nature genetics

Article

https://doi.org/10.1038/s41588-025-02166-6

Pan-genome analysis reveals the evolution and diversity of *Malus*

Received: 26 January 2024

Accepted: 14 March 2025

Published online: 16 April 2025

Check for updates

Wei Li[1,12], Chong Chu[2,12] ✉, Taikui Zhang[3,12], Haochen Sun[1,12], Shiyao Wang[1,12], Zeyuan Liu[4,12], Zijun Wang[1], Hui Li[1], Yuqi Li[1], Xingtan Zhang[5], Zhiqiang Geng[6], Youqing Wang[6], Yi Li[7], Hengtao Zhang[8], Weishu Fan[9], Yi Wang[1], Xuefeng Xu[1], Lailiang Cheng[10], Dehui Zhang[4], Yao Xiong[1], Huixia Li[1,7], Bowen Zhou[1], Qingmei Guan[4] ✉, Cecilia H. Deng[11] ✉, Yongming Han[6] ✉, Hong Ma[3] ✉ & Zhenhai Han[1] ✉

Malus Mill., a genus of temperate perennial trees with great agricultural and ecological value, has diversified through hybridization, polyploidy and environmental adaptation. Limited genomic resources for wild *Malus* species have hindered the understanding of their evolutionary history and genetic diversity. We sequenced and assembled 30 high-quality *Malus* genomes, representing 20 diploids and 10 polyploids across major evolutionary lineages and geographical regions. Phylogenomic analyses revealed ancient gene duplications and conversions, while six newly defined genome types, including an ancestral type shared by polyploid species, facilitated the detection of strong signals for extensive introgressions. The graph-based pan-genome captured shared and species-specific structural variations, facilitating the development of a molecular marker for apple scab resistance. Our pipeline for analyzing selective sweep identified a mutation in *MdMYB5* having reduced cold and disease resistance during domestication. This study advances *Malus* genomics, uncovering genetic diversity and evolutionary insights while enhancing breeding for desirable traits.

Malus Mill. (Rosaceae) is a genus with over 35 species naturally distributed across the temperate Northern Hemisphere, from East Asia and Europe to North America[1,2]. The *Malus* radiation has been frequently driven by natural selection, ecological interaction, hybridization, introgression and polyploidy[3,4], generating vast species diversity within this genus[5,6]. The phylogenetic relationships of *Malus* have yet to be definitely resolved owing to the indistinct genetic background across the genus and the limited available data, primarily derived from plastome and/or nuclear ribosomal DNA sequences, or single-nucleotide polymorphisms (SNPs) from comparison with the genome of domesticated apples[7,8], even with transcriptomic nuclear data from several *Malus* members[9,10].

Whole-genome duplications (WGDs) are found in vertebrate animals[11,12], fungi[13] and especially flowering plants[14–17] and are thought to be a major force in genome and gene functional evolution[18,19]. Although WGDs have been investigated using single or a few sequenced genomes and large-scale phylotranscriptomics[16,20,21], comprehensive analyses of WGDs using pan-genomes have not been reported. A detailed examination of gene evolutionary patterns and genome synteny across multiple genomes of the same group could elucidate the impact of WGDs on gene and functional evolution. *Malus* species have experienced an ancient WGD[22], which provides an opportunity to examine the evolutionary patterns of WGD-derived gene duplicates during the history of *Malus*.

Malus species are economically important and have been extensively grown for their fruits or as ornamentals[5]. Apple (*Malus domestica*) is one of the most widely grown *Malus* species, playing an important role in global deciduous fruit production[22,23]. The domesticated apple

A full list of affiliations appears at the end of the paper. ✉e-mail: chong_chu@hms.harvard.edu; qguan@nwafu.edu.cn; cecilia.deng@plantandfood.co.nz; hanym@mail.buct.edu.cn; hxm16@psu.edu; rschan@cau.edu.cn

https://doi.org/10.1038/s41588-025-02166-6

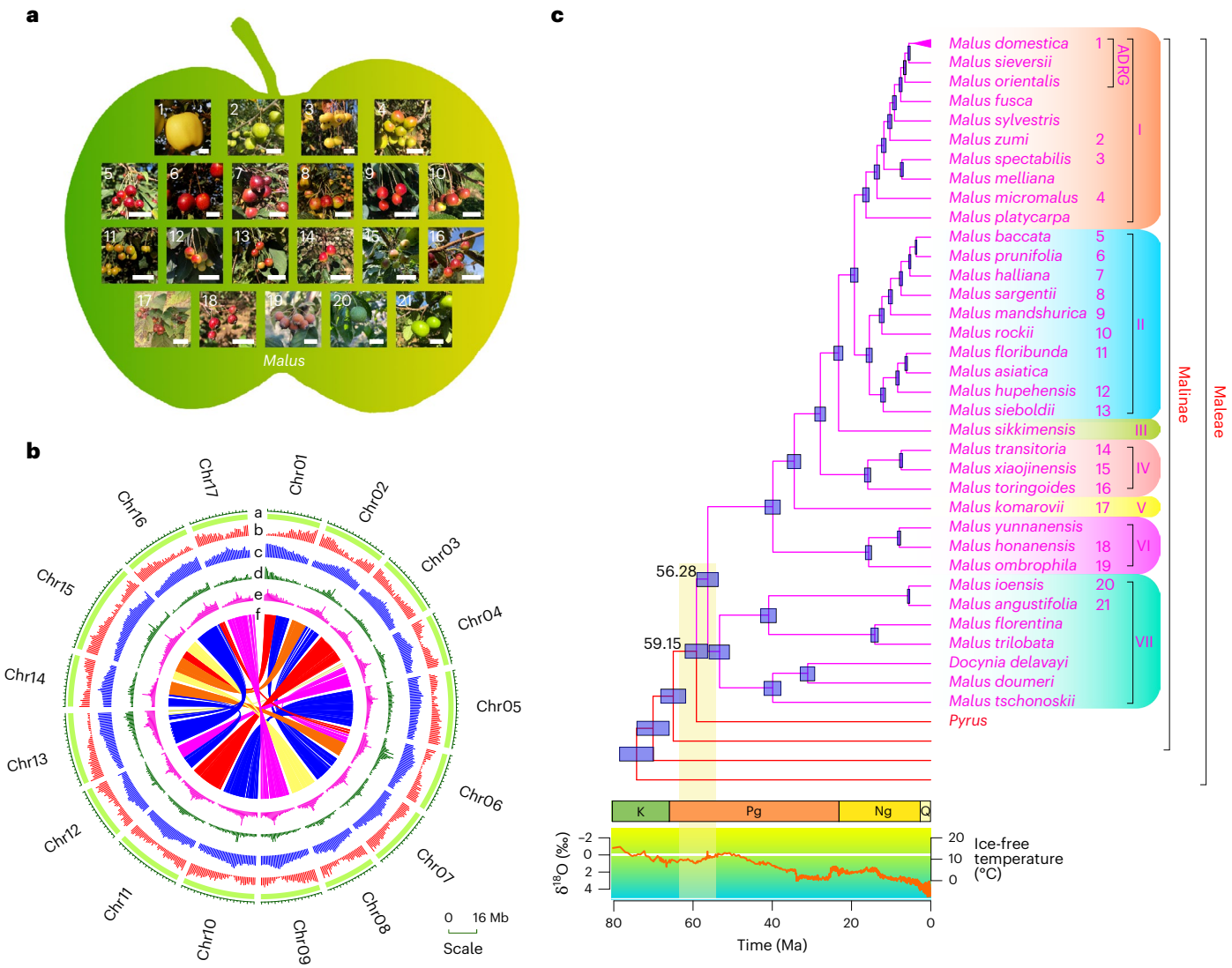


Fig. 1 | The *Malus* evolutionary landscape based on phylogenomics. a, The fruit phenotypes of *Malus* species are highly diverse in size, color and shape. Size bars, 2 cm. Numbers on the photographs are the same as the numbers to the right of the species names in **c**. **b**, A Circos plot of the diploid 'Golden Delicious' apple genome assembly. Tracks from outer to inner (a to f) are chromosomes, gene density, repeat elements density, LTR/Copia, LTR/Gypsy and syntenic regions between chromosomes. **c**, A dated phylogenetic tree of *Malus* simplified from Supplementary Fig. 3. The tree branch length represents the estimated time (the median value) from divergence to the present. Species names are indicated in different colors to the right of terminal branches, with purple for *Malus* species and red for others in the apple tribe (Maleae). The *Malus* topology is a summarized phylogeny from phylogenetic analyses using three different gene sets (Supplementary Fig. 3). In *Malus*, seven clades are designated as I through VII and highlighted with different background colors. Numbers on the nodes represent the divergence times. Horizontal blue bars on each node indicate the 95% confidence intervals of the divergence time in millions of years. The numbers to the right of species are the same as those in **a**. Stratigraphic periods of Cretaceous (K), Paleogene (Pg), Neogene (Ng) and Quaternary (Q) are illustrated by colored boxes below the tree, based on the ages of their boundaries in the International Chronostratigraphic Chart Capella-Gutiérrez (version 2022/02). The orange curve in the graph below the tree indicates the changes in oxygen isotope records of $\delta^{18}O$ (‰), reflecting the temperature changes as indicated by the right axis. The vertical yellow bar indicates the climatic changes during the origination and divergence of *Malus*. ADRG, apple domestication-related group. Ma, millions of years ago.

originated from *M. sieversii* in Central Asia, with introgression from *M. orientalis* and *M. sylvestris*[5,8]. Although *Malus* species have been domesticated and cultivated for thousands of years, only a few closely related cultivars have been extensively used in commercial orchards and industry[6]. The loss of traditional and locally well-adapted cultivars has caused a considerable reduction in the gene pool available for future breeding and improvement practices[24], whereas wild species provide important sources of potentially beneficial genes and genetic variations underlying desirable traits[25,26].

Despite notable advances in assembling the apple genome, previous efforts captured only a limited portion of the genetic diversity within the *Malus* genus[27–39] (Supplementary Note 1). In this study, we de novo assembled 30 high-quality genomes covering *Malus domestica* 'Golden Delicious' and 29 wild *Malus* species, containing diploids and polyploids. The genus-wide high-quality genomes provided an excellent foundation for reconstructing high-resolution phylogenetic relationships among *Malus* species, biogeographical reconstruction of the origins of the *Malus* species and their diversification, and understanding of the *Malus* genome evolution, including WGD events and the important contribution of hybridization to species radiation and polyploidy in *Malus*. Our newly generated graph-based pan-genome enabled the integration of multiple types of genetic variations into

