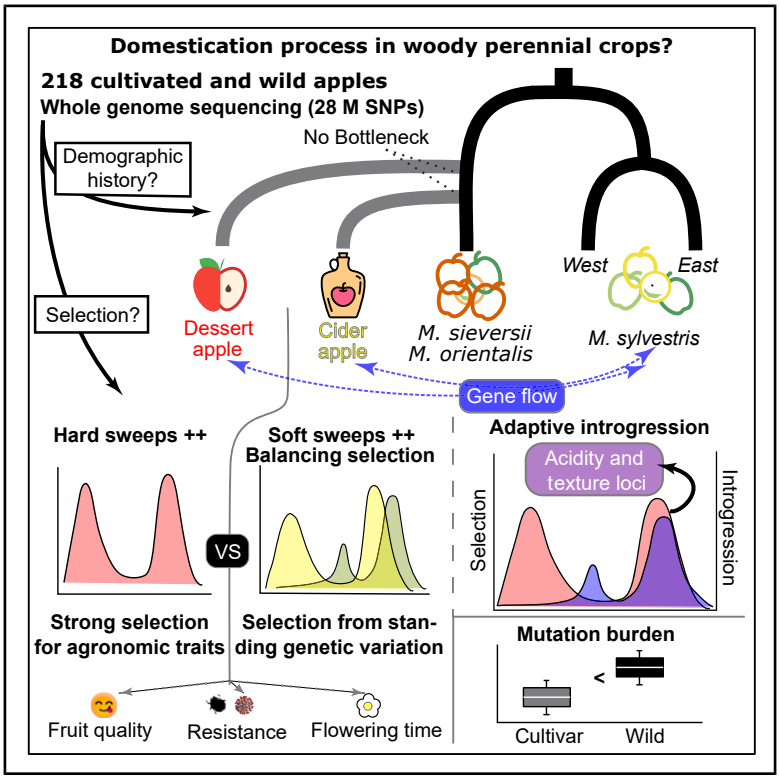
Article

Current Biology

Gene flow from the European wild apple and selection shaped the domesticated apple genome

Graphical abstract



Authors

Xilong Chen, Ronan Dadole, Komlan Avia, ..., Arnaud Lemarquand, Thierry Hance, Amandine Cornille

Correspondence

amandine.cornille@nyu.edu

In brief

Chen et al. examine how gene flow from the European wild apple and selection shape the genome of cultivated apples. Integrating 218 genomes, they reveal distinct gene pools in dessert and cider apples, identify selection signatures, and uncover introgressive patterns from wild relatives, providing insights into the domestication of perennial crops.

Highlights

- Dessert and cider apples form distinct gene pools with gene flow from wild apples
- Dessert apples and cider apples differ in selection modes
- Adaptive introgression from European wild apples shapes the cultivated apple genome
- Cultivated apples have a lower burden of harmful mutations than their wild relatives

Chen et al., 2026, Current Biology *36*, 1–15
April 20, 2026 © 2026 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cub.2026.03.048

CellPress

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

Current Biology

CellPress
OPEN ACCESS

Article

Gene flow from the European wild apple and selection shaped the domesticated apple genome

Xilong Chen,[1,2] Ronan Dadole,[1] Komlan Avia,[3] Anthony Venon,[1,2] Mathieu Brisson,[1] Carine Remoué,[1] Dong Zhang,[4] Ivan Gabrielyan,[5] Anush Nersesyan,[6] Anamaria Roman,[7] Tudor Ursu,[7] Ammar Alhmedi,[8] Dany Bylemans,[8] Tim Beliën,[8] Agnès Rousselet,[1] Martine Le Guilloux,[1] Enrique Dapena,[9] Charles-Eric Durel,[10] Thomas Kirisits,[11] Gayle Volk,[12] Frédérique Didelot,[13] Arnaud Lemarquand,[13] Thierry Hance,[14] and Amandine Cornille[1,2,15,*]

[1]Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE - Le Moulon, 91190 Gif-sur-Yvette, France
[2]Division of Science, New York University Abu Dhabi, Saadiyat Island, PO Box 129188, Abu Dhabi, United Arab Emirates
[3]INRAE, Université de Strasbourg, UMR SVQV, 68000 Colmar, France
[4]College of Horticulture, Yangling Subsidiary Center Project of the National Apple Improvement Center, Northwest A&F University, Yangling, 712100 Shannxi, China
[5]Department of Palaeobotany, A. Takhtajan Institute of Botany, Armenian National Academy of Sciences, Acharyan Street 1, 0063 Yerevan, Armenia
[6]Department of Conservation of Genetic Resources of Armenian Flora, A. Takhtajan Institute of Botany, Armenian National Academy of Sciences, Acharyan Street 1, 0063 Yerevan, Armenia
[7]Institute of Biological Research Cluj-Napoca, Branch of NIRDBS, 48 Republicii Street, 400015 Cluj-Napoca, Romania
[8]Department of Zoology, Fruit Research Center (pcfruit npo), 3800 Sint-Truiden, Belgium
[9]Servicio Regional de Investigación y Desarrollo Agroalimentario, Ctra. AS-267, PK 19, Villaviciosa, 33300 Asturias, Spain
[10]Univ Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, 49000 Angers, France
[11]Institute of Forest Entomology, Forest Pathology and Forest Protection (IFFF), Department of Ecosystem Management, Climate and Biodiversity, BOKU University, Peter-Jordan-Straße 82 (Franz Schwackhöfer-Haus), A-1190 Vienna, Austria
[12]USDA-ARS National Laboratory for Genetic Resources Preservation, 1111 S. Mason Street, Fort Collins, CO 80526, USA
[13]Unité expérimentale Horticole, INRAE, 49000 Angers, France
[14]Earth and Life Institute, Ecology and Biodiversity, Université catholique de Louvain, croix-du-Sud 4-5, 1348 Louvain-la-Neuve, Belgium
[15]Lead contact
*Correspondence: amandine.cornille@nyu.edu
https://doi.org/10.1016/j.cub.2026.03.048

SUMMARY

How selection and demography jointly shape the genomes of perennial crops remains an open question. Apple (*M. domestica*) is a compelling system because its domestication involved multiple wild progenitors and post-domestication admixture across Eurasia. We integrate 218 whole genomes (68 cultivated dessert/cider; 150 wild: *M. sieversii*, *M. orientalis*, *M. sylvestris*), RNA sequencing (RNA-seq), and a genome-wide association study of flowering time to resolve how these forces shaped the cultivated apple genome. Despite weak neutral differentiation and widespread admixture, dessert and cider apples form distinct gene pools that derive primarily from *M. sieversii*-*M. orientalis* rather than European *M. sylvestris*. We find no evidence of a domestication bottleneck, as expected in perennials. Demographic-aware selection scans reveal non-overlapping targets supported by RNA-seq; dessert shows more hard sweeps at fruit quality, disease resistance, and flowering genes, whereas cider shows proportionally more soft sweeps and balancing selection. Wild-to-crop introgression from *M. sylvestris* is extensive but heterogeneous; some introgressed tracts concentrate in hard-sweep regions and approach fixation (consistent with rapid, targeted uptake), whereas others persist at intermediate frequencies with soft-sweep signatures (consistent with diffuse, recurrent introgression of adaptive alleles). The lead chromosome 9 flowering-time association lies within an introgressed segment near a transposable element and is separated from sweep peaks, consistent with regulatory/polygenic control. Cultivated apples carry a lower deleterious load than wild relatives. These results provide a comprehensive genomic portrait of perennial fruit tree domestication, clarifying how selection and adaptive introgression shaped the cultivated apple genome and yielding actionable targets for breeding and conservation.

INTRODUCTION

Domestication provides an experimental setting in evolution, revealing how selection and demography (including gene flow) shape genomes over a relatively short evolutionary timescale. Since Darwin first recognized domestication as a cornerstone model for understanding the origins of biodiversity,[1] it has become clear that plant domestication involves both artificial selection on favorable traits and complex demographic events, including population bottlenecks and gene flow.[2] Perennial fruit

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

trees, however, differ fundamentally from annual crops in their domestication dynamics. Their long generation times, high outcrossing rates, and frequent reliance on clonal propagation lead to milder domestication bottlenecks, retention of gene flow with wild relatives, and fewer generations since domestication onset, resulting in unique genomic outcomes.[3–5]

Despite growing genomic insights into perennials, significant knowledge gaps remain. The extent and dynamics of wild-to-crop gene flow, the nature of selection (positive vs. balancing[4]) acting on domesticated genomes, and the accumulation of deleterious mutations—the so-called cost of domestication—are still poorly understood in perennial fruit trees.[3] While annuals often exhibit elevated genetic load post-domestication, the impact on long-lived, clonally propagated crops is less clear. For example, studies on grapevines suggest that somatic mutations may contribute significantly, but broader patterns across perennials remain elusive.[6,7]

The cultivated apple (*M. domestica* Borkh.) is a powerful model for studying the domestication of perennial plants. Its extraordinary phenotypic diversity—ranging in fruit color, texture, and flavor—has led to its classification into dessert and cider cultivars,[8,9] with sometimes overlapping phenotypic characteristics.[10] Genomic analyses reveal a continuum between dessert and cider apples, with two primary gene pools that are admixed yet genetically structured.[11,12] Previous studies have traced the origin of cultivated apples primarily to *M. sieversii* in Central Asia, with additional contributions from *M. orientalis* in the Caucasus and the European crabapple, *M. sylvestris*, through wild-to-crop introgression.[13] Thus, the domestication history of cider and dessert apples, as well as wild-to-crop gene flow, is relatively well understood.[11,14] However, the genomic architecture and evolutionary consequences of this admixture remain poorly resolved, leaving key questions about the relative roles of selection, introgression, and demographic history in shaping apple domestication.[14–17]

Here, we aim to disentangle the evolutionary processes that shaped the cultivated apple genome, asking how selection, introgression, and demographic history jointly influenced the process of perennial domestication. To address these questions, we analyzed a genome-wide dataset of 218 apple accessions (including 140 newly sequenced), including dessert and cider cultivars as well as their closest wild relatives (*M. sieversii*, *M. orientalis*, and *M. sylvestris*). Unlike previous studies that focused primarily on Central Asian *M. sieversii*, our work is distinguished by the inclusion of an unprecedentedly comprehensive collection of European *M. sylvestris* sampled across its full geographic range. This is crucial because *M. sylvestris* has been a major contributor to recent wild-to-crop introgressions, making this dataset ideal for evaluating the adaptive role of introgression in the domestication of this perennial species. Specifically, we sought to (1) characterize the population structure and origin of apple genetic diversity; (2) identify genomic signatures of selection—both positive and balancing—associated with domestication; (3) investigate the extent, sources, and adaptive roles of wild-to-crop introgression; and (4) evaluate the patterns and consequences of deleterious mutation load. To anchor these inferences at the trait and expression levels, we tested flowering-time differentiation by using a genome-wide association study (GWAS) and profiled the expression of selection candidates with RNA sequencing (RNA-seq), linking population genomic signals to phenotypic and transcriptional divergence. By integrating population genomics, selection scans, ancestry inference, gene expression analyses, and GWAS, this study provides a comprehensive view of the evolutionary forces—wild introgression, selection, and demography—that forged the cultivated apple genome, offering new insights into the evolutionary dynamics of perennial crop domestication.

RESULTS

Cider and dessert apples formed distinct gene pools, admixed with wild relatives

We re-sequenced or retrieved whole-genome short-read data from 218 apple accessions, including 45 dessert and 23 cider cultivars and 150 wild individuals, i.e., 87 *M. sylvestris* (Europe), 40 *M. sieversii* (Kazakhstan and China), 10 *M. orientalis* (the Caucasus), and 13 *M. baccata* (China). After quality control (removing clones and individuals with SNP missing rate >30%), 201 accessions were retained in which we identified 28,377,551 SNPs (Figure S1; Data S1A and S1B).[18,19]

Population structure and admixture analyses based on 31,300 unlinked synonymous SNPs revealed six major genetic clusters: dessert apples (orange), cider apples (pink), and four wild groups corresponding to their geographic origins (Figures 1A and 1B). *Malus sieversii* and *M. orientalis* were not separated at any *K* value (Figure S1). Neighbor-net analysis and principal-component analysis (PCA) confirmed that *M. baccata* was highly differentiated from the other species (Figures 1C and 1D) and therefore used as the outgroup. Western and Eastern European *M. sylvestris* were genetically distinct from *M. sieversii*-*M. orientalis*, and they formed two clearly differentiated genetic clusters corresponding to Western and Eastern European individuals, as supported by PCA and fastSTRUCTURE analysis (Figures 1A and 1D). Both cider and dessert apples were closer to *M. sieversii*-*M. orientalis*, rather than *M. sylvestris*, as indicated in a previous study.[20] In addition, the cider apple gene pool (pink) was substantially admixed with the Western *M. sylvestris* gene pool (blue, mean membership coefficients = 3.58%) but not with the Eastern *M. sylvestris* gene pool. Dessert apples were also predominantly admixed with the Western *M. sylvestris* gene pool (2.76%) but with a small but measurable admixture with the Eastern *M. sylvestris* gene pool (0.57%) (Data S1C).

FastSTRUCTURE membership coefficients[21] were used to assign genotypes to a given population (i.e., group of individuals including genotypes with membership coefficient >0.8 to any cluster; Data S4B). Six populations were defined: DomC (cider apples), DomD (dessert apples), SiOr (*M. sieversii* + *M. orientalis*), SylW (Western *M. sylvestris*), SylE (Eastern *M. sylvestris*), and Bacc (*M. baccata*). Austrian *M. sylvestris* samples showed admixed ancestry between the Western and Eastern European clusters and were excluded from population-based selection analyses.

These results confirmed that cider and dessert apples represent genetically distinct yet admixed gene pools, both of which are more closely related to Central Asian and Caucasian wild apples than to European *M. sylvestris*. Admixture patterns also

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

CellPress
OPEN ACCESS

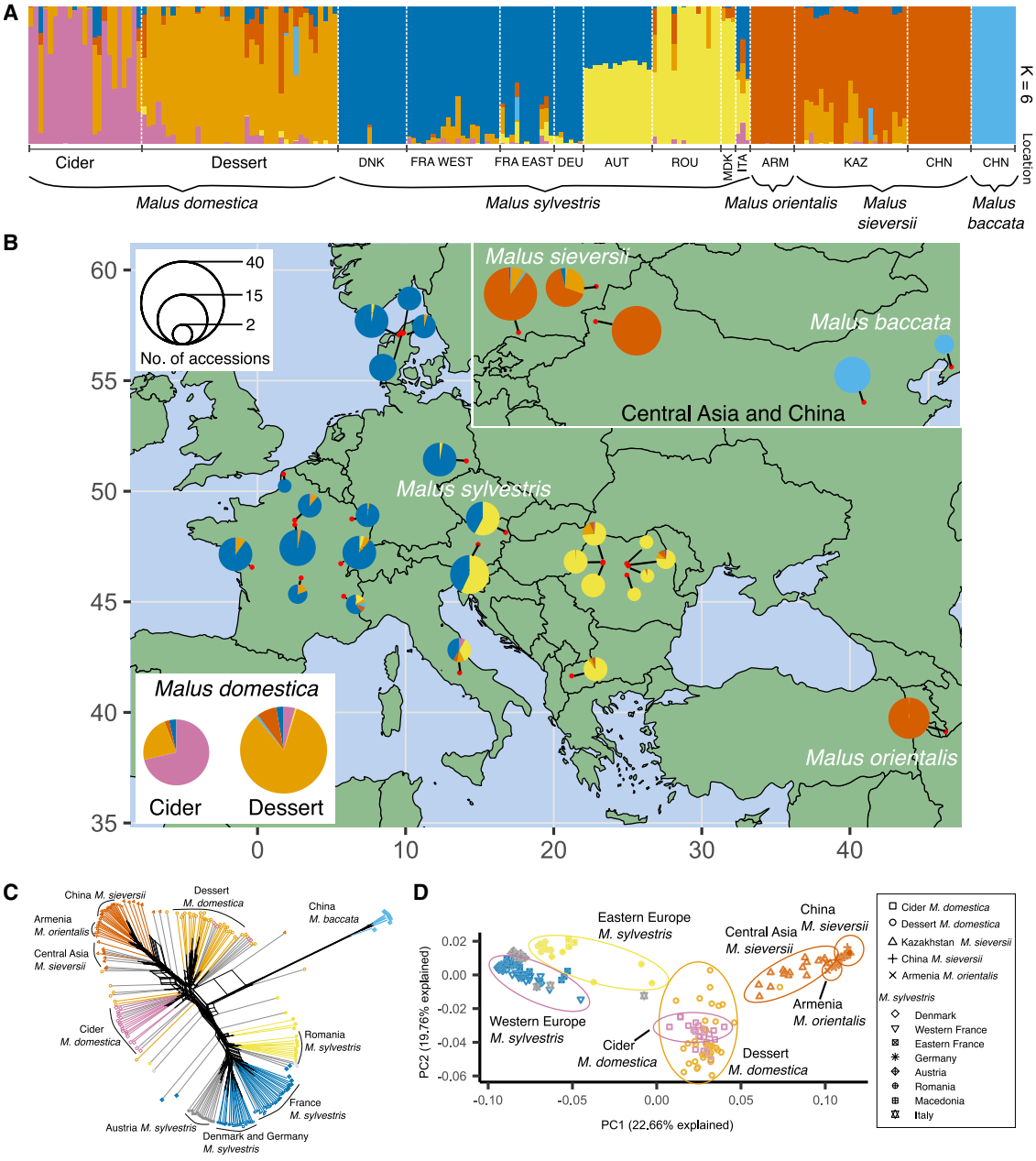


Figure 1. Population genetic structure and variation between cultivated and wild apples in Eurasia

(A) Population structure of cultivated and wild apple accessions inferred by fastSTRUCTURE with $K = 6$ ($n = 201$ accessions). Each individual is represented by a vertical bar, partitioned into K segments corresponding to the proportion of genetic ancestry from the K clusters. Wild accessions are sorted geographically from west to east in the plot, with the countries of origin listed: DNK, Denmark; FRA WEST, Western France; FRA EAST, Eastern France; DEU, Germany; AUT, Austria; ROU, Romania; MKD, Macedonia; ITA, Italy; ARM, Armenia; KAZ, Kazakhstan; and CHN, China. Distinct genetic clusters comprise cider apples (pink), dessert apples (orange), Western *M. sylvestris* (blue), Eastern *M. sylvestris* (yellow), *M. sieversii* and *M. orientalis* (red), and *M. baccata* (green).

