The banana (Musa acuminata) genome and the evolution of monocotyledonous plants

Angélique D’Hont1*, France Deneuque2,3,4*, Jean-Marc Aury2, Fraiz Christophe Baurens3, Françoise Carreel1,5, Olivier Garsmeur1, Benjamin Noel2, Stéphanie Bocs1, Gaëtan Droch1, Mathieu Roudier4, Corinne Da Silva2, Kamel Jabbari2,3,4, Céline Cardi1, Julie Poulaing1, Marlène Souquet1, Karine Labadie2, Cyril Jourda1, Juliette Lengelle1, Marguerite Rodier-Gould1, Adriana Alberti2, Maria Bernard2, Margot Correa2, Saravanaraj Ayyampalayam7, Michael R. McKain7, Jim Leebens-Mack7, Diane Burgess8, Mike Freeling8, Didier Mbégoué-A-Mbégoué1, Matthieu Chabannes3, Thomas Wicker10, Olivier Panaud11, Jose Barbosa9, Eva Hirboiva1, Pat Heslop-Harrison1, Rémy Habas2, Ronan Rivallan3, Philippe Francois1, Claire Poiron1, Andrzej Kilian4, Dheema Burtihia1, Christophe Jenny1, Frédéric Bakry1, Spencer Brown4, Valentin Guignon5, Gert Kema6, Miguel Díaz5, Cees Waalwijk16, Steeve Joseph1, Anne Dievart1, Olivier Jaillon2,3,4, Julie Leclercq1, Xavier Argout1, Eric Lyons17, Ana Almeida8, Mouna Jeridi1, Jaroslav Dolezel12, Nicolas Roux6, Ange-Marie Risterucci1, Jean-Christophe Glaszmann1, Francis Quétier18, Nabila Yahiaoui1 & Patrick Wincker2,3,4.

Bananas (Musa spp.), including dessert and cooking types, are giant perennial monocotyledonous herbs of the order Zingiberales, a sister group to the well-studied Poales, which include cereals. Bananas are vital for food security in many tropical and subtropical countries and the most popular fruit in industrialized countries1. The Musa domestication process started some 7,000 years ago in Southeast Asia. It involved hybridizations between diverse species and subspecies, fostered by human migrations2, and selection of diploid and triploid seedless, parthenocarpic hybrids thereafter widely dispersed by vegetative propagation. Half of the current production relies on somaclones derived from a single triploid variety Fig. 1). Here we describe the draft sequence of the 523-megabase genome of a Musa acuminata doubled-haploid genotype, providing a crucial stepping-stone for genetic improvement of banana. We detected three rounds of whole-genome duplications in the Musa lineage, independently of those previously described in the Poales lineage and the one we detected in the Arecales lineage. This first monocotyledon high-continuity whole-genome sequence reported of the subspecies malaccensis of the subspecies M. acuminata genomes of Cavendish7. A total of 235 microRNAs from 37 families were identified, including only one of the eight microRNA gene (MIR) families found so far solely in Poaceae6. Viral sequences related to the banana streak virus (BSV) dsDNA plant pararetrovirus were found to be integrated in the Pahang genome, with 24 loci spanning 10 chromosomes (Supplementary Text and Supplementary Fig. 2). They belonged to a badnavirus phylogenetic group that differed from the endogenous BSV species (eBSV) found in M. balbisiana8 and most of them formed a new subgroup (Supplementary Fig. 3). Importantly, all of the integrations were highly reorganized and fragmented and thus did not seem to be capable of forming free infectious viral particles, contrary to the eBSV described in M. balbisiana9.

