SCIENTIFIC REPORTS

OPEN

Inferences on specificity recognition at the *Malus*×*domestica* gametophytic self-incompatibility system

Received: 19 September 2017
Accepted: 9 January 2018
Published online: 29 January 2018

Maria I. Pratas[1,2], Bruno Aguiar[1,2], Jorge Vieira[1,2], Vanessa Nunes[1,2], Vanessa Teixeira[1,2], Nuno A. Fonseca[3], Amy Iezzoni[4], Steve van Nocker[4] & Cristina P. Vieira[1,2]

In *Malus* × *domestica* (Rosaceae) the product of each *SFBB* gene (the pollen component of the gametophytic self-incompatibility (GSI) system) of a *S*-haplotype (the combination of pistil and pollen genes that are linked) interacts with a sub-set of non-self S-RNases (the pistil component), but not with the self S-RNase. To understand how the *Malus* GSI system works, we identified 24 *SFBB* genes expressed in anthers, and determined their gene sequence in nine *M. domestica* cultivars. Expression of these *SFBBs* was not detected in the petal, sepal, filament, receptacle, style, stigma, ovary or young leaf. For all *SFBBs* (except *SFBB15)*, identical sequences were obtained only in cultivars having the same *S-RNase*. Linkage with a particular *S-RNase* was further established using the progeny of three crosses. Such data is needed to understand how other genes not involved in GSI are affected by the *S*-locus region. To classify SFBBs specificity, the amino acids under positive selection obtained when performing intra-haplotypic analyses were used. Using this information and the previously identified S-RNase positively selected amino acid sites, inferences are made on the S-RNase amino acid properties (hydrophobicity, aromatic, aliphatic, polarity, and size), at these positions, that are critical features for GSI specificity determination.

Gametophytic self-incompatibility (GSI), the most common reproductive system in flowering plants (see Fig. 1 in Igic *et al.*[1]), is a pre-zygotic genetic mechanism that prevents self-fertilization and promotes out-crossing, by enabling the pistil to reject pollen from genetically related individuals[2]. In this system, to preserve functional incompatibility, there are two distinct components, one that determines the pistil specificity and another that determines the pollen specificity, called *S*- genes. The locus that contains the genes determining GSI specificity is called the *S*-locus.

The pistil specificity component in Rosaceae, Rubiaceae, Solanaceae and Plantaginaceae species, is an extracellular ribonuclease, called S-RNase[3–5]. Since RNase activity is needed for inhibition of self-pollen tube growth[6], it has been assumed that degradation of pollen tube RNAs in the self-pollen tube is part of the biochemical mechanism of self-incompatibility (SI). According to the phylogeny of this gene and the conserved structure (conserved and hypervariable regions, intron number and position) *RNase* based GSI has evolved only once, before the separation of the Asterideae and Rosideae, about 120 million years ago[5,7–9]. Nevertheless, in Rosaceae, Pyrinae (*Malus*, *Pyrus* and *Sorbus*) and *Prunus S-RNase* based GSI evolved from paralogous genes, according to phylogenetic analyses of the *S-RNase* and *S*-pollen lineage genes. *Malus* and *Prunus* GSI genes belong to distinct gene lineages, and only *Prunus* GSI -lineage genes are present in *Fragaria*, that is an out-group to both species[10].

The pollen specificity component encodes a F-box protein(s), and varies from one gene in *Prunus* (called *SFB*, S-haplotype specific F-box gene)[11–18], to multiple genes in *Malus*, *Pyrus*, *Sorbus* (called *SFBB*s, *S*-locus F-box brothers), *Petunia*, and *Nicotiana* (Solanaceae; called *SLF*s, S-locus F-box)[19–28]. *Prunus SFB* and *SFBBs*/*SLFs* genes

