Three New Genome Assemblies Support a Rapid Radiation in *Musa acuminata* (Wild Banana)

Mathieu Rouard[1,*], Gaetan Droc[2,3], Guillaume Martin[2,3], Julie Sardos[1], Yann Hueber[1], Valentin Guignon[1], Alberto Cenci[1], Björn Geigle[4], Mark S. Hibbins[5,6], Nabila Yahiaoui[2,3], Franc-Christophe Baurens[2,3], Vincent Berry[7], Matthew W. Hahn[5,6], Angelique D'Hont[2,3], and Nicolas Roux[1]

[1]Bioversity International, Parc Scientifique Agropolis II, Montpellier, France

[2]CIRAD, UMR AGAP, Montpellier, France

[3]AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, France

[4]Computomics GmbH, Tuebingen, Germany

[5]Department of Biology, Indiana University

[6]Department of Computer Science, Indiana University

[7]LIRMM, Université de Montpellier, CNRS, Montpellier, France

*Corresponding author: E-mail: m.rouard@cgiar.org.

Accepted: October 10, 2018

Data deposition: Raw sequence reads for de novo assemblies were deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) (BioProject: PRJNA437930 and SRA: SRP140622). Genome Assemblies and gene annotation data are available on the Banana Genome Hub (Droc G, Lariviere D, Guignon V, Yahiaoui N, This D, Garsmeur O, Dereeper A, Hamelin C, Argout X, Dufayard J-F, Lengelle J, Baurens F–C, Cenci A, Pitollat B, D'Hont A, Ruiz M, Rouard M, Bocs S. The Banana Genome Hub. Database (2013) doi:10.1093/database/bat035) (http://banana-genome-hub.southgreen.fr/species-list). Cluster and gene tree results are available on a dedicated database (http://panmusa.greenphyl.org) hosted on the South Green Bioinformatics Platform (Guignon et al. 2016). Additional data sets are made available on Dataverse: https://doi.org/10.7910/DVN/IFI1QU.

Abstract

Edible bananas result from interspecific hybridization between *Musa acuminata* and *Musa balbisiana*, as well as among subspecies in *M. acuminata*. Four particular *M. acuminata* subspecies have been proposed as the main contributors of edible bananas, all of which radiated in a short period of time in southeastern Asia. Clarifying the evolution of these lineages at a whole-genome scale is therefore an important step toward understanding the domestication and diversification of this crop. This study reports the de novo genome assembly and gene annotation of a representative genotype from three different subspecies of *M. acuminata*. These data are combined with the previously published genome of the fourth subspecies to investigate phylogenetic relationships. Analyses of shared and unique gene families reveal that the four subspecies are quite homogenous, with a core genome representing at least 50% of all genes and very few *M. acuminata* species-specific gene families. Multiple alignments indicate high sequence identity between homologous single copy-genes, supporting the close relationships of these lineages. Interestingly, phylogenomic analyses demonstrate high levels of gene tree discordance, due to both incomplete lineage sorting and introgression. This pattern suggests rapid radiation within *Musa acuminata* subspecies that occurred after the divergence with *M. balbisiana*. Introgression between *M. a.* ssp. *malaccensis* and *M. a.* ssp. *burmannica* was detected across the genome, though multiple approaches to resolve the subspecies tree converged on the same topology. To support evolutionary and functional analyses, we introduce the PanMusa database, which enables researchers to exploration of individual gene families and trees.

Key words: banana, *Musa* ssp., incomplete lineage sorting, phylogenomics, genome assembly.

Introduction

Bananas are among the most important staple crops cultivated worldwide in both the tropics and subtropics. The wild ancestors of bananas are native to the Malesian Region (including Malaysia and Indonesia) (Simmonds 1962) or to northern Indo-Burma (southwest China). Dating back to the

early Eocene (Janssens et al. 2016), the genus *Musa* currently comprises 60–70 species divided into two sections, Musa and Callimusa (Häkkinen 2013). Most of modern cultivated bananas originated from natural hybridization between two species from the section Musa, *Musa acuminata*, which occurs throughout the whole southeast Asia region, and *Musa balbisiana*, which is constrained to an area going from east India to south China (Simmonds and Shepherd 1955). While no subspecies have been defined so far in *M. balbisiana*, *M. acuminata* is further divided into multiple subspecies, among which at least four have been identified as contributors to the cultivated banana varieties, namely *banksii*, *zebrina*, *burmannica*, and *malaccensis* (reviewed in Perrier et al. 2011). These subspecies can be found in geographical areas that are mostly nonoverlapping. *Musa acuminata* ssp. *banksii* is endemic to New Guinea. *Musa a.* ssp. *zebrina* is found in Indonesia (Java island), *M. a.* ssp. *malaccensis* originally came from the Malay Peninsula (De Langhe et al. 2009; Perrier et al. 2011), while *M. a.* ssp. *burmannica* is from Burma (today's Myanmar) (Cheesman 1948).

While there are many morphological characters that differentiate *M. acuminata* from *M. balbisiana*, the subspecies of *M. acuminata* have only a few morphological differences between them. For instance, *M. a.* ssp. *burmannica* is distinguished by its yellowish and waxless foliage, light brown markings on the pseudostem, and by its compact pendulous bunch and strongly imbricated purple bracts. *Musa a.* ssp. *banksii* exhibits slightly waxy leaf, predominantly brownblackish pseudostems, large bunches with splayed fruits, and nonimbricated yellow bracts. *Musa a.* ssp. *malaccensis* is strongly waxy with a horizontal bunch, and bright red nonimbricated bracts, while *M. a.* ssp. *zebrina* is characterized by dark red patches on its dark green leaves (Simmonds 1956).

Previous studies based on a limited number of markers have been able to shed some light on the relationships among *M. acuminata* subspecies (Sardos et al. 2016; Christelová et al. 2017). Phylogenetic studies have been assisted by the availability of the reference genome sequence for a representative of *M. acuminata* ssp. *malaccensis* (D'Hont et al. 2012; Martin et al. 2016) and a draft *M. balbisiana* genome sequence (Davey et al. 2013). However, the availability of large genomic data sets from multiple (sub)species are expected to improve the resolution of phylogenetic analyses, and thus to provide additional insights on species evolution and their specific traits (Bravo et al. 2018). This is especially true in groups where different segments of the genome have different evolutionary histories, as has been found in *Musaceae* (Christelová et al. 2011). Whole-genome analyses also make it much easier to distinguish among the possible causes of gene tree heterogeneity, especially incomplete lineage sorting (ILS) and hybridization (Folk et al. 2018).

