

A haplotype-resolved graph pangenome of sweet cherry reveals structural variation shaping fruit traits and genome divergence across breeding levels

Claudio Urra

Universidad Mayor

Ismaël Blanchard

University of Bordeaux

Véronique Decroocq

University of Bordeaux

Benjamin Linard

Université de Toulouse

Caixi Zhang

Shanghai Jiao Tong University

José Quero-García

University of Bordeaux

Elisabeth Dirlewanger

University of Bordeaux

Andrea Miyasaka Almeida

Universidad Mayor

Bénédicte Wenden

University of Bordeaux

Anthony Bernard

University of Bordeaux

Quynh-Trang Bui

`quynh-trang.bui@inrae.fr`

University of Bordeaux

Research Article

Keywords: Structural variants, fruit quality, genome architecture, *Prunus avium*

Posted Date: April 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9237087/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Structural variation represents a substantial fraction of genetic diversity but remains incompletely characterized using single-reference genomes. In sweet cherry (*Prunus avium* L.), breeding has focused on a few loci controlling fruit quality, yet the associated structural variation and genome-wide patterns of diversity remain poorly understood.

Results

We constructed a chromosome-scale haplotype-resolved graph pangenome from 27 high-quality genome assemblies representing diverse breeding levels and geographic origins. The pangenome comprised 1.42 Gb of graph sequence with a small core and large accessory fraction, indicating an open and diverse genome despite chromosome-scale collinearity among accessions. We identified over 150,000 structural variants along with millions of SNPs and short indels, revealing extensive polymorphism across chromosomes. Integrating SNPs and structural variants within the graph framework enabled genome-wide association analyses for major fruit traits, including fruit weight and maturity date, identifying candidate loci on chromosomes 2 and 4 overlapping previously reported QTL hotspots. Several associated structural variants were located near genes involved in cellular growth and ripening processes. Genome-wide differentiation between landrace and modern germplasm revealed localized differences in genome architecture in regions enriched for defense and specialized metabolism genes, with limited overlap between differentiation signals and trait-associated loci.

Conclusions

These findings suggest that modern breeding may have contributed to differentiation in a limited subset of genomic regions, while substantial diversity persists elsewhere in the genome. This haplotype-resolved pangenome provides a comprehensive framework for studying genome and trait architectures, and genomic differentiation in sweet cherry, illustrating the value of graph-based genomic resources for crop breeding.

Background

Genomics has profoundly improved our understanding of plant domestication and the selective processes that shape crop diversity. While early domestication initiated the divergence of cultivated taxa from their wild ancestors, modern breeding has become a major driver of genetic change in crops [1]. Intensive selection for yield, quality, stress tolerance, and uniformity has led to strong and often rapid shifts in allele frequencies, leaving distinct genomic signatures associated with breeding practices. Identifying these genomic regions is essential not only to retrace the evolutionary trajectory from domestication to improvement, but also to disentangle the effects of recent selection pressures. Moreover, linking these regions to phenotypic variation provides valuable insights into the genetic architecture of key agronomic traits and facilitates the identification of candidate loci for breeding programs. This knowledge is particularly critical for designing improved cultivars that can meet the challenges of climate change, sustainability, and increasing production demands [2, 3]. Large-scale resequencing projects in major crops have enabled the identification of loci controlling key agronomic traits [4]. A major limitation of traditional genomic studies is their reliance on a single reference genome, which cannot capture the full extent of genetic diversity present within a species [4]. In particular, structural variants (SVs), including insertions, deletions, duplications, inversions, and translocations, are often underrepresented or entirely absent from a single reference sequence [3]. This reference bias can reduce the power of quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS) and may contribute to the problem of “missing heritability” observed for many complex traits [5, 6]. For example, analyses in grapevine have shown that a single reference genome can miss more than 10% of the genes present in heterozygous cultivars, highlighting the importance of considering presence–absence variation and structural diversity in crop genomics [7]. The development of pangenomics has shifted the paradigm from single-reference analyses to a more comprehensive representation of the genomic diversity of a species [3, 8]. A pangenome integrates multiple genome assemblies to distinguish between a core genome shared by all individuals and a variable or accessory genome present only in a subset of genotypes [8, 9]. Early plant pangenomes were often constructed using short-read assemblies, but recent advances in long-read sequencing technologies combined with graph-based data structures, now enable the generation of high-quality haplotype-resolved pangenomes that more accurately represent allelic and structural variation [10, 11]. These approaches have been successfully applied to several crops, where they have revealed novel genes, large structural variants, and previously undetected alleles associated with yield, stress tolerance, and quality traits [2].

The need for pangenomic resources is particularly acute in perennial fruit trees, which present specific biological and breeding challenges compared with annual species. Fruit trees often exhibit long juvenile phases, high levels of heterozygosity, and complex reproductive

systems, resulting in breeding cycles that can span decades and are technically demanding [12, 13]. Recent pangenomes developed for fruit crops such as apple (*Malus domestica*), citrus (*Citrus* spp.), and grapevine (*Vitis vinifera*) have demonstrated that a substantial fraction of intraspecific variation resides in accessory genomic regions and structural variants, many of which are associated with traits of commercial importance, including fruit color, flavor, acidity, and disease resistance [7, 14, 15]. These results highlight the importance of capturing structural variation to fully understand the genetic architecture of complex traits in perennial crops. Structural variants are increasingly recognized as major drivers of phenotypic diversity in plants [16, 17], including in fruit trees. In peach (*Prunus persica*), an integrated SV map generated from hundreds of genomes identified causal variants associated with important fruit traits, including a deletion in the promoter of PpMYB10.1 controlling flesh color [18]. In Chinese plum (*Prunus mume*), an insertion in the promoter of PmPH4 was shown to regulate citric acid accumulation in fruit [19]. These studies demonstrate that SVs of different sizes can underlie key agronomic traits and that their systematic identification through pangenome-based approaches is essential to fully characterize the genetic basis of phenotypic variation in fruit crops.

Among species of the genus *Prunus*, sweet cherry (*Prunus avium* L.) is an economically important fruit crop, with global production exceeding 3 million tonnes annually (FAOSTAT, 2025), and Türkiye, Chile, and the United States among the leading producers. Sweet cherry fruit is mainly consumed fresh but is also widely used for processed products, including juices, preserves, and confectionery, making fruit quality a primary breeding objective. Fruit quality is a complex trait influenced by multiple genetic and environmental factors and includes attributes such as fruit size, firmness, skin and juice color, soluble solids content, titratable acidity, and susceptibility to rain-induced cracking. QTL mapping and, more recently, GWAS have identified genomic regions associated with several of these traits [20, 21, 22, 23, 24], including recurrent hotspots on chromosomes 2 and 4 that have been repeatedly targeted in breeding programs. Population genomic studies have also reported signatures of domestication and selection in sweet cherry, with reduced diversity in modern cultivars compared with wild or landrace germplasm, suggesting that improvement has focused on a limited number of genomic regions while other parts of the genome retain substantial variation [25]. However, the causal variants underlying these loci, particularly structural variants and presence-absence variation, remain largely unknown, and the absence of a comprehensive pangenomic resource limits the ability to fully capture genomic diversity in cultivated and wild-derived germplasm.

To investigate genome structural diversity across cultivated sweet cherry (*Prunus avium*), we constructed a haplotype-resolved, chromosome-scale graph pangenome from 27 high-quality genome assemblies representing diverse breeding histories and geographic origins. This dataset includes 11 *de novo* assemblies generated in this study, two high-quality genome assemblies, and two previously published assemblies, encompassing modern cultivars, early selections, and landrace germplasm. Using this resource, we characterized sequence and structural variation across the species, identified differentiation in genome architecture between landrace and modern accessions, and performed genome-wide association analyses integrating single-nucleotide polymorphisms and structural variants for major fruit quality traits. This pangenome framework provides new insights into the genetic architecture of agronomic traits in sweet cherry and establishes a foundation for the efficient use of genomic diversity in future breeding programs.

Results

Genome assemblies and haplotype resolution for 11 sweet cherry genotypes

High-fidelity (HiFi) PacBio long-read sequencing was performed for 11 sweet cherry accessions representing the nine genetic groups from the study of Campoy et al. [26], where they characterized the population structure of 210 sweet cherry accessions from 16 countries, identifying nine distinct genetic groups that distinguish between landraces, early selections (selections made quite early) and modern cultivars (from modern breeding), while reflecting their eco-geographic distribution. This breeding level classification was mainly based either on information coming from literature or on information gathered in collaboration with the 'Centre National de Pomologie' (<http://pomologie.fr/>). Selecting representatives from each of these groups ensures that the resulting pangenome captures genetic diversity and structural variation across a large germplasm, providing a comprehensive genomic representation of the species population structure. Sequencing depth ranged from 23X to 98X, providing sufficient coverage to generate good quality haplotype-resolved assemblies (Table 1).

De novo genome assemblies were generated using the Asm4pg v1.1.0 pipeline, which integrates assembly, scaffolding, and polishing steps optimized for plant genomes. Assembly with hifiasm produced two haplotype-resolved assemblies per accession, yielding 22 haplotype assemblies in total. Contigs were scaffolded using RagTag against the *P. avium* 'Tieton' v2.0 reference genome [27], followed by anchoring and ordering into chromosome-scale pseudomolecules using ALLMAPS, based on multiple high-density genetic linkage maps.

The resulting assemblies showed high structural continuity and genome completeness. Assembly sizes ranged from 312.2 Mb to 374.9 Mb, with a mean assembly size of 345.0 Mb, consistent with the expected genome size of sweet cherry. Assembly fragmentation varied

among haplotypes, with the number of sequences ranging from 65 to 584 per haplotype. Contiguity metrics further reflected the quality of the assemblies, with N50 values ranging from 37.3 kb to 75.1 kb (mean 51.5 kb).

Each haplotype assembly contained eight chromosome-scale scaffolds, corresponding to the eight chromosomes of sweet cherry, together with additional unscaffolded sequences of variable size. Whole-genome alignments against the Tieton v2.0 reference genome confirmed the overall conservation of chromosome structure across accessions. These alignments also enabled the identification of several local orientation discrepancies, which were corrected by reverse-complementing sequences using EMBOS revseq before reintegration into the assemblies.

To ensure high structural completeness for downstream pangenome analyses, only haplotypes with at least 80% of their assembled sequence anchored to chromosome-scale scaffolds were retained. The proportion of anchored sequence across haplotypes ranged from 82.0% to 96.8%, with a mean of approximately 89.5%. One haplotype (V2775 hap1) exhibited a lower anchoring rate of 77.85% due to its low sequencing depth and was consequently excluded from further analyses.

After filtering, 21 haplotype assemblies were retained for pangenome construction. For consistency, only chromosome-level sequences corresponding to the eight pseudomolecules were included in the final dataset. Unplaced and unlocalized scaffolds were excluded to maintain a consistent coordinate system and to avoid artifacts associated with fragmented or ambiguously positioned sequences.

Assembly completeness was further assessed using BUSCO v5.3.1 with the eudicots_odb10 lineage dataset, with all assemblies achieving completeness scores exceeding 96% (Additional file 1: Table S2), confirming the high completeness of the assemblies and supporting their suitability for downstream genomic analyses.

Together, these chromosome-scale haplotype assemblies provide a good quality representation of structural diversity across sweet cherry genomes and constitute a robust foundation for the construction of a haplotype-resolved sweet cherry pangenome, enabling the systematic characterization of structural variation and understanding of genomic diversity across this important fruit species.

Table 1
Genotypes used to construct the sweet cherry pangenome

Accession name	Assembly label	Breeding level	Origin	Group of diversity ²	Sequencing	Depth sequencing	Reference
'Cypres'	V0088	Landrace	France	8	PacBio HiFi	52X	in this study
'Kassins Frühe'	V0175	Early selection	Germany	1	PacBio HiFi	30X	in this study
'Bigarreau Hatif Burlat'	V0370	Early selection	France	3	PacBio HiFi	98X	in this study
'Cristobalina'	V0897	Landrace	Spain	7	PacBio HiFi	38X	in this study
'Pontavium'	V1813	Modern breeding ¹	France	-	PacBio HiFi	34X	in this study
'Vittoria'	V2049	Modern breeding	Italy	5	PacBio HiFi	28X	in this study
'Bada'	V2076	Modern breeding	United States	6	PacBio HiFi	40X	in this study
'Garnet'	V2155	Modern breeding	United States	4	PacBio HiFi	38X	in this study
'Rubin'	V2775	Modern breeding	Romania	2	PacBio HiFi	23X	in this study
'Fertard'	V3382	Modern breeding	France	4	PacBio HiFi	29X	in this study
'Abouriou'	V4098	Landrace	France	9	PacBio HiFi	48X	in this study
'Santina'	Santina	Modern breeding	Canada	-	PacBio HiFi, ONT, Hi-C	-	[28]
'Regina'	Regina	Modern breeding	Germany	-	PacBio HiFi, ONT, Hi-C	-	[28]
'Tieton'	Tieton	Modern breeding	United States	-	ONT, Hi-C	-	[27]
'Satonishiki'	Satonishiki	Early selection	Japan	-	Illumina	-	[29]

The table summarizes the set of *Prunus avium* accessions included in the pangenome analysis, detailing their assembly identifiers used throughout this article, breeding level (landrace, early selection, or modern breeding), and geographic origin. Genetic diversity groups correspond to groups defined by discriminant analysis of principal components (DAPC) as reported by Campoy et al. [26]. Sequencing platforms and approximate sequencing depth are provided for each accession. Modern reference cultivars (Santina, Regina, and Tieton) were assembled using a combination of long-read (PacBio HiFi and/or Oxford Nanopore Technologies) and Hi-C data, whereas other accessions were primarily sequenced using PacBio HiFi. Missing values are indicated where information was not available.

