ch/popgen/softwares/fstat.htm/; Goudet, 2002). Probability of identity, observed heterozygosity (H_o) and expected heterozygosity (H_e) were estimated for each locus using the (GIMLET software, http://pbil.univ-lyon1.fr/software/Gimlet/gimlet.htm/; Gimlet, 2002). The number of private alleles identified for each geographical origin was calculated using the (GENALEX software, http://biology.anu.edu.au/GenALEx/; Peakall and Smouse, 2006). Genetic differentiation among the geographical origins was determined based on Nei's (1987) coefficient of genetic diversity and differentiation using the (FSTAT software, http://www2.unil.ch/popgen/softwares/fstat.htm/; Goudet, 2002).

https://doi.org/10.1017/S1479262114001026 Published online by Cambridge University Press

Establishment of the core collection

Here, the term 'accession' refers to a single palm in the entire collection, and the term 'entry' refers to an accession selected for the inclusion in the core collection. The cluster-based stratified sampling strategy was used to establish the core collection (van Hintum *et al.*, 2000; Zewdie *et al.*, 2004), which included two steps. The first step involved the determination of clusters based on the following two methods: cluster analysis and sorting method. The cluster analysis was carried out based on Nei and Li's (1979) coefficient of genetic similarity and the unweighted pair group method with arithmetic mean (UPGMA) clustering method using the NTSYSpc software (Rohlf, 2000). The sorting method consisted of a multiple correspondence analysis (MCA) using the coefficient of genetic similarity based on the proportion of shared alleles (PS), according to the formula proposed by Bowcock *et al.* (1994):

$$\mathrm{PS} = \frac{\sum_{j=1}^{r} S_j}{2r},$$

where S_j is the number of shared alleles by two individuals at locus j ($j = 1$ to r).

The second step was the selection of individuals to determine the number of entries to be included in each cluster, which constituted the core collection. Therefore, three methods were employed for the selection of entries. Two of these methods have been proposed previously by Brown (1989), who stated that the number of accessions within clusters is based on the proportion (P) or the logarithmic (L) ratio of the size of each cluster in the whole collection. Finally, for the third selection method of entries, which contains all genotypic features in the core collection, a heuristic algorithm was used as proposed by Kim *et al.* (2007). This analysis was carried out using the (PowerCore software, http://www.genebank.go.kr/eng/PowerCore/powercore.jsp/). Indicators of genetic diversity in the whole and core collections were calculated using the (GENALEX software, http://biology.anu.edu.au/GenALEx/; Peakall and Smouse, 2006). Genetic diversity was quantified by the number of alleles (A) and expected heterozygosity (H_e) retained in the core collection with respect to the whole collection.

Size of the core collection

Brown (1989), Spagnoletti (1993), Li *et al.* (2002), Zewdie *et al.* (2004) and Yan *et al.* (2007) suggested that 10% of the entire collection could be an acceptable sample size for the core collection, and that this fraction could contain 70% of the total diversity. In this study, according to the sampling strategies P and L, the core collection contained 78 oil palm accessions (*E. guineensis* Jacq.) that constituted a sample size of 10%.

Results

Allelic diversity

The SSR molecular markers used in this study were polymorphic among the accessions evaluated, producing a total of 271 alleles (Table 1). The number of alleles identified in the analysed SSR loci ranged from 2 (sEg00126) to 18 (mEgCIR3362). Table 1 shows the SSR loci ordered according to the expected heterozygosity (H_e), where the first locus (mEgCIR3362) was most informative and most likely to detect the differences between two individuals, and the last marker (sEg00126) was least informative and least likely to detect the differences. The mEgCIR3362 locus exhibited the highest values for allelic richness (R_a = 13.69), observed heterozygosity (H_o = 0.82) and expected heterozygosity (H_e = 0.90). Of the identified alleles, 30% corresponded to the private alleles. Here, the term 'private alleles' refers exclusively to the alleles that were present in only one of the three countries. The 455 accessions of oil palm from Angola contained 36 private alleles with frequencies ranging from 0.001 to 0.188; the 311 accessions of oil palm from Cameroon contained 26 private alleles with frequencies ranging from 0.002 to 0.080; and the 22 accessions of oil palm from Indonesia contained 20 private alleles, of which 6 were identified using the mEgCIR3543 locus. The frequencies of those 20 private alleles ranged from 0.023 to 1.000, indicating that despite the limited number of oil palm accessions analysed, the samples from Indonesia exhibited the highest number of private alleles.