Moreover, the availability of multiple reference genome sequences opens the way to so-called pangenome analyses, a concept coined by Tettelin et al. (2005). The pangenome is defined as the set of all gene families found among a set of phylogenetic lineages. It includes 1) the core genome, which is the pool of genes common to all lineages, 2) the accessory genome, composed of genes absent in some lineages, and 3) the species-specific or individual-specific genome, formed by genes that are present in only a single lineage. Identifying specific compartments of the pangenome (such as the accessory genome) offers a way to detect important genetic differences that underlie molecular diversity and phenotypic variation (Morgante et al. 2007).

Here, we generated three de novo genomes for the subspecies *banksii*, *zebrina* and *burmannica*, and combined these with existing genomes for *M. acuminata* ssp. *malaccensis* (D'Hont et al. 2012) and *M. balbisiana* (Davey et al. 2013). We thus analyzed the whole genome sequences of five extant genotypes comprising the four cultivated bananas' contributors from *M. acuminata*, that is, the reference genome "DH Pahang" belonging to *M. acuminata* ssp. *malaccensis*, "Banksii" from *M. acuminata* ssp. *banksii*, "Maia Oa" belonging to *M. acuminata* ssp. *zebrina*, and "Calcutta 4" from *M. acuminata* ssp. *burmannica*, as well as *M. balbisiana* (i.e., "Pisang Klutuk Wulung" or PKW). We carried out phylogenomic analyses that provided evolutionary insights into both the relationships and genomic changes among lineages in this clade. Finally, we developed a banana species-specific database to support the larger community interested in crop improvement.

Materials and Methods

Plant Material

Banana leaf samples from accessions "Banksii" (*Musa acuminata* ssp. *banksii*, PT-BA-00024), "Maia Oa" (*Musa acuminata* ssp. *zebrina*, PT-BA-00182), and "Calcutta 4" (*Musa acuminata* ssp. *burmannica*, PT-BA-00051) were supplied by the CRB-Plantes Tropicales Antilles CIRAD-INRA field collection based in Guadeloupe. Leaves were used for DNA extraction. Plant identity was verified at the subspecies level using SSR markers at the *Musa* Genotyping Centre (MGC, Czech Republic) as described in (Christelová et al. 2011) and passport data of the plant is accessible in the Musa Germplasm Information System (Ruas et al. 2017). In addition, the representativeness of the genotypes of the four subspecies was verified on a set of 22 samples belonging to the same four *M. acuminata* subspecies of the study (supplementary fig. 3, Supplementary Material online).

Sequencing and Assembly

Genomic DNA was extracted using a modified MATAB method (Risterucci et al. 2000). DNA libraries were constructed and sequenced using the HiSeq2000 (Illumina) technology at BGI (supplementary table 1, Supplementary Material online). "Banksii" was assembled using

SoapDenovo (Luo et al. 2012), and PBJelly2 (English et al. 2012) was used for gap closing using PacBio data generated at the Norwegian Sequencing Center (NSC) with Pacific Biosciences RS II. "Maia Oa" and "Calcutta 4" were assembled using the MaSuRCA assembler (Zimin et al. 2013) (supplementary table 2, Supplementary Material online). Estimation of genome assembly completeness was assessed with BUSCO plant (Simão et al. 2015) (supplementary table 3, Supplementary Material online).

Gene Annotation

Gene annotation was performed on the obtained de novo assembly for "Banksii," "Maia Oa," and "Calcutta 4," as well as on the draft *Musa balbisiana* "PKW" assembly (Davey et al. 2013) for consistency and because the published annotation was assessed as low quality. For structural annotation we used EuGene v4.2 (http://eugene.toulouse.inra.fr/) (Foissac et al. 2008) calibrated on *M. acuminata malaccensis* "DH Pahang" reference genome v2, which produced similar results (e.g., number of genes, no missed loci, good specificity, and sensitivity) as the official annotation (Martin et al. 2016). EuGene combined genotype-specific (or closely related) transcriptome assemblies, performed with Trinity v2.4 with RNAseq data sets (Sarah et al. 2017), to maximize the likelihood to have genotype-specific gene annotation (supplementary table 4, Supplementary Material online**)**. The estimation of gene space completeness was assessed with Busco (supplementary table 3, Supplementary Material online). Because of its high quality and to avoid confusing the community, we did not perform a new annotation for the *M. a. malaccensis* "DH Pahang" reference genome but used the released version 2. Finally, the functional annotation of plant genomes was performed by assigning their associated generic GO terms through the Blast2GO program (Conesa et al. 2005) combining BLAST results from UniProt (E-value 1e-5) (Magrane and UniProt Consortium 2011).

Gene Families

Gene families were identified using OrthoFinder v1.1.4 (Emms and 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 (Janssens et al. 2016; Sardos et al. 2016). Those two studies included all *M. acuminata* subspecies, and had the same tree topology (supplementary fig. 1, Supplementary Material online). In the first study, Sardos et al. (2016) computed a Neighbor-Joining tree from a dissimilarity matrix using biallelic 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 (supplementary fig. 2, Supplementary Material online). 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 and 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 and Zdobnov 2010). 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 and Warnow 2015; Zhang et al. 2018). 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 multiple copies), and accessory genes for which only one representative species was missing (except outgroup species). Finally, single-copy OGs (CDS only) were used to generate a concatenated genome-scale alignment with FASconCAT-G (Kück and Longo 2014) and a tree was constructed using PhyML (NNI, HKY85, 100 bootstrap).

Search for Introgression

Ancient gene flow was assessed with the ABBA-BABA test or *D*-statistic (Green et al. 2010; Durand et al. 2011) and computed on the concatenated multiple alignment converted to the MVF format and processed with MVFtools (Pease and Rosenzweig 2018), similar to what is described in Wu et al. (2017) (https://github.com/wum5/JaltPhylo). The direction of introgression was further assessed with the D_2 test (Hibbins and Hahn 2018). The D_2 statistic captures differences in the heights of genealogies produced by introgression occurring in alternate directions by measuring the average divergence between species A and C in gene trees with an ((A, B), C) topology (denoted $[d_{AC}|A, B]$), and subtracting the average A–C divergence in gene 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 statistic

is significantly positive, it means that introgression has either occurred in the B→C direction or in both directions. D_2 significance was assessed by permuting labels on gene trees 1,000 times and calculating p values from the resulting null distribution of D_2 values. The test was implemented with a Perl script using distmat from EMBOSS (Rice et al. 2000) with Tajima–Nei distance applied to multiple alignments associated with gene trees fitting the defined topologies above (https://github.com/mrouard/perl-script-utils).