[1]Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto, Rua Alfredo Allen 208, 4200-135 Porto, Portugal. [2]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Rua Alfredo Allen 208, 4200-135 Porto, Portugal. [3]European Bioinformatics Institute (EMBL-EBI,) Welcome Trust Genome Campus, CB10 1SD, Cambridge, United Kingdom. [4]Michigan State University, East Lansing, MI, 48824-1325, USA. Maria I. Pratas, Bruno Aguiar and Jorge Vieira contributed equally to this work. Correspondence and requests for materials should be addressed to C.P.V. (email: cgvieira@ibmc.up.pt)

are not orthologous[10,25,28–31]. Therefore, it is not surprising that in *Prunus* a self-recognition mechanism is used for S-RNase inhibition[13,32,33], while in the species presenting multiple *S*-pollen genes, each *S*-protein recognizes and interacts with a sub-set of non-self S-RNases, to mediate their degradation[19–22,24,34–36]. In *Petunia*, transgenic experiments were performed to address the function of genes involved in pollen specificity[37]. Diploid pollen carrying two different functional *S*-haplotypes or haploid pollen that carries a duplicated *S*-locus region of a different *S*-haplotype caused breakdown of self-incompatibility[38–40]. Nevertheless, this was not always observed, implying additional *S*-pollen genes determining pollen specificity[24]. Furthermore, coimmunoprecipitation results showed non-self interactions between *S*-pollen proteins and the S-RNases in SI responses[24]. These observations led to the collaborative non-self recognition model, that takes into account the involvement of multiple *S*-pollen proteins in pollen specificity. In Pyrinae (*Malus*, *Pyrus*, and *Sorbus*), such transformation methods are not possible since these species are trees. Nevertheless, in *Malus*, yeast two-hybrid analysis indicated that SFBBs interact mostly with non-self S-RNases[41]. Other sequences assigned as SFBB-like, however, also show a similar pattern. These SFBB-like sequences may be encoded by *SFBB* genes since they are expressed in pollen only and are located in the vicinity of the *S-RNase*, some of them in between recognized *SFBB* genes. They have been assigned as SFBB-like because the authors were not able to show *S*-haplotype linkage. This, however, may be due to difficulties in designing specific primers, since *SFBB* genes can present low nucleotide divergence. Moreover, using the predicted tertiary structure of S-RNases and SFBBs and their binding energies, based on the Wilcoxon rank-sum test, when the hypervariable region of the S-RNase is considered, it has been shown that SFBBs of a *S*-haplotype interact more strongly with non-self than with self S-RNases[42]. Therefore, it seems that in *Malus* the GSI system works in a similar way to that of *Petunia*.

Because the selective pressures in recognition mechanisms with one or multiple *S*-pollen genes are different, the *S*-pollen genes show distinct evolutionary patterns. In *Prunus* the two *S*-genes must co-evolve for specificity recognition and, thus, both genes present similar levels of diversity and number of amino acids under positive selection (those that in principle are involved in specificity determination)[12,43]. In the collaborative non-self -recognition model each *S*-pollen protein recognizes a sub-set of non-self- S-RNases, and levels of diversity at these genes are, at least 2.5 times lower than those of the *S*-pistil gene[20,22,24,28,34,44]. Levels of intra-haplotype divergence are, however, similar to the *S-RNase* diversity[22,34], and amino acids under positive selection have been identified in *Sorbus* (Pyrinae, Rosaceae) when intra haplotypic analyses are performed[22].

Essential for the understanding of the collaborative non-self -recognition model is knowledge of how many *S*-pollen genes exist in a *S*-haplotype. In *Petunia*, anthers transcriptomes of two homozygous plants (*S2S2*, and *S3S3*) revealed 17 *S*-pollen genes for both *S*-haplotypes[21]. 10 of these *S*-pollen genes were previously identified[24,30,45–47], and for eight, transgenic functional assays, have been performed to show that they are involved in *S*-pollen specificity[21,24,37]. Moreover, all 17 SLF proteins of both *S*-haplotypes, using co-immunoprecipitation and mass spectrometry assays, have been shown to be assembled into similar canonical SCF complexes to the eight SLFs confirmed to be involved in GSI[36]. Furtermore, in *Petunia*, the study of 12 homozygous plants, using a combination of next-generation sequencing (from mature pollen and unopened mature anthers) and PCR techniques, revealed that the number of *SLF* genes per *S*-haplotype varies from 16 to 20[20]. These genes define 18 specificity types, and within each type, variation in terms of copy number and amino acid sequence polymorphism was found. Then, variation was used to predict the target S-RNase(s) of each type of SLF, using the rule put forth by Kubo *et al.*[20]: "Sx-RNase is a target of SLFn if the Sx-allele of SLFn is diverged or deleted". For eight *S*-haplotypes, predictions were made regarding the SLF types that recognise seven of the S-RNases. Five of these predictions are supported by experimental data.