Transposable elements account for almost half of the Musa sequence (Supplementary Text and Supplementary Tables 1 and 8–10). Long terminal repeat retrotransposons represent the largest part, with Copia elements being much more abundant than Gypsy elements (25.7–11.6%) (Supplementary Fig. 4). No major recent wave of long terminal repeat retrotransposon insertions appears to have occurred in the Musa lineage. Fewer than 1% of the long terminal repeat retrotransposons are complete and their median insertion is around 4 Myr ago, corresponding to the half-life of this type of sequence errors. The assembly consisted of 24,425 contigs and 7,513 scaffolds with a total length of 472.2 Mb, which represented 90% of the estimated DH-Pahang genome size. Ninety percent of the assembly was in 647 scaffolds, and the N50 (the scaffold size above which 50% of the total length of the sequence assembly can be found) was 1.3 Mb (Supplementary Text and Supplementary Tables 1–3). We anchored 70% of the assembly (332 Mb) along the 11 Musa linkage groups of the Pahang genetic map. This corresponded to 258 scaffolds and included 98.0% of the scaffolds larger than 1 Mb and 92% of the annotated genes (Supplementary Text, Supplementary Table 4 and Supplementary Fig. 1). We identified 36,542 protein-coding gene models in the Musa genome (Supplementary Tables 1 and 5). A total of 235 microRNAs from 37 families were identified, including only one of the eight microRNA gene (MIR) families found so far solely in Poaceae6 (Supplementary Tables 6 and 7).

1Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR AGAP, F-34938 Montpellier, France. 2Commissariat à l’Énergie Atomique (CEA), Institut de Génomique (IG), Genoscope, 2 rue Gaston Crémieux, B.P. 706, 91057 Évry, France. 3Centre National de Recherche Scientifique (CNRS), UMR 8030, BP 706, Évry, France. 4Université d’Évry, UMR 8030, BP706, Évry, France. 5Centre National de Recherche Scientifique (CNRS), UMR 8030, Campus international de Baillarguet, F-34398 Montpellier, France. 6Biodiversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France. 7Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA. 8Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA. 9CIRAD, UMR QUALISUD Station de Neufchâtel, Sainte-Marie, 97130 Capoësterre-Belle-Eau, France. 10Institute of Plant Biology, University of Zurich, CH-8008 Zurich, Switzerland. 11Laboratoire Génomique et Développement des Plantes, UMR 5096 CNRS-UPVD, 66000 Perpignan, France. 12Centre of the Region Hana for Biotechnological and Agricultural Research, Institute of Experimental Botany, Sokolovska 6, CZ-77200 Olomouc, Czech Republic. 13Department of Biology, University of Leicester, Leicester LE1 7RH, UK. 14Diversity Arrays Technology, Yarralumla, Australian Capital Territory 2600, Australia. 15Institut de Sciences du Végétal, CNRS UPR 2355 et FR3315, 91198 Gil-sur-Yvette, France. 16University of Wageningen, Plant Research International, E700 AA Wageningen, Netherlands. 17Department of Plant Sciences, University of Arizona, Tucson, Arizona, USA. 18Département de Biologie, Université d’Évry Val d’Essonne, Évry, France.

*These authors contributed equally to this work.

©2012 Macmillan Publishers Limited. All rights reserved
transposable element14 (Supplementary Fig. 5). Long interspersed elements (LINEs) represent 5.5% of the genome. The banana genome is exceptional in the composition of its class 2 element population, which represents only about 1.3% of the genome. The only superfamilies identified were hAT, followed by Harbinger and Mutator. Only the first family was significantly represented and had non-autonomous deletion derivatives. The superfamilies CACTA and Mariner, which have been found in high copy numbers in all angiosperm genomes studied so far, are absent from the banana genome. Gene-rich regions are mostly located on distal parts of chromosomes, as observed in other plant genomes (Fig. 1 and Supplementary Fig. 1). There is, however, a particularly sharp transition between gene-rich and transposable-element-rich regions. This observation is confirmed by the pattern observed after genomic in situ hybridization, which shows that transposable elements are typically concentrated around centromeres in *Musa*15 (Supplementary Fig. 6). The asymmetric transposable element distributions along the chromosomes indicated that chromosomes 1 and 2 are acrocentric in DH-Pahang (Fig. 1). Long terminal repeat retrotransposons are particularly abundant in centromeric and pericentromeric chromosome regions. Their accumulation in these regions, particularly for the oldest ones, suggests that they are preferentially eliminated from gene-rich regions15 (Supplementary Fig. 5). Remarkably, typical short tandem centromeric repeats were not found in *Musa*. However, one long interspersed element (named *Nanica*) identified in the unassembled reads was localized by fluorescence in situ hybridization in the centromeric region of all *Musa* chromosomes (Supplementary Fig. 7 and Supplementary Table 10).