(B) Map showing the average proportions of fastSTRUCTURE membership coefficients at $K = 6$ for all wild individuals from each sampling location (Data S1) and for cider or dessert varieties (lower left corner, outside the map, for cultivated apples). The size of each pie chart is proportional to the sample size.

(C) Neighbor-net tree depicting relationships among wild and cultivated apple accessions and their associated populations, as identified by fastSTRUCTURE with $K = 6$. Colors correspond to the genetic groups inferred with $K = 6$, and admixed individuals (membership coefficient < 80% for any given gene pool) are shown in gray.

(D) PCA illustrating the genetic variation among wild and cultivated apple accessions. The total variance explained by each of the first two components is indicated along the respective axis. All analyses used the 31,300 unlinked, synonymous SNPs across 201 apple genotypes.

See also Figure S1 and Data S1A and S1B.

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

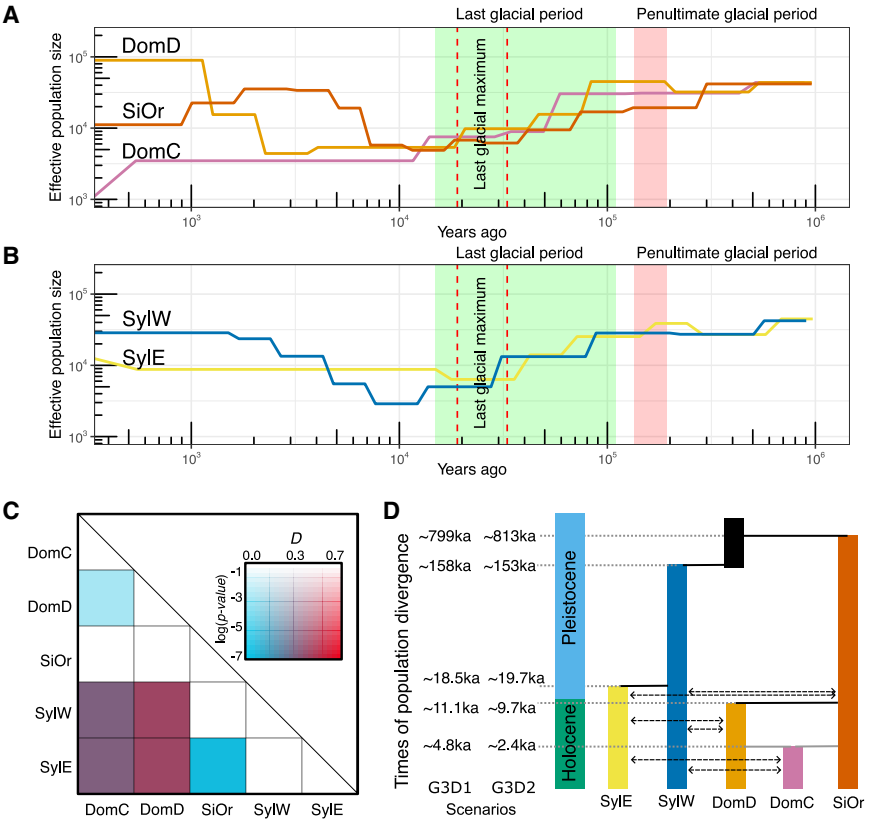


Figure 2. Demographic and divergence histories of wild and cultivated apple populations

(A) Changes in effective population size of cider apples (DomC), dessert apples (DomD), and *M. sieversii* (SiOr) over time, as inferred by SMC++.

(B) Changes in the effective population size of Western (SylW) and Eastern (SylE) European wild apples over time, as inferred by SMC++. Major climatic events are shown: the penultimate glacial period (red), the last glacial period (green), and the last glacial maximum (dashed vertical lines).

(C) Patterson's *D*-statistic (ABBA-BABA test) was estimated between five apple populations. DomC, cider apples; DomD, dessert apples; SiOr, *M. sieversii* and *M. orientalis*; SylE, Eastern European *M. sylvestris*; SylW, Western European *M. sylvestris*. Bacc (*M. baccata*) was used as the outgroup.

(D) Schematic graph of the apple demographic history. Demographic parameters were estimated by fastsimcoal2. Solid lines represent divergence events (gray solid lines indicate the two alternative DomC divergence events in the statistically indistinguishable scenarios), while dashed lines with arrows represent gene flow. The left side indicates the timing of divergence events (ka, thousands of years ago) for G3D1 and G3D2 scenarios (G3 includes gene flow between the ancestors of DomC/DomD and SiOr, gene flow between the ancestors of SylW/SylE and SiOr, and gene flow between DomC/DomD and SylW/SylE populations. Divergence scenarios include D1, where DomC and DomD both diverge from SiOr, with DomC diverging more recently and SylE diverging from SylW; D2, where DomD diverges from SiOr, followed by DomC diverging from DomD, and SylE diverging from SylW.).

See also Figures S2 and S3 and Data S1D–S1F.

highlighted the distinct and asymmetric contributions of the two *M. sylvestris* gene pools (Eastern and Western) to the cider and dessert cultivated apple gene pools.

Divergence and demographic histories of wild and cultivated apples

Genetic differentiation estimates confirmed that Bacc was the most divergent (Figure S2A; Data S1D). Using Bacc as an outgroup, the species-tree topology and coalescent-based ASTRAL phylogeny tree (Figures S2B and S2C) showed that SylW and SylE formed a clade, but this could not resolve whether DomC was closer to SiOr or DomD. Repeated ASTRAL analyses, using the same number of individuals per population, consistently recovered two alternative species-tree topologies (Figure S2D). SylE, SylW, SiOr, DomC, and DomD displayed weak genetic differentiation, further consistent with a recent divergence and/or gene flow (Figure S2E; Data S1E).

Cultivated apples (DomC and DomD) exhibited higher genetic diversity than the wild populations SiOr and SylW (Data S1D), indicating a lack of bottleneck during domestication. Demographic inference revealed a gradual decline in effective population size (N_e) from the penultimate glaciation to the mid-Holocene (~10,000 years ago), followed by divergent trends (Figures 2A, 2B, and S2F). SiOr and SylW expanded post-glacially, whereas SylE maintained a relatively stable N_e (Figure 2B). Among cultivated apples, DomD showed a demographic expansion ~2,500 years ago, whereas DomC showed a contraction (Figure 2A).

Post-glacial climate likely shaped the demography of wild apples; however, we caution that the long-term declines inferred from Markovian coalescent models may also arise from historical population structure rather than from current reductions in N_e.[22] The expansion of DomD reflects its historical spread. DomD shows a demographic expansion consistent with its documented historical spread. By contrast, DomC displays a recent decrease in estimated N_e, which could reflect genetic homogenization associated with modern breeding and/or the dominance of a limited number of elite cultivars in this study, rather than an actual contraction in census size. These demographic changes were taken into account in the selection analyses below.

Genome-wide introgression from European wild apples to cultivated apples

Patterson's *D*-statistics[23,24] and f_4-ratios,[25] using Bacc as the outgroup, revealed significant introgression from SylW into both DomD and DomC, with weaker signals from SylE. Minor gene flow was detected between SylE and SiOr, indicating elevated gene flow from an ancestor of *M. sylvestris* into the *M. domestica* gene pool (Figures 2C and S2E; Data S1E). The

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

CellPress
OPEN ACCESS

massive "proto-gene flow" from ancestral *M. sylvestris* could also explain the observed N_e expansion above.[16,26]

Five-taxon *D*-statistics (D_{FOIL})[27] and topology weighting analyses[28] confirmed contributions from both Western and Eastern *M. sylvestris* to DomD and DomC (Data S4C and S4D). The dominant species-tree topology indicated introgression from SylW into the common ancestor of DomC and DomD. This was further supported by *f*-branch statistics[29] and the best-fit qpGraph model,[30] which inferred multiple admixture events, including one predating the DomC-DomD split (Data S3C and S4E). Among eight demographic models (Figure S8A) tested with fastsimcoal2,[31] the best-fitting scenarios (G3D1 and G3D2) incorporated gene flow from SiOr into SylW/SylE and from SylW/SylE into DomC and DomD, with stronger introgression into DomD (Figure S8B). However, the models did not allow a clear resolution of the origin of DomC (Figure 2D; Data S1F and S4F). Overall, these approaches provided broadly consistent evidence for introgression involving SylW/SylE and the domesticated groups DomC and DomD. Minor differences among methods were mainly reflected in the inferred phylogenetic placement of DomC and DomD. These discrepancies likely reflect both the short divergence times of DomC and DomD and the varying temporal sensitivities of the methods to gene flow.

DomD exhibited stronger overall introgression than DomC, as reflected by higher genome-wide f_4-ratios (DomD = 0.5011; DomC = 0.4570). Per-chromosome *D*-statistics and topology weighting revealed heterogeneous wild-to-crop introgression (Data S4G and S4H).[32,33] In DomC, chromosome 13 showed the highest SylE/SylW admixture, whereas in DomD, chromosomes 16 and 14 were most affected. Sliding-window f_d scans[34] indicated highly introgressed regions (Data S4I), which were enriched for genes linked to nitrogen metabolism, terpene biosynthesis, and steroid biosynthesis (Figure S3B). A subsequent Venn comparison of genes in these regions revealed both shared and unique loci between DomC and DomD, highlighting potential functional roles for the introgressed genes (Figure S3A).

Divergent positive and balancing selection in cider, dessert, and wild apples

Selection scans integrated multiple complementary approaches to detect different modes of selection: OmegaPlus and RAiSD for hard sweeps, G12/G123 for both hard and soft sweeps, and BetaScan2 for balancing selection. Together, these methods provide a comprehensive view of selective pressures by capturing distinct and sometimes overlapping signals across populations (Figures 3B and 3C; Table 1).

Most positive selection genes were population specific; however, 27 were shared between DomC and DomD, including genes linked to fruit quality, disease resistance, and flowering time (Figure 3A; Data S2A and S2B). We also identified 81 genes under balancing selection, 44 of which were shared among populations (Figure 3C; Data S2C). These included stress-response genes such as *MdDREB2* (Figure S4B).[35] Furthermore, 23 balancing selection genes were specifically shared between DomC/DomD and *M. sylvestris* (SylE or SylW), including *MdVIP5* (MD04G1191400), a flowering-related gene[36] (Data S2C). Cider apples (DomC) displayed more soft sweeps and balancing selection, whereas dessert apples (DomD) showed more hard sweeps (Table 1). DomC lacked the fruit size gene *fs15.1*[15,20] but had unique balancing selection targets such as *RPW8-NBS* (Figure S4C).[37] DomD contained two disease-resistance genes and eight flowering genes under hard sweeps.

Wild populations, especially SylW, had more genes under balancing selection than cultivated apples (Table 1; Data S2C). Positive selection in wild groups targeted multiple NBS-LRR genes (Figure S5).

Selection during domestication acted on both a small set of shared domestication-related traits and many population-specific targets. DomC exhibited a higher proportion of soft sweeps and balancing selection, consistent with the retention of broader genetic diversity, whereas DomD experienced stronger directional selection. To corroborate the functional relevance of selection signals, we integrated transcriptomic data from controlled conditions in which wild and cultivated genotypes were grafted onto the same rootstock and grown under identical conditions. Genes differentially expressed between cultivated (DomC or DomD) and wild apples (SiOr or SylW) were cross-referenced with genes under positive selection to identify domestication-related candidates, and 88 sweep genes were differentially expressed in cultivated apples compared with their wild relatives (Data S2A), providing orthogonal validation of the selected candidates.

Adaptive introgression from distinct donors shaped fruit-quality loci in cider and dessert

Local ancestry inference using RFMix[38] revealed enrichment for DomC and DomD in SiOr ancestry. Still, DomC carried more SylW ancestry in hard sweeps, whereas DomD carried more SylE ancestry across sweep types (Data S4J and S4K). $Q95_{SiOr,\ DomC/DomD,\ SylW/SylE\ (10\%,\ 100\%)}$ statistics (i.e., 95th percentile of derived frequencies in DomC/DomD of SNPs that have a derived allele frequency 100% in a donor panel—SylW/SylE—but where the derived allele is at a frequency smaller than 10% in SiOr)[39] showed a bimodal frequency distribution in DomC and lower peaks in DomD (Data S4N). These patterns reflect distinct adaptive introgression histories.

Integrating f_d statistics, topology weights (Data S4M), RFMix, and Q95 (Data S4N), we identified 14 adaptive introgression regions in DomC and 23 in DomD. Haplotypes were clustered using PCA (dudi.pca) followed by hierarchical clustering (hclust) using Genotype Plot[40] to assess grouping patterns and infer shared ancestry with wild donors (Data S2D and S5). Notably, DomC haplotypes clustered with multiple SylW individuals on chromosome 17 overlapping a fruit texture quantitative trait locus (QTL)[41] (Figures 4A and 4B). In contrast, DomD haplotypes clustered with one SylW individual localized on chromosome 16 and included the acidity marker *Ma*[42] (Figures 4C and 4D).

Adaptive introgression shaped key fruit-quality loci, with cider and dessert apples drawing on different donor ancestries and experiencing distinct selection dynamics.