Results

Assemblies and Gene Annotation

We generated three de novo assemblies belonging to *M. acuminata* ssp. *banksii*, *M. a.* ssp. *zebrina*, and *M. a.* ssp. *burmannica*. The *M. a.* ssp. *zebrina* and *M. a.* ssp. *burmannica* assemblies contained 56,481 and 47,753 scaffolds (N50 scaffold of 37,689 and 22,183 bp) totaling 623 Mb and 526 Mb, respectively. The *M. a.* ssp. *banksii* assembly, which benefited from long-read sequencing, contained 9,467 scaffolds (N50 scaffold of 435,833 bp) for a total of 464 Mb (78.2% of the genome) (supplementary tables 1 and 2, Supplementary Material online).

The number of predicted protein coding genes per genome within different genomes of *Musa* ranges from 32,692 to 45,069 (supplementary table 3, Supplementary Material online). 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 (supplementary table 4, Supplementary Material online), 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 (supplementary table 5, Supplementary Material online). 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 (fig. 1). Among these, 8,030 are composed of only one sequence in each lineage (i.e., are likely single-copy orthologs). A set of 1,489 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.

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 (fig. 2*A* and supplementary data 1, Supplementary Material online). 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 topologies (table 2). We also examined topologies at loci that had bootstrap support >90 for all nodes, also finding all 15 different topologies (table 2). Among trees constructed from whole genes, topologies ranged in frequency from 13.12% for the most common tree to 1.92% for the least common tree (table 2) with an average length of the 1,342 aligned nucleotide sites for CDS and 483 aligned sites for proteins. Based on these results, gene tree frequencies were used to calculate concordance factors on the most frequent CDS gene trees (table 2), demonstrating that no split was supported by >30% of gene trees (fig. 2*B*). Therefore, in order to further gain insight into the subspecies phylogeny, we used a combination of different approaches described in the next section.

Inference of a Species Tree

We used three complementary methods to infer phylogenetic relationships among the sampled lineages. First, we concatenated nucleotide sequences from all single-copy genes (totaling 11,668,507 bp). We used PHYML to compute a maximum likelihood tree from this alignment, which, as expected, provided a topology with highly supported nodes (fig. 3*A*). Note that this topology (denoted topology number 1 in table 2) is not the same as the one previously proposed in the literature (denoted topology number 7 in table 2) (supplementary figs. 1 and 2, Supplementary Material online).

Next, we used a method explicitly based on individual gene tree topologies. ASTRAL (Mirarab and Warnow 2015) infers the species tree by using quartet frequencies found in gene trees. It is suitable for large data sets and was highlighted as one of the best methods to address challenging topologies with short internal branches and high levels of discordance (Shi and Yang 2018). ASTRAL found the same topology using ML gene trees from single-copy genes obtained from protein sequences, CDSs, and genes (fig. 3*C*).

Finally, we ran a supertree approach implemented in PhySIC_IST (Scornavacca et al. 2008) on the single-copy genes and obtained again the same topology (fig. 3*B*). PhySIC_IST first collapses poorly supported branches of the gene trees into polytomies, as well as conflicting branches of the gene

Table 1
Summary of the Gene Clustering Statistics Per (Sub)Species

	Musa 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

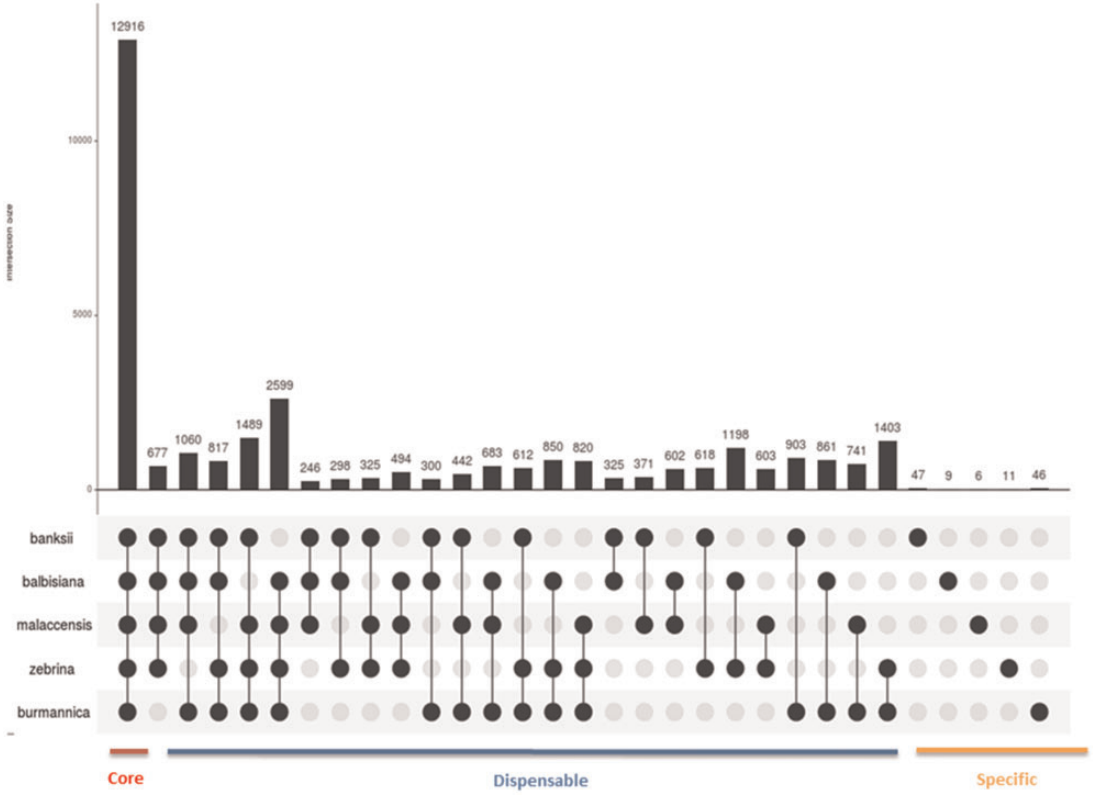

Fig. 1.—Intersection diagram showing the distribution of shared gene families (at least two sequences per OG) among *M. a. banksii* "Banksii," *M. a. zebrina* "Maia Oa," *M. a. burmannica* "Calcutta 4," *M. a. malaccensis* "DH Pahang," and *M. balbisiana* "PKW" genomes. The figure was created with UpsetR (Lex et al. 2014).

