Musa balbisiana genome reveals subgenome evolution and functional divergence

Zhuo Wang1,2, Hongxia Miao1,2, Juhua Liu1,2,12, Biyu Xu1,2, Xiaoming Yao3,12, Chunyan Xu3,12, Shancen Zhao4, Xiaodong Fang5, Caihong Jia1, Jingyi Wang1, Jianbin Zhang1, Jingyang Li2, Yi Xu2, Jiashui Wang2, Weihong Ma2, Zhangyan Wu3, Lili Yu3, Yulan Yang3, Chun Liu3, Yu Guo3, Silong Sun3, Franc-Christophe Baurens5,6, Guillaume Martin5,6, Frederic Salmon6,7, Olivier Garsmeur5,6, Nabila Yahiaoui5,6, Catherine Hervouet5,6, Mathieu Rouard6,8, Nathalie Laboureau9, Remy Habas9,10, Sebastien Ricci6,7, Ming Peng1, Anping Guo1, Jianghui Xie1, Yin Li11, Zehong Ding1, Yan Yan1, Weiwei Tie1, Angélique D’Hont5,6, Wei Hu1,2,1* and Zhiqiang Jin8,9,10.*

Banana cultivars (Musa ssp.) are diploid, triploid and tetraploid hybrids derived from Musa acuminata and Musa balbisiana. We presented a high-quality draft genome assembly of Musa balbisiana with 430 Mb (87%) assembled into 11 chromosomes. We identified that the recent divergence of Musa acuminata (A-genome) and Musa balbisiana (B-genome) occurred after lineage-specific whole-genome duplication, and that the B-genome may be more sensitive to the fractionation process compared to the A-genome. Homoeologous exchanges occurred frequently between A- and B-subgenomes in allopolyploids. Genomic variation within progenitors resulted in functional divergence of subgenomes. Global homoeologue expression dominance occurred between subgenomes of the allotriploid. Gene families related to ethylene biosynthesis and starch metabolism exhibited significant expansion at the pathway level and wide homoeologue expression dominance in the B-subgenome of the allotriploid. The independent origin of 1-aminocyclopropane-1-carboxylic acid oxidase (ACO) homoeologue gene pairs and tandem duplication-driven expansion of ACO genes in the B-subgenome contributed to rapid and major ethylene production post-harvest in allotriploid banana fruits. The findings of this study provide greater context for understanding fruit biology, and aid the development of tools for breeding optimal banana cultivars.

Bananas (Musa ssp.) are large herbaceous plants that are perennial but monocarpic. They originated in Southeast Asia and the Western Pacific and were one of the first crops to be domesticated, about 7,000 years ago, in Southeast Asia2. Bananas are widely distributed throughout the tropics and subtropics, where they are a staple food and fruit for millions of people2,3. Moreover, bananas are distributed throughout the tropics and subtropics, where they are a staple food and fruit for millions of people2,3. Moreover, bananas are an essential food resource and have important socioeconomic and ecological roles.

The genus Musa belongs to the monocotyledon Musaceae family along with the genus Ensete. Its wild species have traditionally been subdivided into four sections: Eumusa (x = 11; x represents the number of chromosomes), Rhodochlamys (x = 11), Australimusa (x = 10) and Callimusa (x = 9 or 10)10-12, and refined recently to two sections in which the Rhodochlamys and Australimusa were merged into the Eumusa and Callimusa, respectively4. Most edible bananas belong to the Eumusa (or Musa) section, and are categorized into the dessert or cooking group based on their usage. Furthermore, bananas of this section are distinguished based on their genetic background as Musa acuminata (A-genome, 2n = 2x = 22; n represents the haploid chromosome number), Musa balbisiana (B-genome, 2n = 2x = 22), Musa schizocarpa (S-genome, 2n = 2x = 22) and Australimusa species (T-genome, 2n = 2x = 20)13. The majority of edible cultivated bananas originated from intraspecific or interspecific hybridization between wild diploid Musa acuminata (A-genome) and Musa balbisiana (B-genome) species. Combinations of these A- and B-genomes have resulted in various genotypes of cultivated edible bananas, including diploid (AA, BB and AB), triploid (AAA, AAB and ABB) and tetraploid (AAAB, ABBB, ABBB) variants4. The triploid genotype variants constitute the predominant cultivated varieties that are planted worldwide.

Genome sequencing of the A-genome banana has provided insights into the evolution of monocotyledonous plants13. Although the A-genome represents a crucial step in the genetic improvement of banana, the lack of a high-quality B-genome sequence greatly hinders germplasm characterization and the molecular breeding of banana. A draft B-genome has previously been reported, but exhibited low quality, based on assembly and annotation via mapping reads to the A-genome14. Here, we sequenced the genome of the

*Key Laboratory of Biology and Genetic Resources of Tropical Crops, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Agricultural Sciences, Haikou, China. 1Key Laboratory of Genetic Improvement of Bananas, Hainan province, Haikou Experimental Station, China Academy of Tropical Agricultural Sciences, Haikou, China. 2BGI Genomics, BGI-Shenzhen, Shenzhen, China. 3BGI Institute of Applied Agriculture, BGI-Shenzhen, Shenzhen, China. 4CIRAD, UMR AGAP, Montpellier, France. 5AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France. 6CIRAD, UMR AGAP, Guadeloupe, France. 7Bioversity International, Montpellier, France. 8CIRAD, UMR BGPI, Montpellier, France. 9BGPI, CIRAD, INRA, Montpellier SupAgro, Montpellier, France. 10Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. 11These authors contributed equally: Zhuo Wang, Hongxia Miao, Juhua Liu, Biyu Xu, Xiaoming Yao, Chunyan Xu. 12e-mail: dhont@cirad.fr; huwei2010916@126.com; 18689846976@163.com
double haploid of the wild diploid genotype Pisang Klutuk Wulung (DH-PKW, 2n = 2x = 22), belonging to the species *M. balbisiana* that contributed the B-subgenome to cultivated allotriploid bananas. We further compared the B- and A-genomes to investigate subgenome evolution, genetic diversity and the functional divergence of subgenomes in polyploid bananas. Our analyses provide insights into the evolution and regulation of fruit-ripening processes in bananas. In particular, the results highlight a significant contribution of the B-genome towards ethylene biosynthesis and starch metabolism during fruit ripening.

