1 **Three new genome assemblies support a rapid radiation in *Musa acuminata***
2 **(wild banana)**

3 Rouard M [1*], Droc G [2,3], Martin G [2,3], Sardos J [1], Hueber Y [1], Guignon V [1], Cenci A [1], Geigle B

4 [4], Hibbins M S [5], Yahiaoui N [2,3], Baurens F-C [2,3], Berry V [6], Hahn M W [5], D'Hont A [2,3] and

5 Roux N [1]

6 [1] Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France. [2]

7 CIRAD, UMR AGAP, F-34398 Montpellier, France; [3]: AGAP, Univ Montpellier, CIRAD,

8 INRA, Montpellier SupAgro, Montpellier, France;[4] Computomics GmbH, Tuebingen, Germany [5]

9 Department of Biology and Department of Computer Science, Indiana University, Bloomington,

10 Indiana [6] LIRMM, CNRS – Univ. Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5,

11 France.

12

13 *Corresponding author: Mathieu Rouard (m.rouard@cgiar.org)

14

15 **Key words:** Banana, *Musa ssp*., Incomplete lineage sorting, Phylogenomics, Genome assembly

1

Abstract

16   **Abstract**

17   Edible bananas result from interspecific hybridization between *Musa acuminata* and *Musa*

18   *balbisiana*, as well as among subspecies in *M. acuminata*. Four particular *M. acuminata*

19   subspecies have been proposed as the main contributors of edible bananas, all of which radiated

20   in a short period of time in southeastern Asia. Clarifying the evolution of these lineages at a

21   whole-genome scale is therefore an important step toward understanding the domestication and

22   diversification of this crop. This study reports the *de novo* genome assembly and gene annotation

23   of a representative genotype from three different subspecies of *M. acuminata*. These data are

24   combined with the previously published genome of the fourth subspecies to investigate

25   phylogenetic relationships and genome evolution. Analyses of shared and unique gene families

26   reveal that the four subspecies are quite homogenous, with a core genome representing at least

27   50% of all genes and very few *M*. *acuminata* species-specific gene families. Multiple alignments

28   indicate high sequence identity between homologous single copy-genes, supporting the close

29   relationships of these lineages. Interestingly, phylogenomic analyses demonstrate high levels of

30   gene tree discordance, due to both incomplete lineage sorting and introgression. This pattern

31   suggests rapid radiation within *Musa acuminata* subspecies that occurred after the divergence

32   with *M. balbisiana*. Introgression between *M. a.* ssp. *malaccensis* and *M. a.* ssp. *burmannica* was

33   detected across a substantial portion of the genome, though multiple approaches to resolve the

34   subspecies tree converged on the same topology. To support future evolutionary and functional

35   analyses, we introduce the PanMusa database, which enables researchers to exploration of

36   individual gene families and trees.

37   **Background**

38   Bananas are among the most important staple crops cultivated worldwide in both the tropics and

39   subtropics. The wild ancestors of bananas are native to the Malesian Region (including Malaysia

40   and Indonesia) (Simmonds 1962) or to northern Indo-Burma (southwest China). Dating back to

41   the early Eocene (Janssens et al. 2016), the genus *Musa* currently comprises 60 to 70 species

42   divided into two sections, Musa and Callimusa (Häkkinen 2013). Most of modern cultivated

43   bananas originated from natural hybridization between two species from the section Musa, *Musa*

44   *acuminata*, which occurs throughout the whole southeast Asia region, and *Musa balbisiana*,

45   which is constrained to an area going from east India to south China (Simmonds & Shepherd

46   1955). While no subspecies have been defined so far in *M. balbisiana*, *M. acuminata* is further

47   divided into multiple subspecies, among which at least four have been identified as contributors

48   to the cultivated banana varieties, namely *banksii*, *zebrina*, *burmannica*, and *malaccensis*

49   (reviewed in Perrier et al. 2011). These subspecies can be found in geographical areas that are

50   mostly non-overlapping. *Musa acuminata* ssp. *banksii* is endemic to New Guinea. *M. a.* ssp*.*

51   *zebrina* is found in Indonesia (Java island), *M. a.* ssp. *malaccensis* originally came from the

52   Malay Peninsula (De Langhe et al. 2009; Perrier et al. 2011), while *M. a.* ssp. *burmannica* is

53   from Burma (today's Myanmar) (Cheesman 1948).

54   While there are many morphological characters that differentiate *M. acuminata* from *M.*

55   *balbisiana*, the subspecies of *M. acuminata* have only a few morphological differences between

56   them. For instance, *M. a.* ssp. *burmannica* is distinguished by its yellowish and waxless foliage,

57   light brown markings on the pseudostem, and by its compact pendulous bunch and strongly

58   imbricated purple bracts. *M. a.* ssp. *banksii* exhibits slightly waxy leaf, predominantly brown-

59   blackish pseudostems, large bunches with splayed fruits, and non-imbricated yellow bracts. *M. a.*

60   ssp. *malaccensis* is strongly waxy with a horizontal bunch, and bright red non-imbricated bracts,

61   while *M. a.* ssp. *zebrina* is characterized by dark red patches on its dark green leaves (Simmonds

62   1956).

63   Previous studies based on a limited number of markers have been able to shed some light

64   on the relationships among *M. acuminata* subspecies (Sardos et al. 2016; Christelová et al.

65   2017). Phylogenetic studies have been assisted by the availability of the reference genome

66   sequence for a representative of *M. acuminata* ssp. *malaccensis* (D'Hont et al. 2012; Martin et al.

3

67   2016) and a draft *M. balbisiana* genome sequence (Davey et al. 2013). However, the availability

68   of large genomic datasets from multiple (sub)species are expected to improve the resolution of

69   phylogenetic analyses, and thus to provide additional insights on species evolution and their

70   specific traits (Bravo et al. 2018). This is especially true in groups where different segments of

71   the genome have different evolutionary histories, as has been found in *Musaceae* (Christelová,

72   Valárik, et al. 2011). Whole-genome analyses also make it much easier to distinguish among the

73   possible causes of gene tree heterogeneity, especially incomplete lineage sorting (ILS) and

74   hybridization (Folk et al. 2018).

75   Moreover, the availability of multiple reference genome sequences opens the way to so-

76   called pangenome analyses, a concept coined by Tettelin et al. (2005). The pangenome is defined

77   as the set of all gene families found among a set of phylogenetic lineages. It includes i) the core

78   genome, which is the pool of genes common to all lineages, ii) the accessory genome, composed

79   of genes absent in some lineages, and iii) the species-specific or individual-specific genome,

80   formed by genes that are present in only a single lineage. Identifying specific compartments of

81   the pangenome (such as the accessory genome) offers a way to detect important genetic

82   differences that underlie molecular diversity and phenotypic variation (Morgante et al. 2007).

83   Here, we generated three *de novo* genomes for the subspecies *banksii*, *zebrina* and

84   *burmannica*, and combined these with existing genomes for *M. acuminata* ssp. *malaccensis*

85   (D'Hont et al. 2012) and *M. balbisiana* (Davey et al. 2013). We thus analyzed the whole genome

86   sequences of five extant genotypes comprising the four cultivated bananas' contributors from *M.*

87   *acuminata,* i.e. the reference genome 'DH Pahang' belonging to *M. acuminata* ssp. *malaccensis*,

88   'Banksii' from *M. acuminata* ssp. *banksii*, 'Maia Oa' belonging to *M. acuminata* ssp. *zebrina*,

89   and 'Calcutta 4' from *M. acuminata* ssp. *burmannica*, as well as *M. balbisiana* (i.e. 'Pisang

90   Klutuk Wulung' or PKW). We carried out phylogenomic analyses that provided evolutionary

91   insights into both the relationships and genomic changes among lineages in this clade. Finally,

92   we developed a banana species-specific database to support the larger community interested in

93   crop improvement.

Results

Assembly and gene annotation

We generated three *de novo* assemblies belonging to *M. acuminata* ssp. *banksii*, *M. a.* ssp. *zebrina* and *M. a.* ssp. *burmannica* (**S Table 1 & 2**). The number of predicted protein coding genes per genome within different genomes of *Musa* ranges from 32,692 to 45,069 (**S Table 4**). Gene number was similar for *M. a.* ssp. *malaccensis* 'DH Pahang', *M. balbisiana* 'PKW' and *M. a.* ssp. *banksii* 'Banksii' but higher in *M. a.* ssp. *zebrina* 'Maia Oa' and *M. a.* ssp. *burmannica* 'Calcutta 4'. According to BUSCO (**S. Table 3**), the most complete gene annotations are 'DH Pahang' (96.5%), 'Calcutta 4' (74.2%) and 'Banksii' (72.5%), followed by 'PKW' (66.5%) and 'Maia Oa' (61.2%).