Genetic differentiation and diversity

According to Nei's (1987) coefficients of genetic diversity and differentiation, a low coefficient of genetic differentiation (G_{ST} = 0.022) was observed between the oil palm accessions from Angola and those from Cameroon, which exhibited a total genetic diversity (H_T) value of 0.658 (Table 2). The distribution pattern of the accessions obtained from the MCA did not reflect a separation in terms of the geographical origins from which the palms were collected (Fig. 1). The three dimensions of the MCA explained 100% of the variation in the accessions analysed, in which each dimension explained 33.3% of the variation. This lead to the occurrence of potential introgression between the accessions from Angola and those from Cameroon, and to low genetic differentiation between them. The inclusion of the 22 accessions from Indonesia in the analysis resulted in a high coefficient of genetic differentiation (G_{ST} = 0.178–0.251), indicating that the accessions from Indonesia were genetically

https://doi.org/10.1017/S1479262114001026 Published online by Cambridge University Press

Table 1. Allelic diversity in microsatellite loci evaluated in oil palm accessions (*Elaeis guineensis*)

			Private alleles by geographical origin							
Locus	Repeat motif	Size range (bp)	Angola ($n = 455$)	Cameroon ($n = 311$)	Indonesia ($n = 22$)	Total number of alleles	R_a	H_o	H_e	PI
mEgCIR3362	$(GA)_{19}$	146–180	2			18	13.69	0.82	0.90	1.82×10^{-2}
mEgCIR3292	$(GA)_{20}$	160–200	3	1		15	11.64	0.78	0.89	2.06×10^{-2}
mEgCIR3886	$(GA)_5GT(GA)_{20}$	174–194	2		2	13	9.20	0.73	0.88	2.73×10^{-2}
mEgCIR3546	$(GA)_{15}$	270–310	4	1		15	10.65	0.72	0.86	3.51×10^{-2}
mEgCIR1772	$(GT)_{22}$	172–195		2		10	8.36	0.47	0.86	3.59×10^{-2}
mEgCIR0802	$(GA)_{12}$	218–328	4	1	1	14	9.21	0.35	0.85	3.73×10^{-2}
mEgCIR3363	$(GA)_{17}$	148–204	4	5		16	9.17	0.72	0.85	4.03×10^{-2}
mEgCIR0067	$(GA)_{17}$	145–175	1	3	1	14	7.98	0.70	0.83	5.20×10^{-2}
mEgCIR1730	$(CT)_{17}(GT)_5$	245–274	1	3		11	7.51	0.70	0.83	5.37×10^{-2}
mEgCIR0254	$(TA)_4(GA)_{18}$	148–170	2	1	1	12	7.59	0.66	0.82	6.31×10^{-2}
mEgCIR3785	$(GA)_{21}$	260–290	1	1		9	7.08	0.68	0.80	8.36×10^{-2}
mEgCIR3282	$(GA)_{20}$	218–245			1	10	7.86	0.72	0.76	1.10×10^{-1}
mEgCIR3543	$(GA)_{17}$	208–240		2	6	14	6.77	0.70	0.74	1.16×10^{-1}
mEgCIR0008	$(GA)_{18}$	200–225		1		8	6.72	0.65	0.72	1.22×10^{-1}
mEgCIR0219	$(GA)_{17}$	210–229		2		8	5.93	0.57	0.71	1.29×10^{-1}
mEgCIR0465	$(CCG)_6$	125–137	1		1	6	5.63	0.56	0.71	1.30×10^{-1}
mEgCIR0230	$(TA)_6GAG(GA)_{18}$	320–350	1			7	5.27	0.59	0.71	1.33×10^{-1}
mEgCIR0173	$(GA)_{18}$	104–144	2	3		13	5.73	0.39	0.70	1.77×10^{-1}
mEgCIR0009	$(GA)_{20}$	164–185			2	7	3.61	0.61	0.65	2.10×10^{-1}
mEgCIR1753	$(GT)_{21}$	282–335	1			5	3.96	0.40	0.62	2.64×10^{-1}
mEgCIR0221	$(GA)_{11}$	195–210			1	5	4.18	0.46	0.55	2.94×10^{-1}
sEg00067	$(TGTA)_6$	235–260			2	5	3.66	0.57	0.49	4.24×10^{-1}
sEg00126	$(CGC)_7$	213–215				2	2.00	0.29	0.49	4.63×10^{-1}
Total	29		36	26	20	271				

n, number of palms; R_a, allelic richness; H_o, observed heterozygosity; H_e, expected heterozygosity; PI, probability of identity.

