This work was funded by European Research Council (DOUBLE-TROUBLE 833522).

Origin and evolution of the triploid cultivated banana genome

Xiuxiu Li[1#], Sheng Yu[2#], Zhihao Cheng[3#], Yingzi Yun[1#], Xuequn Chen[1], Xiaohui Wen[4], Hua Li[5], Wenjun Zhu[6], Shiyao Xu[5], Yanbing Xu[5], Xianjun Wang[1], Chen Zhang[6,7], Qiong Wu[3], Xingtan Zhang[2], Zhenguo Lin[8], Jean-Marc Aury[9], Yves Van de Peer[10, 11,12*], Zonghua Wang[1,7*], Xiaofan Zhou[13*], Peitao Lü[1, 5*], Liangsheng Zhang[1,4, 14*]

[1]Haixia Institute of Science and Technology, State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, College of Plant Protection, Fujian Agriculture and Forestry University, Fuzhou 350002, China.

[2]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

[3]Haikou Experimental Station, Chinese Academy of Tropical Agricultural Sciences, Haikou 571101, China

[4]Genomics and Genetic Engineering Laboratory of Ornamental Plants, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China

[5]College of Horticulture, Fujian Agriculture and Forestry University, Fuzhou 350002, China.

[6]College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China.

[7]Fuzhou Institute of Oceanography, Minjiang University, Fuzhou 350108, China.

[8]Department of Biology, Saint Louis University, St. Louis, MO, United States.

[9]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

[10]Department of Plant Biotechnology and Bioinformatics, Ghent University, and VIB Center for Plant Systems Biology, Ghent, 9052, Belgium.

[11]Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, 0028, South Africa.

[12]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, 210095, China.
[13]Guangdong Laboratory for Lingnan Modern Agriculture, Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Center, South China Agricultural University, Guangzhou 510642, China
[14]Hainan Institute of Zhejiang University, Sanya 572025, China

[#] Shared first authors

***Corresponding authors:** E-mail: yves.vandepeer@psb.ugent.be, wangzh@fafu.edu.cn, xiaofan_zhou@scau.edu.cn, ptlv@fafu.edu.cn, zls83@zju.edu.cn

Abstract

Bananas (*Musa* ssp.) are not only a major staple crop in tropical and sub-tropical regions, but also the most productive species of fruit tree in the world. Most fresh bananas belong to the Cavendish and Gros Michel subgroups (both triploids, AAA genome). Here, we report chromosome-scale genome assemblies of Cavendish (1.48 Gb) and Gros Michel (1.33 Gb), defining three sub-genomes, Ban, Dh, and Ze, with *M. acuminata* ssp. *banksii*, *malaccensis*, and *zebrina* as their major ancestral contributors, respectively. Based on high-quality genome assemblies of Cavendish and Gros Michel, the insertion of repeat sequences in the *Fusarium oxysporum* f. sp. *cubense* (*Foc*) tropical race 4 (TR4) resistance gene analog 2 (*RGA2*) promoter were identified in most diploid and triploid bananas. The absence of the receptor like protein (RLP) locus in the Gros Michel Ze sub-genome leads to the lack of key *Foc* race 1-resistant genes. We identified two NAP (NAC-like, activated by apetala3/pistillata) transcription factor homologs retained after polyploidy in the ancestor of Musaceae, which are highly expressed, bind directly to the promoters of many fruit ripening genes, and are critical for fruit ripening. Our data reveal the banana cultivars' origins, disease resistance and fruit ripening mechanism and should facilitate molecular breeding and super-domestication of bananas.

Main

Bananas (*Musa* ssp.) are large perennial herbs that are not only a staple crop, but also the most productive fruit in the world

(https://www.statista.com/statistics/264001/worldwide-production-of-fruit-by-variety/). Most modern cultivated bananas originated from natural hybridization between *M. acuminata* (A genome, 2n = 22) and *M. balbisiana* (B genome, 2n = 22)[1,2]. Genome assemblies of several A- and B-genome bananas have provided insights into the genetic diversity and functional divergence of polyploid banana sub-genomes[1,3-14]. However, no high-quality triploid banana reference genome has been reported, although triploid bananas are the predominant cultivars and fresh bananas, including the Cavendish and Gros Michel subgroups (both AAA genome)[15].

Most fresh bananas belong to the Cavendish, which is the most important cultivar. Before Cavendish bananas became so popular, the Gros Michel cultivar was the most popular type of banana. However, the Fusarium wilt pathogen *F. oxysporum* f. sp. *cubense* (*Foc*) race 1 has led to the near complete replacement of Gros Michel with Cavendish, which is resistant to *Foc* race 1. Recently though, the Cavendish cultivar has been seriously threatened by *Foc* tropical race 4 (TR4), suggesting this subgroup could become threatened by virtual extinction[16,17]. Gros Michel bananas have a rich creamy texture and are tastier than Cavendish bananas. The Cavendish cultivar has less sweet fruit and a thinner peel than Gros Michel, leading to its greater susceptibility to bruising and a shorter shelf life. The origin of the three A sub-genomes in the cultivated banana triploids Gros Michel and Cavendish has been unclear, and high-quality reference genomes of cultivated bananas are needed to understand the genomic ancestry of current triploid cultivars, which will be essential to guide the selection of parents in banana breeding programs, such as developing disease-resistant, shelf-stable, and flavorful bananas.

