Annals of Botany Page 1 of 10
doi:10.1093/aob/mcw170, available online at www.aob.oxfordjournals.org

ANNALS OF BOTANY
Founded 1887

DArT whole genome profiling provides insights on the evolution and taxonomy of edible Banana (*Musa* spp.)

J. Sardos[1,*], X. Perrier[2], J. Doležel[3], E. Hřibová[3], P. Christelová[3], I. Van den houwe[4], A. Kilian[5] and N. Roux[1]

[1]Bioversity International, Parc Scientifique Agropolis II, 1990 boulevard de la Lironde, 34397 Montpellier Cedex 5, France, [2]CIRAD, UMR AGAP, 34398 Montpellier, France, [3]Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, 78371 Olomouc, Czech Republic, [4]Bioversity International, Willem De Croylaan 42, 3001 Leuven, Belgium and [5]Diversity Arrays Technology Pty Ltd, Building 3, University of Canberra, Bruce, ACT 2617, Australia

**For correspondence. E-mail j.sardos@cgiar.org*

Received: 19 February 2016 Returned for revision: 20 April 2016 Accepted: 17 June 2016

• **Background and Aims** Dessert and cooking bananas are vegetatively propagated crops of great importance for both the subsistence and the livelihood of people in developing countries. A wide diversity of diploid and triploid cultivars including AA, AB, AS, AT, AAA, AAB, ABB, AAS and AAT genomic constitutions exists. Within each of this genome groups, cultivars are classified into subgroups that are reported to correspond to varieties clonally derived from each other after a single sexual event. The number of those founding events at the basis of the diversity of bananas is a matter of debate.
• **Methods** We analysed a large panel of 575 accessions, 94 wild relatives and 481 cultivated accessions belonging to the section *Musa* with a set of 498 DArT markers previously developed.
• **Key Results** DArT appeared successful and accurate to describe *Musa* diversity and help in the resolution of cultivated banana genome constitution and taxonomy, and highlighted discrepancies in the acknowledged classification of some accessions. This study also argues for at least two centres of domestication corresponding to South-East Asia and New Guinea, respectively. Banana domestication in New Guinea probably followed different schemes that those previously reported where hybridization underpins the emergence of edible banana. In addition, our results suggest that not all wild ancestors of bananas are known, especially in *M. acuminata* subspecies. We also estimate the extent of the two consecutive bottlenecks in edible bananas by evaluating the number of sexual founding events underlying our sets of edible diploids and triploids, respectively.
• **Conclusions** The attribution of clone identity to each sample of the sets allowed the detection of subgroups represented by several sets of clones. Although morphological characterization of some of the accessions is needed to correct potentially erroneous classifications, some of the subgroups seem polyclonal.

Key words: *Musa acuminata*, *Musa balbisiana*, *Musa* spp., banana, DArT, domestication, taxonomy, classification, domestication.

INTRODUCTION

Banana, including cooking banana, is a vegetatively propagated crop of great importance for the subsistence of small-scale farmers in developing countries. This fruit and starchy crop is grown in more than 130 countries, mainly tropical, and is a major staple food for millions of people. In addition, more than 19 million tonnes of bananas, i.e. 13 % of the total global production, are exported (http://faostat3.fao.org/faostat-gateway/go/to/home/E). This makes banana critical for both the food security and the economy of many developing countries.

Banana, *Musa* spp., is a monocotyledon. With the exception of Australimusa Fe'i banana, not considered in this paper, it carries four known genomes, A, B, S and T, which correspond to the species *Musa acuminata*, *M. balbisiana*, *M. schizocarpa* and *M. textilis*, respectively. No hybridization among B, T or S genomes has been observed independently of the A genome but *M. acuminata* hybridizes with any of the three other species. However, there are few cultivated bananas composed of S and T genomes. The two main progenitor species of the domesticated forms of bananas are thus *M. acuminata* and *M. balbisisana*. Although no subdivision exists within *M. balbisiana* taxonomy, based on different observed chromosome structures *M. acuminata* has been divided into at least seven subspecies with different geographical distributions (Simmonds and Shepherd, 1955; Shepherd, 1999).

The four species at the origin of cultivated bananas have combined to generate a wide diversity of diploid and triploid cultivars with diverse genetic make-ups varying from AA, AB, AS, AT, AAA, AAB, ABB, AAS to AAT. Within each of these genome groups, cultivars are classified into subgroups that are considered to correspond to groups of varieties clonally derived from each other after a single sexual event. The most well known of the subgroups of banana are seedless triploids, such as the commercially important Cavendish dessert banana (AAA) and the staple cooking African Plantains (AAB), which have importance for food security. However, quite a high number of diploid cultivars are also cultivated, especially in the centre of origin, i.e. the South-East Asia/Melanesia region

(Simmonds, 1962; Lebot, 1999). The origin of modern bananas, especially of the commercial triploids, has been investigated and domestication schemes have been proposed (De Langhe *et al*., 2009; Perrier *et al*., 2011). The emergence of triploid cultivars is believed to have ensued from a multi-stepped process. Modern edible diploids may have been preceded by what De Langhe *et al*. (2009) named 'cultiwilds', i.e. pre-domesticated forms of bananas that might have been devoted to uses other than food, exhibiting intermediate levels of parthenocarpy and occasionally producing seeds. These cultiwilds, originating from different subspecies of *M*. *acuminata*, then probably diffused through exchanges between human communities and/or following human migrations. Once brought into contact, they are thought to have hybridized with local genepools and to have given rise to edible diploids. Due to parental chromosomal rearrangements and unbalanced meiosis in these hybrids, diploid gametes were sometimes formed, so that in some cases the occurrence of sexual reproduction between them led to the emergence of triploid cultivars (reviewed by Perrier *et al*., 2011). The most striking example is the likely resolution of the direct ancestry of the Cavendish AAA sub-groups: restriction fragment length polymorphism (RFLP) and simple sequence repeat (SSR) markers have revealed that two AA landraces originating from the Mlali and Khai clusters were the most likely providers of their AA and A parental gametes, respectively (Carreel *et al*., 2002; Raboin *et al*., 2005; Perrier *et al*., 2009; Hippolyte *et al*., 2012).

A range of molecular markers have been used to characterize and study banana diversity: amplified fragment length polymorphism (AFLP) (Ude *et al*., 2002), RFLP (Carreel *et al*., 2002; Raboin *et al*., 2005) and more recently microsatellites (Perrier *et al*., 2009; Christelová *et al*., 2011; Hippolyte *et al*., 2012; de Jesus *et al*., 2013; Irish *et al*., 2014). Originally developed for rice, diversity arrays technology (DArT) markers (Jaccoud *et al*., 2001) are most widely used. They were designed to enable whole-genome profiling without the need of sequence information. Due to their high polymorphism information content (PIC), DArT has been successfully applied to various crops, from wheat (Akbari *et al*., 2006) and sorghum (Bouchet *et al*., 2012) to chickpea (Roorkiwal *et al*., 2014). In banana, DArT has already been used for a range of applications, from diversity studies (Amorim *et al*., 2009; Risterucci *et al*., 2009) to genetic mapping (Hippolyte *et al*., 2010; D'Hont *et al*., 2012).