In Pyrinae (*Malus*, *Pyrus*, and *Sorbus*), 16 *SFBB*-like genes have been characterized from the sequencing of both BAC clones containing the *S*-locus, and PCR products obtained from genomic DNA using primers for conserved regions[22,23,27,44,48,49]. All these genes, as expected for *S*-pollen genes, are expressed in pollen only, and for all, except *SFBB15*, linkage with the *S-RNase* has been confirmed. Because of the methodologies used, the number of *SFBB*s in Pyrinae could be underestimated. Such data is needed to determine the size of the *S*-locus region and its effect on other genes unrelated to self-incompatibility that are located in the same region[50]. Therefore, in this work we sought to determine the number of *SFBB*s associated with *S*-haplotypes in *M. domestica* and to use this information to provide insights into GSI in *M. domestica* by addressing how copy number variation and amino acid sequence polymorphism at the amino acids under positive selection can be used to predict *S*-pollen specificity. Furthermore, we address which S-RNase amino acid characteristics, at those sites under positive selection, are involved in *S*-pollen specificity recognition. To identify the *SFBB*s, we used an approach similar to that used in *Petunia*, that was a combination of anthers transcriptome of nine *M. domestica* cultivars ['Fuji' (*S1, S9*), 'Northern Spy' (*S1*, *S3*), 'Gala' (*S2, S5*), 'Golden Delicious' (*S2, S3*), 'Honeycrisp' (*S2, S24*), 'Idared' (*S3, S7*), 'Red Delicious' (*S9, S28*), 'McIntosh' (*S10, S25*), and 'Empire' (*S10, S28*)] covering 10 *S*-haplotypes, and a PCR approach using genomic DNA to determine the number of *SFBB* genes in *Malus*.

Results

Assessing Transcriptome Coverage. Because the main goal of this work was to identify as many as possible candidate *SFBB* genes involved in pollen specificity through transcriptome sequencing of anthers, we first assessed the coverage of the transcriptomes used (Supplementary Table S1). According to the accumulation curve obtained for the nine anthers transcriptomes, the number of expressed *Malus* CDS detected in the sample increases at a slower rate after 6000000 paired reads (Supplementary Fig. S1), suggesting that the sampling is sufficient for the discovery of new *SFBB* genes. Moreover, the annotated *SFBB* genes on the *M. domestica* genome[50] are identified in the anthers transcriptomes having *S2*- or *S3*-haplotypes (Supplementary Table S2), providing additional support that the coverage of the transcriptomes is sufficient for the identification of new *SFBB* genes. Furthermore, the 13 *S3*-, 14 *S9*-, and the six *S10*-haplotype *SFBB* genes previously reported[23,27,44], are identified