Results

Cavendish and Gros Michel genome assembly

We generated two high-quality genome assemblies for the two representative AAA triploid varieties: *M. acuminata* cv. Cavendish and *M. acuminata* cv. Gros Michel. The Cavendish genome was sequenced using a combination of PacBio sequencing, Illumina sequencing, and Hi-C technologies (Supplementary Fig. 1 and Supplementary Table 1). The Gros Michel genome was sequenced using the PacBio HiFi sequencing method (Supplementary Table 1). The Cavendish and Gros Michel genome assemblies possess

6,765 contigs (N50 = 241.2 kb) and 6,423 contigs (N50 = 1,038.0 kb) spanning 1.48 Gb and 1.33 Gb, respectively (Extended Data Table 1). A total of 106,540 and 120,653 high-confidence protein-coding genes were predicted for Cavendish and Gros Michel, respectively (Extended Data Table 1). The completeness of the Cavendish and Gros Michel assemblies was estimated to be 97.0% and 96.9% using single-copy and conserved genes (BUSCO), respectively (Extended Data Fig. 1a). The high rate of duplicated genes reported by BUSCO (Cavendish, 82.7%; Gros Michel, 87.7%) indicated that most sequences were retained in multiple copies, largely due to the auto-triploid or allo-triploid nature of the genomes (Extended Data Fig. 1a). Compared with the Cavendish genome (Cavendish*) assembled using short reads[29], our Cavendish assembly showed significant improvements in assembled size (Cavendish vs. Cavendish*, 1.48 Gb vs. 0.96 Gb), completeness (BUSCO of Cavendish vs. Cavendish*, 97.0% vs. 21.8%) and contiguity (Cavendish vs. Cavendish* for contigs N50, 241.2 kb vs. 1.4 kb) (Extended Data Fig. 1a). Several large complex regions composed of multiple resistance (*R*) genes, such as the nucleotide-binding site-leucine-rich repeat (NBS-LRR), receptor like protein (RLP), and receptor like kinase (RLK), were assembled in our Cavendish assembly, but not in the Cavendish* genome (Extended Data Fig. 4a-d). In addition, we assembled a new high-quality *M. acuminata* ssp. *zebrina* genome (v2.0) based on nanopore long-reads (Supplementary Note 2) for comparative analysis in this study.

Using Hi-C data, 1.23 Gb (83.4%) of Cavendish and 1.32 Gb (99.1%) of Gros Michel contig sequences were anchored onto 33 chromosomes (Supplementary Fig. 1 and Extended Data Table 1). All ancestor-specific K-mers of both Cavendish and Gros Michel were traced back to five possible wild diploid ancestors, namely *M. acuminata* ssp. *banskii* (Banskii), *malaccensis* (DH-Pahang), *zebrina* (Zebrina), *burmannica* (Calcutta 4) and *M. schizocarpa*. Three sub-genomes, uncovered in both Cavendish and Gros Michel, defined as Ban (Banksii), Dh (DH-Palang), and Ze (Zebrina), were found to be the top donors. Macro-syntenic comparisons of the genomes of Cavendish and Gros Michel with the haploid genomes of the wild diploid Banskii and DH-Pahang revealed that the triploid and haploid genomes are collinear, with a 3:1 correspondence across the three sub-genomes (Extended Data Figs. 1b, 2 and 3).

Origin of AAA triploid banana

A phylogeny of the three triploid sub-genomes and their four possible wild ancestors (Banskii, DH-Pahang, Zebrina and Calcutta 4) was established using *M. schizocarpa* as the outgroup. The result indicated that the sub-genomes Ban, Dh, and Ze are most closely related to Banskii, DH-Pahang and Zebrina, respectively (Fig. 1a). We estimated synonymous substitution rates (Ks) between AAA–AA paired coding sequences to identify the closest ancestor of each sub-genome. The smallest Ks peaks were found in paired species of Ban vs. Banksii (Cavendish, 0.004; Gros Michel, 0.005), Dh vs. DH-Pahang (Cavendish, 0.010; Gros Michel, 0.011), and Ze vs. Zebrina (Cavendish, 0.005; Gros Michel, 0.005) (Fig. 1b–d). The peak Ks values between the three sub-genomes and *M. schizocarpa* and Calcutta 4 ranged from 0.018–0.020 and 0.015–0.018 (Fig. 1b–d), suggesting that neither *M. schizocarpa* nor Calcutta 4 was the primary contributor to Cavendish and Gros Michel.

To map the origins of chromosomal segments, both Cavendish and Gros Michel assemblies were split into 2-kb fragments and aligned to the genome assemblies of Banksii, DH-pahang, Zebrina, Calcutta 4, and *M. schizocarpa* (Fig. 1e,f and Supplementary Table 2). In the Ban sub-genome, most segments (Cavendish: ~370.4 Mb, 85.22%; Gros Michel: ~357.5 Mb, 78.16%) are concordant with Banksii. Similarly, most segments in the Dh sub-genomes of Cavendish (~322.7 Mb, 87.82%) and Gros Michel (~381.8 Mb, 91.76%) appear to be derived from DH-Pahang. The total length of the Ze sub-genome segments assigned to Zebrina is ~371.9 Mb (85.82%) for Cavendish and ~358.2 Mb (80.23%) for Gros Michel. In addition, a few segments of Cavendish (~16.7 Mb) and Gros Michel (~26.0 Mb) were found to be assigned to *M. schizocarpa*, which is consistent with *M.schizocarpa* (SS genome) as another possible introgression of Cavendish and Gros Michel[30]. Together the results of genome segment comparison, phylogenetic analysis, and distribution of Ks values are consistent with previous reports that the three diploids are the main contributors to the A-genome of cultivated bananas[8,18-20].

Resistance genes/QTLs for Fusarium wilt

The Fusarium wilt pathogen *Foc* TR4 affects more than 80% of banana cultivars, with the Cavendish subgroup suffering particularly severely[21]. Several genes conferring

Foc TR4 resistance have been cloned in bananas[21], including the NBS-LRR gene *RGA2* cloned from the *Foc* TR4-resistant banana plants *M. acuminata* ssp. *malaccensis* accession 850, which has been transformed into Cavendish varieties to promote *Foc* TR4 resistance[22,23]. Here, we identified *RGA2* in Cavendish, Gros Michel, and several wild diploids (Fig. 2a). *RGA2* was found to be a single-copy gene in all triploid sub-genomes and diploids, and is highly conserved at the sequence level, with high amino-acid sequence identity (>97%) and coverage (>99%). We examined the 1 kb upstream promoter region of *RGA2* and found that repeat-sequence insertions of >200 bp were prevalent in many diploid and triploid bananas (Fig. 2a and Supplementary Fig. 2).