In this study, we propose to use a batch of 498 polymorphic previously developed DArT markers (Risterucci *et al*., 2009) to explore the genetic diversity of a large sample composed of 575 accessions of bananas, covering most of the known diversity of wild and cultivated diploids and triploids from the section Eumusa. The accessions are conserved for distribution in Bioversity International's Global Collection of Banana, the International Transit Center (ITC), hosted in the Catholic University of Leuven, Belgium. These accessions originate from diverse field collections and collecting missions (accessed through MGIS, http://www.crop-diversity.org/banana/) and constitute a good representation of the existing diversity of *Musa* worldwide. The results obtained allowed us to provide a global image of *Musa* diversity and to validate the accuracy of DArT markers in detecting genome composition and revealing clustering in banana accessions. Secondly, we discuss the extent of the consecutive bottlenecks that underpinned banana domestication. Finally, we argue for the anchorage of the taxonomy of cultivated bananas within an evolutionary perspective.

MATERIALS AND METHOD

Plant material

A total of 575 accessions were obtained from the ITC's *in vitro* genebank. The sample set was composed of 94 wild accessions and of 481 cultivated accessions, including 208 diploids, 269 triploids and four mixoploids, i.e. accessions exhibiting diploid and triploid cells while measured with flow cytometry. The numbers of individuals per different species and genome groups are summarized in Table 1.

DNA extraction and DArT procedure

DNA was extracted from lyophilized samples provided by ITC following the protocol described at https://www.diversityarrays.com/files/DArT_DNA_isolation.pdf

Development of the DArT assay and DArT array was described by Risterucci *et al*. (2009). Briefly, each DNA sample was digested with a combination of *Pst*I and *Taq*I restriction enzymes, the adapter for *Pst*I overhang was ligated and fragments with *Pst*I ends that are missing the *Taq*I internal restriction site were amplified using primers targeted to the *Pst*I adapter sequence. Genomic representations thus created in that manner (targets) were quality-checked through gel electrophoresis and then fluorescently labelled with either Cy3 or Cy5 fluorescent dye. Labelled targets and FAM-labelled internal control (polylinker of the cloning vector used for DArT library construction) were hybridized to a banana array containing 6144 DArT clones printed in duplicate for 16 h at 62 °C. Slides were subjected to four washes of increasing stringency with a final rinse in water followed by drying. Slides were scanned using a Tecan laser scanner at three wavelengths matching emission of the three fluorescent dyes used in hybridization. The images generated by the scanner were stored in DArTdb (http://www.diversityarrays.com/dart-technology-package-dartDb) and used in marker data extraction. More detailed descriptions of the lab techniques are given by Kilian *et al*. (2012).

DArT analysis

Markers were scored '0' for absence and '1' for presence of the restriction fragment corresponding to DArT probe in the genomic representation. DArTsoft v.7.4 (Diversity Arrays Technoogy P/L, Canberra, Australia) was used to automatically identify and score polymorphic markers. The threshold criteria of call rate and reproducibility were set to be higher than 80 and 97 %, respectively.

Statistical analysis of DArT data

Global representation and structure of Musa *diversity*. Darwin 5.0 (Perrier and Jacquemoud-Collet, 2006; Perrier *et al*., 2003) was used to calculate genetic distances between pairs of the 575 accessions. To do so, both modalities (0,1) were given equal weight using the Sokal and Michener (1958) dissimilarity

TABLE 1. *Composition of the sample by species and genome groups*

		Diploids		Triploids		Mixoploids*	
Wild	94	*M. acuminata*	64	NA		NA	
		M. balbisiana	11				
		M. schizocarpa	11				
		M. acuminata × *M. schizocarpa*	8				
Cultivated	481	AA	199	AAA	140	AB+ABB	4
		AB	2	AAB	84		
		AS	6	ABB	39		
		AT	1	AAA/AAB	3		
		Total cultivated	208	AAT	3		
Total	575		302		269		4

*Mixoploid refers to accessions exhibiting cells with different numbers of chromosomes, here 22 and 33 (measured by flow cytometry, source: MGIS). NA, not applicable.

index as the proportion of unmatching markers. The dissimilarities matrix was first used to perform a principal coordinate analysis (PCoA).

A Bayesian Markov chain Monte Carlo (MCMC) approach was then used to detect genetic clusters within diploids. This model-based analysis was run using the program STRUCTURE version 2.3.3. (Pritchard *et al.*, 2000). We used the admixture model along with the assumption of correlated allele frequencies between groups (Falush *et al.*, 2003) and the optimal value of K was then determined by examining the posterior probabilities Ln $P(D)$, the partitioning of individuals across the K clusters and ΔK (Evanno *et al.*, 2005) as implemented in the web software STRUCTURE HARVESTER (Earl and vonHoldt, 2012). STRUCTURE then partitioned individuals of the sample according to the membership coefficient Q, which ranges from 0 (lowest affinity to the group) to 1 (highest affinity to a group), across the pre-defined K groups. Taking into account that the models implemented within STRUCTURE pre-supposed panmixia, we first analysed seeded accessions and then the edible diploid accessions as they exhibit a higher chance to meet this criterion than triploids. For each analysis, we ran ten replicates of each value of K ranging from 1 to 10 with a burn-in length of 400 000 followed by 1 000 000 iterations of each chain.

Clonal diversity of edible banana. The number of distinct multilocus genotypes (MLGs) present in the cultivated component of our sample (G) was determined using the software GenoType (Meirmans and Van Tienderen, 2004) based on the genetic distances matrix generated by DARwin. GenoType allows choosing a threshold (Th), i.e. the maximum pairwise genetic distance allowed between individuals to belong to the same clonal lineage, or to be clonemates, and then assigns a clonal identity to each individual. We ran two different datasets. The first involved cultivated diploid individuals only (208 samples) and led to the identification of 115 distance classes. The second involved cultivated triploid individuals only (273 samples including 269 triploids and four mixoploids) and led to the identification of 157 distance classes.

We then followed Douhovnikoff and Dodd (2003) to determine the threshold that would enable us to delimit banana clonesets through the observation of the frequency histogram of distances.

RESULTS

Global structure of Musa *diversity*

The PCoA performed on the distance matrix between genotypes of the whole sample is presented in Fig. 1. Factors 1 and 2 represented 52.67 % of the total variation observed. Axis 1, counting for 44.92 % of the variation observed, clearly discriminates accessions according to the proportion of the B genome involved in their genomic composition, going from pure B at the left to pure A at the right.

The discrimination displayed by Axis 2, accounting for 7.75 % of the variation observed, correlates to some extent with the geographical origins of the cultivated accessions, going from the North, e.g. ABB subgroups Pome and Mysore originating from India at the bottom of the graph, to the South with the cultivated AA originating from Papua at the top. However, this pattern does not fit with *M. acuminata* subspecies: if *banksii* is located at the top of the graph near the cultivated diploids from Papua New Guinea, the diversity of the main South-East Asia subspecies, *zebrina* from Java, *malaccensis* from the Malay-Thai peninsula and *burmannica* from Myanmar, is not structured according to geography. Interestingly, none of the subspecies included in this study clusters at the bottom of the graph where there is a large group of cultivated diploids and triploids including the AAA Cavendish and Gros Michel.