Gene families

The percentage of genes in orthogroups (OGs), which is a set of orthologs and recent paralogs (*i.e.* gene family), ranges from 74 in *M. a. zebrina* 'Maia Oa' to 89.3 in *M. a. malaccensis* 'DH Pahang' with an average of 79.8 (**Table 1**). Orthogroups have a median size of 4 genes and do not exceed 50 (**S. Table5**). A pangenome here was defined on the basis of the analysis of OGs in order to define the 1) core, 2) accessory, and 3) unique gene set(s). On the basis of the five genomes studied here, the pangenome embeds a total of 32,372 OGs composed of 155,222 genes. The core genome is composed of 12,916 OGs (**Figure 1**). Among these, 8,030 are composed of only one sequence in each lineage (*i.e.* are likely single-copy orthologs). A set of 1489 OGs are specific to all subspecies in *M. acuminata*, while the number of genes specific to each subspecies ranged from 14 in the *M. acuminata* 'DH Pahang' to 110 in *M. acuminata* 'Banksii' for a total of 272 genes across all genotypes. No significant enrichment for any Gene Ontology (GO) category was detected for subspecies-specific OGs (**S. data 1**).

Variation in gene tree topologies

Phylogenetic reconstruction performed with single-copy genes (*n*=8,030) showed high levels of discordance among the different individual gene trees obtained, both at the nucleic acid and protein levels (**Figure 2A**). Considering *M. balbisiana* as outgroup, there are 15 possible bifurcating tree topologies relating the four *M. acuminata* subspecies. For all three partitions of the data - protein, CDS, and gene (including introns and UTRs) - we observed all 15 different

5

123   topologies (**Table 2**). We also examined topologies at loci that had bootstrap support greater than

124   90 for all nodes, also finding all 15 different topologies (**Table 2**). Among trees constructed from

125   whole genes, topologies ranged in frequency from 13.12% for the most common tree to 1.92%

126   for the least common tree (**Table 2**) with an average length of the 1342 aligned nucleotide sites

127   for CDS and 483 aligned sites for proteins. Based on these results, gene tree frequencies were

128   used to calculate concordance factors on the most frequent CDS gene trees (**Table 2**),

129   demonstrating that no split was supported by more than 30% of gene trees (**Figure 2B**).

130   Therefore, in order to further gain insight into the subspecies phylogeny, we used a combination

131   of different approaches described in the next section.

132   **Inference of a species tree**

133   We used three complementary methods to infer phylogenetic relationships among the sampled

134   lineages. First, we concatenated nucleotide sequences from all single-copy genes (totaling

135   11,668,507 bp). We used PHYML to compute a maximum likelihood tree from this alignment,

136   which, as expected, provided a topology with highly supported nodes **(Figure 3A).** Note that this

137   topology (denoted topology number 1 in **Table 2**) is not the same as the one previously proposed

138   in the literature (denoted topology number 7 in **Table 2**) (**S. Figure 1 & 2**).

139   Next, we used a method explicitly based on individual gene tree topologies. ASTRAL

140   (Mirarab & Warnow 2015) infers the species tree by using quartet frequencies found in gene

141   trees. It is suitable for large datasets and was highlighted as one of the best methods to address

142   challenging topologies with short internal branches and high levels of discordance (Shi & Yang

143   2017). ASTRAL found the same topology using ML gene trees from single-copy genes obtained

144   from protein sequences, CDSs, and genes (**Figure 3C**).

145   Finally, we ran a supertree approach implemented in PhySIC_IST (Scornavacca et al.

146   2008) on the single-copy genes and obtained again the same topology (**Figure 3B**). PhySIC_IST

147   first collapses poorly supported branches of the gene trees into polytomies, as well as conflicting

148   branches of the gene trees that are only present in a small minority of the trees; it then searches

149   for the most resolved supertree that does not contradict the signal present in the gene trees nor

150   contains topological signal absent from those trees. Deeper investigation of the results revealed

151   that ~ 66% of the trees were unresolved, 33% discarded (pruned or incorrectly rooted), and

6

152   therefore that the inference relied on fewer than 1% of the trees. Aiming to increase the number

153   of genes used by PhySIC_IST, we included multi-copy OGs of the core genome, as well as some

154   OGs in the accessory genomes using the pipeline SSIMUL (Scornavacca et al. 2011). SSIMUL

155   translates multi-labeled gene trees (MUL-trees) into trees having a single copy of each gene (X-

156   trees), i.e. the type of tree usually expected in supertree inference. To do so, all individual gene

157   trees were constructed on CDSs from OGs with at least 4 *M. acuminata* and *M. balbisiana* genes

158   (n=18,069). SSIMUL first removed identical subtrees resulting from a duplication node in these

159   trees, it then filtered out trees where duplicated parts induced contradictory rooted triples,

160   keeping only coherent trees. These trees can then be turned into trees containing a single copy of

161   each gene, either by pruning the smallest subtrees under each duplication node (leaving only

162   orthologous nodes in the tree), or by extracting the topological signal induced by orthology

163   nodes into a rooted triplet set, that is then turned back into an equivalent X-tree. Here we chose

164   to use the pruning method to generate a dataset to be further analyzed with PhySIC_IST, which

165   lead to a subset of 14,507 gene trees representing 44% of the total number of OGs and an

166   increase of 80% compared to the 8,030 single-copy OGs. This analysis returned a consensus

167   gene tree with the same topology as both of the previous methods used here (**Figure 3B**).

168   **Evidence for introgression**

169   Although much of the discordance we observe is likely due to incomplete lineage sorting, we

170   also tested for introgression between subspecies. The ABBA-BABA test (Green et al. 2010) was

171   conducted to detect an excess of either ABBA or BABA sites (where "A" corresponds to the

172   ancestral allele and "B" corresponds to the derived allele state) in a four-taxon phylogeny

173   including three *M. acuminata* subspecies as ingroups and *M. balbisiana* as outgroup. Because

174   there were five taxa to be tested, analyses were done with permutation of taxa denoted P1, P2

175   and P3 and Outgroup (**Table 3**). Under the null hypothesis of ILS, an equal number of ABBA

176   and BABA sites are expected. However, we always found an excess of sites grouping

177   *malaccensis* ('DH') and *burmannica* ('C4') (**Table 3**). This indicates a history of introgression

178   between these two lineages.

179   To test the direction of introgression, we applied the D_2 test (Hibbins and Hahn,

180   unpublished). While introgression between a pair of species (e.g. *malaccensis* and *burmannica*)

181   always results in smaller genetic distances between them, the D_2 test is based on the idea that

182   gene flow in the two alternative directions can also result in a change in genetic distance to other

183   taxa not involved in the exchange (in this case, *banksii*). We computed the genetic distance

184   between *banksii* and *burmannica* in gene trees where *malaccensis* and *banksii* are sister (denoted

185    $d_{AC}|A,B$) and the genetic distance between *banksii* and *burmannica* in gene trees where

186   *malaccensis* and *burmannica* are sister (denoted $d_{AC}|B,C$). The test takes into account the genetic

187   distance between the species not involved in the introgression (*banksii*) and the species involved

188   in introgression that it is not most closely related to (*burmannica*). We identified 1454 and 281

189   gene trees with $d_{AC}|A,B=1.15$ and $d_{AC}|B,C = 0.91$, respectively, giving a significant positive

190   value of $D_2=0.23$ (*P*<0.001 by permutation). These results support introgression from

191   *malaccensis* into *burmannica*, though they do not exclude the presence of a lesser level of gene

192   flow in the other direction.

PanMusa, a database to explore individual OGs

194   Since genes underlie traits and wild banana species showed a high level of incongruent gene tree

195   topologies, access to a repertoire of individual gene trees is important. This was the rationale for

196   constructing a database that provides access to gene families and individual gene family trees in

197   *M. acuminata* and *M. balbisiana*. A set of web interfaces are available to navigate OGs that have

198   been functionally annotated using GreenPhyl comparative genomics database (Rouard et al.

199   2011). PanMusa shares most of the features available on GreenPhyl to display or export

200   sequences, InterPro assignments, sequence alignments, and gene trees (**Figure 4**). In addition,

201   new visualization tools were implemented, such as MSAViewer (Yachdav et al. 2016) and

202   PhyD3 (Kreft et al. 2017) to view gene trees.

Discussion

M. acuminata subspecies contain few subspecies-specific families

205   In this study, we used a *de novo* approach to generate additional reference genomes for the three

206   subspecies of *Musa acuminata*; all three are thought to have played significant roles as genetic

207   contributors to the modern cultivars. Genome assemblies produced for this study differ in

208   quality, but the estimation of genome assembly and gene annotation quality conducted with

209   BUSCO suggests that they were sufficient to perform comparative analyses. Moreover, we

210   observed that the number of genes grouped in OGs were relatively similar among subspecies,

211   indicating that the potential over-prediction of genes in 'Maia Oa' and 'Calcutta 4' was mitigated

212   during the clustering procedure. Indeed, over-prediction in draft genomes is expected due to

213   fragmentation, leading to an artefactual increase in the number of genes (Denton et al. 2014).

214   Although our study is based on one representative per subspecies, *Musa* appears to have a

215   widely shared pangenome, with only a small number of subspecies-specific families identified.

216   The pangenome analysis also reveals a large number of families shared only among subsets of

217   species or subspecies (**Figure 1**); this "dispensable" genome is thought to contribute to diversity

218   and adaptation (Tettelin et al. 2005; Kahlke et al. 2012). The small number of species-specific

219   OGs in *Musa acuminata* also supports the recent divergence between all genotypes including the

220   split between *M. acuminata* and *M. balbisiana*.