Foc race 1 has devastated large areas of Gros Michel plantations[24]. However, few *Foc* race 1-resistant genes/loci have been identified[25], and the genetic basis of Cavendish resistance to *Foc* race 1 strains is unknown. A QTL (the RLP locus) associated with *Foc* race 1 resistance has been reported, which contains a *RLP* gene cluster[25]. We performed a comparative analysis of this RLP locus between Cavendish, Gros Michel, and their three ancestors. Each sub-genome of Cavendish has one RLP locus, containing 4 (Ban), 13 (Dh), and 15 (Ze) *RLP* genes (Fig. 2b–d, Extended data Fig. 5 and Supplementary Table 3). Two of the three RLP loci are also present in the Ban and Dh sub-genomes of Gros Michel, and most *RLP* genes have one-to-one orthologous relationships with *RLP* genes in the Cavendish Ban and Dh sub-genomes (Fig. 2b,c). However, the Ze sub-genome RLP locus is absent in the Gros Michel Ze sub-genome (Fig. 2d). The Cavendish Ze sub-genome contains at least four Cavendish-specific RLP alleles that are absent in all sub-genomes of *Foc* race 1-susceptible Gros Michel (Fig. 2d, Extended data Fig. 5 and Supplementary Table 3).

Genes controlling banana fruit ripening

Compared with other climacteric fruits such as tomato and peach, banana ripening involves two positive feedback loops, with the NAC transcription factor (TF) *MaNAP* (*M. acuminata* NAC-like, activated by apetala3/pistillata) being the coupling node between the two loops[26]. Here, we found five and four *MaNAP* homologues in Cavendish and Gros Michel, respectively (Fig. 3a and Supplementary Fig. 3). They are distributed on two clades (named Clade A and B), which are preserved due to the

polyploidy of the *Musaceae* ancestor (Fig. 3a, Supplementary Fig. 3 and Supplementary Fig. 4). In Clade A, the three Cavendish genes *MaNAP1-3* are orthologues of *MaNAP* (Fig. 3a and Supplementary Fig. 3), and these genes were induced during both fruit ripening and leaf senescence (Fig. 3b). However, in Clade B, the two other Cavendish *NAP* homologues, *MaNAP4* and *MaNAP5*, were specifically expressed at a high level during fruit ripening, while their expression level was very low in leaves (Fig. 3c). And the same expression pattern of *MaNAP4*/*5* were also found in Fenjiao (ABB) (Supplementary Fig. 5). Since Cavendish should have three copies after triploidization, one copy was lost in the Ze sub-genome after triploidization.

To identify the genes to which MaNAP4/5 bind, we performed Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) assay in ripening Cavendish fruit tissues using a MaNAP4/5-specific antibody. We defined a *de novo* binding motif with high sequence identity to the known NAC motif (Fig. 3d), and identified 16,997 binding sites, which were associated with 8,907 genes (Fig. 3e,g and Supplementary Table 4). Many of the genes directly bound by MaNAP4/5 were highly expressed in ripe fruit tissues, with promoter chromatin becoming accessible during ripening (Fig. 3f), suggesting MaNAP4 and MaNAP5 play a key role in banana fruit ripening. We infer that these genes are directly regulated by MaNAP4/5 and are key to the fruit ripening process (Extended data Fig. 6). The genes include those involved in ethylene biosynthesis, and a large number of well-known ripening-related genes, such as those involved in fruit characteristic pigments synthesis (*lycopene β-cyclase* (*LCYB*)), cell wall modifications (*expansin* (*EXP*)), starch to sugars conversion (*α-amylase* (*AMY*), *β-amylase* (*BMY*) and *invertase* (*INV*)) and aroma volatiles (Fig. 3g and Supplementary Table 5). We built co-expression networks of MaNAP4/5-binding genes and identified four key ripening-related modules, including 1,602 genes (M2-M5, Supplementary Fig. 6 and Supplementary Table 5). We identified 135 genes (fold change (peel stage 4/old leaf) >10) specifically and highly expressed during fruit ripening, including 24 known genes, such as ethylene-biosynthesis and ripening-related genes, and 111 new genes (unknown function genes) may be involved in fruit ripening, such as glycoside hydrolase family 17, plant invertase/pectin methylesterase inhibitor (PMEI),

Homeodomain-Leucine Zipper (HD-ZIP) transcription factors (Supplementary Table 5). Our results suggest that these 135 genes, including many new genes, are critical for fruit ripening.