Selection and introgression reduced deleterious mutation load in domesticated apples

No highly constrained sites with a GERP-RS score > 2[43] were detected in the apple genome (Data S4O), a threshold commonly

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

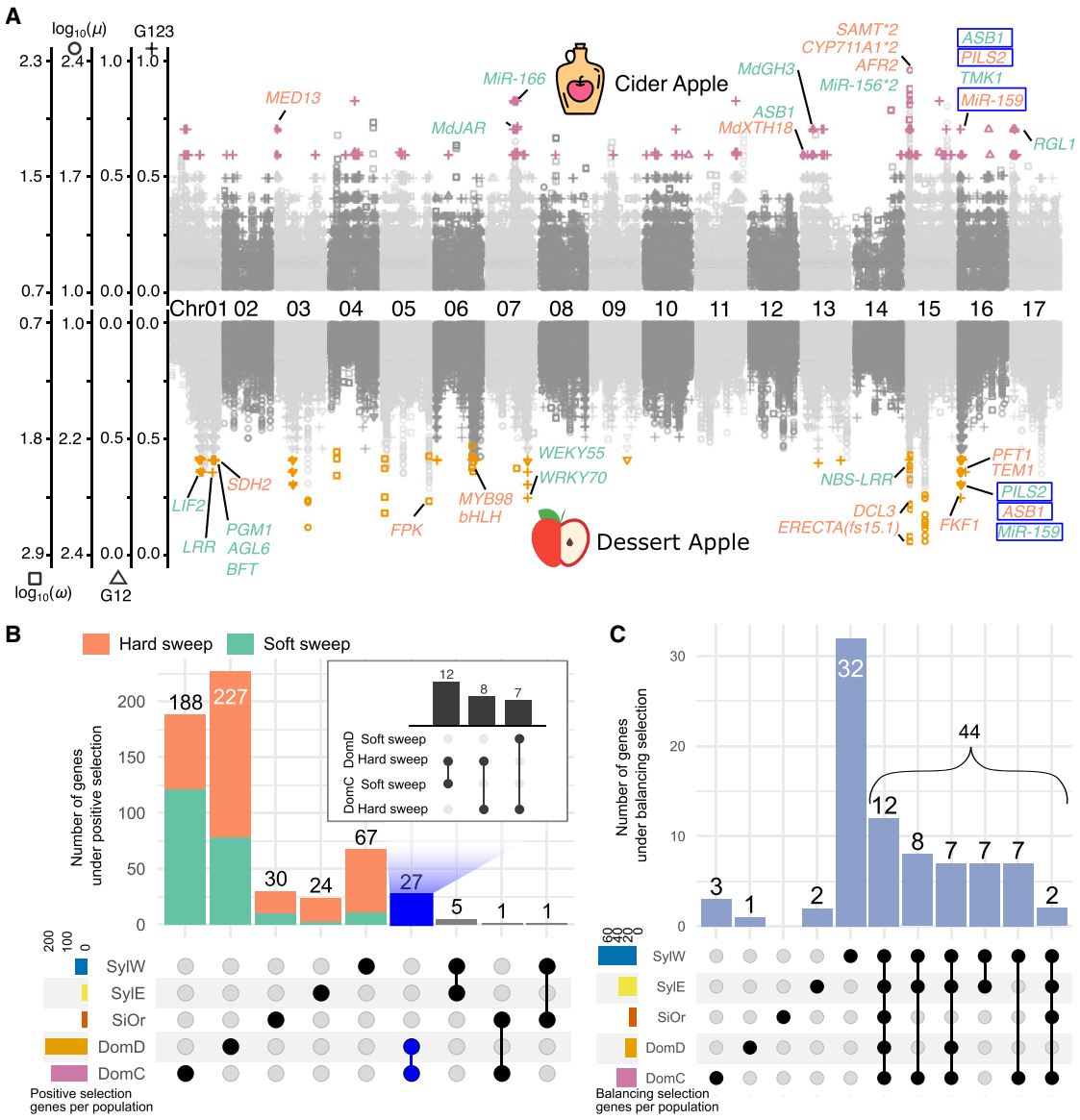


Figure 3. Positive and balancing selection of cultivated and wild apples

(A) Positive selection along the genomes of cider and dessert cultivated apples. Different symbol shapes and *y* axes represent different types of selection tests. From left to right, $\log_{10}(\omega)$ (squares), $\log_{10}(\mu)$ (circles), G12 (triangles), and G123 (crosses) values are shown. Significant selective sweep signals are represented in pink (DomC) or orange (DomD). The positive selection patterns of the gene symbols are represented in red (indicating hard sweeps) and green (indicating soft sweeps), consistent with those depicted in (B). The blue text box represents the positive selection genes shared by DomC and DomC.

(B) UpSet plot summarizing genes under positive selection (hard and soft sweeps) identified within and across the five apple populations. The horizontal bars (bottom left) show the total number of selected genes detected in each population. The matrix (bottom right) indicates which populations contribute to each intersection set, with black circles marking population membership and connecting lines showing shared sets across two or more populations; orange and green squares in the inset indicate hard and soft sweeps, respectively. The upper histogram shows the number of selected genes for each population or population intersection. The blue-highlighted bar represents the set of genes jointly under positive selection in DomC and DomD.

(C) Upset plot of genes under balancing selection across the five apple populations. The horizontal bars (bottom left) display the total number of genes under balancing selection in each population. The matrix (bottom right) shows the populations under consideration as black circles. When an intersection between two or more populations is displayed, these populations are connected by a solid black line. The bar graph (top) represents the number of genes under balancing selection in a population or the number of intersections between two or more populations. DomC, cider apples; DomD, dessert apples; SiOr, *M. sieversii* and *M. orientalis*; SylE, Eastern European *M. sylvestris*; SylW, Western European *M. sylvestris*.

See also Figures S4 and S5 and Data S2A–S2C.

used to identify deleterious mutations in constrained genomic regions.[44,45] Because GERP estimates based on closely related Rosales species can underestimate long-term constraint, mutation load was primarily estimated using SIFT 4G, a phylogeny-independent and widely validated predictor of functional deleteriousness.[46] Cultivated apples carried a lower

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

CellPress
OPEN ACCESS

Table 1. Size and gene count of regions under positive selection and balancing selection in the five cultivated and wild apple populations

	Positive selection						Balancing selection	
	Hard sweep		Soft sweep		Total			
Population	Region size (kb)	Number of genes	Region size (kb)	Number of genes	Region size (kb)	Number of genes	Region size (kb)	Number of genes
DomC	604.27	83	1,124.19	133	1,728.46	216	82.43	39
DomD	1,092.84	169	532.95	85	1,625.79	254	36.14	22
SiOr	276.31	21	74.71	11	351.02	32	21.11	14
SylE	335.73	26	17.64	3	353.37	29	62.56	36
SylW	644.75	62	123.93	12	768.68	74	139.04	75

DomC, cider *M. domestica*; DomD, dessert *M. domestica*; SiOr, *M. sieversii* and *M. orientalis*; SylE, Eastern European *M. sylvestris*; SylW, Western European *M. sylvestris*. See also Data S2A–S2C.

heterozygous and homozygous relative deleterious burden than their wild relatives (Figure 5A). The heterozygous mutation burden is higher in DomC compared with DomD. In contrast, the homozygous mutation burden shows no significant difference (Figures S6A and S6B). Most deleterious variants were heterozygous across populations (Figure 5A).

Introgression from *M. sylvestris*, especially SylE into DomD and SylW into DomC/DomD, was associated with reduced homozygous deleterious mutations (Figure 5B). Non-introgressed regions under hard and soft sweeps also had fewer deleterious mutations than controls (Figure 5C). Hard sweep introgressed genes were depleted of deleterious mutations in DomD but not DomC.

Domesticated apples—especially DomD—carry a lower deleterious load than wild relatives, likely due to the combined effects of selection and introgression.

GWAS supports phenotypic divergence between cider and dessert cultivated apple

To evaluate whether phenotype-associated variants overlap with regions under selection or introgression, we ran a flowering-time GWAS in GAPIT v3 (BLINK; K and K+Q) on best linear unbiased predictions (BLUPs) from a subset of cider and dessert.[47,48] Two SNPs surpassed false discovery rate (FDR) and Bonferroni thresholds: Chr09:24,880,275 ($p = 1.11 \times 10^{-16}$; PVE = 92.4%) and Chr08:21,795,100 ($p = 1.51 \times 10^{-13}$; PVE = 87.7%) (Figures 6A and 6B). The Chr08 signal lies ~2.6 kb from *ACC1* (MD08G1178000), implicated in very-long-chain fatty acid biosynthesis and cold response,[49] while the Chr09 SNP falls ~10 kb from an O-fucosyltransferase gene (MD09G1220000; SPINDLY-like hormone signaling[50]) in a region containing a HODOR transposable element (TE) (ms537607_Chr09_HODOR-RLC_denovoMDO_kr-B-P814.288-Map5_reversed),[18] suggesting potential regulatory effects. Genotype distributions at Chr09:24,880,275 clearly separate DomC from other groups, particularly *M. sylvestris* (Figures 6C and 6D; Data S2E). Consistent with a wild-to-crop origin, f_d for DomC-SylW at this locus exceeds the genome-wide Q3 (the upper quartile of the f_d distribution), indicating localized introgression (Figure S7). These two loci do not overlap sweep peaks, implying that flowering-time differences may stem from regulatory variation or polygenic bases not captured by classical selection scans.

DISCUSSION

Selection, introgression, and demography jointly forged the domesticated apple genome. Despite weak genome-wide differentiation at neutral sites and extensive shared admixture, cider and dessert cultivars of *M. domestica* form distinct gene pools that derive primarily from *M. sieversii-M. orientalis* rather than European *M. sylvestris*.[13,20] Notably, although these two cultivar groups are genetically differentiated, the origin of cider apples remains uncertain because alternative evolutionary scenarios cannot be clearly distinguished from one another based on current data. We find no evidence of a domestication bottleneck in *M. domestica*, as previously observed in other perennials.[3–5] Instead, demographic inference indicates a recent expansion in desert *M. domestica* and relative contraction in cider, alongside post-glacial expansions in *M. sieversii-M. orientalis* and Western *M. sylvestris*, with relative stability in Eastern *M. sylvestris*.[51,52] For *M. sylvestris*, this pattern is consistent with its post-glacial phylogeographic history; during the last glacial maximum, *M. sylvestris* was confined to multiple southern and eastern refugia (e.g., south France, the Balkans, and Carpathian regions), followed by northward and westward recolonization as climates warmed, leading to a large, expanding Western European genetic group but more demographically stable eastern populations that persisted closer to refugial areas. Such a structure reflects the interplay of historical climate change and natural gene flow in shaping population divergence across Europe.[53] We accounted for these histories when inferring selection and in GWAS to limit confounding by relatedness and structure.[47,48]

Our results revealed contrasting selection modes in a perennial crop. Cider *M. domestica* shows proportionally more soft sweeps and balancing selection—consistent with broader retained diversity. In contrast, dessert *M. domestica* exhibits more hard sweeps on fruit-quality, disease-resistance, and flowering genes.[15,54–58] The contrasting selective sweep patterns observed between dessert and cider groups are broadly consistent with domestication trajectories reported in other perennial fruit crops. For example, in grapevines, table and wine grapes show partially overlapping but distinct sets of selective sweeps, reflecting divergent selection targets related to fruit size, texture, sugar content, and post-harvest qualities in table grapes vs.

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

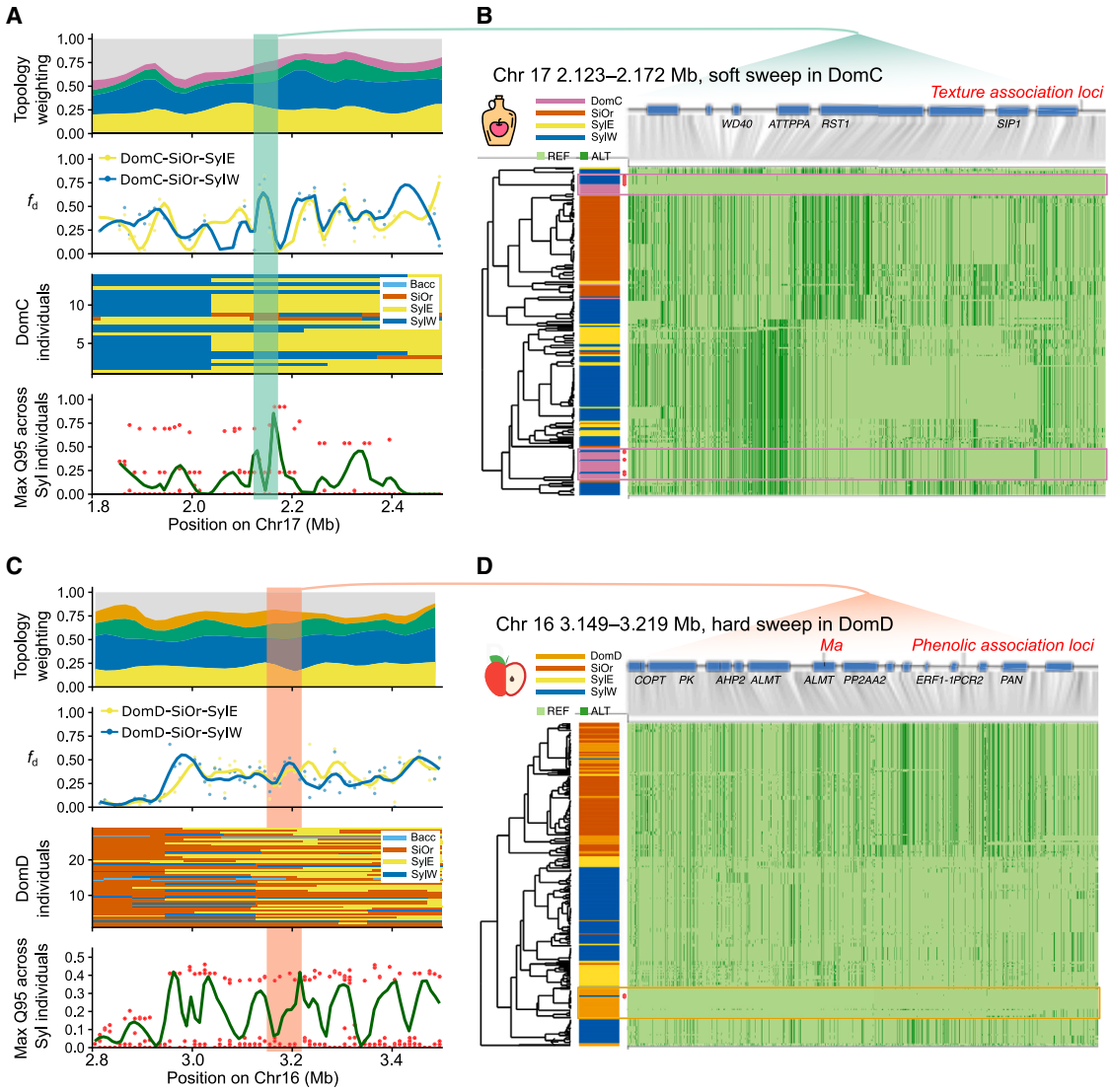


Figure 4. Adaptive introgression signals and typical haplotypes structures across positive sweep regions in cider and dessert apples

(A) Signals of introgression patterns across Chr17 (1.8–2.5 Mb) in the DomC population. From top to bottom: topology weighting, f_d, local ancestry painting, and maximum Q95 value across all Syl donor individuals. The highlighted region indicates the putative adaptive introgression region.

(B) Zoom-in of the highlighted region in (A) showing haplotype structure, spanning from 2.123 to 2.172 Mb, with a soft sweep in the cider apple genome.

(C) DomD, Chr17 (2.8–3.5 Mb). Same analyses as in (A). The highlighted region indicates the putative adaptive introgression region.

