the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2020) **102,** 1008–1025 doi: 10.1111/tpj.14683

Genome ancestry mosaics reveal multiple and cryptic contributors to cultivated banana

Guillaume Martin[1,2,*], Céline Cardi[1,2], Gautier Sarah[2], Sébastien Ricci[2,3,4], Christophe Jenny[1,2], Emmanuel Fondi[3], Xavier Perrier[1,2], Jean-Christophe Glaszmann[1,2], Angélique D'Hont[1,2] and Nabila Yahiaoui[1,2]

[1]*CIRAD, UMR AGAP, F-34398 Montpellier, France,*

[2]*AGAP, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France,*

[3]*CARBAP, Rue Dinde, No. 110, Bonanjo, BP 832 Douala, Cameroon, and*

[4]*CIRAD, UMR AGAP, F-97130 Capesterre Belle Eau, France*

Received 4 July 2019; revised 18 December 2019; accepted 2 January 2020; published online 13 January 2020.
*For correspondence (e-mail guillaume.martin@cirad.fr).

SUMMARY

Hybridizations between closely related species commonly occur in the domestication process of many crops. Banana cultivars are derived from such hybridizations between species and subspecies of the *Musa* genus that have diverged in various tropical Southeast Asian regions and archipelagos. Among the diploid and triploid hybrids generated, those with seedless parthenocarpic fruits were selected by humans and thereafter dispersed through vegetative propagation. *Musa acuminata* subspecies contribute to most of these cultivars. We analyzed sequence data from 14 *M. acuminata* wild accessions and 10 *M. acuminata*-based cultivars, including diploids and one triploid, to characterize the ancestral origins along their chromosomes. We used multivariate analysis and single nucleotide polymorphism clustering and identified five ancestral groups as contributors to these cultivars. Four of these corresponded to known *M. acuminata* subspecies. A fifth group, found only in cultivars, was defined based on the 'Pisang Madu' cultivar and represented two uncharacterized genetic pools. Diverse ancestral contributions along cultivar chromosomes were found, resulting in mosaics with at least three and up to five ancestries. The commercially important triploid Cavendish banana cultivar had contributions from at least one of the uncharacterized genetic pools and three known *M. acuminata* subspecies. Our results highlighted that cultivated banana origins are more complex than expected – involving multiple hybridization steps – and also that major wild banana ancestors have yet to be identified. This study revealed the extent to which admixture has framed the evolution and domestication of a crop plant.

Keywords: admixture, genome ancestry, hybridization, *Musa acuminata*, diversity.

INTRODUCTION

Hybridization between species and subspecies is a widespread evolutionary process in plants and is associated with the domestication and diversification of some major crops [e.g. wheat (McFadden and Sears, 1946); citrus (Wu *et al.*, 2014); date palm (Flowers *et al.*, 2019)], including bananas (Simmonds, 1962; Perrier *et al.*, 2011).

Banana cultivars are the result of hybridization between *Musa* species and subspecies that have diverged in different tropical Southeast Asian regions and islands and western Melanesia (Dodds *et al.*, 1948; Simmonds and Shepherd, 1955; Simmonds, 1962; Carreel *et al.*, 1994; Carreel *et al.*, 2002; Boonruangrod *et al.*, 2008; Perrier *et al.*, 2009; Perrier *et al.*, 2011). The species *Musa acuminata* (A genome, 2n = 22) is involved in the origin of all cultivars, with the exception of the Fe'i vitamin-rich bananas (De Langhe *et al.*, 2009). *M. balbisiana* (B genome, 2n = 22) is associated with *M. acuminata* in many cultivars, while *M. textilis* (T) and *M. schizocarpa* (S) are suggested to have contributed to only a few cultivars. *M. acuminata* has been subdivided into several subspecies with a Northeast India to New Guinea distribution range, including: *siamea*, *burmannica*, *burmannicoïdes*, *malaccensis*, *truncata*, *errans*, *microcarpa*, *zebrina* and *banksii* (Simmonds, 1962; Perrier *et al.*, 2009). Large chromosomal structural variations between the genomes of some of these species and subspecies have been reported (Shepherd, 1999; Martin *et al.*, 2017; Baurens *et al.*, 2019).

In the prevalent banana domestication scenario, fertile plants from geographically isolated *M. acuminata*

subspecies were brought into contact by humans in Southeast Asia and western Melanesia (De Langhe *et al.*, 2009; Kennedy, 2009; Perrier *et al.*, 2011). This gave rise to intersubspecific hybridizations which first resulted in diploid hybrids with reduced fertility. Additional hybridizations within *M. acuminata* or with other *Musa* species, sometimes involving 2n gamete formation, led to the current diploid and triploid banana cultivar diversity. These cultivars were selected for seedless and parthenocarpic fruits, a condition for edibility. They were classified – based on morphology and ploidy – into genomic groups ('AA', 'AAA', 'AB', 'AAB', 'ABB', 'AS', 'AT' and 'AAT') to reflect the main species contributing to their genomes (Simmonds and Shepherd, 1955). The selected cultivars have high levels of sterility and have thus been vegetatively propagated for centuries or millennia, resulting in phenotypic somaclonal variants. The variants that are presumed to have derived from a single original zygote form a subgroup.

The most commercially important banana cultivars are triploid. Among these, the Cavendish subgroup of dessert bananas ('AAA' genome) represents almost half of the world's production (Lescot, 2018). Diploid cultivars derived from *M. acuminata* ('AAcv') are mainly found in New Guinea, but a few of them are present in Malaysia, Indonesia, the Philippines and East Africa (Stover and Simmonds, 1987; Perrier *et al.*, 2019).

Information regarding the contributions of different *Musa* species and subspecies to cultivated banana are essential to gain insight into the domestication process. They are also important for banana breeding strategies that are generally geared towards producing polyploid cultivars with agronomic characteristics similar to those of current elite cultivars, in addition to some improved traits, particularly disease resistance. Three genetically well differentiated subspecies, that is *M. acuminata* ssp. *banksii*, ssp. *zebrina* and ssp. *malaccensis*, were suggested to be the main contributors to the A genome of cultivated bananas (Carreel *et al.*, 2002; Perrier *et al.*, 2009; Perrier *et al.*, 2011; Christelová *et al.*, 2017). Their contributions were found to be dependent on the geographical origins of the cultivated accessions, with an eastern Indonesian and New Guinean pool based mainly on ssp. *banksii* and *zebrina* and a greater ssp. *malaccensis* contribution further westwards (Perrier *et al.*, 2009). It seems that there has just been a minor contribution of the closely related *burmannica*, *burmannicoïdes* and *siamea* subspecies to cultivars (Carreel *et al.*, 1994; Perrier *et al.*, 2011). The possible contributions of *M. acuminata* ssp. *errans* and *microcarpa* and the delimitation of these subspecies are less clear (Carreel, 1994; Carreel *et al.*, 2002; Perrier *et al.*, 2009). Parental donors have been proposed for some commercially important triploid cultivars (Raboin *et al.*, 2005; Perrier *et al.*, 2009; Hippolyte *et al.*, 2012). Triploid Cavendish dessert bananas, for instance, are considered to have resulted from a cross between a diploid accession from the East African Mchare (formerly Mlali) subgroup as 2n gamete donor and an n gamete donor that is related to the 'Pisang Madu' and 'Pisang Pipit' accessions.

Beyond global ancestry estimates, local ancestry patterns along chromosomes provide more precise information on the genomic composition of hybrids and how they were formed, as shown in *Citrus* sp. or cassava (Wu *et al.*, 2014; Bredeson *et al.*, 2016). Banana cultivars – having low fertility and being vegetatively propagated – are generally believed to have resulted from a limited number of crosses and thus meioses since the first inter(sub)specific events. Their genomes are expected to be mosaics of large segments from different origins (i.e. ancestries). Such mosaics are currently being unravelled at the interspecies resolution level for A/B interspecific hybrids (Baurens *et al.*, 2019), but they have yet to be characterized for A genome based hybrids.

In this study, we used single nucleotide polymorphism (SNP) data to characterize the ancestral origins of genome segments along banana chromosomes for a set of 24 diploid banana accessions, including 15 wild (i.e. seedy) banana accessions and nine diploid cultivars (i.e. parthenocarpic accessions) sampled across 'AAcv' diversity. The results showed a variety of genome mosaics, sometimes resulting from more than four ancestral contributors and including unknown ancestry components. The mosaic structure of the triploid 'Grande Naine' accession belonging to the dessert banana Cavendish subgroup was also investigated.

RESULTS

Highly variable heterozygosity and accession-specific allele proportions among accessions

RNA-seq data were obtained from a set of 14 wild and nine diploid cultivated banana accessions (Table 1). The wild accessions included representatives of *M. acuminata* subspecies reported to be involved in the origin of hybrid 'AAcv' cultivars. The diploid cultivars represented different geographical origins and were selected among the 'AAcv' diversity described by Perrier *et al.* (2011). Genomic data from the triploid 'Grande Naine' accession representing the Cavendish cultivars were also used. *M. balbisiana* was used as an outgroup with the draft genome data (Davey *et al.*, 2013) from the 'Pisang Klutuk Wulung' (hereafter 'PKW') accession.

Depending on the accessions, 13.9–94.4 mapped million reads were obtained for polymorphism detection, representing 16.6 to 58.9-fold the *M. acuminata* predicted transcriptome or reference genome coverage per accession (Table S1). A set of 191 876 high confidence nuclear SNP sites, with a distribution reflecting gene density (Figure S1), was selected from these sequences.

Table 1 List of accessions used in the study

Species or group (subgroup)	Origin	Code	Accession name	Geographical origin	Status	Ploidy
M. balbisiana	SRA	SRX339427	Pisang Klutuk Wulung	–	Wild	2x = 22
M. acuminata						
ssp. *siamea*	CRB-PT	PT-BA-00147	Khae (Phrae)	Thailand (collection)	Wild	2x = 22
ssp. *siamea*	CRB-PT	PT-BA-00263	Pa Rayong	Thailand (collection)	Wild	2x = 22
ssp. *burmannicoides*	CRB-PT	PT-BA-00051	Calcutta 4	Botanical Garden India	Wild	2x = 22
ssp. *burmannica*	CRB-PT	PT-BA-00178	Long Tavoy	Myanmar	Wild	2x = 22
ssp. *malaccensis*	CRB-PT	PT-BA-00267	PT-BA-00267	–	Wild	2x = 22
ssp. *malaccensis*	CRB-PT	PT-BA-00363	Selangor	Malaysia	Wild	2x = 22
ssp. *malaccensis*	CRB-PT	PT-BA-00390	THA 018	Thailand (collection)	Wild	2x = 22
ssp. *microcarpa*	CRB-PT	PT-BA-00040	Borneo	Borneo	Wild	2x = 22
ssp. *microcarpa*	CRB-PT	PT-BA-00204	Microcarpa	Malaysia	Wild	2x = 22
ssp. *zebrina*	CRB-PT	PT-BA-00182	Maia Oa	Martinique	Wild	2x = 22
ssp. *zebrina*	CRB-PT	PT-BA-00212	Monyet	Indonesia	Wild	2x = 22
ssp. *banksii*	CARBAP	CMR00429	Banksii ITC0853	Papua New Guinea	Wild	2x = 22
ssp. *banksii*	CARBAP	CMR00427	Banksii ITC0620	Papua New Guinea	Wild	2x = 22
ssp. *errans*	CRB-PT	PT-BA-00008	Agutay	Philippines	Wild	2x = 22
AAcv	CRB-PT	PT-BA-00190	Manang	Philippines	Cultivar	2x = 22
AAcv	CRB-PT	PT-BA-00108	Gu Nin Chiao	Singapore	Cultivar	2x = 22
AAcv (Figue/Sucrier)	CRB-PT	PT-BA-00154	Kirun	Papua New Guinea	Cultivar	2x = 22
AAcv	CRB-PT	PT-BA-00366	SF 215	Papua New Guinea	Cultivar	2x = 22
AAcv (Mchare)	CRB-PT	PT-BA-00056	Chicame	Comoros	Cultivar	2x = 22
AAcv	CARBAP	GAL	Galeo	Papua New Guinea	Cultivar	2x = 22
AAcv	CARBAP	GUY	Guyod	Philippines	Cultivar	2x = 22
AAcv	CRB-PT	PT-BA-00304	Pisang Madu	Sarawak, Malaysia	Cultivar	2x = 22
AAcv	CARBAP	MALA	Mala	Papua New Guinea	Cultivar	2x = 22
AAAcv (Cavendish)	CRB-PT	PT-BA-00104	Grande Naine	–	Cultivar	3x = 33

Heterozygous sites were generally found in higher proportions for cultivated versus wild accessions, in agreement with their presumed hybrid origins. The percentage of heterozygous sites ranged from 7.2% for the 'Guyod' accession, up to 16.1% for 'Pisang Madu' and 16.4% for triploid 'Grande Naine' (Table S2 and Figure 1a). The proportions of heterozygous sites for three of the wild accessions ('Microcarpa', 'Agutay PT-BA-00008' and 'PT-BA-00267') were within the range of those of cultivated hybrids, suggesting a hybrid status. Heterozygosity levels were very low for ssp. *banksii* accessions (Table S2 and Figure 1a), in accordance with their preferentially autogamous reproduction mode, which differs from the allogamy of other *M. acuminata* subspecies.