https://doi.org/10.1038/s41588-025-02166-6

Table 1 | Summary of haploid consensus genome assembly and annotation

Ploidy	Accession	Estimated heterozygosity rate	Assembly size (Mb)	Contig N50 (Mb)	Maximum contig length (Mb)	Scaffold N50 (Mb)	BUSCO completeness of assembly (%)	LAI	Repeat content (%)	Number of genes	BUSCO completeness of annotation (%)
Diploid	*Malus domestica* 'Golden Delicious'	0.0117	676.41	36.69	55.49	37.92	99.0	22.14	56.95	49,981	96.6
	Malus orientalis	0.0139	667.89	32.33	55.68	37.60	98.8	21.72	56.54	48,239	98.1
	Malus fusca	0.0224	687.14	36.21	58.57	36.21	98.8	21.12	57.48	48,916	97.4
	Malus zumi	0.0244	675.11	36.16	55.00	36.84	99.1	21.06	56.82	47,911	98.2
	Malus melliana	0.0304	706.82	32.91	53.79	35.99	98.9	20.48	56.77	49,798	97.8
	Malus spectabilis	0.0312	708.06	32.54	66.58	36.12	98.8	20.40	57.69	48,886	95.9
	Malus micromalus	0.0303	700.02	36.63	56.42	36.63	98.8	19.87	57.44	50,237	98.3
	Malus prunifolia	0.0307	699.84	23.46	54.37	36.15	99.1	20.42	57.65	49,171	97.7
	Malus baccata	0.00989	652.15	35.19	52.92	35.99	99.0	18.21	57.55	43,441	96.7
	Malus halliana	0.0305	664.20	34.63	52.99	36.14	99.1	18.39	58.38	44,289	97.8
	Malus sargentii	0.0304	673.62	33.83	53.89	36.37	99.1	20.46	56.83	48,790	96.0
	Malus mandshurica	0.00829	668.11	8.37	31.12	37.21	98.8	19.00	57.39	46,079	96.1
	Malus floribunda	0.0279	699.95	30.42	54.41	36.50	99.0	18.02	57.47	49,163	97.8
	Malus komarovii	0.00995	627.43	32.32	49.80	35.82	98.7	21.24	55.82	44,320	95.9
	Malus yunnanensis	0.00943	632.33	26.91	50.16	34.55	98.9	18.85	55.65	45,643	97.7
	Malus honanensis	0.00705	627.10	21.19	40.84	35.58	98.5	20.46	55.48	45,101	97.2
	Malus ombrophila	0.00959	632.57	29.91	38.98	34.41	98.9	18.09	55.72	44,546	97.7
	Malus florentina	0.00737	767.88	19.12	50.23	42.38	99.1	22.45	60.75	48,472	97.7
	Malus ioensis	0.00969	747.33	33.79	56.62	38.52	98.8	22.02	61.59	45,800	98.1
	Malus doumeri	0.012	592.22	33.28	47.69	33.27	99.1	17.40	53.54	43,341	97.4
Triploid	*Malus platycarpa*	0.0597	1,182.87	14.82	44.31	31.38	99.2	16.26	56.12	91,738	95.8
	Malus hupehensis	0.015	1,168.99	15.25	53.64	33.67	99.0	14.00	58.12	80,001	96.1
	Malus transitoria	0.0148	1,259.45	20.73	56.16	36.22	99.3	15.33	58.67	83,196	97.7
	Malus angustifolia	0.0126	1,216.72	3.10	18.29	28.78	98.6	19.07	62.38	73,356	96.7
Tetraploid	*Malus rockii*	0.0122	1,296.64	2.37	13.66	30.82	99.1	14.70	56.54	90,653	97.8
	Malus asiatica	0.0175	1,081.27	2.73	17.25	28.84	98.9	15.41	55.87	76,680	97.2
	Malus sieboldii	0.0182	1,371.37	3.20	12.62	16.61	99.4	15.69	58.80	94,365	97.8
	Malus sikkimensis	0.022	1,271.66	5.36	27.03	30.10	99.2	15.18	56.95	90,418	98.2
	Malus xiaojinensis	0.0283	1,393.30	12.20	32.06	33.50	99.2	15.23	58.09	95,938	96.5
	Malus toringoides	0.0254	1,270.97	13.85	47.72	32.36	99.0	15.14	57.28	89,315	98.6

N50 is the shortest contig length for which longer and equal-length contigs together make up at least half of the entire assembly.

a comprehensive genome graph, constructed a pan-*Malus* genome portrait and uncovered pan-*Malus* structural variations (SVs) with a focus on comparison with the *M. domestica* genome, demonstrating clear advantages over a linear reference genome from a single species/accession in capturing selective sweeps for important agronomic traits.

Results

The *Malus* evolutionary genomic landscape

We sequenced accessions from 30 species, de novo assembled high-quality genomes and anchored contigs to chromosomal-scale assemblies (Fig. 1a, Table 1, Supplementary Figs. 1 and 2 and Supplementary Table 1). The assembled *Malus* genomes contain 53.54–62.38% repetitive sequences (Supplementary Table 2). Long terminal repeat retrotransposons (LTR-RTs) are the most prevalent transposable elements, making up 36.05–49.63% of the assembled genomes, as illustrated by the example of the diploid 'Golden Delicious' apple genome (Fig. 1b and Supplementary Table 2). In addition, the *Malus* genomes were predicted to have an average of 47,106, 82,073 and 89,562 protein-coding genes for the 20 diploid, 4 triploid and 6 tetraploid species, respectively (Table 1). Moreover, our completeness evaluation of the assembled genomes and gene predictions achieved values of the LTR assembly index (LAI) ranging from 14.00 to 22.45

(>20 for 12 genomes; Table 1) and ≥98.5% of Benchmarking Universal Single-Copy Orthologs (BUSCOs) in assemblies and ≥95.8% of BUSCOs in the predicted genes across all species (Table 1).

To resolve the phylogenetic relationships among *Malus* species, we used OrthoFinder[40] and HaMStR[41] to identify 998 single-copy genes, constructed gene trees and inferred a coalescent species-tree (Methods, Supplementary Fig. 3a and Supplementary Table 3). Within the *Malus* clade, there are 5 *M. domestica* accessions, 33 additional *Malus* species and a closely related species, *Docynia delavayi*. Our results support a monophyletic *Malus* (Fig. 1c) with the merge of *Docynia* into a more broadly circumscribed *Malus*[9,42] (see details and discussions in Supplementary Fig. 3a–c). Our fully resolved *Malus* phylogeny with high bootstrap (BS) values supports seven *Malus* clades (I–VII; Fig. 1c and Supplementary Fig. 3). To provide a context of geological times for *Malus* evolution, we used a Bayesian approach with 12 fossil calibrations (Methods) and found that the estimated age (56.28 (53.60–58.96) Ma) of crown *Malus* was within the geological period of dramatically changing temperature (Fig. 1c and Supplementary Fig. 3d). Using the new *Malus* nuclear phylogeny reconstructed here and the geographical distribution of *Malus* species, we performed a biogeographical reconstruction of the *Malus* origins (Methods and Supplementary Fig. 4) and found that *Malus* probably originated in Asia and diversified in Asia during the early history of the genus (from 39.84 to 56.28 Ma).

Clade I contains *M. domestica* and nine closely related species (Fig. 1c and Supplementary Fig. 3a–c). Within clade I, *M. orientalis* and *M. sieversii* are strongly supported as successively sisters to *M. domestica*, and these three species together form an apple domestication-related group. Clade II contains *M. baccata*, a common rootstock for grafting apple, and nine relatives. Outside clades I + II, the clades III (*M. sikkimensis*), IV, V (*M. komarovii*), VI and VII (Fig. 1c) occupy progressively more distant lineages. The separation of major *Malus* lineages (for example, the stem lineages of clade III–VI; Fig. 1c and Supplementary Fig. 3d) during decreasing temperatures and subsequent divergence in the cooled climates suggests a likely adaptation of *Malus* to cooler global environments, consistent with the distribution of *Malus* species to high altitude and latitude habitat. Biogeographical reconstruction analyses support the idea that the shared ancestor of the North American and European lineages migrated from Asia to a region including the Mediterranean, Europe and the Caucasus (Supplementary Fig. 4). In addition, multiple recent dispersal events led to the spread of several *Malus* species to North America, northern Eurasia and Africa (Supplementary Fig. 4).

Genome duplications and *Malus* diversity

To address the uncertainties regarding the phylogenetic position of the WGD detected in the *M. domestica* genome and by previous phylotranscriptomic analyses[10,22,43,44] and to investigate the retention patterns of gene duplications (GDs) during the early history of *Malus*, we performed phylogenomic analyses (Methods) and placed 6,519 GDs at the most recent common ancestor (MRCA) of Malinae, the subtribe containing *Malus* and other genera that produce fleshy fruit (node N1; Fig. 2a,b and Supplementary Fig. 5a). Among them, 4,660 GDs have the (AB)(AB) retention type with both duplicates found in both *Crataegus pinnatifida* and one of the other Malinae lineages (Supplementary Fig. 6), providing strong evidence for a WGD event in the early evolution of Maleae[10,22,43,44]. Further syntenic support of the early Maleae WGD was obtained with the detection of anchor genes in the newly sequenced *Malus* genomes (for example, 4,655 of the 4,660 (AB)(AB)-type GDs and 1,852 of the 1,859 (AB)A-type GDs included anchor genes from at least one species; Supplementary Fig. 6). Furthermore, among the 1,852 (AB)A-type GDs with anchor genes, the anchor genes of 1,816 GDs were located in the syntenic blocks that also contain anchor genes corresponding to the GDs of the (AB)(AB) type mapped at Malinae (Supplementary Fig. 6). These results support the idea that most of the GDs in both the (AB)(AB) and (AB)A types were derived from the same WGD mapped to the Malinae MRCA.

In addition to the GDs mapped at Malinae, several clusters with ≥500 GDs were mapped at the nodes within Malinae (N2–N6; Fig. 2c and Supplementary Figs. 5–11). To investigate whether these GD bursts represent additional WGD events, we identified the anchor genes that matched these GD bursts, and discovered that 84.7–98.3% GDs mapping at different phylogenetic positions are matched by the Malinae WGD-derived anchor genes in syntenic blocks, strongly supporting the idea that the GD clusters at N2–N6 are GDs from the Malinae WGD (Fig. 2c and Supplementary Figs. 7–11). One explanation for the mapping of GDs from the same syntenic block to later phylogenetic nodes in Malinae could be that the corresponding gene pairs experienced gene conversion at the MRCA of the related clade and became more similar in sequence than the original paralogs (see, for example, ref. 45).

The tribe Gillenieae were hypothesized to be one of the parental lineages of the ancestral Maleae[44]. We examined the GD distribution for evidence to support this hypothesis and found that 4,076 GDs are mapped at the MRCA of Malinae/Maleae and Gillenieae (Fig. 2a) and include 3,939 GDs branching as ((Malinae/Maleae, Gillenieae), Malinae/Maleae) (Supplementary Fig. 12). This finding suggests that Gillenieae could be a parental lineage, which experienced a hybridization event resulting in the ancestor of Malinae (possibly Maleae) (Fig. 2d). In addition, 3,931 of 3,939 GDs matched anchor genes in the Malinae WGD-related syntenic blocks in one or more of the *Malus* genomes (Supplementary Fig. 12), supporting the allotetraploidization event resulting from a hybridization between Gillenieae and an unknown lineage. Molecular dating suggested that this WGD event might have contributed to the survival of multiple *Malus* groups through a decrease in paleo-temperature (Fig. 2d and Supplementary Fig. 5b).

Multiple introgressions in *Malus*

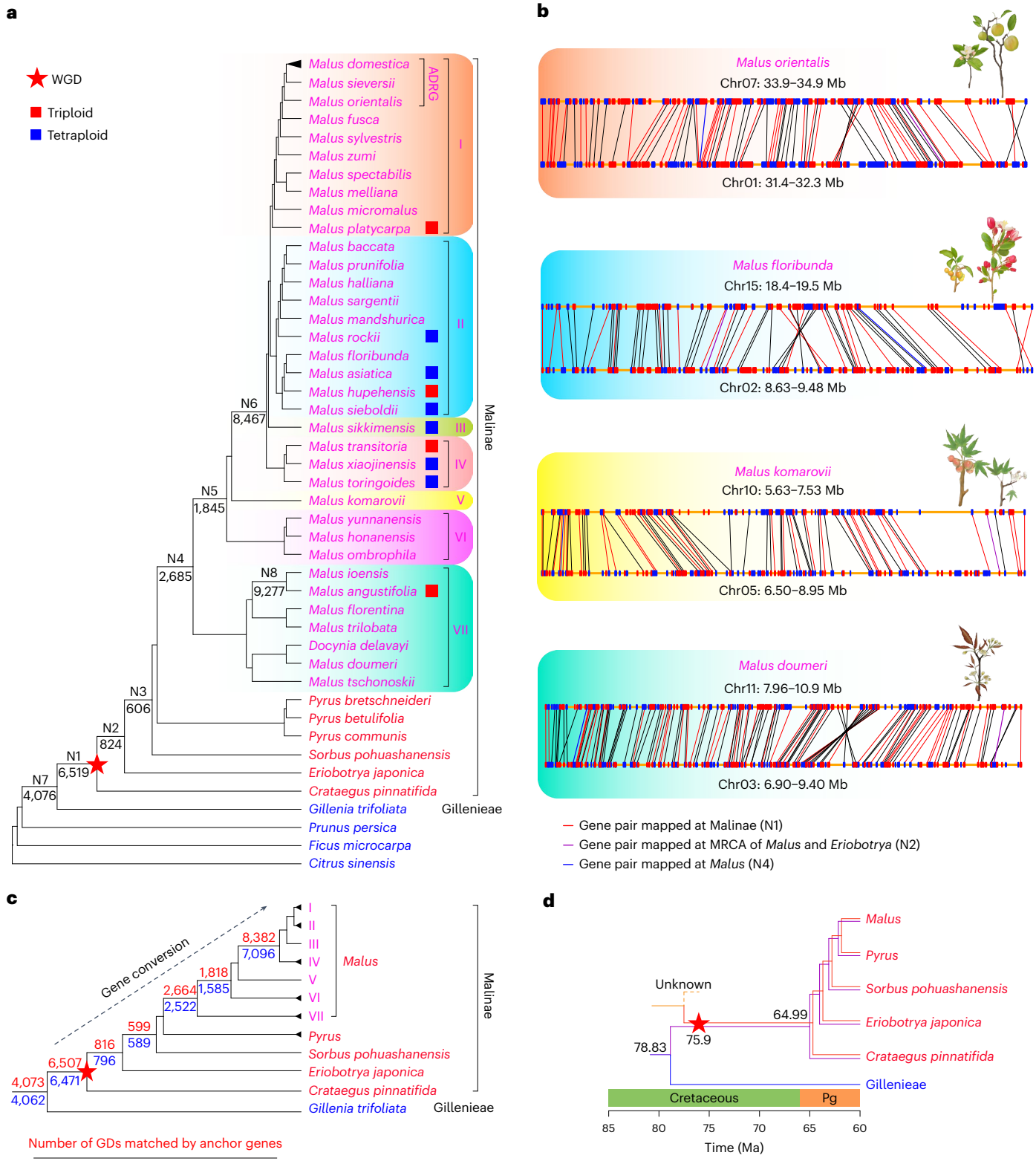
Hybridization or introgression has often been invoked to explain the detection of topological differences between plastid-based species-trees and nuclear-gene-based species-trees (see, for example, refs. 46,47). Phylogenomic investigation of allopolyploids and their