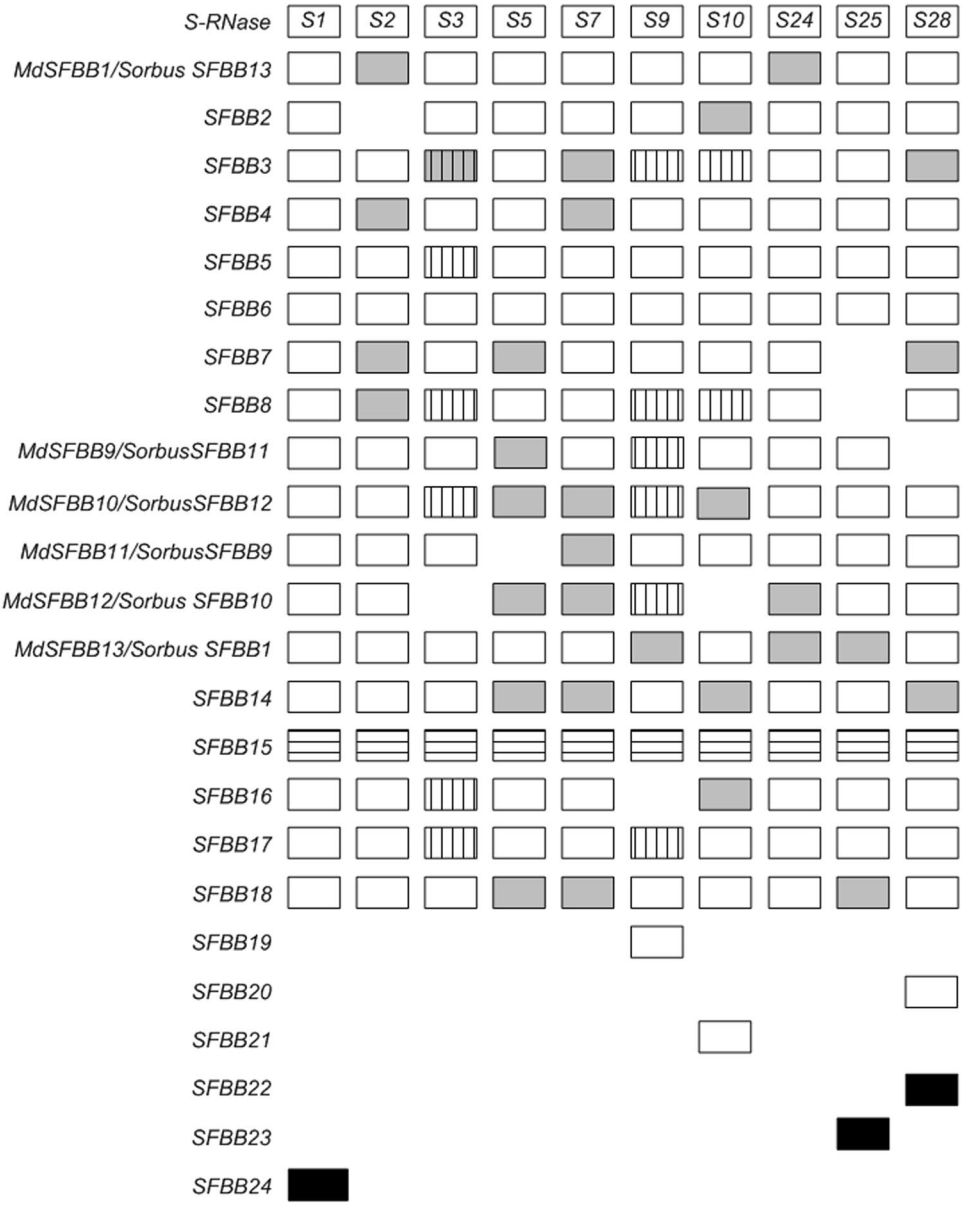

Figure 1. *SFBB* genes in the 10 *S*-haplotypes analysed. White boxes represent sequences obtained using primers SFBBgenF and SFBBgenR, grey boxes represent sequences obtained with specific primers for that particular gene, black boxes represent sequences obtained from Edena contigs. Boxes with vertical lines represent sequences described in the literature[23,27,44] not amplified with primers SFBBgenF and SFBBgenR. Boxes with horizontal lines represent sequences that are identical in cultivars not sharing a *S-RNase*. The star indicates a *SFBB* sequence that presents stop codons in the putative coding region, obtained from 'Golden Delicious' (*S2, S3*), and 'Honeycrisp' (*S2, S24*), that is also present in the *Malus* genome (NW_007545880.1- 1139053… 1137851).

in anthers transcriptomes of the cultivars having these *S*-haplotypes (Supplementary Fig. S2), suggesting that the transcriptome coverage is enough for the identification of new *SFBB* genes.

