

Genomic architecture of the self-incompatibility locus in apple provides insights into the evolution of collaborative non-self recognition

Aurélie Mesnil^{1,2}, Vincent Castric³, Gianluca Lombardi⁴, Marie Monniaux³, Somia Saidi^{1,5}, Xilong Chen^{1,2}, William Marandé⁶, Johann Confais⁵, Anne-Laure Fuchs¹, Enrique Dapena⁷, Anthony Venon^{1,2}, Takashi Tsuchimatsu⁸, Alessandra Carbone^{4,9}, Xavier Vekemans³, Amandine M. Cornille^{1,2}

1 Université Paris Saclay, INRAE, CNRS, AgroParisTech, GQE-Le Moulon, Gif-sur-Yvette, France

2 Division of Science, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

3. Univ. Lille, CNRS, UMR 8198 – Evo-Eco-Paleo, F-59000 Lille, France

4 Sorbonne University, CNRS, IBPS, Computational, Quantitative and Synthetic Biology Department, UMR 7238, Paris, France

5 Université Paris-Saclay, INRAE, BioinfOmics, URGI, 78026 Versailles, France

6 CNRGV, INRAE Toulouse, France

7 SERIDA, Villaviciosa, Spain

8 Graduate School of Science, The University of Tokyo, Japan

9 Institut Universitaire de France

Abstract

Self-incompatibility (SI) systems prevent self-fertilization, thereby maintaining genetic diversity in flowering plants. Among them, collaborative non-self recognition (CNSR) is the most widespread, yet its genomic architecture remains poorly understood. Using 27 haplotype-resolved genomes from wild and domesticated apples (*Malus* spp.), we dissected the structure and evolution of the S-locus. We identified 17 S-RNase alleles and 500 pollen-expressed S-locus F-box brother (SFBB) genes across 18 families. The S-locus exhibits hallmark features of a supergene, including suppressed recombination, TE accumulation, and an excess of non-synonymous polymorphisms. Nonetheless, long-term balancing selection preserves both the integrity and diversity of entire S-haplotypes. Recurrent associations between SFBB families and transposable elements, coupled with gene conversion signatures, suggest a dynamic mechanism for reshuffling recognition specificities. Our results provide a detailed view of how a multigenic, tightly linked recognition system is maintained in a woody perennial, and highlight the

interplay between structural constraints and selection in sustaining functional diversity and reproductive barriers.

Main

Self-incompatibility (SI) prevents self-fertilization and enforces outcrossing in many flowering plants, thereby maintaining genetic diversity and long-term adaptation^{1–3}. In the Maleae tribe of the Rosaceae family, containing *Malus* species (the apple trees), SI is primarily controlled by a gametophytic self-incompatibility locus (S-locus) that encodes the S-RNase, acting as a pistil-side toxin, and a multigene family of S-locus F-box brothers (SFBBs), which serve as pollen-side anti-toxins^{4,5}. This collaborative non-self recognition (CNSR) system necessitates multiple SFBB proteins to recognize and neutralize non-self S-RNases, allowing successful fertilization under outcrossing^{6,7}. While empirical investigations into this mechanism have advanced in model species from other flowering plant families, such as *Petunia* (Solanaceae)⁸ and *Antirrhinum* (Plantaginaceae)⁹, the complex genomic organization and evolutionary dynamics of the CNSR system of Maleae within Rosaceae remain incompletely understood.

Building on this framework, previous studies indicate that the S-locus may develop features of a supergene, such as structural rearrangements, recombination suppression, and long-term balancing selection that maintain co-adapted gene complexes^{10–12}. However, the extent to which these features apply to SI loci remains poorly documented. In particular, the mechanisms behind SFBB expansion, diversification, and maintenance—including tandem duplications, gene conversion, transposable elements (TEs), and selective pressures—are still unclear^{13,14}. Theoretical investigations suggest that natural selection will favor S-haplotypes that are as "complete" as possible, meaning that they possess pollen anti-toxins covering all S-RNase alleles found in the population, except the one carried by its own haplotype^{15,16}. Hence, it is expected that genomic processes will tend to reach this optimal. However, some recent realistic simulations suggest that the dynamic nature of these processes could lead to substantial variation in the anti-toxin composition of S-haplotypes encoding the same S-RNase¹⁷. Here, we present the first haplotype-resolved characterization of the S-locus in *Malus*, leveraging 27 high-quality phased S-haplotypes from both wild and cultivated apple genomes. We annotated 500 SFBB genes, revealing 18 distinct phylogenetic families, and examined their patterns of duplication, diversification, and co-association with S-RNases. Using structural modeling, codon-level selection analysis, and genome-wide TE mapping, we reveal that tandem duplications, positive selection, and

gene conversion — possibly involving nearby TEs — contribute to SFBB turnover across haplotypes. Despite extensive divergence, S-haplotypes preserve conserved *S*-RNase–SFBB combinations, consistent with a non-self recognition model under balancing selection.

Our findings highlight the *Malus* S-locus as a compelling model to study how genomic structure and evolutionary processes interact to maintain multigenic recognition systems. They also raise broader questions about the genomic limits of diversification and recognition capacity in long-lived perennials under strong selective pressures.

Results

Conserved genomic architecture of SI loci with the same S-allele

To investigate the genome architecture and evolution of the S-locus in *Malus*, we analyzed 27 fully assembled S-haplotypes (Supplementary Table 1, Supplementary Fig. 1), including 14 from *M. domestica* (12 dessert, two cider) and 13 from wild relatives (*M. sylvestris*, *M. sieversii*, *M. orientalis*). At the S-locus located in chromosome 17, each haplotype contained a single S-RNase gene, which grouped into 17 clearly distinct allelic lineages (Supplementary Table 3, Fig. S1A). A 100% amino acid identity was typically observed between gene copies within alleles (Supplementary Fig. 1B), indicating strong purifying selection. One exception, haplotype h32 from *M. orientalis*, exhibited only 66.9% similarity with two MdS9 allele sequences from *M. domestica*. Controlled pollinations are necessary to confirm phenotypic recognition equivalence. Several wild haplotypes carried previously uncharacterized alleles (h24, h25, h29, h31). The *S*-RNase phylogeny (Supplementary Fig. 2a) was incongruent with the species tree (based on chromosome 17 single-copy genes), showing extensive interspecific allele sharing. This pattern is consistent with ancestral trans-species polymorphism, which is expected for an S-locus subject to long-term balancing selection¹⁸, or with adaptive introgression between species^{19,20}.

We identified 500 intron-less SFBB genes (17–20 per haplotype; ~1.2 kb each), carrying a 5' F-box domain and a 3' FBA domain (Supplementary Text 1, Supplementary Table 6). These gene sequences clustered into 18 phylogenetic families, labeled SFBB1 to SFBB18 (Supplementary Text 1, Supplementary Table 6, Supplementary Fig. 3). We compared the presence/absence patterns of each SFBB family across different S-haplotypes, defined by the evolutionary tree of *S*-RNase genes (Fig. 1a, b). Each S-haplotype had a specific combination of SFBB families, which was strongly associated with its *S*-RNase allele (Mantel test, *p*-value = 0.024, Fig. 1c). For example, the *M. orientalis* h32 SFBB repertoire was nearly identical to that of the two *M. domestica* MdS9 haplotypes, except for a duplication

of the SFBB3 family. Moreover, the SFBB gene order was highly conserved within haplotypes carrying the same *S*-RNase (Fig. 1c), with some minor differences (e.g., h3 vs. other *MdS2* haplotypes) that may reflect assembly quality (Supplementary Fig. 4, see quality assembly index (LAI)). Large syntenic blocks with low overall structural divergence were observed between haplotypes sharing the same allele even when they belonged to different species (Fig. 2a), suggesting strong constraints on SFBB organization.

In contrast, haplotypes with different *S*-RNases exhibited more divergent architectures (Fig. 1d), although some gene blocks (e.g., SFBB4 to SFBB6 and SFBB5 to SFBB14) were consistently located at the 5' and 3' ends, respectively. Central regions were more structurally variable and harbored more inversions and unaligned segments (Fig. 2b, Supplementary Fig. 5). Haplotypes sharing *S*-RNases had significantly longer syntenic blocks ($p < 2.2e-16$). At the same time, those with different alleles had larger inversion sizes ($p = 1.3e-4$), supporting the role of structural variation in maintaining *S*-haplotype divergence.

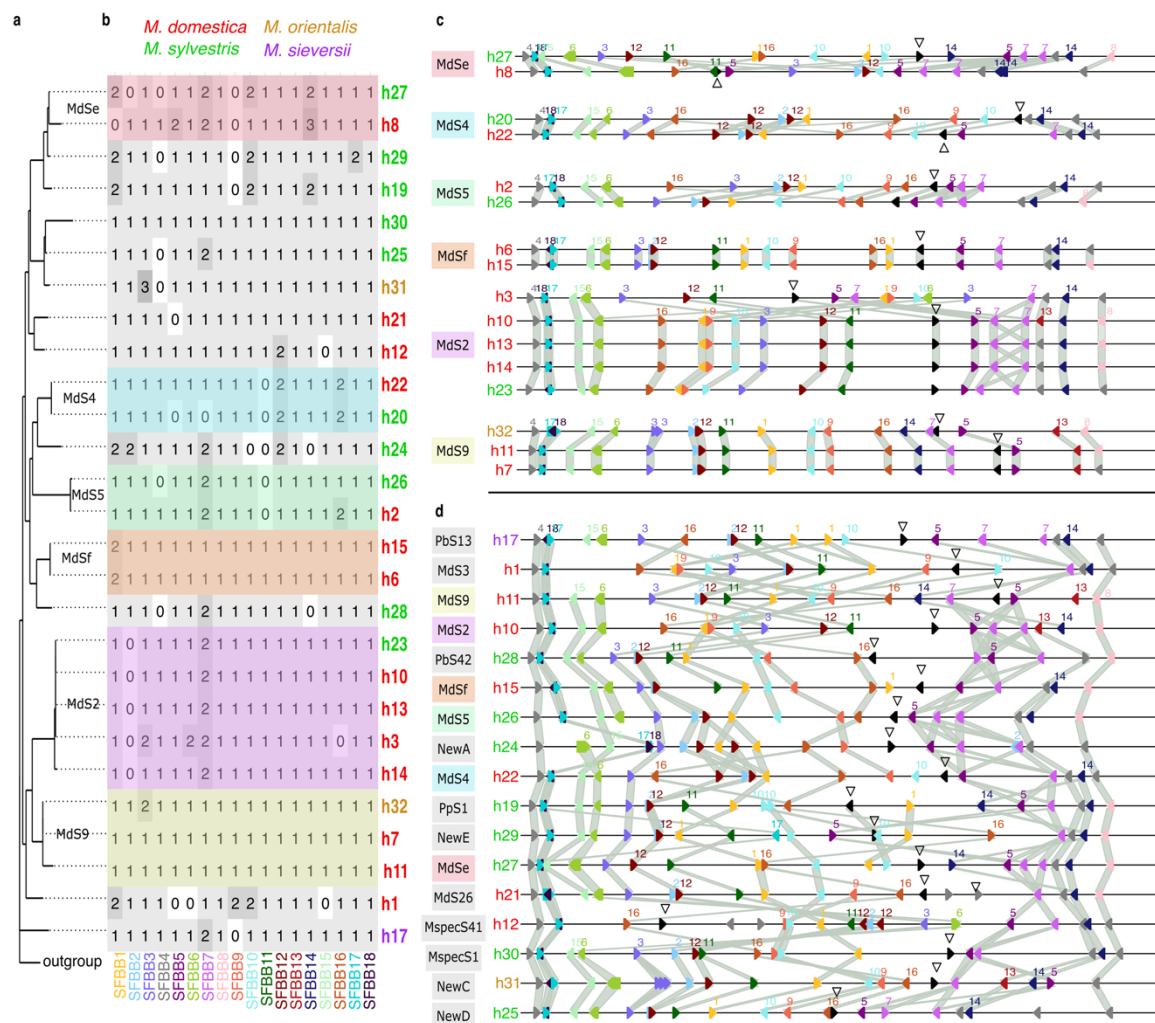


Fig. 1: Evolution, diversity, and organization of the S-locus in wild and cultivated apples.

(a) Phylogenetic tree of S-RNase protein sequences across 27 S-haplotypes (labeled h1–h30).

(b) Heatmap showing SFBB family counts per S-haplotype (rows) and gene family (columns).

(c) Gene organization of SFBBs (colored by family) and S-RNase (black arrowhead) within the S-locus across haplotypes sharing the same S-RNase allele. Gray ribbons link orthologous SI genes.

(d) S-locus structure across haplotypes with different S-RNase alleles; locus lengths are normalized to h30 for comparison.

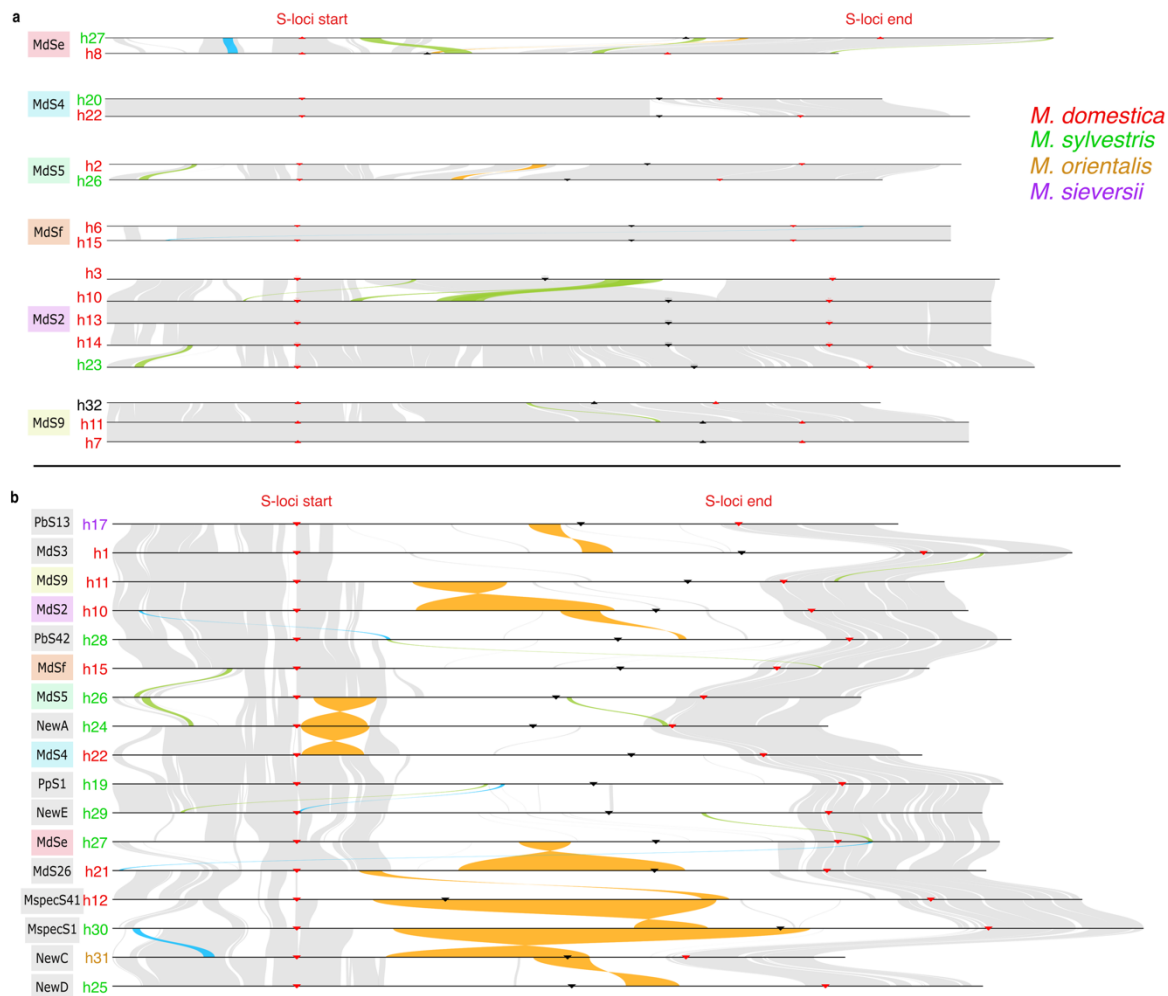


Fig. 2: Structural variation at the *Malus* S-locus.

(a) Whole-locus alignments of haplotypes sharing the same S-RNase (as in Fig. 2c) showing conserved synteny and structural rearrangements detected by SyRI (Synteny and Rearrangement Identifier). Colored curves represent syntenic blocks and rearrangements.

(b) Whole-locus alignments of haplotypes carrying different S-RNase alleles (as in Fig. 1d), highlighting greater structural divergence. Structural variant contributions are quantified in Supplementary Fig. 5. Alignments extend 500 kb upstream and downstream of the S-locus. Red triangles mark locus boundaries; black triangles mark S-RNase gene positions.

Each S-haplotype includes one S-RNase and several SFBB genes, which are key parts of the SI system in the *Malae* tribe⁵. We also identified the limits of the S-locus using conserved flanking genes,

including an upstream NF-Y gene and two downstream PI3/PI4 kinase family genes. The locus averaged 1.35 Mb (± 0.11). It contained about 42 annotated genes, approximately three times fewer than observed in matched control regions (Supplementary Fig. 6). Consequently, it was highly enriched in TEs, particularly LTRs ($\sim 85\%$ vs. 70% genome-wide, Supplementary Fig. 7a,c, Wilcoxon-Mann-Whitney test, p -value = $2.9e-10$). *S*-locus annotated genes were grouped into 55 orthologous clusters using Syntenet²¹, classified into four categories (Supplementary Fig. 8, Supplementary Table 7). Apart from SFBB and *S*-RNase genes, 38 clusters encoded diverse functions (e.g., kinases, transcription factors, transporters), suggesting possible recruitment or retention due to structural dynamics or regulation. Gene content clustering across haplotypes revealed both conserved and variable clusters (Supplementary Fig. 2b), with high TE content and variable gene presence, reinforcing the view of the *S*-locus as a structurally dynamic region.

Genetic diversity (π)^{22,23} computed in sliding windows peaked in the *S*-locus region in all four species, especially in *M. sylvestris* and *M. sieversii*, and was significantly enriched in the top 1% π windows (empirical $p < 0.05$; Supplementary Fig. 9), supporting long-term balancing selection. To avoid confounding effects of positive selection acting on SFBB genes, variants located within SFBB coding regions were excluded from subsequent analyses of coding variation. Even after this filtering, we observed an excess of non-synonymous substitutions ($nS/S > 1$; Supplementary Fig. 10), consistent with supergene-like behavior, where recombination suppression reduces purifying selection and promotes mutation accumulation²⁴.

Finally, comparing across species, with different CNSR systems, the number of SFBB genes and families in *Malus* (~ 18) was similar to that in *Petunia* (~ 18 SLFs)¹³ but lower than in *Antirrhinum* (~ 32 SLFs)⁹ (Table 1).

Table 1. Comparative architecture of the S-locus across flowering plants with CNSR SI systems.

Summary of key genomic features of the S-locus across species with collaborative non-self recognition (CNSR)-based self-incompatibility. Shown are the number of pollen-expressed genes (SFBBs or SLFs), S-locus size (Mb), number of gene families (from sequence/function-based clustering), chromosomal position (when known), and references. Data are drawn from published studies and the present work (*Malus*). In species like *Petunia* and *Citrus*, the S-locus was inferred from BACs or partial assemblies.

Model	Plant family	Number of pollen genes	Mean size of the S-locus (Mb)	Number of pollen gene families	Location	References
Apple (<i>Malus</i>)	Rosaceae	16–20 SFBBs	1.35	18	Peri-telomeric, long arm of chromosome 17 (17 chromosomes)	This study
Snapdragon (<i>Antirrhinum</i>)	Plantaginaceae	32 SLF/SFB	1.20	31	Peri-telomeric, short arm of chromosome 8 (8 chromosomes)	^{9,25}
Potatoes	Solanaceae	12 SLFs	13	8	Chromosome 1 (12 chromosomes)	²⁶
<i>Petunia</i>	Solanaceae	16–20 SLFs	Unknown	18	Sub-centromeric, chromosome 3 (7 chromosomes)	^{13,27}
<i>Coffea</i>	Rubiaceae	Unknown	0.008	Unknown	Chromosome 5 (11 chromosomes)	²⁸
<i>Citrus</i>	Rutaceae	11–17 SLFs	0.198 to 0.370 (BAC)	12	Unknown	^{14,29}
<i>Pyrus</i> and <i>Malus</i>	Rosaceae	17–21 SFBBs	1.35 to 2.64	22	Peri-telomeric, long arm of chromosome 17 (17 chromosomes)	³⁰

Pineapple (<i>Ananas</i>)	Bromeliaceae (Monocot)	Unknown	Unknown	Unknown	LG02 (25 chromosomes)	³¹
--------------------------------	---------------------------	---------	---------	---------	--------------------------	---------------

157

Multiple evolutionary and genomic mechanisms contribute to *SFBB* gene diversification and maintenance within the *Malus* S locus

A major question regarding S-locus architecture is how new recognition specificities are generated and maintained with multiple pollen determinant genes and how they spread across S-haplotypes with different *S-RNase* alleles^{15–17}. We obtained several lines of evidence suggesting that multiple evolutionary forces shape the diversity of *SFBB* genes within the *Malus* S-locus. Together, our results support a model where tandem duplication, gene conversion, and positive selection promote *SFBB* diversification (Fig. 3f), while structural and functional constraints maintain recognition compatibility with resident S-RNase alleles, preserving a diversified but tightly co-adapted SI system in *Malus*.

First, tandem and proximal duplications were found to be the dominant modes contributing to *SFBB* expansion (Fig. 3a), contrasting with segmental duplications that dominated elsewhere in chromosome 17. These local duplications likely provided the raw material for functional diversification of S-alleles. Secondly, signatures of positive selection ($dN/dS > 1$, Supplementary Table 8) were detected in *SFBB*s, with CodeML identifying codons under strong positive selection (Fig. 3b, Supplementary Table 9). Several of these residues were conserved across multiple S-haplotypes and were significantly associated with specific *SFBB* families (Table 2, Supplementary Table 10). These amino acids (aa) clustered in regions of predicted surface interaction based on Boltz-1 structural modeling³² (Supplementary Fig. 13) and exhibited high MuLAN local attention scores, suggesting functional relevance³³ (Supplementary Fig. 12). Multivariate analysis of aa positions across families supported this pattern of divergence (Supplementary Fig. 11).

In parallel, we observed strong sequence conservation within *SFBB* families despite their occurrence on divergent S-haplotypes (Fig. 3d, Supplementary Table 11), consistent with inter-haplotype gene conversion. A quasi-Poisson model identified allele pairs with significantly elevated conversion rates (MdS2–MspecS1, MdS2–MdSe, MdS2–MdSf), independent of their phylogenetic distance (Supplementary Fig. 14). Gene conversion events were also inferred from the recurrent association of specific *SFBB* families with nearby TEs, within 4 kb (Fig. 3e, Supplementary Fig. 15, Supplementary Table 12). These *SFBB*–TE pairs were significantly more frequent than expected under random association models (permutation test, $p \leq 0.01$) and were observed across multiple S-haplotypes carrying distinct S-RNase alleles. The TE consensus sequences enriched for such associations do not belong to any specific classification (Supplementary Table 13), but all of these consensus sequences correspond to incomplete copies, with match on RepBase³⁴. Such conserved *SFBB*–TE co-occurrences, along with similar distances to TE consensus and similar insertion ages (Supplementary Fig. 15a–b),

190 support the hypothesis of sequence block transfers across haplotypes, via gene conversion. Importantly,
191 SFBB–TE associations remained correlated with specific S-RNase alleles, indicating that only
192 compatible combinations were retained—suggesting strong post-conversion selection filters.
193

Table 2 | Amino acid sites associated with SFBB gene families in Malus S-haplotypes.

Summary of amino acid (aa) positions showing significant associations with SFBB gene families across 27 phased S-haplotypes. For each position, the number of aa variants, mean intra-family diversity, mean number of families per site, and FDR-adjusted p-values from multinomial regression are shown. FBA: F-box-associated domain. Positively associated sites highlight candidates for SFBB diversification and allele-specific recognition. Full data in Supplementary Tables 10 and 13.

Amino acid position	SFBB domain	Total number of aa variants	aa diversity within families	Family diversity per aa	<i>p</i> -value	Adjusted <i>p</i> -value
297	–	10	1.33	2.40	1.1e–244	5.7e–244
299	–	10	1.44	2.60	3.0e–188	4.2e–188
336	FBA	11	2.22	3.64	5.0e–98	5.0e–98
356	FBA	7	1.50	3.86	4.9e–215	1.6e–214
358	FBA	11	1.28	2.09	2.2e–263	2.2e–262
421	FBA	11	1.83	3.00	9.0e–191	1.5e–190
429	FBA	10	1.39	2.50	3.5e–199	7.0e–199
476	FBA	7	1.39	3.57	3.8e–203	9. 5e–203
543	–	12	2.00	3.00	8.1e–175	1.0e–174
549	–	8	1.61	3.63	3.6e–166	4.0e–166

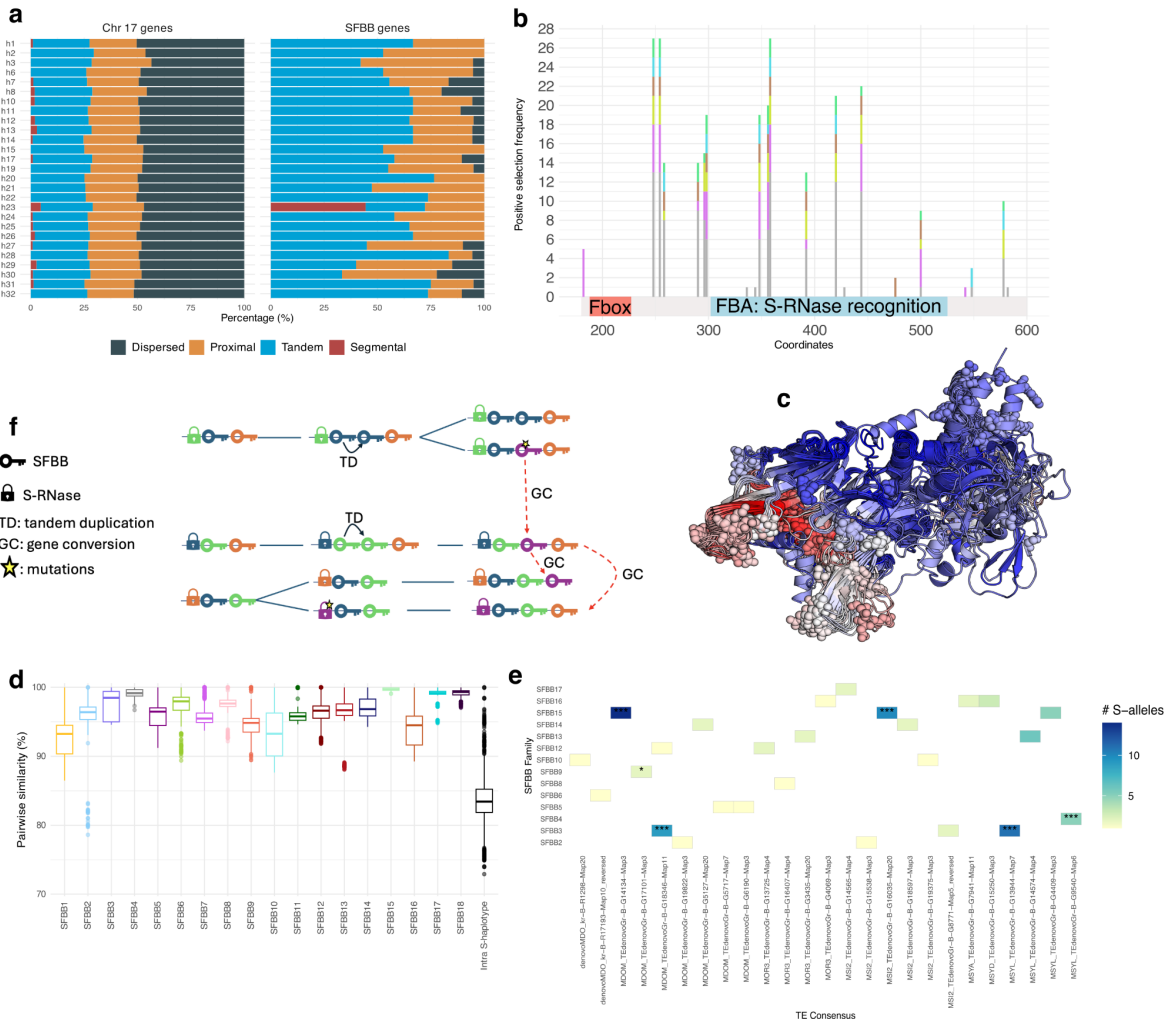


Fig. 3. Evolutionary forces shaping SFBB diversity at the *Malus* S-locus.