(D) Zoom-in of the highlighted region in (C) showing haplotype structure, spanning from 3.148 to 3.220 Mb, the hard-sweep region of the dessert apple genome. The different colored bars on the left represent individuals from different populations. The left panel shows a dendrogram of haplotype clustering by PCA (dudi.pca),[41] and the bar plot shows haplotypes from different populations in different colors. The dark green and light green colors indicate the reference and alternative alleles, and long continuous blocks indicate shared introgressed haplotypes under selection. The clustered selection haplotype and the donor wild individuals are highlighted by red dots. The upper part shows the genes in this region, and the gene IDs come from the annotations of the GDDH13 v1.1 genome.[18] The QTL and GWAS regions are labeled in red above the gene models. DomC, cider apple; DomD, dessert apple; SiOr, *M. sieversii* and *M. orientalis*; SylE, Eastern European *M. sylvestris*; SylW, Western European *M. sylvestris*; Bacc, *M. baccata.*

See also Data S2D and S5.

fermentation-related traits in wine grapes.[59] Although historical records suggest that cider apples were used earlier than dessert apples,[60] both recent studies[13] and our results indicate that cider and dessert apples are genetically very close. Our demographic inference further suggests that the divergence of cider apples may be more recent than that of dessert apples (Figure 2D). Following their more recent divergence, cider apples not only show a slightly lower number of positive selection genes, compared with dessert apples, but also exhibit more soft sweeps that can reach fixation relatively quickly. These observations suggest that differences in the number of detected selective sweeps may not solely reflect the temporal history of use but could also result from the recent divergence, ongoing gene flow, and the specific traits targeted in each population. Also,

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

CellPress
OPEN ACCESS

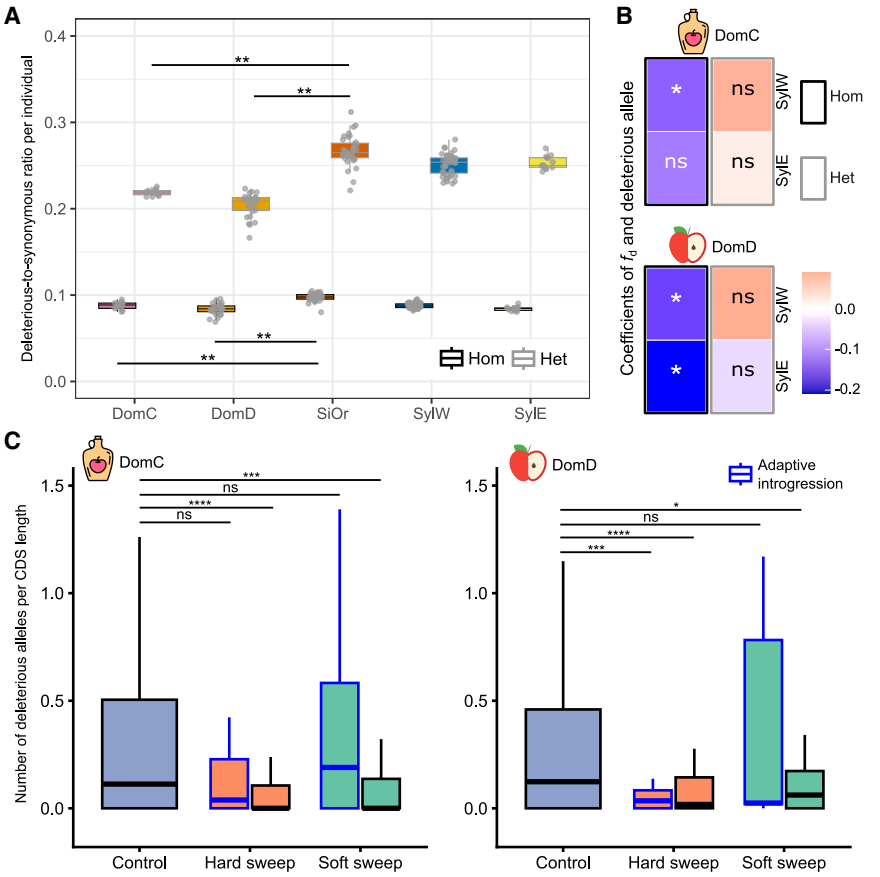


Figure 5. Distribution of mutation burden in cultivated and wild apple populations and the effect of introgression in cider and dessert apples

(A) Homozygote and heterozygote mutation burden across five apple populations, measured as the number of deleterious-to-synonymous mutations in individuals within each population.

(B) Heatmap of coefficients between f_d and the deleterious alleles. The population labels at the bottom and on the left are designated as P2 and P3 of f_d analysis.

(C) Mutation burden in the cider (DomC) and dessert (DomD) apple populations across control genes, hard-sweep genes, and soft-sweep genes. Adaptive introgression genes are indicated with a blue border. The *y* axis shows the number of deleterious alleles per 1,000 bp of coding sequence (CDS). DomC, cider apple; DomD, dessert apple; SiOr, *M. sieversii* and *M. orientalis*; SylE, Eastern European *M. sylvestris*; SylW, Western European *M. sylvestris*. (ns, $p > 0.05$; *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.) Hom, homozygous state; Het, heterozygous state. See also Figure S6.

wild apple populations, to a much larger extent than cider, carry massive balancing selection signals, as expected in natural systems.[61] Transcriptome profiling supports these inferences; a substantial subset of sweep candidates is differentially expressed between cultivated and wild apples, indicating that selection has remodeled both their sequences and their expression.

In addition to selection, introgression from wild relatives played a central role in shaping diversity and adaptation. Local ancestry reveals a clear asymmetry; hard-sweep regions in cider are enriched for Western *M. sylvestris* ancestry, whereas dessert *M. domestica* carries more Eastern *M. sylvestris* ancestry across both hard and soft sweeps. Frequency spectra of introgressed variants further indicate contrasting dynamics. In cider apples, *M. sylvestris*-derived haplotypes cluster in hard-sweep tracts and show a bimodal distribution (many near fixation and many still rare), consistent with strong selection on a few advantageous tracts; in dessert apples, introgressed alleles more often remain at intermediate frequency and co-occur with soft sweeps, a pattern consistent with diffuse, polygenic adaptation in which recombination breaks up haplotypes and limits single-tract fixation.[13,62] A representative case is a lead flowering-time SNP on chromosome 9 that is located near an O-fucosyltransferase (SPINDLY-like) gene—an effector of hormone signaling linked to flowering pathways[50]—embedded in a cider-Western *M. sylvestris* introgressed tract and adjacent to a HODOR long terminal repeat (LTR) element in the GDDH13 reference,[18] suggesting *cis*-regulatory variation of wild origin may underlie phenological divergence. Notably, these two GWAS hits do not coincide with sweep peaks, implying that flowering-time differences may arise from regulatory variation or polygenic effects not captured by classical selection scans. This highlights that GWAS can reveal loci that may be missed by sweep-based approaches, providing complementary insights into the genetic architecture of domestication-related traits.

Cultivated apples carry fewer predicted deleterious alleles than wild relatives, contrary to trends in many annual crops and grapevines,[3,6,7,63,64] but mirror observations in other perennial fruit trees such as the domesticated pear.[65] Depletion of harmful variants in hard-sweep introgressed genes in apples, together with higher burdens in soft-sweep regions, suggests that selection and wild gene flow jointly reduced genetic load while enabling adaptive change. While long-term clonal propagation could mask some deleterious alleles, the concordance between ancestry, selection mode, and load reduction argues for genuine purging in key regions. The depletion of harmful variants in hard-sweep introgressed regions, together with comparatively higher burdens in soft-sweep regions, suggests that positive selection and wild gene flow jointly shaped adaptive change while reducing genetic load. Although long-term clonal propagation could, in principle, mask recessive deleterious alleles, the consistent concordance among ancestry, sweep mode, and load reduction supports genuine purging in key genomic regions rather than simple sheltering effects. These patterns highlight fundamental differences between perennial and annual domestication models. In perennial fruit trees, domestication typically involves prolonged phenotypic evaluation over multiple years and repeated cycles of selection before genotypes are retained

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

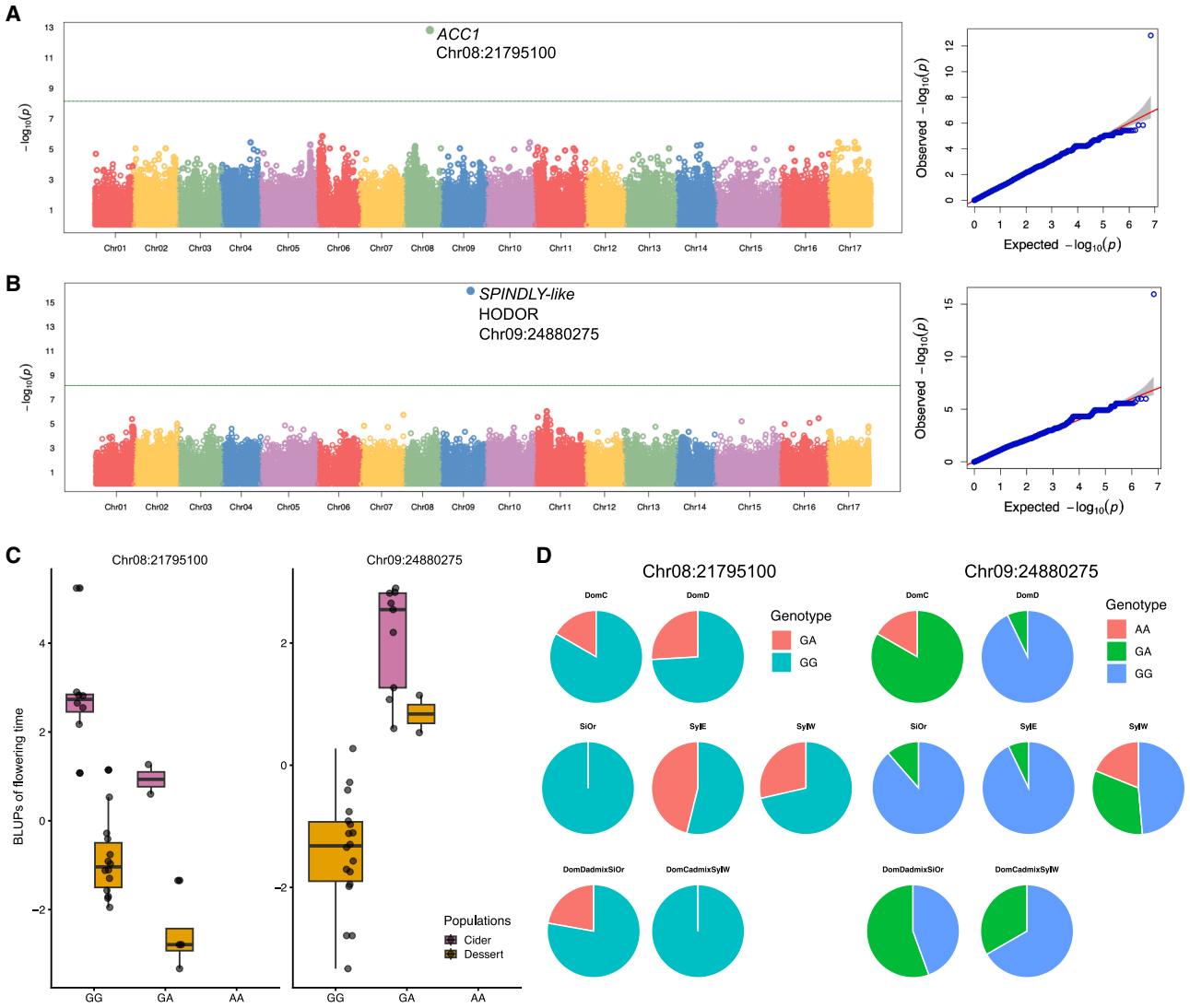


Figure 6. Genome-wide association analysis identifies loci associated with flowering time in cultivated apples

(A) Manhattan plot of $-\log_{10}(p)$ values based on the K+Q model and quantile-quantile (Q-Q) plot comparing the observed and expected $-\log_{10}(p)$ values based on the K+Q model.

(B) Manhattan plot of $-\log_{10}(p)$ values based on the K model. Q-Q plot comparing the observed and expected $-\log_{10}(p)$ values based on the K model. The solid horizontal line indicates the genome-wide significance threshold after Bonferroni correction ($p = 7.22 \times 10^{-9}$). The dashed horizontal line represents the FDR-adjusted significance threshold at 0.05.

(C) Boxplots of flowering-time BLUPs across genotypes and cider and dessert apple populations. The *y* axis shows BLUPs of flowering time, and the *x* axis indicates different genotypes. Box colors correspond to different subpopulations.

(D) Genotype distributions of two SNP loci detected with GWAS across wild and cultivated apple populations. Each pie chart represents the proportion of genotypes for a given SNP. Different colors indicate different subpopulations. DomC, cider *M. domestica*; DomD, dessert *M. domestica*; SiOr, *M. sieversii* and *M. orientalis*; SylE, Eastern Europe *M. sylvestris*; SylW, Western Europe *M. sylvestris*; BAcc, *M. baccata*; XadmixX, group of individuals classified as admixed between two populations (i.e., individuals assigned with membership coefficient < 0.8 to any given cluster at $K = 6$ with fastSTRUCTURE analyses).

See also Figure S7 and Data S2E.

and propagated. Combined with rapid linkage disequilibrium decay and large effective population sizes,[66] this extended selective filtering is expected to weaken Hill-Robertson interference by reducing linkage between adaptive alleles and nearby deleterious variants.[67] As a result, purifying selection can more efficiently eliminate deleterious mutations even in genomic regions affected by strong selective sweeps. We therefore propose that the distinctive domestication dynamics of perennial fruit trees, together with recurrent introgression from wild relatives, facilitate adaptive evolution while limiting the accumulation of linked deleterious variation, explaining the depletion of genetic load observed in sweep regions in DomC and DomD.

Together, our integrative analyses—enabled by broad sampling of *M. sylvestris* across its range—refine wild-to-crop donor histories[13,62,68] and demonstrate that apple domestication combines classic hard sweeps (Mendelian-like) with widespread

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

polygenic adaptation from standing and introgressed variation, much of which escapes detection by standard sweep scans. Future steps include measuring genome-wide polygenic shifts among cultivar groups with a more comprehensive sampling of *M. domestica*, particularly in Western and Northeastern Europe.[62,69] Concurrently, integrating allele-specific expression and methylation marks of nucleotide and structural variants—including TEs—will aid in distinguishing regulatory from coding regions involved in perennial domestication and in improving the prioritization of alleles for breeding and conservation amid global change.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Amandine Cornille (amandine.cornille@nyu.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- All sequencing data generated in this study have been deposited in the NCBI under BioProject accession NCBI: PRJNA1252109 (whole-genome sequencing, Illumina platform) and NCBI: PRJNA1297608 (RNA-seq data, Illumina platform).
- The analysis code and workflow are available at https://github.com/CornilleEclecticLab/Apple-domestication_population_genomics.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The authors thank the INRAE Biological Resource Center "Pome Fruits and Roses" (RosePom: https://www6.angers-nantes.inrae.fr/irhs/Ressources-mutualisees/Ressources-genetiques/CRB-Fruits-a-pepins-et-rosier) and associated staff, especially the INRAE Horticulture Experimental Facility (Unité Expérimentale Horticole: https://horti.angers-nantes.hub.inrae.fr), for maintaining the plant materials and associated datasets used in the present study and all samplers for their help (Data S1A). The authors thank Simon Martin for insightful discussions on detecting gene flow in genomes. We are also grateful to Henri Fourey, Tatiana Griaud, and Marina Olonova for preliminary input and discussions. We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://bioinfo.genotoul.fr) for providing help, computing, and/or storage resources. USDA is an equal opportunity provider, employer, and lender. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. This work was funded by ATIP-CNRS Inserm, IDEEV, LabEx BASC, Tamkeen, under the New York University; Abu Dhabi Research Institute grant AD 454; European Agricultural Fund for Rural Development (EAFRD/FEADER) through the LEADER programme (Île-de-France region), under grant agreement no. RIDF190219CR0110026; Interreg Project France-Wallonia-Flanders PROVERBIO, agreement number: 1.1.359; ANR-21-CE20-0005 PLEASURE; and Ministry of Research, Innovation and Digitization, CNCS/CCCDI-UEFISCDI, project number PN-IV-P8-8.1-PRE-HE-ORG-2024-0223, within PNCDI IV.

AUTHOR CONTRIBUTIONS

Experiment design, A.C.; materials collection, all authors; sample preparation and sequencing, X.C., A.V., A.C., C.R., and M.B.; validation, all authors; formal analysis, X.C., A.C., R.D., and K.A.; writing – original draft, X.C. and A.C.; writing – review and editing, X.C. and A.C., with all authors; visualization, X.C. and A.C.; funding acquisition, A.C. and T.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - DNA extraction, and sequencing
  - Variant calling and filtering
  - Population structure, genetic variation, and differentiation
  - Population genetic diversity and differentiation
  - Divergence time estimates and effective population size (*Ne*) over time
  - Genome-wide estimates of gene flow among wild and cultivated populations
  - Demographic modeling based on site frequency spectrum (SFS)
  - Detection of signatures of positive and balancing selection
  - Control of demographic history for selection analyses
  - Gene function annotation and enrichment analysis
  - RNA extraction, library preparation, and sequencing
  - RNA-sequenced data processing and differential expression analysis
  - Genomic landscape of introgressions and adaptive introgressions
  - Mutation load estimates
  - Genome-wide association study (GWAS) of flowering time in cultivated apples
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2026.03.048.