Finally, the clustering of the main cultivated subgroups is consistent with the accepted taxonomy of the samples.

Number of genetic clusters identified in the wild samples

The overall results obtained from STRUCTURE on the set of 93 wild samples are displayed in Fig. 2. Examining the posterior probabilities of the data for each K, here called Ln $P(D)$, along with their variance across runs, and Evanno *et al.*'s (2005) ΔK (Fig. 2A), we noticed that the highest peak of ΔK appears for $K = 2$. However, the occurrence of smaller peaks along the graph suggests additional levels of clustering, notably for $K = 3$, $K = 4$ and $K = 8$, all corresponding to stable values of Ln $P(D)$ across runs. As the over-representation of the subspecies *banksii* probably introduced some bias into the results, we investigated the partitioning of the individuals across genetic clusters for all these putative values of K (Fig. 2B). The first level of clustering allows a clear discrimination of

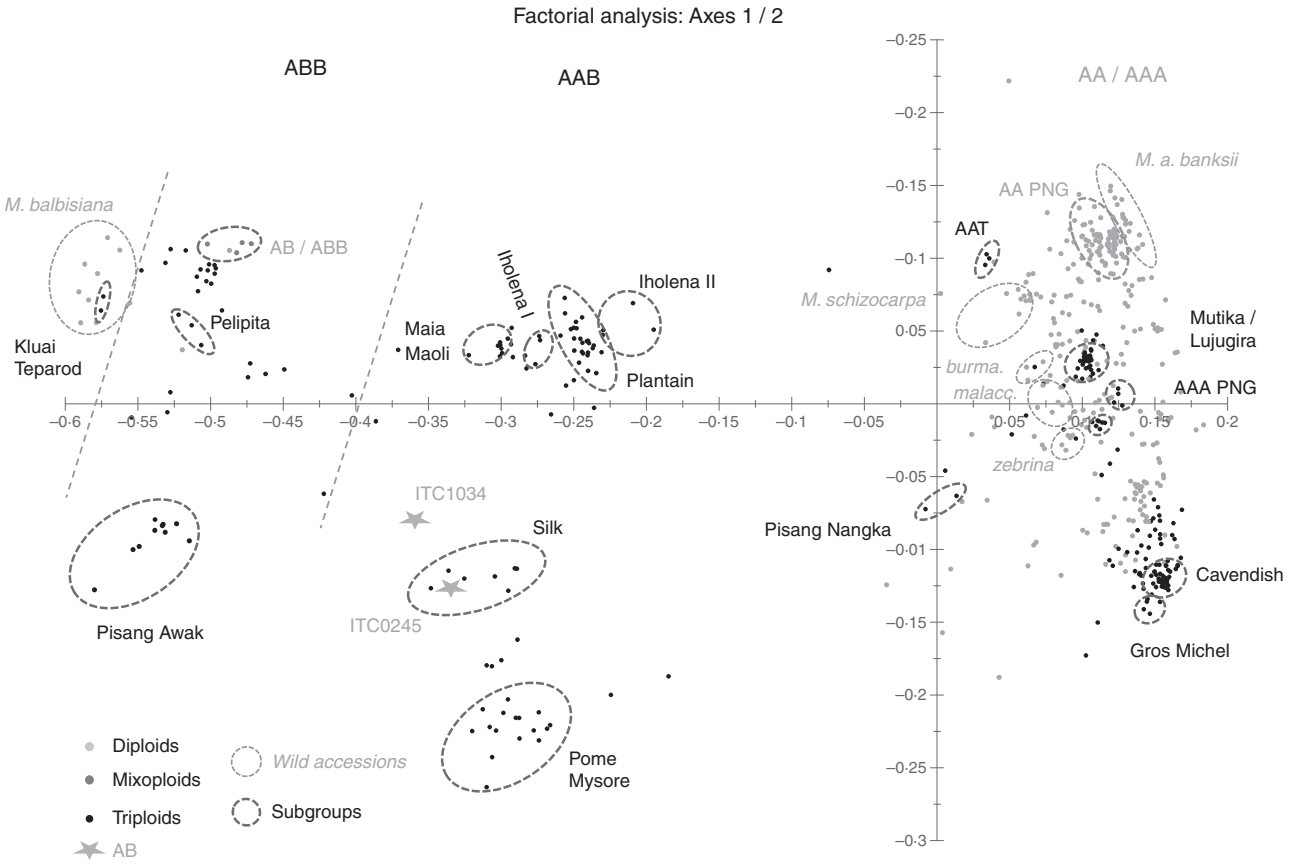

FIG. 1. PCoA performed on the Sokal and Michener dissimilarity matrix obtained from the genotyping of 575 *Musa* accessions with 498 DArT markers.

M. balbisiana from the *M. acuminata*/*M. schizocarpa* samples. The second level of clustering, $K = 3$, allows the further discrimination of *M. acuminata burmannica*/*M. schizocarpa* from *M. acuminata banksii*. The other subspecies from South-East Asia are considered as admixed accessions at this stage. For $K = 4$, *M. schizocarpa* clusters apart from any *M. acuminata* subspecies while the South Asian subspecies appear as a homogeneous group with punctual *banksii* introgressions. The pattern displayed for $K = 8$ is more complex but also allows the discrimination of South-East Asian *M. acuminata* subspecies *burmannica*, *malaccensis* and *zebrina*. In addition, it also provides three accessions classified as *malaccensis* with a hybrid status between *malaccensis* and *zebrina*. However, of the eight putative genetic clusters identified by STRUCTURE, only six display fully assigned individuals ($Q > 0{\cdot}8$).

Number of genetic clusters identified within the cultivated diploid sample

The Evanno *et al.* (2005) method applied to the results obtained from the analysis of the set of 208 cultivated diploid accessions with STRUCTURE (Fig. 3) suggests $K = 2$ as the real value of K even though a secondary peak of ΔK exists at $K = 3$. As we suspected a bias due to the probable overrepresentation of accessions collected in Papua, we also investigated the different clusters detected for $K = 3$ (data not shown), but the pattern displayed for $K = 2$ was the most convincing. The partitioning of individuals across the different clusters identified for $K = 2$ according to their countries of origin is presented in Table 2. Cluster 1 is composed of 50 accessions, mostly originating from South-East Asia, and cluster 2 is composed of 84 accessions, of which 82 originate from Papua, the two other accessions being ITC0299 'Guyod' from the Philippines and ITC1253 'Mjenga' which probably originated from Zanzibar (J. P. Horry, CIRAD, pers. comm.). Seventy-four accessions are admixed between both groups, i.e. $Q < 0{\cdot}8$. A majority of these admixed accessions originate from Papua (42) and the Philippines (11).

Considering the accessions fully assigned to a given cluster only, South-East Asia countries exhibited mainly accessions belonging to cluster 1 while the majority of the accessions collected in Papua belonged to cluster 2.

Clonal diversity of edible banana

We investigated the number of distinct MLGs, or clones, identified in the two cultivated datasets, diploids and triploids (including mixoploids).