The proportion of accession-specific alleles ranged from 0.04% for 'Chicame' to 5.8% for 'Pisang Madu' within *M. acuminata* accessions and reached 24.7% for *M. balbisiana* 'PKW' (Table S2 and Figure 1b). The extremely low proportion of specific alleles in the 'Chicame' accession indicated that almost all of its alleles were shared with one or more other accessions in this setting. Similarly, the two *M. acuminata* ssp. *banksii* accessions shared almost all of their alleles with each other and/or with other accessions in this setting.

Global assessment of ancestral contributions to banana accessions

A global genetic structure analysis of the 24 diploid accessions using the ADMIXTURE program (Alexander *et al.*, 2009) suggested that $K = 6$ was the best number of ancestries fitting our dataset (Table S3).

At $K = 6$, 14 accessions were homogeneous for one of the six ancestries (Figure 1c). They corresponded to 13 of the 15 wild accessions and to cultivar 'Pisang Madu'. Unexpectedly, this cultivar was found to be homogenous for one ancestry that was detected only in cultivated accessions. The remaining ancestries generally corresponded to current knowledge on the banana genetic diversity represented here. *M. balbisiana*, *M. acuminata* ssp. *zebrina* and ssp. *malaccensis* each represented one ancestry, whereas ssp. *burmannica*, ssp. *burmannicoides* and ssp. *siamea* belonged to a single ancestry. The ssp. *banksii* accessions along with ssp. *microcarpa* 'Borneo' were found to be homogenous for one common ancestry, while the 'Microcarpa' accession was a hybrid. Surprisingly, the 'Agutay' accession – which was chosen to represent ssp. *errans* and expected to be closely associated with ssp. *banksii* (Carreel *et al.*, 1994; Perrier *et al.*, 2009; Christelová *et al.*, 2017) – showed 'malaccensis', 'zebrina' and 'burmannica/siamea' ancestries.

Identification of ancestry informative alleles for six predicted ancestral groups

Correspondence analysis (COA) was performed on the 14 accessions representative of the six ancestries (Figure 2 and Table S4).

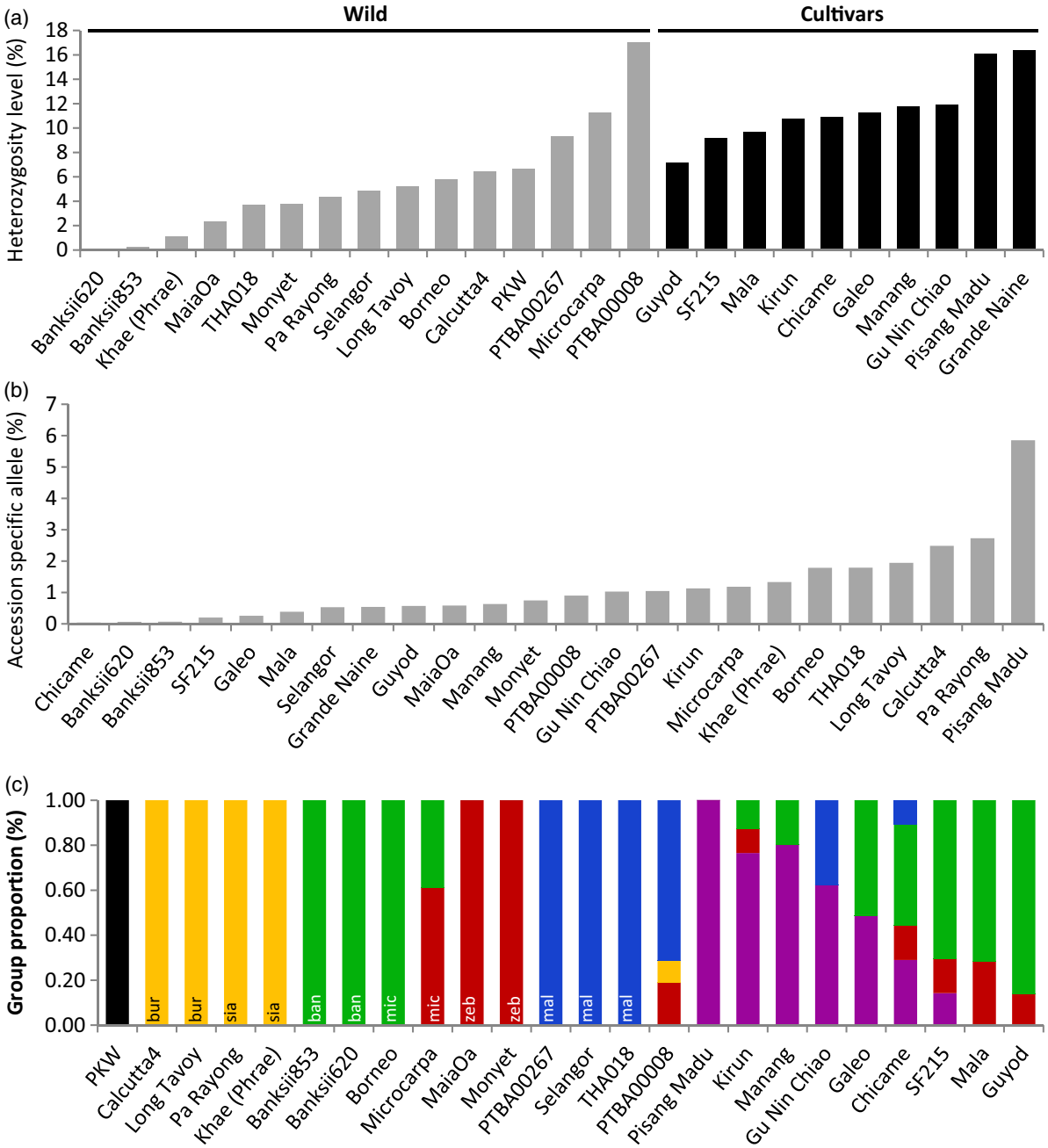

Figure 1. Global genotyping statistics.
(a) Heterozygosity levels among wild and cultivated banana accessions. Heterozygosity was calculated as the number of heterozygous sites within the accession divided by the total number of single nucleotide polymorphism sites in the vcf file.
(b) Proportion of specific alleles in the studied accessions. This proportion was calculated as the percentage of polymorphic sites in which at least one allele was not found in other accessions of the dataset.
(c) Global genetic structure of the dataset obtained via ADMIXTURE analysis with six ancestral populations.

Axes 1–5 allowed discrimination of *M. balbisiana*, of the distinct *M. acuminata* subspecies and of 'Pisang Madu' (Figure 2a,b and Table S4). A total of 74.5% inertia was accumulated on these five axes, in line with the structure found in the ADMIXTURE analysis (Figure 2c–e and Table S4). Axis 6 identified a substructure within *M. acuminata* ssp. *malaccensis* and *M. acuminata* ssp. *burmannica/siamea* clusters, but this resolution level was too fine to be taken into account given our sampling (Figure 2f).

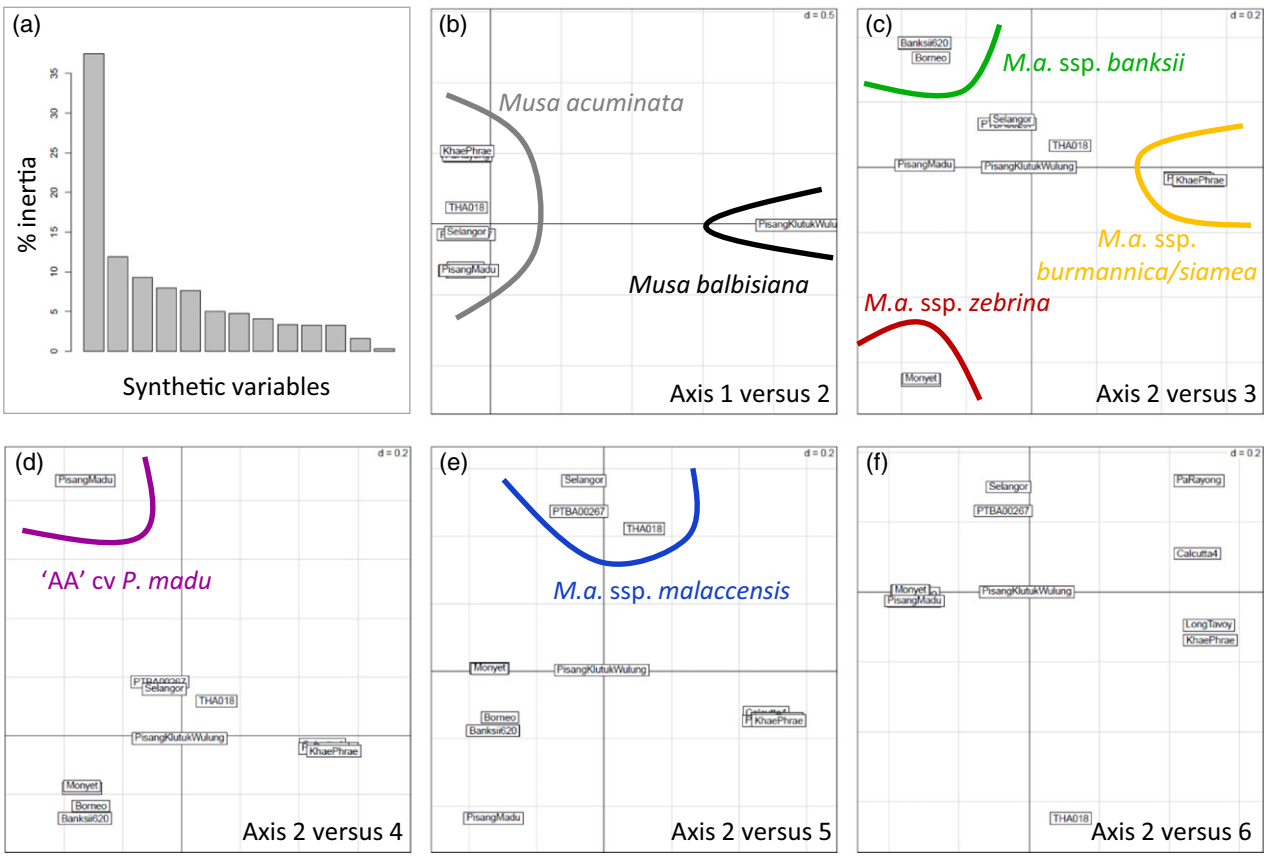

Figure 2. Factorial analysis performed on diploid accessions representing six ancestries. Correspondence analysis was performed only on diploid accessions identified as homogeneous according to ADMIXTURE analysis.
(a) Axis inertia.
(b) Projection of accessions along synthetic axes 1 and 2 discriminating *Musa acuminata* and *M. balbisiana* accessions.
(c) Projection of accessions along synthetic axes 2 and 3 discriminating *M. a.* ssp. *banksii*, *M. a.* ssp. *zebrina* and *M. a.* ssp. *burmannica/siamea* accessions.
(d) Projection of accessions along synthetic axes 2 and 4 discriminating the 'Pisang Madu' accession from other accessions.
(e) Projection of accessions along synthetic axes 2 and 5 discriminating *M. a.* ssp. *malaccensis* accessions from other accessions.
(f) Projection of accessions along synthetic axes 2 and 6.