Central diagram illustrates the interplay of duplication, selection, gene conversion, and transposable elements (TEs) in SFBB diversification.

(a) Duplication modes for SFBB genes across haplotypes, compared with other chromosome 17 genes. Tandem duplications dominate SFBB expansion.

(b) Positively selected codons in SFBB proteins, primarily in the FBA domain but also in the interdomain region. Bar colors reflect associated S-RNase alleles; grey indicates unassigned alleles.

(c) In predicted SFBB model 3D structures (Boltz-1), selected residues cluster near S-RNase interaction surfaces. Superimposed SFBB models are colored with MuLAN attention scores; red-to-white colors correspond to high while blue tones to low scores. High attention scores indicate functional

importance. Note that the longest sequence for each *S*-allele was chosen as representative, totaling 17 sequences, and the resulting structures were aligned in PyMOL to minimize pairwise root mean square deviation of atomic positions.

(d) Sequence similarity of SFBB genes (Clustal Omega), comparing within-family (>90%) versus within-haplotype (~82%) identity.

(e) Heatmap of SFBB–TE co-occurrence across haplotypes. Colors indicate the number of distinct *S*-RNase alleles per association; asterisks denote significant associations (permutation test, 10,000 reps; * $p \leq 0.05$, *** $p \leq 0.01$). See Supplementary Table 14.

Predicting molecular interactions between SFBB and *S*-RNase

S-RNase polymorphisms drive SI specificity and show pronounced hypervariability (Supplementary Fig. 16). Codon-based analyses identified eight positively selected codons conserved within each *S*-RNase allele and distributed across the coding region, not confined to hypervariable segments. Several sites lie outside previously predicted recognition regions^{35,36}, yet 3D models (Boltz-1; Supplementary Fig. 17) indicated surface exposure. MuLAN attention scores were generally low at these positions (Supplementary Fig. 18), suggesting sequence-based models alone incompletely capture functional relevance.

To associate specific SFBB families with their target *S*-RNase(s) pollen determinants, we examined SFBB families consistently absent in haplotypes sharing the same *S*-RNase. SFBB2 was absent from all MdS2 haplotypes (h14, h13, h3, h10, h23, Supplementary Table 13), suggesting its role in detoxifying MdS2 *S*-RNase. SFBB11 was absent from both MdS4 (h20, h22) and MdS5 (h2, h26) haplotypes (Supplementary Table 14), suggesting putative detoxification of multiple targets, as already demonstrated for some SLF families in *Petunia*⁸. Family-specific amino acid deviations at conserved positions also linked SFBB genes to particular *S*-alleles—for example, SFBB12 in MdS2 haplotypes consistently showed a non-consensus residue at position 299 (Supplementary Table 10). SFBB18 was universally present and unmutated at selected sites, suggesting a role in recognizing unrepresented *S*-RNases that were not represented in our study.

Despite overall high confidence scores from Boltz-1 predictions of SFBB–*S*-RNase interactions (pLDDT, pTM, i-pLDDT, i-pTM; pLDDT > 0.6, pTM > 0.3; Supplementary Fig. 19), the results did not allow discrimination between expected and unexpected interactions. Nevertheless, predictions for SFBB proteins alone (i.e., outside the docking context) showed slightly higher confidence values, suggesting that the SFBB structural models themselves were more reliable than the SFBB–*S*-RNase

interaction models. By contrast, the same pipeline applied to *Petunia*—with validated SLF–S-RNase interactions⁸ —produced higher confidence scores (pLDDT > 0.8, i-pLDDT > 0.65, pTM > 0.7; Supplementary Fig. 20) and recovered known interacting pairs. In *Petunia*, positively selected sites mapped to surface regions overlapping high MuLAN scores (Supplementary Fig. 21–22).

Discussion

Higher rates of outcrossing in hermaphrodite plant species foster genetic diversity and adaptive potential in plants^{1–3}, leading to higher macroevolutionary success³⁷. In *Malus*, this is enforced by the S-locus, where S-RNases and SFBB proteins mediate SI. Despite its importance, the genomic organization and evolution of this locus have remained elusive due to the limited availability of reliable haplotype data. Leveraging 27 fully resolved S-haplotypes, we present a comprehensive view of S-locus diversity, highlighting extensive structural variation, dynamics of SFBB gene content, and signatures of evolutionary forces acting on this system.

Architecture and expansion of the *Malus* S-locus

We confirmed the peri-telomeric position of the S-locus on chromosome 17 and estimated its size at ~1.35 Mb, consistent with prior work in *Malus* and *Pyrus*³⁰. Phylogenetic analysis assigned the 27 haplotypes to 17 distinct S-RNase alleles, with considerable redundancy in MdS2 and MdS9. On average, 18.5 SFBB genes were annotated per haplotype, organized into 18 phylogenetic families. Gene order and composition were strongly conserved among haplotypes sharing the same S-RNase, but diverged among different S-RNase haplotypes—suggesting a coupling between S-RNase identity and SFBB gene repertoire, as expected in CNSR systems^{6,17}.

The *Malus* S-locus is markedly smaller than its homologs in other CNSR species (e.g., >13 Mb in *Solanum tuberosum*²⁶ or >3 Mb in *Antirrhinum*⁹), yet it shows signs of substantial local expansion compared to *Gillenia trifoliata*, a Rosaceae species with a 55-kb non-self SI region³⁸. The microsynteny detected between an SFBB-like gene from *Gillenia* and a *Malus* SFBB gene, as well as between a nearby transposable element, suggests a shared ancestral origin between the two loci³⁸. The *Gillenia* region is non-functional, and its smaller size could reflect progressive genomic degeneration in *Gillenia* and/or local duplication events in *Malus*, followed by functional diversification. Together, these processes likely underlie the evolution of the complex multigenic recognition system observed in *Malus*.

Constraints on gene copy number and recognition repertoire

Across CNSR systems, the number of SFBB or SLF genes per haplotype roughly matches the number of distinct gene families: 18 in *Malus*, ~32 in *Antirrhinum*⁹, and 18–20 in *Petunia*¹³. This near one-to-one correspondence may reflect an evolutionary constraint whereby each S-haplotype maintains the whole set of anti-toxins, with the exception of the one targeting its own S-RNase. Indeed, theory predicts that maintaining just enough SFBBs to cover all non-self S-RNases of the population (i.e., reaching maximal "completeness") minimizes fitness costs associated with redundancy and maximizes reproductive success¹⁵.

In our data, some SFBBs, such as SFBB11, exhibited signs of broader recognition. This suggests selection may favor the evolution of multi-target SFBBs to limit the expansion of the locus, while assuring detoxification of all extant S-alleles. Indeed, in *M. domestica*, more than 30 distinct S-alleles have been reported³⁹, while the number of SFBB is limited to 18. Such optimization likely reflects trade-offs between specificity, redundancy, and genome size constraints imposed by recombination suppression at the S-locus. It is not clear whether differences in the total number of SFBB/SLF genes between *Malus* (18) and *Antirrhinum* (32) are due to differences in the average degree of promiscuous interactions in toxin recognition, or whether it is due to differences in number of S-alleles maintained at the species level, which is still unknown in *Antirrhinum*.

Gene duplication, selection, and exchange

SFBB expansion in *Malus* primarily involved tandem and proximal duplications, in contrast to the dispersed duplication patterns seen elsewhere on chromosome 17. These duplications likely facilitated neofunctionalization. Signals of positive selection—particularly in the substrate-binding FBA domain—suggest adaptive diversification to expand recognition specificity.

We also detected evidence of gene conversion and interallelic exchange: high sequence similarity across SFBB families, conserved SFBB–TE pairings, and recurrent TE–SFBB associations across divergent haplotypes. These observations suggest that gene conversion and recombination have helped propagate functional motifs while preserving co-adapted combinations under purifying selection.

While TEs likely did not mediate gene conversion directly, their conserved proximity to SFBBs and the dominance of TIR/MITE elements suggest a role in duplicative reshuffling and S-locus evolution.

Supergene features and recombination suppression

The *Malus* S-locus bears hallmark features of a supergene: tightly linked co-adapted genes, suppression of recombination, TE accumulation, and signs of reduced purifying selection. Haplotypes sharing the same S-RNase allele showed greater collinearity and structural conservation than those carrying different alleles, consistent with recombination barriers that preserve functional S-RNase-SFBB modules.

The buildup of TEs and excess nonsynonymous substitutions suggests reduced selection efficacy due to Hill–Robertson interference⁴⁰, a feature common in other supergenes^{10,11}. Long-term balancing selection, which preserves diverse S-haplotypes over time, coexists with this degenerative context, highlighting the evolutionary tension between maintaining diversity and the mutational burden.

Limits of current interaction models

Using Boltz-1 and MuLAN, we attempted to predict SFBB–S-RNase interactions but failed to distinguish interacting versus non-interacting pairs in *Malus* clearly. In contrast, the same approach successfully recovered validated interactions in *Petunia*⁸. This suggests that SI recognition in *Malus* may involve more complex or subtle molecular interfaces—potentially including conformational dynamics, multimeric assemblies, or regulatory layers (e.g., post-translational modifications or lncRNA-based modulation, as in *Fragaria*)⁴¹.

These findings underscore the limitations of current *in silico* models for decoding the molecular logic of the non-self recognition SI system of Maleae. Functional studies—such as expression assays, CRISPR mutagenesis, or *in vitro* binding experiments—will be essential for validating SFBB–S-RNase interaction predictions and refining models.

Outlook

Our haplotype-resolved characterization reveals the *Malus* S-locus as a dynamic, multigenic, and supergene-like system shaped by local duplications, gene conversion, TE activity, and strong balancing selection. Beyond offering mechanistic insights into SI maintenance, our results establish *Malus* as a model for investigating how genome architecture, selection, and recombination jointly shape complex recognition systems in plants. Future population-scale and structural variant analyses will clarify how SI evolves and breaks down across wild and cultivated *Malus* lineages.

Methods

Dataset description

Twenty-seven assembled haplotypes (Supplementary Table 1) from *Malus* species, including both cultivated apple (*M. domestica*, 16 genotypes, comprising 14 dessert varieties and two cider varieties) and wild relatives (11 genotypes from *M. sylvestris*, one *M. sieversii*, and two *M. orientalis*) were examined in this study. Among the individuals used to produce this dataset, 17 were diploids. In most cases, two haplotype assemblies were obtained per individual (Gala and Fuji cultivars, the cider individual, the *M. orientalis* individual, and the four *M. sylvestris* individuals grown at IDEEV). For the Red Delicious, Honeycrisp, Brown Snout, Costard, Bardsey Island, and Newton cultivars, and for *M. sieversii* and *M. sylvestris*, only one haplotype was kept from previous studies due to the unavailability of the second haplotype or insufficient assembly quality. For Golden Delicious and Hanfu, the haplotypes were derived from a double haploid individual, the GDHH13 *M. domestica* reference genome⁴², and a tri-haploid individual⁴³. This dataset combines 15 already available haplotype assemblies and 12 new sequenced haplotypes from which only chromosome 17 was retained for this study, as this chromosome hosts the SI locus in *Malus*⁴⁴ (Fig. 1).

The 27 haplotypes displayed high assembly completeness, with genome sizes ranging from ~600 Mb to 660 Mb, scaffold N50 values of ~9–39 Mb (average N50 = 33.09 Mb ± 6.86), and BUSCO⁴⁵ completeness scores generally > 95%. Repetitive content varied moderately between haplotypes (65–76%). Details about the genome assemblies and gene annotations are provided in Supplementary Table 1 and Supplementary Text 1. For convenience, all haplotypes were identified with an ID from h1 to h32. To avoid bias due to heterogeneous annotations, all haplotypes were reannotated using a single, standardized gene annotation pipeline.

Construction of an *S-RNase* and *SFBB* gene database

An S-locus gene database, consisting of all *S-RNase* and *SFBB* sequences available in public repositories, was built as a raw material to investigate the genomic architecture of these genes in the genome assemblies described above. Public *S-RNase* DNA sequences^{46–49} and *SFBB* DNA sequences^{50,51} were collected. The homologous sequences of the *S-RNase* and *SFBB* genes were then recovered using BLASTN searches on the GenBank database with the parameter -e-value 0.001 and added to complete the S-locus gene database. The *S-RNase* database was split between exon 1 and 2 sequences to avoid mis-mapping due to high polymorphism of the intron between individuals. The accession numbers of the genes are provided in Supplementary Tables 2 and 5.

S-RNase assignment of each haplotype

Using the database of publicly available *S-RNase* sequences constructed above (Supplementary Table 2), the Maloideae *S-RNase* sequences in the database were aligned to *S-RNase* sequences from the haplotypes using Clustal Omega⁵² to assign a given *S-RNase* allele to each haplotype (Supplementary Table 3 and Supplementary Fig. 1).

Haplotype assembly and gene annotation

Thirty-one haplotype assemblies sourced from publicly available data and newly sequenced genotypes were constructed using PacBio HiFi sequences. Each haplotype was defined as a set of chromosomes transmitted together (Supplementary Table 1 and Supplementary Text 2).

Among these, 15 haplotype assemblies were obtained from the literature. Twelve haplotype assemblies were from 10 *M. domestica* cultivars: double haploid Golden Delicious (one haplotype available, GDDH13), Fuji (two haplotypes), Gala (two haplotypes), Red Delicious (one haplotype available), Honeycrisp (one haplotype), tri-haploid Hanfu (one haplotype available), Brown Snout (one haplotype available), Costard, Bardsey Island (one haplotype available), and Newtons (one haplotype available). Additionally, three haplotype assemblies were used, with one from *M. sieversii* and two from *M. sylvestris*.

The dataset was enriched with six newly sequenced genomes, representing 12 haplotypes. These new genomes represented one *M. domestica* genotype (Xuanina, used for cider production), one *M. orientalis* genotype from Armenia, and four *M. sylvestris* samples, one each from Austria, Denmark, France, and Romania. Each newly sequenced genome contributes two parental haplotypes, resulting in a total of 12 supplementary haplotypes. Details about DNA extraction, sequencing, and assembly of the seven newly sequenced genomes are summarized in Supplementary Table 1 and Supplementary Text 2. Briefly, PacBio sequencing and assembly of the resulting reads generated an average assembly size of 642 Mb, similar to publicly available assemblies (average of 636.5 Mb), with an average scaffold N50 value of 28.12 Mb, which was lower than the average scaffold N50 of publicly available assemblies (average of 37 Mb).

Characterization of *SFBB* gene families

To cluster *SFBB* genes from all haplotypes into families, all complete SFBB protein sequences were aligned using Clustal Omega v1.2.4⁵². The alignment was used as input for IQ-tree⁵³ with ModelFinder⁵⁴

for a first run to determine the best-fitting substitution model using 1,000 bootstrap replicates. Using the resulting phylogenetic tree, the hierarchical clustering method of Ward⁵⁵ was used with the cophenetic distance matrix, using R⁵⁶ and the software packages *ape* v5.7-1⁵⁷, *FactoClass* v1.2.8⁵⁸, and *stats* v4.3.1⁵⁶. Additionally, OrthoFinder⁵⁹ was run on SFBB protein sequence alignments derived from each haplotype. To validate the defined SFBB families, identity percentages were computed from pairwise SFBB alignments (from the same families or different families) using the *pairwiseAlignment* command from the R package *Biostrings* v2.68.1⁶⁰.

The *Malus* S-locus is located in the peri-telomeric region on the long arm of chromosome 17⁴⁴. Therefore, chromosome 17 contigs and annotated genes were extracted using the Helixer/Eggnog pipeline (Supplementary Text 2), followed by analysis of the SI regions to define their size and gene content.

Phylogenetic analysis of S-RNases

To determine the haplotypes of the S-alleles, S-RNase protein sequences were aligned to reference sequences listed in Supplementary Table 2 using Clustal Omega v1.2.4⁵². S-RNase sequences associated with S-alleles from previous studies were used to identify S-alleles in the S-haplotype dataset. Specifically, phylogenetic relationships among all S-RNase sequences were examined using IQ-tree⁵³ with ModelFinder⁵⁴ in the first run to determine the best-fitting substitution model using 1,000 bootstrap replicates.

Sequences and gene synteny within S-loci

To examine the synteny across the S-loci from the 27 *Malus* haplotypes, regions encompassing 500 kb upstream and downstream of S-locus sequences were aligned using the nucmer tool in MUMmer⁶¹. SyRI (Synteny and Rearrangement Identifier) v1.6.3⁶² was then used to identify the syntenic path (longest set of colinear regions), structural rearrangements (inversion, translocations, and duplications), local variations (single nucleotide polymorphisms [SNPs], insertion/deletions [InDels], copy-number variation [CNV]) within syntenic blocks and structural rearrangements, and unaligned regions. SyRI takes alignment blocks identified by the initial aligner and identifies larger regions consisting of multiple consecutive one-to-one alignment blocks. Finally, PlotSR⁶³ was used to visualize synteny using the output from SyRI.

To generate gene synteny plots, S-loci annotation gff files from Apollo⁶⁴ were converted to bed files using *agat_convert_sp_gff2bed.pl* from AGAT⁶⁵, which were used as input for the R package

gggenomes v0.9.9.9000⁶⁶. Identity between S-locus genes was computed using Diamond⁶⁷ based on their corresponding amino acid (aa) sequences.

Syntenic relationships across several S-loci were examined using the network-based approach with the R package Syntenet v1.9.1²¹, which treats anchor pairs (duplicates retained from a large-scale duplication event) from synteny comparisons as connected nodes of an undirected unweighted graph⁶⁸.

Nucleotide diversity analysis with Pixy

Nucleotide diversity (π) across the genome was calculated using Pixy v1.2.6.beta1 software²², which accurately estimates π from VCF files that include invariant sites and accounts for missing data. π was computed in non-overlapping 10-kb windows using the following parameters: `--stats pi --window_size 10000 --populations population_file.txt` to use the species information for each sample in the dataset.

The input VCF file was originally generated in Chen et al., (2025)⁶⁹ and includes multiple phased haploid genomes from several *Malus* species. From this dataset, a subset of 41 samples was extracted, representing four species relevant to our study: *M. domestica* (n=10), *M. orientalis* (n=9), *M. sieversii* (n=10), and *M. sylvestris* (n=12). The analysis was restricted to chromosome 17, which harbours the S-locus. The coordinates of the S-locus in the GDDH13 reference genome⁷⁰, corresponding to haplotype h1 in our study, were defined as 30,032,275 bp and 31,732,190 bp, and windows were classified as either inside or outside the S-locus.

To identify outlier windows with significantly higher π than the remaining genomes, the windows with the top 1% highest- π values within each species were defined independently. To determine whether the S-locus was enriched for high-diversity windows, a permutation-based enrichment test was performed using R⁵⁶. The empirical p -value was calculated as the proportion of permutations in which the number of outlier windows in the S-locus was equal to or greater than the observed count. A significant p -value ($p < 0.05$) indicates that high-diversity windows are non-randomly concentrated in the S-locus, consistent with a signal of balancing selection.

Detection of SNP density and non-synonymous/synonymous ratio along chromosome 17

We used the SnpEff⁷¹-annotated VCF file corresponding to chromosome 17 to investigate the distribution of genetic variation across the S-locus. Single nucleotide polymorphisms (SNPs) were classified according to their predicted functional effects into synonymous or non-synonymous variants. Variants located within SFBB coding regions, where recurrent positive selection is expected, were excluded to avoid upward bias in the nS/S ratio. SNP counts were aggregated in non-overlapping

windows of 5 kb along the chromosome, from which we calculated both the density of SNPs and the ratio of non-synonymous to synonymous substitutions (nS/S). To visualize these patterns, we implemented a custom Python script relying on cyvcf2⁷², pandas⁷³, and matplotlib⁷⁴. The script produced a three-panel figure with (i) the nS/S ratio, (ii) the SNP density per window, and (iii) a gene track. Gene coordinates were retrieved from the GFF3 annotation file, and all annotated genes were represented as blocks along the chromosome, with SFBB genes highlighted in a distinct color.

Signatures of selection at SI genes

To determine whether the *SFBB* genes evolved neutrally or by natural selection (positive or purifying), the codeml program in the PAML package v4.9⁷⁵ was used. Codeml computes the ratio of nonsynonymous (dN) to synonymous (dS) substitutions for various models of sequence evolution. The most likely model explaining the data is chosen based on a likelihood ratio test computed among models. Three pairs of site-specific models allowed us to test for recurrent, diversifying selection (M0 and M3, M1a and M2a, and M7 and M8). One SFBB protein sequence alignment for each haplotype was used. To assess whether aa identity at each aligned position was significantly associated with the *SFBB* gene family, a multinomial regression model was fitted for each position that showed at least two different aa and was present in at least two families. For each position, a null model ($aa \sim 1$) was compared to a full model ($aa \sim \text{Family}$) using a likelihood ratio test. *P*-values were computed from the χ^2 distribution and adjusted for multiple testing using the Benjamini–Hochberg procedure to control the false discovery rate (FDR). Multiple correspondence analysis (MCA), which is specifically designed for categorical data and better captures variation in aa composition across discrete sites, was performed to investigate the clustering of *SFBB* genes based on aa identity at key positions. For each gene, aa residues at conserved positions were extracted and treated as categorical variables. The MCA was conducted using the *FactoMineR* package⁷⁶ in R⁵⁶ with the SFBB family provided as a supplementary qualitative variable. Individuals (genes) were projected onto the first two dimensions, and family labels were used for colouring.

For S-RNases, 27 protein sequences were aligned using Clustal Omega v1.2.4⁵². To identify codon positions likely involved in allele-specific recognition in the S-RNase sequences, evidence from multiple complementary approaches was combined. First, the Fixed Effects Likelihood (FEL) method available on the datamonkey web platform⁷⁷ was applied for codon alignment restricted to one representative sequence per S-allele to focus on substitutions that occurred between allelic groups rather than within them. In parallel, the site models implemented in codeml (PAML⁷⁸) were used to detect

codon sites evolving under positive selection based on likelihood ratio tests comparing models, allowing for $\omega > 1$ (e.g., M2a, M8) against their respective null models. The results from both methods were compared to retain only robust candidates. To assess whether these positively selected sites were conserved within allelic groups, the full codon alignment was examined, including all available sequences per allele. For each candidate site, intra-allelic variation was quantified by counting the number of distinct codons observed within each group. Positions showing no variation within any allele, i.e., monomorphic within-group but polymorphic between groups, were considered strong candidates for functional divergence related to allele-specific interactions.

Search for duplicated genes

To examine how multiple *SFBB* genes were generated and proliferated, the `classify_gene_pairs` command in the R/Bioconductor package *doubletrouble*⁷⁹ was used with the standard scheme (binary = FALSE). In short, intra-haplotypes protein similarity searches were performed using DIAMOND⁶⁷ (sensitive mode, e-value = 1e-10, top hits = 5) to identify the entire set of paralogous gene pairs for each haplotype. Paralogous gene pairs in the same syntenic blocks were classified as arising from segmental duplication. A tandem duplication is a segmental duplication that results from an unequal crossing over^{80,81}. A few genes separate proximal gene pairs⁸² and can originate from ancient tandem duplication events disrupted by the insertion of other genes⁸³. Dispersed duplication occurs through unpredictable and random patterns by mechanisms that remain unclear, generating two gene copies that are neither neighboring nor colinear⁸⁴.

Signatures of gene conversion in *SFBB* genes

Intra-*SFBB* family gene conversion signatures were detected using GENECONV⁸⁵, which scans aligned sequences and identifies segments (tracts) that are (1) identical or nearly identical, (2) longer than expected by chance, and (3) shared between sequences that are otherwise divergent. GENECONV tests gene conversion using 10,000 permutations and assesses significant gene alignment using Bonferroni correction as well as the Karlin-Altschul score. To identify potential preferential gene conversion between specific S-allele pairs, a global table of gene conversion events was compiled, with each event associated with the S-alleles of the two haplotypes involved. For each event, the corresponding S-allele pair was annotated. Because S-alleles were unevenly represented across the 27 S-haplotypes, each S-allele was assigned a weight equal to the inverse of its frequency across the set of assemblies. To enable robust statistical testing, a bootstrap procedure was implemented. In each replicate, 10 S-haplotypes

were randomly selected, and only gene conversion events between S-alleles present in that subset were retained. The number of events per S-allele pair was then counted and associated with the corresponding mean weight. After repeating this procedure 500 times, a dataset was obtained with multiple observations per pair. A quasi-Poisson regression model using the `glm` function from the R package *stats*⁵⁶ was then used to test whether some S-allele pairs exhibited significantly higher or lower gene conversion frequencies than expected. The model was fitted using the number of conversion events as the response variable, the S-allele pair as a predictor, and the mean weight as a statistical weight to account for allelic representation bias.

TE distribution and enrichment analysis around *SFBB* genes

TEs within the S-locus were annotated using a pan-genomic TE library built from a representative set of apple genomes. The TE consensus libraries were first generated using the *TEdenovo* pipeline from REPET v3.0 using Tedenovo⁸⁶, and then using these libraries the TEs of the following genomes were annotated using TEannot⁸⁷: the reference cultivar GDDH13⁷⁰, one haplotype from a cider variety (unpublished data), one haplotype from each European *Malus sylvestris* population (unpublished data), one haplotype of *M. orientalis* (unpublished data), and one haplotype of *M. sieversii*⁸⁸. These libraries were concatenated, and redundancy was removed using PASTEC⁸⁹ with thresholds of at least 98% identity and 95% coverage. Sequences showing high similarity to *SFBB* genes (>90%) or classified as potential host genes (PHGs) by PASTEC were excluded. The resulting curated database was then used to determine shared and non-shared TE insertions across genomes in S-locus regions using *panREPET*⁹⁰.

To identify potentially recurrent associations between TEs and *SFBB* gene families independently of the *S-RNase* allele, a permutation test was performed. For each TE consensus–*SFBB* family pair, the number of distinct S-alleles (nS-alleles) observed among the haplotypes where the association occurred was counted. TE–*SFBB* proximity was restricted to ≤ 4 kb as a conservative threshold. Inter-*SFBB* distances within the S-locus are generally orders of magnitude larger (commonly tens of kb; Supplementary Fig. 23), so a 4-kb window targets elements that are genuinely co-located with an *SFBB* gene, encompassing likely regulatory flanks and the range of plausible short recombination/gene-conversion footprints while limiting noise from distant TEs. This number was then compared to a null distribution obtained by randomly permuting the S-alleles across haplotypes. At each iteration, S-alleles were shuffled across haplotypes while preserving the TE–*SFBB* associations. An empirical *p*-value was calculated as the proportion of permutations in which the number of S-alleles

was equal to or greater than the observed value. This approach allows us to detect TE-*SFBB* associations that recur across multiple S-alleles more frequently than expected by chance, suggesting possible gene conversion or recombination events within the S-locus. These associations were visualized as heatmaps using the R package *ggplot2*⁹¹.

Protein-protein interactions

The sequences from the different families were analysed using MuLAN³³, a sequence-based deep-learning method designed to predict changes in binding affinity induced by mutations affecting protein sequences. In addition, MuLAN can be run on single sequences to compute scores for each residue based on a local attention mechanism, normalized in the range [0, 1], which correlates with interacting or functional regions. Finally, the sequences in each family were aligned. The score profiles at the aligned positions were plotted, highlighting the experimentally determined positive selection sites (Supplementary Fig. 12, Supplementary Fig. 18, and Supplementary Fig. 22). Predicted structures for protein complexes were obtained with Boltz-1³², using the 1.0.0 version of the package. For each complex, multiple sequence alignments were automatically retrieved using the MMseqs2 server and five alternative structures were generated, setting an initial random seed of 42. For each case, we then selected the predicted structure with the highest associated confidence score. We did not employ additional steering potentials as no steric clashes were observed in the model's outputs.

Data availability

The coding sequences and protein sequences of *SFBB* and *S-RNase* genes are available in the ZENODO repository: DOI <https://doi.org/10.5281/zenodo.17226708>. All analysis files and codes are available at <https://github.com/CornilleEclecticLab/Apple-SI/tree/main>.

