nature genetics

Article

https://doi.org/10.1038/s41588-023-01589-3

Origin and evolution of the triploid cultivated banana genome

Received: 14 November 2022

Accepted: 23 October 2023

Published online: 11 December 2023

Check for updates

Xiuxiu Li[1,15], Sheng Yu[2,15], Zhihao Cheng[3,15], Xiaojun Chang[4,15], Yingzi Yun[1,15], Mengwei Jiang[1,15], Xuequn Chen[1], Xiaohui Wen[5,6], Hua Li[1], Wenjun Zhu[1], Shiyao Xu[1], Yanbing Xu[1], Xianjun Wang[1], Chen Zhang[1,7], Qiong Wu[3], Jin Hu[5,6], Zhenguo Lin[8], Jean-Marc Aury[9], Yves Van de Peer[10,11,12] ✉, Zonghua Wang[1,7] ✉, Xiaofan Zhou[13] ✉, Jihua Wang[14] ✉, Peitao Lü[1] ✉ & Liangsheng Zhang[5,6] ✉

Most fresh bananas belong to the Cavendish and Gros Michel subgroups. Here, we report chromosome-scale genome assemblies of Cavendish (1.48 Gb) and Gros Michel (1.33 Gb), defining three subgenomes, Ban, Dh and Ze, with *Musa acuminata* ssp. *banksii*, *malaccensis* and *zebrina* as their major ancestral contributors, respectively. The insertion of repeat sequences in the *Fusarium oxysporum* f. sp. *cubense* (*Foc*) tropical race 4 *RGA2* (resistance gene analog 2) promoter was identified in most diploid and triploid bananas. We found that the receptor-like protein (RLP) locus, including *Foc* race 1-resistant genes, is absent in the Gros Michel Ze subgenome. We identified two NAP (NAC-like, activated by apetala3/pistillata) transcription factor homologs specifically and highly expressed in fruit that directly bind to the promoters of many fruit ripening genes and may be key regulators of fruit ripening. Our genome data should facilitate the breeding and super-domestication of bananas.

Bananas (*Musa* ssp.) are large perennial herbs that are not only a major staple crop in tropical and subtropical regions but also one of the most productive fruits in the world (https://www.statista.com/statistics/264001/worldwide-production-of-fruit-by-variety/). Most modern cultivated bananas originated from natural hybridization between *Musa acuminata* (A genome, $2n = 22$) and *Musa balbisiana* (B genome, $2n = 22$)[1,2]. Genome assemblies of several A-genome and B-genome bananas have provided insights into the genetic diversity and

[1]State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Haixia Institute of Science and Technology, College of Horticulture, Fujian Agriculture and Forestry University, Fuzhou, 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, China. [3]Haikou Experimental Station, National Key Laboratory for Tropical Crop Breeding, Chinese Academy of Tropical Agricultural Sciences, Haikou, China. [4]Laboratory of Medicinal Plant Biotechnology, School of Pharmaceutical Sciences, Zhejiang Chinese Medical University, Hangzhou, China. [5]Zhejiang Provincial Key Laboratory of Horticultural Plant Integrative Biology, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China. [6]Hainan Institute of Zhejiang University, Sanya, China. [7]Fuzhou Institute of Oceanography, Minjiang University, Fuzhou, China. [8]Department of Biology, Saint Louis University, St. Louis, MO, USA. [9]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France. [10]Department of Plant Biotechnology and Bioinformatics, Ghent University and VIB Center for Plant Systems Biology, Ghent, Belgium. [11]Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa. [12]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, China. [13]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, 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, China. [14]Yunnan Seed Laboratory, Kunming, China. [15]These authors contributed equally: Xiuxiu Li, Sheng Yu, Zhihao Cheng, Xiaojun Chang, Yingzi Yun, Mengwei Jiang. ✉e-mail: yvpee@psb.vib-ugent.be; wangzh@fafu.edu.cn; xiaofan_zhou@scau.edu.cn; wjh0505@gmail.com; ptlv@fafu.edu.cn; zls83@zju.edu.cn

functional divergence of polyploid banana subgenomes[1,3–14]. However, no high-quality triploid banana reference genome has been reported, although AAA triploid bananas are the predominant cultivars and fresh bananas, including the Cavendish and Gros Michel subgroups[15].

Most fresh bananas belong to the Cavendish subgroup, which is one of the most important cultivars. Before Cavendish bananas became so popular, the Gros Michel cultivar was the most popular type of banana. However, the Fusarium wilt pathogen *Fusarium 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, the Cavendish cultivar has been seriously threatened by *Foc* tropical race 4 (TR4), suggesting that 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 fruit that is less sweet and has a thinner peel than that of Gros Michel, leading to its greater susceptibility to bruising and a shorter shelf life. The origin of the three A subgenomes of the cultivated bananas 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, in order to develop disease-resistant, shelf-stable and flavorful bananas.

Results

Genome assemblies of triploid Cavendish and Gros Michel

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 (Table 1). A total of 106,540 and 120,653 high-confidence protein-coding genes were predicted for Cavendish and Gros Michel, respectively (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 (Benchmarking Universal Single-copy Orthologs (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 autotriploid or allotriploid nature of the genomes (Extended Data Fig. 1a). Compared with the previous Cavendish genome assembled using short reads[18], our Cavendish assembly showed substantial improvements in assembled size (1.48 Gb versus 0.96 Gb), contiguity for contigs N50 of 241.2 kb versus 1.4 kb) and completeness (Extended Data Fig. 1a). Several large complex regions composed of multiple resistance (*R*) genes, such as those encoding the nucleotide-binding site–leucine-rich repeat (NBS-LRR), receptor-like protein (RLP) and receptor-like kinase (RLK), were assembled in our Cavendish genome, but not in the previous 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 1) 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 Table 1).

Table 1 | Genome assembly and annotation statistics of Cavendish and Gros Michel

Statistic	Cavendish	Gros Michel
Number of contigs	6,765	6,423
Assembled genome (Mb)	1,480	1,331
Contig N50 (kb)	241.2	1,038.0
Contig N50 number	1,256	204
Max contig length (Mb)	8.2	13.2
Mean contig length (kb)	94.1	207.2
G+C content (%)	39.3	39.3
Pseudo-chromosomes	33	33
Total anchored size (Mb)	1,234	1,320
Genome in chromosome (%)	83.4	99.1
Max chromosome length (Mb)	62	78.8
Min chromosome length (Mb)	14.6	7.5
Number of genes	106,540	120,653
Genes anchored on chromosome (%)	97.1	99.8
Genes containing conserved domains (%)	70.7	59.4
Genes classified by GO terms (%)	53.8	45.2
Genes mapping to pathways (%)	15.4	16.4

Origin of AAA triploid cultivated banana

All ancestor-specific *k*-mers (Methods) 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) and *burmannica* (Calcutta 4) and *Musa schizocarpa* (S genome from New Guinea). Three subgenomes were uncovered in both Cavendish and Gros Michel and defined as Ban (Banksii), Dh (DH-Pahang) and Ze (Zebrina), given that Banksii, DH-Pahang and Zebrina were found to be the top three donors. Macrosyntenic comparisons of the genomes of Cavendish and Gros Michel with the haploid genomes of the wild diploid Banskii, DH-Pahang and Zebrina revealed that the triploid and haploid genomes are collinear, with a 3:1 correspondence across the three subgenomes (Extended Data Figs. 1b, 2 and 3).

A phylogeny of the three triploid subgenomes 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 subgenomes Ban, Dh and Ze are most closely related to Banskii, DH-Pahang and Zebrina, respectively (Fig. 1a). We estimated the synonymous substitution rates (K_s) between AAA–AA paired coding sequences to identify the closest ancestor of each subgenome. The smallest K_s peaks were found in paired species of Ban versus Banksii (Cavendish, 0.004; Gros Michel, 0.005), Dh versus DH-Pahang (Cavendish, 0.010; Gros Michel, 0.011) and Ze versus Zebrina (Cavendish, 0.005; Gros Michel, 0.005) (Fig. 1b–d). The peak K_s values among the three subgenomes and *M. schizocarpa* and Calcutta 4 ranged from 0.018 to 0.020 and from 0.015 to 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 subgenome, 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 subgenomes 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 subgenome 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* as another possible introgression of Cavendish and Gros Michel[14]. Together, the results of genome segment comparison, phylogenetic analysis and distribution of K_s values are consistent with previous reports that the three diploids are the main contributors to the A genome of cultivated bananas[8,19,20].