¹ Selection used as rootstock, derived from wild *P. avium* (mazzard). ² Diversity groups based on DAPC analysis from Campoy et al. [26].

Graph-derived genomic relationships define the chromosome-scale sweet cherry pangenome

The dataset comprised 27 high-quality assemblies, including 21 haplotypes-resolved sequences from 11 *de novo* generated diploid assemblies. In addition, the dataset included four haplotype-resolved sequences from two newly available high-quality genomes (Santina and Regina) [28], and two consensus-based sequences from previously published genomes (Tieton and Satonishiki) [27, 29] (Table 1).

We first evaluated genomic relationships among the assemblies incorporated in the sweet cherry pangenome using alignment-free k-mer-based distances. Mash distances were computed independently to each chromosome, and chromosome-specific similarity networks were generated to assess intra- and inter-genome clustering. In all cases, chromosomes clustered tightly by genome of origin, with no cross-

links observed between different accessions or haplotypes. This pattern is consistent with high assembly contiguity and integrity and indicates no evidence of cross-sample contamination in the dataset (Fig. 1d).

Genome-wide pairwise Mash distances were then computed across all 27 assemblies and used to infer a k-mer-based phylogeny. The resulting tree resolved four major genetic groups, broadly corresponding to breeding levels of the sweet cherry accessions. Two groups consist exclusively in modern breeding accessions, whereas the other two groups are composed primarily of early selections or landrace accessions. This phylogeny was subsequently used as the guide tree for progressive chromosome-wise pangenome graph construction with MiniCactus (Fig. 1a).

The resulting graph represents shared and divergent genomic regions across sweet cherry accessions at chromosomes scale. Across all eight chromosomes, the pangenome comprised 15.4 million nodes and 20.8 million edges, representing 1.42 Gb of graph sequence (Additional file 1: Table S3). Node sizes were small (mean length ~ 92 bp), consistent with a graph topology shaped by abundant fine-scale sequence divergence and structural complexity. Pangenome growth analysis identified a core genome of 109.8 Mb, corresponding to 7.7% of the total graph sequence, whereas the remaining 92.3% represented accessory regions variably present among genotypes.

Chromosome-scale comparisons indicated broad conservation of genome structure across sweet cherry accessions. Heatmaps derived from pairwise alignments between each chromosome of every assembly and all chromosomes of the pangenome showed homology signals concentrated along matching chromosomes, consistent with macrosynteny across genomes (Additional file 2: Figures S1-S8). Off-diagonal signals were rare, indicating the absence of major interchromosomal rearrangements. Instead, localized reductions in similarity were observed within otherwise collinear chromosomes, consistent with sequence divergence and structural polymorphism among haplotypes.

To quantify how genomic diversity accumulates with increasing sampling, we examined pangenome growth curves using ordered-hist-growth analyses. Both core and accessory fractions changed as additional genomes were incorporated, with accessory sequence continuing to increase across the sampling range. This pattern indicates that the sweet cherry pangenome remains open and unsaturated (Fig. 1b).

Comparison of core, shared, and private fractions across assemblies revealed marked heterogeneity: Regina and Santina haplotypes contributed the smallest amounts of private sequence, consistent with their roles as high-quality reference-grade assemblies, whereas Satonishiki and Tieton contributed the largest private fractions, reflecting their greater divergence from the remaining genotypes (Fig. 1c).

The sweet cherry pangenome captures extensive sequence and structural variation

Having established the overall architecture of the sweet cherry pangenome, we next characterized sequence and structural variation represented in the pangenome graph. Structural variants (SVs) were identified from graph-derived alignments and classified into insertions, deletions, duplications, inversions, and translocations. A total of 151,758 structural variants (SVs) were identified in the sweet cherry pangenome. Among these, 58.8% corresponded to insertion/deletion-type events (represented as deletions relative to the reference), whereas 41.2% corresponded to duplications.

When the Tieton genome was excluded, insertion/deletion-type variants increased to 67.1%, while duplications decreased to 32.9%, indicating that duplications are enriched in the Tieton genome, widely used to date as the sweet cherry reference genome. SV sizes ranged from 50 bp to > 10 kb (spanning a total of 65.5 Mbp). In this study, the Regina hap1 assembly was used as the reference for SV distribution analyses because it represented a high-quality assembly [28] and showed lower overall divergence from the other genomes than the Tieton reference, reducing potential reference bias. SVs were unevenly distributed along chromosomes and formed localized regions of elevated density when calculated in 100-kb windows across the Regina hap1 reference (Fig. 2a), most prominently in the central region of chromosome 3. The number of SVs varied among genomes, reflecting differing levels of divergence among accessions and haplotypes. Notably, Tieton harbored the largest number of pangenome SVs (Fig. 2d) and the second-highest indel density (Fig. 2c).

Across the pangenome, we identified 11.5 million SNPs and 2.8 million short indels, which were heterogeneously distributed across chromosomes (Additional file 1: Table S4). SNP and indel densities revealed chromosome-specific patterns characterized by hypervariable regions interspersed with segments of reduced diversity (Fig. 2a). As observed for SVs, SNPs were especially abundant in the central region of chromosome 3. Genome-wide variant densities also differed among accessions (Fig. 2b,c), highlighting substantial heterogeneity in sequence diversity across the sweet cherry germplasm represented in the graph.

Together, these results demonstrate that the sweet cherry pangenome captures extensive sequence and structural diversity while maintaining overall chromosome-scale collinearity among accessions, providing a comprehensive framework for studying the genetic basis of key agronomic traits using association analyses.

Genome-wide loci and structural variants associated to fruit weight, maturity date, and additional fruit traits

Genome-wide association analyses were performed using both population SVs and SNP markers. Under a general linear model (GLM), significant associations were detected for fruit weight in both marker datasets (Fig. 3a,b). The strongest signals were concentrated on chromosome 2, within the interval 25.6–26.1 Mb, where several variants exceeded the genome-wide significance threshold. When a mixed linear model (MLM) accounting for relatedness and population structure was applied, the number of significant associations was substantially reduced (Fig. 3c,d), indicating that part of the GLM signal reflected underlying population structure.

In contrast to fruit weight, genome-wide association analyses for fruit maturity date revealed signals distributed across multiple chromosomes (Fig. 4a,b). A shared association peak was detected in the central region of chromosome 4 in both SNP- and SV-based analyses, with SVs concentrated within the interval 14.4–14.7 Mb and SNPs located slightly upstream (14.31–14.37 Mb), indicating partially overlapping signals across marker classes. As observed for fruit weight, application of the MLM reduced the number of significant associations (Fig. 4c,d), consistent with the strong genetic structure present in the panel.

Genome-wide association analyses were also performed for additional fruit traits, including fruit juice color, fruit stem-end cracking, and fruit firmness (Additional file 2: Figure S9). Association signals were detected across multiple chromosomes using both SNP and structural variant (SV) markers under the GLM model, with substantial overlap between SNP- and SV-associated intervals. This overlap was particularly evident for fruit stem-end cracking, where both marker types co-localized within the same genomic region on chromosome 2, and for fruit juice color and fruit firmness, where SNP intervals were largely nested within broader SV-associated regions. Specifically, for fruit juice color, SVs spanned 18.7–27.8 Mb and SNPs 21.0–26.4 Mb on chromosome 3; for fruit stem-end cracking, SVs spanned 23.3–24.8 Mb and SNPs 23.2–24.8 Mb on chromosome 2; and for fruit firmness, SVs spanned 11.4–14.7 Mb and SNPs 13.8–14.8 Mb on chromosome 4.

As observed for fruit weight and maturity date, both the number and magnitude of significant associations were markedly reduced under the MLM model, suggesting that part of the GLM signal likely reflects underlying population structure. Overall, Manhattan plots revealed trait-specific association patterns, with several loci exceeding the genome-wide significance threshold under the GLM model, whereas the MLM results retained fewer but more conservative signals.

Consistent with these patterns, a major association signal was detected on chromosome 2 for fruit weight (25.6–26.1 Mb), involving three SVs and eight associated SNPs, and a prominent peak was identified on chromosome 4 for fruit maturity date (14.4–14.7 Mb), with three SVs and 27 associated SNPs. For fruit weight, SNP- and SV-based analyses showed concordant signals within the same genomic interval, supporting a robust association at this locus. In contrast, for fruit maturity date, SNP associations were located slightly upstream (14.31–14.37 Mb) relative to the SV-associated interval (14.4–14.7 Mb), while still pointing to a shared candidate region.

To further investigate the chromosome 2 association for fruit weight, we examined the genomic context of SVs within the associated interval using the Regina haplotype 1 assembly. Several SVs within this region were located in proximity to annotated genes (Fig. 3e–g). One associated SV was positioned downstream of *NDPK1* (nucleoside diphosphate kinase 1), whereas the deletion DEL00007362 was located approximately 0.5 kb downstream of *RMA1* (E3 ubiquitin-protein ligase). Additional variants were detected near *CDS4* (cytidinediphosphate diacylglycerol synthase 4), indicating that multiple genes reside close to the associated interval. Collectively, these observations define a candidate genomic region spanning approximately 25.6–26.1 Mb on chromosome 2 that likely contributes to fruit weight variation. Although these variants are located near annotated genes, the causal variants underlying the association remain to be determined.

Localized differences in genome architecture between landrace and modern germplasm

To investigate whether loci associated with fruit traits coincide with genomic regions affected by breeding level of accessions, we compared landrace and modern sweet cherry accessions. Genome-wide patterns of differentiation were assessed using SNP-based *F_{st}* estimated in non-overlapping 100-kb windows

Across the genome, 27 windows exceeded the 99th percentile threshold, indicating a limited number of regions with elevated differentiation (Fig. 5a). These outlier windows were unevenly distributed across chromosomes, with the largest numbers detected on chromosomes 2, 3 and 4.

The genome-wide *F_{st}* landscape was characterized by localized peaks rather than broad chromosome-wide shifts in differentiation (Fig. 5b), suggesting that breeding level-related divergence of accessions has affected specific genomic regions rather than extensive genomic segments. In particular, a region at the beginning of chromosome 2 showed a *F_{st}* of more than 0.25 between landrace and modern breeding accessions. Genes located within high-*F_{st}* windows were significantly enriched for several Gene Ontology categories associated with lipid membranes, including triterpenoid biosynthetic process, transmembrane receptor protein tyrosine kinase activity and lipid droplets but also with defense-related processes (Fig. 5c).

The chromosomal distribution of genes contributing to enriched GO terms revealed a clear functional partitioning: genes associated with lipid- and membrane-related processes were predominantly located on chromosomes 3 and 4, whereas genes related to defense responses were mainly distributed across other chromosomes (Fig. 5d). This pattern suggests that distinct biological processes have been differentially targeted by modern breeding, with membrane- and lipid-associated functions concentrated in specific chromosomal regions and defense-related functions shaped more broadly across the genome. In several cases, multiple genes assigned to the same GO term were concentrated within relatively narrow genomic intervals, reinforcing the presence of localized functional differentiation between landrace and modern breeding germplasm (Additional file 1: Table S5).

Key fruit trait loci show regional co-localization with differentiated genomic regions between landrace and modern germplasm

To investigate whether genomic regions showing strong differentiation between landraces and modern cultivars contribute to fruit trait variation, we intersected high-*F_{st}* windows with GWAS signals for fruit traits. To assess co-localization, we defined genomic intervals of ± 250 kb around each significant GWAS SNP and evaluated whether these regions overlapped with *F_{st}* outlier windows (top 1%). The ± 250 kb window was chosen to account for the extent of linkage disequilibrium (LD) decay (~ 100 kb) while accommodating uncertainty in the precise localization of association signals and *F_{st}* windows, thereby capturing broader genomic regions potentially in linkage with causal variants.

Overlap between the two analyses was limited. Although no direct overlap was detected under a strict ± 250 kb criterion, GWAS SNPs associated with fruit weight were located in close proximity (~ 9 kb) to *F_{st}* outlier windows on chromosome 2, indicating regional co-localization between differentiation signals and trait-associated loci. In contrast, two high-*F_{st}* windows located at the distal end of chromosome 4 overlapped with SNP-based GLM association signals for fruit maturity date, although these did not correspond to the principal association peak detected for this trait. On chromosome 2, 61 overlaps were detected between *F_{st}* outlier windows and GWAS SNP regions associated with fruit stem-end cracking when using a ± 250 kb interval around significant markers.

These results indicate that the major locus controlling fruit weight does not coincide with highly differentiated regions between landrace and modern germplasm, whereas some loci associated with fruit maturity date and fruit stem-end cracking fall within regions showing elevated differentiation between the two groups, landrace and modern germplasm.

Discussion

The results presented here provide new insights into the genomic and fruit trait architecture of sweet cherry. In the following sections, we discuss the implications of the haplotype-resolved pangenome for understanding structural variation, trait architecture, and genomic divergence associated with breeding level of accessions.