At Th0, i.e. no difference allowed, GenoType identified 175 different MLGs out of the 208 cultivated diploids and 221 different MLGs out of the 273 cultivated triploids and mixoploids. However, this estimation of the number of different MLGs did

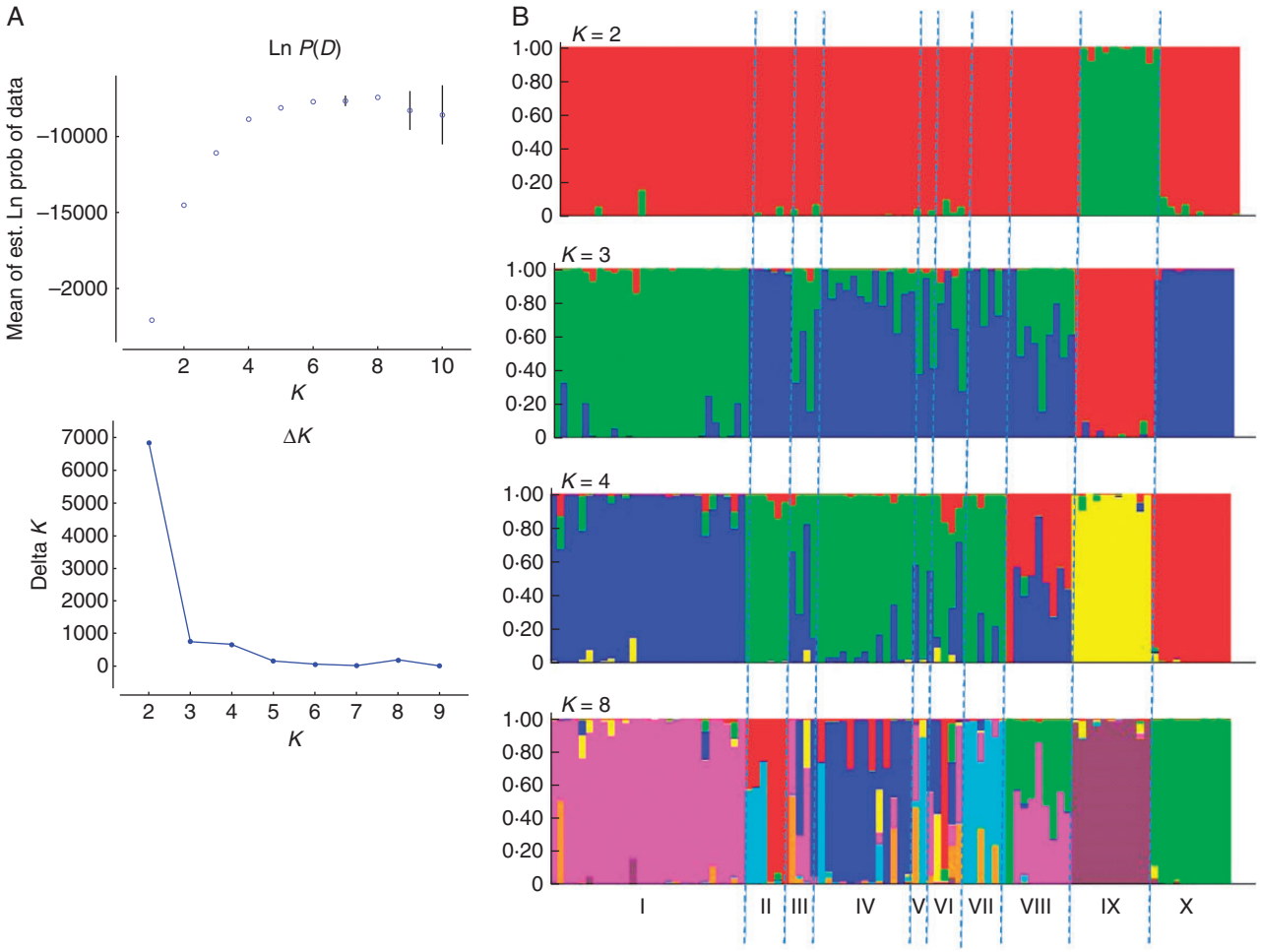

FIG. 2. Results obtained from STRUCTURE for the analysis of the full wild sample (94 individuals) (A) Median Ln(*K*) and median Δ*K* (Evanno *et al*., 2005). (B) Partitioning of the individuals according to their membership coefficient *Q* across the *K* groups for *K* = 2, 3, 4 and 8. Cluster I is composed of 27 *M. acuminata banksii*; cluster II of six *M. acuminata burmannica/burmannicoïdes*; cluster III of one *M. acuminata errans* and three *M. acuminata* qualified as hybrids; cluster IV of 13 *M. acuminata malaccensis*; cluster V of two *M. acuminata microcarpa*; cluster VI of one accession qualified as hybrid, of two *M. acuminata siamea*, of one *M. acuminata truncata* and one *M. acuminata* without known subspecies; cluster VII of seven *M. acuminata zebrina*; cluster VIII of hybrids between *M. acuminata* and *M. schizocarpa*; cluster IX of 11 *M. balbisiana*; and cluster X of 11 *M. schizocarpa*.

not take into account genotyping errors and accumulation of mutations as putative sources of genetic divergence among the accessions. In addition, DArT detects not only DNA sequence variation, but also, at a lower frequency, methylation variation at the *Pst*I site used for the complexity reduction step (Wittenberg *et al*., 2005; Kilian *et al*., 2012). Therefore, the distance estimated based on DArTs not only includes scoring errors, which correspond to a fraction of 1 % given the cutoff of 97 % technical reproducibility and clonal accumulation of mutations, but also epimutations which are likely to accumulate in the meristems of clonally propagated materials. Histograms of the distributions of the pairwise genetic distances for the 50 first classes of these distances (Fig. 4) revealed thresholds that appeared appropriate to evaluate the number of initial founding events, i.e. sexual events, at the origin of each of the sets. The histogram obtained for the cultivated diploids (Fig. 4A) exhibits a clear pattern, with the first peak located at the second distance class. This peak appears to end at the fifth distance class, which we considered to be the threshold value for the cultivated diploids. Therefore, the estimated number of different MLGs in this sample was 117 distinct MLGs out of 208 (see Supplementary Data Table S1).

Of these 117 MLGs, 36 were composed of 2–13 accessions while 81 were composed of unique accessions. However, we suspect that at least seven multi-accession MLGs are composed of duplicates or synonyms (Table S1). It is noticeable that the two AB accessions, in the accepted classification, are classified in the Ney Poovan subgroup, but here are not recognized as belonging to the same clone. Equally, cultivars ensuing from hybridization between *M. acuminata* and *M. schizocarpa* (AS) separate into two different clonal groups.