Fig. 2 | Summary of gene duplications mapped onto the *Malus* phylogeny. **a**, The *Malus* phylogenetic tree with focal clusters of GD events. The left species-tree is a part of that in Supplementary Fig. 3d. Red squares to the right of the species name represent triploids, and blue ones represent tetraploids. GD counts of eight focal GD clusters (N1–N8) are placed below the corresponding branches in the species-tree. Red star indicates an ancient WGD event in the early evolution of the apple tribe. **b**, Illustrations of chromosomal collinearity (synteny) for gene pairs in four *Malus* representatives supporting the WGD event. Red and blue rectangles along a yellow line represent protein-coding genes with the transcriptional direction from 3′ to 5′ and 5′ to 3′, respectively. Red lines represent anchor genes with GD mapped at Malinae; purple lines represent anchor genes with GD mapped at the MRCA of *Malus* and *Eriobotrya*; blue lines represent anchor genes with GD mapped at *Malus*. Black lines represent anchor genes either mapped to a node without focal phylogenetic positions or which were not included into gene trees because the focal gene families require more shared species. **c**, An abbreviated version of the phylogenetic tree shown in **a**, with clades I–VII represented by individual branches. The red star marks the WGD event in **a**. The red number above each branch represents the GDs among those shown in **a** that correspond to syntenic anchor genes present in at least one genome from the species included in the relevant clade. The blue number below each branch represents the number of GDs matching the syntenic anchor genes in those syntenic block(s) that also include gene pairs mapped at Malinae and retained in the (AB)(AB) pattern. Detailed metadata of the number of GDs can be found in Supplementary Figs. 6–12. **d**, A model of introgression of genes from Gillenieae to the ancestor of Malinae and its descendant lineages. Purple and red branches of Malinae represent the possible subgenomes derived from Gillenieae and the other progenitor, respectively. Ma, millions of years ago.

https://doi.org/10.1038/s41588-025-02166-6

possible progenitors through the reconciliation between numerous nuclear gene trees and the species-tree can provide strong evidence for phylogenetic relationships and evolutionary history between progenitors (and/or close relatives) and their hybrid progeny (and/or its descendants)[45]. Our newly assembled *Malus* genomes include four triploids and six tetraploids, providing an opportunity to examine the relationships between genes of different putative subgenomes of the polyploid and those of other *Malus* species.

We performed phylogenomic and syntenic analyses of 32 *Malus* genomes and 7 outgroup genomes and detected the *Malus*



https://doi.org/10.1038/s41588-025-02166-6

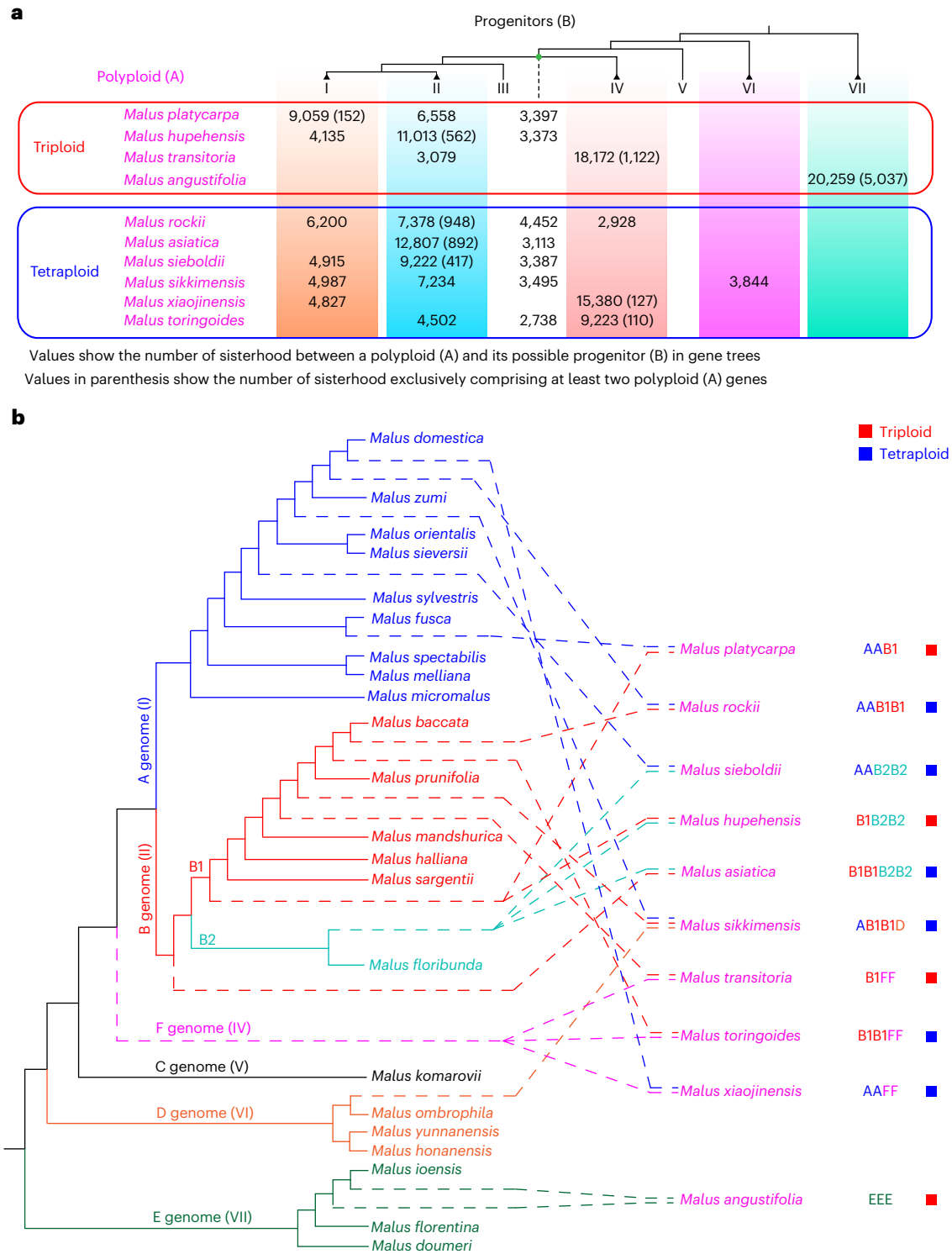


Fig. 3 | Analyses of introgression history in *Malus*. **a**, A summary of the sisterhood of (A, B) in gene trees of the *Malus* orthogroup, where A represents the gene from any focal polyploid and is supported (BS ≥50) as sister to B, which represents the gene(s) from a potential progenitor of the polyploid. At the top is a phylogeny of the *Malus* clades I to VII, derived from the species tree in Fig. 1c. Also in the phylogeny, the dotted line represents the MRCA of clades I to IV, with the numbers below this dotted line for genes from the MRCA of clades I to IV as the potential progenitor. For each of four triploids in the column on the left, up to three numbers are shown below the clade(s) with sisterhood, as support for those clades being potential progenitors. For each of six tetraploids in the left column, up to four numbers are shown below the clade(s) with sisterhood. The number in parenthesis represents the number of gene sisterhoods that have at least two genes from the focal polyploid. The metadata of the sisterhood for each polyploid can be found in Supplementary Figs. 13–22. **b**, An evolutionary model of the polyploids and their possible progenitors. The phylogeny with diploid species (with branches of solid lines) is derived from the *Malus* phylogeny; different colors of the lines correspond to different genome types (A–E genomes; their respective clades in Fig. 1c are shown to the right in parenthesis) represented by closely related diploids, as supported by the tree in Supplementary Fig. 23. Dashed lines correspond to the phylogenetic positions of proposed subgenomes of polyploids. Sister groups of the subgenomes represent their possible progenitors (see support values and phylogeny in Supplementary Fig. 23). The red and blue squares on the right indicate triploids and tetraploids, respectively.

orthogroups (MOGs) (Methods). Here we defined the sisterhood relationship between A and B genes in an MOG of a gene tree when A from one hypothesized subgenome of a *Malus* polyploid is supported (BS ≥ 50) as a sister to B(s) from at least one *Malus* diploid as a possible progenitor. Among the 20,859 identified MOGs, we detected thousands of sisterhood relationships between A genes and B genes of distinct taxa (a single diploid or a group of diploids) (Fig. 3a and Supplementary Figs. 13–22). Moreover, when the MRCA of a clade containing multiple taxa is considered as a possible progenitor, differential lineage-specific gene loss can lead to the recovery of B genes from a subset of taxa derived from the clade; therefore, we included the number of sisterhood relationships with the B(s) from a subset of taxa in one of clades I–VII (see a in Supplementary Figs. 13–22) in the total number of sisterhood relationships matching the MRCA of the entire clade (see b in Supplementary Figs. 13–22). To further estimate phylogenetic relationships between a polyploid and its possible progenitors, we estimated the number of one or more sisterhoods in one MOG and proposed that the top greatest sisterhood(s) relationships supported candidate progenitors for the *Malus* polyploid (Fig. 3a and Supplementary Figs. 13–22).

The identification of genes of polyploids exhibiting close relationships (sisterhoods) with other *Malus* taxa provides an opportunity to re-examine the phylogenetic relationships of putative subgenomes of polyploids with the diploids, including the putative progenitors (Fig. 3a). Specifically, we used the phylogeny of diploid *Malus* species to designate a particular genome type: A through E genomes (Supplementary Fig. 23a). Then the relationships between putative subgenomes of polyploid species and the diploid genomes A–E were analyzed using individual MOG gene trees (Methods), supporting hybridization/introgression events between the A genome and B genome leading to *M. platycarpa*, *M. rockii* and *M. sieboldii* (Supplementary Fig. 23b). Our results also detected introgressions between the species of the B genome (see the B1 and B2 genomic subtypes in Fig. 3b and Supplementary Fig. 23b). In the group with the E genome, our phylogeny and sisterhood analyses supported the diploid *M. ioensis* as a possible progenitor of the triploid *M. angustifolia* (see details in Supplementary Figs. 16c and 23b).