Table 2. Nei's coefficients of genetic diversity and differentiation between oil palm accessions

Groups	N	H_o	H_S	D_{ST}	H_T	G_{ST}	G_{IS}
Angola and Cameroon	766	0.555	0.669	0.015	0.684	0.022*	0.170*
Angola and Indonesia	477	0.560	0.600	0.190	0.789	0.241*	0.067*
Cameroon and Indonesia	333	0.570	0.602	0.202	0.804	0.251*	0.054*
Angola, Cameroon and Indonesia	788	0.561	0.624	0.135	0.759	0.178*	0.100*

N, number of entries; H_o, average of observed genetic diversity; H_S, average of genetic diversity within the groups; H_T, total genetic diversity; D_{ST}, average of genetic diversity among the subgroups; G_{ST}, coefficient of genetic differentiation; G_{IS}, inbreeding coefficient.
Values were statistically significant (* $P = 0.001$).

different from those acquired from Africa (Angola and Cameroon). This analysis was carried out using the same number of accessions for all geographical origins (22 accessions per origin selected randomly), exhibiting a G_{ST} value of 0.249, which also indicated a high coefficient of genetic differentiation between these accessions. The oil palm accessions from Cameroon and Indonesia exhibited the highest genetic diversity ($H_T = 0.804$). The distribution of genotypes within each group of accessions was measured by the inbreeding coefficient (G_{IS}). The values obtained were low (close to zero), positive and significant ($0.054 < G_{IS} > 0.170$) (Table 2), showing that the allelic frequencies of the genotypes evaluated were within the expected proportions that were consistent with Hardy–Weinberg equilibrium.

Establishment of the core collection

Determination of clusters

The dendrogram generated by NTSYSpc using the UPGMA clustering method showed the formation of many clusters with few accessions (see online supplementary Fig. S1). Therefore, we could not identify the distinguishable groups clearly, as the cluster analysis did not determine the optimal number of clusters. Nevertheless, the MCA determined the optimal number of clusters by evaluating the similarities among the individuals based on the χ^2 distances. Thus, the results of the MCA revealed the establishment of 11 groups for 766 accessions (Fig. 2). Cluster I represents 77% of the accessions from Cameroon. Clusters II and III collectively represent 84% of the accessions from Angola. Overall, 80% of the accessions were distributed in clusters I, II and III. Although clusters X and XI represent only one accession each, we decided to include these entries in the core collection. The oil palm accessions from Indonesia were not included in the cluster-based stratified sampling because they may mask the few differences between the accessions from Angola and those from Cameroon; instead, these accessions were considered as an additional cluster (cluster XII; see Table 2).

Allocation of the number of entries in each cluster

A new structure for the whole collection was established, which was divided into 12 clusters. Three sampling methods were used for the allocation of the number of entries. The number of alleles (A) and the expected heterozygosity (H_e) were determined for each group and for the total number of the entries selected. However, only the total number of entries was used to quantify the genetic diversity that could be included in the core collection (Table 3). According to the P and L strategies, approximately 235 alleles could be included in a sample that represented 10% of the entire collection. The L strategy was more effective for the retention of alleles and a high genetic diversity value in clusters with fewer accessions. Two subsets were established using the (PowerCore software, http://www.genebank.go.kr/eng/PowerCore/powercore.jsp/). In the first subset, the entries were selected from each group. As a result, 289 entries were selected, which would be the retention of 271 alleles in a sample representing 37% of the entire collection. In the second subset, the entries were selected from the 788 accessions contained within the whole collection, regardless of the clusters established previously. As a result, 91 entries were selected, representing the 12 groups established previously by the MCA and

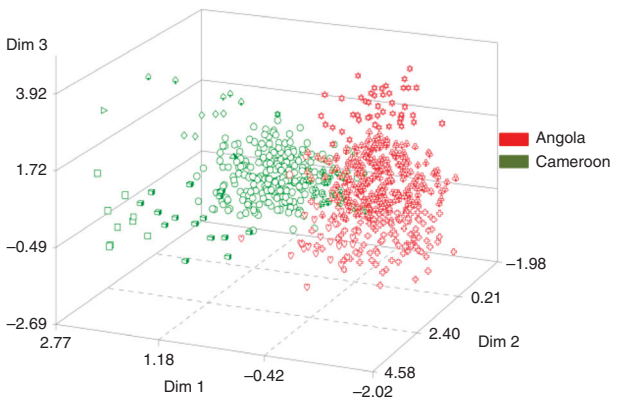

Fig. 1. Three-dimensional (Dim) multiple correspondence analysis of 766 oil palm accessions from Angola and Cameroon.