Haplotype-resolved pangenome captures extensive structural diversity and chromosomal collinearity in sweet cherry

The sweet cherry pangenome reveals extensive sequence and structural diversity while maintaining broad chromosome-scale collinearity and macrosynteny among accessions. In terms of genomic architecture, the sweet cherry graph pangenome, constructed with 25 haplotypes and 2 consensus sequences, is highly expansive, comprising 15.4 million nodes representing 1.42 Gb of graph sequence (Additional file 1: Table S3, Fig. 1). This complexity exceeds that reported for the cultivated peach (*Prunus persica*) pangenome, which contains 10.9 million nodes across a total graph length of approximately 303.5 Mb, constructed with 16 consensus sequences [30]. In addition, the sweet cherry core sequence (109.8 Mb) represents only 7.7% of the total graph sequence, indicating an open and unsaturated pangenome architecture, whereas the peach pangenome has been described as largely saturated, with its core sequence accounting for

nearly half of the total sequence. Extensive genomic diversity has also been reported in other *Prunus* species such as peach and almond [31, 32, 33, 34], although the genetic structure and breeding history differ among species.

These differences likely reflect both biological divergence and methodological factors, as the peach pangenome was primarily constructed from consensus assemblies, whereas the present sweet cherry pangenome integrates haplotype-resolved genomes, enabling a more comprehensive representation of allelic and structural diversity. The implementation of this graph-based framework follows approaches successfully applied in other crops, such as tomato, where graph pangenomes have revealed extensive hidden variation that cannot be captured using a single linear reference genome [34].

Beyond methodological differences between pangenomes, variation in the quality of individual genome assemblies may also influence the representation of structural variation. For example, the Satonishiki genome was generated using earlier short-read sequencing technologies and resulted in a comparatively fragmented assembly [29], which can limit the accurate reconstruction of repetitive or duplicated regions. Conversely, the Tieton v2 genome, to date broadly used as sweet cherry reference genome, shows a substantially higher number of duplication-type structural variants relative to the other genomes included in the pangenome. Although part of this signal may reflect genuine biological variation, it may also be influenced by assembly-specific factors, such as the treatment of repetitive regions or differences in haplotype collapsing during genome construction. Similar effects have been reported in comparative genomic analyses, where variation in sequencing technologies and assembly strategies can affect the detection and classification of structural variants [35, 36]. These methodological differences should therefore be considered when interpreting genome-wide patterns of structural variation within graph-based pangenomes.

The predominance of local structural variants and indels captured in this pangenome (Fig. 2) is consistent with observations in several major crop pangenomes, where structural variation represents a major component of genomic diversity. In maize, millions of non-redundant structural variants have been reported, and the pangenome continues to expand as additional genomes are incorporated, indicating that structural diversity in the species remains far from saturated [37]. Similarly, graph-based pangenomes developed in rice and tomato have revealed extensive presence/absence variation and structural polymorphism, substantially improving the detection of trait-associated loci and the recovery of previously unexplained heritability [34, 38]. In tomato, for example, incorporating structural variants identified through graph-based approaches increased estimated trait heritability by 24%, highlighting the importance of capturing structural variation to resolve incomplete linkage disequilibrium often missed by SNP-based analyses.

Although sweet cherry maintains the strong chromosome-scale synteny characteristic of the *Prunus* genus [39], much of its genomic variability is represented in the accessory fraction of the pangenome graph, which accounts for more than 90% of the total graph sequence. Similar patterns have been observed in other perennial fruit crops. In apple, for example, core genes represent only a subset of the total gene pool, whereas accessory genes contribute substantially to intraspecific diversity [40]. In peach, structural variants such as LTR retrotransposon insertions in promoter regions have been identified as regulators of fruit quality traits, including malate accumulation and flesh coloration [30]. These findings suggest that while sweet cherry preserves overall chromosome-scale macrosynteny, the high density of local variants and the extensive accessory sequence captured through haplotype-resolved assemblies represent major sources of genomic and phenotypic variation.

An illustrative example within the dataset is the genotype V1813 Pontavium, which is derived from mazzard, the wild form of *Prunus avium*. Despite its wild origin, Pontavium does not appear as a strongly divergent lineage in the phylogenetic reconstruction and does not display an unusually high number of structural variants relative to cultivated genotypes. This observation is consistent with previous population genetic studies showing that wild cherries and sweet cherry landraces can be genetically close [25]. These results indicate that wild-derived genotypes may remain closely related to cultivated germplasm at the genome scale, highlighting their potential interest for maintaining genetic diversity in breeding programs. This close relationship between wild and cultivated germplasm may also explain why incorporating diverse genotypes into graph-based pangenomes continues to reveal substantial accessory variation within the species.

Importantly, the extensive structural variation represented in the sweet cherry pangenome provides a powerful framework for investigating the genetic basis of phenotypic traits. By integrating structural variants and sequence polymorphisms within a unified graph representation, the pangenome enables a more comprehensive exploration of genotype–phenotype relationships, including the identification of candidate loci associated with agronomic traits through genome-wide association analyses.

The extensive structural and haplotypic diversity captured by the graph pangenome provides an opportunity to reassess the genetic architecture of agronomic traits. In particular, integrating pangenome variation with genome-wide association analyses allows a more comprehensive evaluation of candidate loci underlying fruit quality traits.

Integrating pangenome variation with GWAS identifies candidate regions underlying fruit traits

The sweet cherry pangenome provided a framework for genome-wide association analyses of five agronomic and fruit quality traits, enabling the integration of SNPs and structural variants (SVs) within a unified genomic representation (Additional file 1: Table S4). Significant associations were detected for all traits evaluated, including fruit weight (Fig. 3), fruit maturity date (Fig. 4), fruit juice color, fruit stem-end cracking, and fruit firmness (Additional file 2: Figure S9). Incorporating SVs alongside SNPs allowed the identification of candidate regions supported by multiple classes of genomic variation, providing increased resolution compared with analyses based on a single linear reference genome.

The integration of structural variants represents an important extension over previous association studies in sweet cherry. For example, Donkpegan et al. [24] showed that SNP–trait associations can depend strongly on the reference genome used, such as Regina or Satonishiki, highlighting the limitations of single-reference analyses. Graph-based pangenomes help mitigate this reference bias by incorporating sequence diversity from multiple genomes into a unified representation. Although genomic coordinates for visualization (e.g., Manhattan plots) are projected onto a linear reference (Regina haplotype 1), short-read mapping and variant detection were performed against the pangenome graph. This allows variants to be identified in sequences that are absent from any single reference genome. In this context, the detection of both SVs and SNPs within the same associated loci provides complementary evidence supporting candidate genomic regions and facilitates the identification of variants that may not be represented in any individual reference assembly.

For fruit weight, the strongest association signals were located in the central region of chromosome 2 between 25.6 and 26.1 Mb (Fig. 3a,b). One candidate variant in this region is the deletion DEL00007362 located downstream of *RMA1* (Fig. 3f). *RMA1* encodes a RING membrane-anchored E3 ubiquitin ligase, a protein class involved in membrane protein turnover and stress-related signaling pathways in plants. Homologs of *RMA1* in several species have been associated with regulation of membrane transport proteins and responses to environmental stimuli, suggesting that variation in this region could indirectly affect cellular processes involved in fruit growth [41, 42, 43].

An additional SV detected in the same interval is located upstream of *NDPK1* (Fig. 3e). *NDPK1* participates in nucleoside triphosphate biosynthesis and has previously been reported in sweet cherry studies related to fruit development [23, 44]. In the present analysis, *NDPK1* was located outside the strict interval defined by associated SVs but in close proximity (~2.3 kb upstream), which explains why this gene was not included among those directly overlapping the associated variants while remaining a plausible nearby candidate.

Interestingly, in apricot, Groppi et al. [45] identified a locus on chromosome 2, between 25.79 and 25.80 Mb, associated with fruit size and development. This region overlaps with the locus detected here in sweet cherry and includes a gene encoding an ABC transporter C family member. Similar transporters have been implicated in fruit development in tomato [46]. Although the presence of a direct ortholog within the sweet cherry associated interval could not be confirmed in the present analysis, the correspondence between these loci across *Prunus* species supports the relevance of this genomic region for fruit size variation. Together, these observations define a candidate locus that may contribute to variation in fruit weight in sweet cherry, although the causal variant remains to be determined.

A reproducible association signal was detected in the central region of chromosome 4 with SVs located between 14.4–14.7 Mb and SNPs slightly upstream (14.31–14.37 Mb) (Fig. 4a,b) for fruit maturity date. This region overlaps with a major QTL previously identified through linkage mapping (qP-FD4.2m / qP-MD4.2m) [22]. Several NAC transcription factors are located within this interval and represent plausible candidate genes. In sweet cherry, NAC family members have been associated with ripening phenology and fruit softening [23], and studies in related species have shown that structural variants affecting regulatory regions of ripening-related genes can influence fruit development and quality traits [30], highlighting the potential functional relevance of structural variation at this locus.

Regarding fruit firmness, GWAS results identified significant signals on chromosome 4 between 11.3 and 14.7 Mb (Additional file 2: Figure S9), aligning with the major QTL qP-FF4.1m reported by Calle & Wünsch [22], which explains up to 64.1% of phenotypic variance (interval of QTL: 10.41–12.57 Mb of peach genome). Previous studies have reported inconsistencies in the physical coordinates of this locus when different reference genomes are used, such as the shifts observed between the Tieton and Satonishiki genomes.

In apricot, Groppi et al. [45] identified four genes on chromosome 4 within the same genomic interval as that detected in our sweet cherry results for fruit maturity date and fruit firmness. The authors proposed several candidate genes, including a NAC-domain containing protein located between 15.97 and 15.98 Mb, suggested to play a role in fruit firmness and ripening, as reported in apple [47]. Consistent with this observation, we identified two NAC-domain containing genes (*NAC098* and *NAC056*) within the fruit firmness-associated interval

on chromosome 4 (11.3–14.7 Mb) in the Regina Hap1 genome, further supporting the potential involvement of NAC transcription factors in regulating fruit firmness in sweet cherry. In this context, the graph-based pangenome provides a unified coordinate framework that facilitates the comparison of association signals across datasets and helps refine the localization of this important breeding hotspot on chromosome 4 associated with fruit maturity date and fruit firmness.

Associations detected for fruit juice color were located near the *MYB10* gene cluster on chromosome 3 (Additional file 2: Figure S9). Previous studies have demonstrated that variation at this locus plays a central role in determining fruit pigmentation in *Prunus* species. In sweet cherry, a deletion of approximately 90 kb encompassing multiple *MYB10* genes has been associated with yellow fruit phenotypes [23]. The ability of pangenome-based analyses to incorporate structural variants alongside SNP markers may therefore facilitate the detection of complex allelic variation at loci where large insertions, deletions, or copy-number changes contribute to phenotypic diversity.

Recent integrative analyses of cherry genomics provide a broader context for these results. Liu et al. [7] compiled genetic linkage maps, QTL studies, GWAS results, and validated candidate genes across edible cherries and projected these data onto a common genomic framework, revealing that key agronomic traits such as fruit weight, firmness, maturity date, color, cracking resistance, and sugar or acid composition frequently map to a limited number of genomic regions detected across independent populations and experimental designs. Several of these loci represent recurrent hotspots, including the well-characterized region on chromosome 4 associated with ripening and firmness, as well as the *MYB10* cluster controlling fruit coloration on chromosome 3. The concordance between previously reported QTL/GWAS intervals and the loci identified in the present study supports the robustness of the associations detected using the pangenome-based framework and suggests that integrating structural variants and haplotype-resolved assemblies can help refine the boundaries of these conserved genetic regions.

More broadly, incorporating structural variants into association analyses is increasingly recognized as essential for improving the genetic resolution of complex traits. In tomato, for example, the inclusion of structural variants identified through graph-based pangenomes increased the estimated heritability of several traits compared with analyses based on a single linear reference genome [34]. Similar effects are likely to occur in perennial fruit crops, where large insertions, deletions, and presence/absence variation represent a substantial fraction of genomic diversity. The present results therefore illustrate how haplotype-resolved pangenomes provide an improved framework for capturing the full spectrum of genetic variation underlying agronomic traits and for identifying candidate loci that may be relevant for marker-assisted selection and breeding in sweet cherry.

Pangenome reveals genome architecture differentiation between landrace and modern accessions at QTL hotspots of chromosomes 2 and 4 in sweet cherry

Several loci identified through the pangenome-based analyses correspond to genomic regions previously reported in sweet cherry as important for breeding and trait variation, particularly on chromosomes 2 and 4, which have been repeatedly described as QTL hotspots controlling major fruit traits [20, 23, 48]. In particular, the genomic interval in the middle of chromosome 2, spanning approximately 6.3 Mb, has been repeatedly associated with traits of agronomic importance, including fruit size [49, 50, 51], fruit firmness [51], and flowering time [52]. Genomic regions presenting differential architecture were also detected on chromosome 4, corresponding to a well-characterized QTL hotspot located within a narrow interval (50–54 cM) previously identified through linkage mapping [22]. This region contains stable QTLs for fruit development time, maturity date, fruit firmness, and soluble solid content. Previous haplotype analyses have suggested reduced diversity in this region in modern cultivars compared with wild or traditional germplasm, particularly for alleles associated with increased fruit firmness in several stone fruit species [20]. Most modern cultivars carry firm-fruit alleles at this locus, whereas wild sweet cherry genotypes (mazzards) are typically homozygous for soft-fruit alleles. Two NAC transcription factors located within this region represent strong candidate genes, as NAC family members are known regulators of fruit ripening and softening in Rosaceae species [53]. The importance of this locus may reflect breeding objectives aimed at combining early maturity with high firmness, as haplotypes conferring early ripening, such as the one present in Cristobalina, are often linked to reduced fruit firmness.