Received: October 18, 2025
Revised: January 17, 2026
Accepted: March 19, 2026

REFERENCES

1. Darwin, C. (1995). The Origin of Species, New Edition (Gramercy).
2. Purugganan, M.D. (2019). Evolutionary insights into the nature of plant domestication. Curr. Biol. *29*, R705–R714. https://doi.org/10.1016/j.cub.2019.05.053.
3. Gaut, B.S., Díez, C.M., and Morrell, P.L. (2015). Genomics and the contrasting dynamics of annual and perennial domestication. Trends Genet. *31*, 709–719. https://doi.org/10.1016/j.tig.2015.10.002.
4. Miller, A.J., and Gross, B.L. (2011). From forest to field: perennial fruit crop domestication. Am. J. Bot. *98*, 1389–1414. https://doi.org/10.3732/ajb.1000522.
5. Xiao, H., Wang, Y., Liu, W., Shi, X., Huang, S., Cao, S., Long, Q., Wang, X., Liu, Z., Xu, X., et al. (2025). Impacts of reproductive systems on grapevine genome and breeding. Nat. Commun. *16*, 2031. https://doi.org/10.1038/s41467-025-56817-7.
6. Gaut, B.S., Seymour, D.K., Liu, Q., and Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication. Nat. Plants *4*, 512–520. https://doi.org/10.1038/s41477-018-0210-1.
7. Zhou, Y., Massonnet, M., Sanjak, J.S., Cantu, D., and Gaut, B.S. (2017). Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. Proc. Natl. Acad. Sci. USA *114*, 11715–11720. https://doi.org/10.1073/pnas.1709257114.

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

8. Cornille, A., Giraud, T., Smulders, M.J.M., Roldán-Ruiz, I., and Gladieux, P. (2014). The domestication and evolutionary ecology of apples. Trends Genet. *30*, 57–65. https://doi.org/10.1016/j.tig.2013.10.002.

9. Cornille, A., Antolín, F., Garcia, E., Vernesi, C., Fietta, A., Brinkkemper, O., Kirleis, W., Schlumbaum, A., and Roldán-Ruiz, I. (2019). A multifaceted overview of apple tree domestication. Trends Plant Sci. *24*, 770–782. https://doi.org/10.1016/j.tplants.2019.05.007.

10. Soomro, T., Watts, S., Migicovsky, Z., and Myles, S. (2022). Cider and dessert apples: What is the difference? Plants People Planet *4*, 593–598. https://doi.org/10.1002/ppp3.10284.

11. Leforestier, D., Ravon, E., Muranty, H., Cornille, A., Lemaire, C., Giraud, T., Durel, C.-E., and Branca, A. (2015). Genomic basis of the differences between cider and dessert apple varieties. Evol. Appl. *8*, 650–661. https://doi.org/10.1111/eva.12270.

12. Lassois, L., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Hibrand-Saint-Oyant, L., Poncet, C., Lasserre-Zuber, P., Feugey, L., and Durel, C.-E. (2016). Genetic diversity, population structure, parentage analysis, and construction of core collections in the French apple germplasm based on SSR markers. Plant Mol. Biol. Rep. *34*, 827–844. https://doi.org/10.1007/s11105-015-0966-7.

13. Cornille, A., Gladieux, P., Smulders, M.J.M., Roldán-Ruiz, I., Laurens, F., Le Cam, B., Nersesyan, A., Clavel, J., Olonova, M., Feugey, L., et al. (2012). New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. PLoS Genet. *8*, e1002703. https://doi.org/10.1371/journal.pgen.1002703.

14. Migicovsky, Z., Gardner, K.M., Richards, C., Thomas Chao, C., Schwaninger, H.R., Fazio, G., Zhong, G.-Y., and Myles, S. (2021). Genomic consequences of apple improvement. Hortic. Res. *8*, 9. https://doi.org/10.1038/s41438-020-00441-7.

15. Chen, P., Li, Z., Zhang, D., Shen, W., Xie, Y., Zhang, J., Jiang, L., Li, X., Shen, X., Geng, D., et al. (2021). Insights into the effect of human civilization on *Malus* evolution and domestication. Plant Biotechnol. J. *19*, 2206–2220. https://doi.org/10.1111/pbi.13648.

16. Spengler, R.N. (2019). Origins of the apple: the role of megafaunal mutualism in the domestication of malus and rosaceous trees. Front. Plant Sci. *10*, 617. https://doi.org/10.3389/fpls.2019.00617.

17. Wedger, M.J., Schumann, A.C., and Gross, B.L. (2021). Candidate genes and signatures of directional selection on fruit quality traits during apple domestication. Am. J. Bot. *108*, 616–627. https://doi.org/10.1002/ajb2.1636.

18. Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., et al. (2017). High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. Nat. Genet. *49*, 1099–1106. https://doi.org/10.1038/ng.3886.

19. Auwera, G.V. der, and O'Connor, B. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, First Edition (O'Reilly Media).

20. Liao, L., Zhang, W., Zhang, B., Fang, T., Wang, X.-F., Cai, Y., Ogutu, C., Gao, L., Chen, G., Nie, X., et al. (2021). Unraveling a genetic roadmap for improved taste in the domesticated apple. Mol. Plant *14*, 1454–1471. https://doi.org/10.1016/j.molp.2021.05.018.

21. Raj, A., Stephens, M., and Pritchard, J.K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics *197*, 573–589. https://doi.org/10.1534/genetics.114.164350.

22. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? Heredity *116*, 362–371. https://doi.org/10.1038/hdy.2015.104.

23. Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Mol. Biol. Evol. *28*, 2239–2252. https://doi.org/10.1093/molbev/msr048.

24. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722. https://doi.org/10.1126/science.1188021.

25. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.

26. Spengler, R.N., Petraglia, M., Roberts, P., Ashastina, K., Kistler, L., Mueller, N.G., and Boivin, N. (2021). Exaptation traits for megafaunal mutualisms as a factor in plant domestication. Front. Plant Sci. *12*, 649394. https://doi.org/10.3389/fpls.2021.649394.

27. Pease, J.B., and Hahn, M.W. (2015). Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol. *64*, 651–662. https://doi.org/10.1093/sysbio/syv023.

28. Martin, S.H., and Van Belleghem, S.M.V. (2017). Exploring evolutionary relationships across the genome using topology weighting. Genetics *206*, 429–438. https://doi.org/10.1534/genetics.116.194720.

29. Malinsky, M., Svardal, H., Tyers, A.M., Miska, E.A., Genner, M.J., Turner, G.F., and Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat. Ecol. Evol. *2*, 1940–1955. https://doi.org/10.1038/s41559-018-0717-x.

30. Maier, R., Flegontov, P., Flegontova, O., Işıldak, U., Changmai, P., and Reich, D. (2023). On the limits of fitting complex models of population history to f-statistics. eLife *12*, e85492. https://doi.org/10.7554/eLife.85492.

31. Excoffier, L., Marchi, N., Marques, D.A., Matthey-Doret, R., Gouy, A., and Sousa, V.C. (2021). fastsimcoal2: demographic inference under complex evolutionary scenarios. Bioinformatics *37*, 4882–4885. https://doi.org/10.1093/bioinformatics/btab468.

32. Adams, R.H., Schield, D.R., and Castoe, T.A. (2019). Recent advances in the inference of gene flow from population genomic data. Curr. Mol. Bio. Rep. *5*, 107–115. https://doi.org/10.1007/s40610-019-00120-0.

33. Cheng, X., and Steinrücken, M. (2024). Population genomic scans for natural selection and demography. Annu. Rev. Genet. *58*, 319–339. https://doi.org/10.1146/annurev-genet-111523-102651.

34. Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol. Biol. Evol. *32*, 244–257. https://doi.org/10.1093/molbev/msu269.

35. Lian, X., Zhao, X., Zhao, Q., Wang, G., Li, Y., and Hao, Y. (2021). MdDREB2A in apple is involved in the regulation of multiple abiotic stress responses. Hortic. Plant J. *7*, 197–208. https://doi.org/10.1016/j.hpj.2021.03.006.

36. Yu, X., and Michaels, S.D. (2010). The Arabidopsis Paf1c complex component CDC73 participates in the modification of FLOWERING LOCUS C chromatin. Plant Physiol. *153*, 1074–1084. https://doi.org/10.1104/pp.110.158386.

37. Orgil, U., Araki, H., Tangchaiburana, S., Berkey, R., and Xiao, S. (2007). Intraspecific genetic variations, fitness cost and benefit of RPW8, a disease resistance locus in Arabidopsis thaliana. Genetics *176*, 2317–2333. https://doi.org/10.1534/genetics.107.070565.

38. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020.

39. Racimo, F., Marnetto, D., and Huerta-Sánchez, E. (2017). Signatures of archaic adaptive introgression in present-day human populations. Mol. Biol. Evol. *34*, 296–317. https://doi.org/10.1093/molbev/msw216.

40. Whiting, J.R. (2022). JimWhiting91/genotype_plot: Genotype Plot. Version v0.2.1 (Zenodo). https://doi.org/10.5281/ZENODO.5913504.

41. Ben Sadok, I., Tiecher, A., Galvez-Lopez, D., Lahaye, M., Lasserre-Zuber, P., Bruneau, M., Hanteville, S., Robic, R., Cournol, R., and Laurens, F. (2015). Apple fruit texture QTLs: year and cold storage effects on sensory and instrumental traits. Tree Genet. Genomes *11*, 119. https://doi.org/10.1007/s11295-015-0947-x.

42. Bai, Y., Dougherty, L., Li, M., Fazio, G., Cheng, L., and Xu, K. (2012). A natural mutation-led truncation in one of the two aluminum-activated

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

malate transporter-like genes at the Ma locus is associated with low fruit acidity in apple. Mol. Genet. Genomics *287*, 663–678. https://doi.org/10.1007/s00438-012-0707-7.

43. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025. https://doi.org/10.1371/journal.pcbi.1001025.
44. Ramu, P., Esuma, W., Kawuki, R., Rabbi, I.Y., Egesi, C., Bredeson, J.V., Bart, R.S., Verma, J., Buckler, E.S., and Lu, F. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat. Genet. *49*, 959–963. https://doi.org/10.1038/ng.3845.
45. Rodgers-Melnick, E., Bradbury, P.J., Elshire, R.J., Glaubitz, J.C., Acharya, C.B., Mitchell, S.E., Li, C., Li, Y., and Buckler, E.S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc. Natl. Acad. Sci. USA *112*, 3823–3828. https://doi.org/10.1073/pnas.1413864112.
46. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nat. Protoc. *11*, 1–9. https://doi.org/10.1038/nprot.2015.123.
47. Wang, J., and Zhang, Z. (2021). GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genom. Proteom. Bioinform. *19*, 629–640. https://doi.org/10.1016/j.gpb.2021.08.005.
48. Huang, M., Liu, X., Zhou, Y., Summers, R.M., and Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience *8*, giy154. https://doi.org/10.1093/gigascience/giy154.
49. Amid, A., Lytovchenko, A., Fernie, A.R., Warren, G., and Thorlby, G.J. (2012). The sensitive to freezing3 mutation of Arabidopsis thaliana is a cold-sensitive allele of homomeric acetyl-CoA carboxylase that results in cold-induced cuticle deficiencies. J. Exp. Bot. *63*, 5289–5299. https://doi.org/10.1093/jxb/ers191.
50. Sun, T.P. (2021). Novel nucleocytoplasmic protein O-fucosylation by SPINDLY regulates diverse developmental processes in plants. Curr. Opin. Struct. Biol. *68*, 113–121. https://doi.org/10.1016/j.sbi.2020.12.013.
51. Cornille, A., Gladieux, P., and Giraud, T. (2013). Crop-to-wild gene flow and spatial genetic structure in the closest wild relatives of the cultivated apple. Evol. Appl. *6*, 737–748. https://doi.org/10.1111/eva.12059.
52. Chen, X., Avia, K., Forler, A., Remoué, C., Venon, A., Rousselet, A., Lucas, G., Kwarteng, A.O., Rover, R., Le Guilloux, M., et al. (2023). Ecological and evolutionary drivers of phenotypic and genetic variation in the European crabapple [Malus sylvestris (L.) Mill.], a wild relative of the cultivated apple. Ann. Bot. *131*, 1025–1037. https://doi.org/10.1093/aob/mcad061.
53. Cornille, A., Giraud, T., Bellard, C., Tellier, A., Le Cam, B., Smulders, M.J.M., Kleinschmit, J., Roldan-Ruiz, I., and Gladieux, P. (2013). Postglacial recolonization history of the European crabapple (*Malus sylvestris* Mill.), a wild contributor to the domesticated apple. Mol. Ecol. *22*, 2249–2263. https://doi.org/10.1111/mec.12231.
54. Meng, D., Cao, H., Yang, Q., Zhang, M., Borejsza-Wysocka, E., Wang, H., Dandekar, A.M., Fei, Z., and Cheng, L. (2023). SnRK1 kinase-mediated phosphorylation of transcription factor bZIP39 regulates sorbitol metabolism in apple. Plant Physiol. *192*, 2123–2142. https://doi.org/10.1093/plphys/kiad226.
55. Yu, J.-Q., Gu, K.-D., Sun, C.-H., Zhang, Q.-Y., Wang, J.-H., Ma, F.-F., You, C.-X., Hu, D.-G., and Hao, Y.-J. (2021). The apple bHLH transcription factor MdbHLH3 functions in determining the fruit carbohydrates and malate. Plant Biotechnol. J. *19*, 285–299. https://doi.org/10.1111/pbi.13461.
56. Yu, J.-Q., Gu, K.-D., Zhang, L.-L., Sun, C.-H., Zhang, Q.-Y., Wang, J.-H., Wang, C.-K., Wang, W.-Y., Du, M.-C., and Hu, D.-G. (2022). MdbHLH3 modulates apple soluble sugar content by activating phosphofructokinase gene expression. J. Integr. Plant Biol. *64*, 884–900. https://doi.org/10.1111/jipb.13236.
57. Kumar, S., Deng, C.H., Molloy, C., Kirk, C., Plunkett, B., Lin-Wang, K., Allan, A., and Espley, R. (2022). Extreme-phenotype GWAS unravels a complex nexus between apple (*Malus domestica*) red-flesh colour and internal flesh browning. FRURES *2*, 1–14. https://doi.org/10.48130/FruRes-2022-0012.
58. Bouché, F., Lobet, G., Tocquin, P., and Périlleux, C. (2016). FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana. Nucleic Acids Res. *44*, D1167–D1171. https://doi.org/10.1093/nar/gkv1054.
59. Dong, Y., Duan, S., Xia, Q., Liang, Z., Dong, X., Margaryan, K., Musayev, M., Goryslavets, S., Zdunić, G., Bert, P.-F., et al. (2023). Dual domestications and origin of traits in grapevine evolution. Science *379*, 892–901. https://doi.org/10.1126/science.add8655.
60. Zohary, D., Hopf, M., and Weiss, E. (2012). Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin, Fourth Edition (Oxford University Press). https://doi.org/10.1093/acprof:osobl/9780199549061.001.0001.
61. Ruzicka, F., Zwoinska, M.K., Goedert, D., Kokko, H., Richter, X.-Y.L., Moodie, I.R., Nilén, S., Olito, C., Svensson, E.I., Czuppon, P., et al. (2025). A century of theories of balancing selection. Biol. Rev. *101*, 804–825. https://doi.org/10.1101/2025.02.12.637871.
62. Urrestarazu, J., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Feugey, L., Poncet, C., Lateur, M., Houben, P., Ordidge, M., et al. (2016). Analysis of the genetic diversity and structure across a wide range of germplasm reveals prominent gene flow in apple at the European level. BMC Plant Biol. *16*, 130. https://doi.org/10.1186/s12870-016-0818-0.
63. Bosse, M., Megens, H.-J., Derks, M.F.L., de Cara, Á.M.R., and Groenen, M.A.M. (2019). Deleterious alleles in the context of domestication, inbreeding, and selection. Evol. Appl. *12*, 6–17. https://doi.org/10.1111/eva.12691.
64. Renaut, S., and Rieseberg, L.H. (2015). The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. Mol. Biol. Evol. *32*, 2273–2283. https://doi.org/10.1093/molbev/msv106.
65. Zhang, X., Song, B., Du, S., Zhang, S., Ren, Y., Xue, C., Xu, S., Zheng, P., Chen, S., Qiao, Z., et al. (2025). Genomic insights into deleterious mutations and their impact on agronomic traits during pear domestication. Hortic. Res. *12*, uhaf140. https://doi.org/10.1093/hr/uhaf140.
66. Petit, R.J., and Hampe, A. (2006). Some evolutionary consequences of being a tree. Annu. Rev. Ecol. Evol. Syst. *37*, 187–214. https://doi.org/10.1146/annurev.ecolsys.37.091305.110215.
67. Barton, N.H. (2010). Genetic linkage and natural selection. Philos. Trans. R. Soc. Lond., B Biol. Sci. *365*, 2559–2569. https://doi.org/10.1098/rstb.2010.0106.
68. Harris, S.A., Robinson, J.P., and Juniper, B.E. (2002). Genetic clues to the origin of the apple. Trends Genet. *18*, 426–430. https://doi.org/10.1016/S0168-9525(02)02689-6.
69. Skytte Af Sätra, J., Garkava-Gustavsson, L., and Ingvarsson, P.K. (2024). Why we thrive beneath a northern sky - genomic signals of selection in apple for adaptation to northern Sweden. Heredity *133*, 67–77. https://doi.org/10.1038/s41437-024-00693-2.
70. Duan, N., Bai, Y., Sun, H., Wang, N., Ma, Y., Li, M., Wang, X., Jiao, C., Legall, N., Mao, L., et al. (2017). Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. Nat. Commun. *8*, 249. https://doi.org/10.1038/s41467-017-00336-7.
71. Sun, X., Jiao, C., Schwaninger, H., Chao, C.T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K., Cheng, L., et al. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat. Genet. *52*, 1423–1432. https://doi.org/10.1038/s41588-020-00723-9.
72. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