Funding

This research was funded by ANR JCJC PLEASURE ANR-21-CE20-0005 and a Tamkeen research fund under AD454 NYUAD led by AMC.

Authors contributions

AMC obtained funding. AV, WM, XC, AM, and AMC generated, assembled, and anchored the genome sequences. AM, GL, SS, ALF, AC, and AMC analyzed the data. All co-authors discussed the results. AM, AMC, VC, MM, XV, and AC wrote the manuscript with critical input from other co-authors.

Conflicts of interest

The authors declare that they have no financial conflict of interest with the content of this article.

References

1. Charlesworth, D. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genet.* **2**, e64 (2006).
2. Glémin, S. & Galtier, N. Genome Evolution in Outcrossing Versus Selfing Versus Asexual Species. in *Evolutionary Genomics* (ed. Anisimova, M.) vol. 855 311–335 (Humana Press, Totowa, NJ, 2012).
3. Wright, S. I., Ness, R. W., Foxe, J. P. & Barrett, S. C. H. Genomic Consequences of Outcrossing and Selfing in Plants. *Int. J. Plant Sci.* **169**, 105–118 (2008).
4. Minamikawa, M. *et al.* Apple S locus region represents a large cluster of related, polymorphic and pollen-specific F-box genes. *Plant Mol. Biol.* **74**, 143–154 (2010).
5. Sassa, H. *et al.* S Locus F-Box Brothers: Multiple and Pollen-Specific F-Box Genes With S Haplotype-Specific Polymorphisms in Apple and Japanese Pear. *Genetics* **175**, 1869–1881 (2007).
6. Fujii, S., Kubo, K. & Takayama, S. Non-self- and self-recognition models in plant self-incompatibility. *Nat. Plants* **2**, 16130 (2016).
7. Williams, J. S., Wu, L., Li, S., Sun, P. & Kao, T.-H. Insight into S-RNase-based self-incompatibility in Petunia: recent findings and future directions. *Front. Plant Sci.* **6**, (2015).

8. Kubo, K. *et al.* Collaborative Non-Self Recognition System in S-RNase-Based Self-Incompatibility. *Science* **330**, 796–799 (2010).
9. Zhu, S. *et al.* The Snapdragon Genomes Reveal the Evolutionary Dynamics of the *S* -Locus Supergene. *Mol. Biol. Evol.* **40**, msad080 (2023).
10. Jay, P. *et al.* Supergene Evolution Triggered by the Introgression of a Chromosomal Inversion. *Curr. Biol.* **28**, 1839-1845.e3 (2018).
11. Jay, P., Jeffries, D., Hartmann, F. E., Véber, A. & Giraud, T. Why do sex chromosomes progressively lose recombination? *Trends Genet.* **40**, 564–579 (2024).
12. Kay, T., Helleu, Q. & Keller, L. Iterative evolution of supergene-based social polymorphism in ants. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210196 (2022).
13. Kubo, K. *et al.* Gene duplication and genetic exchange drive the evolution of S-RNase-based self-incompatibility in Petunia. *Nat. Plants* **1**, 14005 (2015).
14. Liang, M. *et al.* Evolution of self-compatibility by a mutant Sm-RNase in citrus. *Nat. Plants* **6**, 131–142 (2020).
15. Bod'ová, K., Priklopil, T., Field, D. L., Barton, N. H. & Pickup, M. Evolutionary Pathways for the Generation of New Self-Incompatibility Haplotypes in a Nonself-Recognition System. *Genetics* **209**, 861–883 (2018).
16. Harkness, A., Goldberg, E. E. & Brandvain, Y. Diversification or collapse of self-incompatibility haplotypes as outcomes of evolutionary rescue. Preprint at <https://doi.org/10.1101/641613> (2019).

17. Erez, K., Jangid, A., Feldheim, O. N. & Friedlander, T. The role of promiscuous molecular recognition in the evolution of RNase-based self-incompatibility in plants. *Nat. Commun.* **15**, 4864 (2024).
18. Vekemans, X. & Slatkin, M. Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* **137**, 1157–1165 (1994).
19. Castric, V., Bechsgaard, J., Schierup, M. H. & Vekemans, X. Repeated Adaptive Introgression at a Gene under Multiallelic Balancing Selection. *PLoS Genet.* **4**, e1000168 (2008).
20. Charlesworth, D. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genet.* **2**, e64 (2006).
21. Almeida-Silva, F., Zhao, T., Ullrich, K. K., Schranz, M. E. & Van De Peer, Y. syntenet: an R/Bioconductor package for the inference and analysis of synteny networks. *Bioinformatics* **39**, (2023).
22. Korunes, K. L. & Samuk, K. PIXY : Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol. Ecol. Resour.* **21**, 1359–1368 (2021).
23. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76**, 5269–5273 (1979).
24. Jay, P., Véber, A. & Giraud, T. Stepwise expansion of recombination suppression on sex chromosomes and other supergenes through lower load advantage and deleterious mutation sheltering. Preprint at <https://doi.org/10.1101/2025.06.27.661902> (2025).

25. Yang, Q., Zhang, D., Li, Q., Cheng, Z. & Xue, Y. Heterochromatic and genetic features are consistent with recombination suppression of the self-incompatibility locus in *Antirrhinum*. *Plant J.* **51**, 140–151 (2007).
26. Jing, X. *et al.* Genetic diversity of the self-incompatibility locus in diploid potato. *J. Integr. Agric.* S2095311924004209 (2024) doi:10.1016/j.jia.2024.12.011.
27. Harbord, R. M., Napoli, C. A. & Robbins, T. P. Segregation Distortion of T-DNA Markers Linked to the Self-Incompatibility (S) Locus in *Petunia hybrida*. *Genetics* **154**, 1323–1333 (2000).
28. Vázquez, N. *et al.* BDBM 1.0: A Desktop Application for Efficient Retrieval and Processing of High-Quality Sequence Data and Application to the Identification of the Putative *Coffea* S-Locus. *Interdiscip. Sci. Comput. Life Sci.* **11**, 57–67 (2019).
29. Cao, Z.-H. *et al.* An *S* -locus F-box protein as pollen *S* determinant targets non-self S-RNase underlying self-incompatibility in *Citrus*. *J. Exp. Bot.* **75**, 3891–3902 (2024).
30. Gu, C. *et al.* Long-read genome sequencing reveals the sequence characteristics of pear self-incompatibility locus. *Mol. Hortic.* **5**, 13 (2025).
31. Chen, L.-Y. *et al.* The bracteatus pineapple genome and domestication of clonally propagated crops. *Nat. Genet.* **51**, 1549–1558 (2019).
32. Wohlwend, J. *et al.* Boltz-1 Democratizing Biomolecular Interaction Modeling. Preprint at <https://doi.org/10.1101/2024.11.19.624167> (2024).
33. Lombardi, G. & Carbone, A. MuLAN: Mutation-driven Light Attention Networks for investigating protein-protein interactions from sequences. Preprint at <https://doi.org/10.1101/2024.08.24.609515> (2024).

34. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
35. Ishimizu, T. *et al.* Identification of regions in which positive selection may operate in S-RNase of Rosaceae: Implication for S -allele-specific recognition sites in S-RNase. *FEBS Lett.* **440**, 337–342 (1998).
36. Ushijima, K. *et al.* Cloning and characterization of cDNAs encoding S-RNases from almond (*Prunus dulcis*): primary structural features and sequence diversity of the S-RNases in Rosaceae. *Mol. Gen. Genet. MGG* **260**, 261–268 (1998).
37. Goldberg, E. E. *et al.* Species Selection Maintains Self-Incompatibility. *Science* **330**, 493–495 (2010).
38. Wang, R. *et al.* Characterisation of the *Gillenia* S-locus provides insight into evolution of the nonself-recognition self-incompatibility system in apple. *Sci. Rep.* **15**, 14630 (2025).
39. Kim, H.-T. *et al.* Identification and characterization of *S-RNase* genes in apple rootstock and the diversity of *S-RNases* in *Malus* species. *J. Plant Biotechnol.* **43**, 49–57 (2016).
40. Branco, S. *et al.* Multiple convergent supergene evolution events in mating-type chromosomes. *Nat. Commun.* **9**, 2000 (2018).
41. Du, J. *et al.* Molecular characteristics of *S-RNase* alleles as the determinant of self-incompatibility in the style of *Fragaria viridis*. *Hortic. Res.* **8**, (2021).
42. Daccord, N. *et al.* High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).

43. Qin, S. *et al.* A chromosome-scale genome assembly of *Malus domestica*, a multi-stress resistant apple variety. *Genomics* **115**, 110627 (2023).
44. Aguiar, B. *et al.* Convergent Evolution at the Gametophytic Self-Incompatibility System in *Malus* and *Prunus*. *PLOS ONE* **10**, e0126138 (2015).
45. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
46. Dreesen, R. S. G. *et al.* Analysis of *Malus* S-RNase gene diversity based on a comparative study of old and modern apple cultivars and European wild apple. *Mol. Breed.* **26**, 693–709 (2010).
47. Larsen, B., Ørgaard, M., Toldam-Andersen, T. B. & Pedersen, C. A high-throughput method for genotyping S-RNase alleles in apple. *Mol. Breed.* **36**, 24 (2016).
48. Ma, X., Cai, Z., Liu, W., Ge, S. & Tang, L. Identification, genealogical structure and population genetics of S-alleles in *Malus sieversii*, the wild ancestor of domesticated apple. *Heredity* **119**, 185–196 (2017).
49. Vieira, J., Ferreira, P. G., Aguiar, B., Fonseca, N. A. & Vieira, C. P. Evolutionary patterns at the RNase based gametophytic self - incompatibility system in two divergent Rosaceae groups (Maloideae and *Prunus*). *BMC Evol. Biol.* **10**, 200 (2010).
50. Minamikawa, M. *et al.* Apple S locus region represents a large cluster of related, polymorphic and pollen-specific F-box genes. *Plant Mol. Biol.* **74**, 143–154 (2010).

51. Okada, K., Moriya, S., Haji, T. & Abe, K. Isolation and characterization of multiple F-box genes linked to the S 9 - and S 10 -RNase in apple (*Malus × domestica* Borkh.). *Plant Reprod.* **26**, 101–111 (2013).
52. Sievers, F. & Higgins, D. G. Clustal Omega. *Curr. Protoc. Bioinforma.* **48**, (2014).
53. Minh, B. Q. *et al.* Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 2461–2461 (2020).
54. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
55. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
56. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2021).
57. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
58. Campo Elias Pardo *et al.* FactoClass: Combination of Factorial Methods and Cluster Analysis. The R Foundation <https://doi.org/10.32614/cran.package.factoclass> (2008).
59. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
60. H. Pagès, P. A. Biostrings. Bioconductor <https://doi.org/10.18129/B9.BIOC.BIOSTRINGS> (2017).

61. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).
62. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
63. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).
64. Dunn, N. A. *et al.* Apollo: Democratizing genome annotation. *PLOS Comput. Biol.* **15**, e1006790 (2019).
65. Jacques Dainat *et al.* NBISweden/AGAT: AGAT v1.5.0. Zenodo <https://doi.org/10.5281/ZENODO.3552717> (2025).
66. Hackl, T., Ankenbrand, M., van Adrichem, B., Wilkins, D. & Haslinger, K. gggenomes: effective and versatile visualizations for comparative genomics. Preprint at <https://doi.org/10.48550/ARXIV.2411.13556> (2024).
67. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
68. Zhao, T. & Schranz, M. E. Network approaches for plant phylogenomic synteny analysis. *Curr. Opin. Plant Biol.* **36**, 129–134 (2017).
69. Chen, X. *et al.* Gene flow from the European wild apple and selection shaped the domesticated apple (*Malus domestica* Borkh.) genome. Preprint at <https://doi.org/10.1101/2025.09.18.676739> (2025).

70. Daccord, N. *et al.* High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
71. Cingolani, P. *et al.* 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 (Austin)* **6**, 80–92 (2012).
72. Pedersen, B. S. & Quinlan, A. R. cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics* **33**, 1867–1869 (2017).
73. McKinney, W. Data Structures for Statistical Computing in Python. in 56–61 (Austin, Texas, 2010). doi:10.25080/Majora-92bf1922-00a.
74. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
75. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
76. Lê, S., Josse, J. & Husson, F. **FactoMineR**: An R Package for Multivariate Analysis. *J. Stat. Softw.* **25**, (2008).
77. Weaver, S. *et al.* Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Mol. Biol. Evol.* **35**, 773–777 (2018).
78. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

79. Almeida-Silva, F. & Van De Peer, Y. *doubletrouble*: an R/Bioconductor package for the identification, classification, and analysis of gene and genome duplications. Preprint at <https://doi.org/10.1101/2024.02.27.582236> (2024).
80. Freeling, M. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annu. Rev. Plant Biol.* **60**, 433–453 (2009).
81. Hahn, M. W. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *J. Hered.* **100**, 605–617 (2009).
82. Wang, Y., Wang, X. & Paterson, A. H. Genome and gene duplications and gene expression divergence: a view from plants. *Ann. N. Y. Acad. Sci.* **1256**, 1–14 (2012).
83. Freeling, M. *et al.* Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Res.* **18**, 1924–1937 (2008).
84. Wang, Y., Ficklin, S. P., Wang, X., Feltus, F. A. & Paterson, A. H. Large-Scale Gene Relocations following an Ancient Genome Triplication Associated with the Diversification of Core Eudicots. *PLOS ONE* **11**, e0155637 (2016).
85. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* <https://doi.org/10.1093/oxfordjournals.molbev.a040567> (1989)
doi:10.1093/oxfordjournals.molbev.a040567.
86. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE* **6**, e16526 (2011).
87. Quesneville, H. *et al.* Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Comput. Biol.* **1**, e22 (2005).

88. Sun, X. *et al.* Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* **52**, 1423–1432 (2020).
89. Hoede, C. *et al.* PASTEC: An Automatic Transposable Element Classification Tool. *PLoS ONE* **9**, e91929 (2014).
90. Saidi, S., Blaison, M., Rodríguez-Ordóñez, M. D. P., Confais, J. & Quesneville, H. panREPET: a reference-free pipeline for detecting shared Transposable Elements from pan-genomes to retrace their dynamics in a species. Preprint at <https://doi.org/10.1101/2024.06.17.598857> (2024).
91. Wickham, H. ggplot2. *WIREs Comput. Stat.* **3**, 180–185 (2011).

Supplementary Table 1: *Malus* assemblies used in this study

Haplo type	Species	Cultivar / Origin	Ploidy	Sequence length	Scaffold N50 (Mb)	Gene count	Protein coding genes (%)	Repeats (%)	Busco (%)	Chromosome 17 length (Mb)	References	Rq
h1	<i>M. domestica</i>	cult. Golden delicious	Doubled-haploid	656,833,032	37.632	49,762	51.6	65.4	96.9	34.749	Daccord <i>et al.</i> , 2017	Kept for subsequent analysis
h2	<i>M. domestica</i>	cult. Gala	Diploid	657,696,253	37.966	51,889	93.6	74.85	95.7	32.702	Sun <i>et al.</i> , 2020	Kept for subsequent analysis
h3	<i>M. domestica</i>	cult. Gala	Diploid	577,209,808	36.439	45,344	93.3	74.79	86.6	30.262	Sun <i>et al.</i> , 2020	Kept for subsequent analysis
h4	<i>M. domestica</i>	RGTH1	Haploid	754,684,250	35.935				96.7		Tian <i>et al.</i> , 2022	Not kept for subsequent analysis ()
h6	<i>M. domestica</i>	cult. Fuji	Diploid	648,363,766	36.861	56,85	94.1	73.75	97.7	34.006	Li <i>et al.</i> , 2024	Kept for subsequent analysis
h7	<i>M. domestica</i>	cult. Fuji	Diploid	647,946,317	37.293	52,523	93.8	74.15	97.6	34.212	Li <i>et al.</i> , 2024	Kept for subsequent analysis
h8	<i>M. domestica</i>	cult. Red delicious	Diploid	625,992,033	37.632	49,554	93.4	74.16	97.4	33.496	Han <i>et al.</i> , 2017	Kept for subsequent analysis
h9	<i>M. domestica</i>	cult. Honeycrisp	Diploid		31.600				1.9		Khan <i>et al.</i> , 2022	Not kept for subsequent analysis
h10	<i>M. domestica</i>	cult. Honeycrisp	Diploid	648,227,795	32.8	51,523	93.6	74.6	97.6	33.450	Khan <i>et al.</i> , 2022	Kept for subsequent analysis
h11	<i>M. domestica</i>	cult. Hanfu	Tri-haploid	652,502,298	37.139	51,127	93.6	75.26	97.6	33.999	Qin <i>et al.</i> , 2023	Kept for subsequent analysis
h12	<i>M. domestica</i>	cult. Brown Snout	Diploid	645,430,038	37.911	50,416	93.6	75.08	97.7	34.188	Konyves <i>et al.</i> , 2022	Kept for subsequent analysis
h13	<i>M. domestica</i>	cult. Costard	Diploid	645,735,903	37.048	50,379	64.1	75.23	97.6	33.906	Konyves <i>et al.</i> , 2022	Kept for subsequent analysis
h14	<i>M. domestica</i>	cult. Bardsey Island	Diploid	650,288,218	37.701	50,343	93.8	75.28	97.7	34.850	Konyves <i>et al.</i> , 2022	Kept for subsequent analysis
h15	<i>M. domestica</i>	cult. Newtons	Diploid	638,366,891	37.767	50,265	93.8	74.71	97.7	34.042	Konyves <i>et al.</i> , 2022	Kept for subsequent analysis
h16	<i>M. sieversii</i>	Central Asia	Diploid		36.622				0.7		Sun <i>et al.</i> , 2020	Not kept for subsequent analysis
h17	<i>M. sieversii</i>	Central Asia	Diploid	612,178,858	39.79	46,367	93.6	75.96	91.5	32.620	Sun <i>et al.</i> , 2020	Kept for subsequent analysis

h18	<i>M. sylvestris</i>	Germany	Diploid		35.560				0.8		Sun <i>et al.</i> , 2020	Not kept for subsequent analysis
h19	<i>M. sylvestris</i>	Germany	Diploid	601,478,285	35.227	49,446	56.8	74.02	92.7	31.469	Sun <i>et al.</i> , 2020	Kept for subsequent analysis
h20	<i>M. sylvestris</i>	England	Diploid	639,737,487	36.902	50,379	93.8	74.94	97.7	33.873	Ruhsam <i>et al.</i> , 2022	Kept for subsequent analysis
h21	<i>M. domestica</i>	cult. Xuanina	Diploid	644,465,789	29.167	50,762	57.0	74.57	97.1	33.655	This study	Kept for subsequent analysis
h22	<i>M. domestica</i>	cult. Xuanina	Diploid	625,304,225	9.418	48,696	93.7	74.6	95.8	32.333	This study	Kept for subsequent analysis
h23	<i>M. sylvestris</i>	Austria	Diploid	644,132,839	24.315	54,796	93.9	73.79	97.6	35.410	This study	Kept for subsequent analysis
h24	<i>M. sylvestris</i>	Austria	Diploid	633,775,122	18.97	50,424	93.7	74.11	97.3	33.610	This study	Kept for subsequent analysis
h25	<i>M. sylvestris</i>	Denmark	Diploid	638,694,078	24.423	54,188	93.9	74.06	97	34.908	This study	Kept for subsequent analysis
h26	<i>M. sylvestris</i>	Denmark	Diploid	633,317,074	30.454	50,573	93.8	74.25	97.1	32.936	This study	Kept for subsequent analysis
h27	<i>M. sylvestris</i>	France	Diploid	651,453,350	35.932	51,039	93.8	74.19	97.5	35.550	This study	Kept for subsequent analysis
h28	<i>M. sylvestris</i>	France	Diploid	634,091,149	33.34	49,538	93.7	73.75	97.4	32.920	This study	Kept for subsequent analysis
h29	<i>M. sylvestris</i>	Romania	Diploid	649,942,226	35.638	51,404	93.8	73.83	97.8	34.218	This study	Kept for subsequent analysis
h30	<i>M. sylvestris</i>	Romania	Diploid	643,103,033	30.517	50,104	93.7	74.04	97.8	35.246	This study	Kept for subsequent analysis
h31	<i>M. orientalis</i>	Armenia	Diploid	653,674,533	32.637	55,781	93.6	74.32	97.7	35.488	This study	Kept for subsequent analysis
h32	<i>M. orientalis</i>	Armenia	Diploid	652,365,126	32.639	51,253	93.8	74.8	98	33.672	This study	Kept for subsequent analysis

Supplementary Table 2: S-RNase database, public accessions from *Malus* and *Pyrus* genus

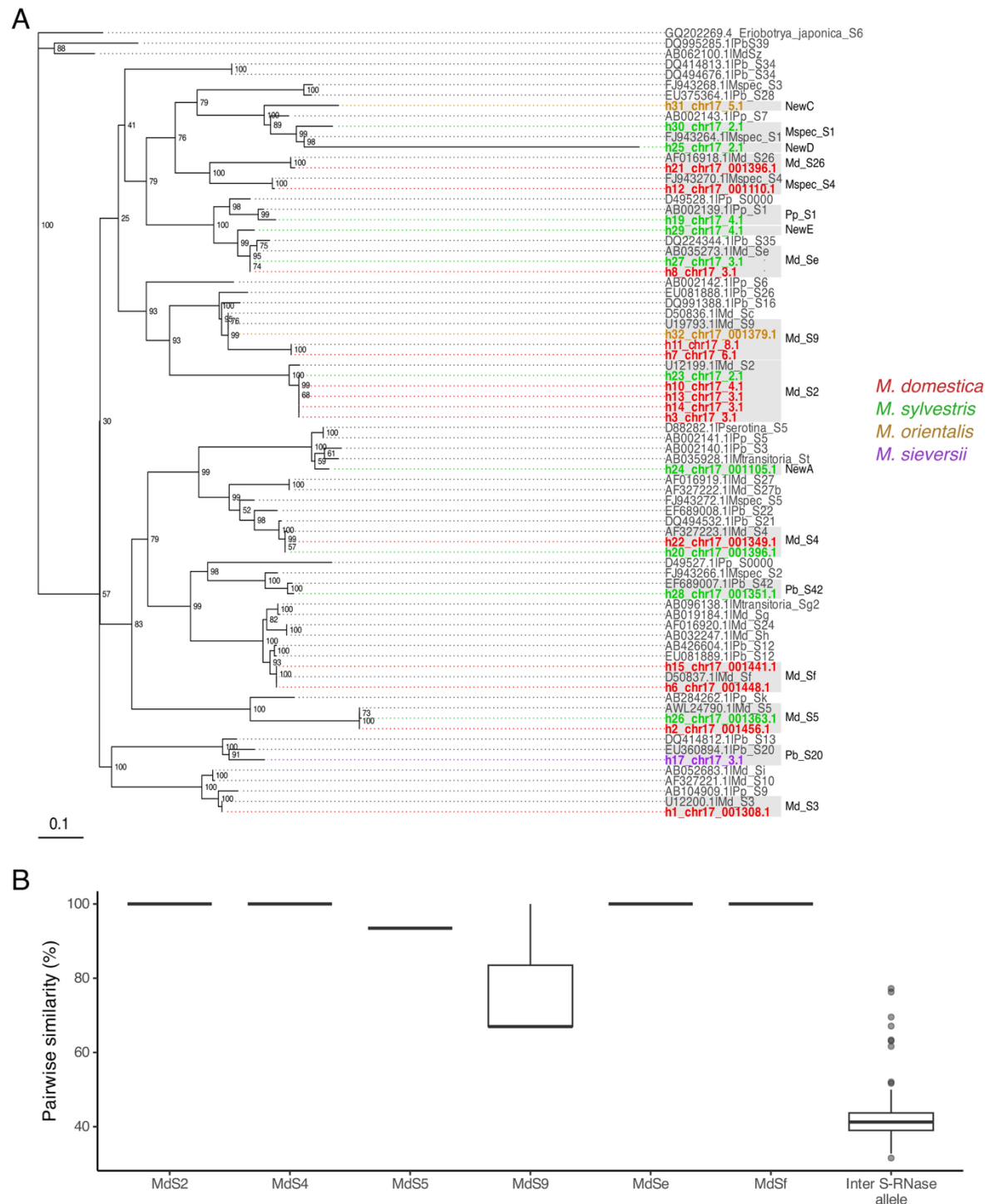
Accession	S allele	Species	Reference
ACS94938.4	EjS6	<i>Eriobotrya japonica</i>	Niska et al., 2010
AB002139.1	PpS1	<i>Pyrus pyrifolia</i>	Ishimizu et al., 1998 https://www.ncbi.nlm.nih.gov/pubmed/9700066
AB002140.1	PpS3	<i>Pyrus pyrifolia</i>	Ishimizu et al., 1998 https://www.ncbi.nlm.nih.gov/pubmed/9700066
AB002141.1	PpS5	<i>Pyrus pyrifolia</i>	Ishimizu et al., 1998 https://www.ncbi.nlm.nih.gov/pubmed/9700066
AB002142.1	PpS6	<i>Pyrus pyrifolia</i>	Ishimizu et al., 1998 https://www.ncbi.nlm.nih.gov/pubmed/9700066
AB002143.1	PpS7	<i>Pyrus pyrifolia</i>	Ishimizu et al., 1998 https://www.ncbi.nlm.nih.gov/pubmed/9700066
AB019184.1	MdSg	<i>Malus domestica</i>	Kitahara et al., 1999
AB032246.1	MdSd	<i>Malus domestica</i>	Kitahara et al., 2000
AB032247.1	MdSh	<i>Malus domestica</i>	Kitahara et al., 2000
AB035273.1	MdSe	<i>Malus domestica</i>	Matsumoto & Kitahara, 2000
AB035928.1	MtSt	<i>Malus transitoria</i>	Matsumoto et al., 2000
AB052683.1	MdSi	<i>Malus domestica</i>	Kitahara & Matsumoto, 2002
AB062100.1	MdSz	<i>Malus domestica</i>	Kitahara & Matsumoto, 2002
AB096138.1	MtSg	<i>Malus transitoria</i>	Matsumoto et al., 2000
AB104908.1	PpS8	<i>Pyrus pyrifolia</i>	Castillo et al., 2001
AB104909.1	PpS9	<i>Pyrus pyrifolia</i>	Castillo et al., 2002
AB236424.2	PcSq	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB236425.1	PcS1	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB236426.2	PcSr	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB236427.1	PcSd	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB236428.1	PcSe	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB236429.1	PcSb	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB236430.1	PcSa	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB236431.1	PcSa	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB236432.1	PcSk	<i>Pyrus communis</i>	Takasaki et al., 2006 https://www.ncbi.nlm.nih.gov/pubmed/16565843
AB258359.1	PcSc	<i>Pyrus communis</i>	Moriya et al., unpublished
AB258360.1	PcSg	<i>Pyrus communis</i>	Moriya et al., unpublished
AB258361.1	PcSi	<i>Pyrus communis</i>	Moriya et al., unpublished
AB258362.1	PcSm	<i>Pyrus communis</i>	Moriya et al., unpublished
AB258363.1	PcSn	<i>Pyrus communis</i>	Moriya et al., unpublished
AB258364.1	PcSp	<i>Pyrus communis</i>	Moriya et al., unpublished
AB258365.1	PcSs	<i>Pyrus communis</i>	Moriya et al., unpublished
AB258366.1	PcSt	<i>Pyrus communis</i>	Moriya et al., unpublished
AB284262.1	PpSk	<i>Pyrus pyrifolia</i>	Kim et al., 2007
AB426604.1	PpS12	<i>Pyrus pyrifolia</i>	Okada et al., unpublished
AB426606.1	PpSk	<i>Pyrus pyrifolia</i>	Okada et al., unpublished
AB540121.1	MsS33	<i>Malus sieversii</i>	Matsumoto et al., 2010
AB540122.1	MdS34	<i>Malus domestica</i>	Matsumoto et al., 2010
AB731592.1	PcS25	<i>Pyrus communis</i>	Takasaki-Yasuda et al., unpublished
AB779646.1	PcS0	<i>Pyrus communis</i>	Takasaki-Yasuda et al., unpublished
AB779647.1	PcS00	<i>Pyrus communis</i>	Takasaki-Yasuda et al., unpublished
AF016918.1	MdS26	<i>Malus domestica</i>	Verdoodt et al., unpublished
AF016919.1	MdS27	<i>Malus domestica</i>	Verdoodt et al., unpublished
AF016920.1	MdS24	<i>Malus domestica</i>	Verdoodt et al., unpublished