https://doi.org/10.1038/s41588-023-01589-3

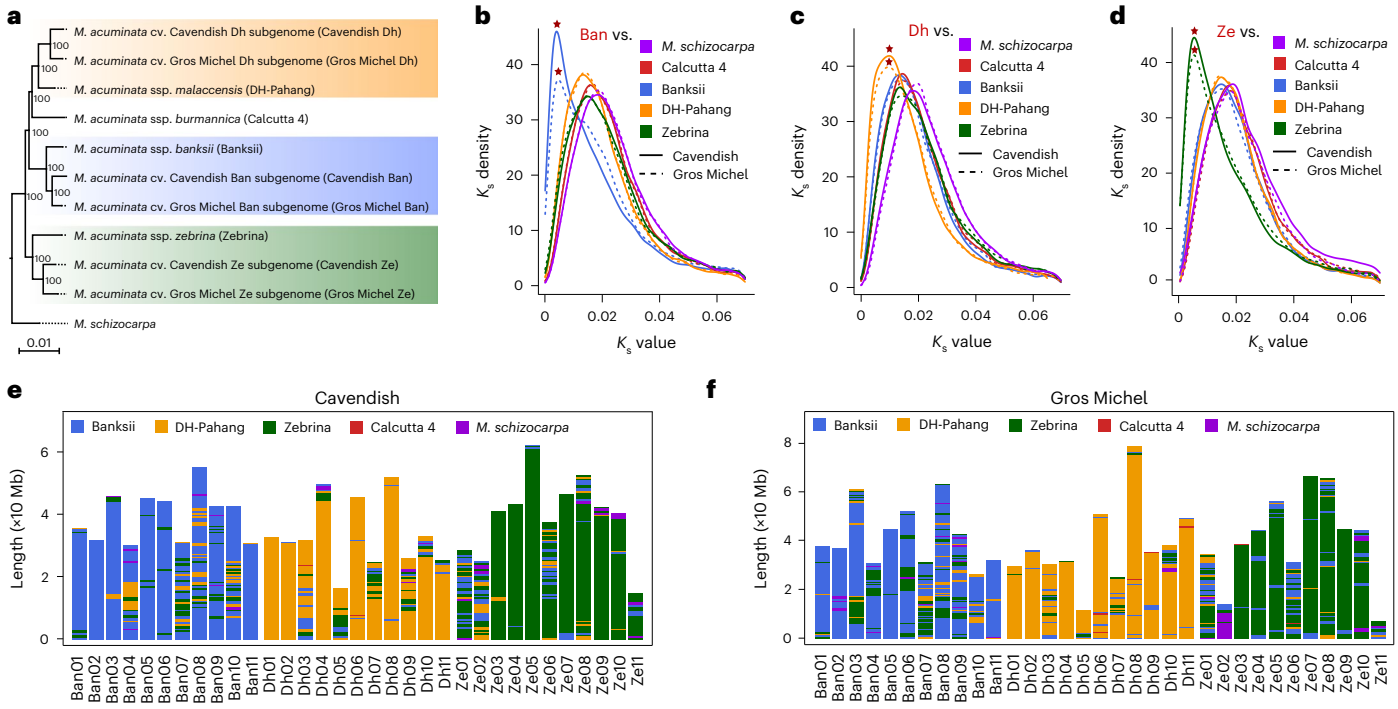


Fig. 1 | Genome evolution of AAA bananas. a, Phylogenetic tree of the three A subgenomes and their four possible original wild diploid ancestors (Banskii, DH-Pahang, Zebrina and Calcutta 4). *M. schizocarpa* serves as an outgroup species. The value above the scale bar is the number of substitutions per site. **b**–**d**, Frequency distribution of K_s among syntenic genes of triploid subgenomes, including Ban (**b**), Dh (**c**) and Ze (**d**), compared with their four wild ancestors and *M. schizocarpa*. The paired species with the smallest average K_s were labeled with red stars. **e**,**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 is colored according to its best alignment with the diploid progenitor species (Banksii, blue; DH-Pahang, orange; Zebrina, green; Calcutta 4, red; *M. schizocarpa*, purple). All diploid banana genome information is listed in Supplementary Table 10.

The evolution of Fusarium wilt resistance genes in banana

The Fusarium wilt pathogen *Foc* TR4 affects more than 80% of banana cultivars; in particular, the Cavendish subgroup suffers severely[21]. Several genes conferring *Foc* TR4 resistance have been cloned in bananas[21], including the NBS-LRR gene resistance gene *RGA2* (resistance gene analog 2) 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, including *Musa itinerans* (known as Yunnan banana, one of the most *Foc* TR4-resistant banana species) and *M. balbisiana* (B genome) (Fig. 2a). *RGA2* was found to be a single-copy gene in all triploid subgenomes and diploids, and is highly conserved at the sequence level with high amino acid sequence identity (>97%). We examined the 1-kb upstream promoter region of *RGA2* and found that repeat sequence insertions of >200 bp were prevalent in most 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 resistance genes/loci have been identified[25], and the genetic basis of Cavendish resistance to *Foc* race 1 strains is unknown. A quantitative trait locus (QTL) (the RLP locus) associated with *Foc* race 1 resistance has been reported that contains an *RLP* gene cluster[25]. We performed a comparative analysis of this RLP locus among Cavendish, Gros Michel and their three ancestors. Each subgenome 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 subgenomes of Gros Michel, and most *RLP* genes have one-to-one orthologous relationships with *RLP* genes in the Cavendish Ban and Dh subgenomes (Fig. 2b, c). However, the Ze-subgenome RLP locus is absent in the Gros Michel Ze subgenome (Fig. 2d). The Cavendish Ze subgenome contains at least four Cavendish-specific *RLP* alleles that are absent in all subgenomes 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 ethylene-dependent ripening fruits (climacteric fruits), such as tomato and peach, banana ripening involves two positive-feedback loops, with the NAC (NAM, ATAF1/ATAF2 and CUC2) transcription factor *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* homologs in Cavendish and Gros Michel, respectively (Fig. 3a and Supplementary Fig. 3). They are distributed on two clades (clade A and clade B), which are preserved due to the polyploidy of the Musaceae ancestor (Fig. 3a and Supplementary Figs. 3 and 4). In clade A, the three Cavendish genes *MaNAP1*–*MaNAP3* are orthologs 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 other two Cavendish *NAP* homologs, *MaNAP4* and *MaNAP5*, were specifically expressed at high levels during fruit ripening, while their expression levels were low in leaves (Fig. 3c). The same expression patterns of *MaNAP4* and *MaNAP5* were also found in Fenjiao (*Musa* spp. ABB) (Supplementary Fig. 5).

To identify the genes to which MaNAP4 and MaNAP5 bind, we performed chromatin immunoprecipitation followed by sequencing (ChIP–seq) assays in ripening Cavendish fruit tissues using a MaNAP4- and MaNAP5-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 or MaNAP5 were highly expressed

https://doi.org/10.1038/s41588-023-01589-3

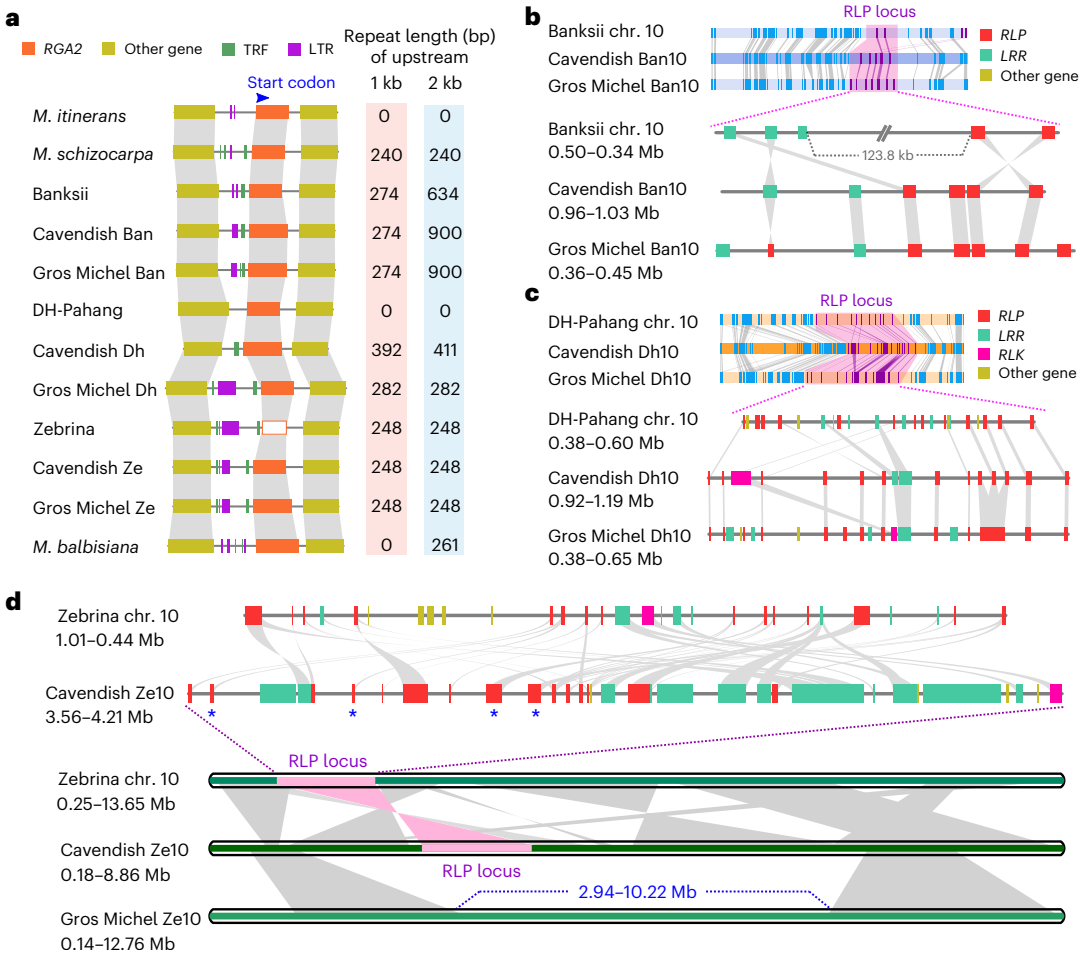


Fig. 2 | The resistance genes/QTLs among Cavendish, Gros Michel and their wild ancestors. a, Comparative analysis of the TR4-resistance gene *RGA2* (ref. 23). Repeat sequences (tandem repeats (TRF) in green, retrotransposons of long terminal repeats (LTR) in purple) in the upstream 1-kb and 2-kb range are plotted and their lengths are shown on the right. The Zebrina genome contains a nearly complete sequence of *RGA2* but lacks a start codon. *M. itinerans*, known as Yunnan banana, is one of the most *Foc* TR4-resistant banana species. *M. balbisiana* is a wild diploid that belongs to the B-genome subgroup. **b**, Microsynteny comparison of the RLP locus[25] among the Cavendish Ban subgenome, Gros Michel Ban subgenome and Banksii. **c**, Microsynteny comparison of the RLP locus[25] among the Cavendish Dh subgenome, Gros Michel Dh subgenome and DH-Pahang. **d**, Comparison of the RLP locus[25] among the Cavendish Ze subgenome, Gros Michel Ze subgenome and Zebrina. The RLP locus in the Cavendish Ze subgenome is absent in the Gros Michel Ze subgenome. Blue stars denote *RLP* genes found only in the Ze subgenome of Cavendish. All abbreviations of banana species refer to Fig. 1a, and all diploid banana genome information is listed in Supplementary Table 10.