To select representative alleles for the ancestral groups discriminated by the COA analysis, alleles were clustered on the first five axes based on their coordinates. Eight clustered allele groups were identified (Figure 3a and Table S5).

Group 1 (53 287 alleles) corresponded to *M. balbisiana* 'PKW' alleles. Alleles of group 2 were mainly derived from ssp. *burmannica/burmannicoides/siamea* accessions ('burmannica/siamea' group; 33 464 alleles). Group 3 (20 067 alleles) corresponded to ssp. *malaccensis* alleles. Group 4 (15 072 alleles) corresponded to ssp. *banksii* and ssp. *microcarpa* 'Borneo' accession alleles. The 'Borneo' accession contributed 35% of the alleles of this group, which we hereafter refer to as 'banksii/Borneo'. Alleles of group 5 (15 280 alleles) were derived from 'Pisang Madu'. Alleles of group 6 (11 640 alleles) corresponded to ssp. *zebrina*. Group 7 (1887 alleles) comprised alleles shared by 'Pisang Madu' and *M. balbisiana*. Finally, group 0 corresponded to central alleles (Figure 3a) for which no specific contribution to ancestral groups could be identified.

The relative proportions of alleles from groups 1 to 7 in all of the studied accessions are presented in Figure 3(b). The accessions selected to determine ancestral groups were clearly defined by their corresponding clustered alleles, but some ssp. *malaccensis* accessions and 'Pisang Madu' were not homogenous. In particular, the 'Pisang Madu' accession showed, as expected, a strong contribution of alleles from group 5, but also carried 'banksii/Borneo' alleles and a lower proportion of 'burmannica' and group 7 alleles. Group 7 alleles were low in number, located at intermediate positions in the COA analysis (in grey in Figure 3a) and their contribution was very limited. This group was not considered in further analysis. Overall, the distribution of six allele groups representing different ancestral origins was congruent with the six ancestries

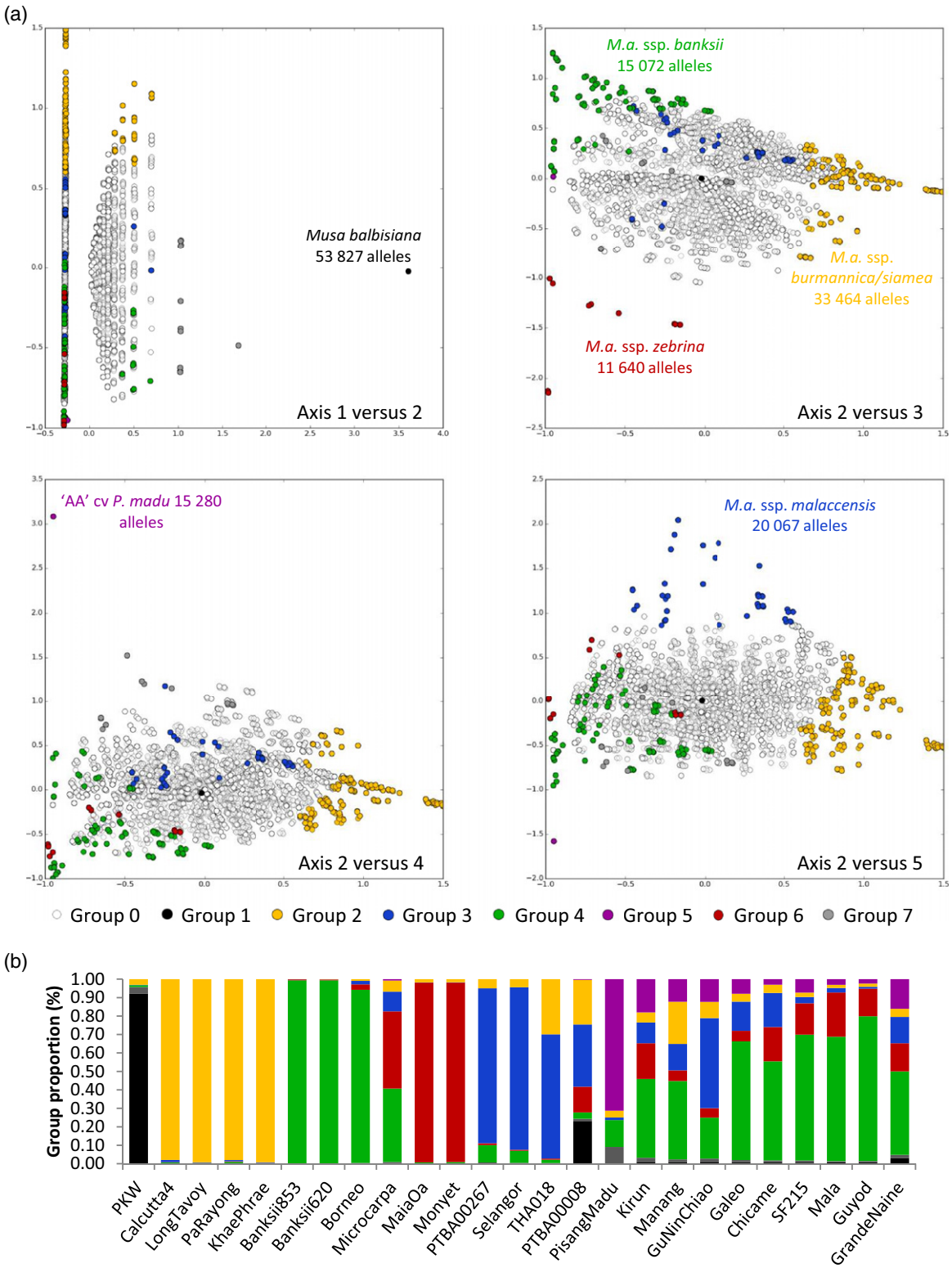

Figure 3. Clustering of ancestry informative alleles.
(a) Alleles corresponding to variables of the correspondence analysis were projected along synthetic axes and clustered using a mean shift algorithm. Eight groups (0, 1, 2, 3, 4, 5, 6 and 7) were identified.
(b) Proportions of alleles from each group in the 25 banana accessions. The proportion was calculated for each accession as the number of alleles specific to a group divided by the total number of grouped alleles in the accession. The central group (0), which corresponded to non-informative alleles, was not used for the proportion estimation.

identified with ADMIXTURE, but it is refined for the different banana hybrid accessions.

'Pisang Madu' local ancestry assignment

Clustered alleles representing groups 1–6 were used to analyze local ancestry along chromosomes of the 25 *Musa* accessions. The 13 accessions found to be the best representatives of the six ancestral groups (Figure 3b) were used for statistical assessment of the expected allele frequency that supported the ancestry assignment along chromosomes. For ancestral groups 1 to 4 and group 6, this probability $P_{i\alpha}$ was estimated by the observed frequency of the group α allele in representative accessions. 'Pisang Madu', although admixed, was the group 5 representative. In that case, we first arbitrarily set $P_{i\alpha} = 1$ if at least one allele of group 5 was found in 'Pisang Madu', and $P_{i\alpha} = 0$ in other cases. In this first round of evaluation of local ancestry, no group 5 contribution was observed to accessions other than 'Pisang Madu' (Figure S2) in contradiction with results presented in Figure 3(b). 'Pisang Madu' chromosome pairs generally showed a group 5 haplotype associated with an unknown haplotype (grey in Figure 4a) and a group 5 haplotype associated with a group 4 haplotype in some chromosome regions. However, as 'Pisang Madu' was a representative of ancestral group 5 in the COA analysis, group 5 alleles should have been present in most of the 'Pisang Madu' genome, except for introgressed regions from other defined ancestral groups.

We considered that the obtained profile could have been related to the high proportion of heterozygous sites in 'Pisang Madu'. The frequencies of heterozygous grouped alleles in all accessions representing ancestral groups were compared (Figure 4b and Table S6). This revealed that nearly all 'Pisang Madu' alleles with a group 5 origin (98%) were in a heterozygous state and suggested that group 5 might represent two different ancestries. Therefore, the local ancestry analysis was performed again on all accessions, taking into account the heterozygosity of group 5 alleles (see Experimental procedures). For 'Pisang Madu', the results showed a group 5 origin for the majority of its chromosome pairs, apart from a few regions that were assigned to group 4 ('Banksii/Borneo') (Figure 4c). In addition, this revealed a contribution of the 'Pisang Madu' group to several cultivars (Figure S3).

Local ancestry patterns identify hybrid or locally admixed genomes in some wild *M. acuminata* accessions

Local ancestry predictions for *M. acuminata* ssp. *zebrina*, *banksii*, *burmannica*/*siamea* accessions (Figures 5 and S3) confirmed the global ancestry results and the absence of recent admixtures with other differentiated subspecies. These accessions were therefore good representatives of their corresponding subspecies. Two ssp. *malaccensis* accessions were found to be admixed: 'THA018' displayed segments assigned to 'burmannica/siamea' on all chromosome sets, and 'PT-BA-00267' had a few large segments assigned to 'banksii/Borneo'.

The ssp. *microcarpa* 'Borneo' accession had a homogeneous ancestry from the 'banksii/Borneo' group. In contrast, the local ancestry mosaic of the ssp. *microcarpa* 'microcarpa' accession revealed its hybrid nature with segments of three different origins: 'zebrina', 'banksii/Borneo' and, to a lesser extent, 'malaccensis'.

The local ancestry mosaic of the 'Agutay' (PT-BA-00008) accession suggested that it was a zebrina/burmannica/malaccensis hybrid, not the expected *M. acuminata* ssp. *errans* accession. This might be due to mislabelling of this accession, which will hereafter be referred to as 'PT-BA-00008'. No segments of *M. balbisiana* origin were predicted in this accession although *M. balbisiana* alleles were detected in the global analysis results (Figure 3b). Alleles of this origin were spread throughout the 'PT-BA-0008' genome either due to the fact that it is a very ancient admixture or due to sequence contamination (Ballenghien *et al.*, 2017). No contribution of the 'balbisiana' group was detected in other accessions.

Local ancestry mosaics in diploid 'AAcv' cultivars

For diploid 'AAcv' cultivars, the local ancestry analysis revealed a mosaic of segments from three to five main ancestral groups (Figures 6 and S3). Based on our approach in which contiguous regions assigned to the same ancestral group were placed next to each other into one 'pseudo-haplotype', large segments of 'banksii/Borneo' ancestry were found at different proportions in all cultivated accessions (including 'Pisang Madu'). In 'Mala', 'SF 215' and 'Guyod' accessions, the 'banksii/Borneo' ancestry was predominant and the second main contributor was 'zebrina', which were found to be present in most or all chromosome sets (Figure 6a,b). In those accessions that were predicted to be 'banksii/zebrina' hybrids based on SSR data (Perrier *et al.*, 2009, 2011), we also detected a few regions from the 'Pisang Madu' and/or 'malaccensis' ancestral groups (Figures 6a,b and S3). For the 'Galeo' accession, two main contributors, that is 'banksii/Borneo' and 'malaccensis', were identified together with a few segments from the 'zebrina' and 'Pisang Madu' ancestral groups.