Whole-genome duplications (WGDs) have played a major role in angiosperm genome evolution14; the first evidence of a WGD event in the *Musa* lineage was reported by Lescot et al.15. We uncovered a complex pattern of paralogous relationships between the 11 *Musa* chromosomes (Supplementary Text and Supplementary Fig. 8). Most paralogous gene clusters shared relationships with three other clusters, suggesting that two WGDs (denoted as α and β) occurred (Supplementary Fig. 9). Based on Ks and synteny relationships, duplicated gene clusters were tentatively assembled into 12 *Musa* ancestral blocks representing the ancestral genome before the α/β duplications (Figs 1 and 2 and Supplementary Figs 10–12). The duplicated segments included in the *Musa* ancestral blocks cover 222 Mb (67% of the anchored assembly) and contain 26,829 genes (80% of the anchored genes) (Supplementary Table 11). The Ks distribution among pairs of paralogous gene clusters dated the two WGDs at a similar period around 65 Myr ago (Supplementary Fig. 13), consistent with the WGDs that occurred in many different plant lineages near the Cretaceous–Tertiary boundary14 (Fig. 3). Additional paralogous relationships between the 12 *Musa* ancestral blocks displaying higher Ks values suggested that an additional, more ancient duplication event (denoted as γ) occurred around 100 Myr ago (Fig. 3 and Supplementary Figs 10, 11, 13 and 14).

In the grass lineage, it is well established that one WGD (denoted as ρ) occurred around 50–70 Myr ago, after Poales separated from other monocotyledon orders16,17. Evidence was reported on an additional WGD (denoted as σ) earlier in the monocotyledon lineage, but after its divergence from the eudicotyledons18. Our comparison of the *Musa* ancestral blocks with the Poaceae σ and α ancestral blocks as defined by Tang et al.19 revealed that genes from segments of different σ blocks (corresponding to one σ block) have orthologous relationships with the same *Musa* regions, showing that the σ Poaceae event is not shared with *Musa*. Reciprocally, genes from *Musa* α/β paralogous segments corresponding to the 12 *Musa* α/β ancestral blocks (right). LINEs, long interspersed elements.

**Figure 1** | Chromosomal distribution of the main *M. acuminata* genome features. Distribution of genes and transposable elements (left) and paralogous relationships between the 11 chromosomes indicated with 12 distinct colours.

**Figure 2** | Whole-genome duplication events. a, Paralogous relationships between chromosome segments from *Musa* α/β ancestral blocks 2 (red) and 8 (green). The 12 *Musa* α/β ancestral blocks are shown in different colours on the circle. b, Orthologous relationships of *Musa* ancestral blocks 2 and 8 with rice ancestral blocks ρ2, ρ5 and ρ6. We did not observe a one-to-one relationship between, for instance, *Musa* α/β ancestral block 2 and one ρ ancestral block, which suggests that the γ and σ duplications are two separate events. c, Representation of the deduced WGD event.
have orthologous relationships with the same ρ and σ regions, showing that the earliest duplication (γ) we identified in the *Musa* lineage is not shared with Poaceae (Fig. 2 and Supplementary Fig. 15).

Independent phylogenomic analyses performed on 3,553 gene families, including genes mapped to syntenic ancestral blocks, generated further evidence (98.7–77.6% of the gene trees, Supplementary Text) that the three rounds of palaeopolyploidization identified in the *Musa* genome and the two previously reported in the *Musa* lineage occurred independently after the Poales and Zingiberales divergence estimated at 109–123 Myr ago19 (Fig. 3 and Supplementary Fig. 16).

Resolution of the Zingiberales relationship relative to Poales and Arecales (palms) has been problematic (see, for example, Givnish et al.20), but our analysis of 93 single-copy nuclear genes suggested that the palms are more closely related to Zingiberales (including *Musa*) than to Poales (Fig. 3, Supplementary Text and Supplementary Figs 20–23 and Supplementary Table 18), which confirms preliminary evidence that placed *Musa* in an intermediate position15. This feature was shared with ginger (Zingiberales) and contrasts with the unimodal GC distribution of date-palm coding sequences (Supplementary Fig. 21).