221   ***M. acuminata* subspecies show a high level of discordance between individual gene trees**

222   By computing gene trees with all single-copy genes OG, we found widespread discordance in

223   gene tree topologies. Topological incongruence can be the result of incomplete lineage sorting,

224   the misassignment of paralogs as orthologs, introgression, or horizontal gene transfer (Maddison

225   1997). With the continued generation of phylogenomic datasets over the past dozen years,

226   massive amounts of discordance have been reported, first in *Drosophila* (Pollard et al. 2006) and

227   more recently in birds (Jarvis et al. 2014), mammals (Li et al. 2016; Shi & Yang 2018) and

228   plants (Novikova et al. 2016; Pease et al. 2016; Choi et al. 2017; Copetti et al. 2017; Wu et al.

229   2017). Due to the risk of hemiplasy in such datasets (Avise et al. 2008; Hahn & Nakhleh 2016),

230   we determined that we could not accurately reconstruct either nucleotide substitutions or gene

231   gains and losses among the genomes analyzed here.

232   In our case, the fact that all possible subspecies tree topologies occurred, and that ratios

233   of minor trees at most nodes were equivalent to those expected under ILS, strongly suggests the

234   presence of ILS (Hahn & Nakhleh 2016). Banana is a paleopolyploid plant that experienced

235   three independent whole genome duplications (WGD), and some fractionation is likely still

236   occurring (D'Hont et al. 2012) (**S. Table 6**). But divergence levels among the single-copy OGs

237   were fairly consistent **(Figure 2A)**, supporting the correct assignment of orthology among

238   sequences**.** However, we did find evidence for introgression between *malaccensis* and

239   *burmannica*, which contributed a small excess of sites supporting one particular discordant

240   topology (**Table 3**). This event is also supported by the geographical overlap in the distribution

241   of these two subspecies (Perrier et al. 2011).

242   The species tree topology supported by all methods used here is different from the tree

243   previously proposed in the literature (**S. Figure 1**). 'Calcutta 4' as representative of *M.*

244   *acuminata* ssp. *burmannica* was placed sister to the other *Musa acuminata* genotypes in our

245   study, whereas several studies have reported proximity between *burmannica* and *malaccensis*,

246   here represented by 'DH Pahang' (Janssens et al. 2016; Christelová et al. 2017). Multiple

247   previous studies have attempted to resolve the topology in the Musaceae, but did not include all

248   subspecies considered here, and had very limited numbers of loci. In Christelova et al. (2011), a

249   robust combined approach using maximum likelihood, maximum parsimony, and Bayesian

250   inference was applied to 19 loci, but only *burmannica* and *zebrina* out of the four subspecies

251   were included. Jarret et al. (1992) reported sister relationships between *malaccensis* and *banksii*

252   on the basis of RFLP markers, but did not include any samples from *burmannica* and *zebrina*. It

253   is worth noting that, on the bases of our resolved topology, introgression from *malaccensis* to

254   *burmannica* was detected, and could explain the relationships described previously (Janssens et

255   al. 2016; Sardos et al. 2016).

256   More strikingly considering previous phylogenetic hypotheses, *malaccensis* appeared

257   most closely related to *banksii*, which is quite distinct from the other *M. acuminata* spp.

258   (Simmonds & Weatherup 1990) and which used to be postulated as its own species based on its

259   geographical area of distribution and floral diversity (Argent 1976). On the bases of genomic

260   similarity, all our analyses support *M. acuminata* ssp. *banksii* as a subspecies of *M. acuminata*.

261   **Gene tree discordance supports rapid radiation of *Musa acuminata* subspecies**

262   In their evolutionary history, *Musa* species dispersed from 'northwest to southeast' into

263   Southeast Asia (Janssens et al. 2016). Due to sea level fluctuations, Malesia (including the

264   nations of Indonesia, Malaysia, Brunei, Singapore, the Philippines, and Papua New Guinea) is a

265   complex geographic region, formed as the result of multiple fusions and subsequent isolation of

266   different islands (Thomas et al. 2012; Janssens et al. 2016). Ancestors of the Callimusa section

267   (of the *Musa* genus) started to radiate from the northern Indo-Burma region towards the rest of

268   Southeast Asia ~30 MYA, while the ancestors of the Musa (formerly Eumusa/Rhodochlamys)

269   section started to colonize the region ~10 MYA (Janssens et al. 2016). The divergence between

10

270   *M. acuminata* and *M. balbisiana* has been estimated to be ~5 MYA (Lescot et al. 2008).

271   However, no accurate dating has yet been proposed for the divergence of the *Musa acuminata*

272   subspecies. We hypothesize that after the speciation of *M. acuminata* and *M. balbisiana* (circa 5

273   MYA) rapid diversification occurred within populations of *M. acuminata*. This hypothesis is

274   consistent with the observed gene tree discordance and high levels of ILS. Such a degree of

275   discordance may reflect a near-instantaneous radiation between all subspecies of *M. acuminata.*

276   Alternatively, it could support the proposed hypothesis of divergence back in the northern part of

277   Malesia during the Pliocene (Janssens et al. 2016), followed by introgression taking place among

278   multiple pairs of species as detected between *malaccensis* and *burmannica*. While massive

279   amounts of introgression can certainly mask the history of lineage splitting (Fontaine et al.

280   2015), we did not find evidence for such mixing.

281   Interestingly, such a broad range of gene tree topologies due to ILS (and introgression)

282   has also been observed in gibbons (Carbone et al. 2014; Veeramah et al. 2015; Shi & Yang

283   2018) for which the area of distribution in tropical forests of Southeast Asia is actually

284   overlapping the center of origin of wild bananas. Moreover, according to Carbone et al. (2014),

285   gibbons also experienced a near-instantaneous radiation ∼5 million years ago. It is therefore

286   tempting to hypothesize that ancestors of wild bananas and ancestors of gibbons faced similar

287   geographical isolation and had to colonize and adapt to similar ecological niches, leading to the

288   observed patterns of incomplete lineage sorting.

289   In this study, we highlighted the phylogenetic complexity in a genome-wide dataset for

290   *Musa acuminata* and *Musa balbisiana*, bringing additional insights to explain why the Musaceae

291   phylogeny has remained controversial. Our work should enable researchers to make inferences

292   about trait evolution, and ultimately should help support crop improvement strategies.

Material and Methods

Plant material

295   Banana leaf samples from accessions 'Banksii' (*Musa acuminata* ssp. *banksii*, PT-BA-00024),

296   'Maia Oa' (*Musa acuminata* ssp. *zebrina*, PT-BA-00182) and 'Calcutta 4' (*Musa acuminata* ssp.

297   *burmannica*, PT-BA-00051) were supplied by the CRB-Plantes Tropicales Antilles CIRAD-

298   INRA field collection based in Guadeloupe. Leaves were used for DNA extraction. Plant identity

299   was verified at the subspecies level using SSR markers at the *Musa* Genotyping Centre (MGC,

11

300   Czech Republic) as described in (Christelová, Valarik, et al. 2011) and passport data of the plant

301   is accessible in the Musa Germplasm Information System (Ruas et al. 2017). In addition, the

302   representativeness of the genotypes of the four subspecies was verified on a set of 22 samples

303   belonging to the same four *M. acuminata* subspecies of the study (**S. Figure 3**).

Sequencing and assembly

305   Genomic DNA was extracted using a modified MATAB method (Risterucci et al. 2000) . DNA

306   libraries were constructed and sequenced using the HiSeq2000 (Illumina) technology at BGI (**S.**

307   **Table 1**). 'Banksii' was assembled using SoapDenovo (Luo et al. 2012), and PBJelly2 (English

308   et al. 2012) was used for gap closing using PacBio data generated at the Norwegian Sequencing

309   Center (NSC) with Pacific Biosciences RS II. 'Maia Oa' and 'Calcutta 4' were assembled using

310   the MaSuRCA assembler (Zimin et al. 2013) (**S. Table 2**). Estimation of genome assembly

311   completeness was assessed with BUSCO plant (Simão et al. 2015) (**S. Table 3**).

Gene annotation

313   Gene annotation was performed on the obtained *de novo* assembly for 'Banksii', 'Maia Oa' and

314   'Calcutta 4,' as well as on the draft *Musa balbisiana* 'PKW' assembly (Davey et al, 2013) for

315   consistency and because the published annotation was assessed as low quality. For structural

316   annotation we used EuGene v4.2 (http://eugene.toulouse.inra.fr/) (Foissac et al. 2008) calibrated

317   on *M. acuminata malaccensis* 'DH Pahang' reference genome v2, which produced similar results

318   (e.g. number of genes, no missed loci, good specificity and sensitivity) as the official annotation

319   (Martin et al. 2016). EuGene combined genotype-specific (or closely related) transcriptome

320   assemblies, performed with Trinity v2.4 with RNAseq datasets (Sarah et al. 2016), to maximize

321   the likelihood to have genotype-specific gene annotation (**S. Table 4**). The estimation of gene

322   space completeness was assessed with Busco (**S. Table 3**). Because of its high quality and to

323   avoid confusing the community, we did not perform a new annotation for the *M. a. malaccensis*

324   'DH Pahang' reference genome but used the released version 2. Finally, the functional

325   annotation of plant genomes was performed by assigning their associated generic GO terms

326   through the Blast2GO program (Conesa et al. 2005) combining BLAST results from UniProt (E-

327   value 1e-5) (Magrane & Consortium 2011) (**S. Data 1**).