The 'banksii/Borneo' and 'zebrina' ancestry, predicted using SSRs (Perrier *et al.*, 2009), was confirmed for the 'Chicame' accession representing the East African Mchare subgroup, but a 'malaccensis' ancestry was also detected here in almost all chromosome sets. In the remaining three accessions ('Gu Nin Chiao', 'Kirun' and 'Manang'), chromosome segments from 'banksii/Borneo', 'malaccensis', 'zebrina', and 'Pisang Madu' ancestral groups were found, with the 'malaccensis' contribution being particularly predominant in 'Gu Nin Chiao'. The 'Manang' accession was

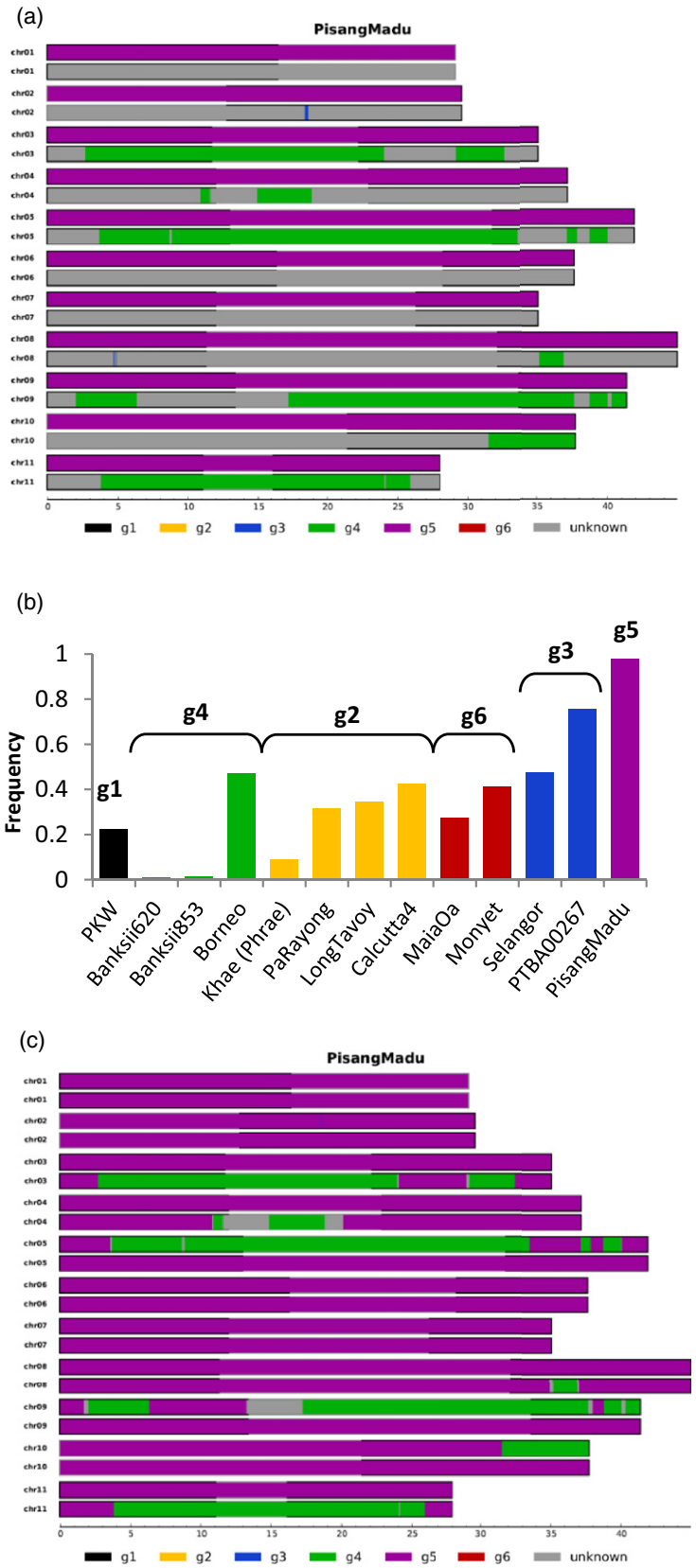

Figure 4. 'Pisang Madu' local ancestry estimation.
(a) Local ancestry estimation for the 'Pisang Madu' accession based on the first assumption that group 5 corresponded to a single ancestor.
(b) Frequency of heterozygous alleles of the main ancestral group present in each representative accession. For each representative accession, the frequency was calculated as the number of sites having exactly one allele of the group over the number of sites with at least one allele of the group. Very high heterozygous frequency for group 5 suggested the contribution of two different ancestries to this group. This led to a new estimation of group 5 ancestry in the studied genotypes.
(c) 'Pisang Madu' local ancestry mosaic based on the assumption that group 5 consisted of two distinct genetic pools.

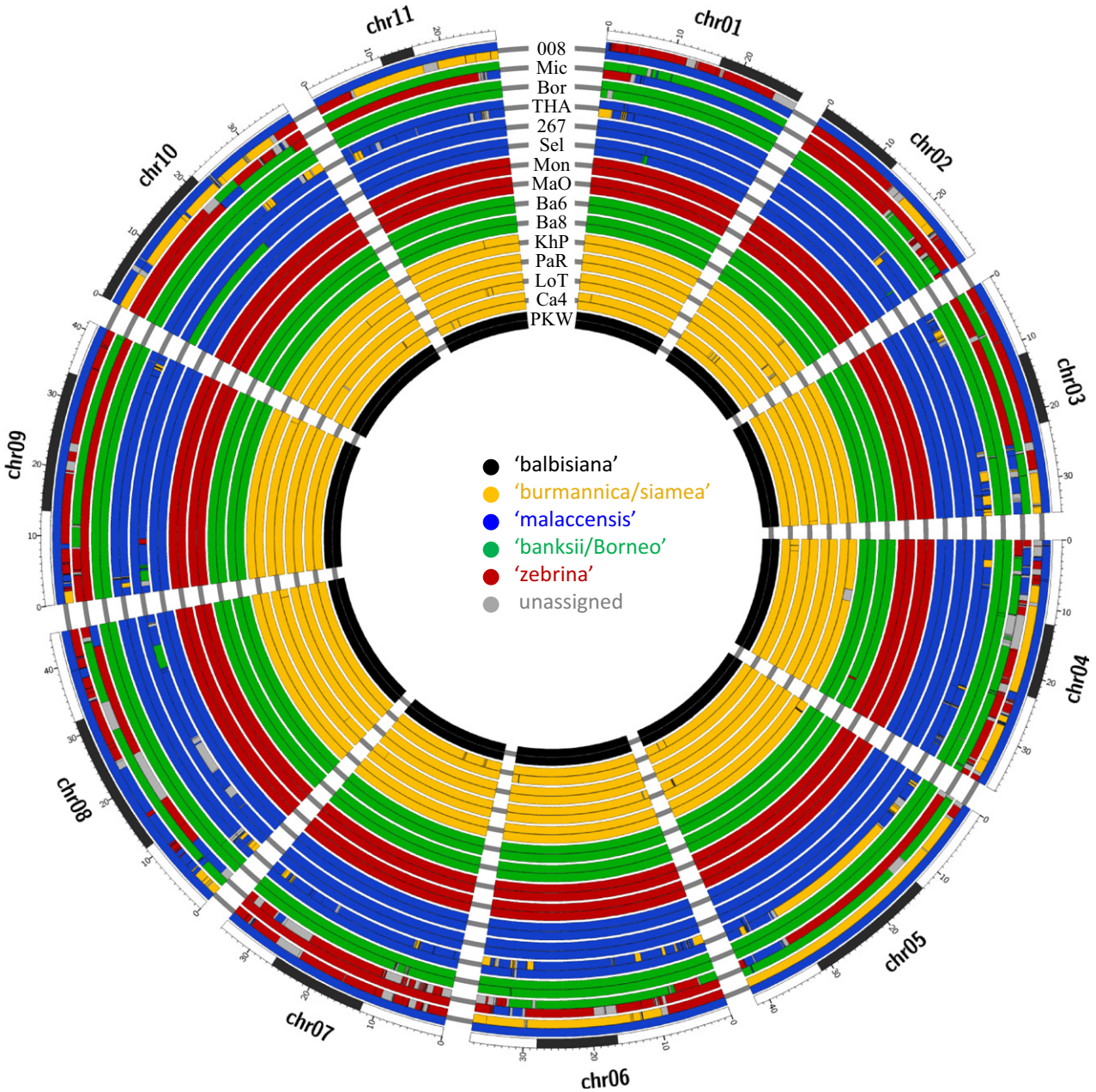

Figure 5. Circular representation of the local ancestry mosaic in wild *M. acuminata* accessions. The outer circle represents the 11 chromosomes of the *M. acuminata* reference genome with dark coloured centromeric regions. Inner circles represent, for each studied accession, the two predicted ancestry pseudo-'haplotypes'. Assigned ancestries are represented by coloured blocks: black, group 1 'balbisiana' (*M. balbisiana*); yellow, group 2 'burmannica/siamea'; blue, group 3 'malaccensis'; green, group 4 'Banksii/Borneo'; purple, group 5 'Pisang Madu' and red, group 6 'zebrina'. Unassigned regions are in grey. Accession names are abbreviated: 008, 'PT-BA-00008'; Mic, 'Microcarpa'; Bor, 'Borneo'; THA, 'THA018'; 267, 'PT-BA-00267'; Sel, 'Selangor'; Mon, 'Monyet'; MaO, 'Maia Oa'; Ba6, 'Banksii ITC0620'; Ba8, 'Banksii ITC0853'; KhP, 'Khae Phrae'; PaR, 'Pa Rayong'; LoT, 'Long Tavoy'; Ca4, 'Calcutta 4'; and PKW, *M. balbisiana*, Pisang Klutuk Wulung'.

the only one that displayed several megabases from the 'burmannica/siamea' ancestral group on different chromosomes and a mosaic of segments from at least five distinct ancestral groups (Figure 6c).

Some regions remained unassigned in all of these accessions (grey in Figures 5–7 and S3). They were often, yet not always, in centromeric regions where marker density was lower.

A 'Mchare' accession and the 'Pisang madu' ancestral group contribute to the triploid 'Grande Naine' accession genome

The local ancestry mosaic of the triploid 'Grande Naine' accession, representing Cavendish cultivars, showed the contributions of four ancestral groups: 'banksii/Borneo', 'zebrina', 'malaccensis' and 'Pisang Madu' (Figure 7). Most

chromosome sets showed contributions from all four of these ancestral groups. The 'Pisang Madu' group contribution was predicted in large segments of several megabases on chromosome sets 1, 6, 7 and 10, and in several discontinuous segments on other chromosomes sets, but it was not detected on chromosome 11 (Figure 7).

To assess the previously proposed contribution of a Mchare clonal group member such as 'Chicame' as 2n donor of 'Grande Naine' and of 'Pisang Madu' as n donor (Raboin *et al.*, 2005; Perrier *et al.*, 2009; Hippolyte *et al.*, 2012), the distributions of polymorphic alleles in these accessions were compared (Figure 7). From 51 602 polymorphic sites among 'Chicame', 'Grande Naine' or 'Pisang Madu', a total of 49 566 (95.9%) sites was consistent with a 2n contribution of 'Chicame' to the 'Grande Naine' genome, whereas 2036 (4.1%) sites were not consistent. These results confirmed that a Mchare accession, highly similar to 'Chicame', contributed a 2n gamete to the 'Grande Naine' genome.

The remaining SNP distribution was consistent with a 'Pisang Madu' origin for chromosomes 1, 2, 7 and 10 (Figure 7). On other chromosomes, regions consistent with a 'Pisang Madu' origin were found together with large regions of several megabases that were not consistent with this origin (light grey bars in Figure 7).

DISCUSSION

We identified five ancestral groups involved in the origin of the analyzed *M. acuminata* accessions based on sequence data from 14 *M. acuminata* wild and nine 'AAcv' cultivated banana diploid accessions, one triploid 'AAA' cultivar and one diploid *M. balbisiana* accession as outgroup. Four of these ancestral groups corresponded to previously characterized *M. acuminata* subspecies, i.e. ssp. *banksii*/ssp. *microcarpa*, ssp. *malaccensis*, ssp. *zebrina* and ssp. *burmannica*/ssp. *siamea*. The fifth one, found only in cultivars, corresponded to one or two new unknown origins. Our results revealed diverse and always more complex than anticipated inter(sub)specific chromosome mosaic patterns, involving four to five ancestral groups for some cultivars, and implying multiple hybridization steps.