Sub-genome dominance in triploid bananas

In polyploids, one of the sub-genomes, referred to as the dominant sub-genome, can have significantly greater gene content and higher homoeolog expression[27]. We found the Ban sub-genome to exhibit significant dominance, with more retained ancestral genes, higher homoeolog expression, and more DNase-hypersensitive sites (DHS) compared with the other two sub-genomes (Extended data Fig. 7a–c, Supplementary Fig. 7, Supplementary Tables 6–8, and Supplementary Note 3). We also investigated whether MaNAP4/5 binding, as revealed by our ChIP-Seq assay, was biased among sub-genomes. The Ban sub-genome possessed more MaNAP4/5 binding sites (6,989) and associated genes (3,650) than the other sub-genomes (the Dh and Ze sub-genome have only 4,510 and 5,095 binding sites, and 2,426 and 2,750 associated genes, respectively) (Extended data Fig. 7e,f and Supplementary Table 4). The fraction of motifs bound by MaNAP4/5 was small and varied across sub-genomes (Ban: 299,723, Dh: 259,777, and Ze: 300,922; Extended data Fig. 7d), implying that other factors are involved in MaNAP4/5 binding. In addition, we did not find a clear relationship between the number of MaNAP4/5 binding sites in promoter regions and the expression patterns of homoeologs in the dominant and suppressed triads, although more binding sites were found in the dominant homoeologs of Ban and Dh (χ^2 P = 0.001; Supplementary Fig. 8). Of the MaNAP4/5 binding genes, 20% were up-regulated in at least one stage during fruit ripening, and more up-regulated genes belong to the Ban sub-genome (718, 39.1%) than either the Dh (523, 28.4%) or Ze sub-genomes (597, 32.2%) (Extended data Fig. 7f and Supplementary Table 5). These results suggest that Ban plays a dominant role in the regulation of fruit ripening.

Because banana production is threatened by several agricultural diseases, the major family of plant resistance genes was analyzed using the RGAugury pipeline[28]. In the Cavendish and Gros Michel subgroups, we identified 291 and 177 *NBS-LRR* genes, 226 and 206 *RLP* genes, and 1,836 and 1,774 *receptor like kinase* (*RLK*) genes, respectively (Extended data Fig. 7g, Supplementary Fig. 9 and Supplementary Table

9). All *R* genes are biased toward the Ze sub-genome of both Cavendish and Gros Michel (Extended data Fig. 7g and Supplementary Table 9). These results indicate that the Ze sub-genome modulates disease resistance more than the other two sub-genomes.

Discussion

Here, we identified *M. acuminata* ssp. *banksia* (Ban), *malaccensis* (Dh) and *zebrina* (Ze) as major contributors to the three sub-genomes of cultivated bananas, with the contribution of the Ban sub-genome to fruit ripening and the Ze sub-genome to disease resistance suggesting sub-genome functional divergence in triploid bananas. We compared the *Foc* TR4 resistance gene *RGA2* between cultivars and wild species and found that many bananas had a >200 bp of repeat sequence insertion upstream of *RGA2*. The degree of *Foc* TR4 protection was found to be strongly correlated with the expression level of *RGA2*[23]. Our results provide a direction to unravel the molecular mechanisms underlying the variation in expression level of endogenous *RGA2* between TR4-resistant and -susceptible bananas. Furthermore, we found that the loss of the RLP locus in the Ze sub-genome of Gros Michel leads to the lack of key *Foc* 1-resistant genes, which partially explains the susceptibility of Gros Michel to *Foc* 1. This RLP locus is probably derived from the Zebrina genome. The Ze sub-genomes of Cavendish and Gros Michel may have derived from different wild ancestors, or that the Ze sub-genome of Gros Michel has lost *Foc* 1-resistant RLP locus.

We find two novel NAP homologs (*MaNAP4* and *MaNAP5*) highly and specifically expressed in fruit, which bind to known fruit ripening-related genes (e.g. *ACS* and *EXP*) and many unknown function genes, and suggest many unknown function genes are critical for fruit ripening. *MaNAP* orthologues (*MaNAP1-3*) were induced during both fruit ripening and leaf senescence, while *MaNAP4/5* were specifically expressed at high levels during fruit ripening. We hypothesize that *MaNAP4/5* may be specifically involved in the positive feedback dual-loop[26] for banana fruit ripening.

The two high-quality AAA genome assemblies should serve as references for the application of functional genomics and comparative genome analysis to identify, clone, and characterize genes responsible for agronomic traits including fruit quality and disease resistance. Our results provide candidate genes for the improvement of

agriculturally important traits and even *de novo* domestication of polyploid bananas by focusing selection on transcriptionally dominant genes or sub-genomes.

References

1. Rouard, M. *et al.* Three new genome assemblies support a rapid radiation in *Musa acuminata* (wild banana). *Genome Bio. Evo.* **10**, 3129-3140 (2018).
2. Langhe, E.D., Vrydaghs, L., Maret, P.D., Perrier, X. & Denham, T. Why bananas matter: An introduction to the history of banana domestication. *Ethnobotany Research and Applications* **7**, 322-326 (2008).
3. D'Hont, A. *et al.* The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213-217 (2012).
4. Wang, Z. *et al. Musa balbisiana* genome reveals subgenome evolution and functional divergence. *Nat. Plants* **5**, 810-821 (2019).
5. Davey, M.W. *et al.* A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids. *BMC Genomics* **14**, 683 (2013).
6. de Jesus, O.N. *et al.* Genetic diversity and population structure of *Musa* accessions in ex situ conservation. *BMC Plant Biol.* **13**, 41 (2013).
7. Martin, G. *et al.* Genome ancestry mosaics reveal multiple and cryptic contributors to cultivated banana. *Plant J.* **102**, 1008-1025 (2020).
8. Kallow, S. *et al.* Maximizing genetic representation in seed collections from populations of self and cross-pollinated banana wild relatives. *BMC Plant Biol.* **21**, 415 (2021).
9. Martin, G. *et al.* Chromosome reciprocal translocations have accompanied subspecies evolution in bananas. *Plant J.* **104**, 1698-1711 (2020).
10. Baurens, F.C. *et al.* Recombination and large structural variations shape interspecific edible bananas genomes. *Mol. Biol. Evol.* **36**, 97-111 (2019).
11. Martin, G. *et al.* Evolution of the banana genome (*Musa acuminata*) is impacted by large chromosomal translocations. *Mol. Biol. Evol.* **34**, 2140-2152 (2017).
12. Belser, C. *et al.* Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Commun. Biol.* **4**, 1047 (2021).
13. Belser, C. *et al.* Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **4**, 879-887 (2018).