trees that are only present in a small minority of the trees; it then searches for the most resolved supertree that does not contradict the signal present in the gene trees nor contains topological signal absent from those trees. Deeper investigation of the results revealed that ~ 66% of the trees were unresolved, 33% discarded (pruned or incorrectly rooted), and therefore that the inference relied on fewer than 1% of the trees. Aiming to increase the number of genes used by PhySIC_IST, we included multicopy OGs of the core genome, as well as some OGs in the accessory genomes using the pipeline SSIMUL (Scornavacca et al. 2011). SSIMUL translates multilabeled gene trees (MUL-trees) into trees having a

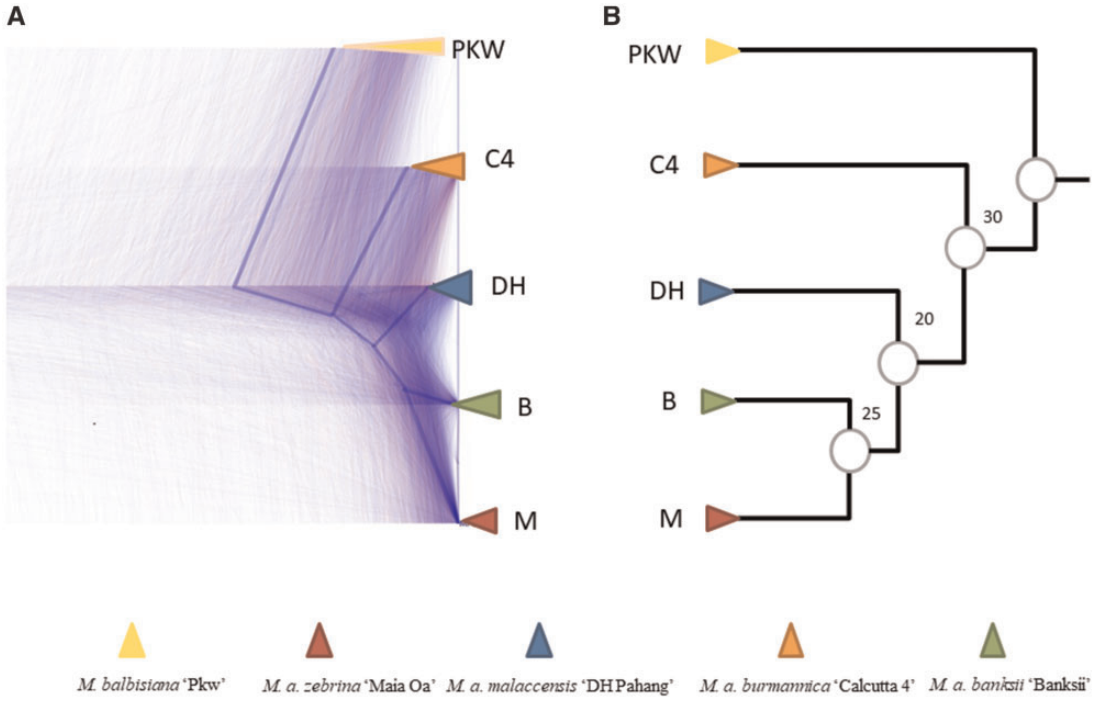

Fig. 2.—Illustration of gene tree discordance. (*A*) Cloudogram of single copy OGs (CDS) visualized with Densitree. The blue line represents the consensus tree as provided by Densitree. (*B*) Species tree with bootstrap-like support based on corresponding gene tree frequency from table 2 (denoted topology number 2). PKW, *M. balbisiana* "PKW"; C4, *M. acuminata burmannica* "Calcutta 4"; M, *M. acuminata zebrina* "Maia Oa"; DH, *M. acuminata malaccensis* "DH Pahang"; B, *M. acuminata banksii* "Banksii".

Table 2

Frequency of Gene Tree Topologies of the 8,030 Single Copy OGs

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

Note.—In bold, the most frequent topology.

PKW, *Musa balbisiana* "PKW"; C4, *Musa acuminata burmannica* "Calcutta 4"; M, *Musa acuminata zebrina* "Maia Oa"; DH, *Musa acuminata malaccensis* "DH Pahang"; B, *Musa acuminata banksii* "Banksii".

single copy of each gene (X-trees), that is, the type of tree usually expected in supertree inference. To do so, all individual gene trees were constructed on CDSs from OGs with at least 4 *M. acuminata* and *M. balbisiana* genes (n = 18,069). SSIMUL first removed identical subtrees resulting from a duplication node in these trees, it then filtered out trees where duplicated parts induced contradictory rooted triples, keeping only coherent trees. These trees can then be turned into trees containing a single copy of each gene, either by pruning the smallest subtrees under each duplication node (leaving only