AF327221.1	MdS10	<i>Malus domestica</i>	Van Nerum et al. unpublished
AF327222.1	MdS27b	<i>Malus domestica</i>	Van Nerum et al. unpublished
AF327223.1	MdS4	<i>Malus domestica</i>	Van Nerum et al. unpublished
AY249428.2	PpS13	<i>Pyrus pyrifolia</i>	Tan et al., unpublished
AY250989.3	PbS21	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
D49527.1	PpS0000	<i>Pyrus pyrifolia</i>	Norioka et al., 1995
D49528.1	PpS000	<i>Pyrus pyrifolia</i>	Norioka et al., 1995
D50836.1	MdSc	<i>Malus domestica</i>	Sassa et al., 1996
D50837.1	MdSf	<i>Malus domestica</i>	Sassa et al., 1996
D88282.1	PsS5	<i>Pyrus serotina</i>	Sassa et Hirano., 1997
DQ135990.1	MdS31	<i>Malus domestica</i>	Kim et al., unpublished
DQ135991.1	MdS32	<i>Malus domestica</i>	Kim et al., unpublished
DQ224344.1	PbS35	<i>Pyrus bretschneideri</i>	Yang & Tan, unpublished
DQ414812.1	PbS13	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
DQ414813.1	PbS34	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
DQ494532.1	PbS21	<i>Pyrus bretschneideri</i>	Zhang et al., 2006
DQ494676.1	PbS34	<i>Pyrus bretschneideri</i>	Zhang et al., 2006
DQ839240.1	PuS35	<i>Pyrus ussuriensis</i>	Wuyun et al., unpublished
DQ991388.1	PbS16	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
DQ995285.1	PbS39	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EF566872.1	PsS28	<i>Pyrus sinkiangensis</i>	Zhang et al., unpublished
EF643630.1	PpS15	<i>Pyrus pyrifolia</i>	Zhang et al., unpublished
EF643631.2	PbS38	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EF643635.1	PbS16	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EF643636.1	PbS18	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EF643638.1	PbS19	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EF643639.1	PsS22	<i>Pyrus sinkiangensis</i>	Tan et al., unpublished
EF643640.1	PbS27	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EF643641.1	PuS30	<i>Pyrus ussuriensis</i>	Tan et al., unpublished
EF689006.1	PuS42	<i>Pyrus ussuriensis</i>	Tan et al., unpublished
EF689007.1	PbS42	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EF689008.1	PbS22	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU081888.1	PbS26	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU081889.1	PbS12	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU101462.1	PbS29	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU101463.1	PbS26	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU101464.1	PuS40	<i>Pyrus ussuriensis</i>	Tan et al., unpublished
EU101465.1	PpS5a	<i>Pyrus pyrifolia</i>	Tan et al., unpublished
EU101466.1	PbS17	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU117115.1	PpS12	<i>Pyrus pyrifolia</i>	Tan et al., unpublished
EU336979.1	PuS32	<i>Pyrus ussuriensis</i>	Tan et al., unpublished
EU336980.1	PbS39	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU360894.1	PbS20	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU375364.1	PbS28	<i>Pyrus bretschneideri</i>	Tan et al., unpublished
EU477839.1	PcS21	<i>Pyrus communis</i>	Sanzol, https://www.ncbi.nlm.nih.gov/pubmed/19096853
EU477840.1	PcS21	<i>Pyrus communis</i>	Sanzol, https://www.ncbi.nlm.nih.gov/pubmed/19096853
FJ943264.1	MspecS1	<i>Malus spectabilis</i>	Zhang et al., unpublished
FJ943266.1	MspecS2	<i>Malus spectabilis</i>	Zhang et al., unpublished
FJ943268.1	MspecS3	<i>Malus spectabilis</i>	Zhang et al., unpublished

FJ943270.1	MspecS4	<i>Malus spectabilis</i>	Zhang et al., unpublished
FJ943272.1	MspecS5	<i>Malus spectabilis</i>	Zhang et al., unpublished
FJ946628.1	PpS46	<i>Pyrus pyrifolia</i>	Tan et al., unpublished
GQ180466.1	PcS4	<i>Pyrus communis</i>	Sanzol, https://www.ncbi.nlm.nih.gov/pubmed/20701704
KF468727.1	MsiS1	<i>Malus sikkimensis</i>	Gu, unpublished
KF468730.1	MsiS7	<i>Malus sikkimensis</i>	Gu, unpublished
KF588567.1	PcS126	<i>Pyrus communis</i>	Nikzad Gharehaghaji et al., unpublished
U12199.1	MdS2	<i>Malus domestica</i>	Broothaerts et al., 1995
U12200.1	MdS3	<i>Malus domestica</i>	Broothaerts et al., 1995
U19793.1	MdS9	<i>Malus domestica</i>	Broothaerts et al., 1995

Supplementary Table 3: S-loci coordinates, predicted SFBB number and S-alleles in studied haplotypes

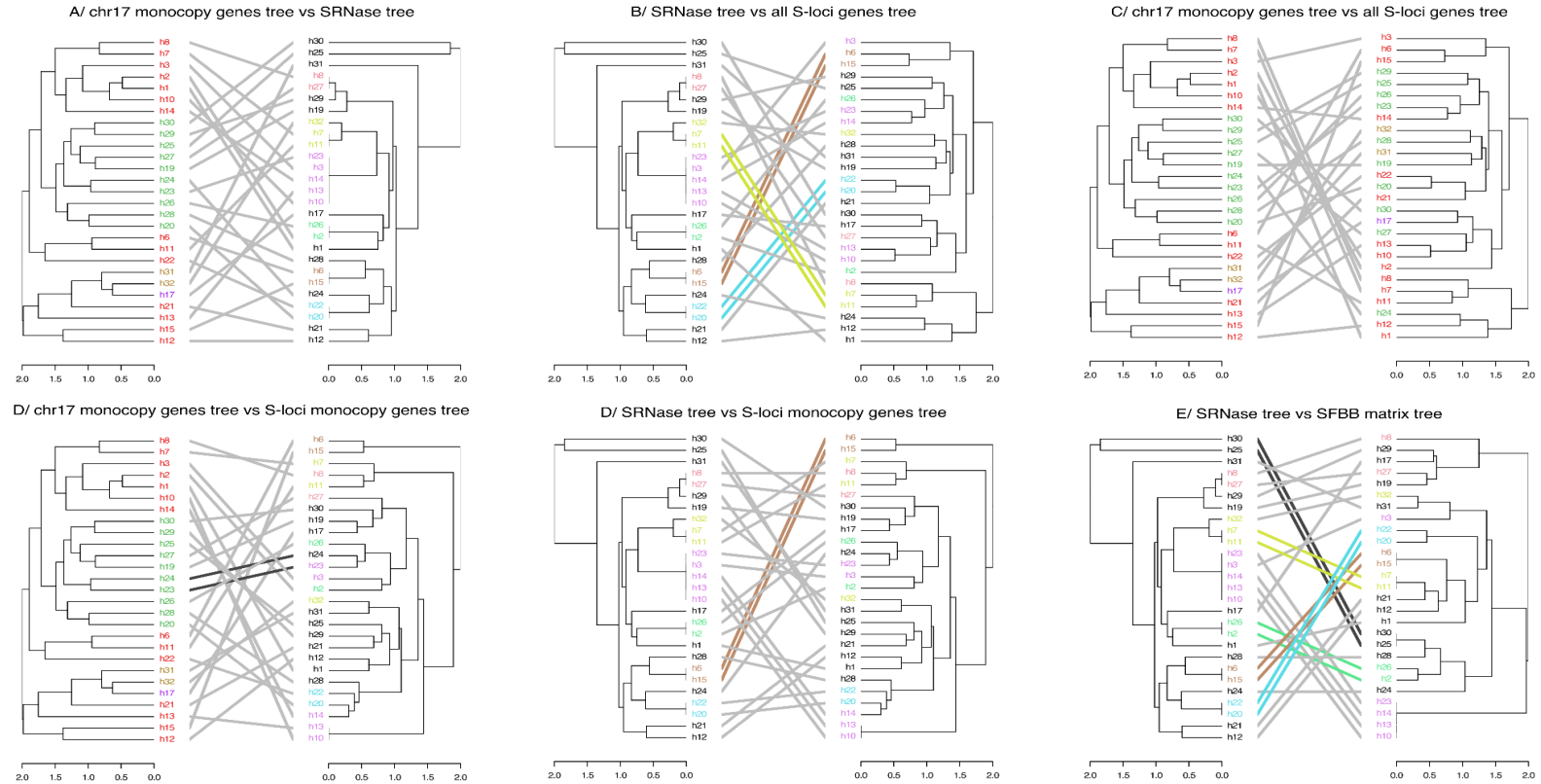
Haplotype	Species	Cultivar/ origin	S-loci start on chr17	S-loci end on chr17	S-loci length	Predicted S-loci gene number	Predicted SFBBs number (complete sequences)	S-Rnase ID	Predicted S-RNase allele
h14	<i>M. domestica</i>	Bardsey Island	30701244	32097319	1396076	39	17	h14_chr17_3.1	MdS2
h12	<i>M. domestica</i>	Brown snout	29604507	31323943	1719437	43	18	h12_chr17_001110.1	MspeS4
h13	<i>M. domestica</i>	Costard	29699373	31095434	1396062	38	17	h13_chr17_3.1	MdS2
h7	<i>M. domestica</i>	Fuji	30197526	31518020	1320495	40	17	h7_chr17_6.1	MdS9
h6	<i>M. domestica</i>	Fuji	29893470	31195680	1302211	48	17	h6_chr17_001448.1	MdSf
h3	<i>M. domestica</i>	Gala	26135286	27539957	1404672	51	17	h3_chr17_3.1	MdS2
h2	<i>M. domestica</i>	Gala	28665960	29985017	1319058	49	17	h2_chr17_001456.1	MdS5
h1	<i>M. domestica</i>	Golden delicious double haploid (GDDH13)	30032275	31732190	1699916	33	16	h1_chr17_001308.1	MdS3
h11	<i>M. domestica</i>	Hanfu	29830584	31150956	1320373	42	18	h11_chr17_8.1	MdS9
h10	<i>M. domestica</i>	Honeycrisp	29310889	30706999	1396111	39	17	h10_chr17_4.1	MdS2
h15	<i>M. domestica</i>	Newtons	29920233	31222329	1302097	43	18	h15_chr17_001441.1	MdSf
h8	<i>M. domestica</i>	Red delicious	29883495	30810778	927284	48	17	h8_chr17_3.1	MdSe
h21	<i>M. domestica</i>	Xuanina	29544619	30982290	1437672	42	18	h21_chr17_001396.1	MdS26
h22	<i>M. domestica</i>	Xuanina	28259688	29525297	1265610	40	18	h22_chr17_001349.1	MdS4
h32	<i>M. orientalis</i>	Armenia	29847786	30941904	1094119	36	19	h32_chr17_001379.1	MdS9
h31	<i>M. orientalis</i>	Armenia	31699461	32754781	1055321	45	19	h31_chr17_5.1	NewC
h17	<i>M. sieversii</i>	Kazakstan	28636332	29834722	1198391	36	17	h17_chr17_3.1	PbS13
h20	<i>M. sylvestris</i>		30157545	31217444	1059900	37	16	h20_chr17_001396.1	MdS4
h19	<i>M. sylvestris</i>		27161358	28640677	1479320	39	20	h19_chr17_4.1	PpS1
h23	<i>M. sylvestris</i>	Austria	30930378	32432552	1502175	47	16	h23_chr17_2.1	MdS2
h24	<i>M. sylvestris</i>	Austria	29911385	30929757	1018373	40	14	h24_chr17_001105.1	NewA
h26	<i>M. sylvestris</i>	Denmark	29352325	30455556	1103232	38	17	h26_chr17_001363.1	MdS5
h25	<i>M. sylvestris</i>	Denmark	30768546	32202276	1433731	42	18	h25_chr17_2.1	NewD
h28	<i>M. sylvestris</i>	France	28763989	30262854	1498866	37	17	h28_chr17_001351.1	PbS42
h27	<i>M. sylvestris</i>	France	31333056	32800349	1467294	41	19	h27_chr17_3.1	MdSe
h29	<i>M. sylvestris</i>	Romania	29817762	31260259	1442498	49	19	h29_chr17_4.1	NewE
h30	<i>M. sylvestris</i>	Romania	30593311	32469098	1875788	47	17	h30_chr17_2.1	MspeS1



Supplementary Fig. 1: Phylogeny and similarity patterns of Malus S-RNases. (A) Maximum-likelihood phylogeny of S-RNase protein sequences including both annotated haplotype sequences from the present dataset and publicly available S-RNases. Species of origin are indicated: *Md* = *Malus domestica*; *Mspec* = *M. spectabilis*; *Msie* = *M. sieversii*; *Pb* = *Pyrus breitschneideri*; *Pc* = *P. communis*; *Pp* = *P. pyrifolia*; *Pu* = *P. ussuriensis*. (B) Distribution of pairwise protein sequence similarities among S-RNases, computed in R based on amino acid identity (1 – distance). Boxplots show similarity values for S-RNases belonging to the same S-allele (intra-allelic comparisons) versus those belonging to different alleles (inter-allelic comparisons).

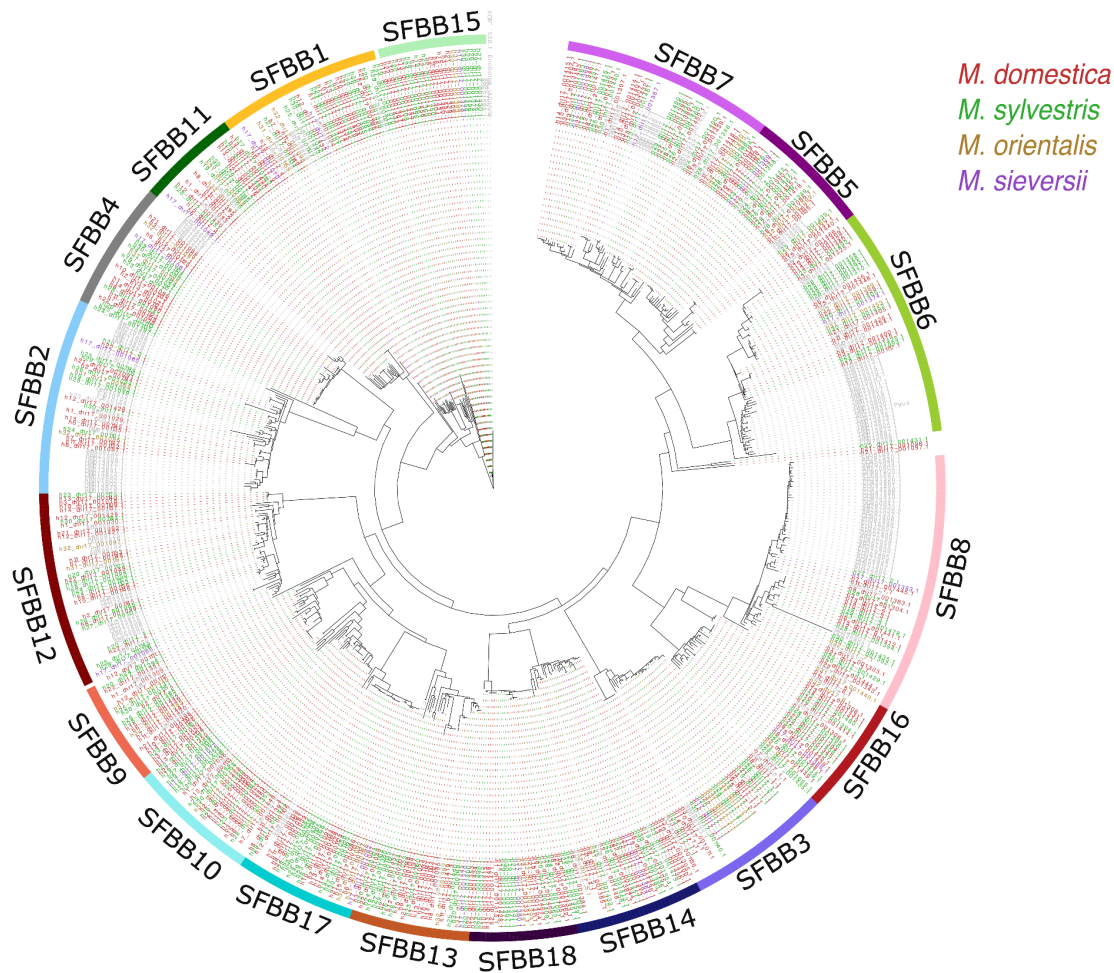
Supplementary Table 4: Expected and observed S-RNase alleles in *M. domestica* S-haplotypes from our dataset

Cultivar	Expected S allele (references)	Observed S alleles
Golden delicious	MdS2/MdS3 (Schneider et al., 2001 ; Ryan et al., 2020)	h1 (MdS3)
Fuji	MdS1/MdS9 (Kasajima et al., 2017 ; Ryan et al., 2020) or MdSf (Matsumoto et al., 1999a) or MdSc/MdSf (Sassa et al., 1994)	h6 (MdSf) / h7 (MdS9)
Honeycrisp	MdS2/MdS24 (Ryan et al., 2020)	h10 (MdS2)
Gala	MdS2/MdS5 (Matsumoto et al., 1999b) or MdS3/MdS5 (Ryan et al., 2020)	h2 (?) / h3 (MdS2)
Red delicious	MdS9/MdS28 (Ryan et al., 2020) MdSe = MdS28 (Kasajima et al., 2017)	h8 (?)
Newtons	Unknown	h15 (MdSf)
Brown snout	Unknown	h12 (MspeS4)
Hanfu	Unknown	h11 (MdS9)
Costard	Unknown	h13 (MdS3)
Bardsey Island	Unknown	h14 (MdS3)



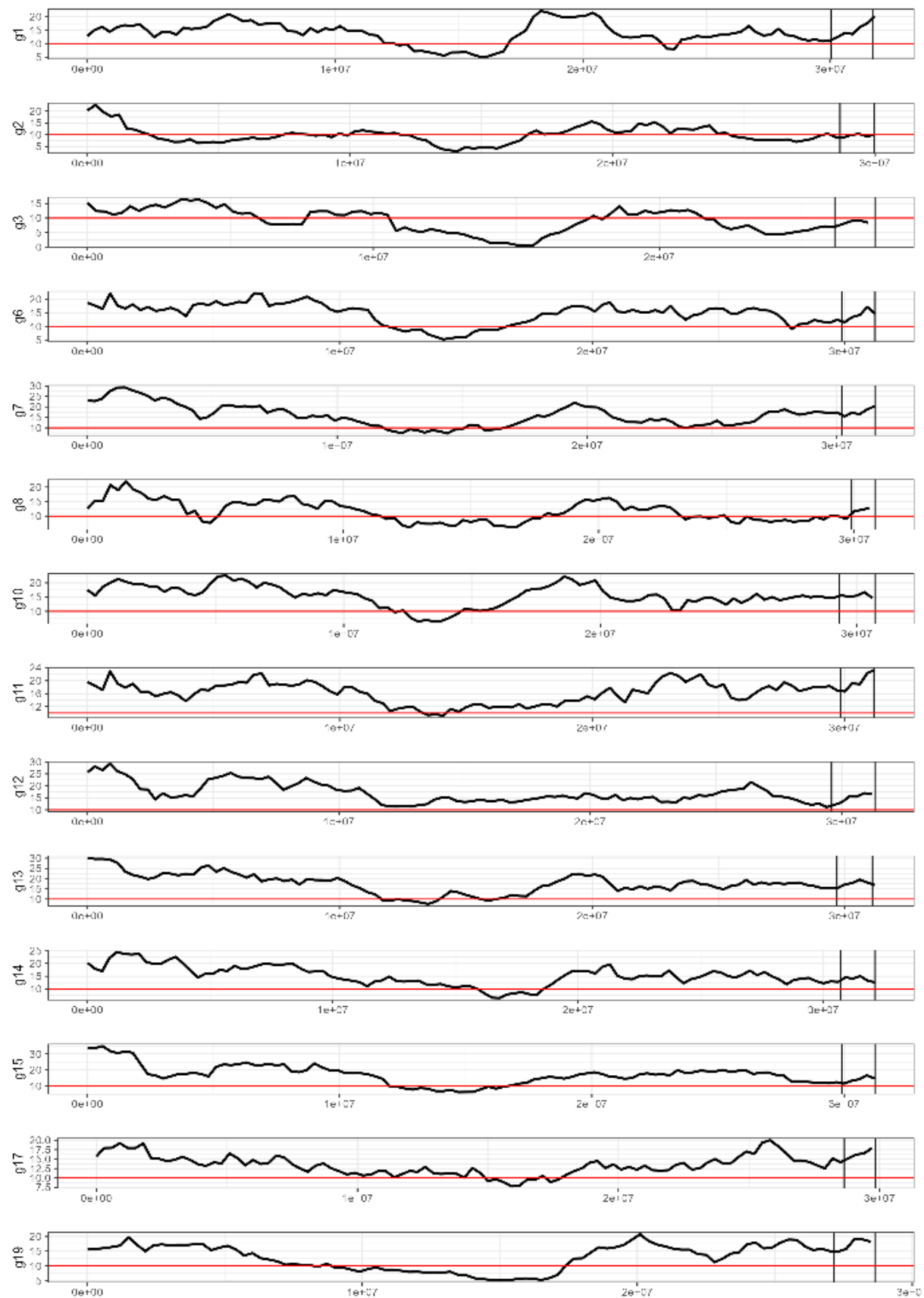
Supplementary Fig. 2: Tanglegram comparisons of phylogenetic trees related to S-locus components in apple trees. Tanglegrams illustrate topological comparisons between hierarchical clustering trees derived from different genomic features: **(A)** Species tree vs. *S-RNase* gene tree ; **(B)** *S-RNase* gene tree vs. tree based on all S-locus genes (including mono- and multi-copy genes) ; **(C)** Species tree vs. all S-loci gene tree **(D)** Species tree vs. tree based on S-locus single-copy genes only ; **(E)** *S-RNase* gene tree vs. S-locus single-copy genes tree ; **(F)** *S-RNase* gene tree vs. tree derived from presence/absence matrix of SFBB genes. Each line links identical haplotypes (h1 to h32) between the trees being compared. Colored connectors highlight haplotypes exhibiting strong or consistent topological congruence across comparisons. Horizontal axes indicate phylogenetic distances. Tree topologies were computed using hierarchical clustering based on pairwise genetic distances. Two independent color schemes are used to facilitate interpretation: 1. Species colors: Red = *M. domestica*; Green = *M. sylvestris*; Yellow = *M. orientalis*; Purple = *M. sieversii*. 2. S-allele colors: Colors correspond to specific S-alleles: Pink: MdS2 ; Brown: MdSf ; Light green: MdS5 ; Blue: MdS4 ; Light yellow: MdS9 ; Salmon: MdSe.

Supplementary Table 5: SFBB database, public accessions from *Malus* and *Pyrus* (excel file).



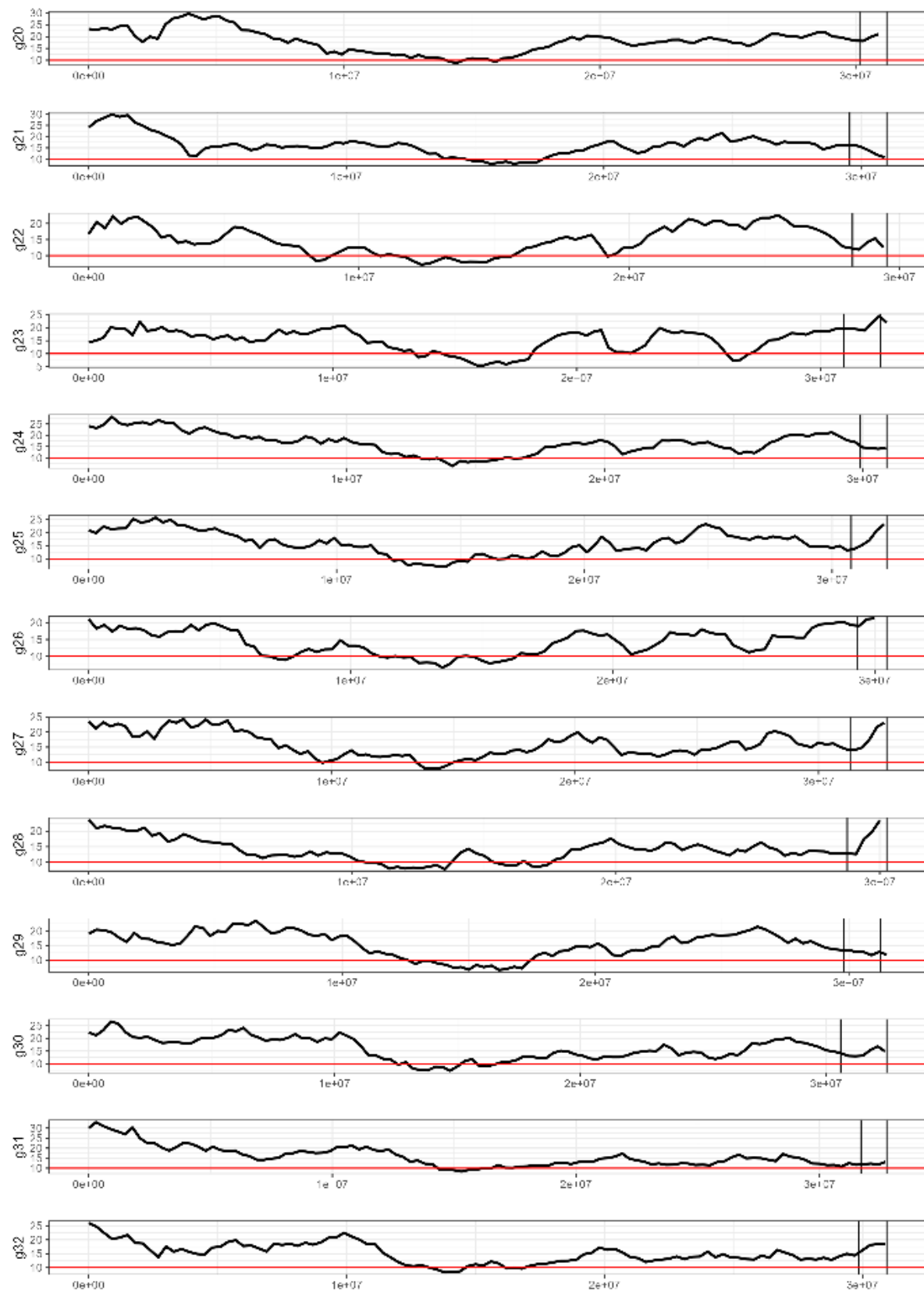
Supplementary Fig. 3: SFBB genes phylogeny and families determined using combination of classification methods. Phylogenetic tree based on the alignment of SFBB protein sequences from *Malus* S-haplotypes included in this study, along with publicly available SFBB sequences. Protein sequences were aligned using Clustal Omega (Sievers & Higgins, 2014), and the tree was inferred with IQ-TREE (Minh et al., 2020) using 1000 bootstrap replicates. Hierarchical clustering using Ward method was combined to Orthogroups identification using Orthofinder (Emms & Kelly, 2019) to determine SFBB families, which were compared to those of publicly available SFBB sequences. Tip labels are color-coded according to the species origin of the S-haplotype from which each SFBB sequence was derived.

Supplementary Table 6: SFBB from assembly set, coordinates and families (excel file).