Our phylogeny *Malus* diploid genomes and related polyploid subgenomes supported a sisterhood relationship between two clades: one clade is the MRCA of diploids with A or B genomes as well as the A- or B-related putative subgenomes of polyploids; the other clade is the MRCA of a putative subgenome (F-related genome) of each of three polyploids (including two subgenomes in each of a triploid *M. transitoria* and two tetraploids, *M. toringoides* and *M. xiaojinensis*) (Fig. 3b and Supplementary Fig. 23b). This result suggests that these three polyploid species probably shared a common (tetraploid) progenitor that originated before the divergence of the A and B genomes and after the split of the C genome from the A and B genomes. Our phylogeny also supported species with the D genome (or its close relative) as one possible parent for the tetraploid *M. sikkimensis* (Fig. 3b and Supplementary Fig. 23b). The E genome was supported as a sister to other *Malus* species (Fig. 3b and Supplementary Fig. 23b). In general, our results provide a detailed phylogenetic relationship among diploid and (some of) the subgenomes of polyploid *Malus* species and support a model for the history of the *Malus* genome evolution with proposed progenitors and their polyploid descendants and implications for genome evolution and apple breeding.

Malus pan-genome and structural variation

To investigate the genetic diversity of the *Malus* genomes, we assembled haplotype-resolved (phased) genomes for all 20 diploid species and achieved total sizes of 1.18–1.48 Gb, which were approximately twice the size of the haploid consensus genomes (Supplementary Table 4). BUSCO analyses supported 98.8–99.3% completeness of the 1,614 single-copy Embryophyta genes in the 20 phased diploid assemblies (Supplementary Table 4).

We first detected SVs by comparing 52 haplotype-resolved assemblies from 26 diploid accessions (20 from this study and 6 from previous studies[30,39]) and the consensus assemblies from 10 polyploid accessions to the 'Fuji' consensus genome[39]. The SVs are dispersed throughout euchromatic regions, and the SV density can be elevated around centromeric regions on certain chromosomes (Fig. 4a). Of the SVs, 33.62% were distributed in intergenic regions (Supplementary Fig. 24a,b). We identified 468,006 presence–absence variations (267,667 insertions and 200,339 deletions; ≥50 bp), 2,591 duplications and 1,990 inversions (Fig. 4b). The presence–absence variations had an accumulated length of 1.62 Gb and accounted for 96.58% of all variations. We found a total of 34,930 SVs shared by all seven clades in the *Malus* species tree (Fig. 4c), with only 18.52% located in intergenic regions (Supplementary Fig. 24c). The total number of SVs increased when additional genomes were added and eventually plateaued, suggesting that the vast majority of SVs in the *Malus* genus have been captured (Fig. 4d). Further analysis revealed that 88.54% of these SVs consist of repetitive sequences, mainly LTR-RTs (Supplementary Fig. 25), suggesting that transposable elements probably drive SV formation in *Malus* genomes.

Pan-genome graphs have the power to uncover SVs that may not be captured by traditional linear comparative genomics methods[48,49]. To further explore the genetic diversity of the *Malus* pan-genome, we constructed a pan-*Malus* genome graph containing 3.15 million segments with an accumulated length of 3.18 Gb, including 1.83 million nonreference segments (unaligned to the 'Fuji' reference genome) of 2.45 Gb (76.85% of total). Overall, 117,246 SVs were detected using this graph-based approach (see Methods for the definition of SVs), which were categorized into four types: insertion (31,989), deletion (21,969), divergent (9,577) and multi-allelic (53,711) (Supplementary Fig. 24d). A detailed characterization of accession-specific segments or SVs may provide important guidance for the genetic study of desirable agronomic traits. For instance, apple scab is one of the most serious diseases all over the world[50]. *M. floribunda* has been extensively utilized in scab resistance breeding[51]. In particular, two *M. floribunda*-specific segments (s2410094 and s2410096) were identified in the *Rvi6* region, a scab resistance locus that contains the *HcrVf2* gene (*Mfloribunda-hap2_g48469*) conferring scab resistance[51,52] (Fig. 4e). Notably, the two segments were present in only one haplotype of the phased diploid genome of *M. floribunda* and were absent in the haploid consensus genome, highlighting the advantages of using phased haplotype-resolved assemblies to better understand the functional and structural diversity of *Malus* genomes. We examined the presence of one *M. floribunda*-specific segment, s2410094 (*SCAB-R*), in 16 scab-resistant cultivars derived from crosses with *M. floribunda*. All 16 cultivars were found to carry the *SCAB-R* segment, indicating a strong co-segregation between *SCAB-R* and scab resistance (Supplementary Fig. 26). This finding suggests that *SCAB-R* could be effectively used in marker-assisted selection of scab resistance, particularly when the resistance trait segregates among progenies involving crosses with *M. floribunda* as a parental stock.

Pan-genome graph tool captures selective sweeps for traits

Utilizing diverse genomes in selective sweep analyses can minimize the impacts of genome-specific biases and enhance the detection of sweeps that could be missed with a single reference genome. Here, we developed IntervalConvertor, a tool that converts the selective sweep regions detected from different reference genomes to a standard genome coordinate using the pan-genome graph, allowing comparisons and reconciliations of sweeps from various genome sources in a consistent and uniform manner (Supplementary Fig. 27).

The resequencing data for 337 accessions (247 domesticated apple accessions and 90 wild accessions) were collected for selective sweep analyses to identify genomic regions under selection during apple domestication[7,8,30,53] (Supplementary Table 5). We utilized five genome assemblies as references, including two cultivars ('Fuji' and 'Golden

https://doi.org/10.1038/s41588-025-02166-6

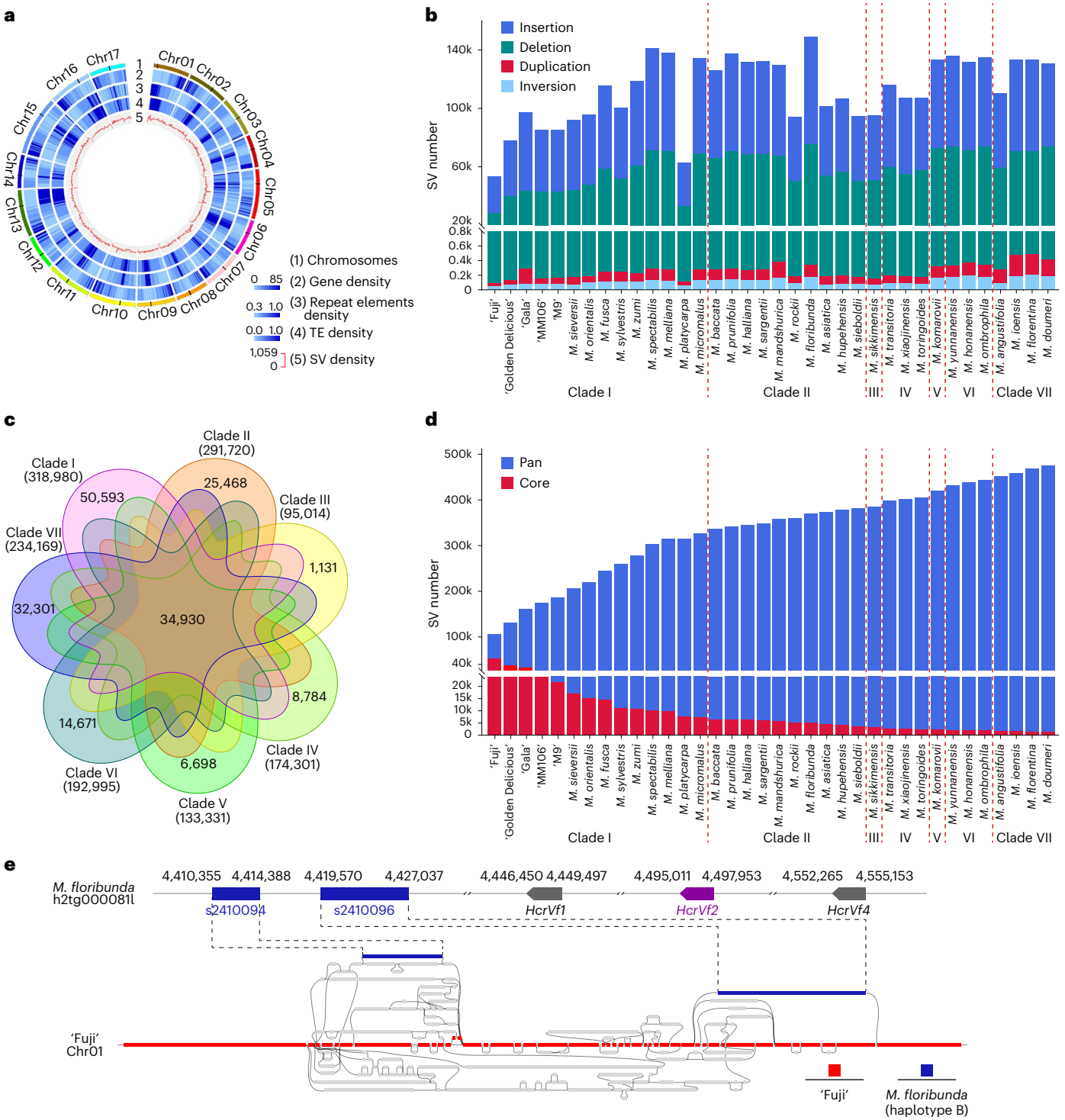


Fig. 4 | Genetic SVs of *Malus* pan-genome. a, A Circos plot of the 'Fuji' genome with genomic features. From outer to inner tracks: (1–4) chromosomes, gene density, repeat elements density and transposable element (TE) density of 'Fuji' genome; (5) density distribution of pan-SV features detected from 36 *Malus* accessions, which were combined and showcased in a pan-*Malus* SV set. The potential centromeric regions on each chromosome are depicted as black bars on track 1. **b**, The composition of SVs in each *Malus* accession. **c**, A seven-way Venn diagram showing shared and clade-specific SVs. Each clade is presented by a distinct color, and the numbers within the colored sections indicate the count of SVs specific to that clade. The number in parentheses under each clade represents the total number of detected SVs in that clade (observed in at least one haplotype of one accession in the clade). A total of 34,930 SVs are commonly detected in all seven clades. **d**, Trend of core- and pan-SVs; the SVs called from additional accessions in the listed order were added iteratively. **e**, Schematic diagram of pan-genome graph paths at the *Rvi6* region of *M. floribunda* haplotig. Red represents the backbone reference genome. Alternative paths are shown in gray. *M. floribunda*-specific segments are shown in blue. The position and orientation of the *HcrVf2* gene are indicated in purple.

Delicious') and three wild relatives (*M. sieversii*, *M. micromalus* and *M. prunifolia*). On the basis of the pan-genome graph, our tool IntervalConvertor converted genomic regions identified from the other four reference accessions to the 'Fuji' genome coordinates. This enabled the identification of selective sweeps and associated genes specific to each assembly (Supplementary Fig. 28a,b), which illustrated that using a single reference genome has limitations in capturing genetic diversity.

Accessions of wild species could exhibit advantages over domesticated accessions in cold tolerance and defense against pathogen invasions. Among those assembly-specific selective sweep regions, we found the *MdMYB5* gene located in one region that could only be detected when using *M. prunifolia* or *M. micromalus* as the reference genome (Fig. 5a). The MYB transcription factors have been shown to mainly regulate the biosynthesis of secondary metabolites, which play key roles in plant stress resistance[54]. When tested in 46 *Malus* accessions, significantly higher expression of *MdMYB5* was observed in wild accessions than in domesticated accessions (Fig. 5b and Supplementary Table 6). One SNP located 260 bp upstream of the ATG start codon of *MdMYB5* was identified to have the C (cytosine) allele instead of T (thymine), with a frequency of homozygous (C/C) being 42.22% in 90 wild accessions but with a relative low frequency of 0.81% in 247 domesticated accessions. To further investigate the impacts of the SNP on *MdMYB5* expression, we transformed apple calli with $MdMYB5^{C}_{pro}$::*GUS* and $MdMYB5^{T}_{pro}$::*GUS*. The β-glucuronidase (GUS) activity of $MdMYB5^{C}_{pro}$::*GUS* was much higher than that of $MdMYB5^{T}_{pro}$::*GUS* (Fig. 5c). The repressive histone mark H3K27me3 was significantly more enriched in the promoter region of *MdMYB5* in domesticated apple accessions with the T/T genotype, compared with wild accessions with the C/C genotype (Supplementary Fig. 28c and Supplementary Table 7). As H3K27me3 is typically associated with gene silencing, this enrichment in domesticated apples aligns with the observed lower expression of *MdMYB5*. These findings reflect a selection pressure on *MdMYB5* during apple domestication, and the identified SNP might be a functional polymorphism related to the variation of *MdMYB5* expression.