https://doi.org/10.1017/S1479262114001026 Published online by Cambridge University Press

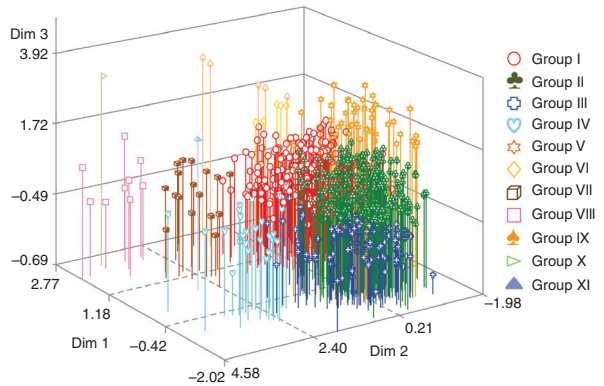

Fig. 2. Three-dimensional (Dim) multiple correspondence analysis showing the classification of 766 oil palm accessions from Angola and Cameroon into 11 clusters.

containing 271 retained alleles with a total H_e value of 0.72 in a sample representing 11% of the entire collection.

Finally, fully random sampling from a different set of SSRs was used (Table 4), with the aim of comparing the methods that have been used previously. For set 1, the total number of SSRs was used. For set 2, the following SSRs were used: mEgCIR0067; mEgCIR0219; mEgCIR0230; mEgCIR1772; sEg00125; sEg00126; sEg00127; mEgCIR 0173; mEgCIR3282; mEgCIR3292; mEgCIR3543; mEgCIR3546; mEgCIR3785; mEgCIR3886. For set 3, the following SSRs were used: mEgCIR0008; mEgCIR0009; mEgCIR0221; mEgCIR0254; mEgCIR0437; sEg00066; sEg00067; sEg00090; sEg00140; mEgCIR0802; mEg CIR1730; mEgCIR1753; mEgCIR3292; mEgCIR3362. Consequently, the number of alleles (A) and the expected heterozygosity (H_e) were less in the three random sets than in the PowerCore, P and L strategies.

Discussion

The accessions of oil palm (*E. guineensis*) evaluated in this study exhibited significant genetic diversity for this species. Therefore, these accessions are considered as a useful genetic resource for increasing the genetic base of oil palm breeding populations. The introduction of new germplasm sources is crucial to the development of new varieties with competitive advantages for the plant breeding sector. A total of 271 alleles were obtained from the 29 SSRs evaluated and private alleles were identified for each geographical origin, which was reflected in the high genetic diversity values obtained for the 788 oil palm accessions evaluated. Similar genetic diversity values for oil palms (*E. guineensis*) have been reported previously by Billotte *et al.* (2001), Montoya *et al.*, (2005), Singh *et al.* (2008), Arias *et al.* (2013a,b). Although microsatellites have been demonstrated to be versatile molecular markers for genetic diversity analysis in oil palms, it is important to consider that they may impose some limitations on data analysis. Microsatellites generally reflect the repeat size ranging from 2 to 5 bp of DNA. In practice, the occurrence of other alleles that correspond to changes outside of the repeat unit is likely, leading to other variations in repeat size, including single-base differences in some microsatellites. The detection of such differences would be difficult with the gel system used in this study, for which the resolution is around 2 bp, which could complicate the interpretation of microsatellite allele frequencies and thus make the estimates of relatedness faulty.

The parameters of genetic differentiation and the MCA determined in this study revealed a low coeffi-

Table 3. Number of entries selected from each cluster using several sampling strategies

				Core collection											
	Base collection			Proportion (P)			Logarithmic (L)			PowerCore[a]			PowerCore[b]		
Groups	*N*	*A*	H_e	*N*	*A*	H_e	*N*	*A*	H_e	*N*	*A*	H_e	*N*	*A*	H_e
I	241	209	0.68	24	171	0.68	12	150	0.67	61	209	0.68	23	183	0.69
II	291	218	0.67	29	183	0.66	13	153	0.67	68	218	0.67	22	184	0.70
III	123	205	0.64	12	142	0.63	11	135	0.63	56	205	0.64	10	148	0.64
IV	26	150	0.55	3	75	0.55	7	108	0.55	22	150	0.55	3	79	0.57
V	47	176	0.62	4	92	0.63	9	127	0.64	31	176	0.64	9	141	0.64
VI	5	114	0.66	1	49	–	4	104	0.66	5	114	0.66	3	93	0.66
VII	17	125	0.56	2	56	0.48	6	104	0.60	15	125	0.56	2	74	0.70
VIII	9	107	0.50	1	37	–	5	78	0.45	9	107	0.50	6	93	0.50
IX	5	106	0.65	1	47	–	4	92	0.65	5	106	0.65	2	71	0.60
X	1	46	–	–	–	–	–	–	–	1	46	–	1	46	–
XI	1	48	–	–	–	–	–	–	–	1	48	–	1	48	–
XII	22	99	0.53	2	66	0.50	7	90	0.53	15	99	0.53	9	97	0.54
Total	788	271	0.75	78	235	0.68	78	234	0.70	289	271	0.75	91	271	0.72