Plant conserved non-coding sequences (CNSs)—a type of phylogenetic footprint—are enriched in known transcription factors or other cis-acting binding sites, and are usually clustered around regulatory genes, supporting their functionality26. Starting with a collection of 16,978 CNSs conserved in Poaceae, we used the *Musa* genome to identify the 116 most deeply conserved regulatory binding sequences (Supplementary Table 22). The banana genome also served as a stepping-stone to finding CNSs conserved beyond monocotyledons (Supplementary Table 19 and Supplementary Table 20, and Supplementary Fig. 24). Deeply conserved CNSs in commelinids were frequently found located 5′ to genes encoding transcription factors, and were significantly enriched in WRKY motifs (Supplementary Table 21). After WGD, genes associated with deeply conserved CNSs were found to be retained as duplicates more often than genes with less deeply conserved CNSs (Supplementary Table 22). The banana genome also served as a stepping-stone to finding CNSs conserved beyond monocotyledons, including 18 CNSs that were found in this study to be conserved in the expected syntenic position in eudicotyledons as well (Supplementary Table 23). This evolutionary distance is not unusual for vertebrate CNSs (detectable after more than 400 million years of divergence)27, but it surpasses the findings of previous plant whole-genome surveys28. Plant deeply conserved CNSs are therefore rare but do exist, and are short compared with those of animals27, and must be at least as old as monocotyledon–eudicotyledon divergence (more than 130 million years of divergence).

---

Figure 3 | Timing of whole-genome duplications relative to speciation events within representative monocotyledons and eudicotyledons. Boxes indicate WGD events. Green boxes indicate WGD events analysed in this paper. All nodes have 100% bootstrap support in a maximum likelihood analysis. Branch lengths (synonymous substitution rate) are indicated. The timing of the β WGD event relative to the Musaceae/Zingiberaceae split remains to be clarified.

Figure 4 | Six-way Venn diagram showing the distribution of shared gene families (sequence clusters) among *M. acuminata*, *P. dactylifera*, *Arabidopsis thaliana*, *Oryza sativa*, *Sorghum bicolor* and *Brachypodium distachyon* genomes. Numbers of clusters are provided in the intersections. The total number of sequences for each species is provided under the species name (total number of sequences/total number of clustered sequences).
The reference Musa genome sequence represents a major advance in the quest to unravel the complex genetics of this vital crop, whose breeding is particularly challenging. Having access to the entire Musa gene repertoire is a key to identifying genes responsible for important agronomic characters, such as fruit quality and pest resistance. Bananas are exported green and then ripened by application of ethylene. RNA-Seq analysis indicated strong transcriptional reprogramming in mature green banana fruits after ethylene treatment (Supplementary Text, Supplementary Tables 24–26 and Supplementary Fig. 25). Transcription factors were particularly involved with 597 differentially regulated genes. Various modifications confirmed the biochemistry of the banana ripening process, such as highly upregulated genes encoding cell-wall modifying enzymes, three downregulated starch synthase genes and one upregulated β-amylase gene. Two WGD-derived paralogous vacuolar invertase genes involved in sucrose conversion displayed opposite expression profiles, suggesting subfunctionalization and possible contribution to the soluble sugar balance in ripening bananas (Supplementary Fig. 26). The race against pathogen evolution is particularly critical in clonally propagated crops such as banana. Up to 50 pesticide treatments a year are required in large plantations against black leaf streak disease, a recent pandemic caused by Mycosphaerella fijiensis. Moreover, outbreaks of a new race of the devastating Panama disease fungus (Fusarium oxysporum f. sp. cubense) are spreading in Asia. Among defence-related genes, those encoding nucleotide-binding site leucine-rich repeat proteins were found to be little represented in the Musa sequence (89 genes) (Supplementary Table 27). RNA-Seq analysis showed that receptor-like kinase genes were upregulated in a partially resistant interaction with M. fijiensis (Supplementary Text, Supplementary Table 28 and Supplementary Fig. 27). Interestingly, direct links between basal plant immunity triggered by receptor-like kinase proteins and quantitative trait loci for partial resistance have been recently established in several plant species (see, for example, Poland et al.29). In addition, we showed that DH-Pahang is highly resistant to the new broad-range Fusarium oxysporum race 4 (Supplementary Text and Supplementary Fig. 28), thus conferring additional specific value to the DH-Pahang sequence.