Gene families

Gene families were identified using OrthoFinder v1.1.4 (Emms & Kelly 2015) with default parameters based on BLASTp (e-value 1e-5). Venn diagrams were made using JVenn online (http://jvenn.toulouse.inra.fr) (Bardou et al. 2014) and alternate visualization was produced with UpsetR (https://gehlenborglab.shinyapps.io/upsetr) (Lex et al. 2014).

Tree topology from literature

A species tree was initially identified based on previous studies (Sardos et al. 2016; Janssens et al. 2016). Those two studies included all *M. acuminata* subspecies, and had the same tree topology (**S. Figure 1**). In the first study, Sardos et al, (2016) computed a Neighbor-Joining tree from a dissimilarity matrix using bi-allelic GBS-derived SNP markers along the 11 chromosomes of the *Musa* reference genome. Several representatives of each subspecies that comprised genebank accessions related to the genotypes used here were included (Sardos et al. 2016). We annotated the tree to highlight the branches relevant to *M. acuminata* subspecies (**S. Figure 2**). In the second study, a maximum clade credibility tree of Musaceae was proposed based on four gene markers (*rps16*, *atpB-rbcL*, *trnL-F* and internal transcribed spacer, ITS) analyzed with Bayesian methods (Janssens et al. 2016).

Genome-scale phylogenetic analyses and species tree

Single-copy OGs (i.e. orthogroups with one copy of a gene in each of the five genotypes) from protein, coding DNA sequence (CDS), and genes (including introns and UTRs) were aligned with MAFFT v7.271 (Katoh & Standley 2013), and gene trees were constructed using PhyML v3.1 (Guindon et al. 2009) with ALrT branch support. All trees were rooted using *Musa balbisiana* as outgroup using Newick utilities v1.6 (Junier & Zdobnov 2010) (**S. Data 2**). Individual gene tree topologies were visualized as a cloudogram with DensiTree v2.2.5 (Bouckaert 2010).

Single-copy OGs were further investigated with the quartet method implemented in ASTRAL v5.5.6 (Mirarab & Warnow 2015). In parallel, we carried out a Supertree approach following the SSIMUL procedure (http://www.atgc-montpellier.fr/ssimul/) (Scornavacca et al. 2011) combined with PhySIC_IST (http://www.atgc-montpellier.fr/physic_ist/) (Scornavacca et al. 2008) applied to a set of rooted trees corresponding to core OGs (including single and

357 multiple copies), and accessory genes for which only one representative species was missing

358 (except outgroup species). Finally, single-copy OGs (CDS only) were used to generate a

359 concatenated genome-scale alignment with FASconCAT-G (Kück & Longo 2014) and a tree was

360 constructed using PhyML (NNI, HKY85, 100 bootstrap).

Search for introgression

362 Ancient gene flow was assessed with the ABBA-BABA test or *D*-statistic (Green et al. 2010;

363 Durand et al. 2011) and computed on the concatenated multiple alignment converted to the MVF

364 format and processed with MVFtools (Pease & Rosenzweig 2017), similar to what is described

365 in Wu et al. (2017) (https://github.com/wum5/JaltPhylo) . The direction of introgression was

366 further assessed with the D_2 test (Hibbins and Hahn, unpublished). The D_2 statistic captures

367 differences in the heights of genealogies produced by introgression occurring in alternate

368 directions by measuring the average divergence between species A and C in gene trees with an

369 ((A,B),C) topology (denoted [d_{AC}|A,B]), and subtracting the average A-C divergence in gene

370 trees with a ((B,C),A) topology (denoted [d_{AC}|B,C]), so that $D_2 = (d_{AC}|A,B) - (d_{AC}|B,C)$. If the

371 statistic is significantly positive, it means that introgression has either occurred in the B→C

372 direction or in both directions. D_2 significance was assessed by permuting labels on gene trees

373 1000 times and calculating *P*-values from the resulting null distribution of D_2 values. The test

374 was implemented with a Perl script using distmat from EMBOSS (Rice et al. 2000) with Tajima-

375 Nei distance applied to multiple alignments associated with gene trees fitting the defined

376 topologies above (https://github.com/mrouard/perl-script-utils).

14

Data availability

377 Raw sequence reads for *de novo* assemblies were deposited in the Sequence Read Archive (SRA)
378 of the National Center for Biotechnology Information (NCBI) (BioProject: PRJNA437930 and
380 SRA: SRP140622). Assembly and gene annotation data are available on the Banana Genome
381 Hub (Droc et al, 2013) (http://banana-genome-hub.southgreen.fr/species-list). Cluster and gene
382 tree results are available on a dedicated database (http://panmusa.greenphyl.org) hosted on the
383 South Green Bioinformatics Platform (Guignon et al. 2016).

Acknowledgments

395

Authors contribution

397 MR, NR and AD set up the study and MR coordinated the study. AD and FCB provided access
398 to plant material and DNA. NY provided access to transcriptome data and GM to repeats library
399 for gene annotation. BG performed assembly and gap closing. MR, GD, GM, YH, JS and AC
400 performed analyses. VB, MSH, and MWH provided guidance on methods and helped with result
401 interpretation. VG and MR set up the PanMusa website. MR wrote the manuscript with
402 significant contributions from MWH, VB, and JS, and all co-authors commented on the
403 manuscript

15

404   **Figures and Legends**

405   **Figure 1. Five-way Venn diagram showing the distribution of shared gene families** (at least

406   two sequences per OG) among *M. a. banksii* 'Banksii', *M. a. zebrina* 'Maia Oa', *M. a.*

407   *burmannica* 'Calcutta 4', *M. a. malaccensis* 'DH Pahang' and *M. balbisiana* 'PKW' genomes. On

408   the right, number of orthologous groups by species and pangenome category. At the bottom,

409   same dataset visualized with UpsetR (Lex et al. 2014).

410   **Figure 2. Illustration of gene tree discordance.** A) Cloudogram of single copy OGs (CDS)

411   visualized with Densitree. The blue line represents the consensus tree as provided by Densitree

412   B) Species tree with bootstrap-like support based on corresponding gene tree frequency from

413   Table 2 (denoted topology number 2). (PKW = *M. balbisiana* 'PKW', C4 = *M. acuminata*

414   *burmannica* 'Calcutta 4, M= *M. acuminata zebrina* 'Maia Oa', DH= *M. acuminata malaccensis*

415   "DH Pahang', B = *M. acuminata banksii* 'Banksii')

416   **Figure 3. Species topologies computed with three different approaches** A) Maximum

417   likelihood tree inferred from a concatenated alignment of single-copy genes (CDS). B)

418   Supertree-based method applied to single and multi-labelled gene trees C) Quartet-based model

419   applied to protein, CDS, and gene alignments.

420   **Figure 4. Overview of available interfaces for the PanMusa database**. A. Homepage of the

421   website. B. List of functionally annotated OGs. C. Graphical representation of the number of

422   sequence by species. D. Consensus InterPro domain schema by OG. E. Individual gene trees

423   visualized with PhyD3. F. Multiple alignment of OG with MSAviewer.

424   **Figure 5. Area of distribution of Musa species in Southeast Asia** as described by Perrier et al,

425   2011; including species tree of *Musa acuminata* subspecies based on results described in Figure

426   4. Areas of distribution are approximately represented by colors; hatched zone shows area of

427   overlap between two subspecies where introgression may have occurred.

16

428 **Table 1. Summary of the gene clustering statistics per (sub)species.**

	M. acuminata malaccensis 'DH Pahang'	*M. acuminata burmannica* 'Calcutta 4'	*M. acuminata banksii* 'Banksii'	*M. acuminata zebrina* 'Maia Oa'	*M. balbisiana* 'PKW'
# genes	35,276	45,069	32,692	44,702	36,836
# genes in orthogroups	31,501	34,947	26,490	33,059	29,225
# unassigned genes	3,775	10,122	6,202	11,643	7,611
% genes in orthogroups	89.3	77.5	81	74	79.3
% unassigned genes	10.7	22.5	19	26	20.7
# orthogroups containing species	24,074	26,542	21,446	25,730	23,935
% orthogroups containing species	74.4	82	66.2	79.5	73.9
# species-specific orthogroups	6	46	47	11	9
# genes in species-specific orthogroups	14	104	110	23	21
% genes in species-specific orthogroups	0	0.2	0.3	0.1	0.1

429

17

430 **Table 2.** Frequency of gene tree topologies of the 8,030 single copy OGs. (PKW = *Musa balbisiana*

431 'PKW', C4 = *Musa acuminata burmannica* 'Calcutta 4, M= *Musa acuminata zebrina* 'Maia Oa', DH=

432 *Musa acuminata malaccensis* "DH Pahang', B = *Musa acuminata banksii* 'Banksii'). In bold the most

433 frequent topology.