14. Cenci, A. *et al.* Unravelling the complex story of intergenomic recombination in ABB allotriploid bananas. *Ann. Bot.* **127**, 7-20 (2021).
15. Lescot, T. Genetic diversity of banana in figures. *FruiTrop* **189**, 58-62 (2008).
16. Stokstad, E. Banana fungus puts Latin America on alert. *Science* **365**, 207-208 (2019).
17. Maxmen, A. CRISPR might be the banana's only hope against a deadly fungus. *Nature* **574**, 15 (2019).
18. Carreel, F. *et al.* Ascertaining maternal and paternal lineage within *Musa* by chloroplast and mitochondrial DNA RFLP analyses. *Genome* **45**, 679-692 (2002).
19. Xavier, P. Combining biological approaches to shed light on the evolution of edible bananas. *Ethnobotany Research and Applications* **7**, 199-216 (2009).
20. Christelová, P. *et al.* Molecular and cytological characterization of the global *Musa* germplasm collection provides insights into the treasure of banana diversity. *Biodiver. Conserv.* **26**, 801-824 (2017).
21. Wang, X., Yu, R. & Li, J. Using genetic engineering techniques to develop banana cultivars with *Fusarium wilt* resistance and ideal plant architecture. *Front. Plant Sci.* **11**, 617528 (2020).
22. Stokstad, E. GM banana shows promise against deadly fungus strain. *Science* **358**, 979 (2017).
23. Dale, J. *et al.* Transgenic Cavendish bananas with resistance to Fusarium wilt tropical race 4. *Nat. Commun.* **8**, 1496 (2017).
24. Tripathi, L., Ntui, V.O. & Tripathi, J.N. CRISPR/Cas9-based genome editing of banana for disease resistance. *Cur.r Opin. Plant Biol.* **56**, 118-126 (2020).
25. Ahmad, F. *et al.* Genetic mapping of *Fusarium wilt* resistance in a wild banana *Musa acuminata* ssp. *malaccensis* accession. *Theor. Appl. Genet.* **133**, 3409-3418 (2020).
26. Lü, P. *et al.* Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. *Nat. Plants* **4**, 784-791 (2018).
27. Thomas, B.C., Pedersen, B. & Freeling, M. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* **16**, 934-946 (2006).
28. Li, P. *et al.* RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 852 (2016).

Figures

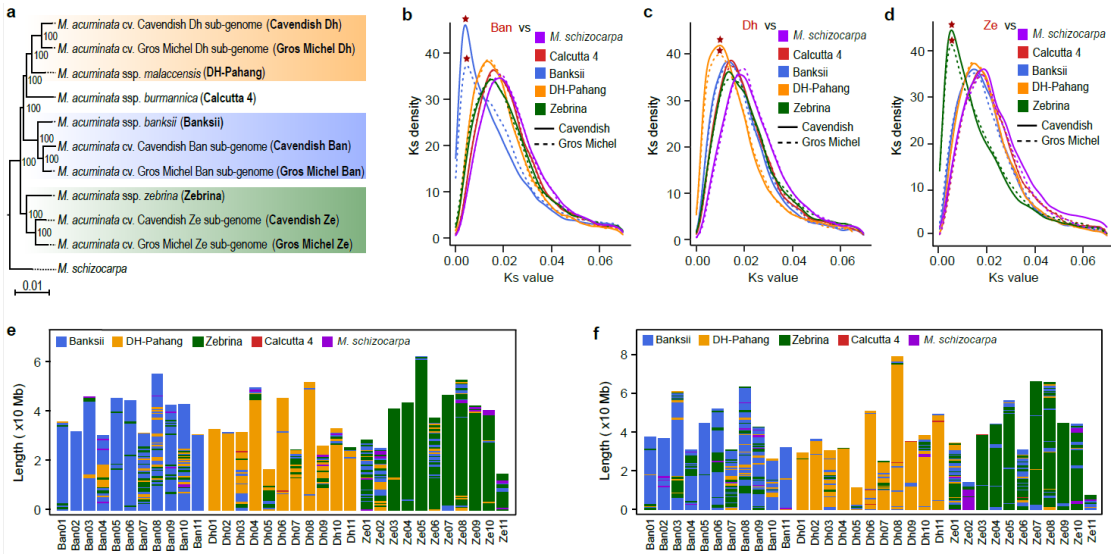

Fig. 1 | Genome evolution of AAA bananas. a, Phylogenetic tree of the three AAA triploid sub-genomes and their four possible original wild diploid ancestors (Banskii, DH-Pahang, Zebrina and Calcutta 4). *M. schizocarpa* serves as an outgroup species. **b**–**d**, Frequency distribution of synonymous substitution rate (Ks) between syntenic genes of triploid sub-genomes, including Ban (B), Dh (C), and Ze (D), compared with their four wild ancestors and *M. schizocarpa*. The paired species with the smallest average Ks were labeled with red asterisks. **e** and **f**, The ancestral origins of Cavendish (e) and Gros Michel (f). Both Cavendish and Gros Michel assemblies were split into 2 kb fragments and aligned to genomes of Banksii, DH-Pahang, Zebrina, Calcutta 4, and *M. schizocarpa*. Each segment was colored according to its best alignment with the diploid progenitor species (Banksii in blue, DH-Pahang in orange, Zebrina in green, Calcutta 4 in red, and *M. schizocarpa* in purple).