To study *MdMYB5* function, we generated *MdMYB5* overexpression (OE) apple lines 2 and 4 with 31- and 28-fold increase in expression and RNA interference (RNAi) lines 5 and 18 with 85% and 71% reduction in the background of GL-3, respectively (Supplementary Fig. 28d,e). Owing to the close association between cuticular wax and the permeability of leaf surfaces[55], we first used Toluidine Blue (TB) staining to assess the cuticular permeability in *MdMYB5* OE and RNAi plants. The results revealed that *MdMYB5* OE plants exhibited significantly less TB staining compared with nontransgenic plants and *MdMYB5* RNAi plants, indicating that *MdMYB5* may alter cuticular permeability (Fig. 5d). We used scanning electron microscopy (SEM) to examine the crystals of cuticular wax and observed a higher density of wax crystals in *MdMYB5* OE plants and a reduced density of wax crystals in *MdMYB5* RNAi plants than in the nontransgenic GL-3 (Fig. 5e). Further investigation of wax composition using gas chromatography-tandem mass spectrometry (GC-MS) analysis revealed similar changes to those of the wax crystals (Fig. 5f and Supplementary Table 8). Nonacosane (C29 alkane) was one of the most abundant components of cuticular wax[56]. When analyzing C29 alkane in 46 apple accessions, we found significantly higher contents in wild accessions than in domesticated accessions (Fig. 5g). Therefore, the domestication process could have selected a *MdMYB5* allele that caused notably decreased cuticular wax biosynthesis.

Cuticular wax on the surface of plants provides essential protection of plants against both biotic and abiotic stresses[57,58]. We first conducted freezing tolerance assays using *MdMYB5* OE and RNAi transgenic plants. Compared with the nontransgenic GL-3 plants, *MdMYB5* OE lines showed increased tolerance to low-temperature stress. In contrast, *MdMYB5* RNAi plants exhibited more severe damage than the nontransgenic plants (Fig. 5h). These results indicate that *MdMYB5* positively regulates freezing tolerance in apple. To further evaluate the biological function of *MdMYB5* against pathogen attack, we infected detached leaves with *Alternaria alternata* (the causal agent for *Alternaria* blotch disease in apple)[59]. After 3 days, the *MdMYB5* OE plants showed a smaller lesion area than the nontransgenic plants. Conversely, the *MdMYB5* RNAi plants had the largest lesion area, indicating that *MdMYB5* plays a positive role in resistance of apple to *A. alternata* infection (Fig. 5i,j). Therefore, the lower expression of *MdMYB5* observed in domesticated accessions, compared with wild accessions, might be a cause of the weakened stress tolerance during the domestication process[54,58,60]. Collectively, the pan-genome graph created in this study, along with the newly developed bioinformatics pipeline, greatly facilitates the utilization of additional reference *Malus* genomes in genome-wide association study or selective sweep analysis, and thus largely increase the efficiency of genetic studies of key agronomy traits.

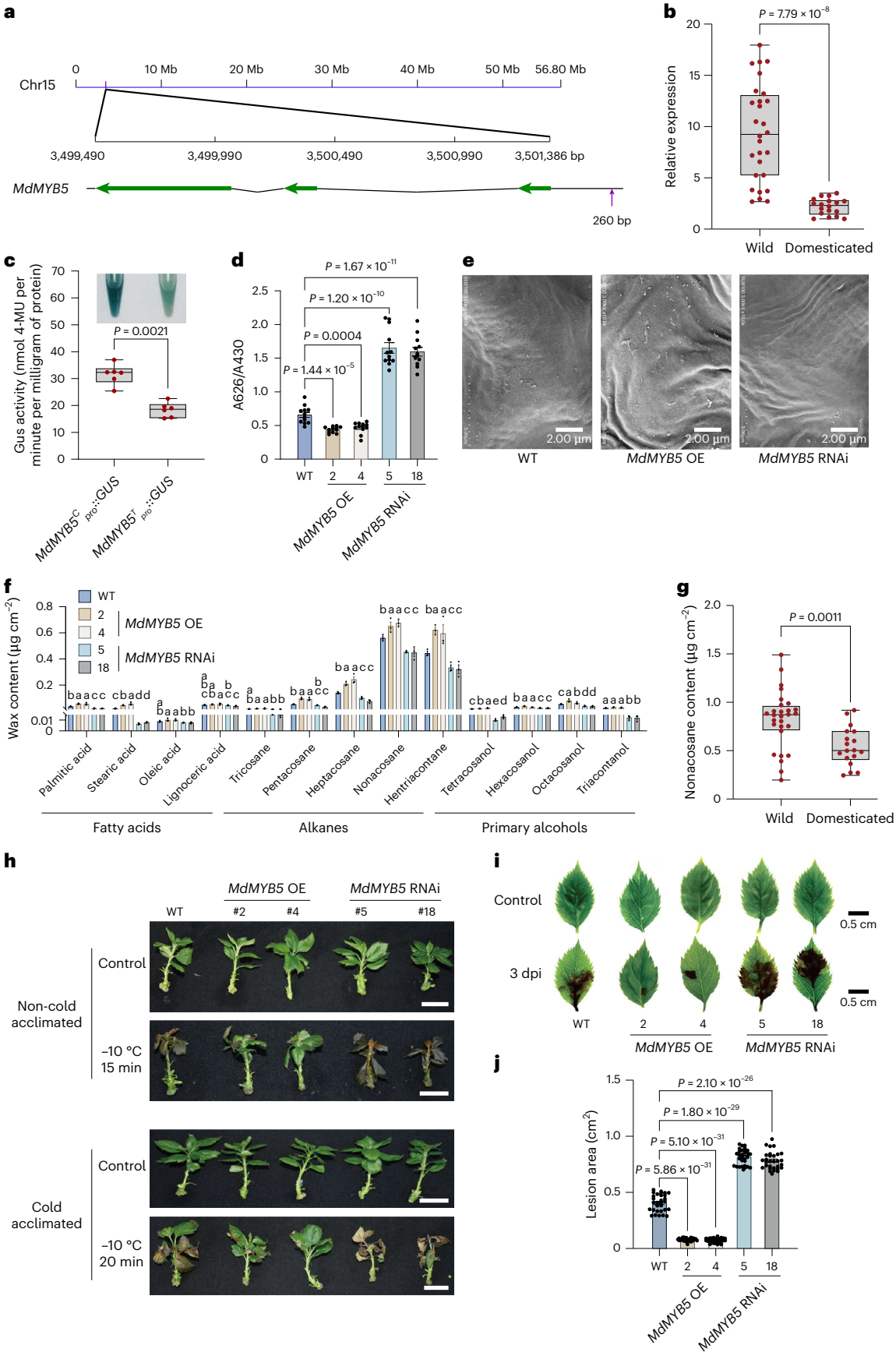
Discussion

The 30 newly assembled *Malus* genomes herein represent the first genus-scale pan-genome for fruit trees (Supplementary Note 2). Many *Malus* species are capable of interspecific hybridization, leading to new germplasm with valuable traits such as disease resistance, dwarfing and ornamental appeal (Supplementary Note 3). Comprehensive genomic analysis of both wild and domesticated *Malus* species enables breeders to effectively exploit interspecific hybridization for the introgression and enhancement of desirable traits in novel apple cultivars. For instance, *Malus floribunda*, known for its resistance to apple scab, has been crossed with *Malus domestica* to create scab-resistant cultivars such as 'Prima', 'Priscilla' and 'Florina'[61,62]. Our pan-genome graph analysis has identified two *M. floribunda*-specific segments, which facilitated the development of a molecular marker for tracking the introgression of scab resistance in interspecific hybridization. This advancement will enable breeders to efficiently incorporate *M. floribunda*'s resistance traits, accelerating the development of scab-resistant apple cultivars.

In this study, we have developed a bioinformatics tool, IntervalConvertor, to promote a multi-genome approach for selective sweep analysis. Our results strongly suggest the benefits and effectiveness

Fig. 5 | *MdMYB5* was under selection pressure during apple domestication and promoted both biotic and abiotic stress tolerances. a, The location of *MdMYB5* in the 'Fuji' genome. The green arrows show protein-coding sequences. The purple arrow indicates the SNP associated with the expression of *MdMYB5*. **b**, The relative expression levels of *MdMYB5* in wild and domesticated apple accessions ($n = 28$ and 18, respectively). **c**, The GUS activity analysis of transgenic apple calli carrying $MdMYB5^{C}_{pro}$::*GUS* or $MdMYB5^{T}_{pro}$::*GUS* ($n = 6$ independent experiments). **d**, Quantification of TB staining. The TB absorbance (A626) is normalized by chlorophyll absorbance (A430) ($n = 12$ independent experiments). **e**, SEM images of leaf cuticular wax crystals from 1-month-old WT, *MdMYB5* OE and RNAi plants. Images are representatives of three independent experiments across three apple plants with similar results. Scale bars, 2 μm. **f**, Leaf cuticular wax compositions and contents of WT, *MdMYB5* OE and RNAi plants. Leaves were harvested from 1-month-old tissue-cultured plants ($n = 3$ independent experiments). The *P* values are listed in Supplementary Table 8. **g**, The leaf nonacosane (C29 alkane) content in wild and domesticated apple accessions ($n = 28$ and 18, respectively). **h**, Freezing tolerance performance of WT, *MdMYB5* OE and RNAi plants with or without cold acclimation. Plants were photographed 5 days after recovery. Scale bars, 2.5 cm. **i**, Disease symptoms of WT, *MdMYB5* OE and RNAi leaves. dpi, days post-infection with *Alternaria alternata*. Scale bars, 0.5 cm. **j**, Quantification analysis of the lesion area for the leaves in **i** ($n = 30$ independently collected leaves from different plants). Error bars represent the standard error on the mean (s.e.m.). WT, wild-type GL-3. For **b**–**d**, **g** and **j**, *P* values were determined using a two-tailed Student's *t*-test. For **f**, one-way analysis of variance (ANOVA) (Tukey's test) was performed for each group. Different letters indicate significant differences ($P < 0.05$). For **b**, **c** and **g**, the center line indicates a median value, the lower and upper hinges represent the 25th and 75th percentiles and the whiskers extend to the minimum and maximum values.

https://doi.org/10.1038/s41588-025-02166-6



of using a graph-based pan-genome in detecting potentially important genomic loci for apple breeding. Our findings suggest that one SNP probably alters the chromatin state of the *MdMYB5* promoter. In wild accessions with the C/C genotype, the chromatin state appears more permissive, thereby enhancing *MdMYB5* expression. In contrast, in domesticated apples with the T/T genotype, the chromatin configuration may be altered by the SNP and results in a decreased *MdMYB5* expression, potentially contributing to reduced cold and disease resistance in domesticated accessions. Moreover, the increase in flavonoid content along with *MdMYB5* overexpression suggests that domestication may have selected against higher *MdMYB5* expression to alter the production of secondary metabolites that affect fruit taste (Supplementary Fig. 29). By selecting for tasty fruit, humans might have inadvertently favored lower *MdMYB5* expression. This hypothesis presents an intriguing area for further research to better understand the role of *MdMYB5* in fruit flavor and the domestication process.

The 30 high-quality chromosomal-level genomes have enabled a comprehensive phylogenomic analysis with insights into genome duplications and multiple introgressions, which could be major drivers for the genome diversity and radiation in *Malus*. Our results also reveal insights into adaptation to the changing environment as a driver for early *Malus* evolution and diversification. In addition, there was phylogenomic support for hybridization between *Malus* lineages for some *Malus* triploid and tetraploid species (Fig. 3 and Supplementary Figs. 13–23). For example, *M. platycarpa* has nearly twice as many genes close to lineage I (subgenome A) as to lineage II (subgenome B1), suggesting that it descended from a hybridization between a diploid with a B1 subgenome and a tetraploid with two copies of the A subgenome. However, this putative tetraploid is possibly extinct or not included in the current study. Similarly, a diploid of lineage II (B1) and a tetraploid with FF genomes might have been the parents of *M. transitoria*. In *M. angustifolia*, a much larger number of genes are close to lineage VII, including an unusually large number of gene trees with two copies, suggesting that its parents might have been a diploid with E and a tetraploid with EE subgenomes. *M. hupehensis* has more than twice the number of genes close to lineage II (B1 and B2) as those close to lineage I (A), implying gene introgression involved in A and B subgenomes. Overall, these findings have advanced our understanding of the *Malus* genome diversity and evolution.