No.	Topology	# CDS (%)	# Protein (%)	# Gene (%)	# Gene bootstrap >90 (%)
1	(PKW,(C4,(M,(DH,B))))	**11.9**	10.58	**13.12**	13.72
2	(PKW,(C4,(DH,(B,M))))	10.8	10.48	11.92	14.88
3	(PKW,((DH,C4),(B,M)))	9.59	7.28	12.73	**17.52**
4	(PKW,(M,(C4,(DH,B))))	9.53	**12.51**	7.78	5.91
5	(PKW,(C4,(B,(DH,M))))	8.02	7.37	8.89	8.44
6	(PKW,((DH,B),(C4,M)))	7.67	6.55	9.16	12.56
7	(PKW,(M,(B,(DH,C4))))	6.66	8.21	5	3.06
8	(PKW,(B,(M,(DH,C4))))	5.58	5.23	4.61	2.53
9	(PKW,(DH,(C4,(B,M))))	5.41	5.21	5.18	4.96
10	(PKW,(B,(C4,(DH,M))))	5.26	4.45	6.2	7.07
11	(PKW,(B,(DH,(C4,M))))	5.02	6.82	3.36	1.9
12	(PKW,(M,(DH,(B,C4))))	4.23	4.68	2.84	1.16
13	(PKW,((DH,M),(B,C4)))	4.037	3.61	4.79	5.06
14	(PKW,(DH,(B,(C4,M))))	3.85	4.18	2.44	0.63
15	(PKW,(DH,(M,(B,C4))))	2.38	2.77	1.92	0.52

18

434 **Table 3.** Four-taxon ABBA-BABA Test (*D*-statistic) used for introgression inference from the well-supported topology from Figure 3.

435 [a] Discordance = (ABBA + BABA) / Total [b] D = (ABBA-BABA) / (ABBA+BABA)

P1	P2	P3	BBAA	ABBA	BABA	Discordance[a]	D[b]	P-value
Malaccensis (DH)	Banksii (B)	Burmannica (C4)	12185	4289	8532	0.51	0.33	0
Malaccensis (DH)	Zebrina (M)	Burmannica (C4)	9622	5400	9241	0.6	0.26	0
Zebrina (M)	Banksii (B)	Burmannica (C4)	11204	6859	6782	0.54	-0.005	0.5
Malaccensis (DH)	Banksii (B)	Zebrina (M)	10450	7119	6965	0.57	-0.02	0.19

436

437   **Additional information**

438   **S. Figure 1. Species tree of Musa acuminata subspecies extrapolated from literature review**

439   **S. Figure 2. Neighbor-Joining tree from 105 M. acuminata and cultivated accessions**

440   **S. Figure 3. Individual ancestries investigated with the Admixture software package**

441   **S. Table 1. Libraries used for the genome assemblies**

442   **S Table 2. Summary of the genome assembly**

443   **S. Table 3. Results of gene space assessment with BUSCO**

444   **S. Table 4. Summary of the genome annotation**

445   **S. Table 5. Global summary of the gene clustering**

446   **S. Table 6. List of 18 phylogenetic informative shared single copy nuclear genes from**
447   **Duarte et al. 2010 mapped to *Musa* genomes.**

448   **S. Data 1. List of Gene ontology mapped by genomes**

449   **S. Data 2. List of gene trees obtained at protein-coding, CDS and gene based level**

References

451 Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual
452 ancestry estimation. BMC Bioinformatics. 12:246. doi: 10.1186/1471-2105-12-246.

453 Argent G. 1976. The wild bananas of Papua New Guinea. Notes Roy Bot Gard Edinb. 35:77–
454 114.

455 Avise JC, Robinson TJ, Kubatko L. 2008. Hemiplasy: A New Term in the Lexicon of
456 Phylogenetics. Syst. Biol. 57:503–507. doi: 10.1080/10635150802164587.

457 Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. 2014. jvenn: an interactive Venn diagram
458 viewer. BMC Bioinformatics. 15:293. doi: 10.1186/1471-2105-15-293.

459 Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics.
460 26:1372–1373. doi: 10.1093/bioinformatics/btq110.

461 Bravo GA et al. 2018. *Embracing heterogeneity: Building the Tree of Life and the future of*
462 *phylogenomics*. PeerJ Inc. doi: 10.7287/peerj.preprints.26449v3.

463 Carbone L et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. Nature.
464 513:195–201. doi: 10.1038/nature13679.

465 Cheesman EE. 1948. Classification of the Bananas. Kew Bull. 3:17–28. doi: 10.2307/4118909.

466 Choi JY et al. 2017. The Rice Paradox: Multiple Origins but Single Domestication in Asian Rice.
467 Mol. Biol. Evol. 34:969–979. doi: 10.1093/molbev/msx049.

468 Christelová P, Valarik M, et al. 2011. A platform for efficient genotyping in Musa using
469 microsatellite markers. AoB Plants. 2011:plr024–plr024. doi: 10.1093/aobpla/plr024.

470 Christelová P et al. 2017. Molecular and cytological characterization of the global Musa
471 germplasm collection provides insights into the treasure of banana diversity. Biodivers. Conserv.
472 26:801–824. doi: 10.1007/s10531-016-1273-9.

473 Christelová P, Valárik M, Hřibová E, De Langhe E, Doležel J. 2011. A multi gene sequence-
474 based phylogeny of the Musaceae (banana) family. BMC Evol. Biol. 11:103. doi: 10.1186/1471-
475 2148-11-103.

476 Conesa A et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in
477 functional genomics research. Bioinforma. Oxf. Engl. 21:3674–3676. doi:
478 10.1093/bioinformatics/bti610.

479 Copetti D et al. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of
480 North American columnar cacti. Proc. Natl. Acad. Sci. 114:12003–12008. doi:
481 10.1073/pnas.1706367114.

21

482   Dagan T, Martin W. 2006. The tree of one percent. Genome Biol. 7:118. doi: 10.1186/gb-2006-
483   7-10-118.

484   Davey MW et al. 2013. A draft Musa balbisiana genome sequence for molecular genetics in
485   polyploid, inter- and intra-specific Musa hybrids. BMC Genomics. 14:683. doi: 10.1186/1471-
486   2164-14-683.

487   De Langhe E et al. 2009. Why Bananas Matter: An introduction to the history of banana
488   domestication. Ethnobot. Res. Appl. 7:165–177. doi: 10.17348/era.7.0.165-177.

489   Denton JF et al. 2014. Extensive Error in the Number of Genes Inferred from Draft Genome
490   Assemblies. PLOS Comput. Biol. 10:e1003998. doi: 10.1371/journal.pcbi.1003998.

491   D'Hont A et al. 2012. The banana (Musa acuminata) genome and the evolution of
492   monocotyledonous plants. Nature. doi: 10.1038/nature11241.

493   Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for Ancient Admixture between
494   Closely Related Populations. Mol. Biol. Evol. 28:2239–2252. doi: 10.1093/molbev/msr048.

495   Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome
496   comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157. doi:
497   10.1186/s13059-015-0721-2.

498   English AC et al. 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-
499   Read Sequencing Technology. PLOS ONE. 7:e47768. doi: 10.1371/journal.pone.0047768.

500   Foissac S et al. 2008. Genome Annotation in Plants and Fungi: EuGene as a Model Platform.
501   Curr. Bioinforma. http://www.eurekaselect.com/82677/article (Accessed March 1, 2018).

502   Folk Ryan A., Soltis Pamela S., Soltis Douglas E., Guralnick Robert. 2018. New prospects in the
503   detection and comparative analysis of hybridization in the tree of life. Am. J. Bot. 0. doi:
504   10.1002/ajb2.1018.

505   Fontaine MC et al. 2015. Extensive introgression in a malaria vector species complex revealed
506   by phylogenomics. Science. 347:1258524. doi: 10.1126/science.1258524.

507   Guignon V et al. 2016. The South Green portal: A comprehensive resource for tropical and
508   Mediterranean crop genomics. Curr. Plant Biol. 7:6–9.

509   Guindon S, Delsuc F, Dufayard J-F, Gascuel O. 2009. Estimating maximum likelihood
510   phylogenies with PhyML. Methods Mol. Biol. Clifton NJ. 537:113–137. doi: 10.1007/978-1-
511   59745-251-9_6.

512   Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evol. Int. J. Org.
513   Evol. 70:7–17. doi: 10.1111/evo.12832.

514 Janssens SB et al. 2016. Evolutionary dynamics and biogeography of Musaceae reveal a
515 correlation between the diversification of the banana family and the geological and climatic
516 history of Southeast Asia. New Phytol. 210:1453–1465. doi: 10.1111/nph.13856.

517 Jarret R, Gawel N, Whittemore A, Sharrock S. 1992. RFLP-based phylogeny of Musa species in
518 Papua New Guinea. Theor. Appl. Genet. 84–84. doi: 10.1007/BF00224155.

519 Jarvis ED et al. 2014. Whole-genome analyses resolve early branches in the tree of life of
520 modern birds. Science. 346:1320–1331. doi: 10.1126/science.1253451.

521 Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree
522 processing in the UNIX shell. Bioinforma. Oxf. Engl. 26:1669–1670. doi:
523 10.1093/bioinformatics/btq243.

524 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
525 improvements in performance and usability. Mol. Biol. Evol. 30:772–780. doi:
526 10.1093/molbev/mst010.