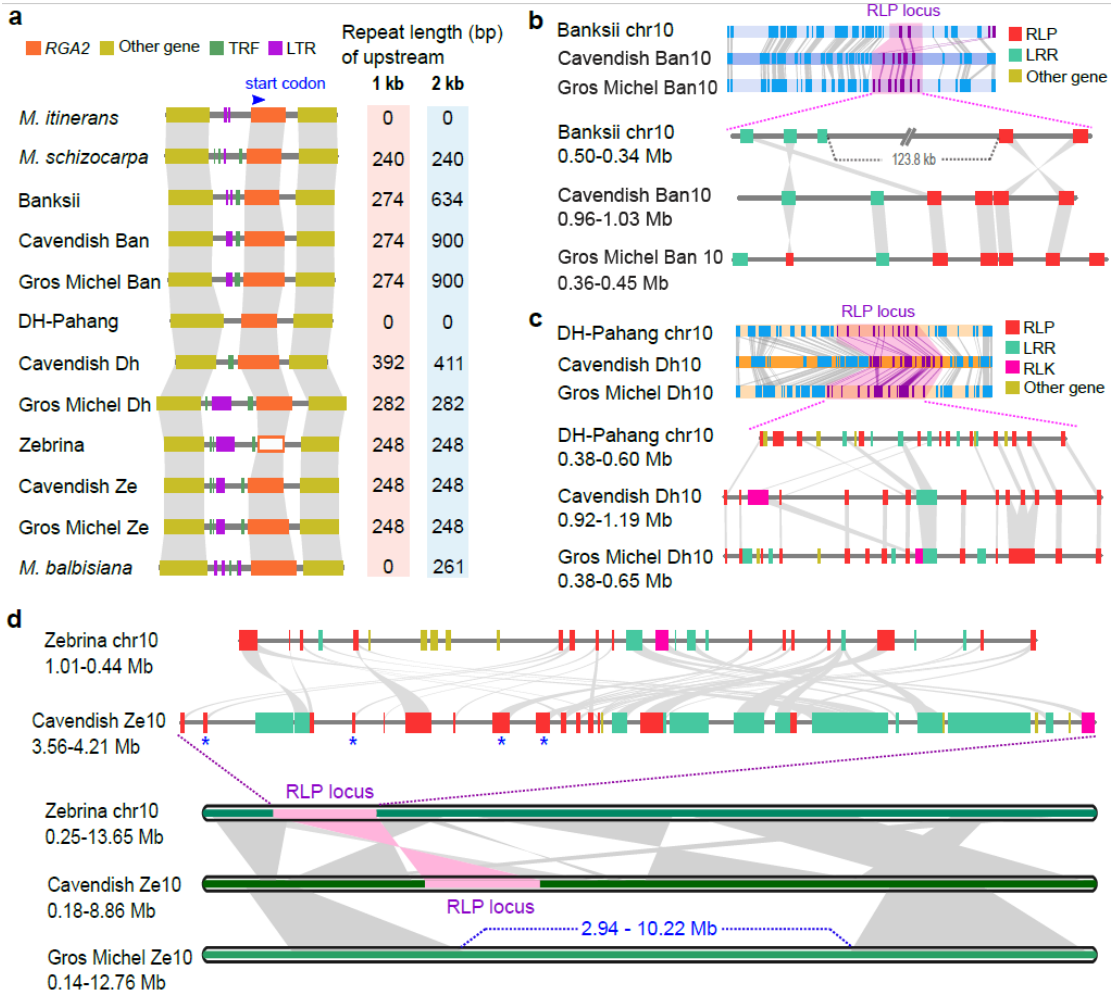

Fig. 2 | Comparative analysis of resistance genes/QTLs against *Foc* race 1 and *Foc* TR4 between Cavendish (AAA), Gros Michel (AAA) and their wild ancestors. a, Comparative analysis of the TR4-resistant gene *RGA2*[22,23]. Repeat sequences (tandem repeats (TRF) in green, retrotransposons of long-terminal-repeats (LTR) in purple) in the upstream 1 and 2 kb range were plotted and their lengths shown on the right. The abbreviations of banana species refer to Fig. 1a. **b**, Micro-synteny comparison of the RLP locus[25] among Cavendish Ban sub-genome, Gros Michel Ban sub-genome, and Banksii. **c**, Micro-synteny comparison of the RLP locus[25] among Cavendish Dh sub-genome, Gros Michel Dh sub-genome, and DH-Pahang. **d**, Comparison of RLP locus[25] among Cavendish Ze sub-genome, Gros Michel Ze sub-genome, and Zebrina. The RLP locus in the Cavendish Ze sub-genome is absent in the Gros Michel Ze sub-genome. Blue asterisks denote RLPs found only in the Ze sub-genome of Cavendish.