N, number of entries; *A*, number of alleles; H_e, expected heterozygosity (Nei, 1987).
[a] Entries were selected from each group. [b] Entries were selected from 788 accessions.

https://doi.org/10.1017/S1479262114001026 Published online by Cambridge University Press

Table 4. Number of entries selected fully randomly from a different set of simple sequence repeats (SSRs)

Random set 1 with 29 SSRs			Random set 2 with 14 SSRs			Random set 3 with 14 SSRs		
N	A	H_e	*N*	A	H_e	*N*	A	H_e
788	271	0.75	788	143	0.69	788	128	0.67
78	214	0.67	78	123	0.69	78	113	0.66
91	220	0.67	91	127	0.69	91	106	0.67
289	234	0.68	289	136	0.69	289	111	0.67

N, number of entries; *A*, number of alleles; H_e, expected heterozygosity (Nei, 1987).

cient of genetic differentiation between the populations of palms collected from Angola and Cameroon. This finding may be attributed to the geographical distribution of oil palms in Africa and their continuous expansion since the early days of the slave trade from the coasts of Cape Verde to Angola, which maintained active seed dispersal, prevented the structuring of populations and promoted genetic diversity of the species (Corley and Tinker, 2003). The results of the present study are consistent with those reported previously by Kularatne *et al.* (2001); Billote *et al.* (2001); Barcelos *et al.* (2002); Hayati *et al.* (2004), Montoya *et al.*, (2005); Maizura *et al.* (2006), Arias *et al.* (2013a,b). These previous studies using different types of molecular markers have reported that, oil palm accessions from different African countries did not exhibit a specific and distinct genetic structure, which was attributed predominantly to the dispersion of the material without any geographical barriers over the African continent. Conversely, Nei's (1987) coefficients of genetic differentiation indicated that Deli *dura* accessions from Indonesia are genetically distinct from oil palm accessions from Africa. Furthermore, these accessions exhibited the greatest number of private alleles compared with those of oil palm from Africa. Although many factors or biological phenomena are known to alter allelic frequencies, the genetic differences found among all the accessions could be attributed to natural processes and some historical events. Historical evidence shows that the oil palm is of African origin, and that its distribution along the equatorial tropics has been the result of the process of domestication by humans. In 1848, four Deli *dura* palms from the Bogor Botanical Gardens were introduced to the Botanical Gardens of Amsterdam and Mauritius. However, the exact origin of these palms remains unknown (Corley and Tinker, 2003). When oil palms were introduced to Southeast Asia, the conditions in Northern Sumatra were found to be more favourable for increasing the production of bunches, the average weight of bunches and the potential of oil production (Gerritsma and Wessel, 1997). The results of the present study are consistent with those reported previously by Kularatne *et al.* (2001), who evaluated oil palm accessions collected from different geographical origins using the amplified fragment length polymorphism (AFLP) markers. The study by Kularatne *et al.* (2001) found that Deli *dura* accessions are significantly different from oil palm populations in Africa, indicating that oil palm accessions from Indonesia and Africa can be treated as separate management units, as they have accumulated significant genetic differences between them.