**Results**

**Genome assembly and annotation.** To reduce heterozygosity, we used the DH-PKW genotype for our genome sequencing and assembly\(^{10}\). A total of 58.99 gigabases (Gb) (113x) of PacBio single-molecule long reads and 86.34 Gb (166x) Illumina paired-end and mate-pair reads were used for assembly (Supplementary Table 1), producing 492.77 megabases (Mb) of scaffolds. The contig N50 and scaffold N50 of the final assembly were 1.83 and 3.05 Mb, respectively (Supplementary Table 2). K-mer analysis suggested that the draft assembly covers approximately 95% of the genome size of DH-PKW (Supplementary Fig. 1). We further evaluated the completeness of the scaffold assembly using the BUSCO (v.3) plants datasets\(^{1} \). Precise exon placement of 91.3% of the total 1,440 single-copy orthologue groups in the embryophyta dataset was identified in the B-genome assembly. To anchor the scaffolds to chromosomes, we constructed high-throughput chromosome conformation capture (Hi-C) libraries of DH-PKW, generating 72 Gb (138x) Hi-C pair-end reads (Supplementary Table 1). Duplicate removal, sorting and quality assessment were performed with HiC-Pro\(^{13}\), and uniquely mapped valid reads were used for Hi-C scaffolding by LACHESIS software\(^{1} \) (Supplementary Fig. 2). As a result, 430 Mb (87.27%) of the assembly and 94.0% of the genes were placed on 11 chromosome groups (Fig. 1 and Supplementary Table 3). The 11 pseudo-molecules were named in accordance with the *M. acuminata* (A-genome) reference sequence\(^{1} \). About 55.75% of the B-genome assembly was composed of repetitive sequences, which is higher than the 41.85% of the A-genome assembly (Supplementary Table 4). This may be due to the spanning of repetitive regions by long reads\(^{14}\). Long terminal-repeat (LTR) retrotransposons represented the most abundant fraction of transposable elements in the A- and B-genomes, among which the families Gypsy and Copia accounted for 12.88 and 28.04% of the B-genome, respectively. DNA transposons comprised 2.12% of the B-genome and 2.03% of the A-genome (Supplementary Table 4). LTR retrotransposons tended to accumulate near the centromeric and pericentromeric regions (Fig. 1). Active insertions of LTR retrotransposons occurred more recently in the B-genome (peak at 0–0.5 million years ago (MYA)) relative to the A-genome (peak at 1.5–2.0 MYA) after their divergence (Supplementary Fig. 3).

**Synthony analysis**

We analysed the genome family clustering and expansion/contraction of banana genomes of *M. acuminata* (A-genome) and *M. balbisiana* (B-genome), compared to 14 other plant genomes using OrthoMCL and CAFE\(^{15} \) (Supplementary Table 7 and Supplementary Fig. 7). We found 9,038 gene families that were conserved in *M. balbisiana*, *M. acuminata*, *Oryza sativa*, *Brachypodium distachyon* and *Vitis vinifera*. In contrast, 348 and 639 gene families were specific to the A- and B-genome, respectively (Supplementary Fig. 8). After their divergence, 1,761 gene families were expanded and 203 were contracted in the A-genome, while 392 gene families were expanded and 1,008 contracted in the B-genome (Supplementary Fig. 9 and Supplementary Tables 8 and 9). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the significantly expanded gene families \((P < 0.05)\) in the B-genome suggested that it was enriched in the photosynthesis and biosynthesis of secondary metabolite pathways, including those associated with the metabolism of inositol, starch and sucrose, linoleic acid and arachidonic acid (Supplementary Fig. 10 and Supplementary Table 10). Plants produce a high diversity of secondary metabolites with prominent functions in defence against a variety of herbivores and pathogens, in addition to the mitigation of various types of abiotic stresses\(^{14}\). Thus, these observations are consistent with the association of the B-genome with improved vigour and tolerance to both biotic and abiotic stresses\(^{16}\).

**Syntenic analysis**

We identified 72 large syntenic blocks between the A- and B-genomes (Supplementary Fig. 11). We identified 72 large syntenic blocks between the A- and B-genomes, including 15 large blocks each containing over 900 gene pairs. These 72 syntenic blocks comprised 75.02% of A-genome space (containing 23% transposable elements) and 68.01% of B-genome space (containing 22% transposable elements) (Supplementary Table 11). We also identified two large translocations and two inversions between the A- and B-genome after their divergence. One large reciprocal translocation comprises 7.09 Mb on chromosome (chr) 1 of the B-genome and 7.03 Mb on chr:3 of the A-genome, and one large inversion of 9.39 Mb on chr:5 of the B-genome and 8.83 Mb on chr:5 of the A-genome (Fig. 1 and Supplementary Fig. 11). These translocations and inversions were also supported by the rearrangements based on genetic mapping\(^{3,15}\), and can serve to introduce novel genetic diversity into the A- and B-genomes\(^{16}\).
Previously, 12 *Musa* ancestral blocks were assembled that represented the ancestral genome before α/β-WGD. We identified 97 syntenic blocks resulting from α/β-WGD events in the B-genome by comparison to the 12 *Musa* ancestral blocks. These blocks contained 24,639 genes and represented 70.10% of all gene models involved in WGD-driven regions. We also identified 100 syntenic blocks that contained 26,780 genes (75.61%) in the A-genome (Supplementary Table 12). Of these ancestral *Musa* α/β blocks, 56.89% (15,236) and 60.59% (14,930) were singletons in the A- and B-genome, respectively, suggesting that genome fractionation (gene loss) and diploidization processes had occurred extensively after WGD events in the *Musa* lineage (Supplementary Table 12).

Taken together, our results indicated that the B-genome exhibited less expansion and more contraction of gene families, less syntenic coverage ratio and a higher singleton ratio in the ancestral blocks compared to the A-genome after divergence. Cycles of WGD followed by diploidization and fractionation have occurred across land plants, and are important in determining chromosome structure and gene content. Consequently, these processes have significantly contributed to the evolutionary success of plants. The diploidization

---

**Fig. 1 | Characterization of *M. balbisiana* (B-genome) and *M. acuminata* (A-genome) chromosomes.** Elements are arranged in the following scheme (from outer to inner). (1) Distribution of Gypsy elements (non-overlapping, window size, 50 kb); (2) distribution of Copia elements (non-overlapping, window size, 50 kb); (3) distribution of orthologous gene pairs between two genomes (non-overlapping, window size, 50 kb); (4) gene density (non-overlapping, window size, 50 kb); (5) syntenic relationships between A- and B-genomes. The connecting blue lines represent alignment blocks, red lines represent inversions, green lines represent translocations and grey lines show small blocks with <30 gene pairs.
and fractionation processes involve a series of evolutionary events, including repetitive DNA loss, chromosome rearrangements and complex patterns of gene loss. The above evidence supports the hypothesis that the B-genome was more sensitive to fractionation than the A-genome after their divergence.