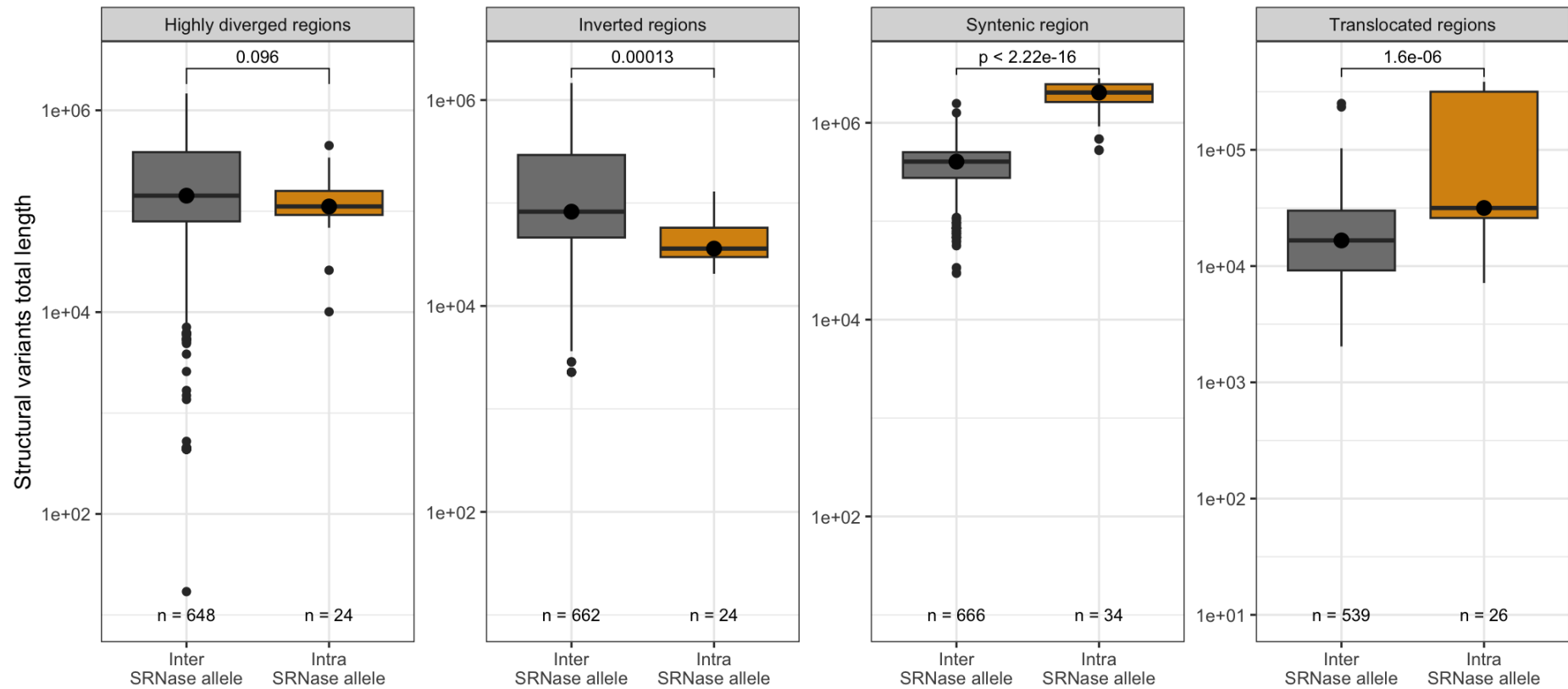


Supplementary Fig. 4: Variation of the LTR Assembly Index (LAI) along chromosome 17 for different *Malus* assemblies. Each panel corresponds to a distinct assembly. Black lines represent local LAI values, while

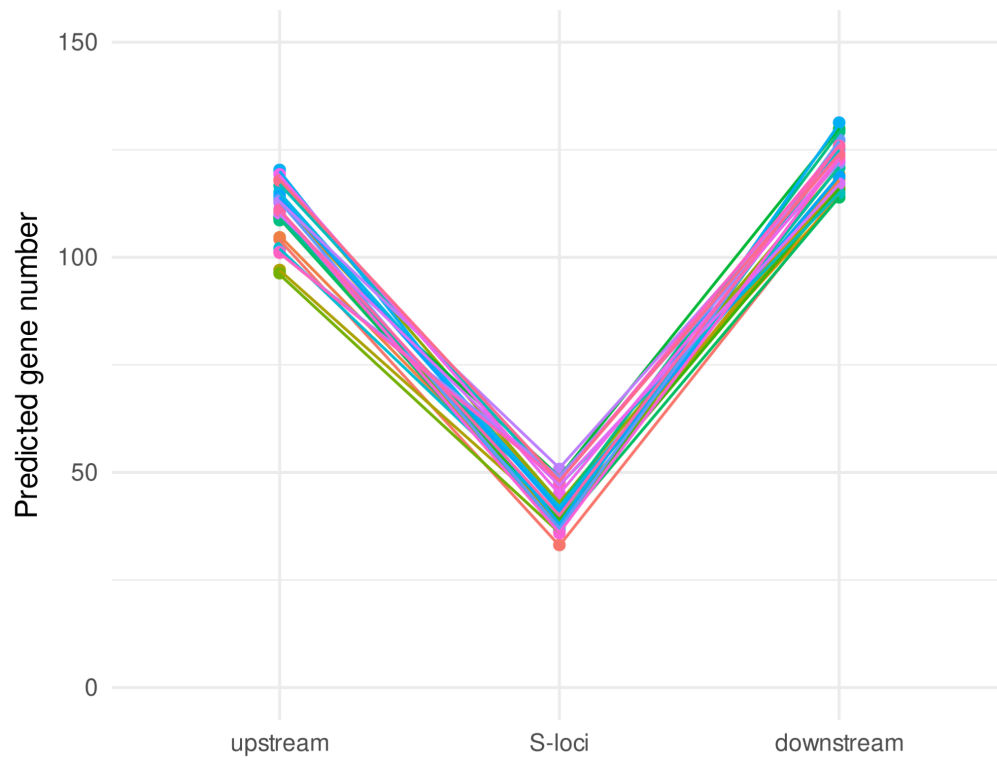
the horizontal red line indicates the quality threshold. The S-locus region is delimited by the two vertical bars at the right end of each plot.



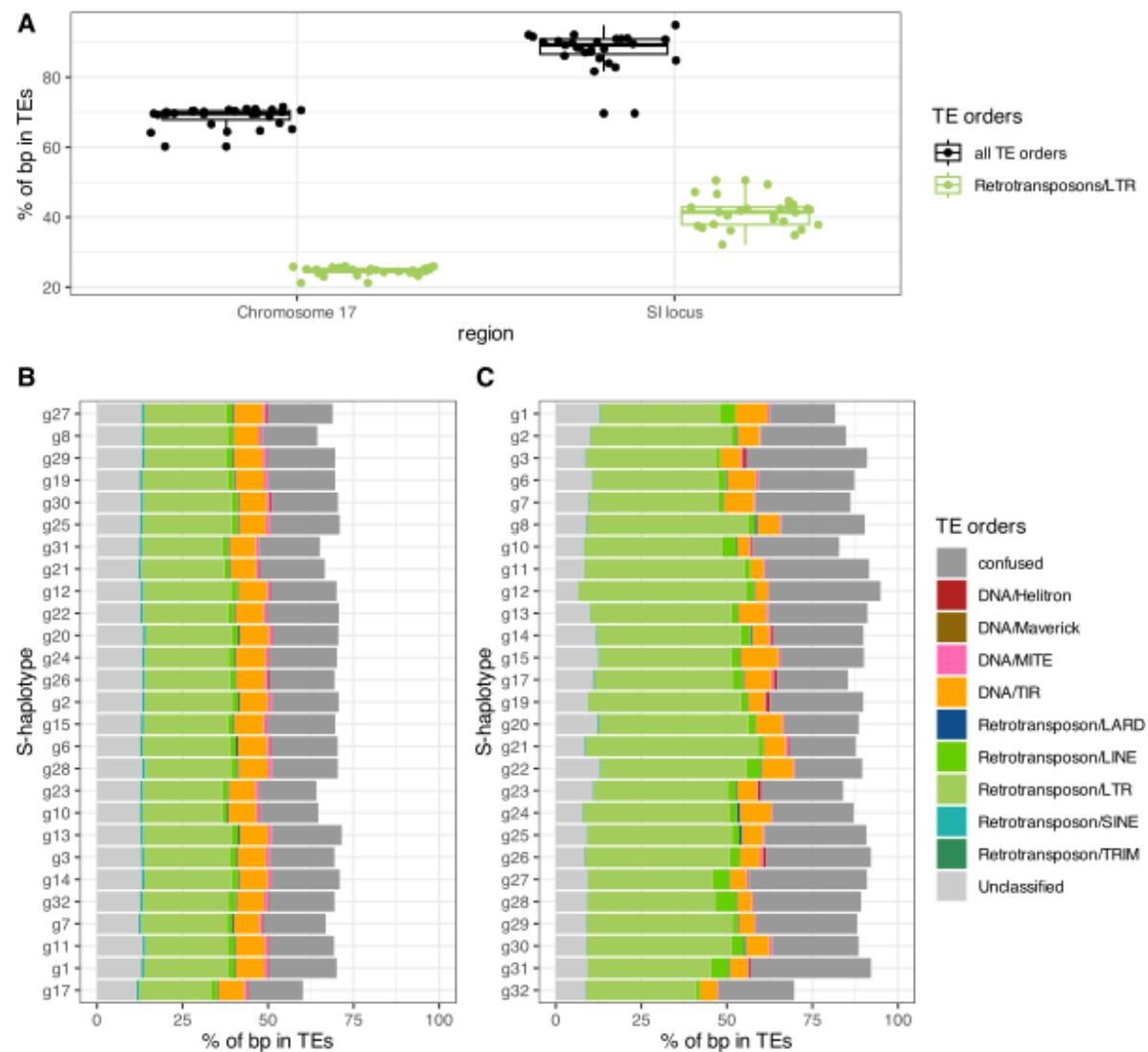
Supplementary Fig. 4 (continued)



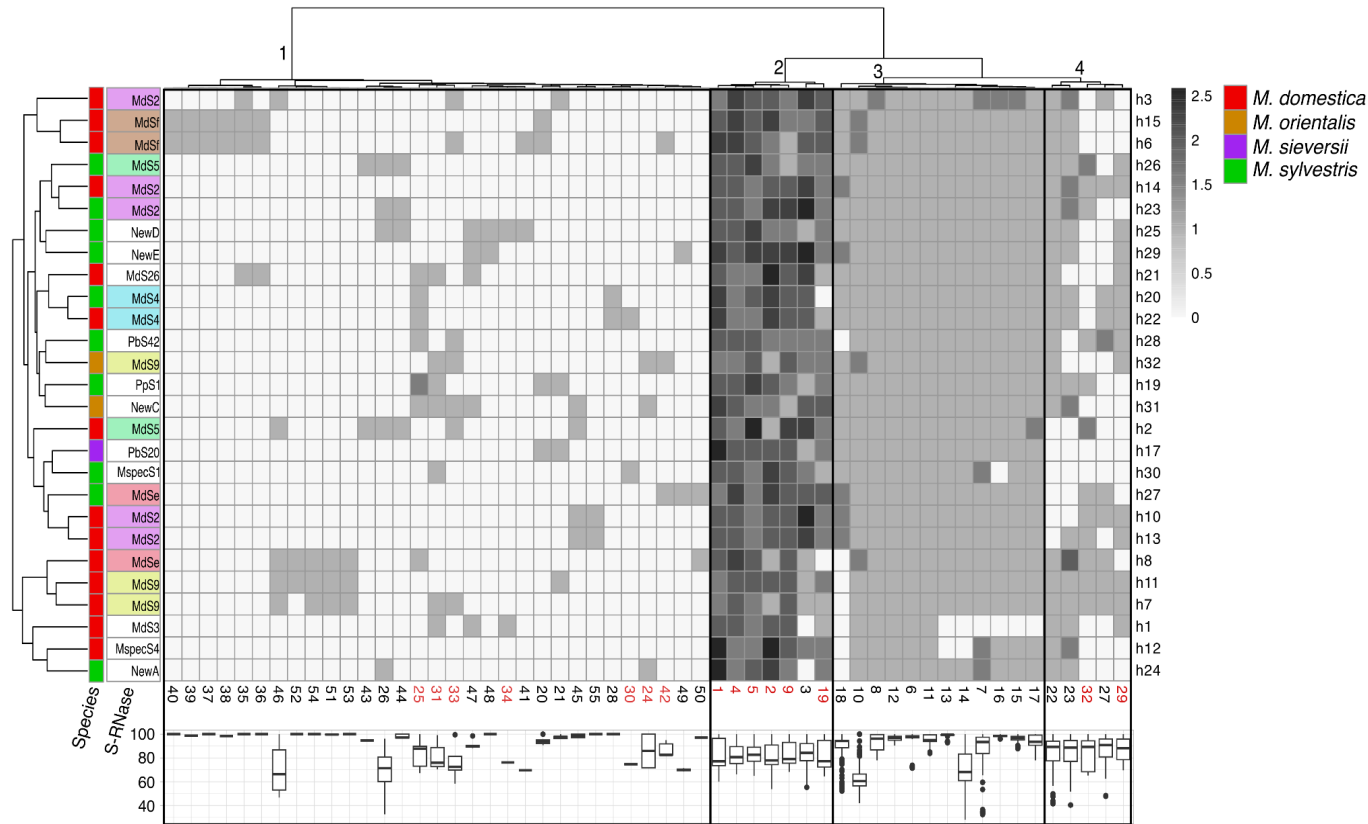
Supplementary Fig. 5: Structural variant and syntenic regions size distribution between S-locus haplotypes sharing the same or different S-RNase alleles. Boxplots showing the distribution of region sizes (in bp, log scale) for different types of structural variants detected in pairwise alignments of S-locus haplotypes, categorized by whether the two haplotypes share the same S-RNase allele (*intra S-RNase allele*, orange) or not (*inter S-RNase allele*, gray). Structural variant classes include inverted regions, translocated regions and highly diverged regions, as defined by SyRI. For each category, the median region size is indicated along with results from a Wilcoxon–Mann–Whitney test comparing intra vs. inter S-RNase groups. W statistics, p -values are reported above each plot. Syntenic regions are significantly larger between haplotypes with the same S-RNase allele (p -value $< 2.2 \times 10^{-16}$), whereas inverted regions are significantly larger (p -value $= 8.9 \times 10^{-5}$) in comparisons between different S-RNases, suggesting increased structural divergence between non-homologous S-haplotypes.



Supplementary Fig. 6: Predicted gene number inside S-loci and in same length regions (1.35 Mb) upstream and downstream. Each point is one haplotype region, and regions from the same haplotypes have the same color and are linked by the same color line.

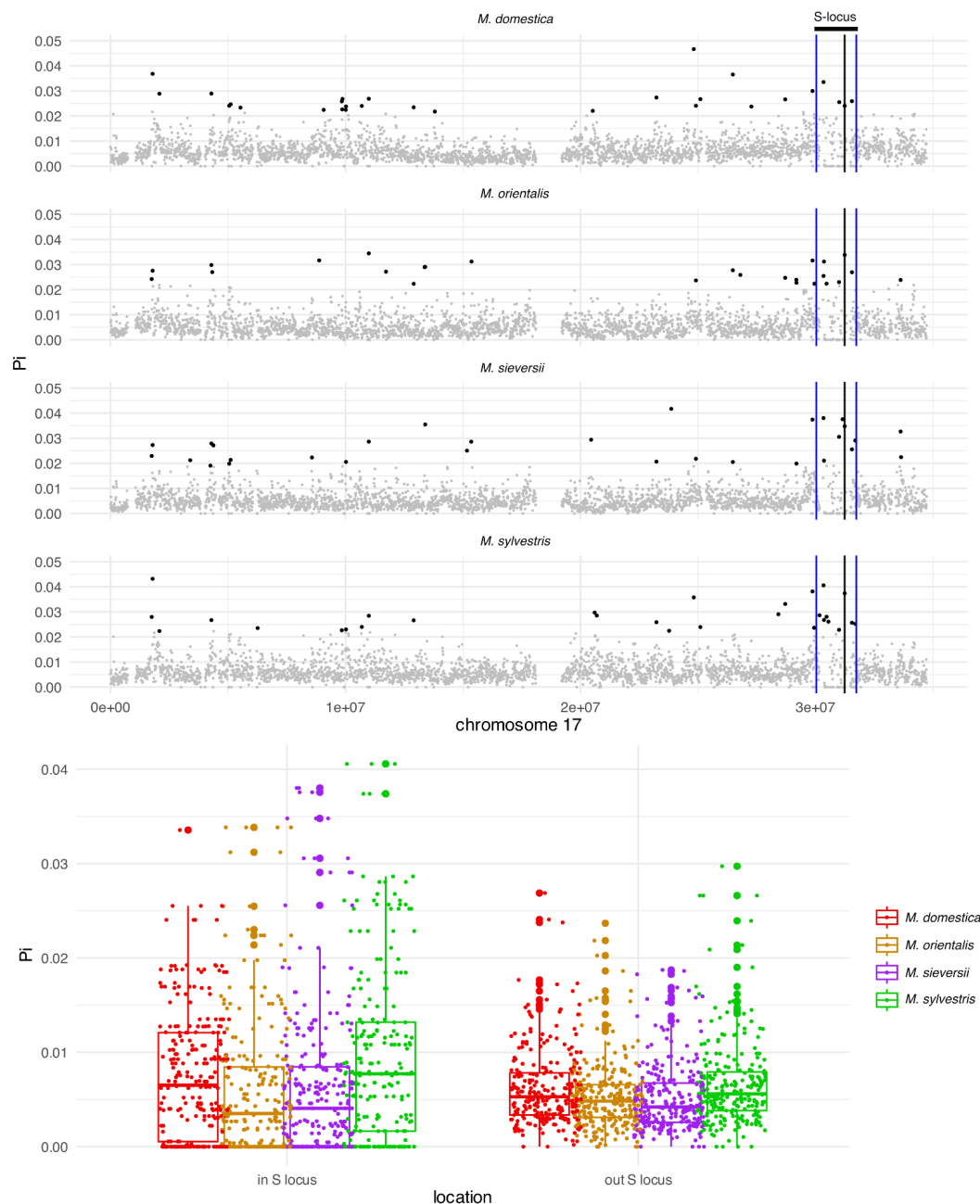


Supplementary Fig. 7: Chromosome 17 and S-loci TEs content in *Malus* haplotypes add7 window size. Base percentage in TE appears higher in S-loci than in chromosome 17 (A, black points) and the difference is more pronounced for LTR retrotransposons (A, green points). TEs families proportion appears conserved in different chromosome 17 (B) and less conserved across SI loci, even with the same S allele (C). The confused category corresponds to TEs for which the classification assigned multiple possible families without a clear preference for one family over another.

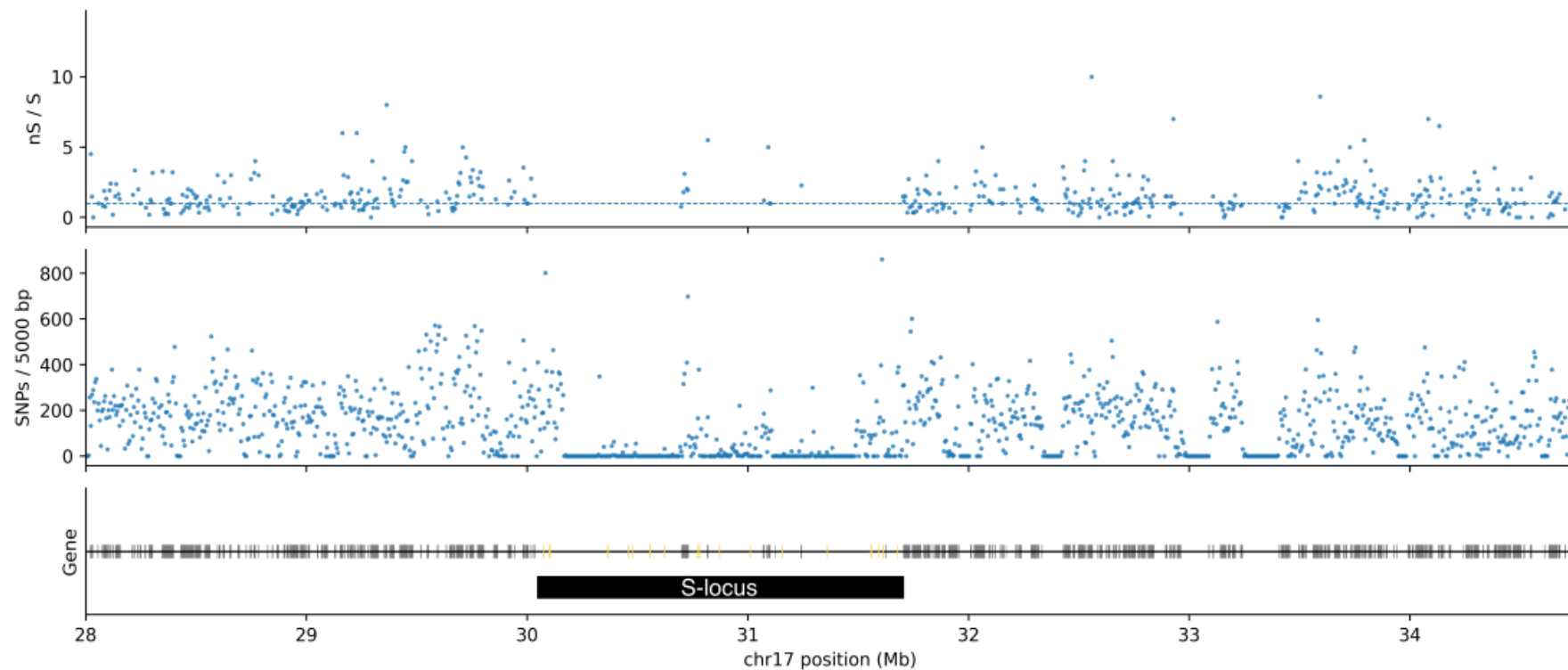


Supplementary Fig. 8: Heatmap showing the number of genes from each synteny cluster present in individual S-haplotypes. Each column represents a gene cluster inferred by syntenet from the *S*-locus region. Each row corresponds to an *S*-haplotype from one of four *Malus* species (*M. domestica*, *M. sylvestris*, *M. sieversii*, *M. orientalis*), annotated by species (left color strip) and *S*-allele (right color strip). The grayscale indicates the log-transformed gene count ($\log_2(n + 1)$) for each cluster within each *S*-haplotype. Cluster IDs are shown below each column; those highlighted in red contain known SFBB genes, which are involved in the self-incompatibility system. Both *S*-haplotypes and gene clusters were hierarchically clustered based on gene content similarity. Based on their distribution across *S*-haplotypes, gene clusters were grouped into four categories. Group 2 includes clusters 19, 3, 9, 2, 4, 1, and 5, which are present in all *S*-haplotypes in multiple copies. This group is composed mostly of SFBB gene clusters, with the exception of cluster 3, which contains genes encoding a serine/threonine-protein kinase and transcription factors. Group 3 contains clusters 15, 8, 16, 17, 7, 14, 13, 11, 6, 12, 10, and 18, which typically include a single gene copy per haplotype. These include the S-RNase gene (cluster 10), as well as genes encoding a DNA ligase, a C2H2-type zinc finger, and a coatomer subunit. Groups 1 and 4 consist of clusters that are not consistently present across all haplotypes. Clusters in group 1 are rarer than those in group 4. Notably, some SFBB gene clusters are also found in these two groups, indicating variable copy number and presence across haplotypes. The full gene composition of each cluster is provided in Supplementary Table 7.

Supplementary Table 7: Gene content of S-locus clusters across the 27 *Malus* haplotypes. For each cluster group and individual cluster, the table lists the gene identifiers present, their putative functions (e.g., SFBB family membership or other annotated domains), and the corresponding S-haplotypes when available (excel file).



Supplementary Fig. 9: Nucleotide diversity (π) on chromosome 17 across four *Malus* species. (Top) Genome-wide distribution of nucleotide diversity on chromosome 17 in *M. domestica*, *M. orientalis*, *M. sieversii* and *M. sylvestris*, calculated in 10 kb non-overlapping windows using pixy. Black dots represent the top 1% windows in terms of π , calculated separately for each species. The region corresponding to the S-locus (as annotated in GDDH13, Daccord et al. 2017, corresponding to haplotype h1 in this study) is highlighted in blue. In all four species, the number of top 1% windows falling within the S-locus was significantly higher than expected by chance based on 10,000 random permutations (*M. domestica*: p-value = $1.07e-2$, *M. orientalis*: $3e-4$, *M. sieversii*: $1e-4$, *M. sylvestris*: $1e-4$). (Bottom) Boxplots of nucleotide diversity for 1,000 randomly sampled windows located inside vs. outside the S-locus, for each species. Each dot represents a 10 kb window. Colors correspond to species.

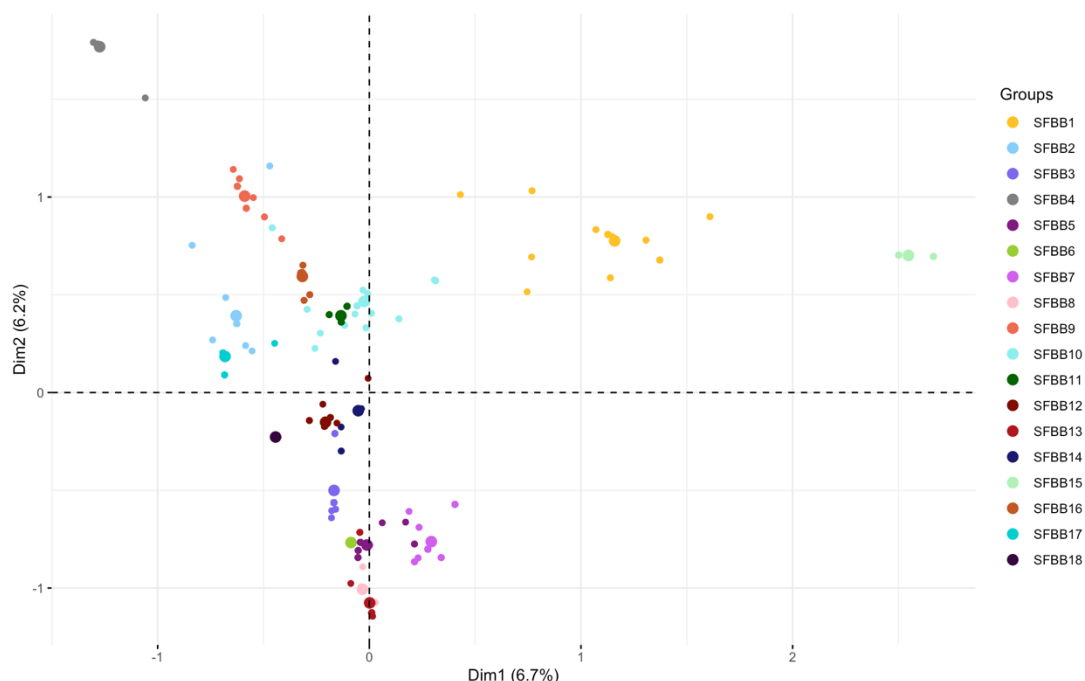


Supplementary Fig. 10: Distribution of SNP density and nS/S ratio along chromosome 17 in *Malus*. SNP density and the ratio of non-synonymous to synonymous variants (nS/S) were calculated in non-overlapping 5 kb windows along chromosome 17. (Top) Ratio of non-synonymous to synonymous substitutions (nS/S). The dashed horizontal line indicates a ratio of 1. (Middle) SNP density per 5 kb window. (Bottom) Gene track showing the position of annotated genes (gray rectangles) and SFBB genes (yellow rectangles) and the coordinates of the S-locus (black bar).

Supplementary Table 8: Detection of positive selection on SFBB genes in each S-haplotype using site models implemented in PAML / CodeML. Results of site models (M0, M1a, M2a, M3, M7, M8) fitted to codon alignments of SFBB genes from haplotypes h1 to h32. For each haplotype, the number of SFBB sequences included (N) and the total alignment length in codons (S) are indicated. **M0** assumes a single ω ratio for all sites and branches. **M1a** (neutral) and **M2a** (selection) compare models with and without a class of positively selected sites ($\omega > 1$). **M3** allows discrete site classes with different ω values. **M7** (beta distribution) and **M8** (beta + $\omega > 1$) are alternative site models for detecting selection. Model parameters shown include the log-likelihood (lnL), site class proportions (p_0, p_1, p_2), and dN/dS ratios ($\omega, \omega_0, \omega_1, \omega_2$) (excel file).