Unknown contributor(s) to cultivated banana genomes

Among the five ancestral groups that we identified as being involved in the origin of banana cultivars from our setting, three (ssp. *banksii*, ssp. *malaccensis*, ssp. *zebrina*) had already been proposed as main contributors to the A genome of cultivated banana, and one (ssp. *burmannica*/ssp. *siamea*) was proposed to be involved in very few cultivars (Carreel *et al.*, 2002; Perrier *et al.*, 2009, 2011; Christelová *et al.*, 2017). The fifth ancestral group, present only in cultivars, was defined based on one of these, i.e. 'Pisang Madu'. The approach we used enabled the assessment of the contribution of this cryptic ancestry, which would have been difficult to achieve with more conventional strategies solely based on the use of fixed alleles.

We showed that this 'Pisang Madu' ancestral group was present in several of the 'AAcv' analyzed here. It was also present in the 'Grande Naine' accession, a representative of the Cavendish dessert banana subgroup that accounts for 50% of the bananas produced worldwide. This ancestral group is thus an important contributor to modern-day cultivars and should be characterized to facilitate its use in banana breeding programmes.

The 'Pisang Madu' accession is one of several apparently similar accessions that were surveyed in Sarawak State on the island of Borneo (Malaysia) and in Perak State in continental Malaysia in 1960/1961 (Rosales *et al.*, 1999). Based on available passport data (https://www.crop-diversity.org/mgis/accession/01GLP005109, (Ruas *et al.*, 2017), (Rosales *et al.*, 1999)), the accession studied here corresponded to one of the Sarawak State accessions. It was described as a 'microcarpa derivative' (Rosales *et al.*, 1999), but the status of this subspecies is still unclear (see below).

We found that the 'Pisang Madu' accession was highly heterozygous, suggesting that it is a hybrid between divergent genetic pools rather than a representative of a population with unfixed alleles. Heterozygosity was noted throughout its chromosomes that could not be explained only by the 'banksii/Borneo' introgressed segments detected in its genome. This suggests that at least two other ancestors contributed to the 'Pisang Madu' genome. The unknown ancestral contributors to the 'Pisang Madu' genome could correspond either to other *M. acuminata* subspecies or to other *Musa* species. *M. acuminata* ssp. *errans* was previously proposed as a contributor to cultivated bananas based on cytoplasmic markers of one representative accession (Carreel *et al.*, 2002). The nuclear genome of this accession was reported to be very close to that of ssp. *banksii* (Carreel *et al.*, 1994; Perrier *et al.*, 2009; Christelová *et al.*, 2017), which was not the case for the 'Pisang Madu' ancestral group we identified. This excludes ssp. *errans* as a major contributor to this ancestral group. The two remaining described *M. acuminata* subspecies, ssp. *truncata* and ssp./var. *sumatrana*, have not been predicted to contribute to cultivars and were not included in our study. Their potential contribution is hard to assess as they are poorly represented in accessible germplasm. *M. schizocarpa*, a species that was recently predicted to contribute to the triploid East African banana genome based on internal transcribed spacer analysis (Nemeckova *et al.*, 2018), should be evaluated as a potential contributor to the 'Pisang Madu' ancestral group.

Sardos *et al.* (2016) used DArT markers to analyze a sample of 576 accessions spanning most of the available banana diversity and identified a large group of 'AA'/'AAA' cultivated bananas, including the Cavendish subgroup, that did not

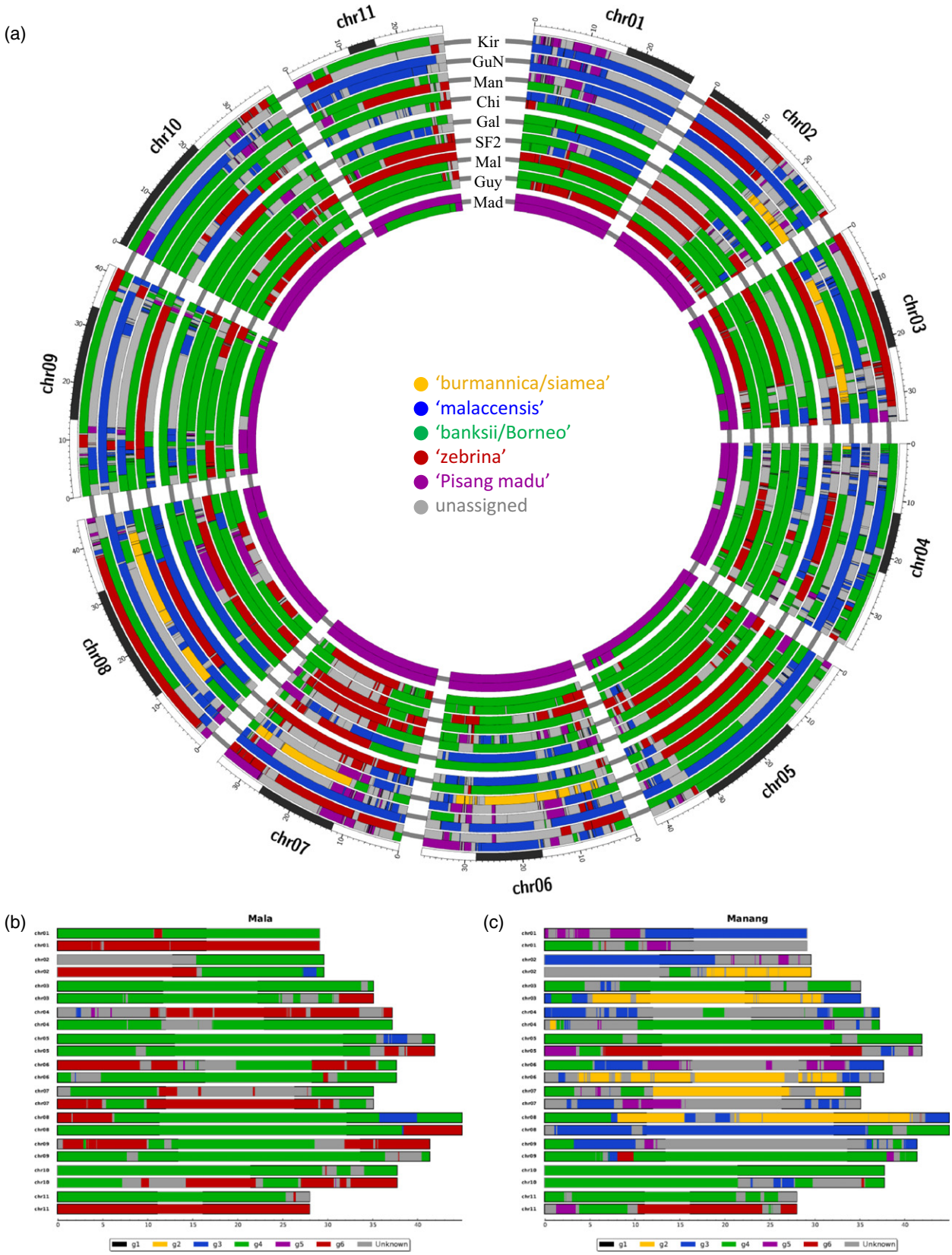

Figure 6. Representation of the local ancestry mosaic in diploid 'AAcv' banana cultivars. Assigned ancestries are represented by coloured blocks: yellow, group 2 'burmannica/siamea'; blue, group 3 'malaccensis'; green, group 4 'Banksii/Borneo'; purple, group 5 'Pisang Madu' and red, group 6 'zebrina'. Unassigned regions are in grey.
(a) Circular representation of the ancestry moaic of diploid 'AAcv' banana cultivars. The outer circle represents the 11 chromosomes of the *M. acuminata* reference genome with dark coloured centromeric regions. Inner circles represent, for each studied accession, the two predicted ancestry pseudo-'haplotypes'. Accession names are abbreviated: Kir, 'Kirun'; GuN, 'Gu Nin Chiao'; Man, 'Manang'; Chi, 'Chicame'; Gal, 'Galeo'; SF2, 'SF.215'; Mal, 'Mala'; Guy, 'Guyod'; Mad, 'Pisang Madu'.
(b) Linear representation of the relatively simple mosaic of the 'Mala' accession.
(c) Linear representation of the mosaic of the 'Manang' accession with at least five ancestral group contributions.

cluster with any of the represented *M. acuminata* subspecies. These results suggested that one or more of the wild contributors to cultivated bananas may not be available in germplasm collections. This may include the cryptic contributor(s) to the 'Pisang Madu' group we identified.

Characteristics of *M. acuminata* subspecies representatives

The ancestry patterns of *M. acuminata* subspecies representatives provided information on the corresponding subspecies. Alleles from ssp. *burmannica*, ssp. *burmannicoides* and ssp. *siamea* accessions were clustered in one ancestral group and accessions from these subspecies had a homogenous local ancestry profile. This confirmed the idea of a single genetic group predicted by SSR and large structural variation analyses (Perrier *et al.*, 2011; Dupouy *et al.*, 2019). The *malaccensis* subspecies is geographically close to ssp. *burmannica*/*siamea*. The existence of hybrids between these subspecies has been mentioned (Perrier *et al.*, 2009). The 'malaccensis' and 'burmannica/siamea' hybrid ancestry mosaic of the 'THA018' accession from Thailand illustrated these genetic contacts. Of the other two ssp. *malaccensis* accessions, the 'PT-BA-000267' accession (named 'Pahang' in the CRB Plantes Tropicales collection) was admixed, which might explain why it was recently found to differ slightly from 'Pahang' accessions of other sources (Noumbissie *et al.*, 2016). Local ancestry analysis could thus identify potential mislabelling cases, as also shown here for the 'PT-BA-00008' accession.

We defined one ancestral group that represented ssp. *banksii* accessions and which also included a significant contribution from the ssp. *microcarpa* 'Borneo' accession. Here it was found that the two representatives of ssp. *microcarpa*, that is 'Borneo' and 'Microcarpa' were different, in line with the findings of previous studies (Carreel *et al.*, 1994; Perrier *et al.*, 2009). The hybrid status of the 'Microcarpa' accession was confirmed, but was found to be more complex, with contributions from three different ancestral groups. No evidence of admixture was found here for the 'Borneo' accession. It was confirmed as being close to ssp. *banksii* (Carreel *et al.*, 1994) with, however, a higher level of heterozygosity than ssp. *banksii* accessions. This raised questions about the status of ssp. *microcarpa*, as already noted (Perrier *et al.*, 2009).

Complex mosaic genomes of 'AAcv' cultivated bananas

The mosaic genome of the studied banana cultivars revealed contributions of at least three and up to five ancestral groups. It highlighted the diversity of genomic constitutions in cultivars originating from different regions. Previous diversity studies on banana accessions had identified different genetic clusters for 'AAcv' (Perrier *et al.*, 2009, 2011; Sardos *et al.*, 2016; Christelová *et al.*, 2017). They included accessions from New Guinea with a dominant 'banksii' contribution, a range of clusters with decreasing 'banksii' and increasing 'zebrina/microcarpa' contributions, and accessions from Southeast Asia that were predicted to be mainly derived from 'malaccensis' and 'zebrina/microcarpa' (Perrier *et al.*, 2009, 2011). We refined these predictions via our approach. We found a predominant 'banksii/zebrina' component in two accessions from Papua New Guinea and accession 'Guyod' from the southern Philippines, thus confirming the general predictions for these accessions. We also observed a predominance of 'malaccensis' ancestry in the Southeast Asian 'Gu Nin Chiao' accession (Singapore) that was consistent with the distribution range of ssp. *malaccensis*. However, all of the analyzed cultivars also carried ancestries other than those previously predicted. Regions of 'malaccensis' or cryptic 'Pisang Madu' origin were found in predicted 'banksii/zebrina' hybrids such as the 'Mala', 'SF215' and 'Guyod' accessions. In the 'Chicame' accession, of the diploid East African banana Mchare subgroup, large 'malaccensis' regions were revealed in addition to the 'banksii/zebrina' ancestry. This suggests a more complex hybridization scheme at their origin than previously hypothesized. Similarly, the genome mosaic of the 'Kirun' representative of the widely cultivated diploid 'Sucrier' subgroup, and the genome mosaics of 'Galeo' and 'Gu Nin Chiao', were found to have resulted from at least four ancestries, while 'Manang' showed at least five different ancestries, implying multiple hybridization events. The mosaic of 'Pisang Madu' from the island of Borneo represented a particular case in which two unknown ancestries seemed to be the main contributors, in addition to admixture with a characterized ('banksii/Borneo') ancestry. Although limited in size, our setting illustrated a high level of diversity and complexity of banana genome composition.