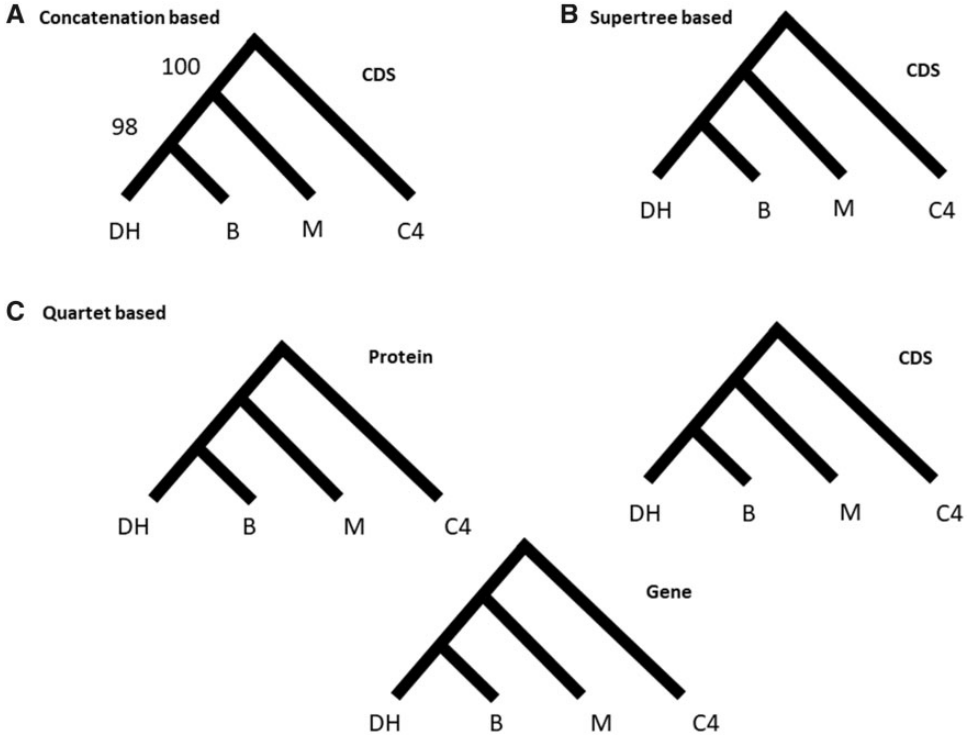

Fig. 3.—Species topologies computed with three different approaches. (*A*) Maximum likelihood tree inferred from a concatenated alignment of single-copy genes (CDS). (*B*) Supertree-based method applied to single and multilabelled gene trees. (*C*) Quartet-based model applied to protein, CDS, and gene alignments.

orthologous nodes in the tree), or by extracting the topological signal induced by orthology nodes into a rooted triplet set, that is then turned back into an equivalent X-tree. Here, we chose to use the pruning method to generate a data set to be further analyzed with PhySIC_IST, which lead to a subset of 14,507 gene trees representing 44% of the total number of OGs and an increase of 80% compared with the 8,030 single-copy OGs. This analysis returned a consensus gene tree with the same topology as both of the previous methods used here (fig. 3*B*).

Evidence for Introgression

Although much of the discordance we observe is likely due to incomplete lineage sorting, we also searched for introgression between subspecies. A common approach, performed in other plant genomes (Eaton and Ree 2013; Eaton et al. 2015; Novikova et al. 2016; Choi et al. 2017), relies on the use of the ABBA-BABA test (or D statistics) (Green et al. 2010). This test allows to differentiate admixture from incomplete lineage sorting across genomes by detecting an excess of either ABBA or BABA sites (where "A" corresponds to the ancestral allele and "B" corresponds to the derived allele state). An excess of each of this patterns is indicative of ancient admixture. Therefore, we applied it in a four-taxon phylogeny including three *M. acuminata* subspecies as ingroups and *M. balbisiana* as outgroup. Because there were five taxa to be tested, analyses were done with permutation of taxa denoted P1, P2, and P3 and Outgroup (table 3). Under the null hypothesis of ILS, an equal number of ABBA and BABA sites are expected. However, we always found an excess of sites grouping *malaccensis* ("DH") and *burmannica* ("C4") (table 3). This indicates a history of introgression between these two lineages.

To test the direction of introgression, we applied the D_2 test (Hibbins and Hahn 2018). While introgression between a pair of species (e.g., *malaccensis* and *burmannica*) always results in smaller genetic distances between them, the D_2 test is based on the idea that gene flow in the two alternative directions can also result in a change in genetic distance to other taxa not involved in the exchange (in this case, *banksii*). We computed the genetic distance between *banksii* and *burmannica* in gene trees where *malaccensis* and *banksii* are sister (denoted $d_{AC}|A, B$) and the genetic distance between *banksii* and *burmannica* in gene trees where *malaccensis* and *burmannica* are sister (denoted $d_{AC}|B, C$). The test takes into account the genetic distance between the species not involved in the introgression (*banksii*) and the species involved in introgression that it is not most closely related to (*burmannica*). We identified 1,454 and 281 gene trees with $d_{AC}|A, B = 1.15$ and $d_{AC}|B, C = 0.91$, respectively, giving a significant positive value of D_2=0.23 ($p < 0.001$ by permutation). These

Table 3
Four-Taxon ABBA-BABA Test (*D*-Statistic) Used for Introgression Inference from the Well-Supported Topology from Fig. 3

P1	P2	P3	BBAA	ABBA	BABA	Disc [a]	*D*[b]	*p* value[c]
Malaccensis (DH)	Banksii (B)	Burmannica (C4)	12185	4289	8532	0.51	−0.33	<2.2e-16
Malaccensis (DH)	Zebrina (M)	Burmannica (C4)	9622	5400	9241	0.6	−0.26	< 2.2e-16
Zebrina (M)	Banksii (B)	Burmannica (C4)	11204	6859	6782	0.54	0.005	0.5097
Malaccensis (DH)	Banksii (B)	Zebrina (M)	10450	7119	6965	0.57	0.02	0.1944

[a]Discordance=(ABBA+BABA)/Total
[b]*D* =(ABBA−BABA)/(ABBA+BABA)
[c]Based on Pearson chi-squared.

results support introgression from *malaccensis* into *burmannica*, though they do not exclude the presence of a lesser level of gene flow in the other direction.

PanMusa, a Database to Explore Individual OGs

Since genes underlie traits and wild banana species showed a high level of incongruent gene tree topologies, access to a repertoire of individual gene trees is important. This was the rationale for constructing a database that provides access to gene families and individual gene family trees in *M. acuminata* and *M. balbisiana*. A set of web interfaces are available to navigate OGs that have been functionally annotated using GreenPhyl comparative genomics database (Rouard et al. 2011). PanMusa shares most of the features available on GreenPhyl to display or export sequences, InterPro assignments, sequence alignments, and gene trees (fig. 4). In addition, new visualization tools were implemented, such as MSAViewer (Yachdav et al. 2016) and PhyD3 (Kreft et al. 2017) to view gene trees.

Discussion

Musa acuminata Subspecies Contain Few Subspecies-Specific Families

In this study, we used a de novo approach to generate additional reference genomes for the three subspecies of *Musa acuminata*; all three are thought to have played significant roles as genetic contributors to the modern cultivars. Genome assemblies produced for this study differ in quality, but the estimation of genome assembly and gene annotation quality conducted with BUSCO suggests that they were sufficient to perform comparative analyses. Moreover, we observed that the number of genes grouped in OGs were relatively similar among subspecies, indicating that the potential overprediction of genes in "Maia Oa" and "Calcutta 4" was mitigated during the clustering procedure. Indeed, overprediction in draft genomes is expected due to fragmentation, leading to an artefactual increase in the number of genes (Denton et al. 2014).