The Musa genome sequence reported here bridges a large gap in genome evolution studies. As such, it sheds new light on the monocotyledon lineage. Several Poaceae-specific characteristics could be highlighted, boosting prospects for analysing the emergence of this very successful family. The Musa genome also enabled identification of deeply conserved CNS within commelinid monocotyledons and between monocotyledons and eudicotyledons, representing an invaluable resource for detecting novel motifs with a gene regulation function. We detected three rounds of polyploidization in the Musa lineage, which were followed by gene loss and chromosome rearrangements, resulting in little synteny conservation between lineages (Supplementary Figs 29 and 30) and over-retention of some gene classes, thus providing ample opportunities for independent diversification. In particular, transcription factor families are strikingly classes, thus providing ample opportunities for independent diversification, resulting in little synteny conservation between lineages.

The Musa genome sequence is therefore an important advance towards securing food supplies from new generations of Musa crops, and provides an invaluable stepping-stone for plant gene and genome evolution studies.

METHODS SUMMARY

Sanger (ABI 3730xd sequencers) and Roche/454 (GSFLX pyrosequencing platform) reads were assembled with Newbler. Scaffolds were anchored to Pahang linkage groups using 652 markers (SSR and DAfT). Protein-coding gene model prediction on the repeat-masked sequence was done with the GAZEl computer framework by combining ab initio gene predictions, protein similarity, existing banana and monocotyledon transcript information and banana RNA-Seq data. A reference library of Musa transposable elements was built based on sequence similarity at the protein and nucleic acid levels and on searches for transposable-element structural signatures. The library was used with the REPET package (http://urgi.versailles.inra.fr/Tools/REPET) to screen the Musa assembly and quantify repeats.

RNA-Seq differential gene expression analysis was performed using Illumina GALIX 76 bases reads that were mapped to the DH-Pahang sequence using SOAP2 (http://soap.genomics.org.cn/).

Received 10 February; accepted 18 May 2012.

Published online 11 July 2012.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was mainly supported by French National Research Agency, Commissariat à l’Energie Atomique and Centre de coopération Internationale en Recherche Agronomique pour le Développement. The Generation Challenge program supported DArT genotyping, and Stichting Het Groene Woud part of BAC-end sequencing. We thank M. Teixeira Souza for authorizing early access to BAC-End sequences, L. Baudouin and T. Hardcastle for their help with the Bayseq analysis, O. Inizan, T. Flutre and F. Choulet for their help in transposable-element mapping. We thank the SouthGreen Bioinformatics Platform (http://southgreen.cirad.fr) for providing us with computational resources. We thank D. Manley for his help with the English in this paper.


Author Information The final assembly and annotation are deposited in DDBJ/EMBL/GenBank under accession numbers CAIC01000001–CAIC01024424 (contigs), HE806462–HE813974 (scaffolds) and HE813975–HE813985 (chromosomes). Genome sequence and annotation can be obtained and viewed at http://banana-genome.cirad.fr. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.D.H. (angelique.d’hont@cirad.fr) or P.W. (pwincker@genoscope.cns.fr).
METHODS
Plant material and DNA preparation. Doubled-haploid Pahang (DH-Pahang ItC1511) was obtained from wild M. acuminate subspecies malaccensis accession ‘Pahang’ through anther culture and spontaneous chromosome doubling31. Genome sizes were estimated by flow cytometry according to Marie and Brown32. High molecular weight DNA was prepared from the youngest fully expanded leaf of DH-Pahang as described in Piffanelli et al.33 with minor modifications (Supplementary Text).

Genome sequencing. The genome was sequenced using a Whole Genome Shotgun strategy combining Sanger, Roche/454 GSFLX and Illumina GAIIx technologies. Sanger sequencing was performed with the ABI 3730d on 10-kilobase (kb) inserts and on two BAC libraries generated with the HindIII and BamHI restriction enzymes resulting in 2.0 million 10-kb fragment-ends and about 90,500 BAC-ends. A total of 27.5 million reads were obtained using Roche/454 GSFLX.