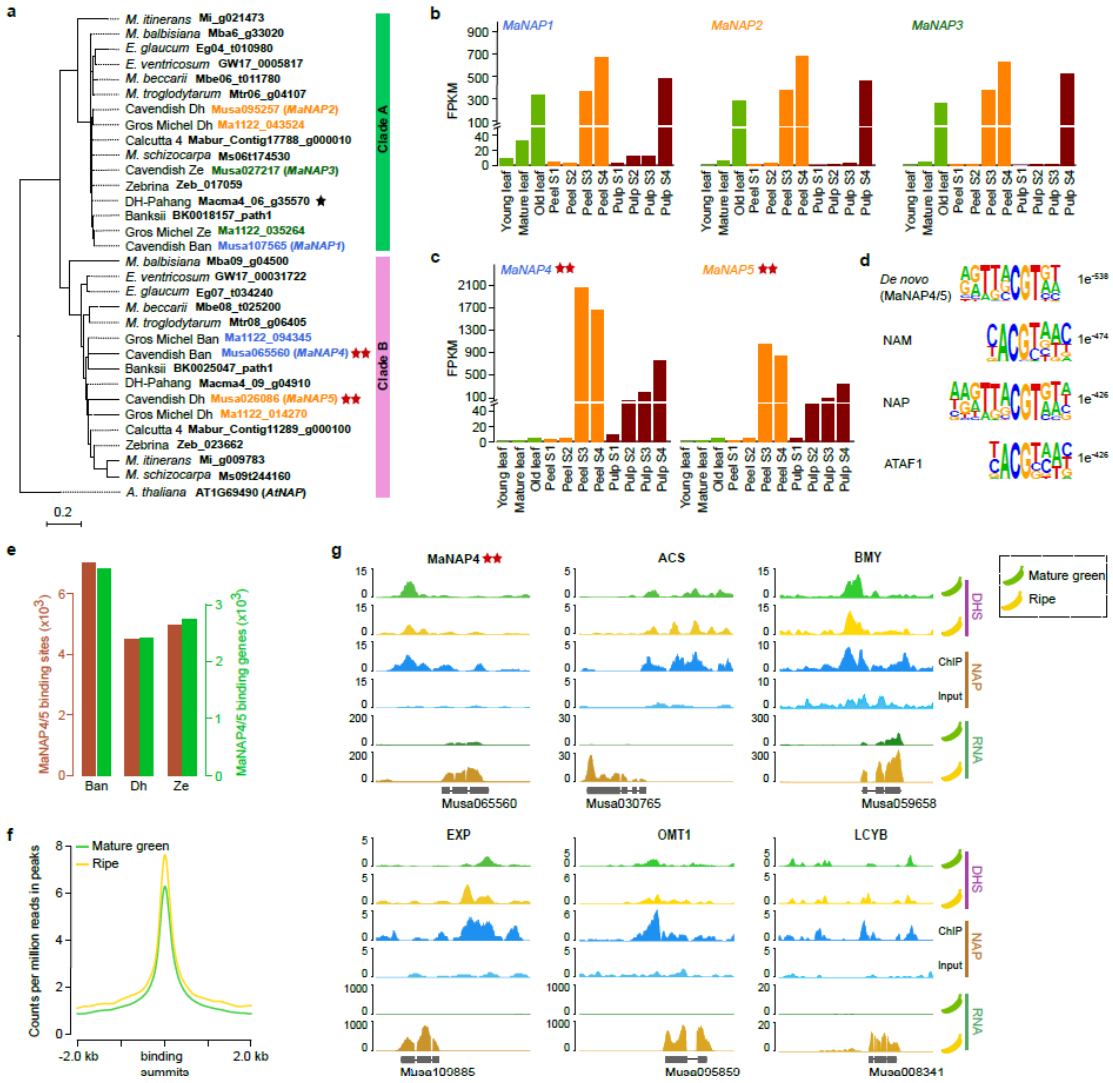

Fig. 3 | Identification of genes controlling ripening in banana. a, Phylogenetic tree of banana NAP homologues associated with climacteric fruit ripening. Black star denotes *MaNAP* in DH-Pahang reported by Lü *et. al.*[26]. Red stars denote two new NAP homologues (*MaNAP4* and *MaNAP5*) in Cavendish. The abbreviations of banana species refer to Fig. 1a. **b**, The expression patterns of *MaNAP1-3* in leaf and fruit of Cavendish cultivar Baxi. S1 to S4 represent the peel and pulp tissues at four developmental stages (stage 1 to 4: fruit set, immature, mature green and fully ripen). **c**, The expression patterns of *MaNAP4* and *MaNAP5* in leaf and fruit of Cavendish cultivar Baxi. **d**, *De novo* motif of MaNAP4/5 binding. NAM, NAP, and ATAF1 belong to NAC transcription factor family in GSE50143. **e**, Sub-genome distribution of MaNAP4/5 binding sites and genes. **f**, Changes in DNase hypersensitivity of

MaNAP4/5 binding sites in pulp tissue of mature green (stage 3) and ripe (stage 4) banana fruit. **g**, Transcription, MaNAP4/5 protein binding and DNase hypersensitivity at several banana ripening related loci. Y axis represents counts per million reads.

Methods

Genome assembly

The PacBio Sequel I sequencing data of Cavendish were assembled using Canu v1.9[29] with the "-pacbio-raw" option. The assembly was then polished with Illumina short-read sequencing data using Pilon v1.23[30] for three iterations. The PacBio HiFi sequencing data of Gros Michel were assembled using Canu v2.1.1[31] with the "-pacbio-hifi" option.

All assembled contigs were assigned to three groups based on ancestor genomic information inspired by the trio-binning algorithm[32]. We first extracted and identified ancestor-specific 27-mers from the three published ancestor banana genomes (Banksii, DH-Pahang and Zebrina). The ancestor-specific K-mers were further traced back to the assembled contigs, which were subsequently partitioned based on the counting of K-mers originating from different ancestral genomes. For instance, we consider a contig originating from Banksii only if the contig contains at least 1.5 x more Banksii-specific Kmers than the other two ancestors (DH-Pahang and Zebrina). Following this criterion, we identified 466.22 Mb, 417.45 Mb, and 447.31 Mb contig sequences in Banksii-originated, DH-Pahang-originated, and Zebrina-originated groups, respectively.

We further anchored these contig sequences in each group onto 11 chromosomes, representing each sub-genome, using two different approaches (reference-guided[33,34] and Hi-C guilded[35,36]). Initially, chimeric contigs were corrected based on abnormal Hi-C contact by ALLHiC_corrector program[35] and the ordering and orientation of these corrected contigs were determined through alignment with reference genomes (i.e., the ancestral genomes) using minimap2 with default parameters[37]. The remaining unanchored contigs were further re-assigned onto each chromosome based on Hi-C contact with anchored sequences by ALLHiC_rescue function[35]. In addition, we

optimized the orders of grouped contigs for each chromosome, resulting in a final release of chromosomal-scale genome assembly.

Evolutionary analyses

The longest proteins of each gene in seven species, including six sub-genomes of Cavendish and Gros Michel, four wild ancestors, and *Musa schizocarpa* (as outgroup), were selected. All those longest proteins were blasted against each other, then clustered with Orthofinder v2.2.7[38]. Then, 2,628 single-copy orthologous genes were used by ProtTest3.0[39], with DT standard, to select the best model (JTT+I+G), and then the phylogenetic trees were constructed with PhyML3.0[40].