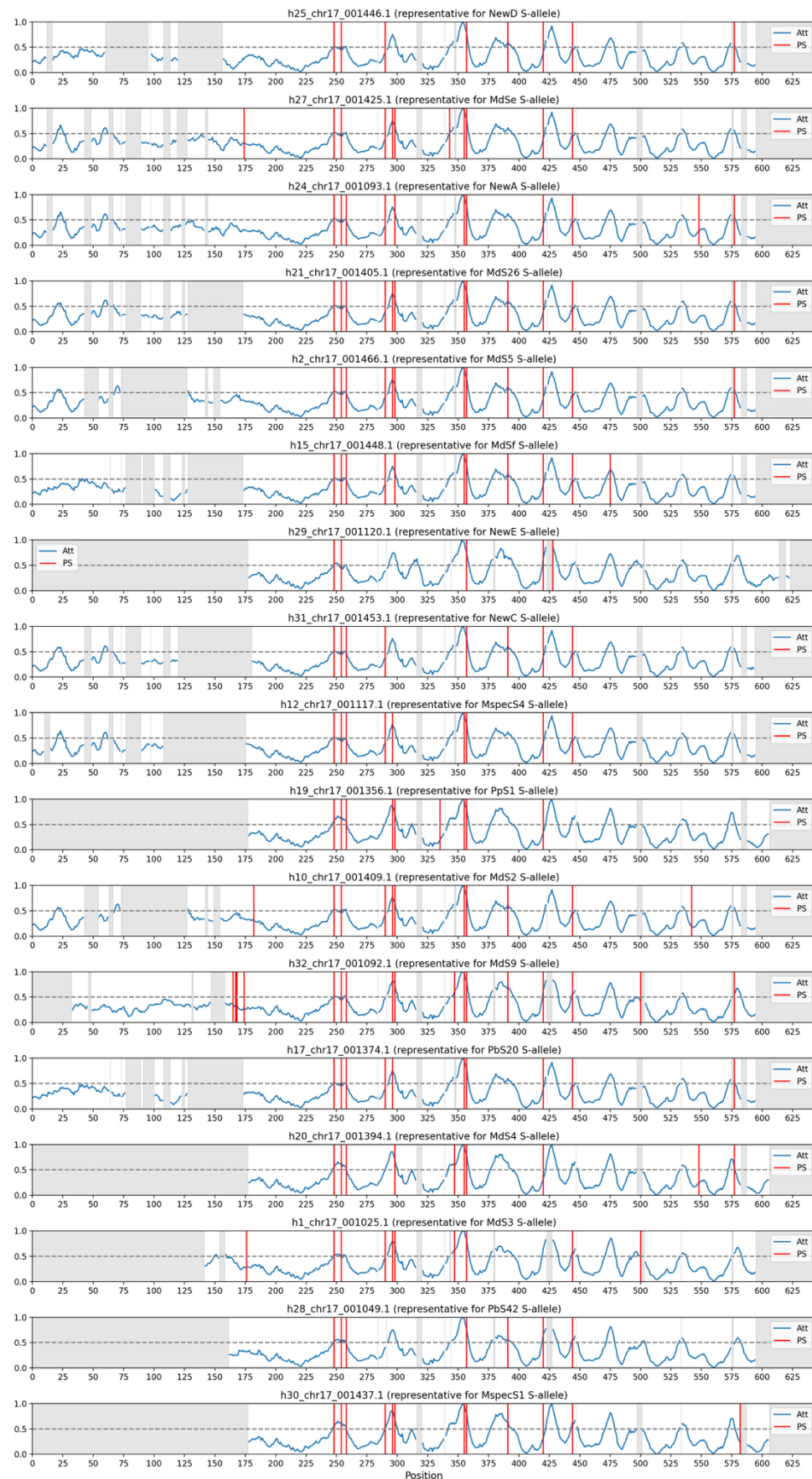
Supplementary Table 9: Likelihood ratio tests (LRTs) for detection of positive selection on SFBB genes in individual S-haplotypes. Results of model comparisons conducted using codeml (PAML) on codon alignments of SFBB genes from haplotypes h1 to h32. For each haplotype, three likelihood ratio tests were performed to evaluate the presence of **codons under positive selection**: **M0 vs M3**: compares a one-ratio model (M0) to a discrete model allowing for multiple site classes (M3), testing for variable selection pressure across sites. **M1a vs M2a**: tests for the presence of positively selected codons ($\omega > 1$) by comparing a neutral model (M1a) to a selection model (M2a). **M7 vs M8**: compares a beta-distributed ω model (M7, without selection) to a model allowing an extra site class with $\omega > 1$ (M8). For each comparison, the table reports log-likelihoods (lnL) of both models, the test statistic $-2\Delta\ell$, degrees of freedom (**K**), significance level (**alpha**), associated p -value, and the chi-squared critical value for 5% significance ($X^2(\alpha)$) (excel file).

Supplementary Table 10: Amino acids observed at positively selected sites across SFBB families in 27 S-haplotypes. For each of the 27 S-haplotypes, the table lists the amino acid residues found at the codon positions most frequently inferred to be under positive selection (posterior probability > 0.95 in PAML analyses). Each cell contains the amino acids present in the given haplotype for a given SFBB family; “absent” indicates that the SFBB gene was not detected in that haplotype (excel file).



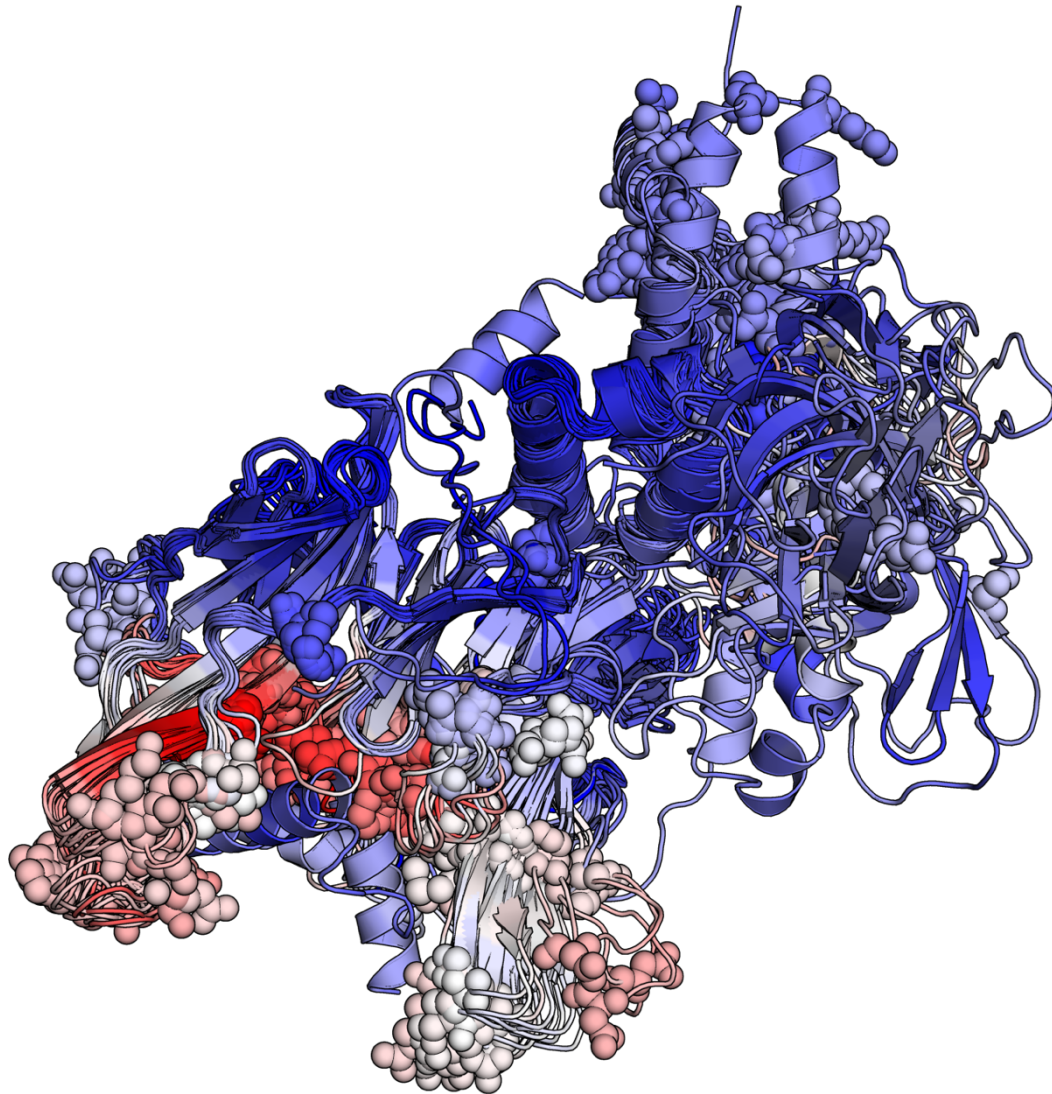
Supplementary Fig. 11. Clustering of SFBB genes based on amino acid identity at conserved positions. Multiple correspondence analysis (MCA) was performed using amino acid residues at selected conserved positions across SFBB genes. Each point represents a gene, colored according to its SFBB family. The first two dimensions, which together explain XX% of the total variance (Dim1: 6.7%, Dim2: 6.2%), reveal a structured separation of gene families. Families with similar amino acid profiles tend to cluster together, while others form

well-defined and distinct groups in the multivariate space. Ellipses represent 95% confidence intervals for each family.



Supplementary Fig. 12: Alignment of MuLAN attention scores and positively selected sites across *Malus* SFBB proteins. For SFBB genes associated with a distinct S-allele, we plotted MuLAN attention scores (blue)

along the amino acid sequence to estimate the predicted functional relevance of each residue. Vertical red lines indicate sites inferred to be under positive selection. Several of these sites co-localize with peaks in attention scores, suggesting that positively selected residues tend to coincide with positions predicted to have high functional impact. Gray-shaded regions represent alignment gaps or poorly aligned regions excluded from the analysis.

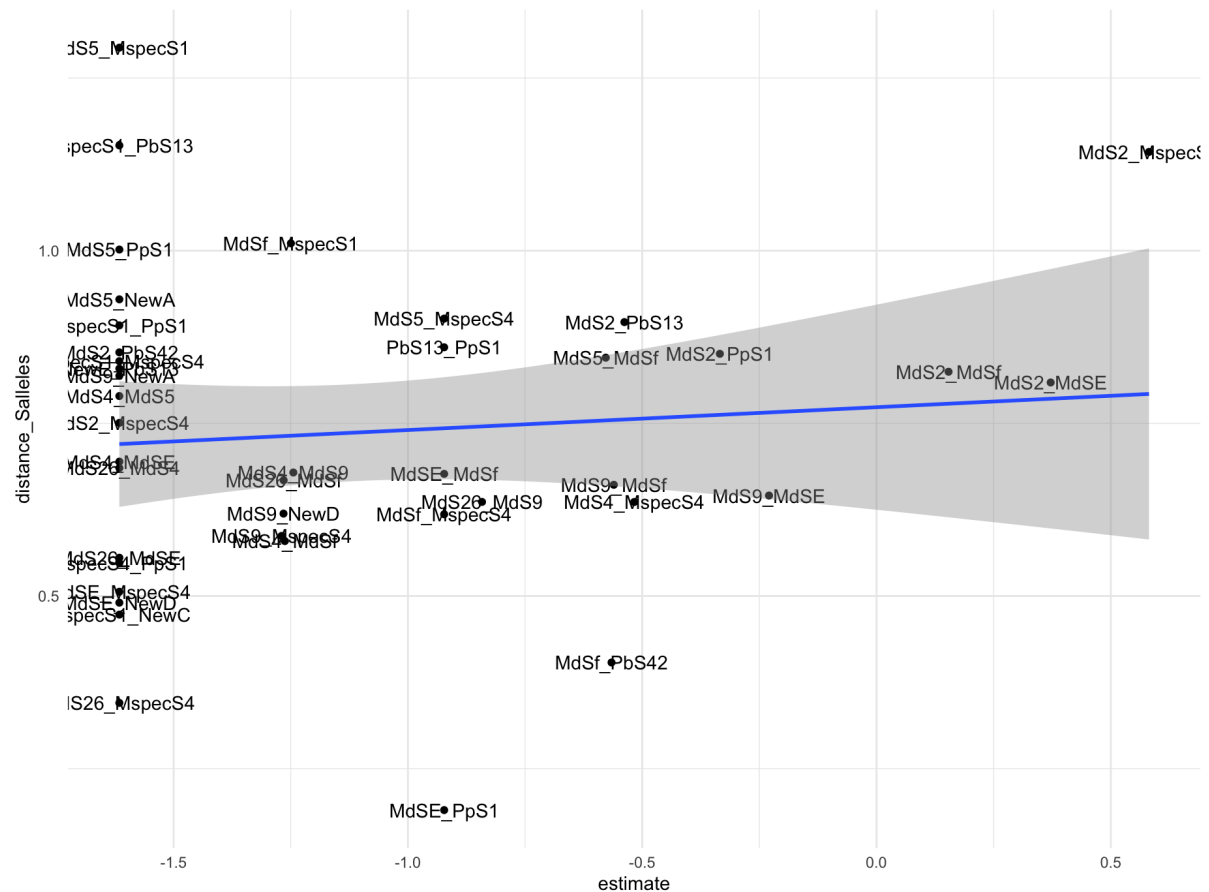


Supplementary Fig. 13: Structural representation of *Malus* SFBBs highlighting positively selected sites. Ribbon diagram of SFBB structures predicted by Boltz-1, showing the typical α/β fold. The longest sequence for each S-allele was chosen as representative, totalling 17 sequences, and the resulting structures were aligned in PyMOL to minimize pairwise root mean square deviation of atomic positions. The spheres indicate amino acid residues inferred to be under positive selection across SFBB families (posterior probability > 0.95, Bayes Empirical Bayes analysis, CodeML). These selected sites are broadly distributed on the surface of the protein, suggesting their potential involvement in allele-specific recognition or interaction with pistil determinants. Structural elements are colored according to MuLAN attention scores. Red-to-white colors correspond to high while blue tones to low scores.

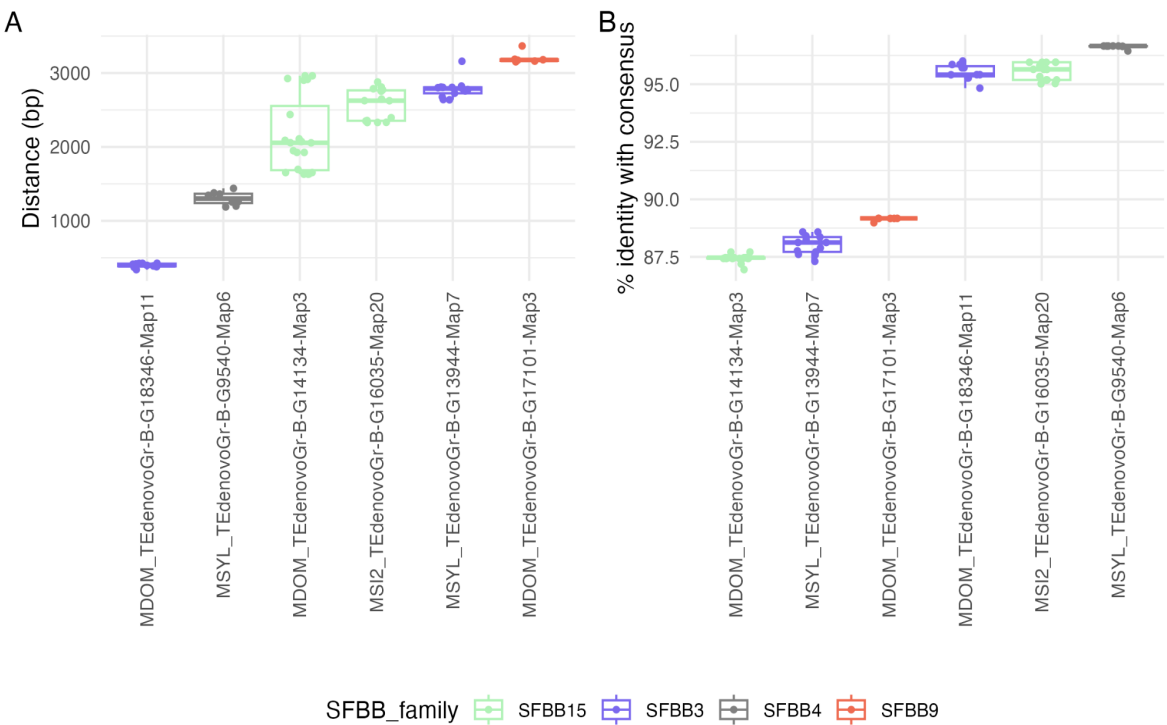
Supplementary Table 11: Amino acid composition at conserved codon positions across SFBB families. Matrix summarizing amino acid (AA) frequencies at ten codon positions previously identified as highly associated with SFBB family identity (see Table 2). Each row corresponds to a codon position, and each column represents one SFBB family. Within each cell, the observed amino acids and their relative frequencies ($\geq 3\%$) are listed, based on all SFBB gene sequences belonging to the corresponding family. Gray cells indicate positions that are invariant within a family (i.e., a single AA observed at 100% frequency). Rows on the right indicate the number and proportion of SFBB families for which each position is invariant, highlighting positions with family-specific conservation. Similarly, the bottom row shows the number and proportion of invariant positions per SFBB family. These results support the presence of family-specific amino acid signatures likely contributing to functional diversification among SFBB proteins.

SFBB family	SFBB 1	SFBB 2	SFBB 3	SFBB 4	SFBB 5	SFBB 6	SFBB 7	SFBB 8	SFBB 9	SFBB10	SFBB 11	SFBB 12	SFBB 13	SFBB 14	SFBB 15	SFBB 16	SFBB 17	SFBB 18	Number of invariant family	Percentage
position 297	F:1	H:0.82 , L:0.14, N:0.05	P:0.74, H:0.26	P:1	N:1	N:1	N:1	E:0.96, G:0.04	F:1	F:1	P:1	H:0.97, Q:0.03	E:0.89, D:0.11	P:1	F:1	F:1	Y:1	P:1	13	72%
position 299	Q:0.61 , E:0.33, H:0.06 , H:0.48 , P:0.42, C:0.03 , D:0.03 , R:0.03	S:0.95, W:0.05	Q:0.81 , H:0.19	V:1	E:1	Q:1	E:0.9, D:0.1	E:1	L:1	L:0.93, M:0.03, Q:0.03	Q:1	L:0.84, I:0.16	E:1	Q:1	E:1	Q:1	L:1	L:1	12	67%
position 336		D:0.86 , N:0.09 , S:0.05	N:0.81 , D:0.19	D:1	D:0.92 , N:0.08	D:1	N:0.67 , D:0.23 , Y:0.1	D:1	D:0.65 , N:0.26 , V:0.09	D:0.63, H:0.17, A:0.07, N:0.07, V:0.07	D:0.5, A:0.27, S:0.18, T:0.05	D:1	N:0.85, D:0.11, S:0.04	D:1	H:1	D:0.9, N:0.1	S:1	D:1	8	44%
position 356	I:0.97, V:0.03	T:0.91, N:0.05 , S:0.05	T:1	S:1	V:0.8, I:0.12, T:0.08	L:1	T:0.82, I:0.18	T:1	S:0.83, T:0.13, N:0.04	T:1	I:1	D:1	T:1	I:1	I:1	T:1	V:0.96, I:0.04	T:1	12	67%
position 358	G:0.85 , A:0.06 , I:0.06, V:0.03 , E:0.82, D:0.06	Q:1	H:1	G:1	K:1	K:1	K:1	K:1	Q:1	Q:0.97, L:0.03	E:1	G:1	K:1	G:0.9, R:0.1	T:1	E:1	R:1	R:1	15	83%
position 421	G:0.06 , K:0.03 , R:0.03 , H:0.88	D:0.95 , Y:0.05	S:0.81, T:0.19	K:1	D:1	D:1	E:1	K:0.96 , I:0.04	T:0.96, R:0.04	Q:0.7, D:0.1, E:0.07, N:0.07, K:0.03, T:0.03	K:1	K:1	K:1	K:1	E:1	T:0.79, N:0.14, K:0.07	K:1	K:1	11	61%
position 429	D:0.06 , Y:0.06	H:0.05 , S:0.05	P:1	L:1	P:1	P:1	P:1	P:1	S:1	H:0.83, S:0.13, T:0.03	Q:1	P:1	P:1	P:1	C:1	S:1	S:0.96, T:0.04	P:1	14	78%
position 476	E:0.88, K:0.12	L:0.95, T:0.05	N:0.81 , K:0.19	Q:1	K:0.96 , E:0.04	K:1	K:0.95 , N:0.05	K:1	R:1	T:0.97, K:0.03	K:1	E:1	K:1	K:0.97, R:0.03	K:1	T:1	K:1	K:1	11	61%

position 543	T:0.85, S:0.15	S:1	C:0.94 , F:0.03, Y:0.03	S:1	C:0.64 , Y:0.2, F:0.16	C:1	C:0.97 , S:0.03	R:0.74 , H:0.19 , C:0.07	S:0.78, N:0.17 , D:0.04	C:0.9, S:0.07, Y:0.03	S:1	C:0.9, G:0.1	H:0.85, C:0.11, Q:0.04	C:0.93, H:0.07	T:0.72, I:0.28	S:0.93, N:0.07	S:1	G:1	6	33%
position 549	V:0.91 , F:0.06, A:0.03	V:0.95 , I:0.05	V:1	A:0.8, T:0.15, V:0.05	L:1	L:1	L:1	I:1	L:1	L:0.9, M:0.07, V:0.03	V:1	I:0.77, L:0.19, F:0.03	I:0.89, L:0.11	L:1	E:1	L:1	L:0.82, I:0.18	L:1	11	61%
Number of invariant positions	1	2	4	9	6	10	5	7	6	2	9	6	6	7	9	7	7	10	113	63%
Percentage	10%	20%	40%	90%	60%	100%	50%	70%	60%	20%	90%	60%	60%	70%	90%	70%	70%	100%		



Supplementary Fig. 14. Relationship between phylogenetic distance of S-RNase alleles and frequency of gene conversion between S-haplotypes. Each point represents a pair of S-haplotypes, for which gene conversion events were inferred. The x-axis shows the estimate of gene conversion frequency (from the quasi-Poisson model), and the y-axis represents the cophenetic distance between the corresponding S-RNase alleles, extracted from the phylogenetic tree (Supplementary Fig. 1). The blue line shows the linear regression fit with 95% confidence interval (grey area). No significant correlation was detected (Pearson's $r = 0.09$, $p = 0.57$), suggesting that phylogenetic relatedness of S-RNases does not predict gene conversion intensity between their associated SFBB genes.

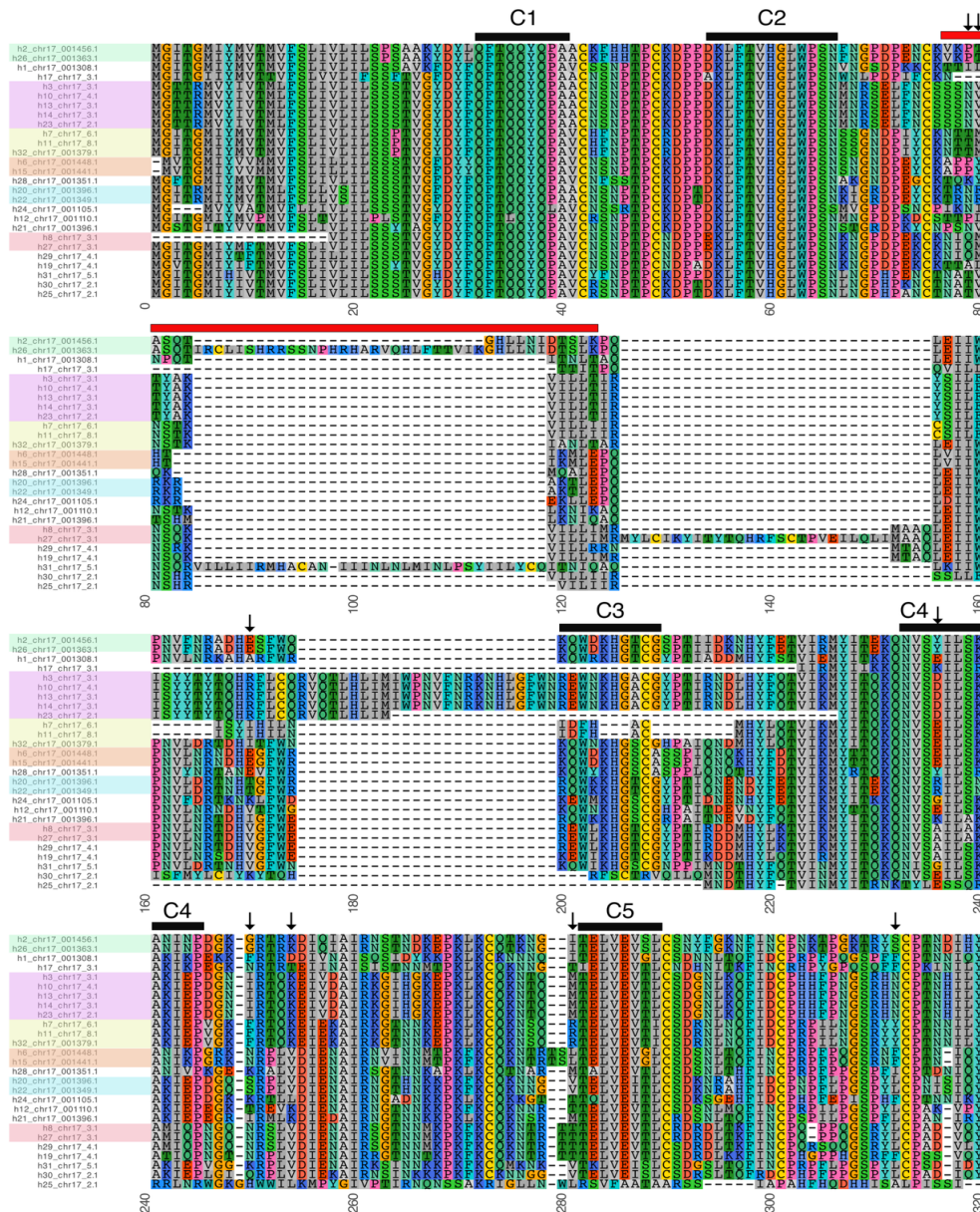


Supplementary Fig. 15: Distance and sequence identity of TE copies linked to SFBB gene families. (A) Distribution of distances (in base pairs) between SFBB genes and their associated TE copies, shown separately for each SFBB family. (B) Sequence identity (%) between TE copies associated with SFBB genes and the corresponding TE family consensus sequence.

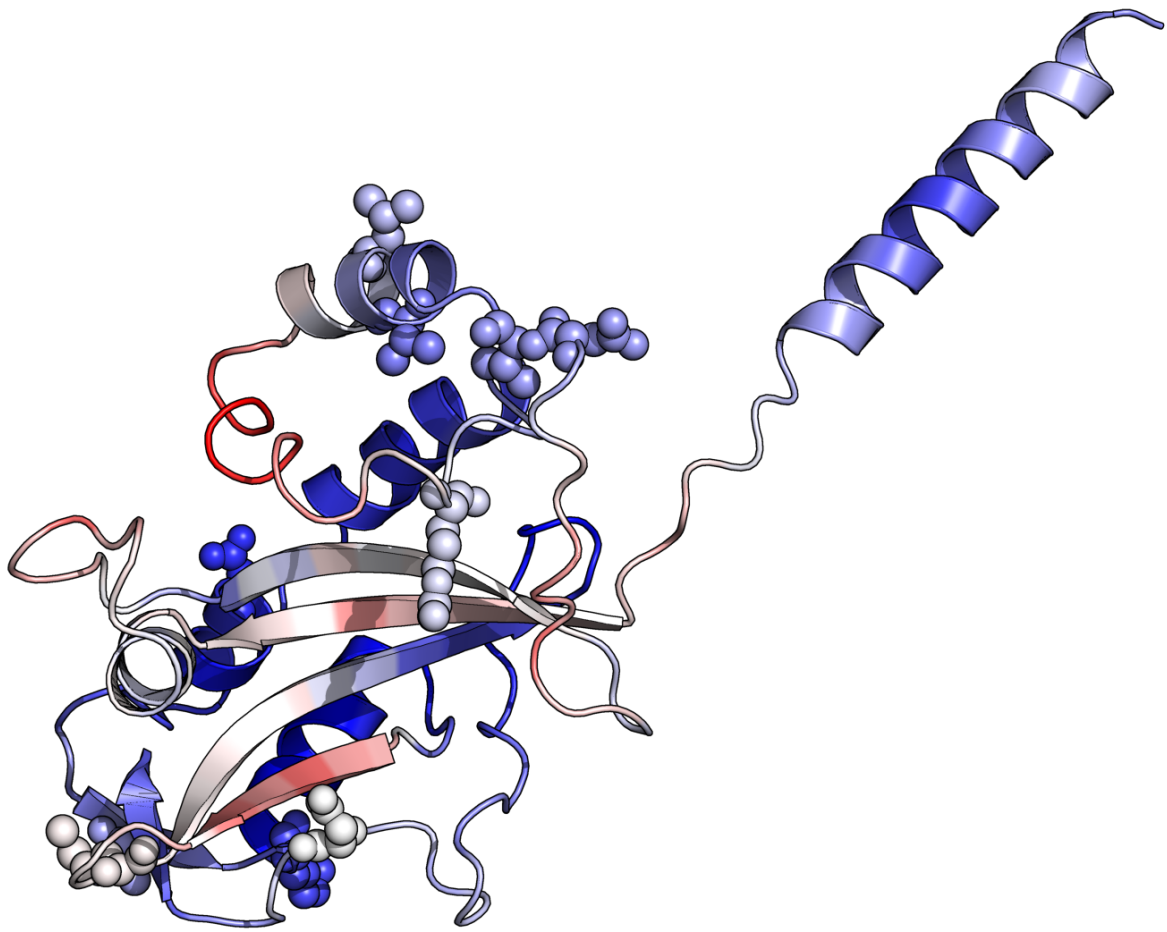
Supplementary Table 13: Transposable elements (TEs) detected in association with SFBB genes in the *Malus S*-locus. Each row represents a TE consensus identified in proximity (< 4kb) to SFBB gene families across the 27 haplotypes analyzed. The columns indicate the class (I = retrotransposon, II = DNA transposon), order (e.g., LTR, TIR, MITE), superfamily (e.g., Ty/Gypsy, hAT, MuDR), and the closest match in RepBase (if any), which provides a putative annotation based on sequence similarity. “unclassified” indicates elements that could not be confidently assigned to a known TE category, and “-” marks missing or uncertain information.