in ripe fruit tissues, with promoter chromatin becoming accessible during ripening (Fig. 3f), suggesting that MaNAP4 and MaNAP5 play a key role in banana fruit ripening. We infer that these genes are directly regulated by MaNAP4 and MaNAP5 and are key to the fruit ripening process (Extended Data Fig. 6). The genes included those involved in ethylene biosynthesis and a large number of well-known ripening-related genes, such as those involved in fruit characteristic pigments synthesis (*LCYB* (*lycopene β-cyclase*)), cell wall modifications (*EXP* (*expansin*)), starch to sugar conversion (*AMY* (*α-amylase*)), (*BMY β-amylase*)) and (*INV (invertase*)) and aroma volatiles production (*OMT1* (*O-methyltransferase*) (Fig. 3g and Supplementary Table 5). We built co-expression networks of MaNAP4- and MaNAP5-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) that may be involved in fruit ripening, such as those encoding glycoside hydrolase family 17, plant invertase/pectin methylesterase inhibitor (PMEI) and homeodomain–leucine zipper (HD-ZIP) transcription factors (Supplementary Table 5). Our results suggest that these 135 genes, including 24 known genes and 111 new genes, are critical for fruit ripening.

Subgenome dominance in triploid bananas

In polyploids, one of the subgenomes, referred to as the dominant subgenome, can have significantly greater gene content and higher homoeolog expression[27]. We found the Ban subgenome to exhibit substantial dominance, with more retained ancestral genes, higher homoeolog expression and more DNase-hypersensitive sites (DHSs) compared with the other two subgenomes (Extended Data Fig. 7a–c, Supplementary Fig. 7, Supplementary Tables 6–8 and Supplementary Note 2). We also investigated whether MaNAP4 and MaNAP5 binding, as revealed by our ChIP–seq assay, was biased among subgenomes. The Ban subgenome possessed more MaNAP4- and MaNAP5-binding sites (6,989) and associated genes (3,650) than the other subgenomes (the Dh and Ze subgenomes have 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 and MaNAP5 was small and varied across subgenomes (Ban, 299,723; Dh, 259,777; Ze, 300,922; Extended Data Fig. 7d), implying that other factors are involved in MaNAP4 and MaNAP5 binding. Of the MaNAP4-

https://doi.org/10.1038/s41588-023-01589-3

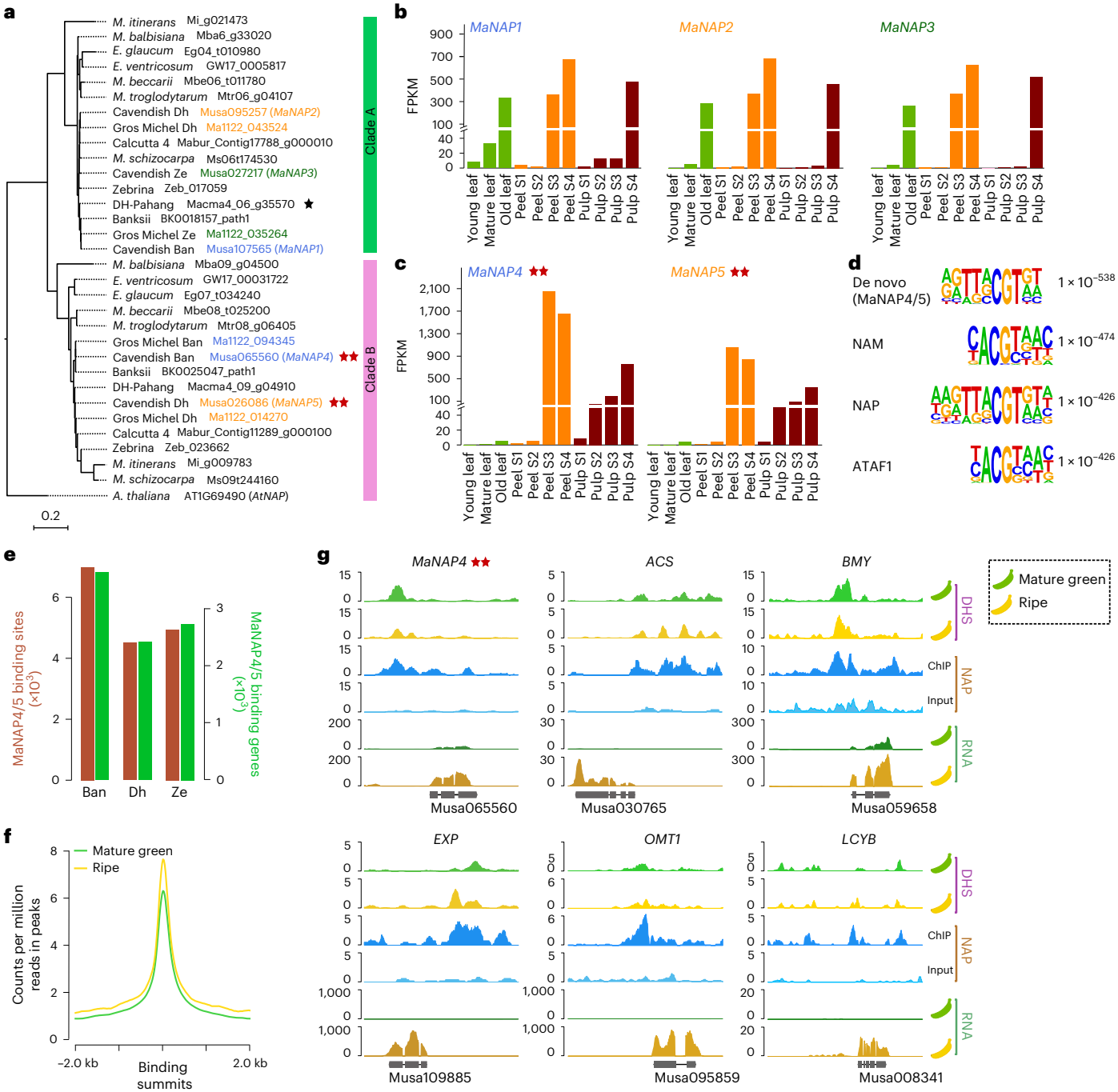


Fig. 3 | Identification of genes controlling ripening in banana. a, Phylogenetic tree of banana NAP homologs associated with climacteric fruit ripening. The black star denotes *MaNAP* in DH-Pahang reported by Lü et al.[26]. Red stars denote two new NAP homologs (*MaNAP4* and *MaNAP5*) in Cavendish. Cavendish should have three copies after triploidization, but one copy was lost in the Ze subgenome after triploidization. The abbreviations of banana species refer to Fig. 1a and all diploid banana genome information is listed in Supplementary Table 10. The value above the scale bar is the number of substitutions per site. **b**, The expression patterns of *MaNAP1*–*MaNAP3* in the leaves and fruit of Cavendish cultivar Baxi. S1 to S4 represent the peel and pulp tissues at four developmental stages (stage 1, fruit set; stage 2, immature; stage 3, mature green; stage 4, ripe). RNA-seq datasets (Sequence Read Archive (SRA)) of Baxi were downloaded from NCBI BioProject PRJNA381300. FPKM, fragments per kilobase per million mapped reads. **c**, The expression patterns of *MaNAP4* and *MaNAP5* in the leaves and fruit of Cavendish cultivar Baxi. **d**, De novo motif of MaNAP4- and MaNAP5-binding. NAM, NAP and ATAF1 belong to the NAC transcription factor family in GSE50143. **e**, Subgenome distribution of MaNAP4- and MANAP5-binding sites and genes. **f**, Changes in DNase hypersensitivity of MaNAP4- and MaNAP5-binding sites in pulp tissue of mature green (stage 3) and ripe (stage 4) banana fruit. **g**, Transcription, MaNAP4 and MANAP5 protein binding and DNase hypersensitivity at several banana ripening-related loci. The *y* axis represents counts per million reads. ACS, 1-aminocyclopropane-1-carboxylic acid synthase; BMY, β-amylase; EXP, expansin; OMT1, *O*-methyltransferase; LCYB, lycopene β-cyclase.

and MaNAP5-binding genes, 20% were upregulated in at least one stage during fruit ripening, and more of the upregulated genes belong to the Ban subgenome (718, 39.1%) than either the Dh (523, 28.4%) or Ze (597, 32.2%) subgenome (Extended Data Fig. 7f and Supplementary Table 5). In addition, we did not find a clear relationship between the number of MaNAP4- and MaNAP5-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). These results suggest that the Ban subgenome has 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. In the Cavendish and Gros Michel subgroups, we identified 296 and 186 *NBS-LRR* genes, 252 and 209 *RLP* genes and 1,868 and 1,815 *RLK* genes, respectively (Extended Data Fig. 7g, Supplementary Fig. 9 and Supplementary Table 9). Among them, *NBS-LRR* genes are significantly biased toward the Ze subgenome of both Cavendish and Gros Michel (Extended Data Fig. 7g and Supplementary Table 9). These results indicate that the Ze subgenome modulates disease resistance more than the other two subgenomes.