Genetic diversity in polyploid bananas and functional divergence of subgenomes. Polyploid species usually exhibit vigorous growth, including high-quality production and high fitness. Most banana cultivars are polyploid and exhibit various levels of ploidy and genomic background. The genetic classification of some bananas is discordant, as is the case for Pelipita. Previous studies have shown that its karyotype comprises eight A and 25 B chromosomes as opposed to the predicted 11 A and 22 B chromosome distribution. Understanding the genetic diversity and genomic constitution of Musa accessions would inform genomic group classifications, in addition to conservation and breeding strategies. Therefore, we resequenced five triploid bananas and four diploid bananas to investigate their genetic diversity (Supplementary Table 13).

Simultaneous alignment of re-sequencing data to the A- and B-subgenomes identified the uniquely mapped reads that were used to analyse coverage depth, variations calling and homoeologous exchanges on each chromosome (Supplementary Table 14). Homeologous exchanges were characterized by read coverage that showed a chromosomal region with a duplicated copy from the corresponding homoeologous subgenome. These analyses confirmed that genome constitutions for the banana accessions were, in most cases, consistent with previous genome group classifications based on morphological traits. We identified 48 segmental homoeologous exchanges in the accession FenJiao, including nine from the B- to the A-subgenome and 39 in the reverse direction (Fig. 2a and Supplementary Table 15). We also found four segmental homoeologous exchanges from the B- to the A-subgenome in the accession Kamaramasenge, and replacement of chromosome 10 of the B-subgenome by the A-subgenome. (Fig. 2b and Supplementary Table 15). For the accession Pelipita, chromosomes 2, 7 and 11 of the A-subgenome were replaced by the B-subgenome and there were 18 segmental homoeologous exchanges on chromosomes 6, 9 and 10 (Fig. 2c and Supplementary Table 15). This indicates the eight A and 25 B chromosome constitution of Pelipita, consistent with previous genomic in situ hybridization studies. This classification is further supported by phylogenetic analyses based on genotyping data (Supplementary Fig. 12). A total of 18,475,661 single-nucleotide polymorphisms (SNPs), 14,053,391 small insertions and deletions and 220,452 structural variations were identified in the samples (Supplementary Tables 16–18). Analysis of gene and SNP density on the chromosomes indicated that SNPs were preferentially located on non-gene-rich regions (Supplementary Fig. 13). There were ~2,5-fold SNPs on the A-genome of Pisang_Mas and Pisang_Kra compared to the B-genome of Balbisiana (Supplementary Table 16). The nucleotide diversity (π, 0.0059) of A-subgenomes was higher than that of the B-subgenomes (0.0031) in accessions Fenjiao, Pelipita and Kamaramasenge.

Gene family expansion and contraction analysis of M. acuminate and M. balbisiana in comparison to other sequenced genomes indicated that there are 83 gene families significantly expanded in the A-genome (and conversely contracted in the B-genome). These families included plant–pathogen interactions, glycosphingolipid biosynthesis-ganglio series and glycosaminoglycan degradation pathways among others (Supplementary Table 19). Conversely, 33 gene families were significantly expanded in the B-genome (and contracted in the A-genome). These families included those involved in photosynthesis, metabolic pathways and ribosome, among others (Supplementary Table 20). This indicates that the A- and B-genomes may have functionally diverged at the genome level during their respective genome evolution.

To explore the transcription of allopolyploid subgenomes, we assessed the expression of homoeologue genes from the A- and B-subgenomes of the triploid FenJiao. Expression levels were measured within different tissues, at different stages of fruit development and ripening and in banana seedlings after abiotic stress treatments (Supplementary Table 21). A total of 25,717 homoeologous
Articles Nature Plants

S-adenosyl-l-methionine synthase (SAMS), 1-aminocyclopropane-1-carboxylic acid synthase (ACS), and the arginine biosynthesis pathway are well characterized, and include 75% of the seven other sequenced plant species among the mono- and eudicots22 (Supplementary Tables 27 and 28). We further characterized 28 homoallelic gene pairs during polyploid fruit ripening remains largely unknown.

We identified 125 SAMS, 11 ACS and 11 ACO genes from the A-genome and 10 SAMS, 11 ACS and 18 ACO genes in the B-genome, which represents a significant expansion compared to the seven other sequenced plant species among the monocots and eudicots22 (Supplementary Tables 27 and 28). We further characterized 28 homoallelic gene pairs from the A- and B-subgenomes, respectively (Supplementary Table 23). KEGG enrichment analysis indicated that genes with expression dominance in the B-subgenome were associated with 2-oxocarboxylic acid metabolism and the arginine biosynthesis pathway (q-value < 0.05) (Supplementary Table 24), whereas those showing expression dominance in the A-subgenome were not significantly enriched in KEGG pathways. Non-synonymous/synonymous substitution (Ka/Ks) ratios were calculated for all homoallelic gene pairs between the A- and B-subgenomes. The Ka/Ks ratios of genes with expression dominance in the A-subgenome (median, 0.157) were slightly lower than those in the B-subgenome (median, 0.196) and non-dominant genes (median, 0.186) (Supplementary Fig. 15).

We then constructed a gene co-expression network for those genes with expression dominance using weighted gene co-expression network analysis (WGCNA)23. The results indicated that 87 and 295 genes with dominance expression interacted with 4,302 and 4,612 genes in the A- and B-subgenomes, respectively. KEGG pathway enrichment analysis suggested that genes in the co-expression network of the A- and B-subgenomes were commonly associated with starch and sucrose metabolism (k00500) and other metabolic pathways. In particular, ubiquinone and other terpenoid-quione biosynthesis, photosynthesis-antenna proteins, carotenoid biosynthesis and other glycan degradation pathways were specially enriched in the A-subgenome (q-value < 0.05) (Supplementary Fig. 16 and Supplementary Table 25). In contrast, selenocompound metabolism and cyanoamino acid metabolism pathways were particularly enriched in the B-subgenome (q-value < 0.05) (Supplementary Fig. 17 and Supplementary Table 26). Overall, these results further support the hypothesis of functional divergence between the A- and B-genomes at the transcriptional level.

Expression dominance of homoallelic gene pairs in the ethylene biosynthesis pathway and expansion of the ACO family affect fruit ripening. Ethylene plays a key role in the regulation of climacteric fruit ripening post-harvest40. The core enzymatic steps in the ethylene biosynthesis pathways are well characterized, and include S-adenosyl-l-methionine synthase (SAMS), 1-aminocyclopropane-1-carboxylic acid synthase (ACS) and ACO41,42 (Fig. 3a). However, the expansion and expression dominance of these homoallelic gene pairs during polyploid fruit ripening remains largely unknown.