Although our study is based on one representative per subspecies, *Musa* appears to have a widely shared pangenome, with only a small number of subspecies-specific families identified. The pangenome analysis also reveals a large number of families shared only among subsets of species or subspecies (fig. 1); this "dispensable" genome is thought to contribute to diversity and adaptation (Tettelin et al. 2005; Medini et al. 2005). The small number of species-specific OGs in *Musa acuminata* also supports the recent divergence between all genotypes including the split between *M. acuminata* and *M. balbisiana*.

Musa acuminata Subspecies Show a High Level of Discordance between Individual Gene Trees

Gene tree conflict has been recently reported in the Zingiberales (Carlsen et al. 2018) and Musa in not an exception. By computing gene trees with all single-copy genes OG, we found widespread discordance in gene tree topologies. Topological incongruence can be the result of incomplete lineage sorting, the misassignment of paralogs as orthologs, introgression, or horizontal gene transfer (Maddison 1997). With the continued generation of phylogenomic data sets over the past dozen years, massive amounts of discordance have been reported, first in *Drosophila* (Pollard et al. 2006) and more recently in birds (Jarvis et al. 2014), mammals (Li et al. 2016; Shi and Yang 2018), and plants (Novikova et al. 2016; Pease et al. 2016; Choi et al. 2017; Copetti et al. 2017; Wu et al. 2017). Due to the risk of hemiplasy in such data sets (Avise et al. 2008; Hahn and Nakhleh 2016), we determined that we could not accurately reconstruct either nucleotide substitutions or gene gains and losses among the genomes analyzed here.

In our case, the fact that all possible subspecies tree topologies occurred, and that ratios of minor trees at most nodes were equivalent to those expected under ILS, strongly suggests the presence of ILS (Hahn and Nakhleh 2016). Banana is a paleopolyploid plant that experienced three independent whole genome duplications (WGD), and some fractionation is likely still occurring (D'Hont et al. 2012) (supplementary table 6, Supplementary Material online). But divergence levels among the single-copy OGs were fairly consistent (fig. 2*A*), supporting the correct assignment of orthology among sequences. However, we did find evidence for introgression between *malaccensis* and *burmannica*, which contributed a

Fig. 4.—Overview of available interfaces for the PanMusa database. (*A*) Homepage of the website. (*B*) List of functionally annotated OGs. (*C*) Graphical representation of the number of sequence by species. (*D*) Consensus InterPro domain schema by OG. (*E*) Individual gene trees visualized with PhyD3. (*F*) Multiple alignment of OG with MSAviewer.

small excess of sites supporting one particular discordant topology (table 3). This event is also supported by the geographical overlap in the distribution of these two subspecies (Perrier et al. 2011).

Previous studies have attempted to resolve the topology in the Musaceae, but did not include all subspecies considered here, and had very limited numbers of loci. In Christelova et al. (2011), a robust combined approach using maximum likelihood, maximum parsimony, and Bayesian inference was applied to 19 loci, but only *burmannica* and *zebrina* out of the four subspecies were included. Jarret et al. (1992) reported sister relationships between *malaccensis* and *banksii* on the basis of RFLP markers, but did not include any samples from *burmannica* and *zebrina*. However, the resolved species tree supported by all methods used here is a new topology compared with species trees comprising at least one representative of our 4 subspecies (Janssens et al. 2016; Sardos et al. 2016; Christelová et al. 2017) (supplementary fig. 1, Supplementary Material online**)**. Indeed "Calcutta 4" as representative of *M. acuminata* ssp. *burmannica* was placed sister to the other *Musa acuminata* genotypes in our study, whereas those studies indicates direct proximity between *burmannica* and *malaccensis*. The detected introgression from *malaccensis* to *burmannica* may be an explanation for the difference observed but increasing the sampling with several genome sequences by subspecies would enable a better resolution.

More strikingly considering previous phylogenetic hypotheses, *malaccensis* appeared most closely related to *banksii*, which is quite distinct from the other *M. acuminata* spp. (Simmonds and Weatherup 1990) and which used to be postulated as its own species based on its geographical area of distribution and floral diversity (Argent 1976) (fig. 5). However, on the bases of genomic similarity, all our analyses support *M. acuminata* ssp. *banksii* as a subspecies of *M. acuminata*.