Discussion

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

We found two novel NAP homologs (*MaNAP4* and *MaNAP5*) highly and specifically expressed in fruit that bind to known fruit ripening-related genes (for example, *ACS* and *EXP*) and many genes of unknown function, suggesting that many genes of unknown function are critical for fruit ripening. *MaNAP* orthologs (*MaNAP1*–*MaNAP3*) were induced during both fruit ripening and leaf senescence, while *MaNAP4* and *MaNAP5* were specifically expressed at high levels during fruit ripening. We hypothesize that *MaNAP4* and *MaNAP5* may be specifically involved in the positive-feedback dual loop[26] for banana fruit ripening.

The two high-quality AAA genome assemblies, together with the newly assembled *M. acuminata* ssp. *zebrina* genome, 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 subgenomes.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01589-3.

References

1. Rouard, M. et al. Three new genome assemblies support a rapid radiation in *Musa acuminata* (wild banana). *Genome Biol. Evol.* **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. *Ethnobot. Res. Appl.* **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. Belser, C. et al. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Commun. Biol.* **4**, 1047 (2021).
12. Belser, C. et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **4**, 879–887 (2018).
13. Cenci, A. et al. Unravelling the complex story of intergenomic recombination in ABB allotriploid bananas. *Ann. Bot.* **127**, 7–20 (2021).
14. Martin, G. et al. Interspecific introgression patterns reveal the origins of worldwide cultivated bananas in New Guinea. *Plant J.* **113**, 802–818 (2023).
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. Busche, M. et al. Genome sequencing of *Musa acuminata* dwarf Cavendish reveals a duplication of a large segment of chromosome 2. *G3* **10**, 37–42 (2020).
19. Carreel, F. et al. Ascertaining maternal and paternal lineage within *Musa* by chloroplast and mitochondrial DNA RFLP analyses. *Genome* **45**, 679–692 (2002).
20. Christelová, P. et al. Molecular and cytological characterization of the global *Musa* germplasm collection provides insights into the treasure of banana diversity. *Biodivers. 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. *Curr. 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).

Methods

Plant materials

The two AAA triploid cultivars 'BaXiJiao' (Baxi) and 'Gros Michel' were selected because they are representative AAA triploids of *M. acuminata* cv. Cavendish and *M. acuminata* cv. Gros Michel, respectively. These two cultivars are and were widely cultivated worldwide and show distinct resistance to *F. oxysporum* f.sp. *cubense* race 1.

Genome assembly

The PacBio Sequel I sequencing data of Cavendish were assembled using Canu[28] (v1.9) with the '-pacbio-raw' option. The assembly was then polished with Illumina short-read sequencing data using Pilon[29] (v1.23) for three iterations. The PacBio HiFi sequencing data of Gros Michel were assembled using Canu[30] (v2.1.1) with the '-pacbio-hifi' option.

All assembled contigs of Cavendish were assigned to three groups based on ancestor genomic information inspired by the trio-binning algorithm[31]. We first extracted and identified ancestor-specific 27-mers from the three published ancestor banana genomes (Banksii, DH-Pahang and Zebrina). Banksii-specific *k*-mers were defined as *k*-mers present in Banksii but not in DH-Pahang and Zebrina. We built ancestor-specific *k*-mers using the *k*-mer counter Merqury[32] (v1.3) with '-M difference'. 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 considered a contig as originating from Banksii only when the contig contained at least 1.5× more Banksii-specific *k*-mers than the other two ancestors (DH-Pahang and Zebrina).

We further anchored the contig sequences in each group onto 11 chromosomes, representing each subgenome, using two different approaches (reference-guided[33,34] and Hi-C-guided[35,36]). Initially, chimeric contigs were corrected based on abnormal Hi-C contact using the ALLHiC_corrector program[35], and the ordering and orientation of these corrected contigs were determined through alignment with reference genomes (that is, the ancestral genomes) using minimap2 with default parameters[37]. The remaining unanchored contigs were reassigned onto each chromosome based on Hi-C contact with anchored sequences using the ALLHiC_rescue function[35]. In addition, we optimized the orders of grouped contigs for each chromosome, resulting in a final release of a chromosomal-scale genome assembly. Using the Cavendish assembly as the reference, we anchored Gros Michel contigs onto 33 chromosomes using RaGOO[33] (v1.11).

The raw nanopore long reads of Zebrina were subjected to self-correction using NextDenovo v2.5.2 (https://github.com/Nextomics/NextDenovo) with the NextCorrect module, and the corrected reads were assembled into contigs. These contigs were then corrected with Illumina short-read sequencing data using NextPolish[38] (v1.4.1) for three rounds. Using the DH-Pahang assembly as the reference, we anchored these corrected contigs onto 11 chromosomes using RaGOO[33] (v1.11).

Gene annotation

Protein-coding genes were annotated using the MAKER[39] (v.2.31.11) genome annotation pipeline, which integrates both ab initio gene predictions generated by AUGUSTUS[40] (v.3.4.0) and GeneMark-EP[41] (v.4.6.3), and homology evidence including plant protein sequences in the OrthoDB v10.1 database[42] as well as a de novo transcriptome assembly generated from 12 RNA-seq datasets download from NCBI SRA (BioProject PRJNA381300) using Trinity[43] (v.2.8.3). For improved results, we performed the MAKER pipeline iteratively (three iterations in total), as recommended[39]. The predicted genes were filtered to remove those with transposable element (TE)-related domains.

The completeness of the final assembly and annotation was evaluated with BUSCO[44] (v.5.3.2) with 1,614 single-copy genes from the lineage database embryophyta_odb10. Protein motifs and domains were annotated using InterProScan[45] (v.5.48). Gene ontology (GO) terms were grouped into plant GO categories based on the results of InterProScan with parameter --GO. The predicted protein-encoding genes were mapped onto KEGG metabolic pathways using the BLAT[46] (v.35) program against the GENES database. The following thresholds were used to find significant matches: amino acid sequence identity ≥30% and length coverage of the query sequence ≥50% with an *E*-value cutoff of 1×10^{-10}.

Repetitive sequence annotation

Repetitive sequences were annotated using a combination of ab initio and homology-based methods. First, an ab initio repeat library was constructed with LTR_FINDER[47] (v1.05) and RepeatModeler[48] (v2.0.2). The predicted repeat library was aligned to the PGSB Repeat Element Database (PGSB-Redat[49]) to classify the repeats into different repeat families. Next, RepeatMasker[50] (open-4.0.7) was applied to perform a homology-based repeat search throughout the whole genome, using both the ab initio repeat database and Repbase[51]. RepeatProteinMask[50] was used to identify any missed repeat related proteins in the previous steps. Finally, overlapping repeats belonging to the same repeat class were combined according to their coordinates in the genome.

Evolutionary analyses

The longest proteins of each gene in seven species, including the six subgenomes of Cavendish and Gros Michel, four wild ancestors and *M. schizocarpa* (as outgroup), were selected. All the longest proteins were compared by BLAST against each other and then clustered with Orthofinder[52] (v2.2.7). Then, 2,628 single-copy orthologous genes were used by ProtTest3.0 (ref. 53) with decision theory (DT) criterion standard to select the best model (JTT+I+G), and phylogenetic trees were constructed with PhyML3.0 (ref. 54).

Paralogous and orthologous gene pairs were identified using MCscan (Python version)[55] with default parameters. Macrosynteny and microsynteny relationships were identified and plotted based on the results of MCscan. K_a and K_s were calculated with the PAML yn00 NG model[56] using coding and protein sequences of orthologous gene pairs, and all zero values were filtered out.

The origins of chromosomal segments were identified based on sequence homology. The genome sequences of five wild diploids, including Banksii, DH-Pahang, Zebrina, Calcutta 4 and *M. schizocarpa*, were merged as the 'reference ancestor genome'. Both Cavendish and Gros Michel assemblies were split into 2-kb fragments and aligned to this 'reference ancestor genome' using BLAT[46] (v.35) '-t = dna -q = dna'. BLAT matches with less than 70% coverage of the query sequence length were filtered out. Based on nucleotide sequence identity, the potential donor of each segment was identified as its best alignment with the diploid progenitor species. We used a 500-kb sliding window without step moving across the whole genomes of Cavendish and Gros Michel and identified the donor of each window based on donor counts for 2-kb segments.

ChIP–seq and data analysis

Banana pulp tissue at stage 4 was fixed with 1% formaldehyde in 1× 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 a Covaris M220 instrument and diluted with low-salt wash buffer (150 mM NaCl in TE) with 1% Triton X-100. The chromatin samples were incubated for 6 h with Dynabeads Protein A/G (Invitrogen) with anti-NAP (DGSSDVHYHL SRQKKP) 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 Tn*5* transposase at 37 °C for 30 min. The beads were then washed with

low-salt and TE buffers and reverse cross-linked for 8 h. The purified DNA was then amplified using N50× and N70× index primers. Two biological replicates were sequenced for the experiment.

Raw reads were first processed using trim_galore[57] (v0.6.7) to trim low-quality or adaptor-originated bases. Trimmed reads were then mapped to the genome reference using bowtie2 (ref. 57) with default parameters. Secondary or supplementary alignments, alignments with mapping quality less than 30 and improperly paired alignments were discarded. PCR duplicates were masked using picard 2.26.8 markduplicates[58]. The resulting alignments of the two replicates were then subjected to MACS2 callpeak[59] for peak calling with the -c parameter specified as the input alignments, -g specified as 557690000 and other parameters set to default. Overlapping peaks between the two replicates were considered as putative protein-binding sites and were used for downstream analysis. Each peak was associated with the nearest gene if it was located within 2 kb upstream to 500 bp downstream of the transcription start site using rtracklayer[60] (v1.60.0) in R.