Paralogous and orthologous gene pairs were identified using MCscan (Python version)[41] with default parameters. Macro- and micro-synteny relationships were identified and plotted based on the results of MCscan. Ka and Ks were calculated with the PAML yn00 NG model[42] using coding and protein sequences of orthologous gene pairs and all zero values were filtered out.

Identified homoeolog expression bias of sub-genomes

The analysis focused exclusively on the gene triads which had a 1:1:1 correspondence across the three homoeologous sub-genomes, including 18,119 syntenic triads and 54,357 homoeologs in total. We defined a triad as expressed when the sum of the Ban, Dh, and Ze sub-genome homoeologs was > 1 FPKM. To standardize the relative expression of each homoeolog across the triad, we normalized the absolute FPKM for each gene within the triad. Then, the homoeolog expression bias categories were identified as described previously for wheat[43].

Chromatin immunoprecipitation sequencing and data analysis

Banana pulp tissue at stage 4 was fixed with 1% formaldehyde in 1x PBS for 15 min under vacuum. Nuclei were purified as previously described[26]. The chromatin was sonicated to 300-500 bp in TE with 0.25% SDS and protease inhibitors using Covaris M220 and diluted with low salt wash buffer (150 mM NaCl in TE) with 1% Triton X-100. The chromatin samples were incubated 6 h with Dynabeads Protein A/G (Invitrogen) with anti-NAP (DGSSDVHYHLSRQKKP) rabbit serum. The purified DNA from the supernatant was used as input. The beads were then washed twice with

low salt buffer (150 mM NaCl in TE) and twice with high salt buffer (250 mM NaCl in TE). The washed magnetic beads were treated with Tn5 transposase at 37 °C for 30 min. The beads were then washed with low salt and TE buffer, and reverse-crosslinked for 8 h. The purified DNA was then amplified using N50x and N70x index primers. Two biological replicates were sequenced for the experiment.

Raw reads were firstly processed using trim_galore[44] to trim low quality or adapter-originated bases. Trimmed reads were then mapped to genome reference using bowtie2[45] with default parameters. Secondary or supplementary alignments, alignments with mapping quality less than 30, and improperly paired alignments were discarded. PCR duplicates were further masked using picard markduplicates[46]. The resulted alignments of two replicates were then subjected to MACS2 callpeak[47] for peak calling with the -c parameter specified as the input alignments, -g specified as 557690000, and other default parameters. Overlapping peaks between two replicates were considered as putative protein binding sites and were used for downstream analysis. Each peak was associated to the nearest gene if it located in 2 kb upstream to 500 bp downstream of transcription start site using rtracklayer in R[48].

Reporting Summary

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

Data availability

Genome assemblies, annotations, and results of ChIP-Seq and DNase-Seq can be accessed at figshare with Digital Object Identifier 10.6084/m9.figshare.21249081. Raw data used for the assemblies including PacBio, Illumina, and Hi-C data are available through the Sequence Read Archive of the National Centre for Biotechnology Information (NCBI) under the BioProject (No. PRJNA889940) with SRA accessions from SRR23425440 to SRR23425471, and from SRR23885547 to SRR23885549. 58 RNA-seq datasets were download from NCBI BioProject: PRJNA381300, PRJNA394594, and PRJNA598018. DNA methylation data were downloaded from the NCBI (BioProject: PRJNA381300).

Code availability

Custom code and scripts for mapping the origins of chromosomal segments are available at figshare with Digital Object Identifier 10.6084/m9.figshare.21249081. All public software used in this study is provided in the accompanying Nature Research Reporting Summary.

References

29. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722-736 (2017).
30. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
31. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291-1305 (2020).
32. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174-1182 (2018).
33. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
34. Schneeberger, K. *et al.* Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 10249-10254 (2011).
35. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants.* **5**, 833-845 (2019).
36. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
38. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol. Evol.* **20**, 238 (2019).
39. Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-5 (2011).
40. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).
41. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486-8 (2008).

42. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555-6 (1997).
43. Ramírez-González, R.H. *et al.* The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018).
44. Stubbs, T.M. *et al.* Multi-tissue DNA methylation age predictor in mouse. *Genome Biol.* **18**, 68 (2017).
45. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421-432 (2019).
46. Institute, B. Picard Toolkit. *Broad Institute, GitHub Repository* https://broadinstitute.github.io/picard/ (2019).
47. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
48. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841-2 (2009).

Acknowledgements

We thank Dr. Guy Riddihough (Life Science Editors) for text editing. This work was supported by the fund for Construction of Plateau Discipline of Fujian Province (102/71201801104). YVdP acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01).

Author contributions

L.Z. conceived and designed the project. P.L., Z.C., Y.Y., W.Z., S.X., Y.X. and H.L. collected the samples and extracted DNA and RNA. L.Z., P.L. and S.Y. coordinated the Illumina and PacBio sequencing. X.Zhou and X.Zhang assembled genomes. X.Zhou, C.Z. and X.Wang conducted protein-coding gene and repetitive sequence annotations. L.Z. and X.L. performed phylogenetic analyses. X.L., X.C. L.Z. performed comparative genomic analysis. X.L., X.Zhou, Q.W., and X.Wen performed the RNA-seq analysis. P.L. and S.Y. performed ChIP-Seq experiments, DNase-seq experiments and bioinformatic analysis of ChIP-Seq DNase-seq, and WGBS data. X.L., P.L., S.Y., X.Zhou and X.Zhang wrote the manuscript draft. L.Z., P.L., S.Y., X.L.,

X.Zhou, YVdP., ZG.L., Z.W. and J.A. reviewed and revised the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.