The pattern of genetic distances for the cultivated triploids, including mixoploids, is different (Fig. 4B): the first peak is also reached at the second genetic distance class but stretches until the eighth distance class. In addition, it is higher than that observed in the diploid accessions, suggesting higher rates of clonal differentiation among the triploids. This first high peak is followed by two lower peaks that suggest the occurrence of

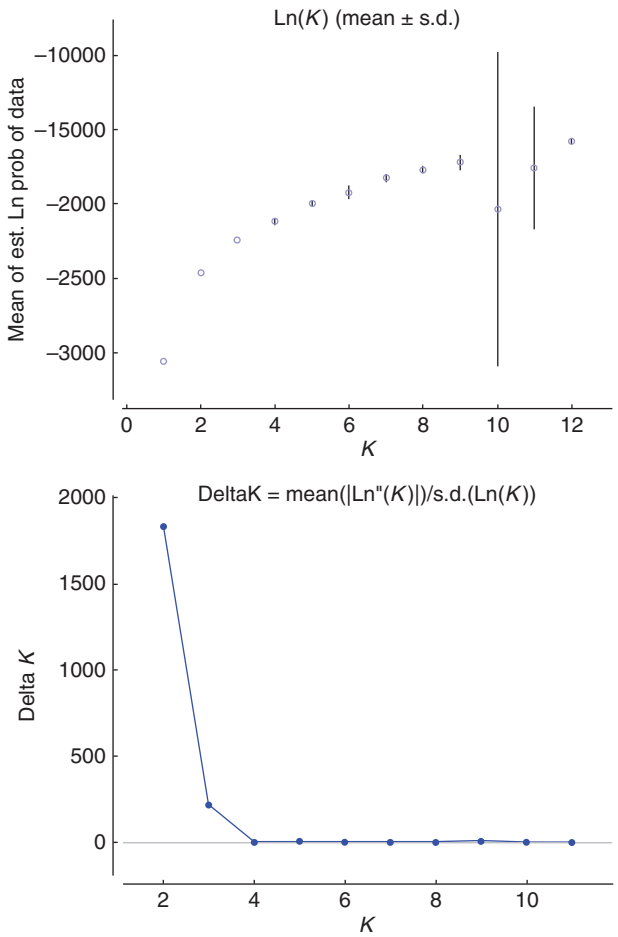

FIG. 3. Methodology from Evanno *et al.* (2005) for the interpretation of STRUCTURE results obtained on a set of 208 cultivated diploids accessions genotyped with 498 DArT markers. Median Ln(*K*), its variance across runs and median Δ*K* are presented for each value of *K*. The two peaks of median Δ*K* at *K* = 2 and 3 indicate two putatively correct values for *K*.

TABLE 2. *Partitioning of 208 edible diploid accessions for the two genetic clusters identified by STRUCTURE; an accession is considered belonging to a cluster when* Q > 0·8

	Cluster 1	Cluster 2	Admixed	Total
India	2	0	0	2
Vietnam	2	0	1	3
Thailand	5	0	0	5
Philippines	3	1	11	15
Malaysia	15	0	5	20
Indonesia/Malaysia	2	0	2	4
Indonesia	6	0	5	11
Papua	10	82	42	134
Madagascar	1	0	0	1
Comoros	0	0	2	2
Zanzibar	0	1	0	1
Tanzania	0	0	2	2
Brazil	1	0	0	1
Trinidad & Tobago	0	0	1	1
Origin not known	3	0	3	6
Total	50	84	74	208

closely related accessions within the sample. We investigated the MLGs clustering at threshold 8 and observed that, for this value, the cultivated triploids displayed 78 different MLGs out of the 273 accessions of the sample (Table S1). Thirty-one of the identified MLGs were composed of 2–44 accessions while 47 were composed of unique accessions. For 27 of the unique MLGs, no taxonomy information was available while 20 were classified as belonging to known subgroups. Noticeably, ITC0686 'Pisang Umbuk', ITC0176 'Lacatan' and ITC0002 'Dwarf Cavendish' classified as Cavendish are here considered as unique clones when 37 accessions classified as Cavendish and Gros Michel are considered as belonging to the same clone. Equally, ITC0060 'Guineo', ITC0170 'Ingarama' and ITC0177 'Makara' are considered unique genotypes but are classified as Mutika/Lujugira when 37 other Mutika/Lujugira accessions are considered as a single clone.

We also noted that AAB Plantain was considered here as a unique clone but Iholena and Silk were composed of two sets of clones each. Most of the Pome and Mysore accessions were considered as a unique clone.

With few exceptions, the results obtained for ABB are consistent with the taxonomy for the subgroups Pisang Awak, Pelipita and Klue Teparod. However, they are not consistent for the accessions classified as Saba, Monthan, Bluggoe, Ney Mannan or Peyan, for which the accessions belong to several MLGs that are themselves a mix of the different subgroups.

DISCUSSION

DArT markers and the characterization of the diversity in Musa

Molecular markers have proved to be useful tools for the resolution of banana taxonomy and management of *ex situ* collections (Hippolyte *et al.*, 2012; de Jesus *et al.*, 2013; Irish *et al.*, 2014). Here we analysed a wide sample of wild and cultivated bananas conserved in the more diverse of the *Musa* genebank, the ITC, with 498 DArT markers. Overall, the clustering of the accessions within our sample is consistent with the acknowledged taxonomy of banana. Compared to a previous study performed with SSR markers (Hippolyte *et al.*, 2012), the clustering of the accessions is consistent and similar. However, the organization of the clusters differs as the tree built with SSR markers did not show an organization of these clusters according to their genomic composition, as is the case here, but according to their common ancestry. Therefore, DArT appears more robust in detecting the genomic composition of accessions, especially in estimating the number of B genomes displayed by each sample (Fig. 1). With regard to the dominant nature of the markers used, the hierarchical clustering of the accessions according to the number of B copies present in their genomic composition is surprising but the same pattern was observed with dominant AFLP markers (Ude *et al.*, 2002). More surprising is the clustering of both accessions classified as Klue Teparod (ABB) within the wild *M. balbisiana* sample. Some authors have claimed the occurrence of parthenocarpic BBB cultivars (Valmayor *et al.*, 2000). Ribosomal DNA analysed for one of these accessions, ITC0652 'Kluaï Tiparot', indeed revealed a B genome component only (Boonruangrod *et al.*, 2008) while internal transcribed spacer (ITS) sequence and cytogenetic analyses of satellite DNA unambiguously confirmed