Genome assembly and automatic error corrections with Solexa/Illumina reads. All reads were assembled with Newbler version MapAsmResearch-03/15/2010. From the initial 29,620,875 reads, 87.8% were assembled. We obtained 24,425 contigs that were linked into 7,513 scaffolds. The contig N50 (the contig size above which 50% of the total length of the sequence assembly is included) was 43.1 kb, and the scaffold N50 was 1.3 Mb. The cumulative scaffold size was 47.2 Mb, about 10% smaller than the estimated genome size of 5.23 Mb. Sequences from each of the scaffolds from the Newbler assembly was improved, as described previously34, by automatic error corrections with Solexa/Illumina reads (50-fold genome coverage), which have a different bias in error type compared with 454 reads. To validate the assembly, we built a unigene set corresponding to 15,017 isoforms that were obtained from the assembly with Newbler (version MapAsmResearch-03/15/2010) of Roche/454 GSFLX reads from six different complementary DNA (cDNA) libraries (829,587 reads, Supplementary Text).

The unigenes were aligned with the assembly using the BLAT algorithm35 with default parameters, and the best match was kept for each unigene. The assembly covers a very large proportion of the euchromatin of the genome35, as 99% of the set of 15,017 unigenes was recovered in the DH-Pahang genome assembly.

Construction of the Pahang genetic map and sequence anchoring. A genetic map was specifically developed for scaffold anchoring and orientation. A total of 2,454 single sequence repeats (SSR) markers and 1,008 polymorphic diversity array technology (DARt) markers were analysed including 1,411 new SSRs defined on sequence contigs and scaffolds. The map used for anchoring was built with 589 SSR and 63 DARt markers that were genotyped on 180 individuals of the Pahang self progeny. Data were analysed using JoinMap 4 (Plant Research International). The 652 markers anchored 258 scaffolds along the 11 linkage groups. Orientation of scaffolds was possible when two or more markers were aligned in the opposite groups of the genetic map. Insertion dates of full-length long terminal repeat retrotransposons were determined as described in Ma et al.36 (Supplementary Text). Orientation of scaffolds was possible when two or more markers were aligned in the opposite groups of the genetic map. Insertion dates of full-length long terminal repeat retrotransposons were determined as described in Ma et al.36 (Supplementary Text). Orientation of scaffolds was possible when two or more markers were aligned in the opposite groups of the genetic map. Insertion dates of full-length long terminal repeat retrotransposons were determined as described in Ma et al.36 (Supplementary Text). Orientation of scaffolds was possible when two or more markers were aligned in the opposite groups of the genetic map. Insertion dates of full-length long terminal repeat retrotransposons were determined as described in Ma et al.36 (Supplementary Text). Orientation of scaffolds was possible when two or more markers were aligned in the opposite groups of the genetic map.

Identification of Musa WGDs and comparative genome analyses. Identification of the Musa WGD was based on sequence similarity at the protein and nucleic-acid levels using blastp and tblastn37 and by de novo identification based on transposable-element structural signatures. Repeats from 1,832,094 remaining unassembled reads were characterized with a blastn ‘walk- ing’ approach38. The obtained reference Musa transposable-element library was used with repet49 to screen the assembly and quantify repeats (Supplementary Text). Insertion dates of full-length long terminal repeat retrotransposons were determined as described in Ma et al.36 with a substitution rate of 9 × 10−8 per site per year, which is twofold higher than that determined for Musa genes by Lescot et al.39.

Identification of Musa WGDs and comparative genome analyses. For the identification of Musa WGD, an all-against-all comparison of Musa protein sequences was performed using the GenomeQuest BLAST package (LASSAP40) and retaining the best hits for each gene. Clusters of paralogues composed of at least 20 genes with a maximal distance of 40 genes between syntenic genes were built with an in-house perl script, using a single linkage clustering with a Euclidian distance based on the gene index order in each chromosome. These clusters were refined using Synmap (http://synteny.cnr.berkeley.edu/CoGe/SynMap.pl) with the BLASTZ algorithm, an average distance expected between syntenic genes of 10, a maximum distance between two matches of 30, a minimum number of aligned gene pairs of 10 and a quota-align ratio of 3 to 3 (Supplementary Text).