73. Andrews, S. (2018). FastQC: a quality control tool for high throughput sequence data, Version 0.11.8. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

74. Vasimuddin, M., Misra, S., Li, H., and Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 314–324. https://doi.org/10.1109/IPDPS.2019.00041.

75. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A.V. der, Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv. https://doi.org/10.1101/201178.

76. Lambert, I., Paysant-Le Roux, C., Colella, S., and Martin-Magniette, M.-L. (2020). DiCoExpress: a tool to process multifactorial RNAseq experiments from quality controls to co-expression analysis through differential analysis based on contrasts inside GLM models. Plant Methods *16*, 68. https://doi.org/10.1186/s13007-020-00611-7.

77. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

78. Siberchicot, A., Bessy, A., Guéguen, L., and Marais, G.A.B. (2017). MareyMap Online: a user-friendly web application and database service for estimating recombination rates using physical and genetic maps. Genome Biol. Evol. *9*, 2506–2509. https://doi.org/10.1093/gbe/evx178.

79. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. *108*, 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005.

80. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., and Yang, T.-L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics *35*, 1786–1788. https://doi.org/10.1093/bioinformatics/bty875.

81. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

82. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

83. Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. *15*, 1179–1191. https://doi.org/10.1111/1755-0998.12387.

84. Francis, R.M. (2017). pophelper: an R package and web app to analyse and visualize population structure. Mol. Ecol. Resour. *17*, 27–32. https://doi.org/10.1111/1755-0998.12509.

85. Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. *23*, 254–267. https://doi.org/10.1093/molbev/msj030.

86. Korunes, K.L., and Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. Mol. Ecol. Resour. *21*, 1359–1368. https://doi.org/10.1111/1755-0998.13326.

87. Terhorst, J., Kamm, J.A., and Song, Y.S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat. Genet. *49*, 303–309. https://doi.org/10.1038/ng.3748.

88. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321. https://doi.org/10.1093/sysbio/syq010.

89. Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. *33*, 1635–1638. https://doi.org/10.1093/molbev/msw046.

90. Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform. *19*, 153. https://doi.org/10.1186/s12859-018-2129-y.

91. Malinsky, M., Matschiner, M., and Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. Mol. Ecol. Resour. *21*, 584–595. https://doi.org/10.1111/1755-0998.13265.

92. Pease, J.B., and Rosenzweig, B.K. (2018). Encoding data using biological principles: the multisample variant format for phylogenomics and population genomics. IEEE/ACM Trans. Comput. Biol. Bioinform. *15*, 1231–1238. https://doi.org/10.1109/TCBB.2015.2509997.

93. Alachiotis, N., Stamatakis, A., and Pavlidis, P. (2012). OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics *28*, 2274–2275. https://doi.org/10.1093/bioinformatics/bts419.

94. Alachiotis, N., and Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. Commun. Biol. *1*, 79. https://doi.org/10.1038/s42003-018-0085-8.

95. Harris, A.M., Garud, N.R., and DeGiorgio, M. (2018). Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity. Genetics *210*, 1429–1452. https://doi.org/10.1534/genetics.118.301502.

96. Siewert, K.M., and Voight, B.F. (2020). BetaScan2: standardized statistics to detect balancing selection utilizing substitution data. Genome Biol. Evol. *12*, 3873–3877. https://doi.org/10.1093/gbe/evaa013.

97. Schiffels, S., and Wang, K. (2020). MSMC and MSMC2: the multiple sequentially Markovian coalescent. Methods Mol. Biol. *2090*, 147–166. https://doi.org/10.1007/978-1-0716-0199-0_7.

98. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.

99. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.

100. Keightley, P.D., and Jackson, B.C. (2018). Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. Genetics *209*, 897–906. https://doi.org/10.1534/genetics.118.301120.

101. Venables, W.N., Ripley, B.D., and Venables, W.N. (2002). Modern Applied Statistics with S, Fourth Edition (Springer). https://doi.org/10.1007/978-0-387-21706-2.

102. Morgan, J. (2013). The New Book of Apples: The Definitive Guide to Over 2000 Varieties, Revised Edition (Ebury Digital).

103. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

104. Di Pierro, E.A., Gianfranceschi, L., Di Guardo, M., Koehorst-van Putten, H.J., Kruisselbrink, J.W., Longhi, S., Troggio, M., Bianco, L., Muranty, H., Pagliarani, G., et al. (2016). A high-density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. Hortic. Res. *3*, 16057. https://doi.org/10.1038/hortres.2016.57.

105. Howard, N.P., Troggio, M., Durel, C.-E., Muranty, H., Denancé, C., Bianco, L., Tillman, J., and van de Weg, E. (2021). Integration of Infinium and Axiom SNP array data in the outcrossing species Malus × domestica and causes for seemingly incompatible calls. BMC Genomics *22*, 246. https://doi.org/10.1186/s12864-021-07565-7.

106. Nei, M., and Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA *76*, 5269–5273. https://doi.org/10.1073/pnas.76.10.5269.

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

107. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x.

108. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80–92. https://doi.org/10.4161/fly.19695.

109. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag). https://doi.org/10.1007/978-0-387-98141-3.

110. Rochette, N.C., Rivera-Colón, A.G., and Catchen, J.M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol. Ecol. *28*, 4737–4754. https://doi.org/10.1111/mec.15253.

111. Pockrandt, C., Alzamel, M., Iliopoulos, C.S., and Reinert, K. (2020). GenMap: ultra-fast computation of genome mappability. Bioinformatics *36*, 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222.

112. Overcast, I. (2023). easySFS: Effective selection of population size projection for construction of the site frequency spectrum. Github. https://github.com/isaacovercast/easySFS

113. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. *5*, e1000695. https://doi.org/10.1371/journal.pgen.1000695.

114. Dong, S.-S., He, W.-M., Ji, J.-J., Zhang, C., Guo, Y., and Yang, T.-L. (2021). LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief. Bioinform. *22*, bbaa227. https://doi.org/10.1093/bib/bbaa227.

115. Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338. https://doi.org/10.1093/bioinformatics/18.2.337.

116. Yu, G., Hu, E., and Gao, C.-H. (2023). enrichplot: visualization of functional enrichment result. Version 1.20.0 (Bioconductor version: Release (3.17)). https://doi.org/10.18129/B9.bioc.enrichplot.

117. Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., and Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. *45*, D1040–D1045. https://doi.org/10.1093/nar/gkw982.

118. Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

119. Toda, N., Rustenholz, C., Baud, A., Le Paslier, M.-C., Amselem, J., Merdinoglu, D., and Faivre-Rampant, P. (2020). NLGenomeSweeper: a tool for genome-wide NBS-LRR resistance gene identification. Genes *11*, 333. https://doi.org/10.3390/genes11030333.

120. Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics *28*, 3211–3217. https://doi.org/10.1093/bioinformatics/bts611.

121. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

122. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. https://doi.org/10.1038/nmeth.4197.

123. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25.

124. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

125. Gao, Y., Yang, Q., Yan, X., Wu, X., Yang, F., Li, J., Wei, J., Ni, J., Ahmad, M., Bai, S., et al. (2021). High-quality genome assembly of "Cuiguan" pear (Pyrus pyrifolia) as a reference genome for identifying regulatory genes and epigenetic modifications responsible for bud dormancy. Hortic. Res. *8*, 197. https://doi.org/10.1038/s41438-021-00632-w.

126. International Peach Genome Initiative, Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M.T., Grimwood, J., and et al.. (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat. Genet. *45*, 487–494. https://doi.org/10.1038/ng.2586.

127. Gao, S., Wang, B., Xie, S., Xu, X., Zhang, J., Pei, L., Yu, Y., Yang, W., and Zhang, Y. (2020). A high-quality reference genome of wild Cannabis sativa. Hortic. Res. *7*, 73. https://doi.org/10.1038/s41438-020-0295-3.

128. Ireland, H.S., Wu, C., Deng, C.H., Hilario, E., Saei, A., Erasmuson, S., Crowhurst, R.N., David, K.M., Schaffer, R.J., and Chagné, D. (2021). The Gillenia trifoliata genome reveals dynamics correlated with growth and reproduction in Rosaceae. Hortic. Res. *8*, 233. https://doi.org/10.1038/s41438-021-00662-4.

129. Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Aiese Cigliano, R., Del Cueto, J., Ricciardi, F., Lotti, C., Ricciardi, L., Dicenta, F., et al. (2019). Mutation of a bHLH transcription factor allowed almond domestication. Science *364*, 1095–1098. https://doi.org/10.1126/science.aav8197.

130. Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S.P., et al. (2011). The genome of woodland strawberry (Fragaria vesca). Nat. Genet. *43*, 109–116. https://doi.org/10.1038/ng.740.

131. Wang, Y., and Paterson, A.H. (2021). Loquat (*Eriobotrya japonica* (*Thunb.*) Lindl.) population genomics suggests a two-staged domestication and identifies genes showing convergence/parallel selective sweeps with apple or peach. Plant J. *106*, 942–952. https://doi.org/10.1111/tpj.15209.

132. Harris, R.S. (2007). Improved pairwise alignment of genomic DNA. PhD thesis (Pennsylvania State University).

133. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

134. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288. https://doi.org/10.1093/bioinformatics/btm098.

135. Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111–120. https://doi.org/10.1007/BF01731581.

136. Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814. https://doi.org/10.1093/nar/gkg509.

137. Simons, Y.B., Turchin, M.C., Pritchard, J.K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. Nat. Genet. *46*, 220–224. https://doi.org/10.1038/ng.2896.

138. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., B *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Leave sampels	This study	See Data S1
NovaSeq 6000	Illumina, Inc.	www.illumina.com/systems/sequencing-platforms/novaseq.html
Agilent 2100 Bioanalyzer	Agilent Technologies	N/A
Chemicals, peptides, and recombinant proteins		
Macherey–Nagel DNA Mini Kit	Macherey-Nagel	N/A
NEBNext DNA Library Prep kits	New England Biolabs	N/A
Nucleospin RNA Plant Kit	Macherey-Nagel	N/A
Novogene NGS Stranded RNA Library Prep Set	Novogene	PT044
Deposited data		
Whole-genome sequencing raw data	This paper	NCBI Bioproject: PRJNA1252109
RNA-seq sequencing raw data	This paper	NCBI Bioproject: PRJNA1297608
GDDH13 v1.1 reference genome	Daccord et al.[18]	https://iris.angers.inra.fr/gddh13/
Cultivar and wild apple whole-genome resequencing data	Duan et al.[70] and Sun et al.[71]	NCBI Bioproject: PRJNA322175 and PRJNA591623
Software and algorithms		
fastp v0.21.0	Chen et al.[72]	https://github.com/OpenGene/fastp
FastQC v0.11.8	Andrews[73]	https://github.com/s-andrews/FastQC
bwa-mem2 v2.1	Vasimuddin et al.[74]	https://github.com/bwa-mem2/bwa-mem2
GATK Best Practices Workflows v4.1.7	Auwera and O'Connor[19] and Poplin et al.[75]	https://github.com/broadinstitute/gatk
DiCoExpress v1	Lambert et al.[76]	https://forge.inrae.fr/GNet/dicoexpress
Plink 2.0 alpha	Chang et al.[77]	https://www.cog-genomics.org/plink/2.0
MareyMap Online	Siberchicot et al.[78]	https://lbbe-shiny.univ-lyon1.fr/MareyMapOnline
Beagle v5.1	Browning et al.[79]	https://faculty.washington.edu/browning/beagle/b5_1.html
PopLDdecay v3.42	Zhang et al.[80]	github.com/BGI-shenzhen/PopLDdecay
Plink v1.9	Purcell et al.[81]	https://www.cog-genomics.org/plink/1.9
vcftools v0.1.16	Danecek et al.[82]	https://vcftools.github.io/man_latest.html
fastSTRUCTURE v1.0	Raj et al.[21]	https://rajanil.github.io/fastStructure/
CLUMPAK v1.1	Kopelman et al.[83]	http://clumpak.tau.ac.il
R package pophelper v2.3.2	Francis[84]	http://github.com/royfrancis/pophelper
Splitstree v4.12.3	Huson and Bryant[85]	http://www.splitstree.org/
Pixy v1.0.0	Korunes and Samuk[86]	https://github.com/ksamuk/pixy
SMC++ v1.15.2	Terhorst et al.[87]	https://github.com/popgenmethods/smcpp
PhyML v3.0	Guindon et al.[88]	https://github.com/stephaneguindon/phyml
ETE toolkit v3.1.3	Huerta-Cepas et al.[89]	https://etetoolkit.org/
ASTRAL-III v5.7.1	Zhang et al.[90]	https://github.com/smirarab/ASTRAL
fastsimcoal2 v2.8	Excoffier et al.[31]	https://cmpg.unibe.ch/software/fastsimcoal28/
Dsuite v0.4	Malinsky et al.[91]	https://github.com/millanek/Dsuite
ADMIXTOOLS v2.0.4	Maier et al.[30]	https://github.com/uqrmaie1/admixtools/

(*Continued on next page*)

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

CellPress
OPEN ACCESS

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Twisst	Martin and Belleghem[28]	https://github.com/simonhmartin/twisst
mvftool v0.6.2.1 (D_{FOIL})	Pease and Hahn[27] and Pease and Rosenzweig[92]	https://github.com/peaselab/mvftools and https://github.com/jbpease/dfoil
RFMix v2.03	Maples et al.[38]	https://github.com/slowkoni/rfmix
bin.vcf2QU.pl (Q95 statistic)	Racimo et al.[39] and this paper	https://github.com/CornilleEclecticLab/Apple-domestication_population_genomics
Genotype Plot v0.2.1	Whiting[40]	https://github.com/JimWhiting91/genotype_plot
OmegaPlus v3.0.3	Alachiotis et al.[93]	https://github.com/alachins/omegaplus
RAiSD v2.9	Alachiotis and Pavlidis[94]	github.com/alachins/raisd
SelectionHapStats (G12 and G123 statistics)	Harris et al.[95]	https://github.com/ngarud/SelectionHapStats
BetaScan2	Siewert and Voight[96]	https://github.com/ksiewert/BetaScan
MSMC2 v2.1.4	Schiffels and Wang[97]	https://github.com/stschiff/msmc2
eggNOG-mapper v2.1.5	Huerta-Cepas et al.[98]	https://github.com/eggnogdb/eggnog-mapper
ClusterProfiler v3.22	Wu et al.[99]	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
GERP++	Davydov et al.[43]	https://github.com/tvkent/GERPplusplus
SIFT4G	Vaser et al.[46]	https://github.com/pauline-ng/SIFT4G_Create_Genomic_DB and https://github.com/pauline-ng/SIFT4G_Annotator
est-sfs v2.0	Keightley and Jackson[100]	https://sourceforge.net/projects/est-usfs/
R package MASS v7.3.61	Venables et al.[101]	https://cran.r-project.org/web/packages/MASS/index.html
R package GAPIT v3.5	Wang and Zhang[47]	https://github.com/jiabowang/GAPIT

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

A total of 140 newly sequenced individuals originated from leaves of 43 cultivated apple (*Malus domestica*) samples, including 21 dessert and 22 cider apple varieties, 75 *M. sylvestris* (European crabapple), nine *M. orientalis*, three *M. sieversii*, and ten *M. baccata*. The status of cider and dessert cultivars was assessed using the New Book of the Apple[102] and the expertise of the ResPom consortium (INRAE Angers) (Data S1A and S4A). Leaf tissues were collected from actively growing shoots for genomic DNA extraction. In addition, publicly available high-coverage genomes[70,71] were incorporated to complement the dataset (Data S3B). For transcriptomic analyses, RNA-seq data were generated from 59 leaf samples from cider *M. domestica* (n = 15, five genotypes × three clonal replicates), dessert *M. domestica* (n = 15, 5 genotypes × three clonal replicates), *M. sylvestris* (n = 15, five genotypes representing SylW × three clonal replicates), and *M. sieversii* (n = 12, four genotypes × three clonal replicates) to investigate domestication-associated expression divergence. All plant material used in this study was collected or sourced in full accordance with applicable international regulations, including the Nagoya Protocol and the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA). Proper permits and agreements were secured for the collection, use, and analysis of both cultivated and wild apple genetic resources.