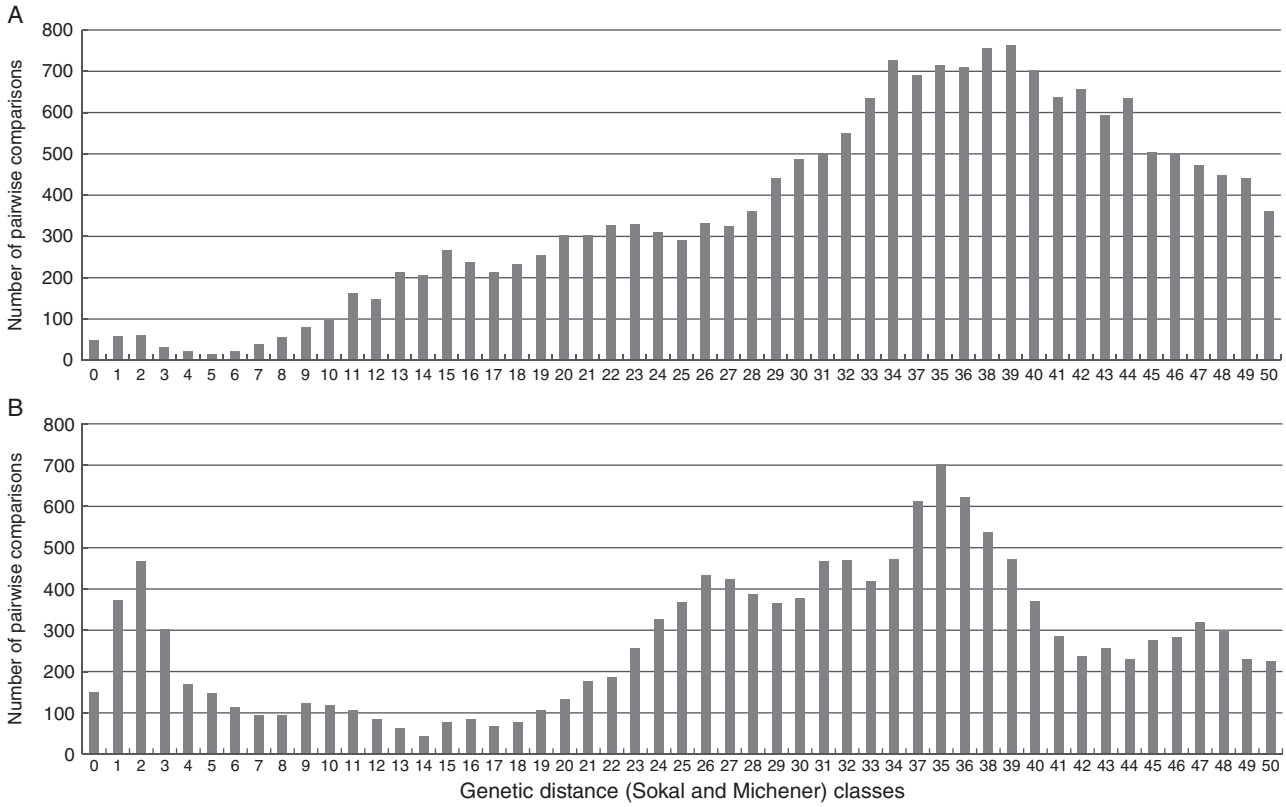

FIG. 4. Histograms of genetic distances for the first 50 classes calculated following the Sokal and Michener similarity index for (A) 208 cultivated diploids and (B) 269 cultivated triploids and four cultivated mixoploids.

the presence of rDNA loci specific to the B as well as to the A genome in the second accessions, ITC0473 'Balonkawe' (Hřibová *et al*., 2011; Čížková *et al*., 2013). Therefore, the potential occurrence of an incomplete A genome within this subgroup needs to be investigated.

Several true duplicates were identified within the MLGs identified by GenoType (see Table S1). However, in most cases they did not cluster together as Th0 (data not shown) revealing that the amount of genetic variation generated by the 'genotyping error' may be equivalent to that between accessions clonally derived from each other. Therefore, although DArT is confirmed as providing reliable markers for estimating and studying the diversity present in a *Musa* germplasm collection, the issue of providing a molecular footprint that would enable the unambiguous discrimination of each particular cultivar cannot be resolved with DArT markers. In such a context, the platform and methodology using a set of SSR markers presented by Christelová *et al*. (2011) is likely to be more accurate.

DArT markers also highlighted discrepancies between the known genetic background of some of the accessions and their clustering in the diversity analysis. For example, ITC1253 'Mjenga' was considered on the basis of morphological and SSR data as a clone belonging to the Mshale subgroup (Hippolyte *et al*., 2012), whereas it clusters here within the Papuan cultivated accessions and not with the other Mshale. We thus suspect a mislabelling problem, either in the ITC or during the DNA processing. On the other hand, the discrepancies observed between the taxonomy of some wild accessions and their clustering in the STRUCTURE analysis may be due either to their erroneous classification or to their hybrid status as explained at $K = 8$ for some of the wild samples (e.g. several *burmannica* accessions). Although such hybridization could be due to the occurrence of natural and regular gene flow between the different genepools of *M. acuminata* (Carreel *et al*., 1994), we cannot exclude that they hybridized and accidentally lost their genetic integrity when maintained in *ex situ* field genebanks (Visser and Ramanatha Rao, 2005) prior their introduction to the ITC. Equally, the patterns displayed by some of the ABB subgroups, in which memberships to sets of clones do not follow the taxonomy provided with the accessions, suggest the erratic classification of these accessions. Both types of discrepancy will be investigated through field verification that will allow the growth, characterization and documentation under standard conditions of the accessions concerned followed by expert consultation (Chase *et al*., 2016). Low cost, fast, accurate and applicable to the whole genome, DArT markers are good tools to help manage *ex situ* collections of banana.

Organization of the diversity and implications for its origin

Wild samples. Examining the successive partitioning displayed by STRUCTURE for the 94 wild accessions according to the number of clusters considered is particularly interesting. As postulated by Meirmans (2015), most of the wild species and

populations exhibit different levels of organization in their genetic structure that can be reflected by different possible values of *K*. With an increase of *K*, we progressively discriminate the different species and subspecies involved in this study consistently with the phylogenetic results published by Janssens *et al.* (2016), the only surprising pattern being the fusion of *M. schizocarpa* and *M. acuminata burmannica* at $K = 3$. Equally, the species *M. schizocarpa* originates in Papua but in the PCoA (Fig. 1) it clusters closer to the South-East Asian subspecies of *M. acuminata* than to the *banksii* from Papua.

Cultivated samples. The rise of cultivated triploid bananas from their direct wild ancestors, *M. acuminata* and *M. balbisiana* among others, can be seen as a three-step process in which the anthropogenic circulation of pre-domesticated forms of diploid bananas extracted from the different wild genepools (Step 1) led to the production of edible and diploid hybrids (Step 2) that occasionally produced unreduced gametes and resulted in the emergence of triploid varieties (Step 3). The founder event that is common to steps 2 and 3 is sexual reproduction. First, sexual recombination led, within cultivated plots, to the birth of diploid specimens suitable for food consumption; second, rare sexual events still occurring among the edible diploids gave birth to triploids (Perrier *et al*., 2011). Therefore, identification of the number of distinct MLGs in both edible diploid and triploid accessions provides an estimation of the number of founding events for each ploidy type of banana and allows us to thus estimate the extent of the two consecutive bottlenecks that gave birth to present-day bananas. Our estimation of the number of MLGs constitutes a straightforward method for such estimations: the sample is wide and takes into account the biological specificity of each sample according to ploidy levels. We estimate that the 208 cultivated diploids of our sample may have arisen from 117 distinct sexual events while 80 sexual events may be at the origins of the 273 triploid accessions. The scores we obtained, in particular for the triploids, are low and highlight the narrowness of the genetic basis of the triploid bananas, despite what was hypothesized by Li *et al.* (2013) based on the study of nucleotide diversity in the *Waxy* and *Adh1* genes. Taking into account that the ITC is seeking the most diverse and rare cultivars for conservation purposes, the estimation given by Bakry and Horry (2016) that 95 % of world banana production relies on 7–14 sexual events is not challenged by our results. It merely highlights the extent of under-utilization of banana genetic resources.