We identified 125 SAMS, 11 ACS and 11 ACO genes from the A-genome and 10 SAMS, 11 ACS and 18 ACO genes in the B-genome, which represents a significant expansion compared to the seven other sequenced plant species among the monocots and eudicots22 (Supplementary Tables 27 and 28). We further characterized 28 homoallelic gene pairs from the A- and B-genomes (Supplementary Table 29). These gene pairs displayed similar expression profiles in the BaXijiao (Musa AAA group, cv. Cavendish, BX), the A-subgenome of FenJiao (Musa ABB group, cv Pisang Awak, FJ) and the B-subgenome of FJ (Fig. 3b and Supplementary Tables 30–32). Interestingly, eight gene pairs exhibited homoallelic expression dominance in the B-subgenome and five gene pairs were dominantly expressed in the A-subgenome of FJ (Fig. 3b and Supplementary Tables 33 and 34). The dominant expression of these genes in various tissues may be related to the fundamental role of ethylene biosynthesis.

Both ACS and ACO have previously been demonstrated as limiting enzymes within ethylene biosynthesis41. The expression abundance of MA-ACS1 (AB021906) and MA-ACO1 (X91076) is consistent with ethylene production during the post-harvest banana-ripening period41. Of the ten ACS gene pairs, MaACS7/MbACS7, which is a homologue of MA-ACS1, exhibited high expression levels during fruit ripening and was dominantly expressed in the B-genome (Fig. 3b). MbACS6 and MbACS7 are paralogous in a large syntenic block (the block contains 19 gene pairs), and maintain synteny and close evolutionary relationships to MaACS6 and MaACS7, respectively, suggesting that these genes duplicated from WGD (Fig. 3c and Supplementary Fig. 18). Of the nine ACO gene pairs, three (MaACO2/MbACO6, MaACO3/MbACO7 and MaACO8/MbACO13) exhibited high expression levels during fruit ripening and were dominantly expressed in the B-subgenome (Fig. 3b). MaACO2/MbACO6 and MaACO3/MbACO7 are in the syntenic block of chr:5 and chr:6 between the A- and B-genome, respectively, and belong to the same phylogenetic clade (Fig. 3d,e and Supplementary Table 35). These results indicated that MaACO2/MbACO6 and MaACO3/MbACO7 originated independently and developed crucial functions during fruit ripening.

We also observed high expression levels (log2(RPKM ≥ 4 in at least one stage of fruit ripening in the B-genome) of homoallelic gene pairs, probably related to fruit softening (pectin methylesterases, galactosidases, expansions and pectate lyase)47, cell wall modification (xyloglucan endotransglycosylase/hydrolases, fascinil-like arabinogalactan proteins and β-d-xilosidas)48-56 and aroma production (alcohol dehydrogenases)48. These genes are closely involved in fruit ripening and are regulated by ethylene41. Almost all of these gene pairs showed expression dominance in the B-subgenome of FJ during fruit ripening, which is the same as the dominance expression of gene pairs related to ethylene biosynthesis (Supplementary Fig. 19 and Supplementary Tables 36 and 37). This co-dominance of homoallelic gene pairs in the B-subgenome further supports the significant contribution of the B-genome to ethylene biosynthesis and fruit ripening.

Gene duplication is a major mechanism that generates new genetic diversity as a basis for evolutionary innovation in eukaryotes47. Compared to the 11 ACO genes within the A-genome, the ACO genes in the B-genome expanded significantly to 18 members. The expansion of ACO genes, including MaACO2, -3, -4, -5 in chr 3, MbACO8, -9, -11 in chr 6 and MbACO16, -18 in scaffolds, was driven by tandem duplications in the B-genome (Fig. 3b,d). Of the ACO genes that were expanded in the B-genome, MbACO2 and MbACO3 showed strong expression levels during the fruit-ripening stages with log2(RPKM > 11 at 6 days post-harvest (DPH) in FJ, which is coincident with the ethylene climacteric period (Fig. 3b,f and Supplementary Table 38). In addition, MbACO8, -9, -11, -16, -18 exhibited high expression levels in roots and fruits at 0 days after flowering (DAF) (Fig. 3b and Supplementary Table 38). These genes belonged to the same cluster and their expression patterns were highly concordant with their duplication (Fig. 3d,e), suggesting that the expansion and evolution of ACO genes in the B-genome contributed to tissue development and fruit ripening.

Ripening of FJ was more rapid than BX during post-harvest ripening. BX required 8 and 14 DPH to reach the more-green-than-yellow and full-yellow stages, respectively, whereas FJ required 3 and 6 DPH to reach these stages, respectively (Supplementary Fig. 20). The dominant expression of ethylene biosynthesis and fruit ripening-related gene pairs and the expansion of the ACO family in the B-genome could have contributed to increased ethylene production and faster fruit ripening in FJ compared to BX.
Fig. 3 | Phylogeny and expression patterns of ethylene biosynthesis genes between M. acuminata (A-genome) and M. balbisiana (B-genome).