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/s41588-025-02166-6.

References

1. Shu, H., et al. *Apple Science* (China Agriculture Press, 1999).
2. Phipps, J. B. *Flora of North America, Vol. 9, North of Mexico Magnoliophyta: Picramniaceae to Rosaceae* (Oxford Univ. Press, 2014).
3. Korban, S. S. Interspecific hybridization in *Malus*. *HortScience* **21**, 41–48 (1986).
4. Harris, S. A., Robinson, J. P. & Juniper, B. E. Genetic clues to the origin of the apple. *Trends Genet.* **18**, 426–430 (2002).
5. Cornille, A., Giraud, T., Smulders, M. J., Roldán-Ruiz, I. & Gladieux, P. The domestication and evolutionary ecology of apples. *Trends Genet.* **30**, 57–65 (2014).
6. Khan, A., Gutierrez, B., Chao, C. T. & Singh, J. in *The Apple Genome* (ed. Korban, S. S.) 383–394 (Springer, 2021).
7. Liao, L. et al. Unraveling a genetic roadmap for improved taste in the domesticated apple. *Mol. Plant* **14**, 1454–1471 (2021).
8. Duan, N. et al. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nat. Commun.* **8**, 249 (2017).
9. Liu, B. B. et al. Phylogenomic conflict analyses in the apple genus *Malus* s.l. reveal widespread hybridization and allopolyploidy driving diversification, with insights into the complex biogeographic history in the Northern Hemisphere. *J. Integr. Plant Biol.* **64**, 1020–1043 (2022).
10. Xiang, Y. et al. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* **34**, 262–281 (2017).
11. Sacerdot, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* **19**, 166 (2018).
12. Xu, P. et al. The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nat. Commun.* **10**, 4625 (2019).
13. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341–345 (2006).
14. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
15. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
16. Zhang, C. et al. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol. Biol. Evol.* **37**, 3188–3210 (2020).
17. Cai, L. et al. Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *N. Phytol.* **221**, 565–576 (2019).
18. Soltis, P. S. & Soltis, D. E. Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* **30**, 159–165 (2016).
19. Cheng, F. et al. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* **4**, 258–268 (2018).
20. Yuan, Z. et al. The pomegranate (*Punica granatum* L.) genome provides insights into fruit quality and ovule developmental biology. *Plant Biotechnol. J.* **16**, 1363–1374 (2018).
21. Chen, D., Zhang, T., Chen, Y., Ma, H. & Qi, J. Tree2GD: a phylogenomic method to detect large-scale gene duplication events. *Bioinformatics* **38**, 5317–5321 (2022).
22. Velasco, R. et al. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
23. Brown, S. in *Fruit Breeding* (eds. Badenes, M. & Byrne, D.) 329–367 (Springer, 2012).
24. Brozynska, M., Furtado, A. & Henry, R. J. Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol. J.* **14**, 1070–1085 (2016).
25. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 (2020).
26. Tang, D. et al. Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).
27. Zhang, L. et al. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit color. *Nat. Commun.* **10**, 1494 (2019).
28. Liu, Y. et al. Multi-omics analyses reveal *MdMYB10* hypermethylation being responsible for a bud sport of apple fruit color. *Hortic. Res.* **9**, uhac179 (2022).
29. Tian, Y. et al. Transposon insertions regulate genome-wide allele-specific expression and underpin flower color variations in apple (*Malus* spp.). *Plant Biotechnol. J.* **20**, 1285–1297 (2022).
30. Sun, X. et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* **52**, 1423–1432 (2020).

31. Li, Z. et al. Chromosome-scale reference genome provides insights into the genetic origin and grafting-mediated stress tolerance of *Malus prunifolia*. *Plant Biotechnol. J.* **20**, 1015–1017 (2022).
32. Li, X. et al. Improved hybrid de novo genome assembly of domesticated apple (*Malus × domestica*). *Gigascience* **5**, 1–5 (2016).
33. Daccord, N. et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
34. Cai, Y. et al. Genome sequencing of 'Fuji' apple clonal varieties reveals genetic mechanism of the spur-type morphology. *Nat. Commun.* **15**, 10082 (2024).
35. Khan, A. et al. A phased, chromosome-scale genome of 'Honeycrisp' apple (*Malus domestica*). *GigaByte* **2022**, gigabyte69 (2022).
36. Wang, T. et al. Pan-genome analysis of 13 *Malus* accessions reveals structural and sequence variations associated with fruit traits. *Nat. Commun.* **14**, 7377 (2023).
37. Mansfeld, B. N. et al. A haplotype resolved chromosome-scale assembly of North American wild apple *Malus fusca* and comparative genomics of the fire blight *Mfu10* locus. *Plant J.* **116**, 989–1002 (2023).
38. Švara, A., Sun, H., Fei, Z. & Khan, A. Advancing apple genetics research: *Malus coronaria* and *Malus ioensis* genomes and a gene family-based pangenome of native North American apples. *DNA Res.* **31**, dsae026 (2024).
39. Li, W. et al. Near-gapless and haplotype-resolved apple genomes provide insights into the genetic basis of rootstock-induced dwarfing. *Nat. Genet.* **56**, 505–516 (2024).
40. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
41. Ebersberger, I., Strauss, S. & von Haeseler, A. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.* **9**, 157 (2009).
42. Liu, G. N. et al. *Malus* includes *Docynia* (Maleae, Rosaceae): evidence from phylogenomics and morphology. *PhytoKeys* **229**, 47–60 (2023).
43. Li, H., Huang, C. H. & Ma, H. in *The Pear Genome* (ed. Korban, S. S.) 279–299 (Springer, 2019).
44. Zhang, T. et al. Cultivated hawthorn (*Crataegus pinnatifida* var. major) genome sheds light on the evolution of Maleae (apple tribe). *J. Integr. Plant Biol.* **64**, 1487–1501 (2022).
45. Zhang, T. et al. Phylogenomic profiles of whole-genome duplications in Poaceae and landscape of differential duplicate retention and losses among major Poaceae lineages. *Nat. Commun.* **15**, 3305 (2024).
46. Zhang, C. et al. Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. *J. Integr. Plant Biol.* **63**, 1273–1293 (2021).
47. Stull, G. W., Pham, K. K., Soltis, P. S. & Soltis, D. E. Deep reticulation: the long legacy of hybridization in vascular plant evolution. *Plant J.* **114**, 743–766 (2023).
48. Leger, A. et al. Genomic variations and epigenomic landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel. *Genome Biol.* **23**, 1–24 (2022).
49. Li, R. et al. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res.* **33**, 463–477 (2023).
50. Bénaouf, G. & Parisi, L. Genetics of host-pathogen relationships between *Venturia inaequalis* races 6 and 7 and *Malus* species. *Phytopathology* **90**, 236–242 (2000).
51. Belfanti, E. et al. The *HcrVf2* gene from a wild apple confers scab resistance to a transgenic cultivated variety. *Proc. Natl Acad. Sci. U.S.A.* **101**, 886–890 (2004).
52. Khajuria, Y. P., Kaul, S., Wani, A. A. & Dhar, M. K. Genetics of resistance in apple against *Venturia inaequalis* (Wint.) Cke. *Tree Genet. Genomes* **14**, 16 (2018).
53. Chen, P. et al. Insights into the effect of human civilization on *Malus* evolution and domestication. *Plant Biotechnol. J.* **19**, 2206–2220 (2021).
54. Zhang, Y. L. et al. The R2R3 MYB transcription factor MdMYB30 modulates plant resistance against pathogens by regulating cuticular wax biosynthesis. *BMC Plant Biol.* **19**, 362 (2019).
55. Tanaka, T., Tanaka, H., Machida, C., Watanabe, M. & Machida, Y. A new method for rapid visualization of defects in leaf cuticle reveals five intrinsic patterns of surface defects in *Arabidopsis*. *Plant J.* **37**, 139–146 (2004).
56. Miller, S. S., Reid, L. M., Butler, G., Winter, S. P. & McGoldrick, N. J. Long chain alkanes in silk extracts of maize genotypes with varying resistance to *Fusarium graminearum*. *J. Agric. Food Chem.* **51**, 6702–6708 (2003).
57. Hietala, T., Mozes, N., Genet, M. J., Rosenqvist, H. & Laakso, S. Surface lipids and their distribution on willow (*Salix*) leaves: a combined chemical, morphological and physicochemical study. *Colloids Surf. B* **8**, 205–215 (1997).
58. Yeats, T. H. & Rose, J. K. The formation and function of plant cuticles. *Plant Physiol.* **163**, 5–20 (2013).
59. Abe, K., Iwanami, H., Kotoda, N., Moriya, S. & Takahashi, S. Evaluation of apple genotypes and *Malus* species for resistance to Alternaria blotch caused by *Alternaria alternata* apple pathotype using detached-leaf method. *Plant Breed.* **129**, 208–218 (2010).
60. Bourdenx, B. et al. Overexpression of Arabidopsis *ECERIFERUM1* promotes wax very-long-chain alkane biosynthesis and influences plant response to biotic and abiotic stresses. *Plant Physiol.* **156**, 29–45 (2011).
61. Bannier, H. J. Moderne Apfelzüchtung: Genetische Verarmung und Tendenzen zur Inzucht. *Erwerbs-Obstbau* **52**, 85–110 (2011).
62. Janick, J. The PRI apple breeding program. *HortScience* **41**, 8–10 (2006).

[1]Institute for Horticultural Plants, China Agricultural University, Beijing, China. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Department of Biology, Eberly College of Science and Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, USA. [4]State Key Laboratory of Crop Stress Biology for Arid Areas/Shaanxi Key Laboratory of Apple, College of Horticulture, Northwest A&F University, Yangling, China. [5]Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Chinese Academy of Agricultural Sciences, Shenzhen, China. [6]College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China. [7]Department of Plant Science and Landscape Architecture, University of Connecticut, Storrs, CT, USA. [8]Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou, China. [9]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China. [10]Section of Horticulture, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA. [11]The New Zealand Institute for Plant and Food Research Limited (Plant and Food Research), Auckland, New Zealand. [12]These authors contributed equally: Wei Li, Chong Chu, Taikui Zhang, Haochen Sun, Shiyao Wang, Zeyuan Liu. ✉e-mail: chong_chu@hms.harvard.edu; qguan@nwafu.edu.cn; cecilia.deng@plantandfood.co.nz; hanym@mail.buct.edu.cn; hxm16@psu.edu; rschan@cau.edu.cn

Methods

Sample collection

The *Malus* species samples used in this study for DNA and RNA sequencing were collected from seven different locations, including: (1) 23 species from the National Germplasm Repository of Pear and Apple (Xingcheng), Huludao, China (40.70° N, 120.75° E), (2) *M. ombrophila* and *M. florentina* from the Horticulture Experimental Station of Northwest A&F University, Yangling, China (34.30° N, 108.08° E), (3) *M. yunnanensis* and *M. tschonoskii* from the Qingdao Academy of Agricultural Sciences, Qingdao, China (36.24° N, 120.55° E), (4) *M. komarovii* from the National Field Genebank for Hardy Fruits (Gongzhuling), Changchun, China (43.50° N, 124.88° E), (5) *M. doumeri* from Shanghai Botanical Garden, Shanghai, China (31.15° N, 121.45° E), (6) *M. xiaojinensis* from China Agricultural University, Beijing, China (40.03° N, 116.29° E) and (7) 'Golden Delicious' from Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou, China (34.72° N, 113.71° E).