DNA methylation analysis

Previously published banana whole-genome bisulfite sequencing (WGBS) data of young leaf, pulp of mature green fruit and pulp of ripe fruit were downloaded from the NCBI SRA using the following accession numbers: SRR6328789, SRR6328804, SRR6328805, SRR6328806, SRR6328788, SRR6328791, SRR6328785 and SRR6328790. Raw reads were first processed using trim_galore[57] (v0.6.7) to trim low-quality or adaptor-originated bases. Subsequent mapping of trimmed reads to the genome reference and measurement of the methylation state were performed using Bismark[61] (v.0.24.0) with default parameters. Averaged methylation levels of 1-kb upstream or downstream regions of genes exhibiting biased expression among subgenomes were calculated using methylKit[62] (v1.26.0) based on the results generated by the Bismark package.

DNase-seq and data analysis

Open chromatin was profiled using DNase-seq as previously described[26]. The purified nuclei of banana pulp were suspended in 500 µl of digestion buffer (30 mM Tris-HCl pH 8.0, 14 mM $MgCl_2$, 0.5% CA-630, 0.2% BSA) containing 0.5 U of DNase (RiboSolutions). After incubation for 3 min at 37 °C, the reaction was immediately terminated by adding a stop buffer (0.1% SDS and 50 mM EDTA). RNase A was added, and the digestion was incubated at 37 °C for 30 min. The DNA fragments were purified by phenol/chloroform and precipitated in sodium acetate and isopropanol. The genomic DNA was sonicated to 75–100 bp and used as control. The purified DNA fragments were run on a 2.5% agarose gel, and the small DNA fragments were purified and converted into Illumina TruSeq-type libraries. Two biological replicates were sequenced.

Trimming, mapping and post-alignment processing of DNase-seq data were performed as for the ChIP–seq data. The resulting alignments were down-sampled to 10 million read counts, and each alignment was transformed by extending 50 bp in the left and right directions from the leftmost 5′ coordinate of the alignment. Each alignment of control data was transformed by resizing to 100 bp from the leftmost 5′ coordinate. The transformed alignments in bed format were subjected to MACS2 callpeak[59] for peak calling with parameter settings -f BED -g 557690000 --nomodel -shift 73 -extsize 147, and in addition, -c was specified as the control alignment. Overlapped peaks between the two replicates were considered as putative DHSs and were used for downstream analysis. DHSs were associated with genes in the ChIP–seq analysis.

RNA-seq data analysis

A total of 58 RNA-seq datasets of banana fruit tissue (SRA accessions downloaded from NCBI BioProject accessions PRJNA381300, PRJNA394594 and PRJNA598018) were used for gene expression analysis. RNA-seq short reads were aligned to the genome with Hisat2 v2.1.0 (ref. 63). The expression level of each gene in terms of fraction count was computed by featureCounts[64] (v2.0.3 –M) such that multiply mapping reads were counted. Finally, all expression levels were summarized from the fraction count to the FPKM. A gene was considered to be expressed if its FPKM was >1.

Identified homoeolog expression bias of subgenomes

The analysis focused exclusively on gene triads that had a 1:1:1 correspondence across the three homoeologous subgenomes, including 18,119 syntenic triads and 54,357 homoeologs in total. We defined a triad as expressed when the sum of the expression of the Ban, Dh and Ze subgenome 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[65].

TE distribution in neighboring regions

We used a 100-bp sliding window with a 10-bp step moving across the 5′ and 3′ flanking regions of genes to estimate the TE density around each gene. In each 100-bp window, we calculated the ratio of TE nucleotides and then averaged the ratio across subsets of the homoeologous genes. The averaged values were plotted as the TE density in the flanking region of these subsets of the homoeologous genes.

Gene family and phylogenetic analyses

The major families of plant resistance genes, including those encoding NBS-LRR, RLP and RLK proteins, were identified using the RGAugury pipeline[66]. Each candidate NBS-LRR sequence was then checked for an NB-ARC domain using HMMER[67] (v3.1b2) with an *E*-value cutoff of 1×10^{-5} and length coverage of the NB-ARC sequence of >50% to remove false-positive NB-ARC domain hits.

NAC transcription factors were identified based on the NAC-type NAM model (PF02365). Briefly, the predicted proteins of Cavendish and Gros Michel were searched for the NAC-type NAM domain using HMMER[67] (v3.1b2). The NAM domain was required to be present with an *E*-value cutoff of 1×10^{-5} for a protein to be identified as a NAC transcription factor.

The phylogenetic tree of NBS-LRR genes was constructed based on the alignment of the NB-ARC domain sequences. The phylogenetic trees of both RLP and NAC genes were constructed based on the alignment of whole-protein sequences. The alignment was input into FastTree[68] (v2.1.11) with the Jones–Taylor–Thornton (JTT) model and visualized using FigTree v1.4.4 (https://github.com/rambaut/figtree) and EvolView[69] (v2).

Statistical analysis

All statistical analyses were performed in R 4.1.1. Significant deviations of expression levels among the homoeologs from the three subgenomes were tested using one-way analysis of variance (ANOVA) with Tukey's HSD test, implemented by using the aov and TukeyHSD functions. Statistical comparisons among expression bias categories of syntenic triad homoeologs were calculated using both a one-way ANOVA with Tukey's HSD test and a two-sample *t*-test. The *t*-test was performed using the t.test function with parameter 'alternative = two.sided, paired = true'. The associations of TE density and averaged methylation of 2-kb upstream regions for each gene with the relative expression of gene triads were tested using the cor.test function with the following settings: method = 'pearson', alternative = 'two.sided'.

Reporting summary

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

Data availability

Genome assemblies of Cavendish, Gros Michel and Zebrina v2.0 have been deposited into NCBI under GenBank numbers JAVVNX000000000, JAVVNW000000000 and JAVVNV000000000 and in the National Genomics Data Center BioProject database (https://ngdc.cncb.ac.cn/bioproject/) under the accession number PRJCA019650. Genome assemblies with annotations and results of ChIP–seq and DNase-seq can be accessed at FigShare (https://figshare.com/projects/Origin_and_evolution_of_the_triploid_cultivated_banana_genome/178041). 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 PRJNA1017453 with SRA accessions from SRR23425440 to SRR23425472 and from SRR23885547 to SRR23885549. Fifty-eight RNA-seq datasets were downloaded from NCBI BioProject accessions PRJNA381300, PRJNA394594 and PRJNA598018. DNA methylation data were downloaded from NCBI BioProject PRJNA381300.

Code availability

Custom code and scripts for mapping the origins of chromosomal segments are available at FigShare (https://doi.org/10.6084/m9.figshare.21229205.v1)[70]. All public software used in this study is provided in the accompanying Nature Portfolio Reporting Summary.

References

28. 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).
29. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
30. 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).
31. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
32. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
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. USA* **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**, 289–293 (2009).
37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
38. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
39. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
40. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
41. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom. Bioinform.* **2**, lqaa026 (2020).
42. Kriventseva, E. V. et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
43. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
44. Simão, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
45. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
46. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
47. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
48. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
49. Spannagl, M. et al. PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* **44**, D1141–D1147 (2016).
50. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, 10.1– 10.14 (2009).
51. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
52. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol. Evol.* **20**, 238 (2019).
53. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
54. 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).
55. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
56. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
57. Stubbs, T. M. et al. Multi-tissue DNA methylation age predictor in mouse. *Genome Biol.* **18**, 68 (2017).
58. Broad Institute. Picard toolkit. *GitHub* https://broadinstitute.github.io/picard (2019).
59. Zhang, Y. et al. Model-based analysis of ChIP–Seq (MACS). *Genome Biol.* **9**, R137 (2008).
60. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
61. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
62. Akalin, A. et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
63. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
64. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
65. Ramírez-González, R. H. et al. The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018).

66. Li, P. et al. RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 852 (2016).
67. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
68. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
69. He, Z. et al. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Res.* **44**, W236–W241 (2016).
70. Li, X. et al. Custom code and scripts for mapping the origins of chromosomal segments. *FigShare* https://doi.org/10.6084/m9.figshare.21229205.v1 (2023).

Acknowledgements

We thank G. Riddihough (Life Science Editors) for text editing. X.L. acknowledges funding from the National Natural Science Foundation of China (32370687). P.L. acknowledges funding from the National Natural Science Foundation of China (32372666) and Construction of Plateau Discipline of Fujian Province (102/71201801104). L.Z. acknowledges funding from the National Natural Science Foundation of China (32272750). Y.V.d.P. 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., J.W. and H.L. collected the samples and extracted DNA and RNA. L.Z., P.L., J.W. and S.Y. coordinated the Illumina and PacBio sequencing. X.Z., M.J. and X. Chang assembled genomes and Hi-C data analyses. X.Z., C.Z. and X. Wang conducted protein-coding gene and repetitive sequence annotations. L.Z. and X.L. performed phylogenetic analyses. X.L., X. Chen and L.Z. performed comparative genomic analysis. X.L., X.Z., 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. and X.Z. wrote the manuscript draft. L.Z., P.L., S.Y., X.L., X.Z., Y.V.d.P., Z.L., Z.W., J.H. and J.-M.A. reviewed and revised the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

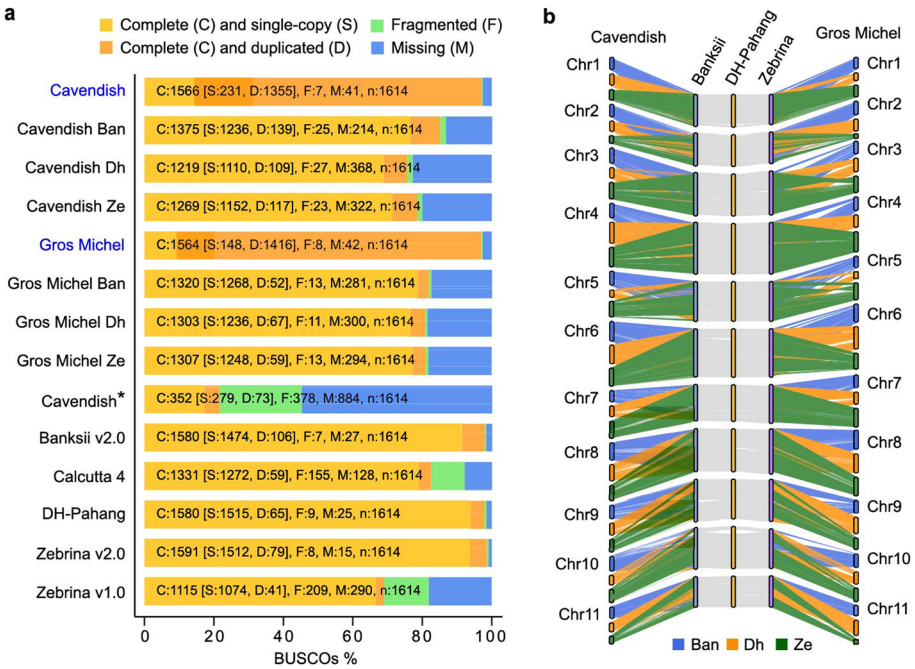
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01589-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01589-3.