a. Overview of the ethylene biosynthesis pathway. b. Expression patterns of SAMS, ACS and ACO family genes in the root and leaf, and at different stages of fruit development and ripening in BX, the A-subgenome of FJ and the B-subgenome of FJ. Genes aligned horizontally in the heat map indicate homoeologue gene pairs between the A- and B-genomes. White boxes with diagonals indicate the lack of homoeologue gene pairs between the A- and B-genomes. Asterisks indicate expression dominance of homoeologue gene pairs between the A- and B-subgenomes of FJ. c, d. Synteny analysis of ACS (c) and ACO (d) families between the A- and B-genomes. Red lines indicate paralogous gene pairs resulting from WGD, blue lines indicate homoeologous gene pairs, purple lines indicate tandem duplication, light blue strips indicate aligned syntenic blocks, light green strip indicates translocation blocks and light red strips indicate inversion blocks. The blocks in outer ring represent location and length of genes; blue blocks represent genes from A-genome and orange blocks represent genes from B-genome. e. Phylogenetic analysis of ACO family genes among nine species: M. acuminata, M. balbisiana, A. thaliana, O. sativa, Sorgum bicolor, Solanum lycopersicum, Phoenix dactylifera, Asparagus officinalis and B. distachyon. f. Ethylene production at different stages of fruit development and ripening in BX and FJ. Error bars show standard error of the mean from three independent experiments (n = 3).
Fig. 4 | Comparison of genomic expansion, evolutionary history and differential expression patterns of the starch metabolic pathway between *M. acuminata* (A-genome) and *M. balbisiana* (B-genome). a, Overview of the starch biosynthesis and degradation pathway. b, Gene families in the starch metabolic pathway that are expanded in *M. acuminata* and *M. balbisiana*. c, Expression patterns of families AMY, BMY and DPE in the starch degradation pathway in BX, the A-subgenome of FJ and the B-subgenome of FJ during fruit-ripening stages. Horizontally oriented genes in the heat map indicate homoeologue gene pairs between the A- and B-genomes. White boxes with diagonals indicate that no homoeologue gene pairs were identified between the A- and B-genomes. Asterisks indicate expression dominance of homoeologue gene pairs between the A-subgenome of FJ and the B-subgenome of FJ. d, e, Synteny analyses of AMYs (d) and BMYs (e) between the A- and B-genomes. Red lines indicate paralogous gene pairs resulting from segmental/WGD-driven duplication, blue lines indicate homoeologous gene pairs, purple lines indicate tandem duplication, light blue strips indicate aligned syntenic blocks, light green strips indicate translocation block and light red strips indicate inversion blocks. The blocks in the outer ring represent location and length of genes; blue blocks represent genes from A-genome and orange blocks represent genes from B-genome. f, Starch contents at different stages of fruit development and ripening in BX and FJ. Error bars show standard error of the mean from three independent experiments (n = 3). g, Scanning electron microscopy of starch granules at different stages of fruit development and ripening in BX and FJ. The experiment was repeated three times independently with similar results.
The active starch metabolic pathway in the B-genome during fruit development and the ripening process. Starch is the most widespread and abundant storage carbohydrate in plants. It is also a major component of cultivated banana, accumulating to high levels (60–75% of dry weight) and leading to the presence of large starch granules (8–30 m) during banana fruit development, along with near-complete conversion to soluble sugars during post-harvest ripening. Thus, banana could serve as an excellent model for the investigation of starch metabolism in fresh starchy fruits. The major enzymes that are responsible for starch biosynthesis (sugars will eventually be exported transporter: SWEET; sucrose transporter: SUT; sucrose synthase: SuSy; UDP-glucose pyrophosphorylase: UGP; ADP-glucose pyrophosphorylase: AGPase; granule-bound starch synthase: GBSS; soluble starch synthase: SSS; starch branching enzyme: SBE; and starch debranching enzyme: DBE) and degradation (α-amylase: AMY; β-amylase: BMY; and starch phosphorylase: DPE) are encoded by multigenic families across land plants, and have significantly contributed to their evolutionary success. Our results indicate that the B-genome may be more sensitive to fractionation than the A-genome after WGD, although the A- and B-subgenomes have diverged very recently. Variation in genomic structure between the A- and B-genomes consists of chromosome rearrangements and gene loss during diploidization, which has resulted in the functional divergence of subgenomes in polyploidy bananas. This divergence is supported by differential enrichment of expanded/contracted gene families between the A- and B-genomes and the expression dominance of homoeologue genes from A- and B-subgenomes in triploids. Although homoeologue expression dominance has been identified in certain polyploid species, the relationship between homoeologue expression dominance and functional divergence (especially in regard to ethylene biosynthesis/starch metabolism) of subgenomes in triploids remains to be elucidated. Thus, these results provide an important basis for the improvement of agriculturally important traits by focusing selection on transcriptionally dominant genes. It is worth noting that homoeologous exchanges may obscure the signal of expression dominance in subgenomes of allopolyploids. The extensive homoeologue exchanges in allopolyploid bananas may remove many progenitor genome conflicts that result in subgenome biases in gene content and expression. Thus, homoeologous exchanges may contribute to the diversity of homoeologue expression dominance and induce a series of rapid genetic and epigenetic modifications for agronomic traits.

Previous studies have suggested an important role of ethylene production in fruit ripening and starch metabolism with regard to fruit quality post-harvest. However, the genetic mechanisms underlying polyploid fruit ripening are less well known. Here, we analysed biological processes related to these two pathways in triploid bananas. We identified significant genomic expansions and dominant expression of homoeologue genes in the B-genome at the pathway level of gene families and, most notably, in the ACO family, known to play a critical role in ethylene production. Our analysis revealed the origin and evolution of crucial gene families in these pathways, particularly for the independent origin of the MaACO2/MaACO6 and MaACO3/MaACO7 gene pairs and their specific function in fruit ripening. Moreover, we identified that this tandem event has led to expansion of ACO genes in the B-genome and to strong expression of these genes during fruit ripening. Our analysis also indicated a potential link between the dominant expression of homoeologue genes and the expansion of gene families with fruit ripening and starch metabolism. Previous studies have demonstrated that the B-genome is associated with improved vigour and tolerance to both biotic and abiotic stresses. Consequently, the
B-genome is a target for banana breeding programmes. Here, we highlight that the B-genome is of great importance in ethylene biosynthesis and post-harvest banana ripening, which will further our understanding of ripening mechanisms in polyplody fruit.

The *M. balbisiana* genome assembly, along with our previously acquired *M. acuminata* genome data, may aid in the discovery of cultivar-specific sequences that are related to important cultivar-specific traits, including shelf life, quality and stress tolerance. Thus, these resources can be used to build molecular characterization strategies for various cultivars to assist in rapid quality control and the conservation of biodiversity. The data from this study also pave the way for whole-genome association studies, germplasm improvement and genetic modification of bananas to meet increasing commercial demands. These genomic resources and results also reinforce the use of the banana as an appropriate model to study subgenomic evolution and fruit biology in triploid variants. Due to the sterility and seedlessness of banana cultivars, further efforts will be needed to leverage the key gene resources from precisely characterized germplasms to achieve effective breeding schemes.

**Methods**

**Sample collection.** A double haploid of the wild diploid genotype PKW; 2n = 2x = 22 was provided by the Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) for genome sequencing. Fresh, unexpanded leaves were harvested and then frozen immediately with liquid nitrogen to preserve genomic DNA for isolation. High-molecular weight genomic DNA was extracted using a standard cetyltrimethyl ammonium bromide (CTAB) method. DNA integrity was tested using the gel electrophoresis (concentration of agarose gel, 1%; voltage, 150 V; electrophoresis time, 40 min). Finally, DNA was purified from the gel using a QIAquick Gel Extraction kit (QIAGEN).

**Library construction and sequencing.** One paired-end and eight mate-pair libraries were constructed for short-read sequencing on the Illumina HiSeq 2000 platform, which generated around 86.34 Gb (166× coverage) of high-quality data. For long-read DNA sequencing, 5.79 million SMRT long reads (58.99 Gb data, 11×coverage) were sequenced using the PacBio Sequel system with libraries of 20-kb insert size; sub-reads had a mean length of 10.2 kb and N50 length of 16.6 kb. One Hi-C library was prepared and sequenced on Illumina NovaSeq 6000 to generate 71.96 Gb (138×coverage) of high-quality data (Supplementary Table 1). Additional details are available in the Supplementary note.