The importance of chromosomes 2 and 4 is further supported by population-scale comparisons of wild genotypes, landraces, and modern cultivars. Pinosio et al. [54] identified several genomic regions showing reduced diversity and increased linkage disequilibrium in domesticated germplasm, including a strong signal on the second arm of chromosome 2. This interval overlaps with the major fruit weight and fruit size hotspot detected in the present study. Similarly, a region on chromosome 4 was reported to contain a NAC domain-containing gene showing negative Tajima's D values and strong expression in fruit tissue [54]. In apricot, Groppi et al. [45] identified differentiated genomic regions containing genes encoding an ABC transporter C family member on chromosome 2 and a NAC transcription factor on chromosome 4. These regions overlap with the genomic intervals detected here for fruit weight and fruit maturity date in sweet cherry. Because strong synteny is conserved among *Prunus* species, the correspondence between these loci suggests that similar genomic regions may contribute to the control of fruit traits across species. Although the present analysis did not directly test for

selective sweeps, the overlap with regions previously reported as targets of selection in related species raises the possibility that these loci have been repeatedly involved in breeding or domestication processes affecting fruit quality traits.

Beyond the major QTL hotspots on chromosomes 2 and 4, the genome-wide differentiation analysis revealed additional regions of elevated genetic divergence between landrace and modern breeding germplasm (Fig. 5b). Functional enrichment analysis of genes located within windows exceeding the 99th percentile of F_{st} identified several Gene Ontology categories associated with plant defense and specialized metabolism (Additional file 1: Table S5). Among the most significantly enriched terms were triterpenoid biosynthetic process, defense response to bacterium, and transmembrane receptor protein tyrosine kinase activity (Fig. 5c). Triterpenoids constitute a large class of plant secondary metabolites involved in protection against pathogens and herbivores [55], whereas receptor-like kinases play central roles in plant innate immunity by recognizing pathogen-associated molecular patterns and activating downstream defense signaling pathways [56].

Notably, many genes contributing to these enriched categories are located on chromosomes 3 and 4 (Fig. 5d), indicating localized regions of functional differentiation between landrace and modern breeding germplasm. Despite the enrichment of defense-related processes, overlap between high- F_{st} regions and the main GWAS peaks for fruit weight and maturity date was limited, suggesting that genomic regions showing differentiation between breeding levels are largely distinct from those associated with major fruit traits.

The absence of direct overlap between GWAS signals for fruit weight and F_{st} outlier windows likely reflects differences in resolution between the two approaches. F_{st} was estimated using fixed genomic windows, whereas GWAS signals correspond to individual variants. As a result, differentiation boundaries may fall just outside the defined windows, leading to near-miss cases despite regional co-localization. The observation that GWAS-associated SNPs lie within a few kilobases of high- F_{st} regions supports the existence of nearby genomic intervals contributing to both differentiation between breeding levels and trait variation.

Similar patterns have been reported in other perennial fruit crops. In apple, genomic regions associated with secondary metabolism and stress responses have been reported to differ between wild relatives and cultivated germplasm [40]. These observations are consistent with the idea that breeding programs focusing on fruit quality traits may affect a limited number of genomic regions, while other parts of the genome show differentiation among germplasm groups.

Overall, the combination of pangenome analysis, association mapping, and population genomics provides a coherent view of genomic differences associated with breeding level in sweet cherry. Together, these results demonstrate that haplotype-resolved pangenomes provide a powerful framework for capturing the full spectrum of genomic variation in sweet cherry, refining the localization of agronomically important loci, and facilitating the study of genome architecture across diverse germplasm. The integration of graph-based genomic resources with association mapping and population genomics will help improve trait discovery, support breeding strategies, and promote the efficient use of genetic diversity in sweet cherry. The relatively limited number of genomes included in the current pangenome may underestimate rare variants, and future expansions including additional wild and landrace accessions will further refine the structure of the sweet cherry pangenome.

Conclusions

The haplotype-resolved graph pangenome developed in this study provides a comprehensive representation of genomic diversity in sweet cherry and demonstrates the value of integrating multiple high-quality assemblies to capture both sequence and structural variation within the species. Despite strong chromosome-scale collinearity among accessions, the pangenome reveals extensive local polymorphism and a large accessory component, indicating that a substantial fraction of intraspecific diversity is not represented in any single reference genome. The incorporation of haplotype-resolved genomes enabled the detection of structural variants at high resolution and provided a unified coordinate system suitable for downstream comparative and association analyses.

By integrating SNPs and structural variants within the graph framework, genome-wide association analyses identified candidate loci for major agronomic traits, including fruit weight, maturity date, firmness, juice color, and cracking susceptibility. Several of these loci are located on chromosomes 2 and 4, which also contain regions showing elevated genetic differentiation between landrace and modern breeding accessions. These results indicate that variation affecting important fruit traits is concentrated in a limited number of genomic regions, while other parts of the genome remain more conserved across breeding levels.

Genome-wide differentiation analyses revealed that divergence between landrace and modern breeding germplasm is localized to a restricted number of genomic windows rather than distributed across entire chromosomes. Genes located within highly differentiated regions were enriched for functions related to membrane components, lipid-associated processes, and defense-related pathways, indicating that differences between breeding levels involve specific functional categories rather than widespread genome-wide changes.

The large accessory fraction identified in the pangenome highlights the extent of structural and sequence diversity present across sweet cherry genomes and underscores the importance of using multiple assemblies to represent the species. This diversity provides a valuable resource for identifying genetic variants associated with agronomic traits and for improving genomic tools for breeding.

Overall, this work illustrates how haplotype-resolved pangenomes provide an improved framework for studying genome diversity, trait architecture, and genomic differences among breeding levels in perennial fruit crops. The integration of graph-based genomic resources with association mapping and population genomic analyses facilitates the identification of candidate regions underlying phenotypic variation and supports the efficient use of genetic diversity in sweet cherry breeding programs. As additional genomes become available, expanding graph-based pangenomes will further refine the characterization of structural variation and trait-associated loci, providing a robust foundation for future genomic studies and breeding applications in sweet cherry and other perennial fruit species.

Methods

Genome dataset

The sweet cherry pangenome was constructed from chromosome-scale *Prunus avium* assemblies representing 15 accessions (Table 1). The dataset comprised eleven newly generated diploid assemblies resolved into phased haplotypes, two previously published haplotype-resolved genomes (Regina and Santina), and two publicly available consensus assemblies (Tieton v2 and Satonishiki). To minimize artifacts associated with fragmented assemblies, only sequences with at least 80% of their assembled length anchored to chromosome-scale pseudomolecules were retained. One haplotype, V2775.hap1, did not meet this threshold and was excluded. After filtering, the final dataset contained 27 chromosome-scale genome sequences, including 25 phased haplotypes and 2 consensus assemblies. Only the eight assembled pseudomolecules were retained for pangenome construction; unplaced and unlocalized scaffolds were excluded. Regina haplotype 1 (Regina Hap1) was used as the reference path for graph construction and downstream coordinate-based analyses.

Haplotype-resolved genome assemblies

High-fidelity (HiFi) PacBio long-read sequencing data were used to generate *de novo* assemblies for eleven sweet cherry genotypes (Additional file 1: Table S1). Assemblies were produced using the Asm4pg v1.1.0 pipeline (<https://forge.inrae.fr/asm4pg/GenomAsm4pg>), which integrates quality control, assembly, and polishing steps optimized for plant genomes.

Genome assembly was performed with hifiasm (v0.24.0-r703) using default parameters to generate two haplotype-resolved assemblies (hap1 and hap2) [54]. Contigs were scaffolded using RagTag (v2.0.1) [36] with the *Tieton Genome v2.0* assembly (<https://www.rosaceae.org/Analysis/9262820>) as reference. Scaffolds were then anchored and ordered into chromosome-scale pseudomolecules using ALLMAPS, based on high-density genetic linkage maps, which were derived from studies by Klagges et al. [58], Castède et al. [52, 59] Calle et al. [21, 22], Quero-García et al. [60], and Branchereau et al. [61].

Each phased assembly comprised eight chromosome-scale pseudomolecules, corresponding to the eight sweet cherry chromosomes, together with unscaffolded sequences retained separately. Whole-genome alignments against the Tieton v2.0 reference were performed using D-GENIES [62] to verify chromosome orientation. Inverted regions were corrected by reverse-complementing sequences using EMBOSS revseq and reintegrating them into the assemblies.

Assembly completeness was assessed using BUSCO (v5.3.1) [63] with the eudicots_odb10 lineage dataset.

Plant material for SV- and SNP-based GWAS using short-reads

Plant material consisted of a panel of 122 sweet cherry accessions (*Prunus avium* L.), belonging to the sweet cherry genetic resources collection maintained by the INRAE *Prunus-Juglans* Biological Resources Center [64]. Trees are grafted, with one replicate per accession, and are located at the Fruit Tree Experimental Unit of INRAE in Bourran, France (GPS coordinates: 44.33463359165044, 0.4125325574726346). The panel was selected in order to: (i) include a maximum of genetic diversity from a collection of 210 accessions based on a previous study [26], and (ii) cover a wide range of phenotypic variability in fruit quality traits. The panel includes accessions from France (44%). The remaining accessions originate mainly from Europe (27%), America (22%), Asia (4%), or are of unknown origin (3%). The accessions can be divided into four breeding categories: landraces (41%), early selections (11%), modern cultivars (44%), and unknown (4%). Full details of the accessions are provided in Additional file 1: Table S6.

Phenotyping and BLUPs calculation

A total of five traits related to fruit quality were phenotyped among the 122 accessions between 2000 and 2019: fruit maturity date, fruit weight, fruit juice color, fruit stem end cracking, and fruit firmness. Fruit maturity date corresponds to when the fruits have reached their final size with advanced coloring (BBCH stage 85) [65], expressed in Julian days (i.e., the sequential day of the year counted from January 1st, ranging from 1 to 365 or 366 in leap years). Fruit weight is the average weight of ten fruits measured with an electronic balance [24], expressed in grams. Fruit juice color is observed on ten fruits, with an ordinal scale from colorless (1) to black red (9), according to the ECPGR *Prunus* Database Descriptor n°35

(https://www.ecpgr.org/fileadmin/templates/ecpgr.org/upload/NW_and_WG_UPLOADS/Prunus/EPDB_New_list_of_descriptors_2011.pdf).

Fruit stem-end cracking was assessed visually as the presence of cracks in the stem cavity on 50 fruits [60], given in percentage. Fruit firmness is measured using a Durofel® texture analyzer on ten fruits, with two measurements per fruit [24]. Raw phenotypic data are provided in Additional file 1: Table S7. The number of years in which at least one accession was phenotyped for each trait is as follows: 18 years for fruit maturity date, 16 years for fruit weight, 12 years for fruit juice color, 8 years for fruit stem-end cracking, and 10 years for fruit firmness.

Because the phenotypes were not available for all accessions and all years, the means of genotypic effects were obtained for each accession (one replication per accession) by adjusting for year using a mixed linear model to produce the Best Linear Unbiased Predictions (BLUPs). The BLUPs were predicted by adjusting for year as a fixed effect as follows:

$$P_{ik} = \mu + Y_i + g_k + e_{ik},$$

where P_{ik} is the observed phenotype of accession k in year i , μ is the overall mean, Y_i is the fixed effect of year, g_k is the random effect of accession, and e_{ik} is the residual error. The BLUPs were estimated using the R package lme4 [66], and the results are provided in Additional file 1: Table S6.

Genomic distance estimation and guide tree inference

A guide phylogeny was inferred to parameterize progressive pangenome construction. A pangenome database was built using PanTools (v4.3.3) [67] from chromosome-scale assemblies, with k-mer counting performed using KMC (v3.2.4) [68]. The pangenome was constructed using a k-mer size of 21. Pairwise genome distances were derived from the k-mer-based representation and used to generate a Newick-formatted guide tree. Regina haplotype 1 (Regina Hap1) was specified as the primary reference lineage for downstream graph construction and reference-based projections.

Alignment-free genome distance estimation and network analysis

Genome-wide similarity across assemblies was additionally assessed using Mash (v2.2.2) [69]. A combined multi-genome FASTA file was generated and sequence identifiers were standardized to encode accession, haplotype, and chromosome. Pairwise distances were computed using a sketch size of 10,000. The resulting Mash distance matrix was converted to a network representation using the mash2net.py utility distributed with PGGB (v0.7.4) [70], retaining edges with Mash distance ≤ 0.05 to visualize community structure and relative genome similarity.

Pangenome construction

A chromosome-scale *Prunus avium* pangenome was constructed using the cactus-pangenome workflow implemented in Cactus/MiniCactus v2.6.13 (71, 72), following the same methodology described by Blanchard et al. [73] for apricot. Progressive multiple-genome alignment was performed independently for each of the eight chromosomes using chromosome-specific dataset files and a guide tree inferred from k-mer-based genomic distances. Regina haplotype 1 was designated as the primary reference and was preserved as the unclipped reference path throughout graph construction.

Graph construction was run with VCF projection enabled to allow reference-based extraction of sequence variants relative to Regina Hap1. Clipping was applied with a maximum unaligned segment length of 10 kb (default), and frequency filtering retained sequences present in at least one genome ($-\text{filter } 1$). Haplotype-aware graph construction was enabled to preserve phased information ($-\text{haplo}$).

For each chromosome, the workflow generated full, clipped, and filtered graph representations in GBZ and GFA formats, as well as ODGI graph representations and vg giraffe-compatible indexes to enable downstream read mapping and variant extraction.