The initial purpose of this study was to develop a core collection using molecular data to maximize the representativeness of genetic diversity. This will certainly allow the efficient use of genetic resources and, at the same time, a reduction in the costs of evaluation and characterization. The core collection can be modified eventually to accommodate new knowledge and new diversity. Cluster-based stratified sampling is one of the most effective methods for the establishment of a core collection, as the variation within clusters is minimized and the variation between clusters is maximized (Diwan *et al.*, 1995; Cochran, 1977). One of the most commonly used clustering methods is the UPGMA method. However, in this study, clearly differentiable clusters could not be identified. Therefore, clustering criteria were established based on a sorting method similar to that shown in the MCA. Results obtained from clustering analysis can be unstable when many clustered individuals are intermediate in terms of their similarity and/or differences compared with the others. In this case, these individuals are not assigned to a specific cluster, which leads either to the formation of several clusters consisting of only a few individuals or to individuals that are not assigned to any cluster at all. This problem can be overcome by using sorting methods because they do not group individuals according to a hierarchical structure. However, these methods may identify the relationships between individuals based on the presence or absence of a common molecular marker. After the identification of the relationship, the array of the individuals according to their similarities or differences can be determined easily. The MCA represents the similarities between individuals based on the χ^2 distances. This is the most appropriate approach for the analysis of molecular data, as it is sufficiently sensitive for detecting the patterns of similarity based on rare alleles that are shared by some genotypes (Laurentin, 2009). The MCA was the most suitable approach for defining the clusters that contain the total genetic diversity and for more effectively establishing an oil palm core collection that is representative of the genetic diversity of the whole collection. The results of this study show that the P strategy was more effective in retaining alleles within clusters with the largest number of accessions, whereas the L strategy was more effective

https://doi.org/10.1017/S1479262114001026 Published online by Cambridge University Press

in retaining alleles within clusters with fewer accessions. The two methods used for assigning the number of entries retained the maximum number of alleles and the observed genetic diversity within a sample representing 10% of the entire collection. This finding is consistent with that reported previously by Brown (1989), who established that these two methods optimize the entry allocation process within a core collection. However, the combination of stratified sampling based on a sorting method and a heuristic algorithm developed by Kim *et al.* (2007) was found to be the most effective method than the P and L strategies for the development of an oil palm core collection set while maintaining allele numbers and genetic diversity. Furthermore, based on the accuracy of the classification of individuals, PowerCore proved to be a powerful tool for streamlining the generation process of the core collection set of oil palms. Therefore, based on the allelic diversity evaluated using the 29 SSR loci in 788 oil palm accessions, two core collections were identified. The first core collection consisted of 289 entries (37% of the entire collection), while the second core collection is a mini core collection consisting of 91 entries (11% of the entire collection). The results of this study are consistent with those reported previously by Chung *et al.* (2009), who found that the heuristic method used for establishing a core collection was more efficient than the P and random core collection strategies. Zhao *et al.* (2010) found similar results for the development of a rice core collection. The use of the heuristic method was significantly more favourable than the random sampling method used for capturing the maximum number of alleles with minimum redundancy. Moe *et al.* (2012) previously demonstrated the efficiency of PowerCore in the development of a core collection using AFLP markers in mung bean plants.

The oil palm is an ideal example of a crop whose yield still responds to improved environments and whose genetic potential has not been fully exploited. Its adaptation to new environments represents a process that will continue to produce diversification according to new demands for domestication. Thus, accessions collected from Indonesia and Africa are a valuable genetic resource that may represent a long-term benefit for the development of resistance to various biotic and abiotic factors. We proposed the development of an oil palm core collection using molecular tools as a measure to increase its economical use and to develop strategies for the conservation of genetic resources. This core collection would be of great interest to oil palm researchers because new strategies for breeding programmes can be developed based on these advances to increase production and introduce new varieties that exhibit excellent nutritional qualities and are resistant to pests, diseases, drought and salinity.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1479262114001026

Acknowledgements

The authors thank the Angolan National Coffee Institute and the Institute for Agricultural Research and Development (IARD) for their collaboration and permission to collect the oil palm material. They thank Dr Leonardo Rey for collecting the oil palm material in Africa. They also thank Myriam Cristina Duque and Juan Bosco, researchers at the International Center for Tropical Agriculture (CIAT), for their assistance in the data analysis. This study was funded by the Oil Palm Development Fund (FFP), managed by Fedepalma.

References

Arias DM, Montoya C and Romero HM (2010) Preliminary results on the molecular characterization of oil palm using microsatellites markers. *PALMAS* 31: 35–45.

Arias D, Montoya C and Romero H (2013a) Molecular characterization of oil palm *Elaeis guineensis* Jacq. materials from Cameroon. *Plant Genetic Resources: Characterization and Utilization* 11(2): 140–148. doi: http://dx.doi.org/10.1017/S1479262112000482.

Arias D, González M, Prada F, Restrepo E and Romero H (2013b) Morpho-agronomic and molecular characterisation of oil palm *Elaeis guineensis* Jacq. material from Angola. *Tree Genetics & Genomes* 9: 1283–1294.