The identification, using STRUCTURE, of two main genepools within the diploid samples, one corresponding to South-East Asia and the other to Papua, is consistent with what was described for other vegetatively propagated crops in the region, such as taro (*Colocasia esculenta* Schott.) (Krieke *et al*., 2004) and great yams (*Dioscorea alata* L.) (K. Abraham, CTCRI, and G. Arnau, CIRAD, pers.comm.), and supports the hypothesis of an independent centre of domestication for some crops, including banana, in Papua (Lebot, 1999). This view therefore challenges the acknowledged representation of banana domestication, for which edible diploid cultivars arose from crosses between the different wild genepools, the structural heterozygosity of the genomes obtained being considered as a major force that underpinned the selection of unseeded cultivars (Perrier *et al*., 2011). We may therefore consider at least two different domestication centres for banana, one in South-East Asia and one in New Guinea, in which the selection forces that applied to domesticated bananas were probably different from the currently accepted representation of banana domestication.

Molecular markers and taxonomic resolution

The results obtained when estimating the putative number of MLGs, i.e. of sexual events that occurred within our sample, are of particular interest for taxonomic purposes. This analysis supports the assumption that the subgroup Plantain originated from the vegetative diversification of a single seed (Noyer *et al*., 2005) as all Plantain are considered a single clone (Table S1). However, it does not discriminate Gros Michel from Cavendish, whereas these two subgroups were hypothesized as siblings with two different *n* gamete donors, 'Khai Nai On' and 'Pisang Pipit', respectively (Hippolyte *et al*., 2012). Despite this supposed difference, the level of genetic divergence assessed with DArT markers between Gros Michel and Cavendish is equivalent to that observed for a monoclonal subgroup. In contrast, subgroups such as AA Pisang Jari Buaya, AB Ney Poovan, AAA Cavendish, AAA Mutika/Lujugira, AAB Silk and AAB Iholena seem to be composed of several clonal entities each. We cannot exclude that this pattern partly results from the potential erroneous classification of some clones, although the recent study of Kagy *et al.* (2016) confirmed the occurrence of polyclonal subgroups. The question raised by such a pattern is the definition of subgroups. Do we consider only monoclonal sets as subgroups *sensus stricto* or do we accept that a subgroup is likely to be composed of different clonal entities? In their paper considering the Iholena subgroup, defined based on its particular fruit and bunch morphology, Kagy *et al.* (2016) observed that this Pacific AAB subgroup was indeed composed of at least two different but related genotypes and postulated that they probably arose from the same restricted subset of parental diploids. Therefore, we may acknowledge that a subgroup could arise from different sexual events that occurred within the same genepools, conditional to morphological similarity. In such a context, molecular markers are of great help in detecting evolutionary differences underlying the emergence of subgroups. However, revising the taxonomy of banana requires joint morphological and molecular characterization of ambiguous accessions to check their classification and, if necessary, to refine the morphological criteria delimitating the subgroups concerned.

CONCLUSIONS

We have conducted one of the largest and most comprehensive studies of the genetic diversity of banana germplasm. We confirmed that DArT markers were good tools both for resolving the taxonomy of accessions and for identifying mislabelling problems. The identification of two main genepools in the cultivated diploid accessions suggests at least two main regions of domestication, one in New Guinea and one, if not more, in South-East Asia. If it is consistent to hypothesize that the Papuan cultivars were domesticated from the local subspecies *M. acuminata banksii*, the South-East Asian domestication scheme is probably far more complex as it involves several

subspecies. These subspecies are far from well known. As we postulate here, many of the accessions conserved in the ITC, and thus in their source collection, are likely to be hybrids between two or more genepools rather than pure representatives of their taxa. Whether hybridization occurred during their conservation in field *ex situ* collections or in the wild prior to being collected is not clear. The poor representation of some of the *M. acuminata* subspecies in *ex situ* genebanks does not help to clarify this issue. A striking example is *Musa acuminata errans* that was described in the Philippines (Valmayor, 2001). Currently, the only available specimen affiliated to this subspecies is ITC1028 'Agutay' and it appears here that it may well be a *banksii* hybrid. It is thus not possible to strictly assess, from this given accession, if *errans* participated in the build-up of cultivated bananas. The large group of AA/AAA cultivated bananas that does not cluster with any of the *M. acuminata* subspecies present in our sample suggests in addition that not all the diversity of the wild *M. acuminata* has been studied. To fill these gaps in both our knowledge and in the available genetic resources, systematic prospecting coupled with thorough phylogenetics and population studies should be undertaken in the future.

SUPPLEMENTARY DATA

Supplementary data are available online at www.aob.oxfordjournals.org and consist of Table S1: assignment of clonal IDs to the sample accessions.

ACKNOWLEDGEMENTS

We wish to thank all the countries, institutions and people who contributed to the ITC, helping it achieve its mission: the long-term security of banana genetic resources and holding the collection in trust for the benefit of future generations. We also thank the volunteer experts of the Taxonomy Advisory Group of MusaNet for their ongoing efforts in addressing issues linked to the taxonomy and classification of banana. Thanks to Rachel Chase for final editing of the manuscript. The Direction générale coopération au développement et Aide humanitaire (DGD, Belgium) supported this work, implemented between 2008 and 2010, which has been critical to the effective management of the ITC. In addition, this work would not have been possible without the support and authorization of RTB, the CGIAR Research Program on Roots, Tubers and Bananas. E.H., P.C. and J.D. were supported by the Czech Ministry of Education, Youth and Sports (grant awards LO1204 from the National Program of Sustainability I, and LG-15017).

LITERATURE CITED

Akbari M, Wenzl P, Caig V, *et al.* 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theoretical and Applied Genetics* **113**: 1409–20.

Amorim, EP, Vilarinhos, AD, Cohen, KO, *et al.* 2009. Genetic diversity of carotenoid-rich bananas evaluated by Diversity Arrays Technology (DArT). *Genetics and Molecular Biology* **32**: 96–103.

Bakry F, Horry JP. 2016. Advances in genomics: applications to banana breeding. *Acta Horticulturae* **1114**: 171–180.

Boonruangrod R, Desai D, Fluch S, Berenyi M, Burg K. 2008. Identification of cytoplasmic ancestor gene-pools of *Musa acuminata* Colla and *Musa balbisiana* Colla and their hybrids by chloroplast and mitochondrial haplotyping. *Theorical and Applied Genetics* **118**: 43–55.

Bouchet S, Pot D, Deu M, *et al.* 2012. Genetic structure, linkage disequilibrium and signature of selection in sorghum: lessons from physically anchored DArT markers. *PLoS One* **7**: e33470. doi:10.1371/journal.pone.0033470.

Carreel F, Fauré S, González de León D, *et al.* 1994. Evaluation de la diversité génétique chez les bananiers diploïdes (*Musa* sp). *Genetics Selection Evolution* **26** (S1): 125s–136s.

Carreel F, Gonzalez de Leon D, Lagoda P, *et al.* 2002. Ascertaining maternal and paternal lineage within *Musa* by chloroplast and mitochondrial DNA RFLP analyses. *Genome* **45**: 679–692.

Chase R, Sardos J, Ruas M, *et al.* 2016. The field verification activity: a cooperative approach to the management of the global *Musa* in-vitro collection at the International Transit Centre. *Acta Horticulturae* **1114**: 61–65.