Graph-derived sequence and structural variation

Two complementary variant datasets were generated in this study. First, sequence and structural variants intrinsic to the pangenome were derived directly from the multiple-genome alignment generated by Cactus/MiniCactus (v2.6.13) [71, 72], capturing differences among the assembled haplotypes included in the graph. Second, population-scale variants were identified from short-read sequencing data across the diversity panel (see sections below) and used for association and population genomic analyses. For each chromosome, VCF projections produced during graph construction were used to extract single-nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) using bcftools [74]. Variant classes were separated using allele-type filters, retaining all bi- and multi-allelic sites. These graph-derived SNPs and indels represent sequence differences among the haplotypes incorporated into the pangenome and were used to quantify chromosome-level and genome-wide sequence diversity across assembled genotypes.

Structural variants describing genomic differences among the assemblies included in the pangenome were obtained directly from the HAL multiple alignment using HAL tools (Cactus module v2.9.8). Branch-specific mutation events were extracted for each genome and haplotype using halBranchMutations, with alignment fragmentation controlled using a maximum gap size of 50 bp and excluding regions containing ambiguous sequence (maxNfraction = 0). Only events ≥ 20 bp were retained. These variants correspond to insertion, deletion, and rearrangement events resolved in the pangenome alignment and represent structural differences among the haplotypes incorporated into the graph.

Chromosome-scale density profiles for SNPs, indels, and structural variants were computed in non-overlapping 100 kb windows. SVs were assigned to non-overlapping 100 kb windows according to their genomic midpoint, and densities were normalized per variant type (p95 for SNPs/indels, p85 for SVs). The resulting tracks were visualized as vertical chromosome heatmaps in R, allowing comparison of large-scale patterns of variation across the eight *P. avium* chromosomes.

Together, these graph-derived SNPs, indels, and SVs describe sequence and structural diversity within the assembled pangenome and are distinct from population-level variants obtained from short-read genotyping. These variants are hereafter referred to as pangenome SVs.

Pangenome growth analysis

Pangenome growth dynamics were analyzed using Panacus v0.4.1 [75]. The ordered-hist-growth module was applied to sequentially incorporate genomes and quantify changes in core and accessory genome fractions. Core genome size was defined as sequences present in at least 90% of genomes, whereas accessory regions were defined as sequences present in fewer than 90% of genomes. Growth curves were used to evaluate whether the *P. avium* pangenome exhibits open or closed dynamics, providing insights into genome diversification during cultivar differentiation.

Short-read alignment to the pangenome and linear projection for population-level genotyping

To characterize structural and sequence variation segregating in the diversity panel used for association analyses, a second dataset of variants was generated from short-read resequencing data mapped to the pangenome graph.

Short-read sequencing data from 122 *Prunus avium* accessions were aligned against the chromosome-scale pangenome graph using vg giraffe (vg v1.65.0) [76]. A combined graph representation was generated from chromosome-level GFA files using vg combine and indexed for giraffe mapping via vg autoindex. The resulting GBZ index was used for haplotype-aware graph-based alignment.

Paired-end reads were mapped with sample-specific read group identifiers, producing GAM-format alignments against the full pangenome graph. To enable consistent coordinate-based variant calling across individuals, alignments were projected onto the linear coordinate system of Regina haplotype 1 (Regina Hap1) using vg surject. Surjection was restricted to the eight Regina Hap1 chromosomal paths to maintain coordinate uniformity across all accessions. Surjected alignments were converted to BAM format, sorted, and indexed using samtools (v1.21) [71]. When multiple sequencing libraries were available for a given accession, replicate BAM files were merged prior to downstream analyses.

For benchmarking purposes, reads were also aligned to the linear reference genome using minimap2 (v2.28) [77] in short-read mode (`-ax sr`). However, all downstream variant discovery and association analyses were performed using graph-based alignments projected onto Regina Hap1 coordinates.

Importantly, variants identified from these read-based alignments represent population-scale polymorphisms segregating across the 122 accessions and are analytically distinct from the SNPs, indels, and structural variants derived directly from the multiple-genome alignment of the assembled pangenome. The latter describe structural and sequence differences among the assemblies used to build the pangenome itself, whereas the read-based variants were used for genome-wide association and selection analyses.

Genome-wide association analyses were performed using the 122 *Prunus avium* accessions described above. SNP and SV discovery was based on short-read alignments mapped to the pangenome graph and projected onto Regina Hap1 coordinates.

SNP discovery and filtering

Cohort SNP calling was performed using bcftools (v1.21) [71]. Genotype likelihoods were computed with bcftools *mpileup* using minimum mapping quality 20 and minimum base quality 20, excluding indels at this stage. Multiallelic variant calling was conducted using bcftools *call*, followed by normalization against the Regina Hap1 reference and decomposition of multiallelic records. Variants were restricted to SNPs and sorted prior to filtering.

Quality filtering retained variants with $QUAL \geq 30$, missing genotype rate ($F_MISSING$) < 0.20 , and allele frequency between 0.01 and 0.99. Allele frequency and missingness statistics were computed using bcftools *+fill-tags*. The final filtered cohort VCF was converted to rMVP-compatible genotype matrices by encoding genotypes as 0, 1, and 2 corresponding to homozygous reference, heterozygous, and homozygous alternate states, respectively. Duplicate SNP identifiers were resolved by generating unique remapped IDs while preserving links to original variant identifiers. Marker position files were exported in both string-based and numeric chromosome formats.

Structural variant discovery and filtering

Population structural variants were identified using DELLY (v1.1.8) [78] from Regina Hap1-aligned BAM files. SVs were first called independently per accession based on paired-end and split-read evidence. A union set of candidate SV sites was generated using *delly* *merge*, followed by joint genotyping across all accessions to ensure consistent genotype calls at shared loci.

Multi-sample VCF files were merged and normalized using bcftools (v1.21). Structural variant length (SVLEN) was calculated as the absolute difference between END and POS coordinates when not explicitly provided. Breakend (BND) variants were excluded from downstream analyses.

Three structural variant panels were generated to assess robustness of association results to call stringency. A permissive panel retained non-BND variants with $|SVLEN| \geq 20$ bp. A high-confidence panel retained non-BND variants with $|SVLEN| \geq 50$ bp that were classified as PRECISE and either passed filtering (FILTER = PASS) or were unfiltered. A size-restricted comparison panel retained non-BND variants between 50 bp and 100 kb. SV genotypes were encoded numerically (0/1/2) for rMVP compatibility, and corresponding marker position files were generated.

Association testing

Association analyses were conducted using the R package rMVP v1.4.5 [79]. SNP and SV genotype matrices were analyzed separately under both general linear models (GLM) and mixed linear models (MLM), allowing comparison between association signals detected without and with explicit correction for relatedness and population structure. Analyses were performed on 122 *P. avium* accessions using pre-calculated BLUP values extracted from the phenotypic dataset. Population structure was controlled using principal component analysis, with the first three principal components included as covariates in both GLM and MLM analyses. The MLM additionally incorporated a genomic relationship matrix calculated using the VanRaden method, as implemented in rMVP, to account for relatedness among accessions..

For both SNPs and SVs, markers were filtered prior to association testing by retaining variants with call rate ≥ 0.8 and minor allele frequency (MAF) ≥ 0.01 . Missing genotypes were imputed using the mean genotype value for each marker. For the SV-based GWAS, the structural variant panel consisted of non-BND variants with absolute SV length ≥ 20 bp. Genome-wide significance thresholds were defined

using a Bonferroni correction based on the number of markers tested in each analysis, and association results were visualized using Manhattan and quantile–quantile plots.

Investigation of genomic regions presenting differential architecture according to breeding level

To investigate genomic regions potentially influenced by breeding level, genetic differentiation between 44 landrace and 61 modern breeding accessions was assessed using SNP-derived F_{st} estimates.

Accessions were classified according to breeding level, and only individuals with complete genotype data were retained for analysis. SNP genotypes were extracted from the filtered cohort VCF used for GWAS and subset to landrace and modern breeding groups. Only biallelic SNPs with sufficient genotype representation were included.

Population differentiation was estimated using Hudson's F_{st} . For each SNP, allele frequencies were calculated separately for landrace and modern breeding groups based on alternate allele counts. Within-population nucleotide diversity and between-population divergence were computed, and per-site F_{st} values were derived accordingly.

To identify genomic regions exhibiting elevated differentiation, SNP-level F_{st} was calculated using 100 kb non-overlapping windows, a size supported by genome-wide LD decay analysis, which showed that linkage disequilibrium decreases to $r^2 \approx 0.2$ at ~ 100 kb (Additional file 2: Figure S10).. For each window, summary statistics including mean and maximum F_{st} were calculated. Empirical thresholds were defined using the genome-wide distribution of window-based mean F_{st} values. Windows exceeding the 99th percentile were considered candidate regions of elevated differentiation.

The chromosomal distribution of high- F_{st} windows was examined to identify enrichment patterns and to assess their positional overlap with loci identified through genome-wide association analyses, with particular attention to fruit maturity date, fruit weight, fruit firmness and fruit stem-end cracking signals detected under the GLM model.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The sweet cherry pangenome graph (HAL format), representing the multiple genome alignment used in this study, has been submitted to the Genome Database for Rosaceae (GDR) (accession ID pending).

The Regina haplotype 1 (HapA) genome assembly and its annotation, used as the reference coordinate system, are available at GDR. Functional analyses were performed using the gene annotation of the Regina haplotype 1 (HapA) assembly.

Phenotypic data for the 122 accessions used for GWAS are provided in Additional file 1: Table S7. All scripts and workflows used in this study are available at https://github.com/ClaudioUPz/Sweet_cherry_pangenome.

Raw PacBio HiFi long reads from 11 sweet cherry accessions used for de novo assembly have been submitted to the European Nucleotide Archive (ENA) under accession numbers ERR16914830, ERR16914829, ERR16913752, ERR16913751, ERR16913700, ERR16913699, ERR16913692, ERR16913691, ERR16913610, ERR16913544, and ERR16908726.

Competing interests

The authors declare that they have no competing interests.

Funding

Long-read sequencing on 11 newly assembled genomes was funded by the French National Research Agency, France 2030 Research Program PEPR Agroécologie et Numérique, AGRODIV (ANR-22-PEAE-0005-AgroDiv) flagship and BRelf (ANR-22-PEAE-0014).

Part of short-read sequencing of the GWAS panel was funded by Pr. Caixi Zhang, at the Department of Plant Science, School of Agriculture and Biology, Shanghai Jiao Tong University, Minhang, Shanghai, 200240, China.

Another part of the short-read sequencing of the GWAS panel was funded by the PrADAm team's own funds. Team "*Prunus*: Adaptation, Diversity, Breeding", belonging to INRAE Nouvelle-Aquitaine Bordeaux Center, BFP Unit "Fruit Biology and Pathology", Villenave d'Ornon, 33140, France.

Long-read and short-read sequencing of Santana and Regina varieties were funded by Agencia Nacional de Investigación y Desarrollo (ANID/Chile) ANID/ACT210007, Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT/Chile) grant 1230163 and ANID/C203020001 (IE2501).

Dr. Claudio Urra and Ismaël Blanchard developed this work during Chile-France bilateral missions financed by Agencia Nacional de Investigación y Desarrollo (ANID/Chile) ANID/ECOS230011 and ECOS Sud-ANID France-Chile action ECOS n° C23B01. Development of the pangenome pipeline benefited of support from the HE FRUITDIV project (#101133964) and conceptual help from the Pangenome network of the PEPR Agroécologie et Numérique, AGRODIV (ANR-22-PEAE-0005-AgroDiv) flagship.

Authors' contributions

C.U. and I.B. contributed equally to this work.

C.U., I.B., Q.-T.B., and A.B. designed the study and coordinated the pangenome and population genomics analyses. I.B. developed the computational pipelines for genome preprocessing, phylogenetic inference, graph pangenome construction using MiniCactus, and GWAS analyses based on short-read data, including read mapping and variant calling. C.U. adapted and implemented these pipelines for sweet cherry and performed downstream analyses, including pangenome-derived variant characterization, Fst analyses, integration of Fst with GWAS signals, and gene annotation-based interpretation.

C.U. performed formal data analyses and prepared the figures.

C.U. and Q.-T.B. curated the data and prepared submissions to public repositories.

Q.-T.B. generated the haplotype-resolved genome assemblies for the 11 newly sequenced accessions. A.B. provided phenotypic characterization and BLUP estimations.

A.M.A., V.D., and B.W. contributed to the interpretation of the results and to the discussion. E.D. and J.Q.-G. provided guidance in the selection of the GWAS panel and the long-read sequencing panel. B.L. contributed to the scripting. C.Z. provided part of the short-read data.

C.U., A.B., and Q.-T.B. wrote the original draft of the manuscript with input from all authors. All authors reviewed and approved the final manuscript.

Acknowledgements

We thank the INRAE Prunus-Juglans Biological Resources Center for maintaining the sweet cherry collection. More information is available at <https://doi.org/10.17180/WN42-3J20>; Prunus-Juglans BRC, member of BRC4Plants, INRAE, 2024, Biological Resource Centers for Plants of AgroBRC-RARE.

We thank the Fruit Tree Experimental Unit of INRAE in Bourran for assistance with leaf collection. More information is available at <https://doi.org/10.17180/9ST1-4J21>; UEA, Arboricultural Experimental Facility, INRAE, 2024.

We thank the GENTYANE platform (INRAE, GENoTYpage et séquençage en Auvergne, 2026; <https://doi.org/10.15454/1.5572409592543596E12>) for performing library preparation and sequencing, and CNRGV French Plant Genomic Resource Center, <http://doi.org/10.15454/1.5572367923221042E12> for high-molecular weight DNA extraction.

We are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, <https://doi.org/10.15454/1.5572369328961167E12>) and the Institut Français de Bioinformatique (IFB) Core Cluster (ANR-11-INBS-0013)

for providing computing and storage resources.