Correspondence and requests for materials should be addressed to Yves Van de Peer, Zonghua Wang, Xiaofan Zhou, Jihua Wang, Peitao Lü or Liangsheng Zhang.

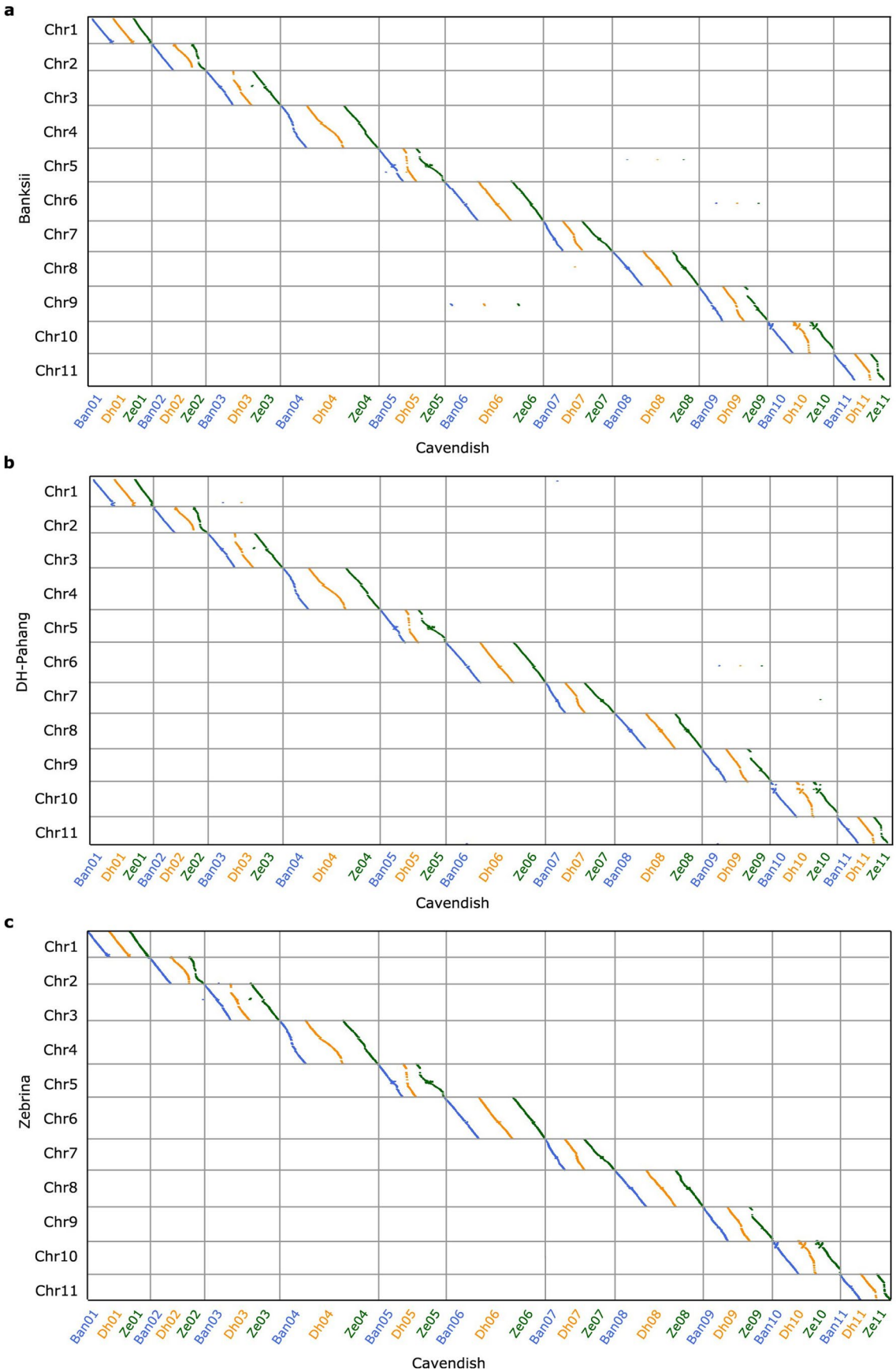
Peer review information *Nature Genetics* thanks Jordi Garcia-Mas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-023-01589-3



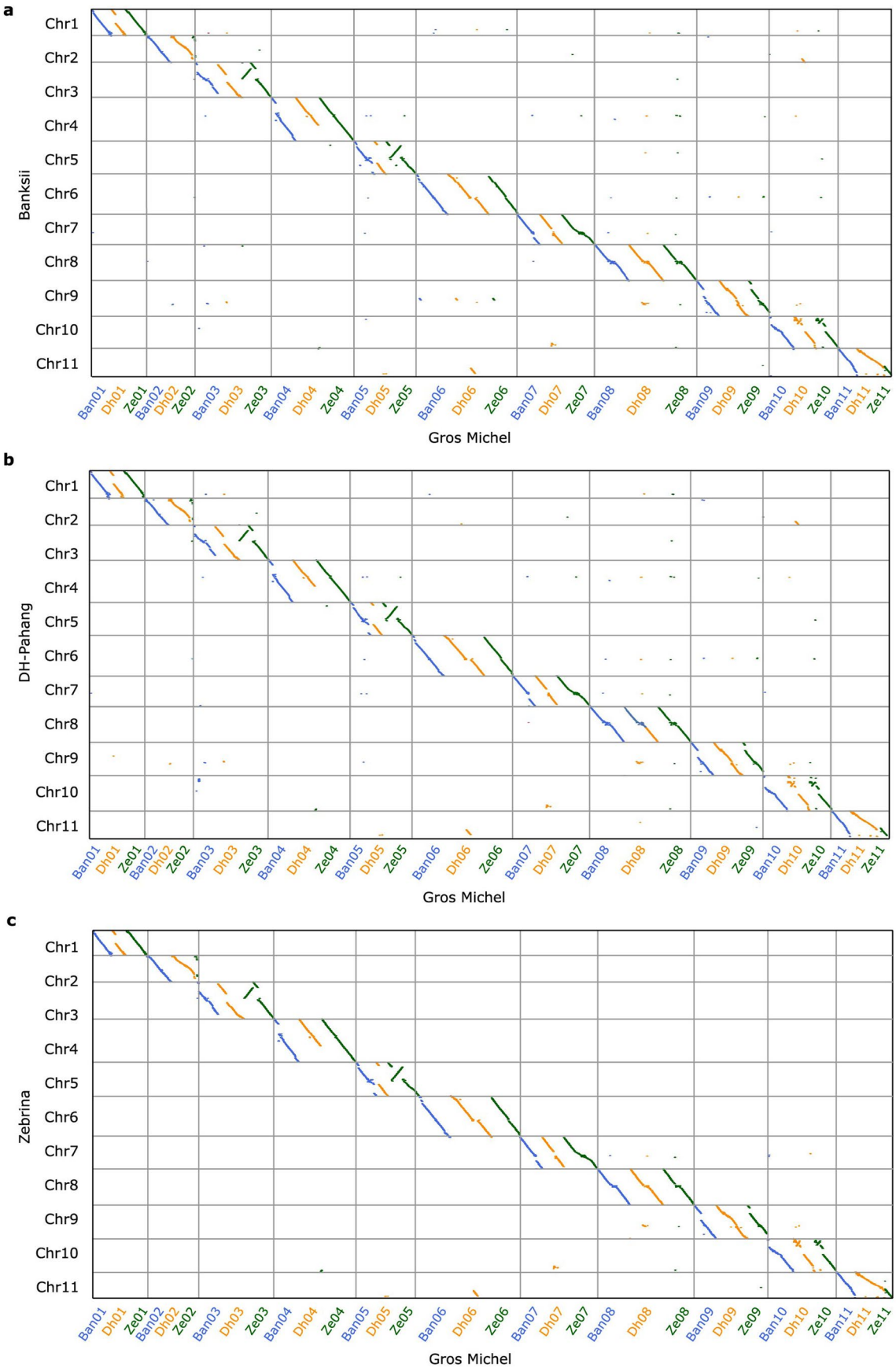
Extended Data Fig. 1 | Genome assemblies of Cavendish and Gros Michel. **a**, BUSCO completeness assessments of the genome assemblies of Cavendish, Gros Michel, and four diploid wild banana species (Banksii, DH-Pahang, Zebrina, and Calcutta 4). Cavendish* was assembled by Busche et al.[18]. Zebrina v1.0 was assembled by Rouard et al.[1], and Zebrina v2.0 was our assembly based on nanopore long-reads. The abbreviations of banana species refer to Fig. 1a. **b**, Macrosyntenic comparison of the entire Cavendish, Gros Michel and three diploid wild banana genomes (Banksii, DH-Pahang, and Zerbina), with each chromosome colored according to sub-genomes (Ban in blue, Dh in orange, and Ze in green).