Christelová P, Valarik M, Hribova E, *et al.* 2011. A platform for efficient genotyping in Musa using microsatellite markers. *AoB PLANTS* 2011 plr024 doi:10.1093/aobpla/plr024.

Čížková J, Hribova E, Humplíková L, Christelová P, Suchánková P, Dolezel J. 2013. Molecular analysis and genomic organization of major DNA satellites in banana (*Musa* spp.). *PLoS One* **8**(1): e54808. doi:10.1371/journal.pone.0054808.

D'Hont A, Denoeud F, Aury JM, *et al.* 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**: 213–217.

de Jesus ON, de Oliveira e Silva S, Amorim EP, *et al.* 2013. Genetic diversity and population structure of *Musa* accessions in *ex situ* conservation. *BMC Plant Biology* **13**: 41.

De Langhe E, Vrydaghs L, De Maret P, Perrier X, Denham TP. 2009. Why bananas matter: an introduction to the history of banana domestication. *Ethnobotany Research and Applications* **7**: 165–177.

Douhovnikoff V, Dodd RS. 2003. Intra-clonal variation and a similarity threshold for identification of clones: application to *Salix exigua* using AFLP molecular markers. *Theorical and Applied Genetics* **106**: 1307–1315.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**: 359–361.

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.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure: extensions to linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Hippolyte I, Bakry F, Séguin M, *et al.* 2010. A saturated SSR/DArT linkage map of *Musa* acuminate addressing genome rearrangements among bananas. *BMC Plant Biology* **10**: 65.

Hippolyte I, Jenny C, Gardes L, *et al.* 2012. Foundation characteristics of edible Musa triploids revealed from allelic distribution of SSR markers. *Annals of Botany* **109**: 937–951.

Hřibová E, Cizkova J, Christelová P, Taudien S, de Langhe E, Dolezel J. 2011. The ITS1–5.8S-ITS2 sequence region in the Musaceae: structure, diversity and use in molecular phylogeny. *PLoS One* **6**: e17863. doi:10.1371/journal.pone.0017863.

Irish BM, Cuevas HE, Simpson SA, *et al.* 2014. *Musa* spp. germplasm management: microsatellite fingerprinting of USDA–ARS national plant germplasm system collection. *Crop Science* **54**: 2140–2151.

Jaccoud D, Peng K, Feinstein D, Kilian A. 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research* **29**: e25. doi:10.1093/nar/29.4.e25.

Janssens SB, Vandelook F, De Langhe E, *et al.* 2016. Evolutionary dynamics and biogeography of Musaceae reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. *New Phytologist* **210**: 1453–1465.

Kagy V, Wong M, Vandenbroucke H, *et al.* 2016. Traditional banana diversity in Oceania: an endangered heritage. *PLoS One* **11**: e0151208. doi:10.1371/journal.pone.0151208.

Kilian A, Wenzl P, Huttner E, *et al.* 2012. Diversity Arrays Technology: a generic genome profiling technology on open platforms. *Methods in Molecular Biology* **888**: 67–89.

Krieke CM, Van Eck HJ, Lebot V. 2004. Genetic of taro, *Colocasia esculenta* (L.) Schott, in Southeast Asia and the Pacific. *Theorical and Applied Genetics* **109**: 761–768.

Lebot V. 1999. Biomolecular evidence for plant domestication in Sahul. *Genetic Resources and Crop Evolution* **46**: 619–628.

Li LF, Wang HY, Zhang C, *et al.* 2013. Origins and domestication of cultivated banana inferred from chloroplast and nuclear genes. *PLoS One* **8**(11): e80502. doi:10.1371/journal.pone.0080502.

Meirmans PG, Van Tienderen PH. 2004. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* **4**: 792–794.

Meirmans PG. 2015. Seven common mistakes in population genetics and how to avoid them. *Molecular Ecology* **24**: 3223–3231.

Noyer JL, Causse S, Tomepke K, Bouet A, Baurens FC. 2005. A new image of plantain diversity assessed by SSR, AFLP and MSAP markers. *Genetica* **124**: 61–69.

Perrier X, Jacquemoud-Collet JP. 2006. *DARwin software version 5.0*. CIRAD: http://www.darwin.cirad.fr/darwin.

Perrier X, Flori A, Bonnot F. 2003. Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC, eds. *Genetic diversity of cultivated tropical plants*. Montpellier: Enfield, Sciences Publisher, 43–76.

Perrier X, Bakry F, Carreel F, *et al.* 2009. Combining biological approaches to shed light on the evolution of edible bananas. *Ethnobotany Research and Applications* **7**: 199–216.

Perrier X, De Langhe E, Donohue M, *et al.* 2011. Multidisciplinary perspectives on banana (*Musa* spp.) domestication. *Proceedings of the National Academy of Sciences of the USA* **108**: 11311–11318.

Pritchard JK, Stephens M, Donnelly PJ. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Raboin LM, Carreel F, Noyer JL, *et al.* 2005. Diploid ancestors of triploid export banana cultivars: molecular identification of 2n restitution gamete donors and n gamete donors. *Molecular Breeding* **16**: 333–341.

Risterucci AM, Hippolyte I, Perrier X, *et al.* 2009. Development and assessment of diversity arrays technology for high-throughput DNA analyses in *Musa*. *Theoretical and Applied Genetics* **119**: 1093–1103.

Roorkiwal M, von Wettberg EJ, Upadhyaya HD, Warschefsky E, Rathore A, Varshney RK. 2014. Exploring germplasm diversity to understand the domestication process in *Cicer* spp. using SNP and DArT Markers. *PLoS One* **9**: e102016. doi:10.1371/journal.pone.0102016.

Shepherd K. 1999. *Cytogenetics of the genus Musa*. Montpellier: INIBAP.

Simmonds NW. 1962. *The evolution of bananas*. London: Longmans.

Simmonds NW, Shepherd K. 1955. The taxonomy and origins of the cultivated bananas. *Journal of the Linnean Society of London* **55**: 302–312.

Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. *Scientific Bulletin of the University of Kansas* **38**: 1409–1438.

Ude G, Pillay M, Nwakanma D, Tenkouano A. 2002. Genetic diversity in *Musa acuminata* Colla and *Musa balbisiana* Colla and some of their natural hybrids using AFLP markers. *Theoretical and Applied Genetics* **104**: 1246–1252.

Valmayor RV. 2001. Classification and characterization of *Musa exotica*, *M. alinsanaya* and *M. acuminata* ssp. *errans*. *Infomusa* **10**: 35–39.

Valmayor RV, Jamaluddin SH, Silayoi B, *et al.* 2000. *Banana cultivars names and synonyms in South-East Asia*. Los Baños: INIBAP.

Visser B, Ramanatha Rao V. 2005. Gene flow and the management of *ex-situ* collections. In: de Vicente C, ed. *Issues on gene flow and germplasm management*. *Topical Reviews in Agricultural Biodiversity*. Rome: IPGRI.

Wittenberg AH, van der Lee T, Cayla C, Kilian A, Visser RG, Schouten HJ. 2005. Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. *Molecular Genetics and Genomics* **274**: 30–39.