TE REPET Consensus ID	Consensus size	class	order	superfamily	Match on RepBase
MSYL_TEdenovoGr-B-G9540-Map6	1818	II	TIR	hAT	HAT-4_Mad
MDOM_TEdenovoGr-B-G18346-Map11	691	II	TIR	hAT	HAT-10_Mad
MSI2_TEdenovoGr-B-G16035-Map20	649	II	TIR	MuDR	MuDR-4_Mad
MSYL_TEdenovoGr-B-G13944-Map7	574	I	LTR	Ty/Gypsy	Gypsy-38_Mad
MDOM_TEdenovoGr-B-G14134-Map3	371	II	MITE	-	-
MDOM_TEdenovoGr-B-G17101-Map3	484	unclassified	unclassified	unclassified	-

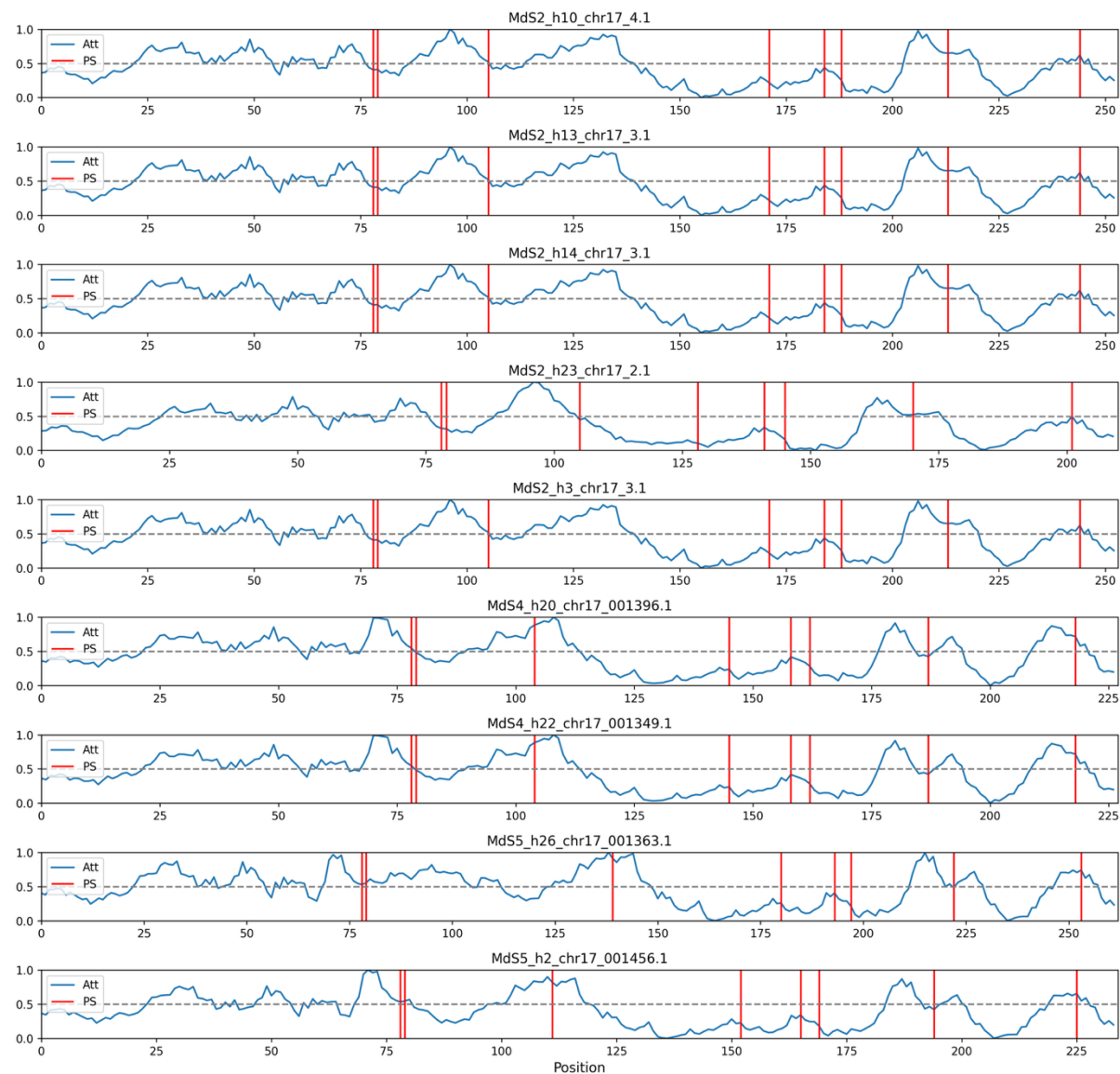
Supplementary Table 12: Contextual information for TE–SFBB associations in the S-locus of *Malus*. This table reports only TE–SFBB pairs identified as significant in the permutation test. Each row corresponds to a TE copy located within 4 kb of an SFBB gene. For each TE, we report its consensus family, TE order, identity to the consensus, and haplotype of origin, along with the nearest upstream and downstream SFBB families and their respective distances. The relative position of the SFBB gene with respect to the TE (upstream, downstream, overlapping, or inside) is indicated. Haplotype metadata include species of origin and associated S-RNase allele. “Enriched” TE consensus and SFBB families pairs (excel file).



Supplementary Fig. 16: Multiple *Malus* S-RNase amino-acids sequence alignment. The background color of sequence names reflects their allele, with identical colors indicating the same S-RNase specificity. Each amino acid is represented by its one-letter code. Background colors indicate amino acid chemical properties, with similar colors assigned to residues of similar physicochemical characteristics (e.g., hydrophobic in grey, aromatic in cyan, acids in red, basics in blue, polar in green, specials in yellow, orange and pink). The alignment is divided into consecutive blocks of 80 amino acids for readability. Positions under diversifying selection are marked with an arrow (↓) above the alignment columns. Gaps introduced during the alignment process are shown in white. The putative hypervariable region (in red) defined by Ishimizu et al. (1998) and Ushijima et al. (1998) and five conserved regions (C1, C2, C3, C4 and C5) are underlined.



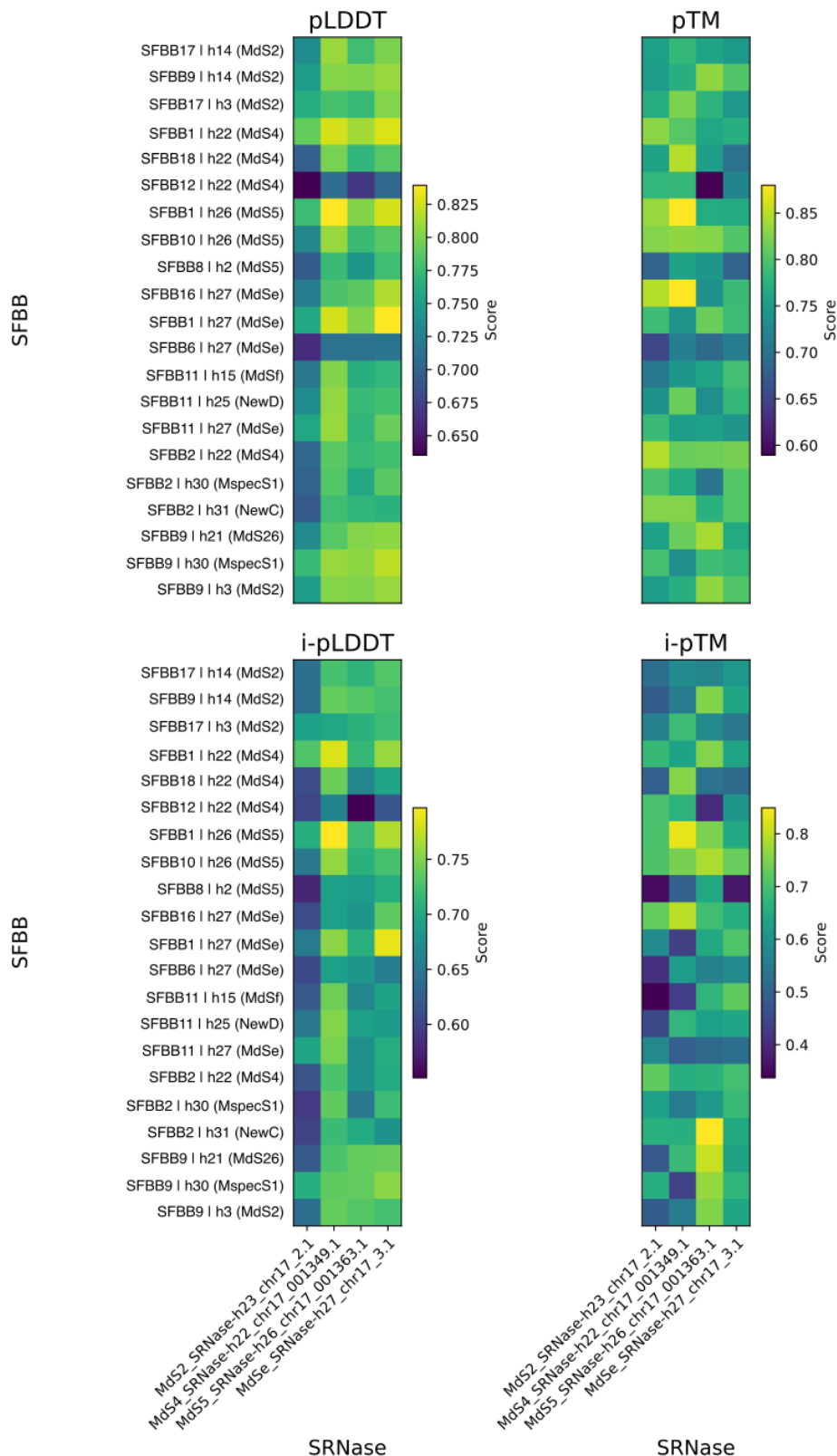
Supplementary figure 17: Structural representation of a *Malus* S-RNase highlighting positively selected sites. Ribbon diagram of a representative S-RNase protein structure (Accession MdS2_h23_chr17_2.1), showing the typical α/β fold. The spheres indicate amino acid residues inferred to be under positive selection across S-haplotypes (posterior probability > 0.95, Bayes Empirical Bayes analysis, CodeML). These selected sites are broadly distributed on the surface of the protein, suggesting their potential involvement in allele-specific recognition or interaction with pollen determinants. Structural elements are colored according to MuLAN attention scores. Red-to-white colors correspond to high while blue tones to low scores.



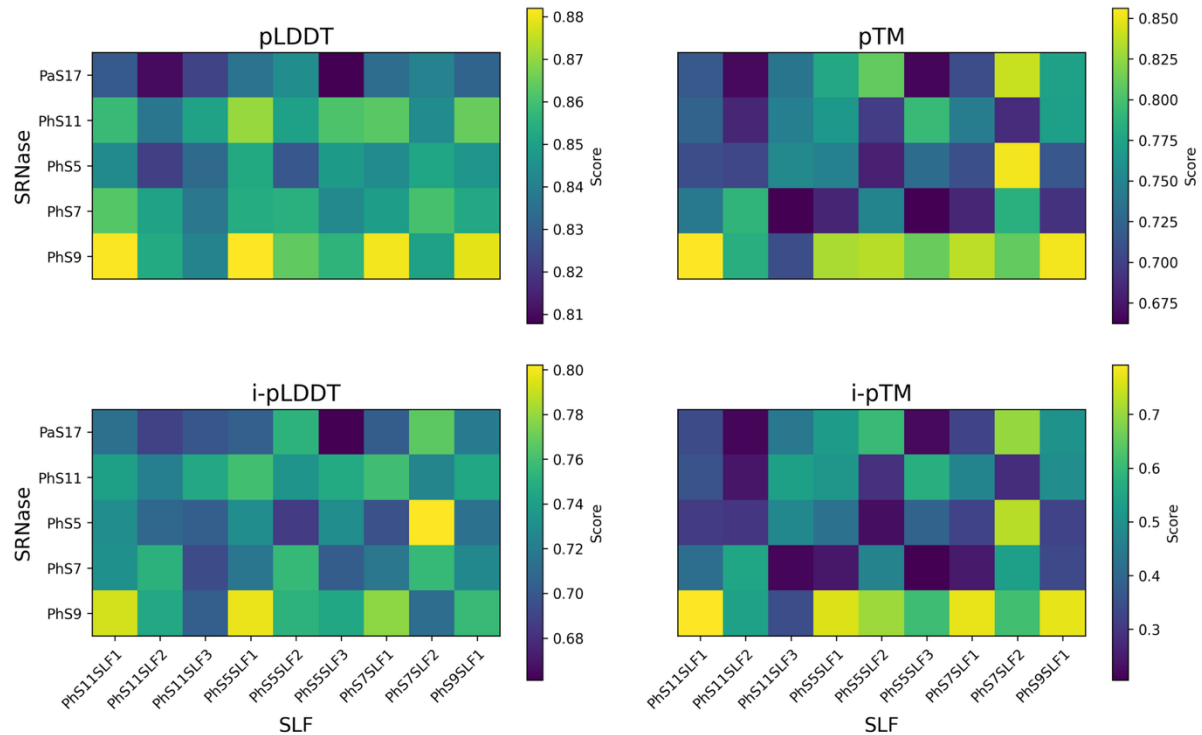
Supplementary Fig. 18: Comparison of MuLAN attention scores and positively selected sites across *Malus* S-RNase sequences. MuLAN attention scores (blue) are shown along the amino acid sequences of several S-RNases (MdS2, MdS4, MdS5). Vertical red lines mark positions inferred to be under positive selection (posterior probability > 0.95, CodeML analysis). While some positively selected sites coincide with peaks in attention scores, others appear in regions with low model-derived attention, suggesting that not all adaptively evolving residues are predicted to be functionally critical by the sequence-based model.

Supplementary Table 14. Putative targeting SFBB families determined for each S-RNases in the *Malus* dataset. Conclusions are based on the absence of SFBB families or on SFBB proteins mutations in S alleles. Sites located in the FBA domain, involved in the interaction with S-RNase are in bold characters.

S allele (S-RNase)	Putative targeting SFBB family (supported by)
MdS2 (5 haplotypes)	SFBB2 (absence from all haplotypes) SFBB3 (consistent AA variations on sites 299, 336 , 421 and 476)
MdS4 (2 haplotypes)	SFBB11 (absence from all haplotypes)
MdS5 (2 haplotypes)	SFBB11 (absence from all haplotypes)
MdS9 (3 haplotypes)	SFBB12 (consistent AA variation on site 543), SFBB13 (consistent AA variation on sites 297, 336 , 543, 549), SFBB16 (consistent AA variation on site 336)
MdSf (2 haplotypes)	SFBB5 (consistent AA variation on site 336), SFBB7 (consistent AA variation on site 476)
MdSe (2 haplotypes)	SFBB9 (absence from all haplotypes)

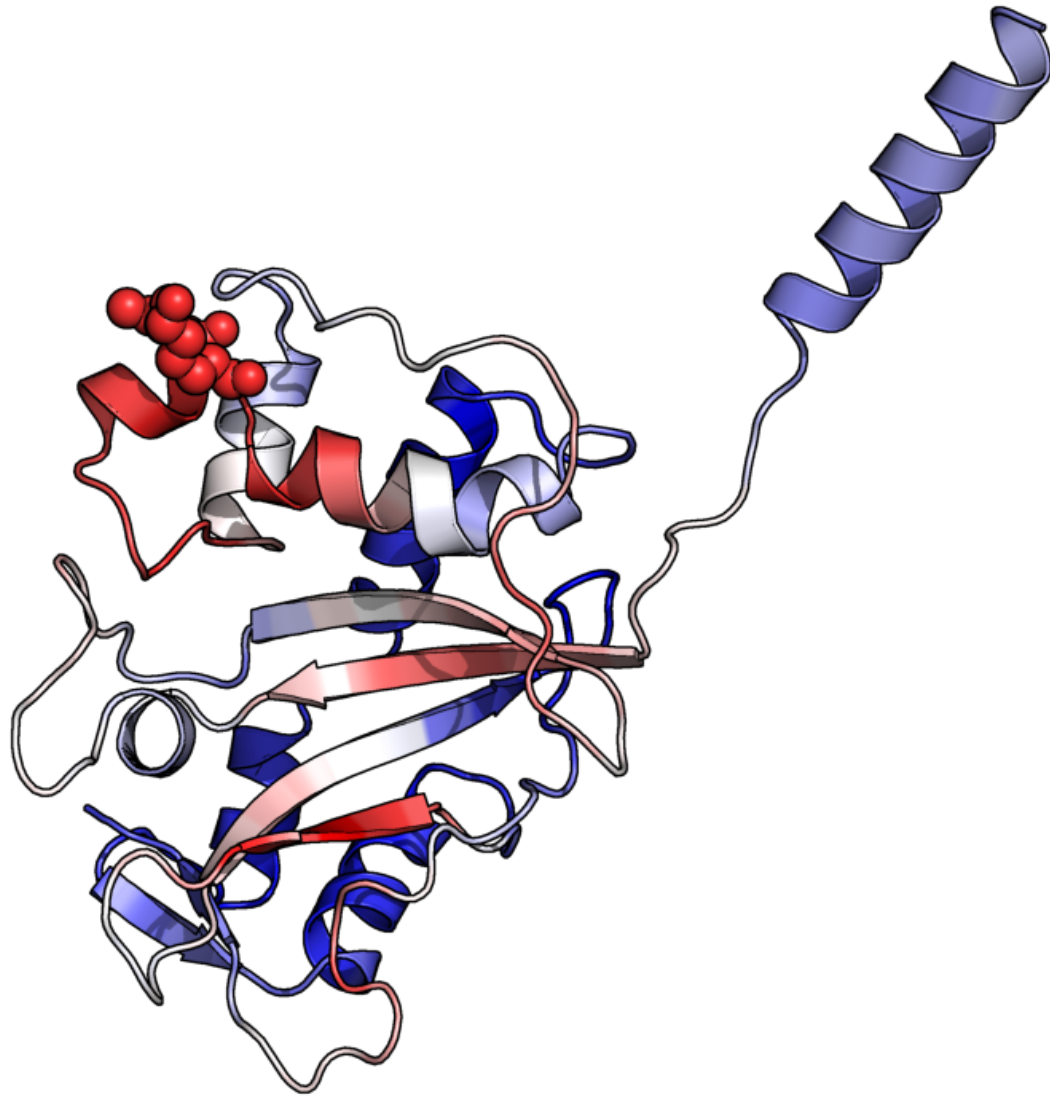


Supplementary Fig. 19: Predicted interaction scores between *Malus* SFBB proteins and S-RNases. Interaction predictions were computed using the Boltz-1 model for all pairwise combinations of 21 SFBB proteins (y-axis) and four S-RNase proteins (x-axis). Four different metrics are shown: pLDDT (predicted local distance difference test, top left), pTM (predicted template modeling score, top right), i-pLDDT (interface pLDDT, bottom left), and i-pTM (interface pTM, bottom right), each representing different confidence measures of the predicted protein-protein interaction. Lighter colors indicate higher predicted interaction scores. SFBB identifiers include family and haplotype. S-RNase identifiers include family, haplotype and gene id.

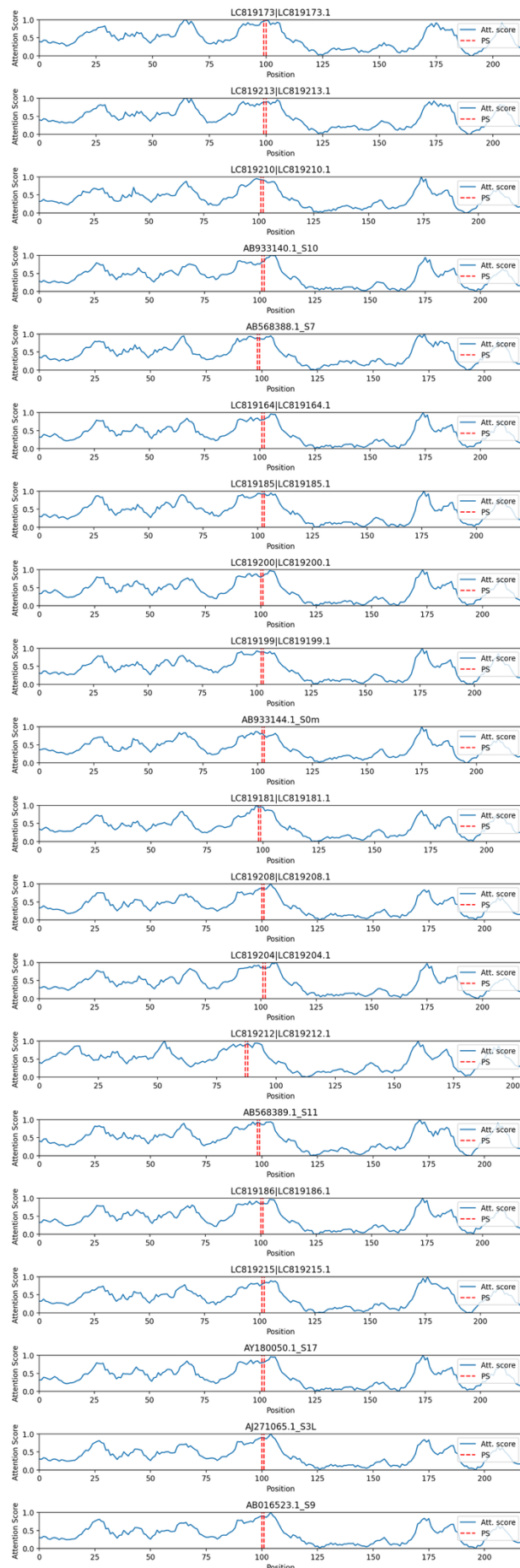


Supplementary Fig. 20: Predicted interaction scores between *Petunia* SLF proteins and S-RNases.

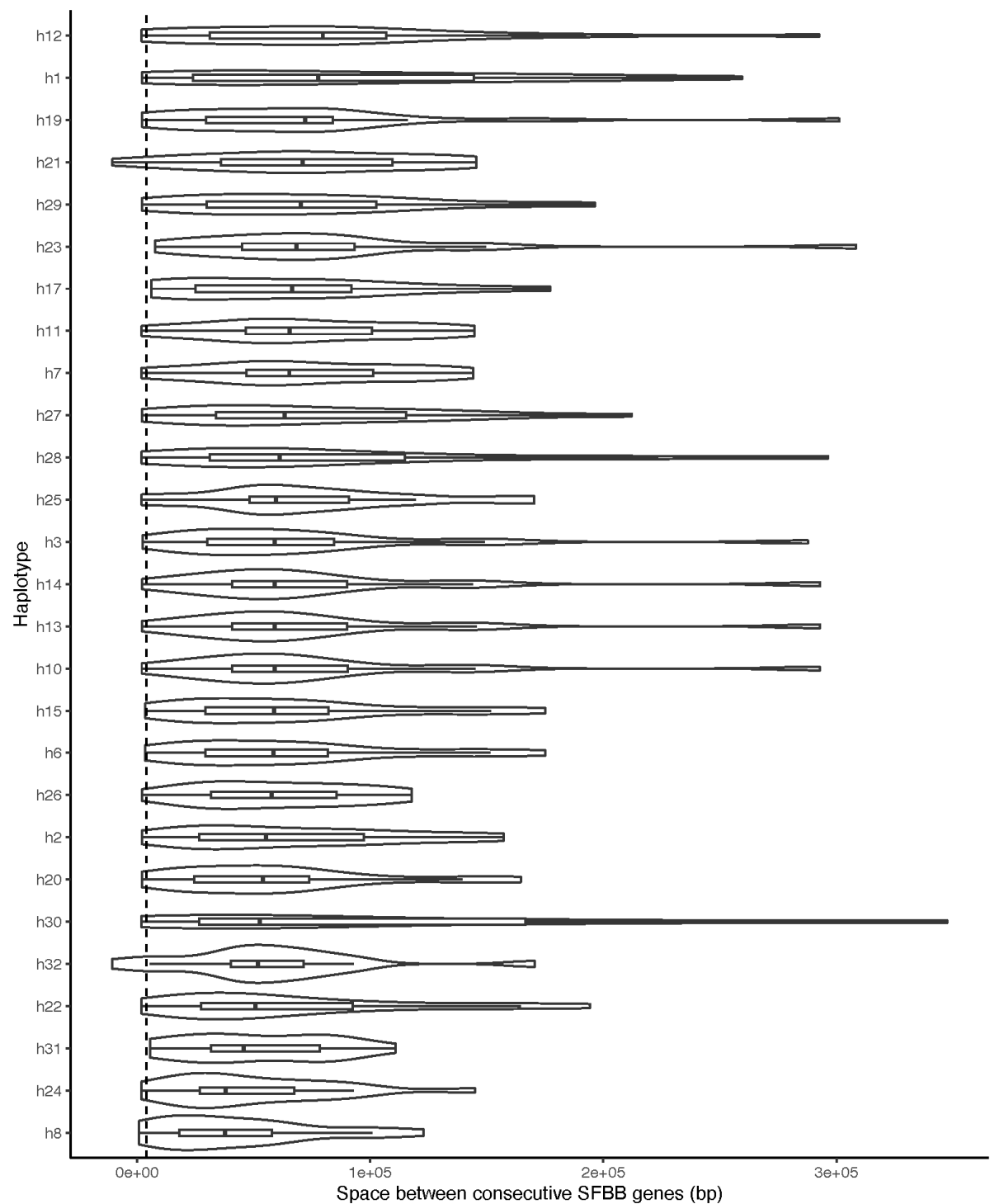
Interaction predictions were computed using the Boltz-1 model for all pairwise combinations of 9 SLF proteins (x-axis) and five S-RNase proteins (y-axis). Four different metrics are shown: pLDDT (predicted local distance difference test, top left), pTM (predicted template modeling score, top right), i-pLDDT (interface pLDDT, bottom left), and i-pTM (interface pTM, bottom right), each representing different confidence measures of the predicted protein-protein interaction. Lighter colors indicate higher predicted interaction scores. SLF and S-RNase identifiers include haplotype and gene id.



Supplementary Fig. 21: Structural representation of a *Petunia* S-RNase highlighting positively selected sites. Ribbon diagram of a representative S-RNase protein structure (Accession LC819199), showing the typical α/β fold. The spheres indicate amino acid residues inferred to be under positive selection across S-haplotypes (posterior probability > 0.95, Bayes Empirical Bayes analysis, CodeML). These selected sites are broadly distributed on the surface of the protein, suggesting their potential involvement in allele-specific recognition or interaction with pollen determinants. Structural elements are colored according to MuLAN attention score. Red-to-white colors correspond to high while blue tones to low scores.