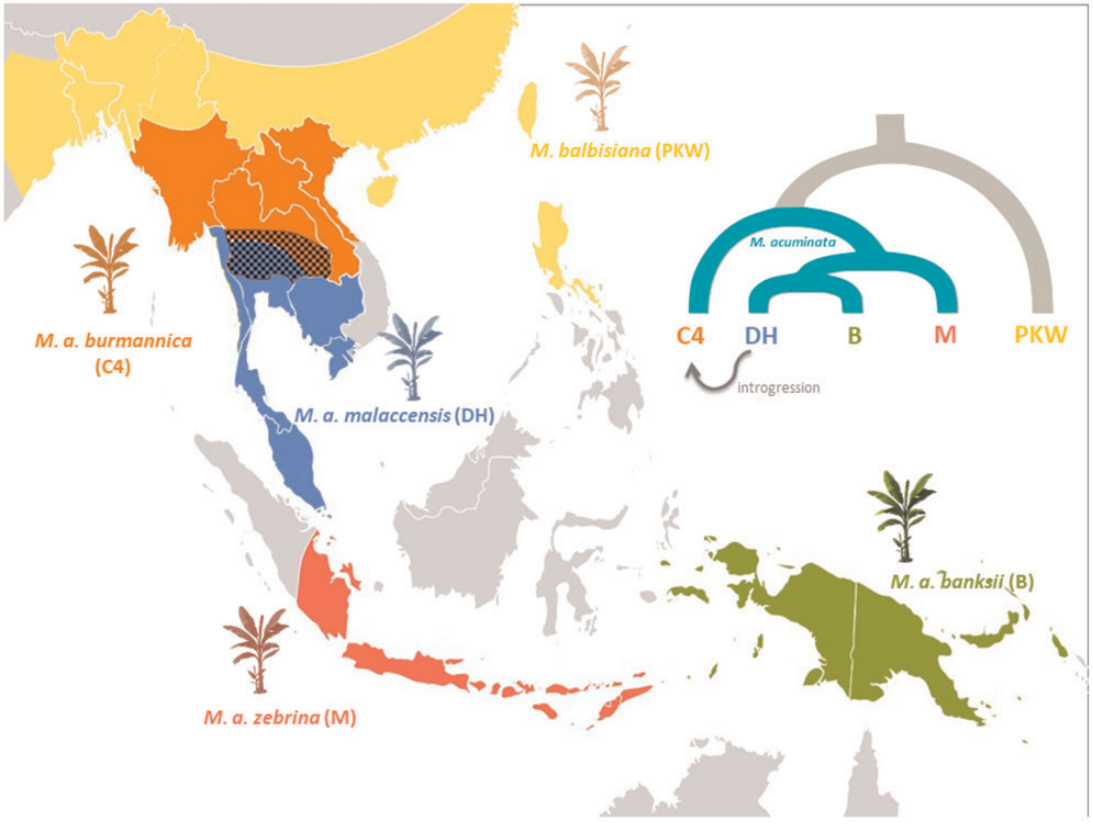

Fig. 5.—Area of distribution of Musa species in Southeast Asia as described by Perrier et al. (2011); including species tree of *Musa acuminata* subspecies based on results described in figure 4. Areas of distribution are approximately represented by colors; hatched zone shows area of overlap between two subspecies where introgression may have occurred.

Gene Tree Discordance Supports Rapid Radiation of *Musa acuminata* Subspecies

In their evolutionary history, *Musa* species dispersed from "northwest to southeast" into Southeast Asia (Janssens et al. 2016). Due to sea level fluctuations, Malesia (including the nations of Indonesia, Malaysia, Brunei, Singapore, the Philippines, and Papua New Guinea) is a complex geographic region, formed as the result of multiple fusions and subsequent isolation of different islands (Thomas et al. 2012; Janssens et al. 2016). Ancestors of the Callimusa section (of the *Musa* genus) started to radiate from the northern Indo-Burma region toward the rest of Southeast Asia ~30 Ma, while the ancestors of the Musa (formerly Eumusa/Rhodochlamys) section started to colonize the region ~10 Ma (Janssens et al. 2016). The divergence between *M. acuminata* and *M. balbisiana* has been estimated to be ~5 Ma (Lescot et al. 2008). However, no accurate dating has yet been proposed for the divergence of the *Musa acuminata* subspecies. We hypothesize that after the speciation of *M. acuminata* and *M. balbisiana* (ca. 5 Ma) rapid diversification occurred within populations of *M. acuminata*. This hypothesis is consistent with the observed gene tree discordance and high levels of ILS. Such a degree of discordance may reflect a near-instantaneous radiation between all subspecies of *M. acuminata*. Alternatively, it could support the proposed hypothesis of divergence back in the northern part of Malesia during the Pliocene (Janssens et al. 2016), followed by introgression taking place among multiple pairs of species as detected between *malaccensis* and *burmannica*. While massive amounts of introgression can certainly mask the history of lineage splitting (Fontaine et al. 2015), we did not find evidence for such mixing.

Interestingly, such a broad range of gene tree topologies due to ILS (and introgression) has also been observed in gibbons (Carbone et al. 2014; Veeramah et al. 2015; Shi and Yang 2018) for which the area of distribution in tropical forests of Southeast Asia is actually overlapping the center of origin of wild bananas. Moreover, according to Carbone et al. (2014), gibbons also experienced a near-instantaneous radiation ~5 Ma. It is therefore tempting to hypothesize that ancestors of wild bananas and ancestors of gibbons faced similar geographical isolation and had to colonize and adapt to similar ecological niches, leading to the observed patterns of incomplete lineage sorting.

In this study, we highlighted the phylogenetic complexity in a genome-wide data set for *Musa acuminata* and *Musa*

balbisiana, bringing additional insights to explain why the Musaceae phylogeny has remained controversial. Our work should enable researchers to make inferences about trait evolution, and ultimately should help support crop improvement strategies.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Noel Chen and Qiongzhi He (BGI) for providing sequencing services with Illumina and Ave Tooming-Klunderud (CEES) for PacBio sequencing services and Computomics for support with assembly. We thank Erika Sallet (INRA) for providing early access to the new version of Eugene with helpful suggestions. We thank the CRB-Plantes Tropicales Antilles CIRAD-INRA for providing plant materials. We would like also to acknowledge Jae Young Choi (NYU), Steven Janssens (MBG), Laura Kubatko (OSU) for helpful discussions and advice on species tree topologies. This work was financially supported by CGIAR Fund Donors and CGIAR Research Programme on Roots, Tubers and Bananas (RTB) and technically supported by the high performance cluster of the UMR AGAP – CIRAD of the South Green Bioinformatics Platform (http://www.southgreen.fr). Finally, this work benefited from the GenomeHarvest project (https://www.genomeharvest.fr/) funded by the Agropolis fondation.

Authors Contribution

M.R., N.R., and A.D. set up the study and M.R. coordinated the study. A.D. and F.C.B. provided access to plant material and DNA. N.Y. provided access to transcriptome data and G.M. to repeats library for gene annotation. B.G. performed assembly and gap closing. M.R., G.D., G.M., Y.H., J.S., and A.C. performed analyses. V.B., M.S.H., and M.W.H. provided guidance on methods and helped with result interpretation. V.G. and M.R. set up the PanMusa website. M.R. wrote the manuscript with significant contributions from M.W.H., V.B., and J.S., and all coauthors commented on the manuscript.