**Genome assembly.** De novo assembly of DH-PKW was performed using wtdbg (v1.1.2.8; https://github.com/ruanjue/wtdbg/) based on PacBio data (only reads longer than 1 kbp were used in assembly). The assembled genome was corrected for two rounds using the ‘wtdbg-cns’ programme in the wtdbg package. We then used the algorithm Arrow (https://github.com/PacificBiosciences/GenomicConsensus), which takes into account all of the underlying data and the raw quality values inherent in SMRT sequencing, to polish the assembly again for the final consensus accuracies. The final consensus contigs were scaffolded using the SPSPACE-standard program of two mate-pair reads from libraries of insert size 2–20 kb. Based on Hi-C data, 430.02-Mb scaffolds were anchored to 11 pseudo-molecules using a standard cetyltrimethyl ammonium bromide (CTAB) method. DNA integrity was tested using the gel electrophoresis (concentration of agarose gel, 1%; voltage, 150 V; electrophoresis time, 40 min). Finally, DNA was purified from the gel using a QIAquick Gel Extraction kit (QIAGEN).

**Library construction and sequencing.** One paired-end and eight mate-pair libraries were constructed for short-read sequencing on the Illumina HiSeq 2000 platform, which generated around 86.34 Gb (166× coverage) of high-quality data. For long-read DNA sequencing, 5.79 million SMRT long reads (58.99 Gb data, 11×coverage) were sequenced using the PacBio Sequel system with libraries of 20-kb insert size; sub-reads had a mean length of 10.2 kb and N50 length of 16.6 kb. One Hi-C library was prepared and sequenced on Illumina NovaSeq 6000 to generate 71.96 Gb (138×coverage) of high-quality data (Supplementary Table 1). Additional details are available in the Supplementary note.

**Evaluation of assembly quality.** BUSCO (v3.2.4) was used to assess assembly completeness. We mapped 29,610 *M. acuminata*-expressed sequence tags (ESTs) to the assembled genome using BLAST (v2.3.0+) with default parameters. In total, 93.59% of the ESTs were aligned to the genome with identity >90%. Additionally, BWA (v0.7.12) (ahn = 35) was used to map 59× Illumina reads to the assembly, and 96.11% of the reads were mapped to the assembled genome. Additional details regarding assembly quality are available in the Supplementary note.

**Genome annotation.** Repetitive sequences within the *M. balbisiana* genome were identified by a combination of homology-based and de novo approaches (Supplementary Table 1). Gene structures were annotated iteratively using three main approaches (ab initio predictions, homologous proteins and transcriptome data) that were combined using MAKER (v2.31.8) (Supplementary Table 5). Gene functions were annotated according to the best match of the alignments using blastp (E-value < 1 × 10−5) against the Swiss-Prot, TREMBL, NR (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/), KOG and KEGG databases. Additional details are available in the Supplementary note.

**Transcription factors.** We used the iTAK programme to identify transcription factors based on the consensus rules that are mainly summarized within PlantTFDB and PlantTFDB2, with families from PlantTFcat and AtTFDB used as supporting evidence. In total, we identified 3,329 transcription factor genes in *M. balbisiana* and 3,899 in *M. acuminata* (Supplementary Table 6).

**Gene family analysis.** A total of 500,142 genes from 16 plant species were used with available whole-genome sequences were used for gene family clustering analysis. BLAST (v2.2.26) was used to generate pairwise protein sequence alignments with E-values < 1 × 10−5. Then, OrthoMCL was used to cluster gene families by setting the main inflation value at 1.5 and using default settings for the other parameters. These analyses resulted in 39,358 gene families comprising 393,700 genes from the 16 species (Supplementary Table 7 and Supplementary Fig. 7).

We identified 519 single-copy gene families shared among the 16 species, and constructed a phylogenetic tree using MrBayes (v3.1.2) software with the geometrically reversible model (Supplementary Fig. 4). For the 16 species were also estimated based on fourfold degenerate sites of all single-copy orthologous genes using the MCMCTree programme in the PAML package (v4.4) (Supplementary Fig. 5).

**CLEF (v2.1) was used to analyse the expansion and contraction of gene families.** A random birth-and-death model was used to classify genes in gene families across the specified phylogenetic tree (Supplementary Fig. 9). Families with P < 0.05 were considered as significant expansion or contraction, and pathway enrichment analysis of these families was conducted using the enrichment pipeline.

**Whole-genome alignment analysis.** MCSCANX (parameters: -a 1 -r 5 -s 5) was used to detect collinearity within *M. acuminata* (A-genome) and *M. balbisiana* (B-genome) and among various species. Syntenic blocks containing at least ten gene pairs were obtained. All of the orthologous and paralogous gene pairs were extracted from the syntenic blocks for calculation of 4dTv distances using the HKY substitution model (Supplementary Fig. 8). Additional details are available in the Supplementary note.

**Orthologous gene pair analysis.** BLAST (v2.2.26) was used to align *M. acuminata* proteins to *M. balbisiana* proteins for identification of orthologous genes. The value 1 × 10−5 was used as a cut-off to define the raw orthologues. We then filtered the BLAST results using two parameters (CIP ≥ 60% and CALP ≥ 60% (ref. 1)). We identified 25,717 orthologous gene pairs (81.83% consistency with syntenic blocks) between *M. acuminata* and *M. balbisiana* using these two parameters (Supplementary Table 22). The orthologous gene pairs were first aligned using MUSCLE (v3.8.31.1), then the Ka/Ks ratio of each gene pair was calculated using KaKs Calculator; (v2.0) with model w80. The significant difference between Ka/Ks values was detected by Student’s t-test.

**Re-sequencing analysis.** Nine different genotypes of banana were used for re-sequencing, including the triploid plants BaXiJiao (subgroup Cavendish, AAA), Gros Michel (subgroup Gros Michel, AAA), Pisang_Mas (subgroup Pisang Awak, ABB), Kamararamsegan (ABB) and Pelipita (ABB), in addition to the diploid plants Pisang_Mas (subgroup Sucrier, AAA), Pisang_Kra (subsp. malaccensis, AA), DH-PKW (BB) and balbisiana (BB) (Supplementary Table 13). BaXiJiao, Gros_Michel and Pelipita were obtained from the Tissue Culture Centre of the Chinese Academy of Tropical Agricultural Sciences (CATAS). Pelipita, Pisang_Mas, Pisang_Kra, *M. balbisiana* and Kamararamsegan were provided by the Bioversity International Musa Transit Centre. Genomic DNA was extracted from fresh leaves of seedlings at the five-leaf stage using the CTAB method.