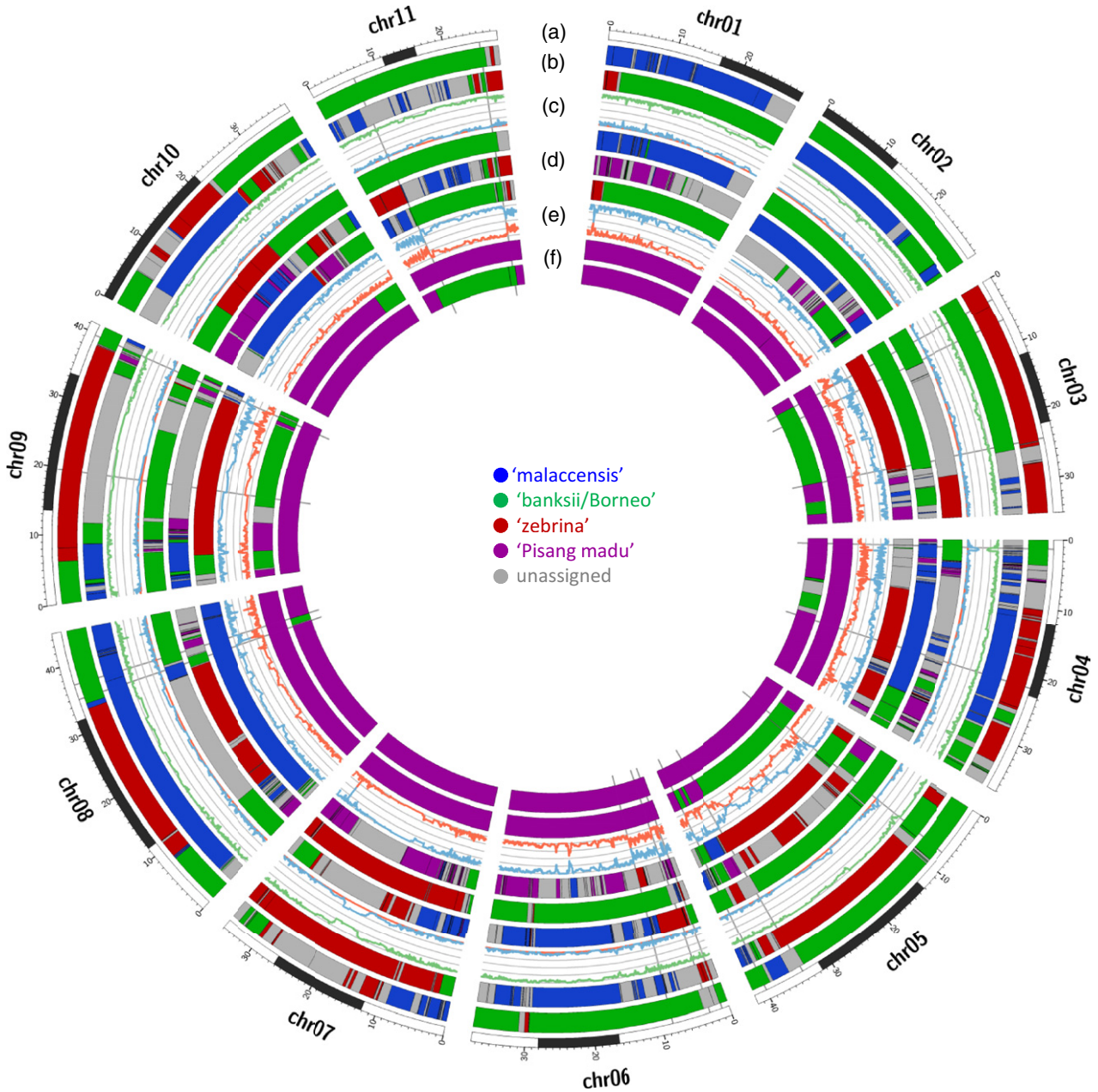

Figure 7. Circular representation of the local ancestry mosaic of the triploid 'Grande Naine' banana cultivar and comparison with candidate 2n and n gamete donors.
From the outer circle to the inner circle: (a) *M. acuminata* reference chromosomes with dark coloured centromeric regions. (b) Predicted mosaic structure of the diploid 'Chicame' accession. (c) Shared allele proportions along chromosomes, between 'Grande Naine' and 'Chicame' accessions. Green, blue and red curves represent the proportion of variant sites in which both one and no 'Chicame' alleles, respectively, were found in 'Grande Naine'. (d) Predicted mosaic structure of the triploid 'Grande Naine' accession. (e) Shared allele proportions between the remaining haplotype of 'Grande Naine' and 'Pisang Madu' accession. Blue and red curves represent the proportion of variant sites in which one and no 'Pisang Madu' alleles, respectively, were found in 'Grande Naine'. (f) Predicted mosaic structure of the 'Pisang Madu' accession. Assigned ancestries are represented by coloured blocks: blue, group 3 'malaccensis'; green, group 4 'Banksii/Borneo'; purple, group 5 'Pisang Madu' and red, group 6 'zebrina'. Unassigned regions are in grey.

Several accessions included chromosome segments of various sizes that were not assigned to any ancestry. These segments sometimes corresponded to centromeric regions of reduced marker density. Moreover, unassigned regions along the chromosomes may have resulted from: (i) incomplete representation of the diversity of known *M. acuminata* groups; (ii) the presence of admixed representatives in our setting, which might have affected the allele clustering

efficiency; (iii) underestimation of the complex 'Pisang Madu' contribution; or (iv) the existence of other ancestries.

The observed genome mosaic of intraspecific cultivars suggests that intermediate banana hybrids may have retained fertility for several generations, leading to a combination of up to five different ancestral genetic pools. These results supported the idea that the banana domestication process has involved more hybridization steps than initially thought, as already proposed for A/B interspecific hybrids (De Langhe *et al.*, 2010; Baurens *et al.*, 2019).

Origin of Cavendish bananas

The genome mosaic of the Cavendish triploid 'Grande Naine' accession showed a contribution of four identified ancestral groups, i.e. 'malaccensis', 'zebrina', 'banksii/Borneo' and 'Pisang Madu'. The significant contribution of the 'Pisang Madu' group suggests that one or both genetic components of this group are major contributors to this commercially important banana cultivar.

The Mchare origin of the 2n gamete donor of Cavendish hypothesized in previous studies (Raboin *et al.*, 2005; Hippolyte *et al.*, 2012) was supported by local ancestry patterns detected and polymorphic allele comparisons of 'Chicame' and 'Grande Naine'. The low proportion (4.1%) of SNPs that were in disagreement with this 2n origin could be explained by possible somaclonal mutation accumulation in representatives of the Mchare and Cavendish subgroups or errors in allele dosage estimations during genotype calling of the triploid. The Mchare 2n gamete donor predicted to contribute 'zebrina' and 'banksii' to the Cavendish genome was also shown here to have contributed a large 'malaccensis' component. A translocation involving chromosomes 1 and 4 present in ssp. *malaccensis* accessions and thought to have originated in ssp. *malaccensis* was detected in Mchare accessions and in Cavendish (Martin *et al.*, 2017). 'Chicame' and 'Grande Naine' bear large 'malaccensis' regions on both chromosomes 1 and 4; this is in line with the presence of these structural variations.

Based on SSR markers, the n donor haplotype of Cavendish was suggested to correspond to a haplotype shared between 'Pisang Madu' and 'Pisang Pipit' accessions (Perrier *et al.*, 2009). Our analysis revealed, in Cavendish, large chromosomal segments that were concordant with 'Pisang Madu' as donor of the n gamete. However, other large regions do not seem to derive from Pisang Madu. These results suggested that a close relative of 'Pisang Madu' contributed the n gamete of Cavendish.

As banana cultivars are highly sterile, breeding cannot efficiently build on several recurrent steps involving cultivars. Conventional banana breeding classically consisted in the reconstruction of a triploid in one hybridization step, whereas it has now evolved towards improvement of diploid parents before reconstruction of a triploid product. Information on the genome ancestry mosaic of current cultivars is essential to guide the choice of parents in these breeding programmes. In addition, our results should help focus future germplasm research on characterizing the unknown contributor(s) we revealed. Finally, our findings suggested that the domestication process of banana cultivars involved more hybridization steps and more ancestral contributors than initially thought.

EXPERIMENTAL PROCEDURES

Transcriptome dataset

A set of 23 diploid *M. acuminata* accessions comprising 14 seedy and nine parthenocarpic accessions (Table 1) was selected to represent the wild *M. acuminata* contributors to cultivated bananas and different groups of diploid cultivars ('AAcv'). Plant materials were collected from field grown plants maintained in the CIRAD collection hosted by CRB Plantes Tropicales Antilles CIRAD-INRA in Guadeloupe (France), except for five accessions that were from the Collection Musacées of CARBAP in Cameroon. Leaf, flower and fruit tissues from each accession were harvested and separately stored in RNA*later*® solution (QIAGEN, https://www.qiagen.com). Total RNA was extracted as described in Jourda *et al.* (2014) with lithium chloride 2 M precipitation for fruit samples. For each accession, one pool of 65% flower RNA, 20% fruit RNA and 15% leaf RNA was used for library construction, except for the 'Banksii ITC0620' accession (for which only flowers and fruit were used) and 'Khae (Phrae)' and 'Pa Rayong' accessions [for which flowers (80%) and leaves (20%) were used]. RNA sequencing was performed using the Illumina Tru-SEQ RNA Kit (Illumina Inc, https://www.illumina.com/) and the Illumina mRNA-seq paired-ends protocol on a HiSeq2000 sequencer for 2 × 100 cycles. Sequence data for the 'Mala' accession were pooled from two sequencing experiments. Data from 10 of the wild accessions were part of the study of Clement *et al.* (2017).

Genomic dataset

Raw reads from the draft B genome sequence of *M. balbisiana* accession 'Pisang Klutuk Wulung' (Davey *et al.*, 2013) were retrieved from the NCBI Sequence Read Archive (ID: SRR956987).

Leaf material of the triploid 'Grande Naine' accession (Cavendish subgroup, 'AAA' genome) was collected from field grown plants hosted in the CRB Plantes Tropicales Antilles CIRAD-INRA collection in Guadeloupe. Total DNA was isolated using a modified MATAB method (Risterucci *et al.*, 2000) and was used to construct a 5-kb insert mate-pair library that was sequenced with the Illumina HiSeq platform at GENOSCOPE (http://www.genoscope.cns.fr) to obtain 2 × 101 bp paired-end reads.

Read filtering

RNA-seq read filtering. For each library, adaptor sequences were removed from raw paired-end reads using the Cutadapt program (Martin, 2011). Cutadapt was also used to remove the first seven bases of each RNA-seq read in which nucleotide-specific mismatch errors may occur (van Gurp *et al.*, 2013). Reads were filtered based on a minimum length of 35 and a mean base quality of 30 using *arcad_hts_2_Filter_Fastq_On_Mean_Quality.pl* and *arcad_hts_3_compare_fastq_paired_v5.pl* programs available on the South Green Bioinformatics platform (https://github.com/SouthGreenPlatform/arcad-hts).

DNA-seq read filtering. For each library, reads were trimmed from both ends until a base quality ≥20 was reached. Reads were then truncated on the second N found in the sequence and only reads of a minimum of 30 bases were kept.

Read mapping, processing, and variant calling

RNA-seq. Filtered RNA-seq reads were aligned to *Musa* reference genomes comprising the *M. acuminata* nuclear reference sequence (D'Hont *et al.*, 2012) version 2 (Martin *et al.*, 2016), 12 mitochondrial scaffolds and the chloroplast genome sequence (Martin *et al.*, 2013) using the STAR v2.5 aligner software (Dobin *et al.*, 2013) with no more than 10 mismatches per paired-end read, an intron size of between 20 bases and 50 kb and no multiple alignment. The mapping process was performed in three steps: (i) first mapping of all paired reads from all libraries; (ii) identification of all splicing sites; and (iii) second mapping of each library independently using splicing site information. In this last mapping step, for each accession, paired-end reads and single-end reads were aligned independently. Resulting 'sam' files were merged using PICARD Tools v2.7.0 (http://broadinstitute.github.io/picard).