527 Kreft L, Botzki A, Coppens F, Vandepoele K, Van Bel M. 2017. PhyD3: a phylogenetic tree
528 viewer with extended phyloXML support for functional genomics data visualization.
529 Bioinforma. Oxf. Engl. doi: 10.1093/bioinformatics/btx324.

530 Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment
531 preparations concerning phylogenetic studies. Front. Zool. 11:81. doi: 10.1186/s12983-014-
532 0081-x.

533 Lescot M et al. 2008. Insights into the Musa genome: Syntenic relationships to rice and between
534 Musa species. BMC Genomics. 9:58. doi: 10.1186/1471-2164-9-58.

535 Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: Visualization of
536 Intersecting Sets. IEEE Trans. Vis. Comput. Graph. 20:1983–1992. doi:
537 10.1109/TVCG.2014.2346248.

538 Li G, Davis BW, Eizirik E, Murphy WJ. 2016. Phylogenomic evidence for ancient hybridization
539 in the genomes of living cats (Felidae). Genome Res. 26:1–11. doi: 10.1101/gr.186668.114.

540 Luo R et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo
541 assembler. GigaScience. 1:18. doi: 10.1186/2047-217X-1-18.

542 Maddison WP. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536. doi:
543 10.1093/sysbio/46.3.523.

544 Magrane M, Consortium U. 2011. UniProt Knowledgebase: a hub of integrated protein data.
545 Database J. Biol. Databases Curation. 2011. doi: 10.1093/database/bar009.

546   Martin G et al. 2016. Improvement of the banana "Musa acuminata" reference sequence using
547   NGS data and semi-automated bioinformatics methods. BMC Genomics. 17:243. doi:
548   10.1186/s12864-016-2579-4.

549   Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome.
550   Curr. Opin. Genet. Dev. 15:589–594. doi: 10.1016/j.gde.2005.09.006.

551   Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many
552   hundreds of taxa and thousands of genes. Bioinforma. Oxf. Engl. 31:i44-52. doi:
553   10.1093/bioinformatics/btv234.

554   Morgante M, De Paoli E, Radovic S. 2007. Transposable elements and the plant pan-genomes.
555   Curr. Opin. Plant Biol. 10:149–155. doi: 10.1016/j.pbi.2007.02.001.

556   Novikova PY et al. 2016. Sequencing of the genus Arabidopsis identifies a complex history of
557   nonbifurcating speciation and abundant trans-specific polymorphism. Nat. Genet. 48:1077–1082.
558   doi: 10.1038/ng.3617.

559   Pease J, Rosenzweig B. 2017. Encoding Data Using Biological Principles: the Multisample
560   Variant Format for Phylogenomics and Population Genomics. IEEE/ACM Trans. Comput. Biol.
561   Bioinform. PP:1–1. doi: 10.1109/TCBB.2015.2509997.

562   Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics Reveals Three Sources of
563   Adaptive Variation during a Rapid Radiation. PLOS Biol. 14:e1002379. doi:
564   10.1371/journal.pbio.1002379.

565   Perrier X et al. 2011. Multidisciplinary perspectives on banana (Musa spp.) domestication. Proc.
566   Natl. Acad. Sci. doi: 10.1073/pnas.1102001108.

567   Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread Discordance of Gene Trees with
568   Species Tree in Drosophila: Evidence for Incomplete Lineage Sorting. PLOS Genet. 2:e173. doi:
569   10.1371/journal.pgen.0020173.

570   Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software
571   Suite. Trends Genet. TIG. 16:276–277.

572   Risterucci AM et al. 2000. A high-density linkage map of <Emphasis Type="Italic">Theobroma
573   cacao </Emphasis>L. Theor. Appl. Genet. 101:948–955. doi: 10.1007/s001220051566.

574   Rouard M et al. 2011. GreenPhylDB v2.0: comparative and functional genomics in plants.
575   Nucleic Acids Res. 39:D1095-1102. doi: 10.1093/nar/gkq811.

576   Ruas M et al. 2017. MGIS: managing banana (Musa spp.) genetic resources information and
577   high-throughput genotyping data. Database. 2017. doi: 10.1093/database/bax046.

578   Sarah G et al. 2016. A large set of 26 new reference transcriptomes dedicated to comparative
579   population genomics in crops and wild relatives. Mol. Ecol. Resour. doi: 10.1111/1755-
580   0998.12587.

581   Sardos Julie et al. 2016. A Genome-Wide Association Study on the Seedless Phenotype in
582   Banana (*Musa* spp.) Reveals the Potential of a Selected Panel to Detect Candidate Genes in a
583   Vegetatively Propagated Crop. PLOS ONE. 11:e0154448. doi: 10.1371/journal.pone.0154448.

584   Sardos J. et al. 2016. DArT whole genome profiling provides insights on the evolution and
585   taxonomy of edible Banana (Musa spp.). Ann. Bot. mcw170. doi: 10.1093/aob/mcw170.

586   Scornavacca C, Berry V, Lefort V, Douzery EJ, Ranwez V. 2008. PhySIC_IST: cleaning source
587   trees to infer more informative supertrees. BMC Bioinformatics. 9:413. doi: 10.1186/1471-2105-
588   9-413.

589   Scornavacca C, Berry V, Ranwez V. 2011. Building species trees from larger parts of
590   phylogenomic databases. Inf. Comput. 209:590–605. doi: 10.1016/j.ic.2010.11.022.

591   Shi C-M, Yang Z. 2017. Coalescent-based analyses of genomic sequence data provide a robust
592   resolution of phylogenetic relationships among major groups of gibbons. Mol. Biol. Evol. doi:
593   10.1093/molbev/msx277.

594   Shi C-M, Yang Z. 2018. Coalescent-Based Analyses of Genomic Sequence Data Provide a
595   Robust Resolution of Phylogenetic Relationships among Major Groups of Gibbons. Mol. Biol.
596   Evol. 35:159–179. doi: 10.1093/molbev/msx277.

597   Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO:
598   assessing genome assembly and annotation completeness with single-copy orthologs.
599   Bioinformatics. 31:3210–3212. doi: 10.1093/bioinformatics/btv351.

600   Simmonds NW. 1956. Botanical Results of the Banana Collecting Expedition, 1954-5. Kew
601   Bull. 11:463–489. doi: 10.2307/4109131.

602   Simmonds NW. 1962. *The evolution of the bananas*. Longmans: London (GBR).

603   Simmonds NW, Shepherd K. 1955. The taxonomy and origins of the cultivated bananas. J. Linn.
604   Soc. Lond. Bot. 55:302–312. doi: 10.1111/j.1095-8339.1955.tb00015.x.

605   Simmonds NW, Weatherup STC. 1990. Numerical taxonomy of the wild bananas (Musa). New
606   Phytol. 115:567–571. doi: 10.1111/j.1469-8137.1990.tb00485.x.

607   Tettelin H et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus
608   agalactiae: Implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A.
609   102:13950–13955. doi: 10.1073/pnas.0506758102.

610   Thomas DC et al. 2012. West to east dispersal and subsequent rapid diversification of the mega-
611   diverse genus Begonia (Begoniaceae) in the Malesian archipelago. J. Biogeogr. 39:98–113. doi:
612   10.1111/j.1365-2699.2011.02596.x.

613   Veeramah KR et al. 2015. Examining Phylogenetic Relationships Among Gibbon Genera Using
614   Whole Genome Sequence Data Using an Approximate Bayesian Computation Approach.
615   Genetics. 200:295–308. doi: 10.1534/genetics.115.174425.

616   Wu M, Kostyun JL, Hahn MW, Moyle L. 2017. Dissecting the basis of novel trait evolution in a
617   radiation with widespread phylogenetic discordance. bioRxiv. 201376. doi: 10.1101/201376.

618   Yachdav G et al. 2016. MSAViewer: interactive JavaScript visualization of multiple sequence
619   alignments. Bioinforma. Oxf. Engl. 32:3501–3503. doi: 10.1093/bioinformatics/btw474.

620   Zimin AV et al. 2013. The MaSuRCA genome assembler. Bioinformatics. 29:2669–2677. doi:
621   10.1093/bioinformatics/btt476.