Bakoumé C, Wickneswari R, Rajanaidu N, Kushairi A, Amblard P, and Billotte N (2007) Allelic diversity of natural oil palm (*Elaeis guineensis Jacq.*) populations detected by microsatellite markers: implications for conservation. *Plant Genetic Resources: Characterization and Utilization* 5(2): 104–107. doi: 10.1017/S1479262107710870.

Barcelos E, Amblard P, Berthaud J and 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.

Billotte N, Risterucci AM, Barcelos E, Noyer JL, Amblard P and Baurens FC (2001) Development, characterisation, and across-taxa utility of oil palm (*Elaeis guineensis* Jacq.) microsatellite markers. *Genome* 44: 413–425.

Billotte N, Jourjon MF, Marseillac N, Berger A, Flori A, Asmady H, Adon B, Singh R, Nouy B, Potier F, Cheah SC, Rohde W, Ritter E, Courtois B, Charrier A and Mangin B (2010) QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis Jacq.*). *Theoretical and Applied Genetics* 120: 1673–1687.

Billotte N, Marseillac N, Risterucci AM, Adon B, Brottier P, Baurens FC, Singh R, Herrán A, Asmady H, Billot C, Amblard P, Durand-Gasselin T, Courtois B, Asmono D, Cheah SC, Rohde W, Ritter E and Charrier A (2005) Microsatellite-based high density linkage map in oil palm

(*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics* 110: 754–765.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455–457.

Brown AHD (1989) Core collections: a practical approach to genetic resources management. *Genome* 31: 818–824.

Chung JW, Kim KW, Chung JW, Lee JR, Lee SY, Dixit A, Kang HK, Zhao W, McNally KL, Hamilton RS, Gwag JG and Park YJ (2009) Development of a core set from a large rice collection using a modified heuristic algorithm to retain maximum diversity. *Journal of Integrative Plant Biology* 51: 1116–1125.

Cochard B, Adon B, Rekima S, Billotte N, Desmier de Chenon R, Koutou A, Nouy B, Omoré A, Purba AR, Glazsmann JC and Noyer JL (2009) Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. *Tree Genetics & Genomes* 5: 493–504.

Cochran WG (1977) *Sampling Techniques*, 3rd edn. New York: John Wiley and Sons.

Corley RHV and Tinker PB (2003) *The Oil Palm*, 4th edn. World Agricultural Series. Oxford UK: Blackwell Publishers Ltd.

Diwan N, McIntosh MS and Bauchan GR (1995) Methods of developing a core collection of annual *Medicago* species. *Theoretical and Applied Genetics* 90: 755–761.

FAOSTAT (2010) FAO Statistical Yearbook. World Food and Agriculture. Food and Agriculture Organization of the United Nations. Available at http://faostat.fao.org/site/ (accessed 11 October 2012).

Gerritsma W and Wessel M (1997) Oil palm: domestication achieved? *Netherlands Journal of Agricultural Science* 45: 463–475.

Gimlet VN (2002) GIMLET: a computer program for analyzing genetic individual identification data. *Molecular Ecology Notes* 2: 377–379.

Goudet J (2002) *Institute of Ecology, Biology Building, UNIL Software (FSTAT), Version 2.9.3.2*. http://www2.unil.ch/popgen/softwares/fstat.htm

Hao CY, Dong YC, Wang LF, You GX, Zhang HN, Ge HM, Jia JZ and Zhang XY (2008) Genetic diversity and construction of core collection in Chinese wheat genetic resources. *Chinese Science Bulletin* 53: 1518–1526.

Hayati A, Wickneswari R, Maizura I and Rajanaidu N (2004) Genetic diversity of oil palm (*Elaeis guineensis* Jacq.) germplasm collections from Africa: implications for improvement and conservation of genetic resources. *Theoretical and Applied Genetics* 108: 1274–1284.

Hodgkin T, Roviglioni R, de Vicente MC and Dudnik N (2001) Molecular methods in the conservation and use of plant genetic resources. *Acta Horticulturae* 546: 107–118.

Kalia RK, Rai MK, Kalia S, Singh R and Dhawan AK (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* 177: 309–334. doi: 10.1007/s10681-010-0286-9.

Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG and Park YJ (2007) PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23: 2155–2162.

Kularatne RS, Shah FH and Rajanaidu N (2001) The evaluation of genetic diversity of Deli *dura* and African oil palm germplasm collection by AFLP technique. *Tropical Agricultural Research* 13: 1–12.