https://doi.org/10.1038/s41588-023-01589-3



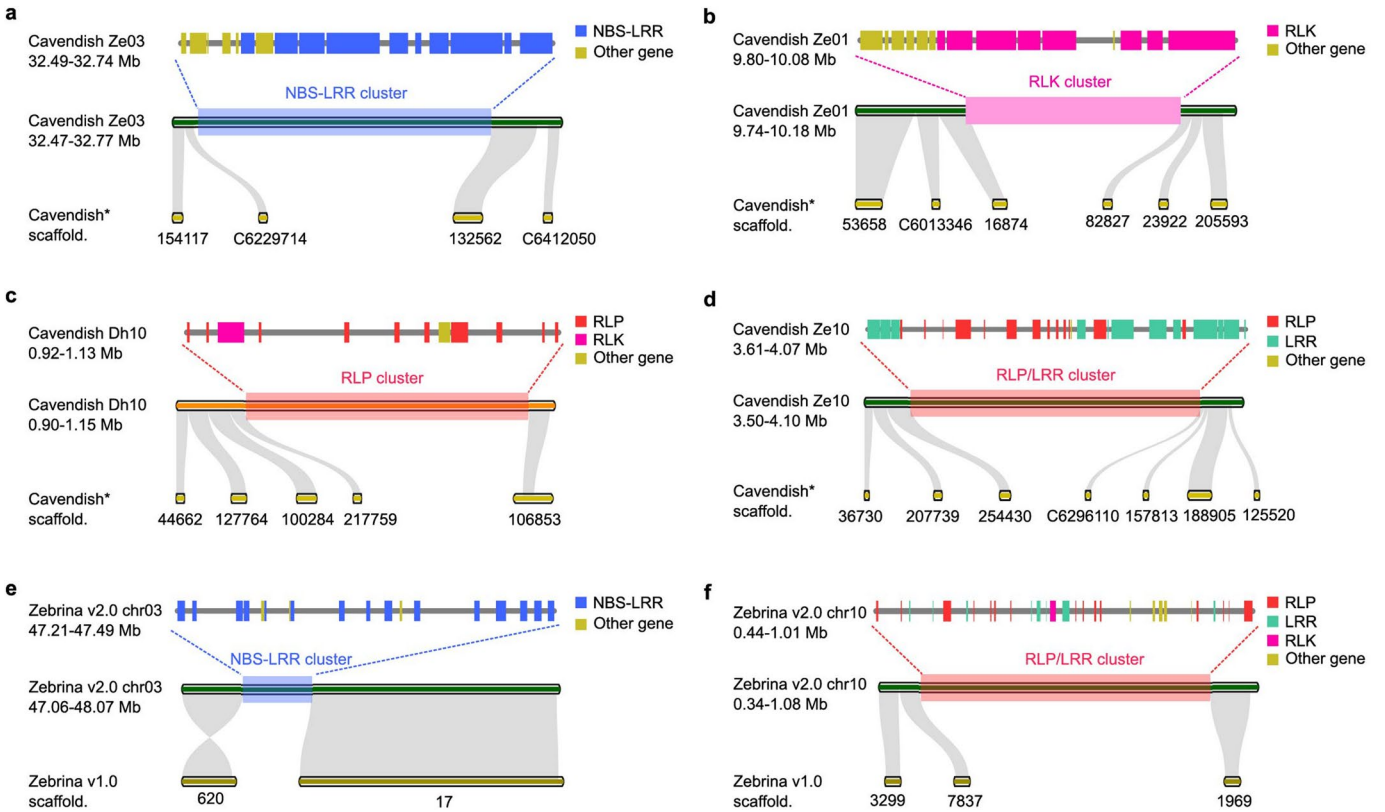
Extended Data Fig. 2 | Macrosyntenic comparison of the entire Cavendish and three diploid wild banana genomes: Banksii (a), DH-Pahang (b), and Zebrina (c). Each chromosome set colored according to sub-genomes (Ban in blue, Dh in orange, and Ze in green). The abbreviations of banana species refer to Fig. 1a.

https://doi.org/10.1038/s41588-023-01589-3



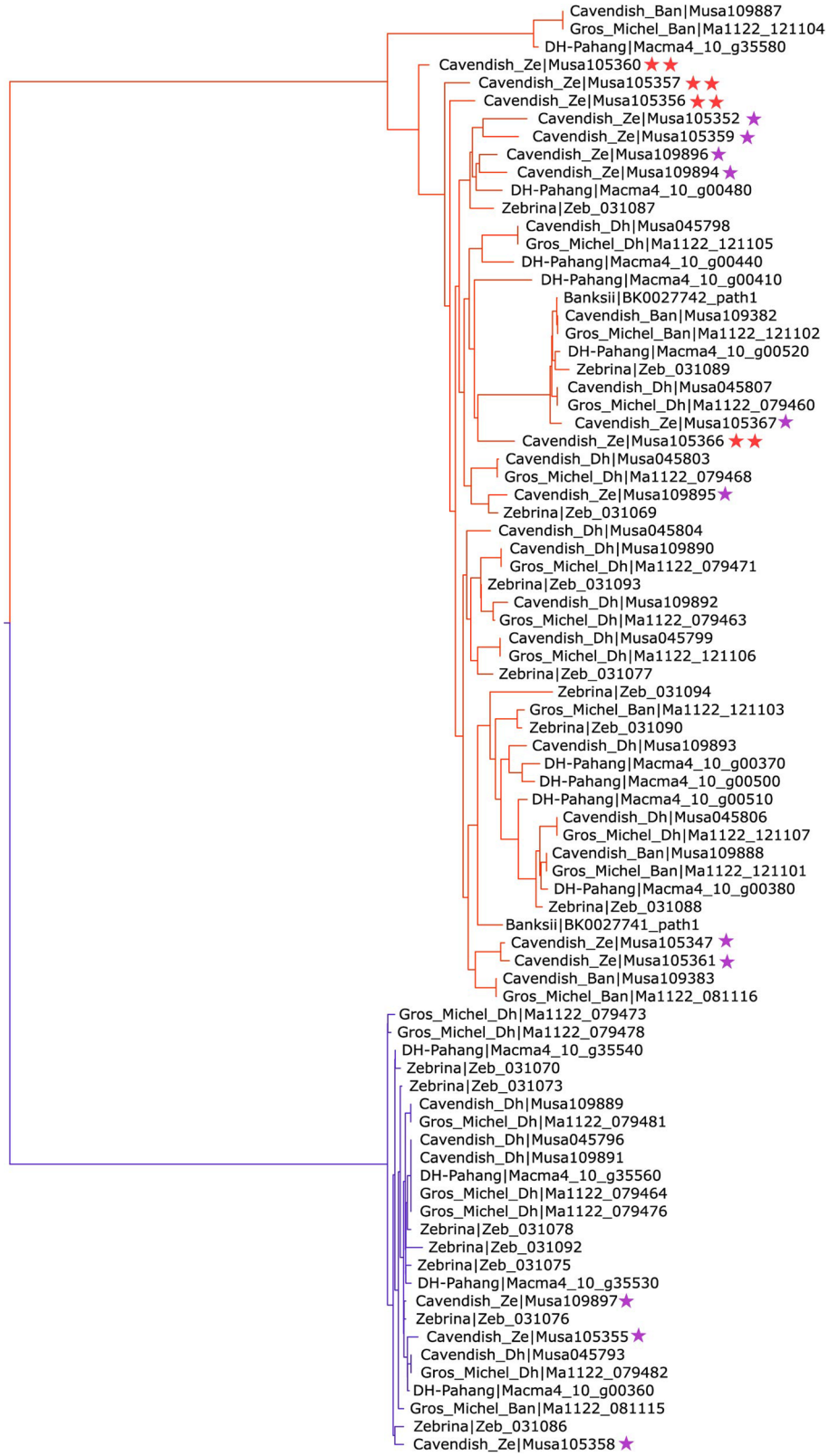
Extended Data Fig. 3 | Macrosyntenic comparison of the entire Gros Michel and three diploid wild banana genomes: Banksii (a), DH-Pahang (b), and Zebrina (c). Each chromosome set colored according to sub-genomes (Ban in blue, Dh in orange, and Ze in green). The abbreviations of banana species refer to Fig. 1a.

https://doi.org/10.1038/s41588-023-01589-3



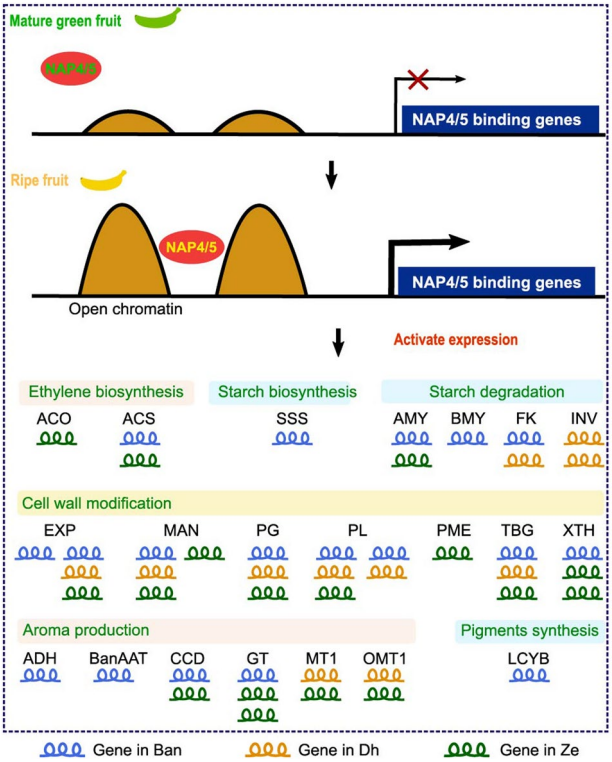
Extended Data Fig. 4 | Examples of high-quality Cavendish and Zebrina genome assemblies. a-d, NBS-LRR cluster, RLK cluster, RLP cluster, and RLP/LRR cluster on Ze03, Ze01, Dh10, and Ze10 of Cavendish, while not assembled in the previously published Cavendish assembly. Cavendish* was assembled by Busche et al.[18]. **e** and **f**, NBS-LRR cluster on chromosome 3 and RLP/LRR cluster on chromosome 10 of our assembled Zebrina v2.0 with length of 280 kb and 370 kb, while being two big gaps in the published Zebrina v1.0 (ref. 1). Each resistance gene was colored on micro-synteny plot (NBS-LRR in blue, RLK in pink, RLP in red, LRR in green, and other gene in yellow). The abbreviations of banana species refer to Fig. 1a.