Literature Cited

Argent G. 1976. The wild bananas of Papua New Guinea. Notes Roy Bot Gard Edinb. 35:77–114.
Avise JC, Robinson TJ, Kubatko L. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. Syst Biol. 57(3):503–507.
Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. 2014. jvenn: an interactive Venn diagram viewer. BMC Bioinformatics 15(1):293.
Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26(10):1372–1373.
Bravo GA, et al. 2018. Embracing heterogeneity: Building the Tree of Life and the future of phylogenomics. PeerJ Preprints 6:e26449v3 https://doi.org/10.7287/peerj.preprints.26449v3
Carbone L, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. Nature 513(7517):195–201.
Carlsen MM, et al. 2018. Resolving the rapid plant radiation of early diverging lineages in the tropical Zingiberales: pushing the limits of genomic data. Mol Phylogenet Evol. 128:55–68.
Cheesman EE. 1948. Classification of the bananas. Kew Bull. 3(1):17–28.
Choi JY, et al. 2017. The rice paradox: multiple origins but single domestication in Asian rice. Mol Biol Evol. 34:969–979.
Christelová P, 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.
Christelová P, et al. 2011. A platform for efficient genotyping in Musa using microsatellite markers. AoB Plants 2011:plr024.
Christelová P, Valárik M, Hřibová E, De Langhe E, Doležel J. 2011. A multi gene sequence-based phylogeny of the Musaceae (banana) family. BMC Evol Biol. 11:103.
Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676.
Copetti D, et al. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc Natl Acad Sci U S A. 114(45):12003–12008.
Davey MW, et al. 2013. A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids. BMC Genomics 14(1):683.
De Langhe E, et al. 2009. Why bananas matter: an introduction to the history of banana domestication. Ethnobot Res Appl. 7:165–177.
Denton JF, et al. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 10(12):e1003998.
D'Hont A, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature. 488:213.
Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28(8):2239–2252.
Eaton DAR, Hipp AL, González-Rodríguez A, Cavender-Bares J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution 69(10):2587–2601.
Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). Syst Biol. 62(5):689–706.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.
English AC, et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7(11):e47768.
Foissac S, et al. 2008. Genome annotation in plants and fungi: EuGene as a model platform. Curr Bioinformatics 3:87–97.
FolkRA, Soltis Pamela S, Soltis Douglas E, Guralnick R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. Am J Bot. 105:364–375.
Fontaine MC, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347(6217):1258524.
Green RE, et al. 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.
Guignon V, et al. 2016. The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics. Curr Plant Biol. 7:6–9.
Guindon S, Delsuc F, Dufayard J-F, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537:113–137.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evol Int J Org Evol. 70(1):7–17.
Häkkinen M. 2013. Reappraisal of sectional taxonomy in Musa (Musaceae). Taxon. 62(1):809–813.
Hibbins MS, Hahn MW. 2018. Population genetic tests for the direction and relative timing of introgression. bioRxiv 328575.
Janssens SB, et al. 2016. Evolutionary dynamics and biogeography of Musaceae reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. New Phytol. 210(4):1453–1465.
Jarret R, Gawel N, Whittemore A, Sharrock S. 1992. RFLP-based phylogeny of Musa species in Papua New Guinea. Theor Appl Genet 84:579–584.
Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346(6215):1320–1331.
Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26(13):1669–1670.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.
Kreft L, Botzki A, Coppens F, Vandepoele K, Van Bel M. 2017. PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. Bioinformatics 33:2946–2947.
Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 11(1):81.
Lescot M, et al. 2008. Insights into the Musa genome: syntenic relationships to rice and between Musa species. BMC Genomics 9(1):58.
Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graph 20(12):1983–1992.
Li G, Davis BW, Eizirik E, Murphy WJ. 2016. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). Genome Res. 26(1):1–11.
Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.
Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46(3):523–536.
Magrane M, UniProt Consortium. 2011. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011:bar009.
Martin G, et al. 2016. Improvement of the banana '*Musa acuminata*' reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genomics 17:243.
Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Curr Opin Genet Dev. 15(6):589–594.
Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31(12):i44–i52.
Morgante M, De Paoli E, Radovic S. 2007. Transposable elements and the plant pan-genomes. Curr Opin Plant Biol. 10(2):149–155.
Novikova PY, et al. 2016. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat Genet. 48(9):1077–1082.
Pease JB, Rosenzweig BK 2018. Encoding Data Using Biological Principles: The Multisample Variant Format for Phylogenomics and Population Genomics. IEEE/ACM Trans Comput Biol Bioinformatics 15:1231–1238.
Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14(2):e1002379.
Perrier X, et al. 2011. Multidisciplinary perspectives on banana (*Musa* spp.) domestication. Proc Natl Acad Sci U S A. 108:11311–11318.
Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2(10):e173.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16(6):276–277.
Risterucci AM, et al. 2000. A high-density linkage map of *Theobroma cacao* L. Theor Appl Genet. 101(5-6):948–955.
Rouard M, et al. 2011. GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res. 39(Suppl_1):D1095–D1102.
Ruas M, et al. 2017. MGIS: managing banana (Musa spp.) genetic resources information and high-throughput genotyping data. Database 2017. doi: 10.1093/database/bax046.
Sarah G, et al. 2017. A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Mol Ecol Resour.17:565–580.
Sardos J, et al. 2016. A genome-wide association study on the seedless phenotype in banana (*Musa* spp.) reveals the potential of a selected panel to detect candidate genes in a vegetatively propagated crop. PLoS One 11(5):e0154448.
Sardos J, et al. 2016. DArT whole genome profiling provides insights on the evolution and taxonomy of edible banana (*Musa* spp.). Ann Bot. mcw170.
Scornavacca C, Berry V, Lefort V, Douzery EJ, Ranwez V. 2008. PhySIC_IST: cleaning source trees to infer more informative supertrees. BMC Bioinformatics 9(1):413.
Scornavacca C, Berry V, Ranwez V. 2011. Building species trees from larger parts of phylogenomic databases. Inf Comput. 209(3):590–605.
Shi C-M, Yang Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. Mol Biol Evol. 35(1):159–179.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Simmonds NW. 1956. Botanical results of the banana collecting expedition, 1954–5. Kew Bull. 11(3):463–489.
Simmonds NW. 1962. The evolution of the bananas.London (GBR): Longmans.
Simmonds NW, Shepherd K. 1955. The taxonomy and origins of the cultivated bananas. J Linn Soc Lond Bot. 55(359):302–312.
Simmonds NW, Weatherup STC. 1990. Numerical taxonomy of the wild bananas (Musa). New Phytol. 115(3):567–571.
Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102:13950–13955.
Thomas DC, et al. 2012. West to east dispersal and subsequent rapid diversification of the mega-diverse genus Begonia (Begoniaceae) in the Malesian archipelago. J Biogeogr. 39(1):98–113.
Veeramah KR, et al. 2015. Examining phylogenetic relationships among Gibbon genera using whole genome sequence data using an approximate Bayesian computation approach. Genetics 200(1):295–308.
Wu M, Kostyun JL, Hahn MW, Moyle L. 2017. Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. bioRxiv 201376.
Yachdav G, et al. 2016. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics 32:3501–3503.
Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19(Suppl 6):153.
Zimin AV, et al. 2013. The MaSuRCA genome assembler. Bioinformatics 29(21):2669–2677.

Associate editor: Laura Rose