Supplementary Fig. 22: Comparison of MuLAN attention scores and positively selected sites across *Petunia* S-RNase sequences. MuLAN attention scores (blue) are shown along the amino acid sequences of several S-RNases. Vertical red lines mark positions inferred to be under positive selection (posterior probability > 0.95, CodeML analysis). Positively selected sites coincide with peaks in attention scores.



Supplementary Fig. 23: Distribution of intergenic distances between adjacent SFBB genes across S-locus haplotypes. Violin plots show the full distribution of distances (in base pairs), while boxplots indicate the median (horizontal line) and interquartile range. Haplotypes are ordered according to their median spacing. The dashed vertical line marks the 4 kb threshold, used as a reference for genomic proximity.

Supplementary text 1: SFBB families' determination and validation

To determine SFBB families of the 500 complete SFBB sequences, the amino acid sequences were aligned with sequences from the database (Supplementary Table 5), and phylogeny was constructed with 1000 bootstraps. *Eriobotya japonica* (Loquat) Fbox sequence (accession AGM21636.1) was used as an outgroup. Then, Ward hierarchical clustering (Ward, 1963) based on cophenetic distances between tree nodes (R package *stats*,) as well as orthogroups determination using Orthofinder (Emms & Kelly, 2019) were applied. Ward clustering segregated SFBB sequences in 18 groups while Orthofinder defined 16 orthogroups (Fig. 1). The main difference is that a big Orthofinder orthogroup contains 3 Ward groups.

We took advantage of publicly available SFBB gene accessions, for which the family they belong is known, to determine the family of SFBB genes identified from the assemblies we used. In most cases, public accessions, tagged with identical SFBB families, clustered together, allowing us, to attribute the SFBB family to all the cluster. But in some cases, we found that public accessions tagged with identical SFBB families, belong to several groups in our phylogeny. For instance, public accessions tagged with SFBB3 family, are found in two distant groups (Fig. 1). Public accessions tagged with SFBB1 family clustered together in our phylogenetic tree (Fig. 1). However, hierarchical clustering and Orthofinder allowed to split the group they form into two groups. The same conclusions can be made for public accessions tagged with SFBB8 family (Fig. 1).

To validate that these families should indeed be divided into two distinct groups, we compared the pairwise similarity distribution among sequences, considering both inter- and intra-group sequence similarity. We observed that sequences within each proposed subgroup displayed significantly higher similarity to one another compared to sequences from the other subgroup, supporting the hypothesis of functional divergence between these groups. Furthermore, when examining gene structure and conserved motifs, we identified differences that were consistent with the proposed division. This evidence collectively supports the reclassification of SFBB1, SFBB3, and SFBB8 families into two distinct families each, respectively SFBB1 and SFBB15, SFBB3 and SFBB4, SFBB8 and SFBB13 (Fig. 2, A,B and C).

On the contrary, families SFBB4 and SFBB12 can be brought together. Computing intra-family identity percentages considering that these two families form only one, showed that all intra-superfamily identity percentages are higher than 90% (Fig. 2,D). For families SFBB9 and SFBB10, Ward hierarchical clustering and Orthofinder defined two groups (Fig. 1), but whose limits are not the same as original SFBB families. The best thing to do is probably to redefine SFBB9 and SFBB10 boundaries, as the Ward groups and orthogroups are supported by branch bootstrap of 100%. Finally, some new families were created because no SFBBs with family associated was present in some groups. Eighteen updated SFBB families were finally defined (Table 1, Figure 1). SFBB genes annotated on *Malus* haplotypes, coordinates and families are synthetized in Supplementary Table 6.

Table 1: Original and updated SFBB families, based on Ward hierarchical clustering, Orthofinder results and phylogenetic relationships among SFBB protein sequences.

Original SFBB families	Updated SFBB families proposition	Supported by
SFBB1	SFBB1	Ward, Orthofinder, phylogeny, identity percentages
	SFBB15	Ward, Orthofinder, phylogeny, identity percentages
SFBB2	SFBB2	Ward, Orthofinder, phylogeny, identity percentages
SFBB3	SFBB3	Ward, Orthofinder, phylogeny, identity percentages
	SFBB4	Ward, Orthofinder, phylogeny, identity percentages
	SFBB17	Ward, Orthofinder, phylogeny, identity percentages
SFBB8	SFBB8	Ward, phylogeny, identity percentages
	SFBB13	Ward, phylogeny, identity percentages
SFBB4	SFBB12	Ward, phylogeny, identity percentages
SFBB12		
SFBB9	~SFBB9	Ward, Orthofinder, phylogeny, identity percentages
SFBB10	~SFBB10	Ward, Orthofinder, phylogeny, identity percentages
SFBB11	SFBB11	Ward, Orthofinder, phylogeny, identity percentages
SFBB5	SFBB5	Ward, phylogeny, identity percentages
SFBB6	SFBB6	Ward, phylogeny, identity percentages
SFBB7	SFBB7	Ward, phylogeny, identity percentages
-	SFBB14	Ward, orthofinder, phylogeny, identity percentages
-	SFBB16	Ward, Orthofinder, phylogeny, identity percentages
-	SFBB18	Ward, Orthofinder, phylogeny, identity percentages

Considering these updated SFBB families, their presence was checked in each haplotype (figure 1, B). In each haplotype, zero to two genes of each family of *SFBB* were identified. Since some S alleles are

represented in this dataset by multiple haplotypes, it was possible to compare them, and observe that some SFBB families are absent from all haplotypes with the same S allele. For instance, SFBB11 is absent from all MdS4 (h20 and h22) and MdS5 (h2 and h26) haplotypes. SFBB2 family is absent from all MdS2 haplotypes except h23 (h3, h10, h13 and h14). Absent SFBB families' products could target these S-RNases. Concerning haplotypes from S allele MdSf, all identified SFBB families are present, suggesting that the SFBB family whose products target this S-RNase was not identified from our dataset, or that one of the SFBB families identified contain unfunctional products (although the protein sequences are complete).

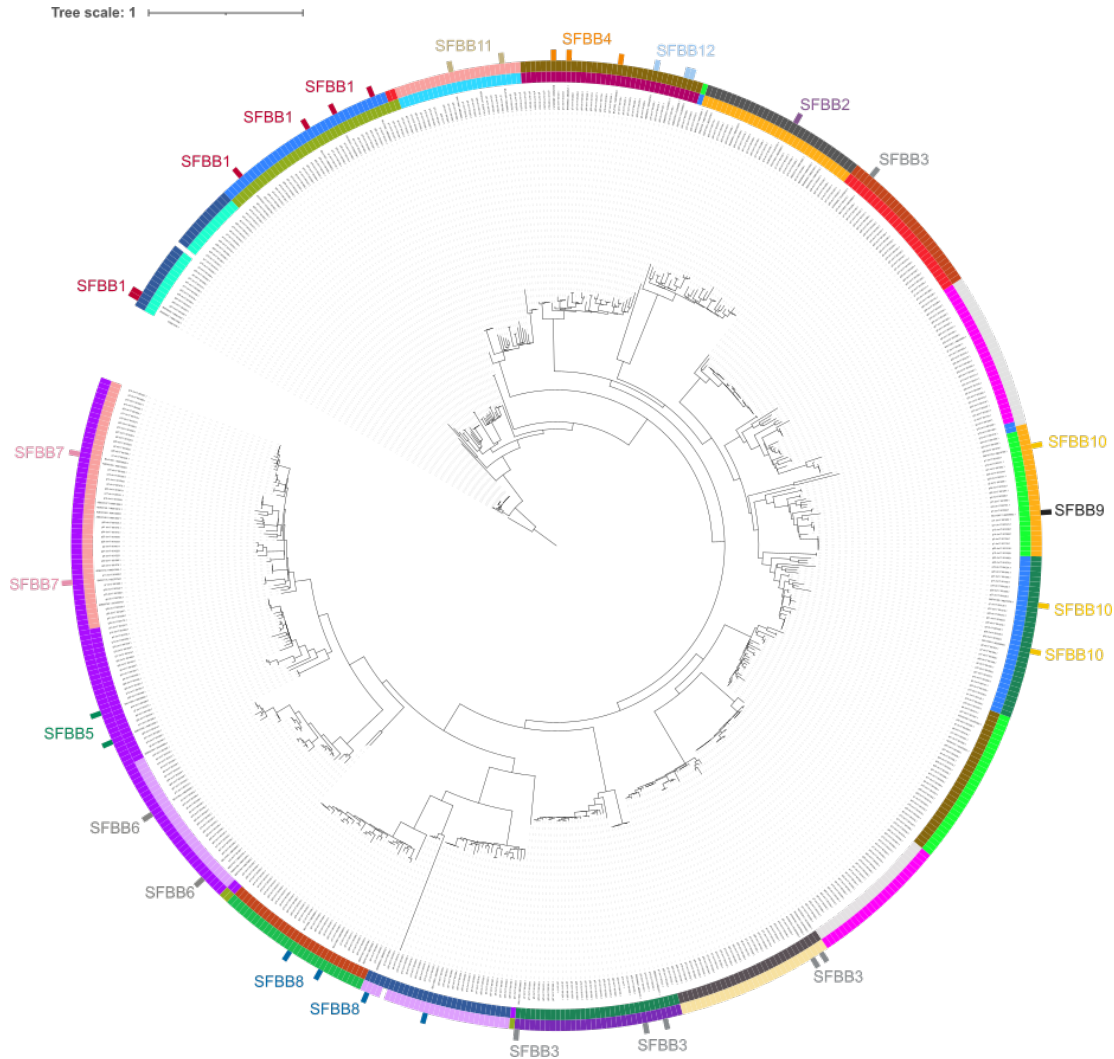


Fig. 1: SFBB clustering results from Ward hierarchical clustering (k=18) (Ward, 1963) and Orthofinder (Emms & Kelly, 2019). In this circular phylogenetic tree, the innermost layer represents the families determined using Ward's method. The intermediate layer corresponds to families identified through Orthofinder. The outermost layer, when available, indicates the known families associated with the public accessions. This visualization highlights the consistency between the different approaches and known classifications.

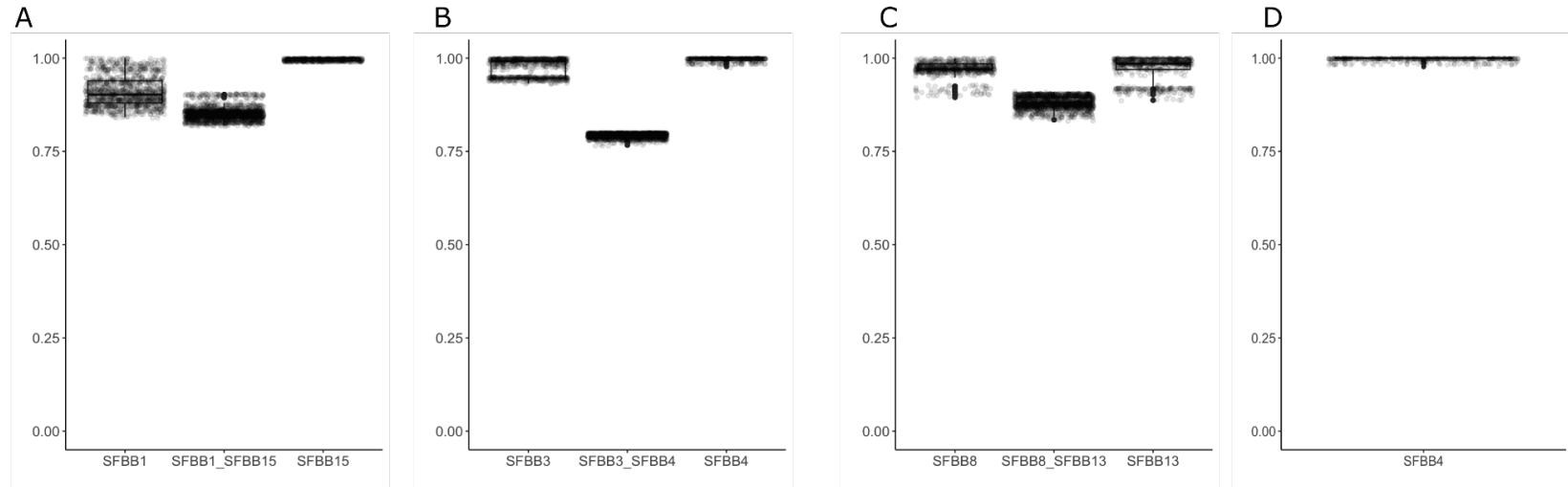


Fig. 2: Pairwise similarity distribution of SFBB gene groups before and after proposed reclassification. Boxplots display the pairwise similarity values among sequences within and between different SFBB groups. (A) SFBB1 group and its proposed subdivision into SFBB1 and SFBB15. (B) SFBB3 group and its proposed subdivision into SFBB3 and SFBB4. (C) SFBB8 group and its proposed subdivision into SFBB8 and SFBB13. (D) SFBB4 group, resulting from the merging of two previously separate families. For (D), the high pairwise similarity observed among sequences supports the merging of these two families into a single SFBB4 family. Each boxplot represents the distribution of pairwise similarity values within the corresponding group or subgroup, with the proposed subdivisions and merging reflecting patterns of sequence similarity consistent with functional or evolutionary divergence.

References:

- Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1). <https://doi.org/10.1186/s13059-019-1832-y>
- Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301), 236–244. <https://doi.org/10.1080/01621459.1963.10500845>

Supplementary text 2: Details on dataset used, DNA extraction, PacBio sequencing, assembly of the new genomes, TEs and genes annotation

Our dataset included 15 already available haplotype assemblies and 12 new sequenced haplotypes from which only chromosome 17 were kept for this study as this chromosome hosts the SI loci in *Malus* (Aguilar et al., 2015). All high-quality assembled genomes of apples available in 2022 were retrieved from public repositories. One haplotype from the *M. domestica* double-haploid GDDH13 reference genome (Daccord et al., 2017), two phased haplotypes of a Gala cultivar (Sun et al., 2020), two available phased haplotypes of a Fuji cultivar (Li et al., 2024) were used. For cultivars Red Delicious (Han et al., 2017), Honeycrisp (Khan et al., 2022), Hanfu (Qin et al., 2023), Brown Snout, Costard, Bardsey Island and Newtons (Könyves et al., 2022), only one phased haplotype was available, and used in this study. In the same way, for individuals *M. sieversii* native to Central Asia (Sun et al., 2020), *M. sylvestris*, native to Germany (Sun et al., 2020), and native to England (Ruhsam et al., 2022), only one phased haplotype was available and collected for this study. In total, 12 *M. domestica*, two *M. sylvestris* and one *M. sieversii* haplotypes were therefore retrieved from public repositories (Supplementary Table 1).

To complete this dataset, *M. sylvestris* individuals from Austria, Denmark, France and Romania, one *M. orientalis* individual native to Armenia, and one Spanish Cider *M. domestica* cultivar, were grown in IDEEV facilities, in order to sequence their genomes sequenced using PacBio Sequel II. Assembly pipeline used allowed us to have the two parental haplotypes for these genotypes, so twelve supplementary haplotypes. Ultimately, 16 *M. domestica*, 11 *M. sylvestris*, one *M. sieversii* and two *M. orientalis* haplotypes, in total, 27 haplotypes, were available for this study.

Most individuals used to make this dataset are diploids and, in most cases, two assembled haplotypes were obtained by individual (double haploid Golden Delicious, tri-haploid Hanfu, Gala and Fuji cultivars, the cider individual, the *M. orientalis* individual, the *M. sieversii* individual and the four *M. sylvestris* individuals grown at IDEEV). For eight diploids genotypes (Double haploid golden delicious, Hanfu, Red delicious, Honeycrisp, Brown snout, Costard, Bardsey Island and Newtons cultivars, *M. sieversii* and *M. sylvestris* from previous studies), we only have one assembled haplotype. In total, 14 haplotypes of *M. domestica*, 10 haplotypes of *M. sylvestris*, one haplotype of *M. sieversii* and two haplotypes of *M. orientalis* were used for further analyses (Supplementary table 1). For convenience, all haplotypes were identified with id from h1 to h32. Finally, two haplotypes were collected from haploids individuals: the double haploid Golden Delicious individual used to construct the GDDH13 *M. domestica* reference genome (Daccord et al., 2017), and the *M. domestica* cultivar Hanfu individual (Qin et al., 2023), which is a tri-haploid.

DNA extraction, libraries constructions and sequencing

For h21 to h32, Genomic DNA was extracted from whole plant material using the QIAGEN Genomic-tips 100/G kit (Cat No./ID: 10243) following the tissue protocol. Briefly, ~ 1g of plant tissue were frozen and ground in liquid nitrogen with a mortar and pestle. After 2h of lysis at 50°C and a single centrifugation step, DNA was bound to the column, washed and subsequently eluted. The eluate was desalted and concentrated by alcohol precipitation and DNA was finally resuspended in EB buffer.

DNA concentration and purity were assessed using a NanoDrop One spectrophotometer (Thermo Scientific) and a Qubit 3 Fluorometer using the Qubit dsDNA BR assay (Invitrogen), respectively. DNA integrity and fragment size distribution were evaluated using a Femto Pulse system (Agilent, Santa

Clara, CA, USA). Then, DNA was submitted to Leiden Genome Technology Center (Netherlands) for PacBio® HiFi library construction and sequencing.

Genome assemblies and scaffold anchoring

Raw PacBio reads were assembled with hifiasm (Cheng et al., 2021), with default parameters, that allowed to obtain two phased haplotypes. TIDK (Brown et al., 2025) was used to detect telomere repeats on resulting contigs. Contigs were anchored to chromosomes with ALLMAPS JCVI (Tang et al., 2015), which allows combining several genetic maps to order and orientate contigs, and using RagTag scaffold (Alonge et al., 2022), which allows to use whole-genome alignments to a reference assembly to scaffold query sequences. The comparison of agp files resulting from ALLMAPS and from RagTag was performed using homemade perl scripts available on https://github.com/CornilleEclecticLab/ART_pipeline/tree/main, and AGPTOOLS (Ricemeyer et al., 2025). Briefly, perl scripts allow to select contigs associated with a specific chromosome in several agp files, and to make a final agp file synthesizing all results. Then, this final agp file was used as input with AGPTOOLS.

Concerning ALLMAPS anchoring, a integrated genetic maps from Di Pierro et al., (2016), and from Howard et al., (2021) with SNP markers from IRSC pear Infinium® II 1 K array (Montanari et al., 2019) and the apple 20K SNP array (Bianco et al., 2014) were used. Concerning RagTag scaffold, GDDH13 genome (Daccord et al., 2017) was used as reference to scaffold assembly contigs. Only chromosome 17 was then kept for further analyses.

Gene and transposable element annotation

In order to eliminate the bias of genome annotation caused by different pipelines, gene annotation was performed using a unique pipeline. First, TEs libraries were constructed using REPET TEde novo (Flutre et al., 2011) using 8 assemblies from several *Malus* species (GDDH13, cider, MSYL_AUT, MSYL_DNK, MSYL_FRA, MSYL_ROU, Msiev, MORI_ARM), in order to capture maximal diversity of TE content. Then, these TEs libraries were merged and redundancy was removed using PASTEC (Hoede et al., 2014). Finally, TEs were annotated in all assemblies using the constructed TEs panlibrary, and REPET TEannot (Quesneville et al., 2005). TE sequences were softmasked from assemblies using BEDOPS (Neph et al., 2012), to convert gff files to bed files and BEDTOOLS maskfasta (Quinlan & Hall, 2010) to softmask TE sequences.

The quality of repetitive genomic regions was assessed using the LTR Assembly Index (LAI) (Quesneville et al., 2005): (1) LTRharvest in GenomeTools v1.6.2 (Ellinghaus et al., 2008) was used to de novo predict the candidate LTR-RTs (full-length LTRs retrotransposon) in the haplotypes assembly sequences for chromosome 17, and (2) LTR_retriever v2.9.5 (Ou & Jiang, 2018) was then used to combine and refactor all the candidates to get the final full-length LTR-RTs. LAI was calculated based on the formula: $LAI = (\text{intact LTR-RT length} / \text{total LTR-RT length}) \times 100$.

LAI values in the S-loci vary among haplotypes but are mostly above 10 (Supplementary fig. 4), except in Gala haplotypes (h2 and h3), where assembly quality may not be sufficient and must lead to be careful with conclusions made on these haplotypes.

To predict genes, Helixer (Stiehler et al., 2021) was run on softmasked haplotype fasta files using land_plant as –lineage model. Gffread (Pertea & Pertea, 2020) allowed to get cds and protein sequences

for predicted genes. Finally, gene functional annotation was performed from protein sequences, using eggNOG-mapper (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2016, 2017).

To complete SFBB and S-RNase gene annotation, manual annotation was performed using Apollo (Dunn et al., 2019) to identify supplementary SFBB and S-RNase genes, and to check SFBB and SRNase genes structures.

References

- Aguiar, B., Vieira, J., Cunha, A. E., Fonseca, N. A., Iezzoni, A., Van Nocker, S., & Vieira, C. P. (2015). Convergent Evolution at the Gametophytic Self-Incompatibility System in *Malus* and *Prunus*. *PLOS ONE*, *10*(5), e0126138. <https://doi.org/10.1371/journal.pone.0126138>
- Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, *23*(1), 258. <https://doi.org/10.1186/s13059-022-02823-7>
- Bianco, L., Cestaro, A., Sargent, D. J., Banchi, E., Derdak, S., Di Guardo, M., Salvi, S., Jansen, J., Viola, R., Gut, I., Laurens, F., Chagné, D., Velasco, R., Van De Weg, E., & Troggio, M. (2014). Development and Validation of a 20K Single Nucleotide Polymorphism (SNP) Whole Genome Genotyping Array for Apple (*Malus × domestica* Borkh). *PLoS ONE*, *9*(10), e110377. <https://doi.org/10.1371/journal.pone.0110377>
- Brown, M. R., Manuel Gonzalez De La Rosa, P., & Blaxter, M. (2025). tidk: A toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics*, *41*(2), btaf049. <https://doi.org/10.1093/bioinformatics/btaf049>
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. <https://doi.org/10.1093/molbev/msab293>

- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2), 170–175.
<https://doi.org/10.1038/s41592-020-01056-5>
- Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., Di Pierro, E. A., Gouzy, J., Rees, D. J. G., Guérif, P., Muranty, H., Durel, C.-E., Laurens, F., Lespinasse, Y., Gaillard, S., ... Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*, 49(7), 1099–1106. <https://doi.org/10.1038/ng.3886>
- 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., Tartarini, S., Letschka, T., Lozano Luis, L., Garkava-Gustavsson, L., Micheletti, D., Bink, M. C., Voorrips, R. E., Aziz, E., Velasco, R., ... Van De Weg, W. E. (2016). A high-density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. *Horticulture Research*, 3(1), 16057. <https://doi.org/10.1038/hortres.2016.57>
- Dunn, N. A., Unni, D. R., Diesh, C., Munoz-Torres, M., Harris, N. L., Yao, E., Rasche, H., Holmes, I. H., Elsik, C. G., & Lewis, S. E. (2019). Apollo: Democratizing genome annotation. *PLOS Computational Biology*, 15(2), e1006790. <https://doi.org/10.1371/journal.pcbi.1006790>
- Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9(1), 18.
<https://doi.org/10.1186/1471-2105-9-18>
- Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE*, 6(1), e16526.
<https://doi.org/10.1371/journal.pone.0016526>
- Han, M., Sun, Q., Zhou, J., Qiu, H., Guo, J., Lu, L., Mu, W., & Sun, J. (2017). Insertion of a solo LTR retrotransposon associates with spur mutations in ‘Red Delicious’ apple (*Malus × domestica*). *Plant Cell Reports*, 36(9), 1375–1385. <https://doi.org/10.1007/s00299-017-2160-x>

- Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., & Quesneville, H. (2014). PASTEC: An Automatic Transposable Element Classification Tool. *PLoS ONE*, 9(5), e91929. <https://doi.org/10.1371/journal.pone.0091929>
- Howard, N. P., Troggio, M., Durel, C.-E., Muranty, H., Denancé, C., Bianco, L., Tillman, J., & 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(1), 246. <https://doi.org/10.1186/s12864-021-07565-7>
- Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, 34(8), 2115–2122. <https://doi.org/10.1093/molbev/msx148>
- Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C., & Bork, P. (2016). eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44(D1), D286–D293. <https://doi.org/10.1093/nar/gkv1248>
- Khan, A., Carey, S. B., Serrano, A., Zhang, H., Hargarten, H., Hale, H., Harkess, A., & Honaas, L. (2022). A phased, chromosome-scale genome of ‘Honeycrisp’ apple (*Malus domestica*). *Gigabyte*, 2022, 1–15. <https://doi.org/10.46471/gigabyte.69>
- Könyves, K., Mian, S., Johns, J., Royal Botanic Garden Edinburgh Genome Acquisition Lab, Royal Botanic Gardens Kew Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Ruhsam, M., Leitch, I. J., & Darwin Tree of Life Consortium. (2022). The genome sequence of the apple, *Malus domestica* (Suckow) Borkh., 1803. *Wellcome Open Research*, 7, 297. <https://doi.org/10.12688/wellcomeopenres.18646.1>
- Li, W., Chu, C., Li, H., Zhang, H., Sun, H., Wang, S., Wang, Z., Li, Y., Foster, T. M., López-Girona, E., Yu, J., Li, Y., Ma, Y., Zhang, K., Han, Y., Zhou, B., Fan, X., Xiong, Y., Deng, C. H., ...

- Han, Z. (2024). Near-gapless and haplotype-resolved apple genomes provide insights into the genetic basis of rootstock-induced dwarfing. *Nature Genetics*, 56(3), 505–516.
<https://doi.org/10.1038/s41588-024-01657-2>
- Montanari, S., Bianco, L., Allen, B. J., Martínez-García, P. J., Bassil, N. V., Postman, J., Knäbel, M., Kitson, B., Deng, C. H., Chagné, D., Crepeau, M. W., Langley, C. H., Evans, K., Dhingra, A., Troggo, M., & Neale, D. B. (2019). Development of a highly efficient Axiom™ 70 K SNP array for *Pyrus* and evaluation for high-density mapping and germplasm characterization. *BMC Genomics*, 20(1), 331. <https://doi.org/10.1186/s12864-019-5712-3>
- Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoiyannopoulos, J. A. (2012). BEDOPS: High-performance genomic feature operations. *Bioinformatics*, 28(14), 1919–1920. <https://doi.org/10.1093/bioinformatics/bts277>
- Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gky730>
- Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology*, 176(2), 1410–1422. <https://doi.org/10.1104/pp.17.01310>
- Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, 9, 304. <https://doi.org/10.12688/f1000research.23297.2>
- Qin, S., Xu, G., He, J., Li, L., Ma, H., & Lyu, D. (2023). A chromosome-scale genome assembly of *Malus domestica*, a multi-stress resistant apple variety. *Genomics*, 115(3), 110627. <https://doi.org/10.1016/j.ygeno.2023.110627>
- Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., & Anxolabehere, D. (2005). Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Computational Biology*, 1(2), e22. <https://doi.org/10.1371/journal.pcbi.0010022>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. <https://doi.org/10.1093/bioinformatics/btq033>

- Ricemeyer, E. S., Carroll, R. A., & Warren, W. C. (2025). Agptools: A utility suite for editing genome assemblies. *Bioinformatics*, 41(7), btaf388.
<https://doi.org/10.1093/bioinformatics/btaf388>
- Ruhsam, M., Bell, D., Hart, M., Hollingsworth, P., Royal Botanic Garden Edinburgh Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, & Darwin Tree of Life Consortium. (2022). The genome sequence of the European crab apple, *Malus sylvestris* (L.) Mill., 1768. *Wellcome Open Research*, 7, 296. <https://doi.org/10.12688/wellcomeopenres.18645.1>
- Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M., & Denton, A. K. (2021). Helixer: Cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics*, 36(22–23), 5291–5298. <https://doi.org/10.1093/bioinformatics/btaa1044>
- Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K., Cheng, L., Zhong, G.-Y., & Fei, Z. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature Genetics*, 52(12), 1423–1432. <https://doi.org/10.1038/s41588-020-00723-9>
- Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., & Lu, J. (2015). ALLMAPS: Robust scaffold ordering based on multiple maps. *Genome Biology*, 16(1), 3. <https://doi.org/10.1186/s13059-014-0573-1>