https://doi.org/10.1038/s41588-023-01589-3



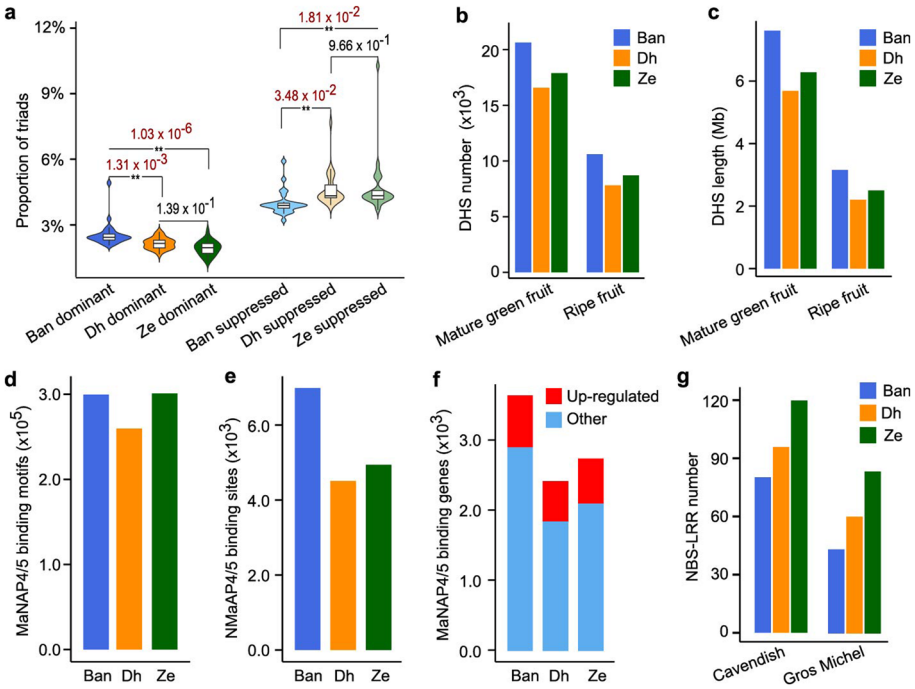
Extended Data Fig. 5 | Phylogenetic tree of banana RLPs involved in *Foc* race1-associated QTL (named as RLP locus)[25]. The purple stars denote RLPs located in the Ze sub-genome, while the two red stars denote RLPs found only in the Ze sub-genome of Cavendish. The abbreviations of banana species refer to Fig. 1a.

https://doi.org/10.1038/s41588-023-01589-3



Extended Data Fig. 6 | A model of MaNAP4/5′ regulation of banana fruit ripening. In the model, these genes directly regulated by MaNAP4/5 are key genes in the fruit ripening process.

https://doi.org/10.1038/s41588-023-01589-3



Extended Data Fig. 7 | Sub-genome dominance in the triploid banana genome. a, Statistical comparison of categories of syntenic triad homoeolog expression bias. P-values were determined by one-way ANOVA with Tukey's HSD test (n = 26 tissues of each category) within the suppression and dominance categories, and P-values less than 0.05 was highlighted in red. For boxplot in this study, the middle line represents the median, the lower and upper edges of the box represent the first and third quartiles, the end of the lower whisker represents the smallest value at most 1.5× inter-quartile range from the lower edge of the box, the end of the upper whisker represents the largest value at most 1.5× inter-quartile range from the upper edge of the box. **b** and **c**, Total number (b) and length (c) of DNase-hypersensitive sites (DHSs) detected in mature green and ripe fruits. **d**-**f**, Sub-genome distribution of MaNAP4/5 binding motifs (d), sites (e) and genes (f). **g**, Distribution of NBS-LRR resistance genes in the sub-genomes.

nature portfolio

Corresponding author(s): Liangsheng Zhang

Last updated by author(s): Jul 4, 2023

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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.

n/a	Confirmed
☐	☒ 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	No software was used for data collection
Data analysis	All routine analysis were performed using publicly available and appropriately cited softwares. Software and algorithm used in this paper are listed as follows: Canu v1.9, Canu v2.1.1, NextDenovo v2.5.2, NextPolish v1.4.1, Pilon v1.23, ALLHiC (https://github.com/tangerzhang/ALLHiC), Merqury v1.3, RaGOO v1.11, MAKER v.2.31.11, AUGUSTUS v.3.4.0, GeneMark-EP v.4.6.3, Trinity v.2.8.3, BUSCO v5.3.2, InterProScan v.5.48, RepeatModeler 2.0.2, LTR_FINDER v1.05, RepeatMasker open-4.0.7, RepeatProteinMask (included in RepeatMasker open-4.0.7), BLAT v.35, Orthofinder v2.2.7, ProtTest3.0, PhyML3.0, MCscan (Python version), Bismark v.0.24.0, MACS2, trim_galore v0.6.7, picard 2.26.8, R package methylKit v1.26.0, R package rtracklayer v1.60.0, bowtie2, Hisat2 v2.1.0, featureCounts v2.0.3, HMMER version 3.1b2, FastTree v2.1.11, FigTree v1.4.4, EvolView v2 (online tool), R 4.1.1, RGAugury pipline and trio-binning algorithm.

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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 PRJNA889940 with SRA accessions from SRR23425440 to SRR23425471, and from SRR23885547 to SRR23885549. 58 RNA-seq datasets were downloaded from NCBI BioProject: PRJNA381300, PRJNA394594, and PRJNA598018. DNA methylation data were downloaded from NCBI BioProject: PRJNA381300. Genome assemblies of Cavendish, Gros Michel, and Zebrina v2.0 have been deposited into NCBI under BioProject PRJNA889940. Genome annotations and results of ChIP-Seq and DNase-Seq can be accessed at Figshare with Digital Object Identifier 10.6084/m9.figshare.21249081.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided 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	Five wild bananas were selected, including Musa schizocarpa (S genome), M. acuminata ssp. banksii (A genome), malaccensis (A genome), zebrina (A genome), and burmannica (A genome), to anchor contigs of triploid Cavendish and Gros Michel into three sub-genomes (haplotypes).
Data exclusions	No data were excluded from the analysis.
Replication	Two biological replicates were used for hromatin immunoprecipitation and DNase hypersensitive sites sequencing and data analysis. All attempts at replication were successful.
Randomization	This is not relevant to this study, as it is about assembly and analysis of two banana varieties from Cavendish and Gros Michel group.
Blinding	No group allocation was needed in this study.

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

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: MaNAP4/5-specific antibody was generated by using synthetic peptide DGSSDVHYHLSRQKKP.

Validation: This is an anti-MaNAP4/5 rabbit serum customized at supplier QWBIO (http://www.qwbio.com). We have tested the antibody with western blotting and used it for ChIP-seq.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Digital Object Identifier 10.6084/m9.figshare.21249081 at Figshare (https://www.digital-science.com/product/figshare)

Files in database submission: ChIP_seq directory in ChIP_Seq_and_DNase_Seq.zip contains six files: 7R3.overlapped.narrowPeak, 7R3_rep1.bw, 7R3_rep1.narrowPeak, 7R3_rep2.bw, 7R3_rep2.narrowPeak, input.bw

Genome browser session
(e.g. UCSC)

no longer applicable for final submission.

Methodology

Replicates: Two biological replicates of ChIP-seq were performed: Baxi pulp at stage 4 (ChIP-seq), one input control was prepared and used for peak calling

Sequencing depth: ChIP-seq replicate 1: total number of reads: 26,523,150; uniquely mapped reads: 14,456,182; PE150 reads
ChIP-seq replicate 2: total number of reads, 29,961,282; uniquely mapped reads: 16,178,046; PE150 reads
Input control: total number of reads: 28,565,670; uniquely mapped reads: 10,774,934; PE150 reads

Antibodies: MaNAP4/5-specific antibody

Peak calling parameters: MACS2 callpeak was used for peak calling with the -c parameter specified as the input alignments, -g specified as 557690000, and other default parameters.

Data quality: Quality of ChIP-seq was checked by manually viewing bw file of ChIP and input control on IGV genome browser. Special attentions were also payed to the number of peaks identified by MACS2 callpeak and whether the shifting model was built normally. At FDR 5% (MACS2 callpeak default settings, -q, minimum FDR (q-value) cutoff for peak detection, 0.05), there were 22,525 and 34,503 peaks were called from ChIP-seq replicate 1 and 2, respectively. The overlapped 16,997 peaks between the two peak sets were used for down stream analysis.

Software: Trim_galore was used to trim low quality or adapter-originated bases from raw reads. Trimmed reads were then mapped to genome reference using bowtie2 with default parameters. Picard markduplicates was used to mask PCR duplicates. Finally, MACS2 callpeak was used for peak calling.