Genome assembly

For diploid accessions, the haploid consensus genome was generated using hifiasm v0.15.2_r334[63] with HiFi data and default settings. HiC-pro v3.1.0[64] was run on a high-performance computing environment using a conda image. Hi-C data were mapped to the de novo assembly, and the contact bin matrix was created. The matrix was fed to EndHiC v1.0[65] for scaffolding purposes. To achieve chromosome-level assembly, manual curation was conducted through visually inspecting Hi-C contact maps using Juicebox v1.11.08[66]. The resulting sequences were aligned to the 'Fuji' chromosomes utilizing 'minimap2 -x asm5' (v2.18-r1015)[67], and the corresponding chromosome number was assigned on the basis of the synteny. For haplotype-resolved contigs, a 'Hi-C integration' mode of hifiasm v0.15.2_r334[63] was employed to phase the assemblies, on the basis of Hi-C data.

For polyploid accessions, the haploid consensus genome was first generated using hifiasm v0.15.2_r334[63] with HiFi data and default settings. The HiFi data were aligned to the de novo assembly using minimap2 v2.18-r1015[67], the alignments were compressed with gzip, the primary assembly and alternative haplotigs were then separated using 'purge_dups -2' (v1.2.5)[68] and the sequences for the purged results were obtained using command 'get_seqs' within the purge_dups package. The primary assembly sequences were aligned to the 'Fuji' chromosomes using 'minimap2 -x asm5' (v2.18-r1015)[67] to confirm that all 17 chromosomes were adequately represented. Following the scaffolding workflow for diploid accessions (combining HiC-Pro, EndHiC, Juicebox and minimap2), chromosome-level consensus assembly was achieved for polyploid accessions. In the case of unanchored contigs and the alternative haplotigs separated in the purge_dups step, the same workflow as described above was used for further scaffolding. The final polyploid assembly included both the chromosomes and the unanchored scaffolds, maximizing the retention of genomics information.

To assess the integrity of the genome assemblies, we employed two methods: BUSCO completeness and LAI (see details in Supplementary Note 4).

Genome annotation

For repetitive elements annotation, RepeatModeler v2.0.1 (https://github.com/Dfam-consortium/RepeatModeler) was run with parameter '-LTRStruct' to construct a de novo high-quality nonredundant transposable elements library. Next, RepeatMasker v4.1.0 (https://www.repeatmasker.org/RepeatMasker/) was used to annotate and mask the repetitive elements of the assemblies by combining the constructed de novo libraries with the Repbase database[69]. To identify centromeric regions, we used multiple approaches (see details in Supplementary Note 5). Gene prediction was performed with BRAKER pipeline v2.1.6 (see details in Supplementary Note 6). For gene function annotation, BLASTP v2.10.1+ was used to query the NCBI-nr database of the predicted genes.

Evolutionary analyses

To explore the evolutionary relationships among *Malus* genomes, we performed phylogenetic analyses (see details in Supplementary Note 7), molecular dating (see details in Supplementary Note 8), biogeographical analyses (see details in Supplementary Note 9), WGD analyses (see details in Supplementary Note 10) and ancestral hybridization analyses (see details in Supplementary Note 11). Briefly, to reconstruct phylogenetic relationships among *Malus* species, we identified single-copy nuclear genes and performed ASTRAL analyses of gene trees using three different single-copy gene datasets (998, 799 and 661 genes). Using the species-tree generated here, we performed Bayesian molecular dating with 12 fossils to date the divergence times in *Malus*. Then, we inferred the possible biogeographical events among *Malus* species according to the *Malus* time-tree and the published data of the biogeographical distribution. In addition, we detected GDs shared by *Malus* genomes and identified syntenic genes in each genome. We estimated the number of GDs matched by syntenic gene pairs to assess the implication of WGD in *Malus* genome evolution. Furthermore, we applied phylogenomic analyses to detect the MOG from gene trees and to infer possible hybridization event(s) among *Malus* species from the MOGs.

Pan-SV detection

Thirty-six accessions were used for pan-SV construction: (1) 20 diploid accessions with haplotype-resolved assemblies, (2) six previously reported chromosome-level haplotype-resolved assemblies ('Gala', *M. sieversii*, *M. sylvestris*, 'Fuji', 'M9' and 'MM106')[30,39] and (3) 10 polyploidy accessions with purged haploid consensus assemblies. Two approaches were utilized to call SVs from the assemblies: reference-based and pan-genome graph-based.

Reference-based SV detection from assemblies. Each of the assemblies was aligned to the 'Fuji' assembly using minimap2 v2.21[67] with parameter '-x asm5 -a --eqx --cs'. The alignments were sorted, compressed and indexed using SAMtools[70]. SVs were called accordingly using svim-asm diploid v1.0.2[71] with parameter '--query_names --interspersed_duplications_as_insertions --min_sv_size 50'. The generated raw VCF files were sorted and indexed using BCFtools[70]. Lastly, we used SURVIVOR v.1.0.7[72] to merge VCF files for each accession to an integrated VCF file.

Pan-genome graph-based SV detection. First, assembled sequences were renamed with a prefix of their Latin names to track the source of variations (for example, Mkomarovii_h1 for *M. komarovii* haplotype A and Mkomarovii_h2 for *M. komarovii* haplotype B). The pan-genome graph was generated using Minigraph v0.15-r426[73] with the parameter '-xggs'. The 'Fuji' haploid consensus assembly was used as the initial backbone. The hapA and hapB from 26 diploid accessions and purged haploid consensus assemblies from 10 polyploid accessions were added iteratively. Bandage v0.8.1[74] was used to visualize the pan-genome graph.

The generated *Malus* pan-genome graph contains chains of bubbles with the 'Fuji' genome as the backbone. Each bubble represented a possible SV. Bubbles were called using gfatools v0.5 (https://github.com/lh3/gfatools), outputting details such as the start and end nodes on the reference sequences as well as the paths connecting these nodes. Four types of SVs were characterized on the basis of the number of paths in a bubble: bi-allelic (insertion/deletion) with two paths, where the shorter path is <50 bp and the longer path is ≥50 bp (defined as insertion or deletion by comparison with the reference path); divergent with two paths, where the shorter path is ≥50 bp; and multi-allelic with more than two paths, where the shortest path is ≥50 bp.

Genotype analysis of scab-resistant apple cultivars

Leaf samples of scab-resistant cultivars were collected from the United States Department of Agriculture Agricultural Research Service (USDA-ARS) Plant Genetic Resources Unit in Geneva, New York, for genomic DNA extraction to analyze the genotype of the 4,034-bp insertion (s2410094). PCR amplification was carried out using 2× DreamTaq Green PCR Master Mix (Thermo Scientific, K1081) with the following conditions: 95 °C for 3 min; 36 cycles of 95 °C for 15 s, 62 °C for 15 s and 72 °C for 15 s, followed by a final extension at 72 °C for 5 min. Primers (as detailed in Supplementary Table 9) were used to generate a 774-bp product corresponding to the 4,034-bp insertion and a 174-bp product for *MdActin* as a positive control. The PCR products were visualized on a 1% agarose gel.

Pan-genome guided identification of selective sweeps

Mapping and variant calling. To identify genomic regions selected during apple domestication, the resequencing data of 337 *Malus* accessions (247 cultivars and 90 wild accessions) were collected from public databases (Supplementary Table 5). Five genome assemblies ('Fuji', 'Golden Delicious', *M. sieversii*, *M. prunifolia* and *M. micromalus*) were employed as the reference genomes, respectively (see details in Supplementary Note 12).

Genome-wide scan for selective sweeps. Genome-wide detection of selective sweeps in each genome was performed based on the filtered SNPs. Nucleotide diversity (π) and population fixation statistics (F_{ST}) were calculated using VCFtools v0.1.16[75] with a sliding window of 20 kb and step size of 1 kb. Reduction of diversity was calculated from the nucleotide diversity (π) across the entire reference genome. The regions with the top 5% for both F_{ST} and reduction of diversity were considered as candidate regions for selective sweeps.

Conversion of selective sweeps. To convert the selective sweeps identified from five different genomes into a consistent coordinate system, we developed this workflow: (1) a pan-genome graph was constructed from the five assemblies using 'Fuji' as the backbone, (2) each of the other four reference genomes was aligned to the constructed genome graph using Minigraph v0.15-r426[73] with parameter '-x lr', and a Graph Alignment Format file was generated and (3) a conversion tool, IntervalConvertor, was developed, and the selective sweeps identified in each genome were converted into coordinates in the 'Fuji' haploid consensus genome, on the basis of the corresponding Graph Alignment Format file (Supplementary Fig. 27).

RNA extraction and qRT-PCR analysis of *MdMYB5*

The cetyltrimethylammonium bromide method was used to extract total RNA from apple leaves of 46 accessions (listed in Supplementary Table 6)[76]. DNA was digested using RNase-free DNase I (Thermo Scientific, EN0521). For quantitative reverse transcription polymerase chain reaction (qRT-PCR) analysis, 1 µg of total RNA was reverse transcribed into first-strand cDNA using the Hifair AdvanceFast One-step RT-gDNA Digestion SuperMix for qPCR kit (YEASEN, 11151ES). The qRT-PCR was performed using Hieff qPCR SYBR Green Master Mix (High Rox Plus) (YEASEN, 11203ES). *MdMDH* served as the reference gene. Three independent experiments were conducted for each accession.

Promoter activity analysis and ChIP-qPCR assay

The promoter (1,869 bp) of *MdMYB5* was amplified from GL-3 and cloned into the binary expression vector pMDC164 (TAIR accession no. 1009003759) through Gateway technology, resulting in the $MdMYB5^{T}_{pro}$::*GUS* vector. The $MdMYB5^{C}_{pro}$::*GUS* vector, which differed only at 260 bp upstream of the ATG start codon of *MdMYB5*, was also prepared in the similar way. Both vectors were used to transform apple calli through *Agrobacterium tumefaciens*-mediated transformation, respectively[76]. GUS staining was conducted[77], and GUS activity was quantified using a GUS Gene Quantitative Detection Kit (Coolaber, SL7161-50T) as per the manufacturer's instructions, with at least six independent experiments performed. A chromatin immunoprecipitation-quantitative PCR (ChIP-qPCR) assay was conducted on apple leaves of 10 distinct apple accessions (see details in Supplementary Note 13 and Supplementary Table 7).

Apple transformation and phenotype analysis

To obtain the *MdMYB5* overexpression vector, the full-length cDNA sequence of *MdMYB5* was introduced into the binary expression vector pK2GW7 (TAIR accession no. 6531113855) driven by the CaMV 35S promoter using the Gateway technology. A 324-bp fragment of the *MdMYB5* coding region was amplified using the PCR primers (listed in Supplementary Table 9) and introduced into the pK7GWIWG2D vector[76] to generate the RNAi plasmid. These plasmids were introduced into *A. tumefaciens* strain EHA105, followed by transformation into apple GL-3 materials[78] (refer to the transformation protocol in Supplementary Note 14). The TB penetration test, SEM analysis, freezing tolerance and disease resistance were conducted based on wild-type GL-3, and *MdMYB5* OE and RNAi plants[79,80] (see details in Supplementary Note 15).

Quantification of leaf cuticular wax

Cuticular wax was extracted from leaves and analyzed using GC-MS coupled with the Trace GC ULTRA/ISQ MS detector (Thermo Scientific)[81]. The conversion of GC-MS data to wax content was performed using the following formula: sample wax component content = $(A \times B)/(C \times D)$ where A represents the internal standard content in micrograms, B is the peak area of the sample, C is the peak area of the internal standard and D is the leaf area in square centimeters. Cuticular wax quantification was carried for wild-type GL-3, *MdMYB5* OE and RNAi plants, as well as 46 apple accessions (listed in Supplementary Table 6).

Flavonoid content analysis

We transformed 'Orin' calli, and used qRT-PCR and western blot to confirm the presence and expression of the transgene[76] (see details in Supplementary Note 16). Flavonoids were extracted from apple calli using 1% (v/v) HCl-methanol[82]. In brief, 1 g of powdered apple calli was incubated in 10 ml of 1% (v/v) HCl-methanol at 4 °C overnight. After centrifugation, the supernatant was collected for the subsequent reaction. The reaction buffer consisted of 100 g l^{-1} $Al(NO_3)_3$, 50 g l^{-1} $NaNO_2$ and 40 g l^{-1} NaOH. Following the reaction, the mixture was centrifuged, and the absorbance of the supernatant was measured at 510 nm using a microplate reader. Experiments were independently performed at least three times.