622

A

B

PKW

C4

DH

B

M

30

20

25
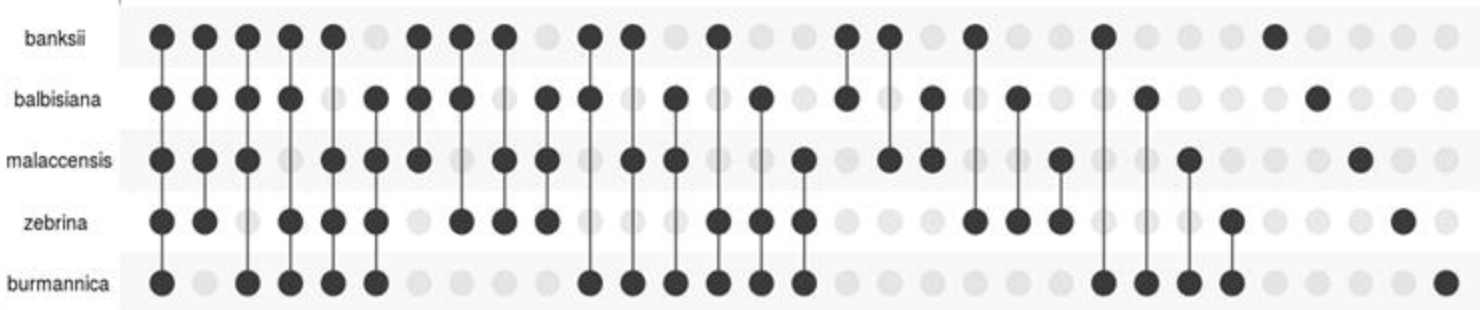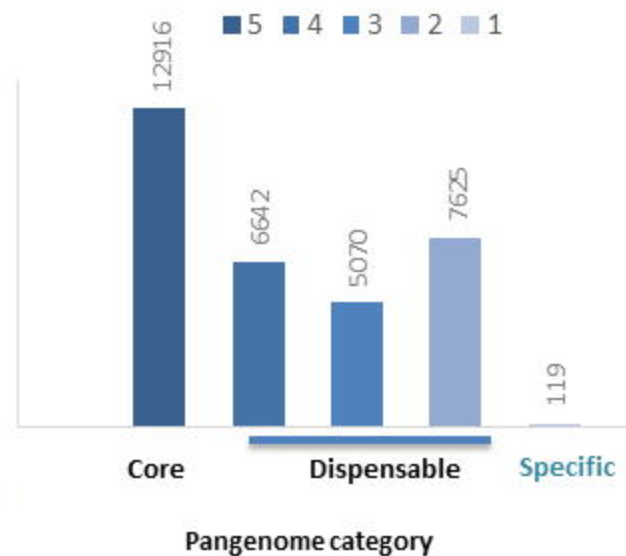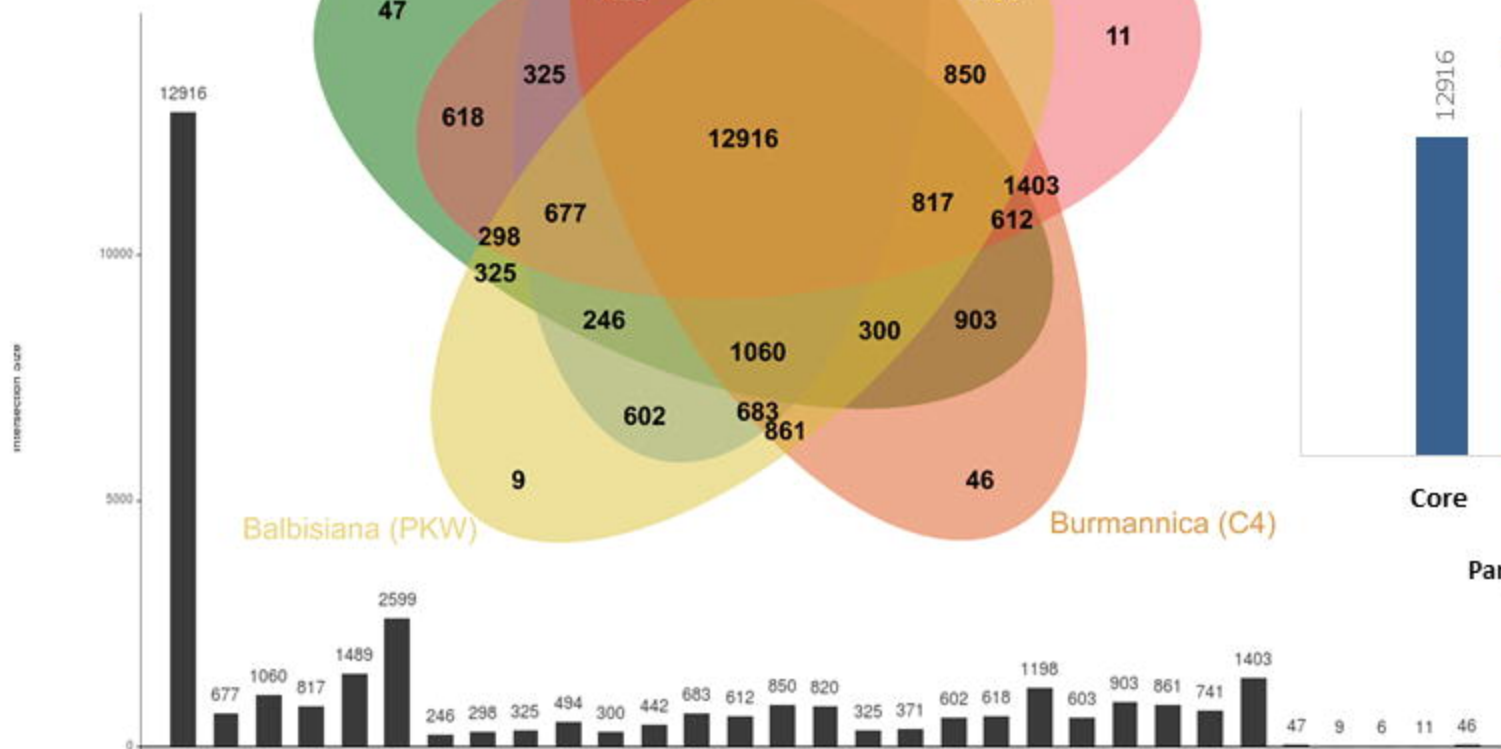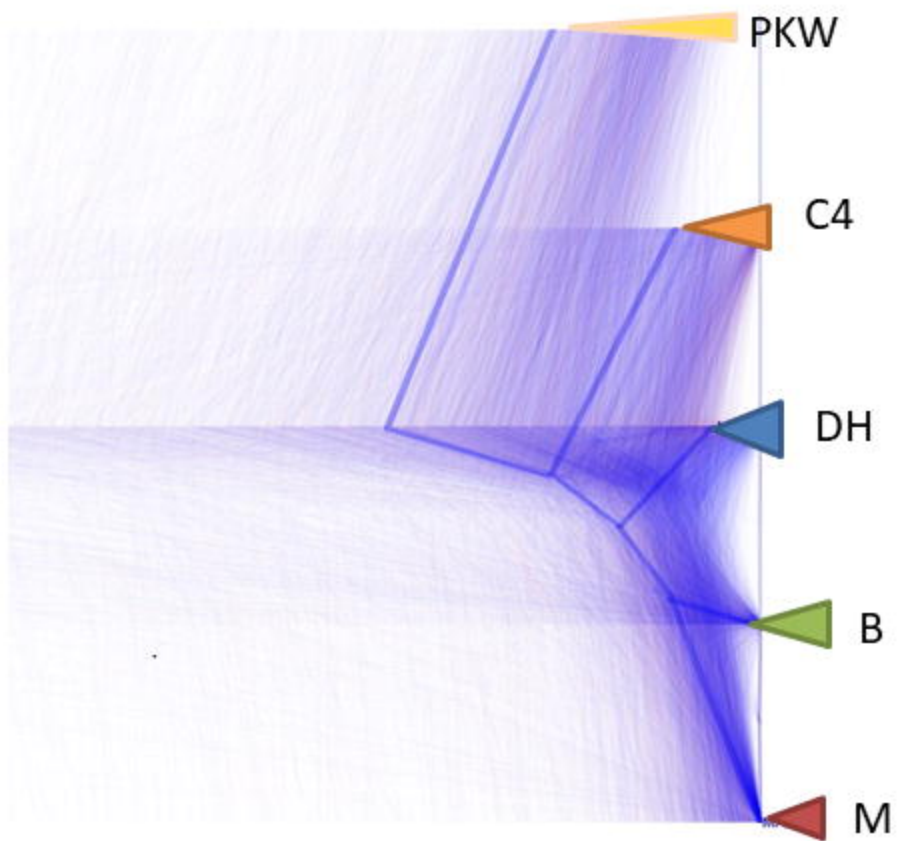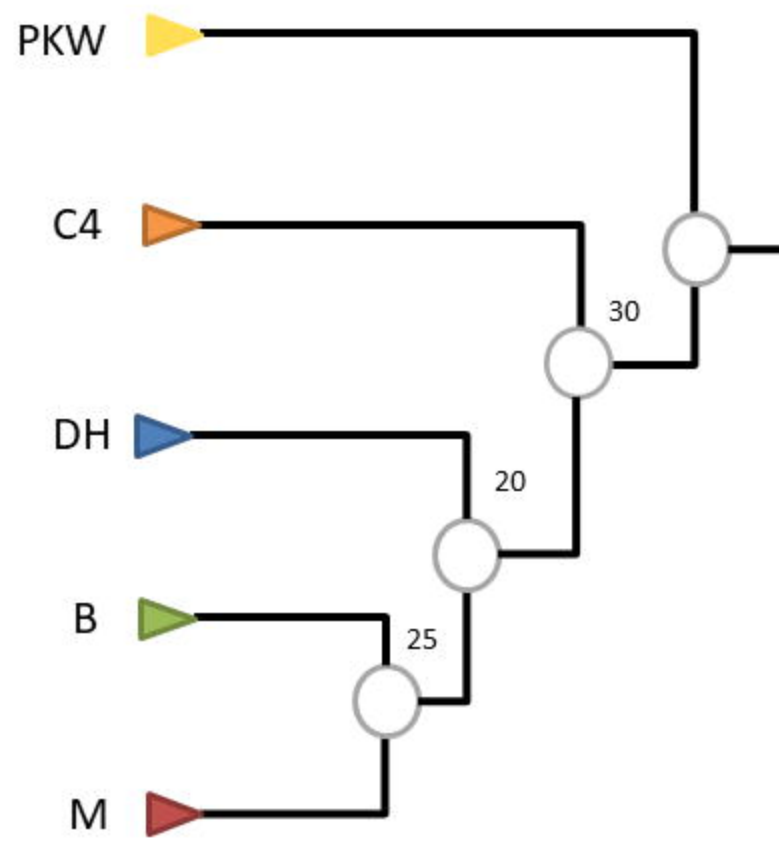
M. balbisiana 'Pkw'   *M. a. zebrina* 'Maia Oa'   *M. a. malaccensis* 'DH Pahang'   *M. a. burmannica* 'Calcutta 4'   *M. a. banksii* 'Banksii'