METHOD DETAILS

DNA extraction, and sequencing

DNA was extracted with the Macherey-Nagel DNA Mini Kit using the manufacturer's instructions. DNA quality controls were performed using two methods: (1) Agarose gel electrophoresis for testing DNA degradation and potential contamination, and (2) Qubit 2.0, which quantifies the DNA concentration precisely. A total of 1.0 μg of DNA per sample was then used as input material for the DNA library preparations. Sequencing libraries were generated using NEBNext DNA Library Prep Kit following the manufacturer's recommendations, and indices were added to each sample. The genomic DNA was randomly fragmented to 350 bp by shearing, then the DNA fragments were end-polished, A-tailed, and ligated to the NEBNext adapter for Illumina sequencing, and further PCR-enriched with P5 and indexed P7 oligos. The PCR products were purified (AMPure XP system), and resulting libraries were

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

analyzed for size distribution by Agilent 2100 Bioanalyzer and quantified using real-time PCR. After pooling, the qualified libraries were loaded into Illumina sequencers based on their effective concentration and expected data volume.

This newly sequenced dataset was complemented with publicly available sequencing data with at least 15× coverage (78 other genotypes in Duan et al. and Sun et al.,[70,71] including 24 dessert and one cider apple cultivars, one *M. orientalis*, 37 *M. sieversii*, 12 *M. sylvestris*, and three *M. baccata*). Therefore, the total dataset included 218 genotypes and was composed of 45 dessert and 23 cider apple varieties and the four main wild relatives of the cultivated apple: *M. orientalis* (n = 10), *M. sieversii* (n = 40), *M. sylvestris* (n = 87), and *M. baccata* (n = 13) (Data S1A and S4A).

Variant calling and filtering

Quality control and pre-processing of the reads were performed using FastQC v0.11.8[73] and fastp v0.21.0,[72] respectively. Variant calling was carried out following the Best Practices procedure.[19] Clean DNA reads were aligned to the GDDH13 Version 1.1 reference genome[18] using bwa-mem v2.1[74] (https://github.com/bwa-mem2/bwa-mem2). Mapping statistics are shown in Data S2. Raw variants were called using GATK HaplotypeCaller[75] (GATK version 4.1.7, https://github.com/broadinstitute/gatk). The raw variants were filtered using the GATK variant filter module with a hard filter (Level 1 filters) and further refined with a custom script (Level 2 filters) (Figure S1). Genomic kinship was calculated for all pairs of individuals using Plink2.[77] Duplicate or clonally related individuals were excluded from further analysis (i.e., individuals with pairwise KING-robust kinship estimates greater than 0.354),[103] using 'Plink2 –make-king'. Individuals with a missing rate > 30% were also removed. All filtering parameters are listed in Data S3A, and the removed samples are detailed in Data S1A, which represents 17 genotypes and retains 201 individuals. The dataset was phased using Beagle v5.1[79] with default settings and genetic maps[104] for each chromosome. The genetic map position of each SNP was obtained by the MareyMap pipeline. The recombination rate variation across the genome was estimated using high-density integrated genetic maps.[104,105] We downloaded marker genetic distance and physical distance data for the GDDH v1.1 reference genome and imported them into MareyMap Online(https://lbbe-shiny.univ-lyon1.fr/MareyMapOnline/), then followed the MareyMap pipeline to obtain recombination rates along the genome.

Pairwise linkage disequilibrium (LD) between SNPs was computed by PopLDdecay v3.42.[80] The decay distance of LD was defined as the physical distance at which r^2 decreased to half of its maximum value. The LD decay distance varied among populations, ranging from 6.6 Kb to 123.5 Kb (Data S3D). For genetic diversity (π),[106] genetic distance (D_{XY}),[106] differentiation (F_{ST}),[107] and f_d scans analysis, a window size of 10 kb was chosen, providing a compromise between resolution and statistical robustness given the SNP density across the genome. For the D_{FOIL} analysis, a window size of 100 kb was used because the information content in five individual sets of 10 kb windows was insufficient.

Linked and non-synonymous SNPs were filtered out for population structure and demographic inference to avoid bias from potential SNPs linked to genomic regions under selection (Level 3 filters, Data S3A). SNPs were annotated with SnpEff v5.0,[108] and only synonymous SNPs were retained. SNPs in LD were removed. Linked SNPs were filtered using Plink v1.9[81] with the '–indep-pairwise' function, applying the following parameters: window = 5 Kb, step = 1 SNP, and r^2 = 0.2. SNPs with distances less than 5 Kb were removed using the 'thin' function of vcftools v0.1.16.[82] In total, 31,300 unlinked and synonymous SNPs were used for analyses of population structure, genetic variation, and differentiation across 201 genotypes.

Population structure, genetic variation, and differentiation

The population structure and admixture were inferred using the Bayesian clustering method implemented in fastSTRUCTURE v1.0,[21] with 31,300 unlinked, synonymous SNPs and 201 genotypes. Analyses were performed with default parameters, running 20 iterations for each K value from K = 1 to K = 10 clusters. The consensus solution for each K value was generated using CLUMPAK,[83] and the optimal K value was determined based on the cross-validation error.[21] The bar plots were generated using the R package pophelper v2.3.1, and the K value that yielded well-assigned individuals across all clusters was selected.

Two additional methods were employed to further explore the genetic variation and differentiation among the genetic groups detected by fastSTRUCTURE. Principal component analysis (PCA) was conducted using the Plink v1.9 PCA module,[81] and the first three principal components were visualized using the R package "ggplot2"[109]. A Hamming distance matrix among individuals was calculated using Plink v1.9,[81] followed by the construction of a neighbor net, which was visualized with SplitsTree v4.12.3.[85] For both PCA and neighbor-net, individuals assigned to a given cluster with a membership coefficient $\geq$ 0.8 (Figure S4) were color-coded according to their genetic cluster inferred with fastSTRUCTURE, while admixed individuals were colored gray. This threshold was selected based on the distribution of maximum membership coefficients inferred by fastSTRUCTURE (see Results, Figure S4B).

Population genetic diversity and differentiation

Descriptive population genetic estimates were computed for each population (i.e., each cluster inferred with fastSTRUCTURE, excluding admixed individuals with a membership coefficient < 0.80). Genetic diversity (π),[106] genetic distance (D_{XY}),[106] and differentiation (F_{ST})[107] between populations were estimated using pixy[86] with the "all-sites" variant calling format (VCF). Population genetic diversity statistics—including the number of private alleles (A_p), number of polymorphic sites (S), observed heterozygosity(H_o), expected heterozygosity (H_e), and inbreeding coefficient (F_{IS})—were calculated for each population using the *populations* module in Stacks.[110] Genome-wide π, D_{XY}, and F_{ST} were calculated using all sites (332,335,748 variant and invariant sites) in 10 Kb window regions of high mappability.

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

CellPress
OPEN ACCESS

Divergence time estimates and effective population size (*Ne*) over time

To infer population divergence history, we first reconstructed a coalescent species tree using ASTRAL-III.[90] ASTRAL is a tool for inferring an unrooted species tree from a set of unrooted gene trees under the multispecies coalescent model. This method is particularly well suited to addressing incomplete lineage sorting, a common phenomenon in tree species.[66] Phylogenetic trees per 100 SNP windows were calculated with PhyML[88] with the GTR model. The output trees were re-rooted with a single BAcc individual using the ETE Toolkit,[89] and then used as input to ASTRAL-III. These trees were also used as input to Twisst[28] to get the most common species-tree topology of cultivated and wild apple populations.

Effective population size (Ne) over time and population divergence estimates were inferred using two methods: SMC++.[87] First, the reference genome mappability scores were calculated with GenMap v1.3.0[111] using a *k*-mer size of 140 (the average length of filtered sequencing reads) for 20 Kb windows with 10 Kb steps. Any mean scores of 20 kb or higher than 0.8 were defined as high-mappability regions. SMC++ demographic inference employed SNPs within high mappability regions with all individuals for each population. A mutation rate of 3.9×10^{-9} per bp per generation and a generation time of 10 years were used to convert generations into years.[13] The divergence time between DomD and SiOr, DomC and SiOr, and SylW and SylE was also inferred by *SMC++* using the "*split*" command.

Genome-wide estimates of gene flow among wild and cultivated populations

D-statistics and f_4-ratio were employed to detect gene flow among cultivated and wild populations. The *D*-statistic (ABBA-BABA test)[23,28] and related statistics, such as the f_4-ratio,[25] are widely used parsimony-like methods for distinguishing gene flow from incomplete lineage sorting. While the *D*-statistic tests for introgression, the f_4-ratio estimates the proportion of admixture or introgression using population allele frequencies. We computed genome-wide four-taxon *D*-statistics and f_4-ratios using the Dtrios command in Dsuite[91] with unlinked 31,300 SNPs. Tests were performed on populations identified by fastSTRUCTURE, excluding admixed individuals with membership coefficients <0.80, and using BAcc as the outgroup. *D*-statistics and f_4-ratios were computed for all possible trios and organized according to the tree topology. The significance of *D*-statistics was assessed using a jackknife across 20 blocks.[23,24] Results were visualized using the Ruby scripts plot_d.rb and plot_f4ratio.rb available from M. Matschiner's repository. After that, *D*-statistics and f_4-ratios were calculated for each branch of the population tree using the "Fbranch" module in Dsuite using the most common species-tree topology and the coalescent species-tree inferred by ASTRAL (Data S4E). The *f*-branch result was visualized using the dtools.py script provided in the Dsuite.

We investigated the directionality and extent of gene flow with *qpGraph* implemented in the R package ADMIXTOOLS 2.[30] As input, allele frequencies and f_2-statistics[25] were calculated for each population using the 31,300 unlinked and synonymous SNPs. The function "find_graphs" in qpGraph was then executed to explore admixture graphs with the assumed number of admixture events ranging from 0 to 4 using BAcc as the outgroup population. A testing procedure was used in which the best-scoring graph among 200 candidate graphs for a given number of migrations was first identified, and graphs with the same number of migrations were not significantly worse (*p-value* > 0.05). It was also tested whether the best-scoring graph for each number of migrations could be significantly rejected in favor of a graph with a higher number of migrations. The test score was calculated by optimizing a topology from a subset of the *f*-statistics and evaluating it on the remaining data. A significance test was performed for each graph using a jackknife approach, and the results were then compared with those from the remaining graphs using a nominal *p-value* of 0.05. The optimal number of admixture events was determined by identifying the graph that was significantly better than all graphs with fewer migrations (Data S3C and S4E).

To further characterize the gene flow between cultivated and wild apple, full genome SNPs (28,377,551) were used for D_{FOIL}, Twisst, and chromosome-level *D*-statistic, and f_d genome scan analysis. The D_{FOIL} analysis[27] was conducted on a selection of five populations (*P1, P2, P3, P4*, and *O*), and the divergence time between P1 and P2 was earlier than that between P3 and P4. The *M. baccata* population was used as the outgroup (*O*). The SMC++ divergence time estimates suggest that the SylW and SylE divergence time was earlier than DomC from SiOr but later than DomD from SiOr. So, all four possible populations set, set1 (P1 = SiOr, P2 = DomC, P3 = SylW, P4 = SylE, O = Bacc), set2 (P1 = SiOr, P2 = DomD, P3 = SylW, P4 = SylE, O = Bacc), set3 (P1 = SylW, P2 = SylW, P3 = SiOr, P4 = DomC, O = Bacc), and set4 (P1 = SylW, P2 = SylW, P3 = SiOr, P4 = DomD, O = Bacc) were used as the input for D_{FOIL} analysis (Data S4C). D_{FOIL} analysis was applied by the mvftool v0.6.2.1[92] 'CalcPatternCount' function. A total of 100 non-repeating individual sets were randomly generated, and the test was applied on 100 KB-sized windows along the genome. The total number of significant introgression windows in each set was then summarized. The Twisst analysis[28] was applied to two groups of five populations: Set DomD, comprising SiOr, DomD, SylE, SylW, and BAcc; and Set DomC, comprising SiOr, DomC, SylE, SylW, and BAcc. The analysis was constrained to a five-taxon topology (Data S4D). Phylogenetic trees were inferred for windows containing 100 SNPs using PhyML,[88] with the same parameters as above, and served as the input data. Chromosome-level topology weighting analysis was applied separately, constrained to a six-taxon topology (Data S4H) by Twisst.[28] Chromosome-level *D*-statistics were calculated for each chromosome using the Dtrios command in Dsuite in four population groups (SiOr-DomC-SylE-BAcc, SiOr-DomC-SylW-BAcc, SiOr-DomD-SylE-Bacc, and SiOr-DomD-SylW-Bacc). To identify introgression within specific genomic regions, we also employed the f_d statistic in 10-kb sliding windows using the ABBABABAwindows.py (https://github.com/simonhmartin/genomics_general/tree/master) on the same four population groups in Chromosome-level *D*-statistics analysis (Data S4G). We identified the potential introgressed regions as those in the top 5% of f_d values across four population groups and extracted the genes located in these regions. The list of genes was then subjected to KEGG enrichment analysis, and the significant KEGG pathways were visualized by R (see details in 'Gene function annotation and enrichment analysis' section).

Demographic modeling based on site frequency spectrum (SFS)

Based on the inferred topology from qpGraph, the most common ASTRAL phylogenetic tree, and the gene flow events identified using *D*-statistics and D_{FOIL}, fastsimcoal v2.8[31] was employed to further investigate the demographic history of apple populations. The analysis considered four possible gene flow scenarios: G0, where no gene flow occurs within populations; G1, involving gene flow between the ancestors of DomC/DomD and SiOr, as well as SylW/SylE populations; G2, where gene flow occurs between DomC/DomD and SylW/SylE populations; and G3, which combines both G1 and G2. Additionally, four divergence scenarios were also included: D1, where DomC and DomD both diverge from SiOr, with DomC diverging more recently and SylE diverging from SylW; D2, where DomD diverges from SiOr, followed by DomC diverging from DomD, and SylE diverging from SylW. This framework resulted in eight distinct scenarios being evaluated (Figure S4F).