Statistical analyses

For statistical analysis, a two-tailed Student's *t*-test was used to evaluate differences in the following: (1) relative expression levels of *MdMYB5*, leaf nonacosane (C29 alkane) content and relative H3K27me3 enrichment of the *MdMYB5* promoter between wild and domesticated apple accessions, (2) GUS activity between $MdMYB5^{C}_{pro}$::*GUS* and $MdMYB5^{T}_{pro}$::*GUS* transgenic apple calli and (3) TB staining quantification, leaf lesion area, *MdMYB5* relative expression levels and flavonoid content between wild-type and transgenic materials. For multiple-group comparisons of leaf cuticular wax content, significance was analyzed using one-way analysis of variance (ANOVA, $P < 0.05$) followed by Tukey's test, with significant differences indicated by different letters.

Reporting summary

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

Data availability

All raw sequencing reads of PacBio HiFi, Hi-C and RNA-seq have been deposited to the Sequence Read Archive at the National Center for

Biotechnology Information (NCBI) database with BioProject no. PRJNA861686. The 30 *Malus* species genome assemblies are available in the NCBI under BioProject no. PRJNA1062995. Additionally, all accession numbers, genome assemblies and annotation data for the 30 *Malus* species can be accessed at the Genome Warehouse in the National Genomics Data Center, China National Center for Bioinformation with BioProject no. PRJCA015452.

Code availability

The source code for IntervalConvertor is publicly available via GitHub at https://github.com/CGotw/IntervalConvertor as well as via Code Ocean at https://doi.org/10.24433/CO.4320631.v2 (ref. 83).

References

63. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
64. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C processing. *Genome Biol.* **16**, 1–11 (2015).
65. Wang, S. et al. EndHiC: assemble large contigs into chromosome-level scaffolds using the Hi-C links from contig ends. *BMC Bioinform.* **23**, 1–19 (2022).
66. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
67. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
68. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
69. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 1–6 (2015).
70. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
71. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinfomatics* **35**, 2907–2915 (2019).
72. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
73. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 1–19 (2020).
74. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
75. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
76. Niu, C. et al. Methylation of a MITE insertion in the *MdRFNR1-1* promoter is positively associated with its allelic expression in apple in response to drought stress. *Plant Cell* **34**, 3983–4006 (2022).
77. Guan, Q., Lu, X., Zeng, H., Zhang, Y. & Zhu, J. Heat stress induction of *miR398* triggers a regulatory loop that is critical for thermotolerance in Arabidopsis. *Plant J.* **74**, 840–851 (2013).
78. Dai, H. et al. Development of a seedling clone with high regeneration capacity and susceptibility to *Agrobacterium* in apple. *Sci. Hortic.* **164**, 202–208 (2013).
79. Yang, S. L., Tran, N., Tsai, M. Y. & Ho, C. K. Misregulation of *MYB16* expression causes stomatal cluster formation by disrupting polarity during asymmetric cell divisions. *Plant Cell* **34**, 455–476 (2022).
80. Shen, X. et al. The RNA-binding protein MdHYL1 modulates cold tolerance and disease resistance in apple. *Plant Physiol.* **192**, 2143–2160 (2023).
81. Cao, F. et al. Natural variation in an HD-ZIP factor identifies its role in controlling apple leaf cuticular wax deposition. *Dev. Cell* **60**, 949–964.e6 (2024).
82. Wang, Y. et al. Brassinolide inhibits flavonoid biosynthesis and red-flesh coloration via the MdBEH2.2-MdMYB60 complex in apple. *J. Exp. Bot.* **72**, 6382–6399 (2021).
83. Li, W. et al. Pan-genome analysis reveals evolution and diversity of *Malus*. *Code Ocean* https://doi.org/10.24433/CO.4320631.v2 (2025).

Acknowledgements

This work was supported by the earmarked fund for the China Agricultural Research System (grant no. CARS-27 to Z.H.), 2115 Talent Development Program of China Agricultural University, the National Natural Science Foundation of China (grant nos. 32172522 to Z.H. and 32422077 to W.L.), the National Key Research and Development Program (grant no. 2022YFD1200501 to Z.H.), Pinduoduo-China Agricultural University Research Fund (grant no. PC2024B01002 to W.L.), 111 Project (grant no. B17043), National Horticulture Germplasm Resources Center (grant no. NHGRC to Z.H.), the China Postdoctoral Science Foundation (grant no. 2019M661344 to T.Z.), the Eberly College of Science and the Huck Institutes of the Life Sciences at the Pennsylvania State University, Key S&T Special Projects of Shaanxi Province, China (grant no. 2020zdzx03–01–02 to Q.G.) and the Pipfruit Technology Development Programme at The New Zealand Institute for Plant and Food Research Limited. We thank B. Gutierrez, curator of the *Malus* collections at the USDA-ARS Plant Genetic Resources Unit in Geneva, New York, for assistance in collecting leaf samples from scab-resistant cultivars. We thank P. Cong, C. Cheng and K. Wang's team from the National Germplasm Repository of Pear and Apple (Xingcheng), F. Ma's team from the Horticulture Experimental Station of Northwest A&F University, G. Sha's team from the Qingdao Academy of Agricultural Sciences, B. Zhang from the National Field Genebank for Hardy Fruits (Gongzhuling) and J. Zhu from the Shanghai Botanical Garden for assistance in providing samples and related pictures.

Author contributions

Z.H., W.L., C.C. and H.M. conceived and designed the project. Z.H., W.L., Hui Li, Yuqi Li, H.S., S.W., Z.W., Yi Wang, H.Z. and L.C. collected and provided plant materials. W.L., C.C., H.S., S.W., Z.W., C.H.D. and Hui Li assembled and annotated the genome. W.L., C.C., Y.H., C.H.D., H.S., S.W., Z.W., Hui Li and Yuqi Li performed and interpreted pan-genome analysis. H.M., T.Z., W.L., H.S., Z.W. and S.W. performed evolution analysis. Q.G., Z.L., W.L., S.W., H.S., Z,W., D.Z. and Huixia Li performed and analyzed the functional verification data. W.L., Z.H., C.C., H.M., C.H.D., Q.G., T.Z., H.S., S.W., Z.W., Z.L. and Y.H. interpreted data and contributed to writing the paper. X.Z., Z.G., Youqing Wang, Yi Li, H.Z., W.F., Yi Wang, X.X., L.C., Y.X. and B.Z. provided suggestions on data analysis and paper editing. All authors read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02166-6.

Correspondence and requests for materials should be addressed to Chong Chu, Qingmei Guan, Cecilia H. Deng, Yongming Han, Hong Ma or Zhenhai Han.

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.

nature portfolio

Corresponding author(s): Zhenhai Han

Last updated by author(s): Feb 18, 2025

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	Data was collected using bash commands to access the web portals of the data sources used in this study.
Data analysis	The following software were used throughout the analysis: hifiasm (v0.15.2_r334), HiC-Pro (v3.1.0), EndHiC (v1.0), Juicebox (v1.11.08), minimap2 (v2.18-r1015 & v2.21), purge_dups (v1.2.5), BUSCO (v5.0.0), LTRharvest (v1.6.1), LTR_FINDER (v1.07), Ltr_retriever (v2.9.0), RepeatModeler (v2.0.1), RepeatMasker (v4.1.0), quarTeT (v1.2.1), Juicer (v1.6), Centromics (v0.3), fastp (v0.20.1 & v0.23.4), HISAT2 (v2.2.1), SAMtools (v1.11), AUGUSTUS (v3.4.0), GenomeThreader (v1.7.1), BRAKER (v2.1.6), BLASTP (v2.10.1+), OrthoFinder (v2.5.5), PASTA (v1.9.0), HMMER (v3.3.2), HaMStR (v13.2.6), Trinity (v2.11.0), TransMCL (v1), MAFFT (v7.505), MUSCLE (v3.8.31), FastTree (v2.1.11), pal2nal (v14), trimAl (v1.4.rev22), RAxML (v8.2.12), ASTRAL (v5.7.8), PAML (v4.9), deeptime (v1.0.1), ggtree (v1.14.6), BioGeoBEARS (v1.1.3), Diamond (v2.1.5), PhyloMCL (v2.0), IQTREE2 (v2.1.2), Tree2GD (v1.0.40), LAST (v921), JCVI (v1.2.20), svim-asm diploid (v1.0.2), BCFtools (v1.8), SURVIVOR (v.1.0.7), Minigraph (v0.15-r426), Bandage (v0.8.1), gfatools (v0.5), BWA (0.7.17-r1188), GATK (v4.2.2.0), Plink (v1.90b6.21), VCFtools (v0.1.16), IntervalConvertor (v2.0), ImageJ (v1.51j8), FloMax (v2.82). The source code for 'IntervalConvertor' is publicly available on GitHub at https://github.com/CGotw/IntervalConvertor as well as from Code Ocean at https://doi.org/10.24433/CO.4320631.v2.

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

All raw sequencing reads of PacBio HiFi, Hi-C, and RNA-seq have been deposited to the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) database with the BioProject number PRJNA861686. The 30 Malus species genome assemblies are available in NCBI under the BioProject number PRJNA1062995. Additionally, all genome assemblies and annotation data for the 30 Malus species can be accessed at the Genome Warehouse in the National Genomics Data Center, China National Center for Bioinformation (https://ngdc.cncb.ac.cn/gwh), with BioProject accession number PRJCA015452.

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	Our sampling strategy focused on capturing the broad genetic diversity of the Malus genus by selecting 30 representative species, encompassing domesticated and wild Malus, as well as both diploids and polyploids. This selection was primarily guided by Alfred Rehder's globally recognized classification system based on morphological traits. To further enrich diversity, we incorporated additional species native to China, following Yu Dejun's expansion of Rehder's framework.
Data exclusions	No data were excluded.
Replication	All experiments were performed with at least three biological replicates, and the presented experimental images and conclusions are supported by at least three consistent independent experiments. All replications were successful and replications for each experiment were clearly stated in the corresponding figure legends or Methods.
Randomization	For each apple individual, the sampling process for genome DNA/RNA sequencing was randomly conducted. All WT and transgenic materials were evaluated under the same growth condition in a random arrangement.
Blinding	The investigators were blinded to group allocation during data collecting and data analysis.

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

anti-MYC Mouse antibody (Abcam, Cat#Ab18185, Clone#Myc.A7)
anti-Actin Mouse antibody (Abmart, Cat#M20009, Clone#26F7)
anti-mouse HRP-conjugated IgG antibody (ABclonal, Cat#AS003)
anti-H3K27me3 Rabbit antibody (Millipore, Cat#17-622)

For Western blot analysis, the following dilutions were used: 1:5000 dilution of anti-MYC Mouse antibody, 1:10000 dilution of anti-Actin Mouse antibody, and 1:5000 dilution of anti-mouse HRP-conjugated IgG antibody.

For ChIP, the following dilution was used: 5 μg of anti-H3K27me3 Rabbit antibody was utilized in conjunction with 10 μg of chromatin.

Validation

All commercial antibodies have been validated by the source. We provide links to the available technical datasheet below:
1) anti-MYC Mouse antibody: https://www.abcam.cn/products/primary-antibodies/myc-tag-antibody-myca7-ab18185.html#lb
2) anti-H3K27me3 Rabbit antibody: https://www.sigmaaldrich.com/TW/en/product/mm/17622
3) anti-Actin Mouse antibody: https://www.ab-mart.com.cn/upload/20170614135558xz.pdf
4) anti-mouse HRP-conjugated IgG: https://abclonal.com.cn/catalog/AS003

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

0.2g fresh samples were placed in a petri dish and 500μl of nuclear extraction solution was added. The sample was shredded with a sharp blade to fully extract the intact nucleus. After 60 seconds in the extraction solution, the mix was filtered into the sample tube with a 50μm filter. 2,000μl of DAPI fluorescent dye was added into the sample tube, and the measurement was carried out after 2 minutes in a shading environment.

Instrument

Nuclei suspension were analyzed by CyFlow Space Flow Cytometer (Sysmex Partec, Muenster, Germany).

Software

FloMax software was used to analysis the data.

Cell population abundance

Not available.

Gating strategy

Not available.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.