Paired-end reads with libraries of 500-bp insert size were aligned to the A- and B-genomes simultaneously using BWA (v0.7.12) with the parameters ‘bwa aln -t 20 -l 35’ (Supplementary Table 14), and only uniquely mapped reads were kept. Potential PCR duplicates were marked using Picard (v1.5.4, https://broadinstitute.github.io/picard/) and indexed using the SAMtools package. The Genome Analysis Toolkit was then used to infer SNPs and InDels. SNP identifications were filtered based on the following parameters: QD > 2.0 || DP < 20.0 || FS < 20.0 || ReadPosRankSum < −20.0. Breakdancer (http://breakdancer.sourceforge.net/) was used to detect structural variations by checking paired-end reads with an abnormal orientation and/or span. The final structural variations were filtered using the following requirements: (1) minimum size of 100 and maximum size of 1,000,000; (2) the minimum score of ≥20; and (3) minimum number of required reads supporting each structural variation ≥5 (Supplementary Table 18). Nucleotide diversity (θ) was analysed using VCFtools (v0.1.13; https://vcftools.github.io/man_latest.html).

**Analysis of homoeologous exchanges.** Assessment of read coverage depth was used to detect homoeologous exchanges between the A- and B-subgenomes. We inferred these based on cases where reads covering a given region on one parental genome was significantly high while the orthologous one was low or had no coverage. High coverage indicates duplication, and low or no coverage indicates loss. The uniquely mapped paired-end reads were used to calculate the coverage depth of each sample on the A- and B-subgenomes (Supplementary Figs. 25–27).
According to coverage depth, we detected homoeologous exchanges in the triploids 'FenJiao (ABB)', 'Pelipita (ABB)' and 'Kamaramasenge (AAB)' (Supplementary Table 15). Additional details are available in the Supplementary note.

Transcriptome analysis. Banana fruits at different stages of development (0, 20 and 80 DAF) and ripening (8 and 14 DPH for BX, 3 and 6 DPH for FJ) were obtained from the banana plantation at the Institute of Tropical Bioscience and Biotechnology (Chengmai, Haian, 20° N, 110° E). Two-month-old BX and FJ banana seedlings were obtained from the Tissue Culture Centre of CATAS and cultured in Hogaoland's solution. Seedlings at the five-leaf stage were treated with 200 mM mannotol for 7 days, 300 mM NaCl for 7 days and low-temperature conditions (4°C) for 22 h. Fruit, root and leaf samples were frozen in liquid nitrogen and stored at −80°C until RNA extraction for transcriptome analysis.

Total RNAs were isolated using a plant RNA extraction kit (TIANGEN). Total RNA (3 µg) from each sample was converted to complementary DNA using a RevertAid First-Strand cDNA Synthesis Kit (Fermentas). cDNA libraries were constructed using TruSeq RNA Library Preparation Kit v2, and were subsequently sequenced on the Illumina HiSeq 2000 platform using the Illumina RNA-Seq protocol. Two biological replicates were used for each sample. A total of 159.14 Gb (Supplementary Table 21) of high-quality clean data were produced. Gene expression levels were calculated as RPKM^{-1}. Differentially expressed genes were identified by methods previously established with the read count of two replicates for each gene (fold change ≥2; false discovery rate ≤0.001)^{14}. Additional details are available in the Supplementary note.

Weighted gene co-expression network analysis. Gene expression patterns for all identified genes were used to construct a co-expression network using WGCNA (v.1.47)^{15}. Genes without expression detected in all tissues were removed before analyses. Soft thresholds were set based on the scale-free topology criterion employed in ref. ^{16}. An adjacency matrix was developed using squared Euclidean distance values, and the topological overlap matrix was calculated for unsigned network detection using the Pearson method. Co-expression coefficients >0.55 between the target genes were then selected. Finally, the network connections were visualized using cytoscape^{16}. Determination of total starch content. Banana pulp was immersed in 0.5% sodium bisulphite for 10 min to prevent browning, and then dried at 60°C for 24 h. Pulp was then ground and centrifuged. The residue was suspended in 5 ml of 80% Ca(NO3)2 in a boiling water bath for 10 min then centrifuged for 4 min at 4000 r.p.m. The supernatant was transferred to a 20-ml volumetric flask and the residue was extracted twice with 80% Ca(NO3)2, which yielded a combined extract volume of 20 ml. All experiments were repeated three times. The total starch content was assessed following methods described by Yang et al.^{17}.

Scanning electron microscopy. Samples were fixed in stubs using double-faced tape and coated with a 10-nm platinum layer in a Bal-tec MED2020 coating system (Kettleshule). The prepared samples were analysed using a FEI Quanta 600 FEG scanning electron microscope (FEI Co.). Observations were performed in secondary electron mode while operating at 15 kV.

Measurement of ethylene production during fruit post-harvest stage. Ethylene production was measured by enclosing fruit samples in an airtight container for 2 h at 25°C. After incubation, 1 ml of the headspace gas was withdrawn and injected into a gas chromatograph (GC-2010, Shimadzu) fitted with a flame ionization detector and an activated alumina column. Ethylene production measurements were performed as recommended by the manufacturer.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw sequence reads for B-genome assembly and transcriptome for all samples were deposited in the CNSA (https://dl.cnbg.org/search/project/CSN0000292) of CNGIdb with accession number CSN0000292 and Sequence Read Archive of the National Centre for Biotechnology Information (NCBI) under the BioProject (No. PRIN4A332894). Genome assembly and annotation of DH-PKW were submitted to NCBI (No. PFTD0000000). Assembly and gene annotation of the A-genome (DH-Pahang) are available on the Banana Genome Hub (http://banana-genome-hub.southgreen.fr/).