Kuroda Y, Tomooka N, Kapa A, Wanigadea SM and Vaughan D (2009) Genetic diversity of wild soybean (*Glycine soja* Sieb. et Zucc.) and Japanese cultivated soybeans [*G. max* (L.) Merr.] based on microsatellite (SSR) analysis and the selection of a core collection. *Genetic Resources and Crop Evolution* 56: 1045–1055.

Laurentin H (2009) Data analysis for molecular characterization of plant genetic resources. *Genetic Resources and Crop Evolution* 56(2): 277–292. doi: 10.1007/s10722-008-9397-8.

Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon AF, Boursiquot JM and This P (2008) Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. sativa. *BMC Plant Biology* 8: 31.

Li ZC, Zhang HL, Zeng YW, Yang ZY, Shen SQ, Sun CQ and Wang XK (2002) Studies on sampling schemes for the establishment of core collection of rice landrace in Yunnan, China. *Genetic Resources and Crop Evolution* 49: 67–74.

Maizura I, Rajanaidu N, Zakri A and Cheah S (2006) Assessment of genetic diversity in oil palm (*Elaeis guineensis* Jacq.) using restriction fragment length polymorphism (RFLP). *Genetic Resources and Crop Evolution* 53: 187–195.

Moe KT, Gwag JG and Park YJ (2012) Efficiency of PowerCore in core set development using amplified fragment length polymorphic markers in mungbean. *Plant Breeding* 131: 110–117.

Montoya C, Arias DM, Rey L, and Rocha PJ (2005) Caracterización molecular de materiales de *E. guineensis* Jacq. Procedentes de Angola. *Fitotecnia Colombiana* 5(2): 1–10.

Nei M and Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences* 76: 5269– 5273.

Nei M (1987) *Molecular Evolutionary Genetics*. New York: Colombia University Press.

Peakall R and Smouse P (2006) GENALEX6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6 288–295.

Qu L, Li X, Wu G and Yang N (2005) Efficient and sensitive method of DNA silver staining in polyacrylamide gels. *Electrophoresis* 26: 99–101.

Rohlf FJ (2000) Statistical power comparisons among alternative morphometric methods. *American Journal Physical Anthropology* 111: 463–478.

Rey L, Gómez PL, Ayala I, Delgado W and Rocha P (2004) Colecciones genéticas de palma de aceite *Elaeis guineensis* (Jacq.) y *Elaeis oleifera* (H.B.K.) de Cenipalma: Características de importancia en el sector palmicultor. *PALMAS* 25: 39–48.

Sangiri C, Kaga A, Tomooka N, Vaughan DA and Srinives P (2007) Genetic diversity of the mungbean (*Vigna radiata,* Leguminosae) gene pool on the basis of microsatellite analysis. *Australian Journal of Botany* 55: 837–847. doi:10.1071/BT07105.

Spagnoletti PL and Qualset CO (1993) Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. *Theoretical and Applied Genetics* 87: 295–304.

Singh R, Mohd N, Ting N, Rosli R, Tan S, Leslie E, Ithnin M and Cheah S (2008) Exploiting an oil palm EST database for the

development of gene-derived SSR markers and their exploitation for assessment of genetic diversity. *Biologic* 63: 227–235 Section Cellular and Molecular Biology. doi: 10.2478/s11756-008-0041-z.

van Hintum T, Brown AHD, Spillane C and Hodgkin T (2000) *Core Collections of Plant Genetics Resources. IPGRI Technical Bulletin No. 3*. Rome, Italy: International Plant Genetic Resources Institute.

Yan WG, Rutger N, Bryant R, Bockelman HE, Fjellstrom RG, Chen MH, Tai T and McClung A (2007) Development and evaluation of a core subset of the USDA rice germplasm collection. *Crop Science* 47: 869–876. doi: 10.2135/cropsci2006.07.0444.

Zewdie Y, Tong N and Bosland P (2004) Establishing a core collection of *Capsicum* using a cluster analysis with enlightened selection of accessions. *Genetic Resources and Crop Evolution* 51: 147–151.

Zhang H, Zhang D, Wang M, Sun J, Qi Y, Li J, Wei X, Han L, Qiu Z, Tang S and Li Z (2011) A core collection and mini core collection of *Oryza sativa* L. in China. *Theoretical and Applied Genetics* 122: 49–61.

Zhao WG, Cho GT, Ma KH, Chung JW, Gwag JG and Park YJ (2010) Development of an allele-mining set in rice using a heuristic algorithm and SSR genotype data with least redundancy for the post-genomic era. *Molecular Breeding* 26: 639–651.