For comparative genome analyses, orthologous gene-pairs were identified using predicted proteomes of M. acuminate, O. sativa (IRGPS/RAP, build 4), Vitis vinifera (http://www.genoscope.cns.fr/externe/Download/Projects/Project_ML/ data/12X/annotation/) and Phoenix dactylifera (draft sequence version 3, http://qatar-well.cornell.edu/research/datepalmGenome/download.html). Alignments were performed using BLASTp (e value 1 × 10−5) and retaining best hits. Syntenic clusters of genes were built using a single linkage clustering with a Euclidian distance. Dot-plots were performed using an in-house perl program allowing the painting of paralogous and orthologous gene clusters. Circle diagrams were made with Circos50.

To calculate the number of synonymous substitutions per site (Ks), ClustalW alignments of paralogous and orthologous protein sequences were used to guide nucleic coding sequence alignments with PAL2NA51. Ks values were calculated using the Yang–Nielson method implemented in PAML52.

Phylogenomic analysis. To infer the timing of genome duplication events relative to speciation events, all annotated Musa genes were sorted based on best BLASTp hits into the gene family clusters circumscribed by Liao et al.33 and the PlantTribes database53 (http://tigr.rosalind.info/tribedit/), including sequenced eudicotyledons and monocotyledons, along with transriptome assemblies for other non-grass monocotyledons (Supplementary Text). Gene family clusters were queried for Sorghum52 and Musa orthologues mapping to syntenic blocks, and maximum likelihood trees were constructed for these gene families using the GTR+Gamma model of molecular evolution in RAxML54. The estimation of divergence times was performed on maximum likelihood trees based on concatenated MAFFT55 alignments for 93 gene families that included only one gene from each of the sequenced genomes (Supplementary Text).

Comparative analysis of gene families. The Musa proteome was globally compared with O. sativa (RGAP version 6.0), S. bicolor (IGI version 1.4), P. dactylifera (IGI version 1.0), P. dactylifera (draft sequence version 3, http://qatar-well.cornell.edu/research/datepalmGenome/download.html) and A. thaliana (TAIR version 9) proteomes filtered of transposable elements and alternative splicing. An all-against-all comparison was performed using BLASTP (1 × 10−10) followed by clustering with OrthoMCL56 (inflation 1.5). Analysis of species-specific sets was made with a Fisher’s exact test (P < 0.0001) on InterPro (version 28) domains. For analyses of specific gene families, the 36,542 Musa protein sequences were inserted in the plant proteome clustering of the GreenPhyl database57. Transcription factor families were mostly retrieved using InterPro domains, using the IPR2genomes tool in GreenPhylDB58 (Supplementary Text). Kinases and nucleotide-binding site proteins were retrieved using hidden markov models (hmmssearch version 3) to search for corresponding Pfam domains (Supplementary Text).

Identification of CNSs. Pan-genomic conservation between rice, sorghum and Brachypodium were prepared using an automated pipeline59. The obtained 16,978 CNSs were used to query Musa using BLATSN (e value < 0.001) following a manual or a semi-automated procedure depending on CNS size (Supplementary Text and Supplementary Fig. 24). The resulting set of CNSs was extensively analysed using GEvo60 (http://synteny.cnr.berkeley.edu/CoGe/GEvo.pl) and the MSU Rice Genome Browser61 (http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/) to remove false positives (Supplementary Table 19). Adding rice and sorghum homologues, Brachypodium and maize orthologues and Arabidopsis ‘best hit orthologues’ to GEvo panels enabled the identification of 18 CNS conserved deeply throughout the monocot kingdom.

Transcriptome sequencing. For RNA-Seq analyses (Supplementary Text), cDNA libraries were sequenced using 76-base length read chemistry in a single-flow cell on the Illumina GA IIx. Reads were mapped against the automatic annotated transcripts with SOAPAligner/Soap2 (2.20, http://soap.genomic.sor.cn/ and only the unique mapped reads were kept. RNA-seq data were statistically analysed with the R packages baySeq version 1.6.0 (ref. 61) and DESeq version 1.5.6 (ref. 62).