A Concatenation based
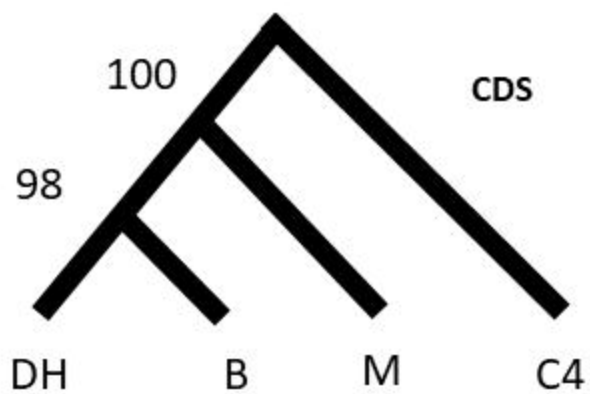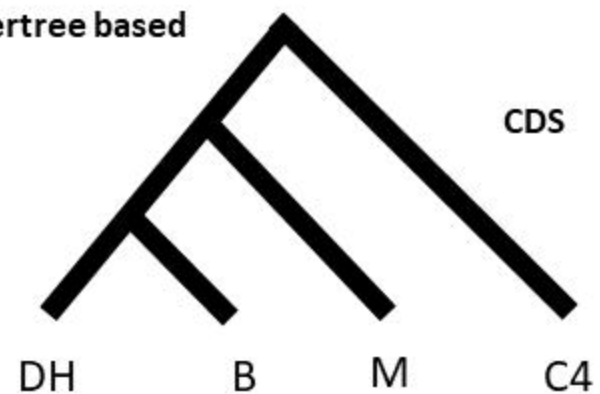
100

98

CDS

DH   B   M   C4

B Supertree based

CDS
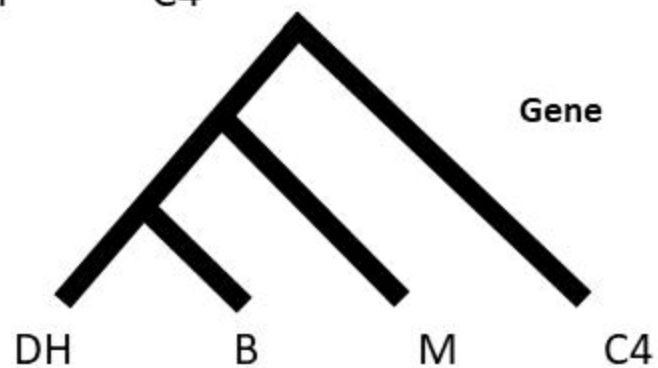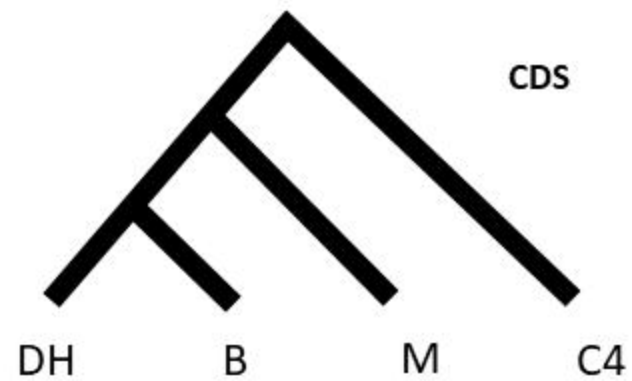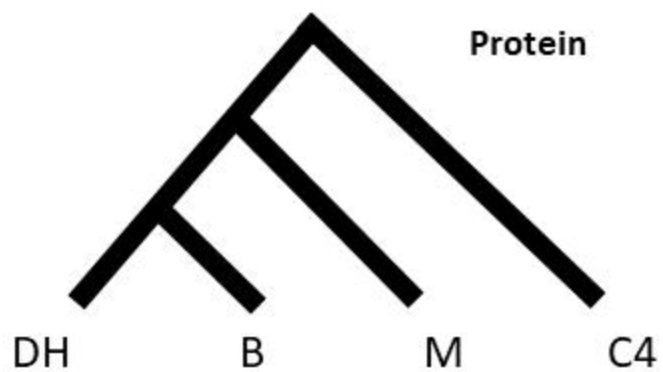
DH   B   M   C4

C Quartet based

Protein

DH   B   M   C4

CDS

DH   B   M   C4

Gene

DH   B   M   C4

A

PanMusa is a web resource designed for comparative and functional genomics in the Musaceae that includes Bananas (Musa spp.). The database contains a catalogue of gene families based on protein coding genes from representative of the 4 ancestral pools of M. acuminata (i.e. 'DH Pahang', 'Banksii', 'Maia Oa' and 'Calcutta 4') and M. balbisiana (i.e. 'Pisang Klutuk Wulung' abbrv. PKW).

Information

- PanMusa version launched

Musaceae — Musa — Eumusa
- Musa balbisiana Pisang Klutuk Wulung
- Musa acuminata malaccensis DH Pahang
- Musa acuminata Banksii Banksii
- Musa acuminata burmanica Calcutta 4
- Musa acuminata zebrina Maia Oa

B
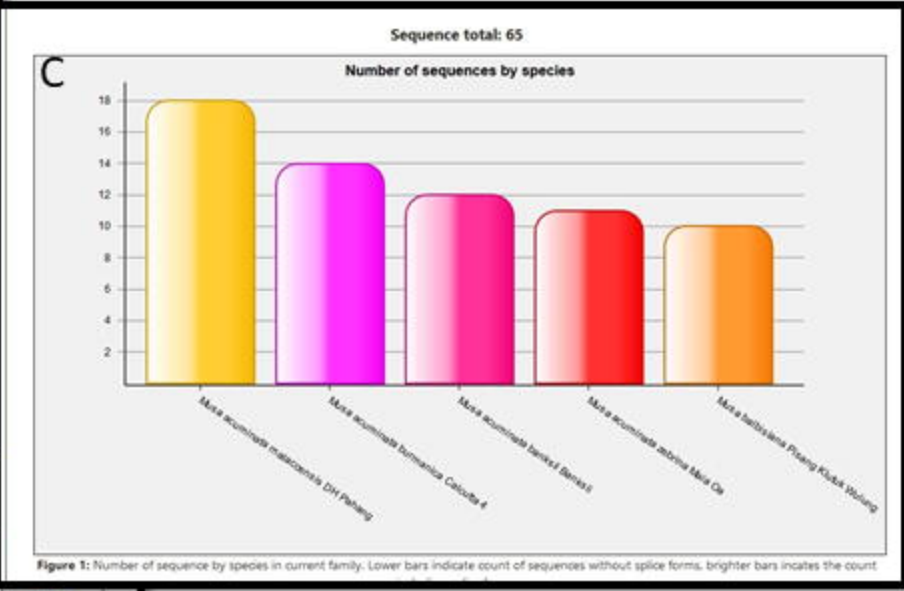
Number of clusters: 32372

	Family Id	Family Name	Number of sequences	Status
	OG0000000	Unannotated cluster	134	🟢
	OG0000001	Actin and Actin-like family	104	🟢
	OG0000006	Unannotated cluster	68	🟢
	OG0000007	ATPase transporter family		🟢
	OG0000008	Tubulin family		
	OG0000012	Unannotated cluster		

C

Sequence total: 65

Number of sequences by species

Figure 1: Number of sequence by species in current family. Lower bars indicate count of sequences without splice forms, brighter bars incates the count

D

Representative InterPro domains - Consensus schema

IPR008250 IPR001757 IPR018303 IPR023299 IPR006534 IPR004014 IPR023298 IPR023214

E

- dynamic node hiding
- invert colors
- show phylogram
- lineup nodes
- show branch length values decimals 3
- show support values decimals 1
- show node names for all nodes
- show node labels
- show additional node information
- show taxonomy
- taxonomy colorization show taxonomy colors table
- node size 12px
- X Y RESET X Y
- show domain architecture
- show domain names
- domain colorization domain scale 100
- P value 1

F

Show/hide alignment

Download alignment file (filtered): [FASTA]

M. balbisiana (PKW)

M. a. burmannica (C4)

M. a. malaccensis (DH)

M. a. zebrina (M)

M. a. banksii (B)

M. acuminata

C4   DH   B   M   PKW

introgression