For each accession, redundant reads were removed using MARKDUPLICATE from PICARD Tools v2.7.0 and reads were split into exon segments using the SPLITNCIGARREADS tool of GATK v3.3 (McKenna *et al.*, 2010). Split reads were locally realigned around indels using GATK INDELREALIGNER.

DNA-seq. Filtered DNA-seq reads were aligned to *Musa* reference genomes comprising the *M. acuminata* nuclear reference sequence version 2 (Martin *et al.*, 2016), 12 mitochondrial scaffolds and the chloroplast genome sequence (Martin *et al.*, 2013) using BWA v0.7.15 with the *mem* algorithm (Li, 2013). Reads aligning at several positions were removed using SAMTOOLS v1.3 (Li *et al.*, 2009).

For each accession, redundant reads were removed using MarkDuplicate from PICARD Tools v2.7.0. Reads were locally realigned around indels using the GATK v3.3 INDELREALIGNER.

Variant calling. For each accession, at each covered position, all mapped bases with a mapping quality ≥ 10 were counted with the bam-readcount program (https://github.com/genome/bam-readcount). Variant sites were filtered according to the following criteria: (i) only data points covered by at least 10 reads were considered; (ii) only variant alleles supported by at least three reads and having a frequency ≥0.05 in at least one accession were kept; and (iii) sites showing at least one variant were kept for variant calling. For each accession and at each variant site, a genotype was called based on the maximum likelihood of all possible genotypes, calculated based on a binomial distribution assuming a sequencing error rate of 0.005. The resulting variant calling file was formatted to vcf format. The complete process was performed using the custom python scripts *process_RNAseq.1.0.py process_reseq_1.0.py* and *VcfPreFilter.1.0* available at https://github.com/SouthGreenPlatform/vcfHunter.

SNP filtering

The vcf file was filtered according to the following parameters: (i) all genotype calls having less than 10-fold total coverage or less than three-fold coverage for each allele were converted to missing data and variant sites having a missing genotype call were filtered out; (ii) SNP clusters of at least three SNPs in a window of 10 bases or two contiguous SNPs, identified using GATK v3.3, were removed; and (iii) only di-allelic variant sites were kept.

Heterozygous sites and accession-specific alleles

The percentage of heterozygous sites for each accession (i.e. percentage of heterozygous sites/total number of sites in the vcf file) and the percentage of accession-specific alleles (i.e. percentage of sites having at least one allele present only in the accession) were calculated on the final vcf file using the *vcf2struct.1.0.py* custom script available at https://github.com/SouthGreenPlatform/vcfHunter.

Global ancestry estimation

To detect genetic clusters, ADMIXTURE software v1.23 (Alexander *et al.*, 2009) was run unsupervised on the set of 24 diploid accessions. Seven values of K were tested (K = 2, 3, 4, 5, 6, 7 and 8) with cv = 10 and four replicates corresponding to different seeds ($seed$ = 10, 100, 500, and 1000). The optimal cluster number was determined by the lowest cross-validation error value. The ancestry fraction (Q) in each individual was plotted for the run with the lowest cross-validation error value.

Identification of ancestry informative alleles

On the basis of the ADMIXTURE analysis, all accessions predicted as homogeneous for a cluster for the best K value were selected to attribute alleles to ancestral groups. Polymorphic SNPs for these accessions were re-coded as follows: for each accession and each allele at a variant site, two lines were generated, a first one in which the allele presence state was coded as 0 or 1 and a second line in which the allele absence state was coded as 0 or 1 (Figure S4). Alleles present or absent in all accessions were discarded. A COA was then performed on the transposed matrix with R software v3.2.2 using the ADE4 package (Dray and Dufour, 2007).

Allele coordinates on the synthetic axes were clustered using the mean shift algorithm implemented in SCIKIT-LEARN (Pedregosa *et al.*, 2011) with a bandwidth of 1.16. Allele grouping along synthetic axes was plotted using MATPLOTLIB (Hunter, 2007). After identification of allele groups that were informative on ancestry, the relative contributions of each group to each of the 25 accessions was checked by calculating grouped allele proportions (i.e. number of grouped alleles from a group/total number of grouped alleles in one accession).

The complete process was performed using custom python scripts *vcf2struct.1.0.py* available at https://github.com/SouthGreenPlatform/vcfHunter.

Local ancestry analysis

Accessions representative of each ancestral group were used to infer, for each position (i) corresponding to an informative grouped allele, the probability $P_{i\alpha}$ of observing an allele from a given ancestral group α if this position is representative of the ancestral group α, and the probability $P_{i\bar{\alpha}}$ of observing an allele from a given ancestral group α if this position is not representative of the ancestral group α (corresponding to noise and/or wrong allele assignment).

The probability $P_{i\alpha}$ is estimated by the observed frequency of the group α allele in accessions that are representative of this group. When there is only one admixed representative of an ancestral group, $P_{i\alpha}$ = 1 if at least one allele of the group is found in the admixed individual, $P_{i\alpha}$ = 0 in other cases.

The probability $P_{i\bar{\alpha}}$ was estimated by the observed frequency of the group α allele in accessions representative of other groups.

The $P_{i\alpha}$ probabilities were then used to estimate the expected number of α specific alleles, $m_{rk\alpha}$, on sliding windows of size n starting at position r [r,$r + n$]. This value $m_{rk\alpha}$ also depends on the number k of haplotypes of an origin α, with $k = 1$ or 2 in a diploid, 1, 2 or 3 in a triploid. The $m_{rk\alpha}$ value is calculated as follows:

$$m_{rk\alpha} = \sum_{i=r}^{r+n} kP_{i\alpha}$$

And an associated variation parameter ($\Delta_{rk\alpha}$) around this value is calculated as follows:

$$\Delta_{rk\alpha} = \sqrt{\sum_{i=r}^{r+n} kP_{i\alpha} * (1 - p_{i\alpha})}$$

The variation parameter in ancestry attribution is estimated based on accessions identified as representatives of the ancestral group. This parameter might be underestimated for some ancestral groups as the sampling was limited here. Therefore, we used a maximal estimation of the variation parameter Δ_{rkg} defined as the maximal value of all $\Delta_{rk\alpha}$ in the given window.

The $P_{i\bar{\alpha}}$ probabilities were used to estimate the expected number of noise specific alleles ($m_{r\bar{\alpha}}$) on sliding windows of size n starting at position r [r, $r + n$]. This maximal value ($m_{r\bar{\alpha}}$) depends on the ploidy β of the accession. This noise expectation ($m_{r\bar{\alpha}}$) is calculated as follows:

$$m_{r\bar{\alpha}} = \sum_{i=r}^{r+N} \beta P_{i\bar{\alpha}}$$

The associated variation parameter ($\Delta_{r\bar{\alpha}}$) around this noise expectation value is calculated as follows:

$$\Delta_{r\bar{\alpha}} = \sqrt{\sum_{i=r}^{r+N} \beta P_{i\bar{\alpha}} * (1 - P_{i\bar{\alpha}})}$$

A global noise variation value ($\Delta_{r\bar{g}}$) was then calculated as the maximal value of all $\Delta_{r\bar{\alpha}}$ in the given window.

All of these values were used to calculate the probability of having at least k haplotypes based on the observed number of alleles from group α in a given window [r, $r + n$] for an accession ($x_{r\alpha}$). This probability ($q_{rk\alpha}$) was calculated as follows (Figure S5a):

$$\text{if}: \quad x_{r\alpha} \geq m_{rk\alpha} - \Delta_{rkg} \text{ then } q_{rk\alpha} = 1.$$

$$\text{else: } P_{rk\alpha} = \frac{\frac{1}{\Delta_{rkg}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_{r\alpha} - \left(m_{rk\alpha} - \Delta_{rkg}\right)}{\Delta_{rkg}}\right)^2}}{\frac{1}{\Delta_{rkg}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\left(m_{rk\alpha} - \Delta_{rkg}\right) - \left(m_{rk\alpha} - \Delta_{rkg}\right)}{\Delta_{rkg}}\right)^2}}$$

which corresponded to the probability density of a normal distribution with a mean value $\mu_{rk\alpha} = m_{rk\alpha} - \Delta_{rk\alpha}$ and $\sigma_{rk\alpha} = \Delta_{rk\alpha}$ normalized so that the highest probability is equal to one. Similarly, the probability ($q_{r\bar{\alpha}}$) of having the observed number of alleles ($x_{r\alpha}$) from group α in a given window [r, $r + n$] in case of no haplotype from this origin was calculated as follows (Figure S5b):

$$\text{if}: \quad x_{r\alpha} \leq m_{r\bar{\alpha}} + \Delta_{r\bar{g}} \text{ then } q_{r\bar{\alpha}} = 1,$$

$$\text{else: } q_{r\bar{\alpha}} = \frac{\frac{1}{\Delta_{r\bar{g}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_{r\alpha} - \left(m_{r\bar{\alpha}} + \Delta_{r\bar{g}}\right)}{\Delta_{r\bar{g}}}\right)^2}}{\frac{1}{\Delta_{r\bar{g}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\left(m_{r\bar{\alpha}} + \Delta_{r\bar{g}}\right) - \left(m_{r\bar{\alpha}} + \Delta_{r\bar{g}}\right)}{\Delta_{r\bar{g}}}\right)^2}}$$

which corresponded to the probability density of a normal distribution with a mean value $\mu_{r\bar{\alpha}} = m_{r\bar{\alpha}} + \Delta_{r\bar{g}}$ and $\sigma_{r\bar{g}} = \Delta_{r\bar{g}}$ normalized, so that the highest probability is equal to 1.

These probabilities were calculated for each accession and along each chromosome on sliding windows of n clustered SNPs sites (Figure S6). The overlap between two windows was $n - 1$. For our analysis, the window was 301 clustered SNPs sites. The probabilities calculated for each window were attached to the central SNP position of the window. Ancestries were assigned to the position based on a major probability rule. Only probabilities >0.1 were considered. For each accession, as the SNPs were not phased, a minimum number of recombination was assumed and therefore contiguous regions of the same ancestry were attributed to one of the homologous chromosomes forming pseudo-haplotypes.

The complete process was performed using custom python scripts *vcf2linear.1.1* available at https://github.com/SouthGreenPlatform/vcfHunter.

'Pisang Madu' local ancestry estimation

As a first round of local ancestry analysis did not allow the full characterization of the 'Pisang madu' group contribution and, given the high level of heterozygosity of group 5 alleles, the local ancestry analysis was performed again on all accessions with a new estimation of $P_{i\alpha}$ for group 5 as follows: for all regions of 'Pisang Madu' displaying a group 5 haplotype and a second haplotype with no assigned ancestry in the first analysis, the probability of emitting a group 5 allele was $P_{i\alpha} = 0.5$ in heterozygous states or 1 in (the rare) homozygous states. The other cases considered were regions with a group 5 haplotype and a haplotype from another ancestral group (i.e. group 4) or those with no group 5 haplotype identified in 'Pisang Madu'. In these regions, $P_{i\alpha} = 1$ if one allele of group 5 was found in the admixed individual, $P_{i\alpha} = 0$ if no group 5 allele was found.

Ancestry analysis of the triploid 'Grande Naine' cultivar

Local ancestry analysis of the triploid 'Grande Naine' accession was performed as described above with the ploidy level parameter set at 3. Polymorphic SNP sites among 'Pisang Madu', 'Chicame' and 'Grande Naine' were selected to verify the predicted parentage of 'Grande Naine'. Then, at each site, the proportion of Chicame alleles also present in Cavendish (i.e. shared alleles) was calculated along the 11 chromosomes, as in the following example: at a given site if the called genotype in 'Grande Naine' is CCT and the 'Chicame' genotype is CC, the proportion of shared alleles is 1 (for first allele C of 'Chicame') + 1 (for second allele C of 'Chicame')/2 (lower ploidy level) = 1; if the 'Grande Naine' genotype is CCT and the 'Chicame' genotype is TT, the proportion of shared alleles is (1 + 0)/2 = 0.5; similarly, the proportion of shared alleles is 0 if 'Chicame' is TT and 'Grande Naine' is CCC.