A folded 2D site-frequency spectrum (SFS) was generated using 31,300 SNPs with the program easySFS (using the -a option).[112,113] Each scenario was optimized through 50 independent runs with specific settings, including 100,000 coalescent simulations per likelihood estimation (-n 100,000), 40 cycles of the conditional maximization algorithm (-L40), and a minimum of 10 observed SFS entries for likelihood computation (-C10). The maximum likelihood values obtained from these runs were used to determine the best-fitting scenario or scenario group. Once the most likely scenario or scenario group was obtained, the parameter median estimates from the 50 runs were calculated, and the 95 percent confidence intervals.

Detection of signatures of positive and balancing selection

Signatures of positive selective sweeps in the genomes of cultivated and wild apples were detected using OmegaPlus[93] and RAiSD,[94] which are linkage disequilibrium (LD) and composite methods, respectively, employing 28,377,551 SNPs (Data S4A). OmegaPlus computes the ω statistic, which can be used for synonymous and non-synonymous linked SNPs while considering the LD between SNPs. For the selective sweep analysis, OmegaPlus was applied to each chromosome individually per population with the parameters -minsnps 5, −minwin 5,000, −maxwin 100,000, and a grid size set to the number of chromosome lengths divided by 1000 bp (chosen given a median gene length of ~2,000 bp, to ensure coverage of all genes for detection). RAiSD calculates the μ statistic, a composite evaluation metric that scores genomic regions by quantifying local reduction in polymorphism, shifts in the SFS towards low- and high-frequency variants, and localized patterns of LD across the genome. We performed μ-statistics using 50 SNPs per chromosome per population. Soft sweeps were detected using the *G12* and *G123* statistics.[95] Positive selection can reduce genomic diversity at both the selected and linked sites, producing a characteristic signature of elevated expected haplotype homozygosity. These selective sweeps can be classified as either hard or soft. In a hard selective sweep, a single adaptive haplotype rises to high population frequency. In contrast, in a soft sweep, multiple haplotypes sweep through the population simultaneously, producing distinct patterns of genetic variation near the selected site. VCF files were converted to the format required by SelectionHapStats.[95] The *G12* and *G123* statistics[95] were calculated for each chromosome of each population using overlapping windows of 200 SNPs with a step size of 50 SNPs. The above threshold definitions for these positive selection signals are referenced in the "Control of demographic history for selection analyses" section.

Signatures of balancing selection were inferred using BetaScan2,[96] which identifies the enrichment of variants in regions with low deviation from allele frequency and a deficit of substitutions. VCF files were converted to the input format required by BetaScan2 using a custom script, excluding folded allele frequencies lower than 0.15 (-m 0.15). The β scores were calculated for all remaining SNPs using a window size of 1000 bp (-w 1000) for each SNP position. LD blocks were defined as regions with LD ($r^2 > 0.8$), computed using LDBlockShow v1.39.[114] As in the positive selection analysis, a neutral model was used to define the threshold for β scores.

Control of demographic history for selection analyses

The effect of population demographic history on SFS, LD, and genetic diversity across the genome was controlled for in all selection analyses presented above. Significance thresholds for deviations from the neutral model in selection analyses were assessed based on simulated datasets under the neutral demographic model. The population history (*Ne*) estimated by MSMC2[97,115] was converted into ms parameters using the msmc2ms.py script (https://github.com/stschiff/msmc-tools). A total of 10,000 datasets, each comprising five populations of equal numbers of pure individuals, as defined by fastSTRUCTURE analyses, were simulated using the ms program. Each dataset consisted of a 1 Mb chromosome fragment from the respective model parameters. Subsequently, 10,000 simulations of a 1 Mb chromosome under the neutral demographic model were conducted to assess the significance of ω, μ, G12/G123, and β statistics using the same parameters for the real genome. The top 1% values of ω, μ, G12/G123, and β were considered to deviate significantly from the neutral model simulations, thereby serving as thresholds for each selection analysis. The low mappability regions (see above) were excluded from the selection signal results.

Gene function annotation and enrichment analysis

Enrichment analyses of genes under selection were then performed. KEGG pathway annotations were obtained using eggNOG-mapper v2.1.8,[98] with the GDDH v1.1 genome gene protein file[18] as input. KEGG pathway names were downloaded from https://www.kegg.jp. Enrichment analysis was conducted using the ClusterProfiler v4.0 R package,[99] with significantly enriched KEGG pathways defined as those with a *q*-value < 0.05. Enrichment plots were generated using the Enrichplot v3.17 R package.[116] Flowering genes were identified by aligning the GDDH v1.1 protein sequence with the Flowering Interactive (FLOR-ID) database.[58] Genes were retained if they exhibited an e-value of less than 1e-5, matched bases of at least 50 base pairs, and an identity of 80% or higher. Transcription factors (TFs) were identified in the GDDH v1.1 genome using PlantTFDB (http://planttfdb.cbi.pku.edu.cn/)[117] by

Please cite this article in press as: Chen et al., Gene flow from the European wild apple and selection shaped the domesticated apple genome, Current Biology (2026), https://doi.org/10.1016/j.cub.2026.03.048

CellPress
OPEN ACCESS

HMMER3.0.[118]The TFs were classified according to the consensus rules, including the required and prohibited protein domains for each TF gene family, summarized on the PlantTFDB website. Resistance gene analogs (Rgene) based on differences in their domains identified by NLGenomeSweeper,[119] combine information from the GDDH v1.1 released annotation.[18]

RNA extraction, library preparation, and sequencing

Leaf samples were immediately flash-frozen in liquid nitrogen after harvesting and subsequently stored at –80°C. Samples were pulverized using a QIAGEN TissueLyser II (R2) with the addition of Polyvinylpyrrolidone (PVP40) to reduce interference from phenolic compounds. Total RNA was extracted using the Nucleospin RNA Plant Kit (Macherey-Nagel, Düren, Germany), following the manufacturer's protocol, which includes RNA binding, DNase treatment, and centrifugation steps.

RNA quality and quantity were initially assessed using a NanoDrop spectrophotometer, with purity determined via absorbance ratios at 260/280 nm and 260/230 nm. Samples with A260/A230 ratios between 1.8 and 2.2 were considered of poor quality and discarded. RNA integrity was further evaluated using a Bioanalyzer (Agilent Technologies, USA). High-quality RNA was used to construct a cDNA library using the Novogene NGS Stranded RNA Library Prep Set (PT044), with rRNA depletion. Sequencing was performed on a NovaSeq 6000 system (Illumina, San Diego, CA, USA) using the S4 Reagent Kit v1.5 (300 cycles), producing 150 bp paired-end reads targeting 20 million reads per sample.

RNA-sequenced data processing and differential expression analysis

Raw reads were assessed using FastQC v0.11.2,[73] followed by adapter trimming and quality filtering with fastp v0.23.2,[72] and ribosomal RNA filtering using SortMeRNA v4.3.2.[120] Cleaned reads were aligned to the *M. domestica* GDDH13 v1.1 reference genome[18] using STAR v2.7.10b,[121] and gene quantification was performed in alignment-based mode using Salmon v1.10.0.[122]

Differential gene expression analysis was conducted with the R-based DiCoExpress pipeline.[76] Read counts were first normalized by collection year and then merged. Low-expression genes were filtered using the "NbConditions" strategy with a CPM_cutoff of 5. TMM normalization was applied,[123] and differential expression was tested using the edgeR method[124] with an FDR threshold (Alpha_DiffAnalysis) of 0.05. Genes differentially expressed between cultivated (cider or dessert) and wild apples were cross-referenced with genes under positive selection in DomC and DomD to identify domestication-related candidate genes.

Genomic landscape of introgressions and adaptive introgressions

The genomic landscape of introgression was investigated as a first step to studying the adaptive introgression. As described in "Genome-wide estimates of gene flow among wild and cultivated populations", introgression regions across the genome were detected using the f_d statistic and topological weighting. A smoothing approach was then applied for topological weighting results, using a 20 Kb window size to refine the inferred topologies. The resulting topology weighting data were categorized into five distinct patterns (Data S4H): no introgression (one topology), DomC/DomD introgressed by the ancestor of SylE and SylW (one topology), SylE introgression (three topologies), SylW introgression (three topologies), and others (seven topologies).

To investigate adaptive introgression further, two new analyses were introduced. The Q95 statistic[39] was calculated in the form of $Q95_{SiOr,\ DomC/DomD,\ SylW/SylE\ (10\%,\ 100\%)}$, which pertains to the 95th quantile frequency of derived alleles in the DomC or DomD populations that have a derived allele frequency of 100% in a donor panel in the SylW and SylE individuals and are present at <10% frequency in the combined SiOr panel (Data S4L). Local ancestry was inferred using RFMix (v2.03)[38] with phased data from all DomC and DomD individuals (Data S4J and S4K). The BAcc, SiOr, SylW, and SylE populations, retaining only very pure individuals (membership coefficient > 0.99 in fastSTRUCTURE analysis), served as reference populations. A comprehensive analysis integrating results from four analytical methods—f_d, topology weighting, Q95 statistic, and local ancestry inference—was conducted to identify potential regions of adaptive introgression. For each selection region, f_d, introgression topology weighting, and SylW/SylE ancestry proportion were calculated. Regions exhibiting outlier values based on the interquartile range (IQR) method were subsequently extracted, with thresholds for extreme values defined as (Q3 + 1.5IQR) to identify outliers (Data S4M and S4N). The distribution of Q95 values exhibited a bimodal pattern, prompting using the k-means algorithm to categorize the signals into two clusters. To prevent overinterpretation of the Q95 statistic's efficacy, the cluster centroid value of the higher Q95 values was utilized as the threshold (Data S4N). Any selected region containing two or more signals (including signals of the same type but from different wild donor populations) was designated as an adaptive introgression area based on the four identified signals. Haplotype plots with dendrograms for candidate adaptive introgression regions were generated using Genotype Plot v0.2.1.[40]

Mutation load estimates

The constrained regions of the apple genome were identified using the GERP++ software.[43] First, genome sequences of eight species within the Rosales order (i.e., apple (*M. domestica*), pear (*Pyrus pyrifolia*), loquat (*Eriobotrya japonica*), Bowman's root (*Gillenia trifoliata*), almond (*Prunus dulcis*), peach (*Prunus persica*), strawberry (*Fragaria vesca*), and marijuana (*Cannabis sativa*))[18,125–131] were collected for multiple sequence alignments. Multiple sequence alignments were performed with LastZ v1.02.[132] A phylogenetic tree for the 4-fold conserved loci was then constructed based on the maximum likelihood method in IQ-TREE v2.1 software.[133] Finally, the number of "rejected substitutions" (GERP RS score) was estimated using the 'gerpcol' module from the GERP++ software.

Since no sites with GERP RS scores> 2 were found in the apple genome (see supplemental information above), we estimated genome-wide mutation load using the number of derived deleterious alleles identified in *Malus* accessions based on Sorting Intolerant From Tolerant 4G (SIFT 4G) scores.[46] We constructed a SIFT 4G database specific to the apple (*Malus domestica*) genome

using the SIFT4G_Create_Genomic_DB tool. The database was built based on the reference genome sequence (FASTA) and genome annotation (GFF3) files. To enhance the evolutionary conservation information, we incorporated the UniRef90 protein clustering database[134] as a source of homologous protein sequences during database construction. This integration provided a comprehensive set of homologs for multiple sequence alignment, thereby improving the accuracy of functional impact predictions. Following database construction, we employed the SIFT4G_Annotator tool to annotate genetic variants (in VCF format) with SIFT scores. Only variants located within coding sequences (CDS) were retained for downstream analyses. The proportion of missing genotypes per individual was low, with an average of 2.4% and a median of 2.1%. We then polarized derived and ancestral alleles using *M. baccata* and *P. pyrifolia* as outgroups. The software est-sfs v2.04[100] was used to estimate the ancestral alleles of *M. domestica, M. sieversii*, *M. orientalis*, and *M. sylvestris* individuals. First, short-read sequences from five *M. baccata* (outgroup 1) and five *P. pyrifolia* (outgroup 2) were mapped to the GDDH v1.1 apple reference genome.[18] Alleles with a frequency greater than 0.8 in the outgroup were considered as putatively ancestral, as high frequency in these related species likely reflects the ancestral state. The Kimura two-parameter model[135] was then applied to estimate ancestral allele probabilities, and only those with an ancestral site probability greater than 0.9 were considered. Candidate deleterious mutations were screened based on the SIFT4G results, with a threshold set at a "SIFT score < 0.05" of any nonsynonymous position.[136] We observed a reference bias, where sites with the derived allele in the reference apple genome were more often categorized as tolerated *versus* sites where the reference was ancestral, similar to findings in human and soybean genomes (Data S4O). We adjusted the derived deleterious allele count at sites where the reference genome contained the derived allele using the probability estimation method described by Simons.[137] The mutation load for each accession was quantified by counting the ratio of derived deleterious to synonymous mutations at the individual level (Data S4O).

To examine the relationship between gene flow or selection sweep and mutation load. The mutation load of each gene was quantified as the average number of derived deleterious alleles per 1,000 bp of the CDS region in DomC or DomD populations, which standardizes for gene length and facilitates comparison among genes. The introgression level (f_d) was calculated for each gene in the four population groups (SiOr-DomC-SylE-BAcc, SiOr-DomC-SylW-BAcc, SiOr-DomD-SylE-BAcc, and SiOr-DomD-SylW-BAcc). The deleterious allele distribution per gene was fitted to a negative binomial distribution, and the coefficients with f_d were estimated by the *glm.nb* function from the MASS package.[101] Genes from various selection regions, including soft or hard sweeps and adaptive introgression or non-introgression, were extracted, while the remaining genes function used as controls. The statistical significance of the difference between the selection genes and the control genes was assessed using a Wilcoxon test.

Genome-wide association study (GWAS) of flowering time in cultivated apples

To investigate phenotypic divergence between cider and dessert apples, we conducted a GWAS for flowering time using a subset of 30 cultivated accessions, each with six replicates in a randomized design measured across three common gardens over three years, planted in 2021. We analyzed genotype–phenotype associations using the GAPIT R package (v3.0)[47] implementing the BLINK method,[48] with kinship (K) and population structure (Q) as covariates. SNPs were filtered for minor allele frequency (MAF $\geq$ 0.05) and missing rate < 20% prior to analysis. Significance thresholds were determined using both false discovery rate (FDR < 0.05) and Bonferroni correction.[138] For significant SNPs, we examined average LD decay distance (10 Kb) nearby genes and transposable elements using the GDDH13 v1.1 genome gene annotation and TE annotations.[18] To assess population-specific differentiation at these loci, we visualized genotype distributions using custom scripts that generated barplots of allele frequencies by individual and genetic group (DomC, DomD, SiOr, SylE, SylW) based on fastSTRUCTURE assignments, excluding individuals with >20% missing data or admixture coefficients < 0.8. We assessed the distribution of flowering-associated alleles across selective sweep and introgressed regions in the DomC and DomD populations.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were conducted in R (v4.4.1). Unless otherwise specified, all statistical tests were two-sided.

Population genetic statistics, including nucleotide diversity (π), genetic differentiation (F_{ST}), divergence (D_{XY}), and linkage disequilibrium decay, were calculated using standard estimators.

The significance of *D*-statistics and related admixture tests was assessed using block jackknife resampling across genomic blocks. Neutral expectations for selection scans were generated under demographic models inferred from coalescent analysis. GWAS were performed using a mixed linear model controlling for kinship and population structure. Significant SNPs were determined using both Bonferroni correction (p_{adj} < 0.05) and false discovery rate (FDR) control (FDR < 0.05). For adaptive introgression analyses, outlier regions were defined using interquartile range (IQR) criteria (outlier regions defined as Q3 + 1.5 * IQR) or clustering-based thresholds where bimodal distributions were observed. Regions supported by multiple independent statistics were considered high-confidence candidates. Differential gene expression significance was determined using adjusted *p-value*s (FDR < 0.05) from the statistical framework applied in the RNA-seq pipeline. Gene functional enrichment analyses were evaluated using hypergeometric tests with FDR correction (FDR < 0.05). Mutation load comparisons and associations between introgression, selective sweeps, and deleterious variation were evaluated using Wilcoxon rank-sum tests and generalized linear models where appropriate.