We thank Vanita Haurheeram (Université Paris-Saclay, INRAE, BioinfOmics, URGI, Versailles, France) for her assistance with ENA data submission.

References

1. Purugganan MD, Fuller DQ. The nature of selection during plant domestication. *Nature*. 2009;457(7231):843–8. 10.1038/nature07895.
2. Petereit J, Bayer PE, Thomas WJW, Tay Fernandez CG, Amas J, Zhang Y, et al. Pangenomics and crop genome adaptation in a changing climate. *Plants (Basel)*. 2022;11(15):1949. 10.3390/plants11151949.
3. Raza A, Li Y, Jan F, Fernandez CGT, Mir RR, Hu Z, et al. From the genome to super-pangenome: a new paradigm for accelerated crop improvement. *npj Sci Food*. 2026;2(1):4. 10.1038/s44383-025-00019-z.
4. Jayakodi M, Shim H, Mascher M. What are we learning from plant pangenomes? *Annu Rev Plant Biol*. 2025;76(1):663–86. 10.1146/annurev-arplant-090823-015358.
5. Kaur H, Shannon LM, Samac DA. A stepwise guide for pangenome development in crop plants: an alfalfa (*Medicago sativa*) case study. *BMC Genomics*. 2024;25(1):1022. 10.1186/s12864-024-10931-w.
6. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet*. 2018;50(2):278–84. 10.1038/s41588-018-0041-z.
7. Liu Z, Wang N, Su Y, Long Q, Peng Y, Shanguan L, et al. Grapevine pangenome facilitates trait genetics and genomic breeding. *Nat Genet*. 2024;56(12):2804–14. 10.1038/s41588-024-01967-5.
8. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. *Nat Plants*. 2020;6(8):914–20. 10.1038/s41477-020-0733-0.
9. Schreiber M, Jayakodi M, Stein N, Mascher M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat Rev Genet*. 2024;25(8):563–77. 10.1038/s41576-024-00691-4.
10. Espinosa E, Bautista R, Larrosa R, Plata O. Advancements in long-read genome sequencing technologies and algorithms. *Genomics*. 2024;116(3):110842. 10.1016/j.ygeno.2024.110842.
11. Hu H, Wang J, Nie S, Zhao J, Batley J, Edwards D. Plant pangenomics, current practice and future direction. *Agric Commun*. 2024;2(2):100039. 10.1016/j.agrcom.2024.100039.
12. Quero-García J, Schuster M, López-Ortega G, Charlot G. Sweet cherry varieties and improvement. In: Quero-García J, Iezzoni A, Puławska J, Lang G, editors. *Cherries: botany, production and uses*. Oxfordshire (UK): CABI; 2017. pp. 60–94.
13. Labbancz J, Dhingra A. Tree fruit and nut crops at the dawn of the pangenomic era. *Horticulturae*. 2025;11(12):1537. 10.3390/horticulturae11121537.
14. Wang T, Duan S, Xu C, Wang Y, Zhang X, Xu X, et al. Pan-genome analysis of 13 *Malus* accessions reveals structural and sequence variations associated with fruit traits. *Nat Commun*. 2023;14(1):7377. 10.1038/s41467-023-43270-7.
15. Huang Y, He J, Xu Y, Zheng W, Wang S, Chen P, et al. Pangenome analysis provides insight into the evolution of the orange subfamily and a key gene for citric acid accumulation in citrus fruits. *Nat Genet*. 2023;55(11):1964–75. 10.1038/s41588-023-01516-6.
16. Gabur I, Chawla HS, Snowdon RJ, Parkin IAP. Connecting genome structural variation with complex traits in crop plants. *Theor Appl Genet*. 2019;132(3):733–50. 10.1007/s00122-018-3233-0.
17. Yuan Y, Bayer PE, Batley J, Edwards D. Current status of structural variation studies in plants. *Plant Biotechnol J*. 2021;19(11):2153–63. 10.1111/pbi.13646.
18. Guo J, Cao K, Deng C, Li Y, Zhu G, Fang W, et al. An integrated peach genome structural variation map uncovers genes associated with fruit traits. *Genome Biol*. 2020;21(1):258. 10.1186/s13059-020-02169-y.
19. Huang X, Lin X, Zhou P, Tan W, Gao F, Ni Z, et al. Pangenome analysis reveals structural variations associated with citric acid accumulation in *Prunus mume*. *Plant Biotechnol J*. 2025;1–19. 10.1111/pbi.70518.
20. Cai L, Quero-García J, Barreneche T, Dirlewanger E, Saski C, Iezzoni A. A fruit firmness QTL identified on linkage group 4 in sweet cherry (*Prunus avium* L.) is associated with domesticated and bred germplasm. *Sci Rep*. 2019;9(1):5008. 10.1038/s41598-019-41484-8.
21. Calle A, Serrano M, Wünsch A. Genetic linkage map and QTL analysis for fruit quality traits in sweet cherry. *Mol Breed*. 2020;40:34. 10.1007/s11032-020-01165-1.

22. Calle A, Wünsch A. Multiple-population QTL mapping of maturity and fruit-quality traits reveals LG4 region as a breeding target in sweet cherry (*Prunus avium* L.). *Hortic Res.* 2020;7:127. 10.1038/s41438-020-00349-2.
23. Holušová K, Čmejlová J, Suran P, Čmejla R, Sedlák J, Zelený L, et al. High-resolution genome-wide association study of a large Czech collection of sweet cherry (*Prunus avium* L.) on fruit maturity and quality traits. *Hortic Res.* 2022;10(1):uhac233. 10.1093/hr/uhac233.
24. Donkpegan ASL, Bernard A, Barreneche T, Quero-García J, Bonnet H, Fouché M, et al. Genome-wide association mapping in a sweet cherry germplasm collection (*Prunus avium* L.) reveals candidate genes for fruit quality traits. *Hortic Res.* 2023;10(10):uhad191. 10.1093/hr/uhad191.
25. Mariette S, Tavaud M, Arunyawat U, Capdeville G, Millan M, Salin F. Population structure and genetic bottleneck in sweet cherry estimated with SSRs and the gametophytic self-incompatibility locus. *BMC Genet.* 2010;11:77. 10.1186/1471-2156-11-77.
26. Campoy JA, Lerigoleur-Balsemin E, Christmann H, Beauvieux R, Girollet N, Quero-García J, et al. Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of *Prunus avium* L. landraces and bred cultivars. *BMC Plant Biol.* 2016;16:49. 10.1186/s12870-016-0712-9.
27. Wang J, Liu W, Zhu D, Hong P, Zhang S, Xiao S, et al. Chromosome-scale genome assembly of sweet cherry (*Prunus avium* L.) cv. Tieton obtained using long-read and Hi-C sequencing. *Hortic Res.* 2020;7(1):122. 10.1038/s41438-020-00343-8.
28. Urrea C, Gaete-Loyola J, Bui QT, Povea P, Carrasco N, Moraga C, et al. Chromosome-scale genome assembly of the Santana and Regina varieties of *Prunus avium*. *Tree Genet Genomes.* 2026;22:6. 10.1007/s11295-026-01732-1.
29. Shirasawa K, Isuzugawa K, Ikenaga M, Saito Y, Yamamoto T, Hirakawa H, et al. The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. *DNA Res.* 2017;24(5):499–508. 10.1093/dnares/dsx020.
30. Chen W, Xie Q, Fu J, Li S, Shi Y, Lu J, et al. Graph pangenome reveals the regulation of malate content in blood-fleshed peach by NAC transcription factors. *Genome Biol.* 2025;26(1):7. 10.1186/s13059-024-03470-w.
31. The International Peach Genome Initiative, Verde I, Abbott A, et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet.* 2013;45(5):487–94. 10.1038/ng.2586.
32. Cao K, Zheng Z, Wang L, Liu X, Zhu G, Fang W, et al. Comparative population genomics identified genomic regions and candidate genes associated with fruit domestication traits in peach. *Plant Biotechnol J.* 2014;12(3):338–50. 10.1111/pbi.12166.
33. Alioto T, Alexiou KG, Bardil A, Barteri F, Castanera R, Cruz F, et al. Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *Plant J.* 2020;101(2):455–72. 10.1111/tpj.14538.
34. Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature.* 2022;606(7914):527–34. 10.1038/s41586-022-04808-9.
35. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods.* 2018;15(6):461–68. 10.1038/s41592-018-0001-7.
36. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell.* 2020;182(1):145–e6123. 10.1016/j.cell.2020.05.021.
37. Yang S, Wang Y, Huang Q, Wang M, Wang S, Fu X, et al. A pangenome of maize provides genetic insights into drought resistance. *Nat Genet.* 2025;57(11):2831–41. 10.1038/s41588-025-02378-w.
38. Yang L, He W, Zhu Y, Lv Y, Li Y, Zhang Q, et al. GWAS meta-analysis using a graph-based pan-genome enhanced gene mining efficiency for agronomic traits in rice. *Nat Commun.* 2025;16(1):3171. 10.1038/s41467-025-58081-1.
39. Aranzana MJ, Decroocq V, Dirlewanger E, Eduardo I, Gao ZS, Gasic K, et al. *Prunus* genetics and applications after de novo genome sequencing: achievements and prospects. *Hortic Res.* 2019;6:58. 10.1038/s41438-019-0140-8.
40. Su Y, Yang X, Wang Y, Li J, Long Q, Cao S, et al. Phased telomere-to-telomere reference genome and pangenome reveal an expansion of resistance genes during apple domestication. *Plant Physiol.* 2024;195(4):2799–814. 10.1093/plphys/kiae258.
41. Lee HK, Cho SK, Son O, Xu Z, Hwang I, Kim WT. Drought stress-induced Rma1H1, a RING membrane-anchor E3 ubiquitin ligase homolog, regulates aquaporin levels via ubiquitination in transgenic *Arabidopsis* plants. *Plant Cell.* 2009;21(2):622–41. 10.1105/tpc.108.061994.
42. Pavan S, Jacobsen E, Visser RGF, Bai Y. Loss of susceptibility as a novel breeding strategy for durable and broad-spectrum resistance. *Mol Breed.* 2010;25(1):1–12. 10.1007/s11032-009-9323-6.
43. Wang Y, Kong L, Wang W, Qin G. Global ubiquitinome analysis reveals the role of E3 ubiquitin ligase FaBRIZ in strawberry fruit ripening. *J Exp Bot.* 2023;74(1):214–32. 10.1093/jxb/erac400.

44. Martínez-Esteso MJ, Sellés-Marchart S, Lijavetzky D, Pedreño MA, Bru-Martínez R. A DIGE-based quantitative proteomic analysis of grape berry flesh development and ripening reveals key events in sugar and organic acid metabolism. *J Exp Bot.* 2011;62(8):2521–69. 10.1093/jxb/erq434.
45. Groppi A, Liu S, Cornille A, Decroocq S, Bui QT, Tricon D, et al. Population genomics of apricots unravels domestication history and adaptive events. *Nat Commun.* 2021;12(1):3956. 10.1038/s41467-021-24283-6.
46. Ofori PA, Geisler M, di Donato M, Pengchao H, Otagaki S, Matsumoto S, et al. Tomato ATP-binding cassette transporter SIABCB4 is involved in auxin transport in the developing fruit. *Plants (Basel).* 2018;7(3):65. 10.3390/plants7030065.
47. Migicovsky Z, Yeats TH, Watts S, Song J, Forney CF, Burgher-MacLellan K, et al. Apple ripening is controlled by a NAC transcription factor. *Front Genet.* 2021;12:671300. 10.3389/fgene.2021.671300.
48. Cai L, Voorrips RE, van de Weg E, Peace C, Iezzoni A. Genetic structure of a QTL hotspot on chromosome 2 in sweet cherry indicates positive selection for favorable haplotypes. *Mol Breed.* 2017;37:85. 10.1007/s11032-017-0689-6.
49. Zhang G, Sebolt AM, Sooriyapathirana SS, et al. Fruit size QTL analysis of an F1 population derived from a cross between a domesticated sweet cherry cultivar and a wild forest sweet cherry. *Tree Genet Genomes.* 2010;6:25–36. 10.1007/s11295-009-0225-x.
50. Rosyara UR, Bink MCAM, van de Weg E, et al. Fruit size QTL identification and the prediction of parental QTL genotypes and breeding values in multiple pedigreed populations of sweet cherry. *Mol Breed.* 2013;32:875–87. 10.1007/s11032-013-9916-y.
51. Campoy JA, Le Dantec L, Barreneche T, et al. New insights into fruit firmness and weight control in sweet cherry. *Plant Mol Biol Rep.* 2015;33:783–96. 10.1007/s11105-014-0773-6.
52. Castède S, Campoy JA, García JQ, Le Dantec L, Lafargue M, Barreneche T, et al. Genetic determinism of phenological traits highly affected by climate change in *Prunus avium*: flowering date dissected into chilling and heat requirements. *New Phytol.* 2014;202(2):703–15. 10.1111/nph.12658.
53. Pirona R, Eduardo I, Pacheco I, Da Silva Linge C, Miculan M, Verde I, et al. Fine mapping and identification of a candidate gene for a major locus controlling maturity date in peach. *BMC Plant Biol.* 2013;13:166. 10.1186/1471-2229-13-166.
54. Pinoso S, Marroni F, Zuccolo A, Vitulo N, Mariette S, Sonnante G, et al. A draft genome of sweet cherry (*Prunus avium* L.) reveals genome-wide and local effects of domestication. *Plant J.* 2020;103(4):1420–32. 10.1111/tpj.14809.
55. Thimmappa R, Geisler K, Louveau T, O'Maille P, Osbourn A. Triterpene biosynthesis in plants. *Annu Rev Plant Biol.* 2014;65:225–57. 10.1146/annurev-arplant-050312-120229.
56. Tang D, Wang G, Zhou JM. Receptor kinases in plant-pathogen interactions: more than pattern recognition. *Plant Cell.* 2017;29(4):618–37. 10.1105/tpc.16.00891.
57. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods.* 2021;18(2):170–75. 10.1038/s41592-020-01056-5.
58. Klagges C, Campoy JA, Quero-García J, Guzmán A, Mansur L, Gratacós E, et al. Construction and comparative analyses of highly dense linkage maps of two sweet cherry intra-specific progenies of commercial cultivars. *PLoS ONE.* 2013;8(1):e54743. 10.1371/journal.pone.0054743.
59. Castède S, Campoy JA, Le Dantec L, Quero-García J, Barreneche T, Wenden B, et al. Mapping of candidate genes involved in bud dormancy and flowering time in sweet cherry (*Prunus avium*). *PLoS ONE.* 2015;10(11):e0143250. 10.1371/journal.pone.0143250.
60. Quero-García J, Letourmy P, Campoy JA, Branchereau C, Malchev S, Barreneche T, et al. Multi-year analyses on three populations reveal the first stable QTLs for tolerance to rain-induced fruit cracking in sweet cherry (*Prunus avium* L). *Hortic Res.* 2021;8(1):136. 10.1038/s41438-021-00571-6.
61. Branchereau C, Quero-García J, Zaracho-Echagüe NH, Lambelin L, Fouché M, Wenden B, et al. New insights into flowering date in *Prunus*: fine mapping of a major QTL in sweet cherry. *Hortic Res.* 2022;9:uhac042. 10.1093/hr/uhac042.
62. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ.* 2018;6:e4958. 10.7717/peerj.4958.
63. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38(10):4647–54. 10.1093/molbev/msab199.
64. Roux-Cuvelier M, Grisoni M, Bellec A, et al. Conservation of horticultural genetic resources in France. *Chron Hort.* 2021;61:21–36.
65. Fadón E, Herrero M, Rodrigo J. Flower development in sweet cherry framed in the BBCH scale. *Sci Hortic.* 2015;192:141–47. 10.1016/j.scienta.2015.05.027.

66. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 2015;67:1–48. 10.18637/jss.v067.i01.
67. Jonkheer EM, van Workum DM, Sheikhezadeh Anari S, Brankovics B, de Haan JR, Berke L, et al. PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics.* 2022;38(18):4403–5. 10.1093/bioinformatics/btac506.
68. Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics.* 2017;33(17):2759–61. 10.1093/bioinformatics/btx304.
69. Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17:132. 10.1186/s13059-016-0997-x.
70. Garrison E, Guarracino A, Heumos S, et al. Building pangenome graphs. *Nat Methods.* 2024;21:2008–12. 10.1038/s41592-024-02430-3.
71. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Cactus: algorithms for genome multiple sequence alignment. *Genome Res.* 2011;21(9):1512–28. 10.1101/gr.123356.111.
72. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature.* 2020;587(7833):246–51. 10.1038/s41586-020-2871-y.
73. Blanchard I, Bui QT, Mergez A, Denni S, Cornille A, Dufau I, et al. Phylogeny-driven pangenome analysis uncovers the genomic landscape of domesticated and wild *Armeniaca* species. *Hortic Res.* 2026. 10.1093/hr/uhag104.
74. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021;10(2):giab008. 10.1093/gigascience/giab008.
75. Parmigiani L, Garrison E, Stoye J, Marschall T, Doerr D. Panacus: fast and exact pangenome growth and core size estimation. *Bioinformatics.* 2024;40(12):btae720. 10.1093/bioinformatics/btae720.
76. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol.* 2018;36(9):875–79. 10.1038/nbt.4227.
77. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–100. 10.1093/bioinformatics/bty191.
78. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28(18):i333–9. 10.1093/bioinformatics/bts378.
79. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, et al. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics Proteom Bioinf.* 2021;19(4):619–28. 10.1016/j.gpb.2020.10.007.

Figures

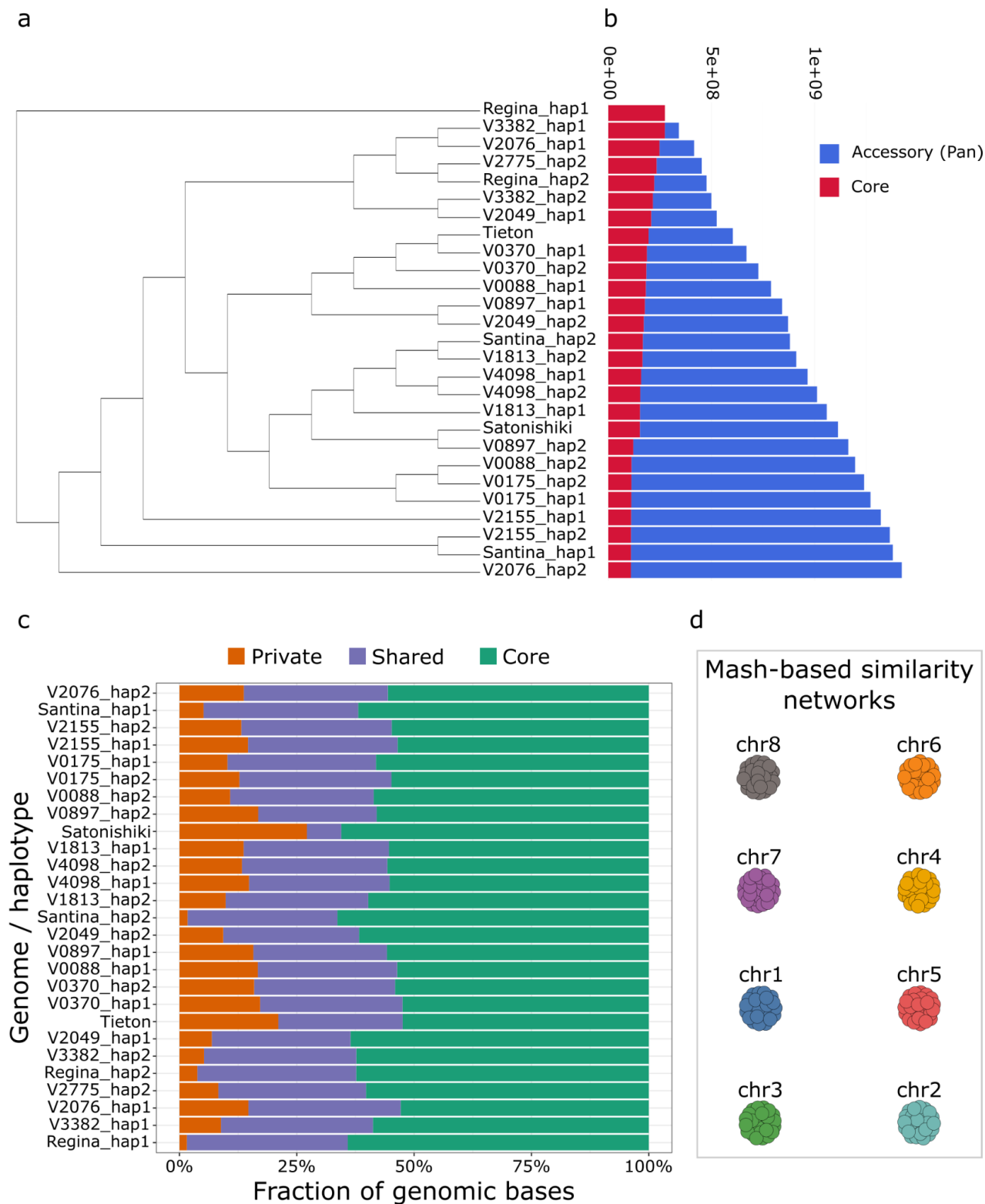


Figure 1

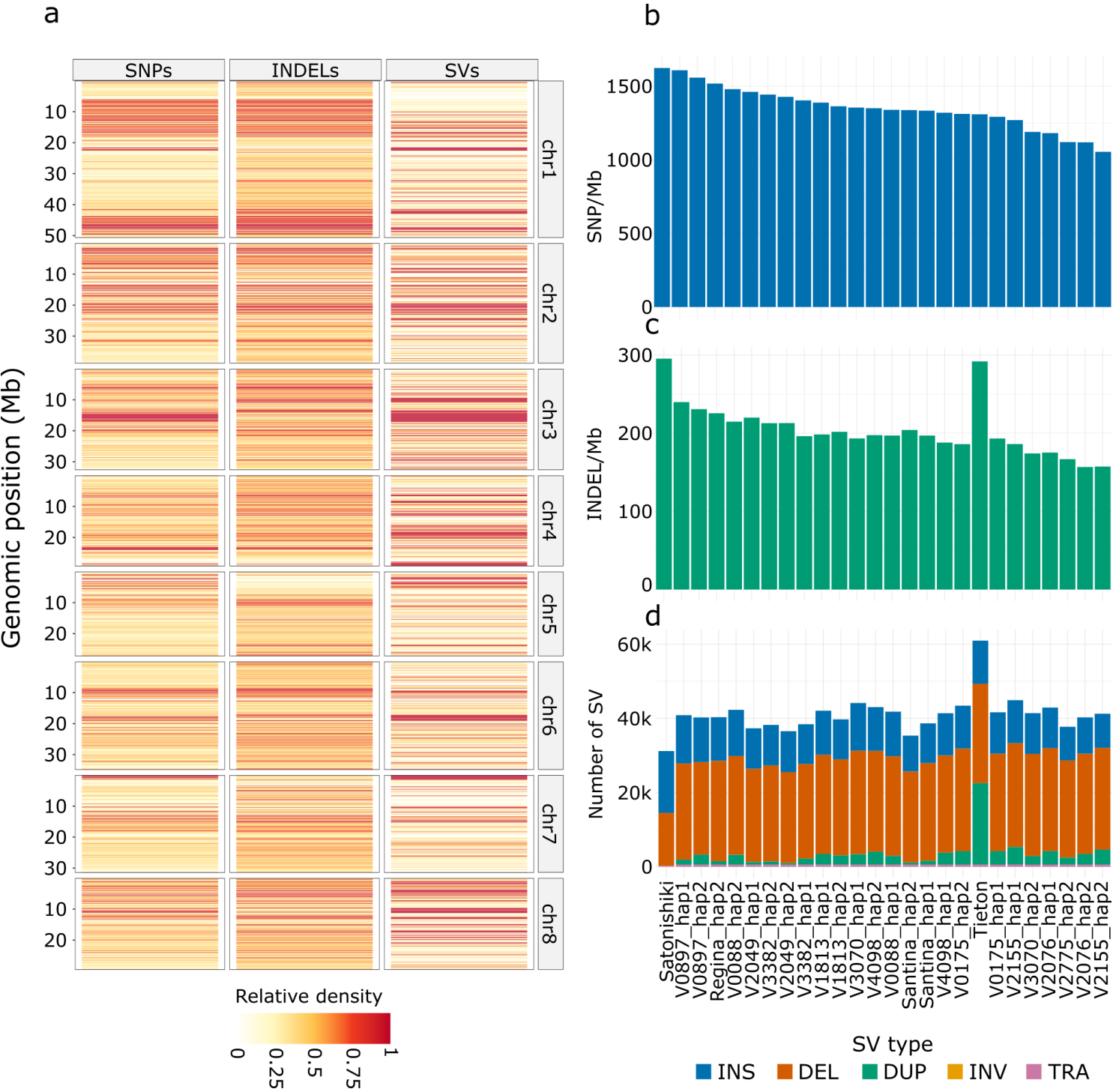
Genomic relationships and pangenome architecture across 27 sweet cherry genomes.

(a) Alignment-free phylogeny of 27 *Prunus avium* genome assemblies representing 15 sweet cherry accessions and haplotypes. Distances were computed from genome-wide k-mer profiles using Mash. The dataset includes 21 haplotypes from 11 newly generated diploid assemblies (one haplotype excluded due to insufficient assembly quality), phased haplotypes from the Regina and Santina genomes, and consensus assemblies from previously published genomes (Tieton and Satonishiki). The resulting tree resolved four major genetic groups and was used as the guide tree for progressive chromosome-wise pangenome graph construction with MiniCactus.

(b) Pangenome growth curves showing the accumulation of core and accessory genomic sequence as assemblies are progressively incorporated (Panacus ordered-hist-growth analysis). Core genome size approaches a plateau with increasing sampling, whereas accessory sequence continues to increase, indicating an open pangenome.

(c) Fraction of genomic bases classified as core, shared, or private for each genome or haplotype based on node-level presence-absence across the pangenome graph. Core regions correspond to sequences present in $\geq 90\%$ of genomes, shared regions to sequences present in multiple but not all genomes, and private regions to sequences unique to a single genome or haplotype.

(d)Chromosome-wise Mash similarity networks for each of the eight *P. avium* chromosomes. Nodes represent chromosome assemblies from individual genomes or haplotypes, and edges connect pairs with Mash distance ≤ 0.05 . Chromosomes cluster by genome of origin, consistent with strong chromosome-scale similarity among assemblies and the absence of cross-sample contamination.



Genome-wide distribution of sequence and structural variation in the sweet cherry pangenome.