In a second step, at sites in which the proportion of 'Chicame' alleles shared with 'Grande Naine' was equal to 1 (95.9% of sites), 'Chicame' alleles were subtracted from the 'Grande Naine' variant calling (i.e. if at a site 'Grande Naine' was CCT and 'Chicame' was CT, the remaining allele is C). This operation, under the hypothesis that 'Chicame' is very close to the 2n gamete donor of 'Grande Naine', allowed access to the n gamete donor alleles distributed along 'Grande Naine' chromosomes. This n gamete allele distribution of 'Grande Naine' was then compared with 'Pisang Madu' according to the same procedure as for 'Grande Naine' and 'Chicame' (i.e. if 'Pisang Madu' is CT and 'Grande Naine' remaining allele is C, the proportion of shared alleles is 1; if 'Pisang Madu' is CC and the remaining 'Grande Naine' allele is T, the proportion of shared alleles is 0).

The complete process was performed using custom python scripts *vcf2struct.1.0.py* (for vcf filtering), *vcfIdent.1.0.py* (for comparison) and *vcfRemove.1.0.py* (for 'Grande Naine' n gamete donor extraction) available at https://github.com/SouthGreenPlatform/vcfHunter.

ACKNOWLEDGEMENTS

This study was supported by funding from the Agropolis Fondation 'ARCAD' project No. 0900-001 and the Agropolis Fondation 'GenomeHarvest' project (ID 1504-006) through the French 'Investissements d'avenir' programme (Labex Agro: ANR-10-LABX-0001-01) and by the CGIAR Research Programme on Roots, Tubers, and Bananas (RTB). The authors thank Laure Sauné, Sylvain Santoni, Jacques David and Sylvain Glémin for RNA-seq libraries and sequencing within the framework of the ARCAD project and Karine Labadie (Genoscope) for DNA sequencing. We also thank CRB Plantes Tropicales Antilles CIRAD-INRA Guadeloupe France and the Collection Musacées CARBAP Cameroon for providing plant materials. This work was also supported by the CIRAD – UMR AGAP HPC Data Center of the South Green Bioinformatics platform (http://www.southgreen.fr/).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

GM developed the bioinformatics tools and performed data analyses, CC performed RNA extractions, GS contributed to data management, JCG contributed to the methodology design, XP contributed to the project design, CJ, SR, EF contributed to plant material collection, AD edited the manuscript, GM and NY designed and developed the project, analyzed the results and wrote the paper.

DATA AVAILABILITY STATEMENT

The RNA-seq reads are available under the NCBI Bioproject PRJNA326055. DNA-seq reads for accession 'Grande Naine' are available under accession number PRJEB32153. The vcf file and raw local ancestry assignment statistics plots are available at the Banana Genome Hub (Droc *et al.*, 2013) under https://banana-genome-hub.southgreen.fr/download

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Single nucleotide polymorphism marker density plot along the 11 chromosomes of the *Musa acuminata* reference genome.

Figure S2. First local ancestry estimation in the 25 studied banana accessions.

Figure S3. Final local ancestry estimation in the 25 studied banana accessions.

Figure S4. Allele coding for correspondence analysis.

Figure S5. Schematic representation of the calculation of probabilities for ancestry assignment.

Figure S6. Example of detailed outputs from the local ancestry estimation.

Table S1. Sequence statistics.

Table S2. Statistics on the vcf used for analysis.

Table S3. ADMIXTURE program analysis: cross-validation error with varying seed and K parameters.

Table S4. Multivariate analysis axis inertia.

Table S5. Number of alleles assigned to each group with the mean shift clustering algorithm.

Table S6. Heterozygous clustered allele ratio in accessions representative of ancestral groups.

REFERENCES

Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664.

Ballenghien, M., Faivre, N. and Galtier, N. (2017) Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biol.* **15**, 25.

Baurens, F.C., Martin, G., Hervouet, C. *et al.* (2019) Recombination and large structural variations shape interspecific edible bananas genomes. *Mol. Biol. Evol.* **36**, 97–111.

Boonruangrod, R., Desai, D., Fluch, S., Berenyi, M. and 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. *Theor. Appl. Genet.* **118**, 43–55.

Bredeson, J.V., Lyons, J.B., Prochnik, S.E. *et al.* (2016) Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* **34**, 562–570.

Carreel, F. (1994) Etude de la diversité génétique des bananiers (genre Musa) à l'aide de marqueurs RFLP: Institut national Agronomique Paris-Grignon, pp. 90.

Carreel, F., Fauré, S., Gonzalez de Leon, D., Lagoda, P., Perrier, X., Bakry, F., Tezenas du Montcel, H., Lanaud, C. and Horry, J.P. (1994) Evaluation de la diversite génétique chez les bananiers diploïdes (*Musa* sp). *Genet. Sel. Evol.* **26**, 125s–136s.

Carreel, F., Gonzalez de Leon, D., Lagoda, P., Lanaud, C., Jenny, C., Horry, J.P. and Tezenas du Montcel, H. (2002) Ascertaining maternal and paternal lineage within Musa by chloroplast and mitochondrial DNA RFLP analyses. *Genome*, **45**, 679–692.

Christelová, P., De Langhe, E., Hřibová, E. *et al.* (2017) Molecular and cytological characterization of the global Musa germplasm collection provides insights into the treasure of banana diversity. *Biodivers. Conserv.* **26**(4), 801–824.

Clement, Y., Sarah, G., Holtz, Y. *et al.* (2017) Evolutionary forces affecting synonymous variations in plant genomes. *PLoS Genet.* **13**, e1006799.

Davey, M.W., Gudimella, R., Harikrishna, J.A., Sin, L.W., Khalid, N. and Keulemans, J. (2013) A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids. *BMC Genom.* **14**, 683.

De Langhe, E., Vrydaghs, L., De Maret, P., Perrier, X. and Denham, T. (2009) Why bananas matter: an introduction to the history of banana domestication. *Ethnobot. Res. Appl.* **7**, 165–177.

De Langhe, E., Hribova, E., Carpentier, S., Dolezel, J. and Swennen, R. (2010) Did backcrossing contribute to the origin of hybrid edible bananas? *Ann. Bot.* **106**, 849–857.

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

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Dodds, K.S., Simmonds, N.W. and Cheesman, E.E. (1948) Genetical and cytological studies of *Musa*. IX. The origin of an edible diploid and the significance of interspecific hybridization in the banana complex. *J. Genet.* **48**, 285–293.

Dray, S. and Dufour, A.-B. (2007) The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20.

Droc, G., Larivière, D., Guignon, V. *et al.* (2013) The banana genome hub. *Database*, **2013**, bat035.

Dupouy, M., Baurens, F.C., Derouault, P. *et al.* (2019) Two large reciprocal translocations characterized in the disease resistance-rich *burmannica* genetic group of *Musa acuminata*. *Ann. Bot.* **124**, 319–329.

Flowers, J.M., Hazzouri, K.M., Gros-Balthazard, M. *et al.* (2019) Cross-species hybridization and the origin of North African date palms. *Proc. Natl Acad. Sci. USA*, **116**, 1651–1658.

van Gurp, T.P., McIntyre, L.M. and Verhoeven, K.J.F. (2013) Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS ONE*, **8**, e85583.

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

Hunter, J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95.

Jourda, C., Cardi, C., Mbéguié-A-Mbéguié, D., Bocs, S., Garsmeur, O., D'Hont, A. and Yahiaoui, N. (2014) Expansion of banana (*Musa acuminata*) gene families involved in ethylene biosynthesis and signalling after lineage-specific whole-genome duplications. *New Phytol.* **202**, 986–1000.

Kennedy, J. (2009) Bananas and people in the homeland of genus *Musa*: Not just pretty fruit. *Ethnobot. Res. Appl.* **7**, 19.

Lescot, T. (2018) Banana. Genetic diversity. *Fruitrop*, 92–96.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.*arXiv:1303.3997*, *[q-bio]*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G.P.D.P. (2009) The Sequence Alignment/Map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* **17**, 10.

Martin, G., Baurens, F.C., Cardi, C., Aury, J.M. and D'Hont, A. (2013) The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledon evolution. *PLoS ONE*, **8**, e67350.

Martin, G., Baurens, F.C., Droc, G. *et al.* (2016) Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genom.* **17**, 243.

Martin, G., Carreel, F., Coriton, O. *et al.* (2017) Evolution of the banana genome (*Musa acuminata*) is impacted by large chromosomal translocations. *Mol. Biol. Evol.* **34**, 2140–2152.

McFadden, E.S. and Sears, E.R. (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J. Hered.* **37**, 81–89, 107–116.

McKenna, A., Hanna, M., Banks, E. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Nemeckova, A., Christelova, P., Cizkova, J., Nyine, M., Van den Houwe, I., Svacina, R., Uwimana, B., Swennen, R., Dolezel, J. and Hribova, E. (2018) Molecular and cytogenetic study of East African Highland banana. *Front. Plant Sci.* **9**, 1371.

Noumbissie, G.B., Chabannes, M., Bakry, F. *et al.* (2016) Chromosome segregation in an allotetraploid banana hybrid (AAAB) suggests a translocation between the A and B genomes and results in eBSV-free offsprings. *Mol. Breeding*, **36**, 38.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) Scikit-learn: machine learning in Python. *J. Mac. Learn. Res.* **12**, 2825–2830.

Perrier, X., Bakry, F., Carreel, F., Jenny, C., Horry, J.P., Lebot, V. and Hippolyte, I. (2009) Combining biological approaches to shed light on the evolution of edible bananas. *Ethnobot. Res. Appl.* **7**, 199–216.

Perrier, X., De Langhe, E., Donohue, M. *et al.* (2011) Multidisciplinary perspectives on banana (*Musa* spp.) domestication. *Proc. Natl Acad. Sci. USA*, **108**, 11311–11318.

Perrier, X., Jenny, C., Bakry, F., Karamura, D., Kitavi, M., Dubois, C., Hervouet, C., Philippson, G. and De Langhe, E. (2019) East African diploid and triploid bananas: a genetic complex transported from South-East Asia. *Ann. Bot.* **123**, 19–36.

Raboin, L.M., Carreel, F., Noyer, J.L., Baurens, F.C., Horry, J.P., Bakry, F., Montcel, H.T.D., Ganry, J., Lanaud, C. and Lagoda, P.J.L. (2005) Diploid ancestors of triploid export banana cultivars: molecular identification of 2n restitution gamete donors and n gamete donors. *Mol. Breeding*, **16**, 333–341.

Risterucci, A.M., Grivet, L., N'Goran, J.A.K., Pieretti, I., Flament, M.H. and Lanaud, C. (2000) A high-density linkage map of *Theobroma cacao* L. *Theor. Appl. Genet.* **101**, 948–955.

Rosales, F., Arnaud, E. and Coto, J.E. (1999) *A Catalogue of Wild and Cultivated Bananas. A Tribute to the Work of Paul Allen*. Montpellier: INIBAP.

Ruas, M., Guignon, V., Sempere, G. *et al.* (2017) MGIS: managing banana (Musa spp.) genetic resources information and high-throughput genotyping data. *Database*, **2017**, bax046.

Sardos, J., Perrier, X., Dolezel, J., Hribova, E., Christelova, P., Van den Houwe, I., Kilian, A. and Roux, N. (2016) DArT whole genome profiling provides insights on the evolution and taxonomy of edible Banana (*Musa* spp.). *Ann. Bot.* **118**, 1269–1278.

Shepherd, K. (1999) *Cytogenetics of the genus Musa* Montpellier. France: INIBAP.

Simmonds, N.W. (1962) *The Evolution of the Bananas*. London, UK: Longmans.

Simmonds, N.W. and Shepherd, K. (1955) The taxonomy and origins of the cultivated bananas. *Bot. J. Linn. Soc.* **55**, 302–312.

Stover, R.H. and Simmonds, N.W. (1987) *Bananas*. Harlow, UK: Longman Scientific & Technical.

Wu, G.A., Prochnik, S., Jenkins, J. *et al.* (2014) Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* **32**, 656–662.