Received: 7 May 2018; Accepted: 20 May 2019; Published online: 15 July 2019

References

1. D’Hont, A. et al. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488, 213–217 (2012).
2. Darmody, M. W. et al. A draft Musa balbisiana genome sequence for molecular genetics in polyploid, inter-and intra-specific Musa hybrids. BMC Genom. 14, 683 (2013).
3. De Langhe, E. et al. Why bananas matter: an introduction to the history of banana domestication. Ethnobot. Res. Appl. 7, 165–177 (2009).
4. Paul, J. et al. Golden bananas in the field: elevated fruit pro-vitamin A from the expression of a single banana transgene. Plant Biotechnol. J. 15, 520–532 (2017).
5. Cheesman, E. E. Classification of the bananas. Kew Bull. 2, 97–117 (1947).
6. Simmonds, N. W. & Shepherd, K. The taxonomy and origins of the cultivated bananas. Bot. J. Linn. Soc. 55, 302–312 (1955).
7. Daniells, J. et al. Diversity in the Genus Musa (INIBAP, 2001).
8. Hakkinen, M. Reappraisal of sectional taxonomy in Musa (Musaceae). Taxon 62, 809–813 (2013).
9. Martin, G. et al. Improvement of the banana ’Musa acuminata’ reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genom. 17, 243 (2016).
10. Assani, A. et al. Production of haploids from anther culture of banana [Musa balbisiana (BB)]. Plant Cell Rep. 21, 511–516 (2003).
11. Simao, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).
12. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16, 259 (2015).
13. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. 31, 1119 (2013).
14. Li, C. et al. Genome sequencing and assembly by long reads in plants. Genes 9, 6 (2017).
15. Campbell, M. S. et al. Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinformatics 48, 4–11 (2014).
16. Zheng, Y. et al. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol. Plant 9, 1667–1670 (2016).
17. Givnish, T. J. et al. Monocot plastid phylogenomics, timeline, net rates of species diversification, the power of multi-gene analyses, and a functional model for the origin of monocots. Am. J. Bot. 105, 1888–1910 (2018).
18. Lescot, M. et al. Insights into the Musa genome: syntenic relationships to rice and between Musa species. BMC Genomics 9, 58 (2008).
19. Christelova, P. et al. A multi gene sequence-based phylogeny of the Musaceae (banana) family. BMC Evol. Biol. 11, 103 (2011).
20. Janssens, S. B. et al. 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, 1453–1465 (2016).
21. Mandáková, T. et al. How diploidization turned a tetraploid into a pseudotriploid. Am. J. Bot. 103, 1187–1196 (2016).
22. Li, L., Stoekert, C. J. & Roos, D. S. OrthoMCL: identification of orthologous groups for eukaryotic genomes. Genome Res. 13, 2178–2189 (2003).
23. De Bie, T. et al. CAFe: a computational tool for the study of gene family evolution. Bioinformatics 22, 1269–1271 (2006).
24. Bennett, R. N. & Wallisgrove, R. M. Secondary metabolites in plant defence mechanisms. New Phytol. 127, 617–633 (1994).
25. Baurens, F. C. et al. Recombination and large structural variations shape interspecific edible bananas genomes. Mol. Biol. Evol. 36, 97–111 (2018).
26. Saxena, R. K., Edwards, D. & Varshney, R. K. Structural variations in plant genomes. Brief. Funct. Genom. 13, 296–307 (2014).
27. Jourdà, C. et al. Expansion of banana (Musa acuminata) gene families involved in ethylene biosynthesis and signalling after lineage specific whole genome duplications. New Phytol. 202, 986–1000 (2014).
28. Garsmeur, O. et al. Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. 31, 448–454 (2013).
29. Dodsworth, S., Chase, M. W. & Leitch, A. R. Is post-polyplidization diploidization the key to the evolutionary success of angiosperms? Bot. T. Linn. Soc. 180, 1–5 (2016).
30. Langham, R. J. et al. Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 166, 935–945 (2004).
31. Ni, Z. et al. Altered circadian rhythms regulate growth vigor in hybrids and allopolyploids. Nature 457, 327 (2009).
32. de Jesus, O. N. et al. Genetic diversity and population structure of Musa accessions in ex situ conservation. BMC Plant Biol. 13, 41 (2013).
33. D’Hont, A. et al. The interspecific genome structure of cultivated banana, Musa spp. revealed by genomic DNA in situ hybridization. Theor. Appl. Genet. 100, 177–183 (2000).
65. Kent, W. J. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002).

Competing interests
The authors decline no competing interests.
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41477-019-0452-6.

Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.D., W.H. or Z.J.
Peer review information: Nature Plants thanks M. Schranz, Robert VanBuren and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature Limited 2019

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

- Homolog data (homolog species genome and gene set) was downloaded from public databases: NCBI, JGI, Uniprot, KEGG, GO and NR.
- Genome data of Musa acuminata was downloaded from http://banana-genome-hub.southgreen.fr/. Genomes of Musa balbisiana (DH-PKW) and nine different accessions, and transcriptome data of Baxijiao and Fenjiao were sequenced by ourselves in this research.

Data analysis

- We used lots of software for data analysis in this paper, and all data and software used was described in Methods section of manuscript.
- Genomic assembly: wtdbg v1.2.8, SSPACE v3, Arrow, LACHESIS, HiC-Pro v2.8.1
- Evaluation of assembly quality: BUSCO v3, BLAT v35, BWA v0.7.12
- Repeat annotation: RepeatMasker v4.0.6, RepeatProteinMask, Repbase v21.01, Piler v1.0, RepeatScout v1.0.5, LTR-FINDER v1.0.5
- Tandem Repeats Finder v4.09
- Gene structure annotation: Blast v2.2.26, Augustus v3.2.1, SNAP, HISAT2 v2.0.1-beta, StringTie v1.2.1, PASA_lite, MAKER v3.31.8
- Genome annotation completeness: BUSCO v3
- Gene function annotation: BLAST v2.2.26, InterProScan v5.16
- ncRNA annotation: tRNAscan-SE v1.23, INFERNAL. BLAST v2.2.26
- Gene family analysis: OrthoMCL v1.4, CAFE v2.1, blast v2.2.26
- Phylogenetic analysis: PAML package v4.4, MrBayes v3.1.2
- Transcription factor prediction: iTAK v1.5
- Resequencing analysis: bwa 0.7.12, GATK v3.3.0, Breakdancer
- Nucleotide diversity: VCFtools v0.1.13
- Genome syntetic analysis: blast v2.2.26, MCscan v1.5.1, Kaks_Calculator v2.0
- Gene co-expression network analysis: R platform v3.2.2, WGCNA package v1.47
- RNA-Seq analysis: SOAPaligner/SOAP2 v2.2.1, DESeq2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequence reads for gene assembly and gene expression for all samples were deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under the BioProject (PRJNA432894). Genome assembly and annotation of DH-PKW was submitted to NCBI (PYDT00000000). A full data availability statement is included in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
One double haploid (DH-PKW) sample from M. balbisiana was used for genome assembly. Nine different genotypes (AAA, ABB, AA, BB, AAB) of banana accessions were used for resequencing. Two cultivated varieties of BaXijiao and FenJiao were used for transcriptomic analysis and forty samples were collected from different tissues and treatments including development and postharvest ripening process of fruits, and osmotic, salt and low temperature treatments. Two biological replicates were used for each sample.

Data exclusions
N/A

Replication
For gene expression profiling of BaXijiao and FenJiao, we produced RNA-seq data of fruits, roots, and leaves with two biological replicates.

Randomization
N/A

Blinding
N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems  Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☑   | Antibodies            |
| ☑   | Eukaryotic cell lines |
| ☑   | Palaeontology         |
| ☑   | Animals and other organisms |
| ☑   | Human research participants |
| ☑   | Clinical data         |

| n/a | Involved in the study |
|-----|-----------------------|
| ☑   | ChIP-seq              |
| ☑   | Flow cytometry        |
| ☑   | MRI-based neuroimaging |