- (a) Genome-wide density of SNPs, indels and pangenome structural variants (SVs) across the Regina hap1 reference genome calculated in 100 kb windows. Each row corresponds to a chromosome and each column to a variant class. Colors represent relative density, where darker colors indicate genomic windows with higher variant density relative to the upper quantile of the distribution for each variant type. SVs correspond to structural variants identified from the pangenome graph and projected onto Regina reference coordinates.
- (b) SNP density per megabase across the genomes included in the pangenome.
- (c) Indel density per megabase across the genomes included in the pangenome.
- (d) Number of structural variants per genome and haplotype, classified by SV type (INS, DEL, DUP, INV and TRA).

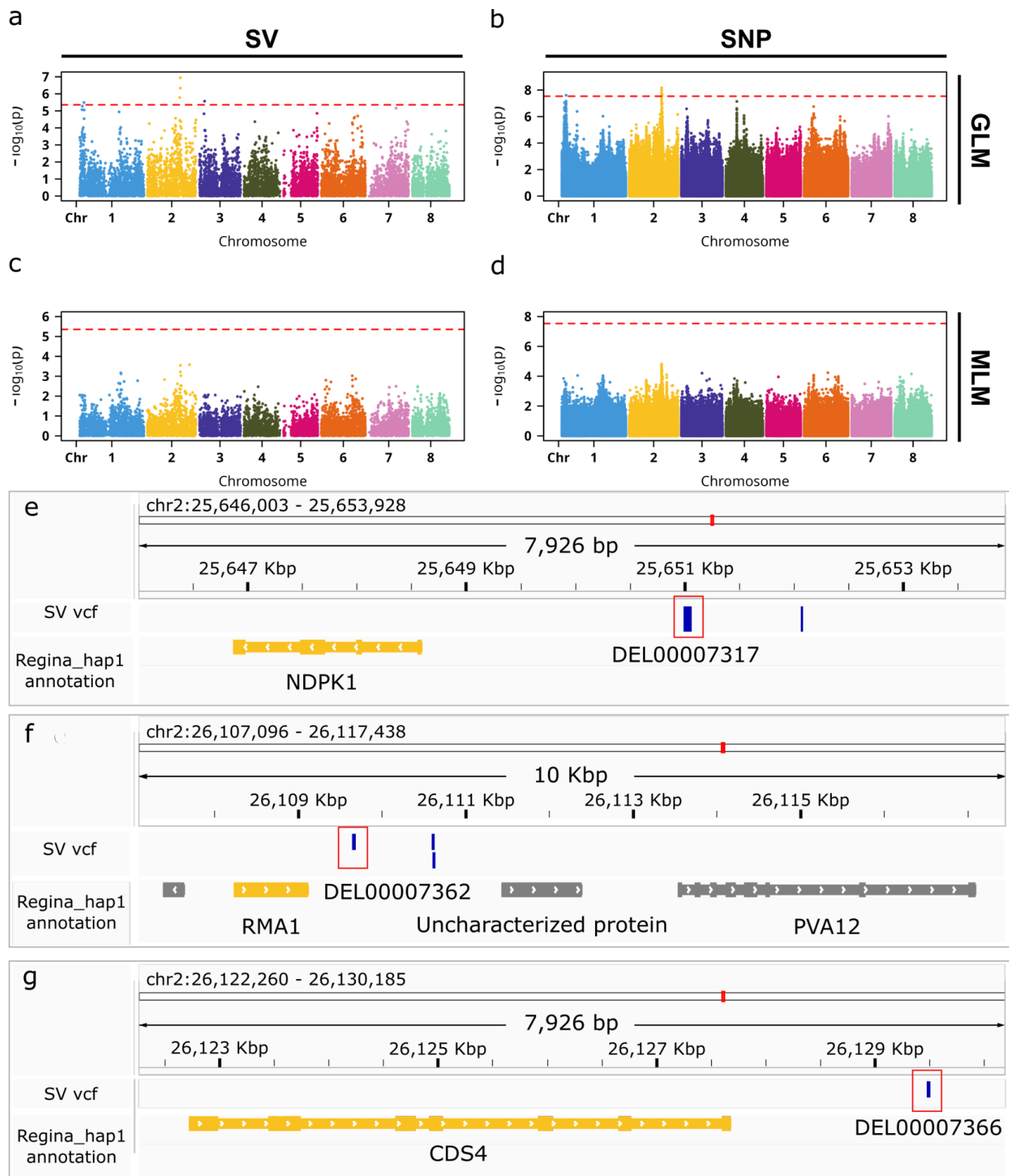


Figure 3

Genome-wide association analysis identifies loci associated with fruit weight on chromosome 2.

(a–b) Manhattan plots showing genome-wide associations for fruit weight using structural variants (SVs) **(a)** and single nucleotide polymorphisms (SNPs). **(b)** under a general linear model (GLM). The red dashed line indicates the genome-wide significance threshold.

(c–d) Manhattan plots obtained using a mixed linear model (MLM) accounting for population structure and relatedness.

(e–g) Genomic context of SVs associated with fruit weight on chromosome 2 relative to gene annotations from the *Regina* haplotype 1 assembly. Red boxes indicate associated SVs.

(e) SV located downstream of **NDPK1**. (f) Deletion **DEL00007362** located ~0.5 kb downstream of **RMA1**, with nearby genes including an uncharacterized protein and **PVA12**. (g) SV located near **CDS4**. Arrows indicate gene orientation.

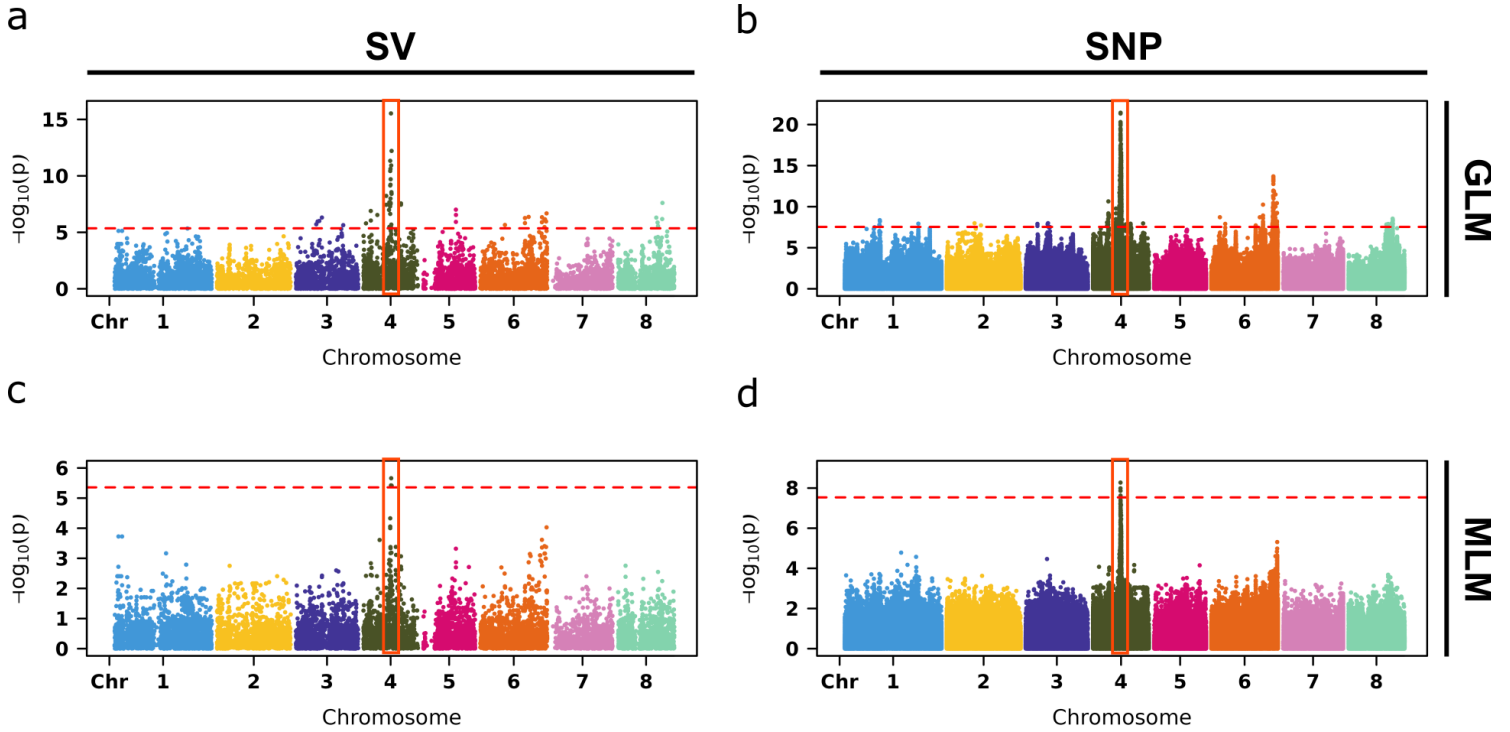


Figure 4

GWAS reveals shared chromosome 4 locus for fruit maturity date.

Manhattan plots showing associations detected using **GLM (a,b)** and **MLM (c,d)** models. Panels **(a,c)** correspond to population **structural variants (SVs)**, whereas **(b,d)** show **SNP associations**. A consistent association peak in the **central region of chromosome 4** (orange box) is detected in both SNP and SV datasets.

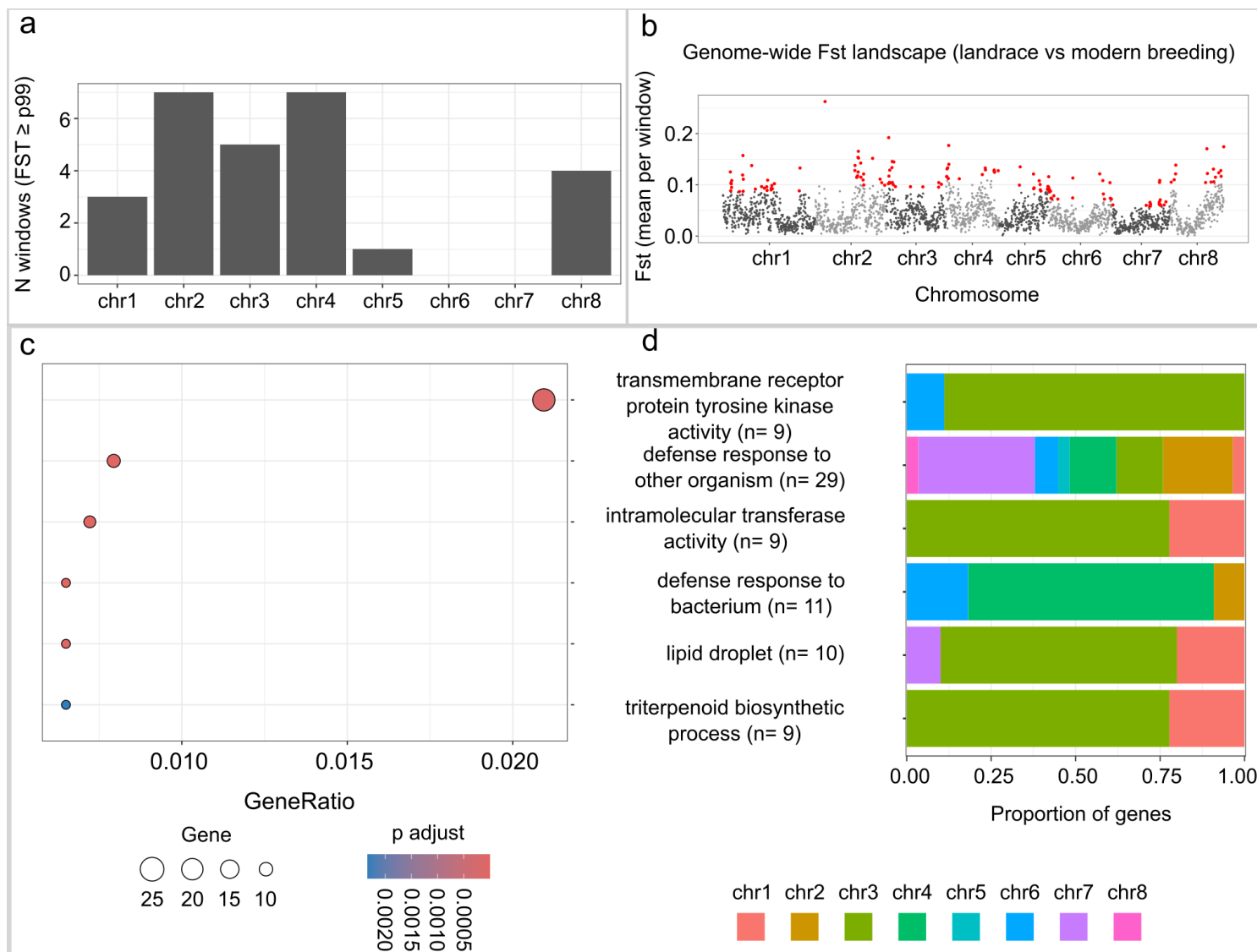


Figure 5

Differential genomic architecture between landrace and modern sweet cherry accessions.

- (a) Number of genomic windows with extreme genetic differentiation ($F_{ST} \geq 99$ th percentile) across chromosomes between landrace and modern breeding groups.
- (b) Genome-wide distribution of mean F_{ST} values calculated in **100 kb windows** between landrace and modern breeding groups. Red points indicate windows exceeding the **99th percentile threshold**.
- (c) **Gene Ontology (GO) enrichment analysis** of genes located within high- F_{ST} windows. Bubble size indicates the number of genes associated with each GO term, and color represents the **adjusted p-value**.
- (d) Chromosomal contribution to enriched GO terms, showing the **proportion of genes per chromosome** associated with each GO category.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.xlsx](#)
- [Additionalfile2.pdf](#)