Identifying *SFBB* genes from Edena assemblies. When the 33 *SFBB* sequences from *S3-, S9-*, and *S10*-haplotype[23,27,44] were searched in the Trinity (Supplementary Table S3) and Edena (Supplementary Table S4) transcriptome assemblies of 'Golden Delicious' (*S2, S3*), 'Northern Spy' (*S1*, *S3*), 'Idared' (*S3, S7*), 'Fuji' (*S1, S9*) and 'Red Delicious' (*S9, S28*), 'McIntosh' (*S10, S25*) and 'Empire' (*S10, S28*), as described in Material and Methods, we found contigs for 28 and 26 *SFBB* sequences, respectively (Table 1). These sequences cover all *SFBB* genes described in the literature[23,27,44]. Nevertheless, these were smaller than 190 bp in the Trinity assembly, and with an average size of 697 bp for the Edena transcriptome assembly. It should be noted that the transcriptomes obtained are from heterozygous individuals (Material and Methods). Given the high level of sequence similarity between *SFBB* genes (Aguiar *et al.*[22], and references therein), it is possible, that ambiguities arise during assembly

Gene	*S3*-haplotype			*S9*-haplotype		*S10*-haplotype	
	GD	NS	Idared	RD	Fuji	Mc	Empire
MdSFBB1/SorbusSFBB13	**144** (2)	—	**254** (1)	—	—	—	—
	556 (2)	—	601 (3)	905 (2)	408 (1)	—	—
SFBB2	**156** (1)	**122** (2)	—	**442** (2)	**285** (4)	n.a.	n.a.
	746 (1)	463 (1)	—	1191 (3)	—	n.a.	n.a.
SFBB3	**159** (1)	**108** (2)	**183** (2)	**255** (2)	**186** (1)	—	—
	—	—	—	542 (1)	—	—	—
SFBB4	—	—	**114** (2)	**183** (1)	—	**259** (1)	—
	781 (2)	546 (2)	1141 (4)	693 (2)	859 (3)	537 (2)	1049 (5)
SFBB5	**102** (1)	**117** (1)	**169** (2)	—	**151** (1)	n.a.	n.a.
	—	238 (1)	—	654 (2)	1175 (3)	n.a	n.a
SFBB6	—	**151** (1)	—	**278** (2)	**157** (1)	**179** (2)	—
	—	—	—	467 (1)	352 (1)	904 (5)	540 (2)
SFBB7	—	—	—	**106** (1)	—	n.a.	n.a.
	324 (2)	1069 (3)	647 (3)	860 (2)	647 (2)	n.a.	n.a.
SFBB8	**136** (2)	**141** (5)	**108** (1)	**323** (3)	**252** (1)	**436** (4)	**181** (4)
	—	—	—	138 (1)	—	—	—
MdSFBB9/SorbusSFBB11	**151** (1)	—	**150** (1)	**163** (1)	—	**105** (1)	**105** (1)
	840 (3)	—	1031 (4)	1171 (4)	710 (3)	747 (3)	1170 (4)
MdSFBB10/SorbusSFBB12	**153** (2)	**258** (1)	**156** (1)	**157** (2)	—	**101** (1)	—
	368 (1)	—	—	1155 (3)	872 (3)	525 (1)	923 (3)
MdSFBB11/SorbusSFBB9	**173** (1)	**331** (1)	**240** (1)	**246** (2)	**234** (1)	n.a.	n.a.
	—	—	—	269 (1)	831 (3)	n.a.	n.a.
SFBB16	**190** (1)	**190** (1)	**205** (2)	—	—	n.a.	n.a.
	722 (1)	—	391 (1)	361 (2)	221 (1)	n.a.	n.a.
SFBB17	—	**115** (2)	—	**181** (2)	—	n.a.	n.a.
	—	—	—	734 (1)	—	n.a.	n.a.

Table 1. Size, in bp, of longest sequence in the Trinity (in bold) and Edena datasets derived from seven *M. domestica* cultivars that match the 33 *SFBB* sequences reported for the *S3*-, *S9*-, and *S10*-haplotypes[23,27,44]. n.a. sequences not reported for *S10*-haplotype[44]. — sequences not present in the dataset. () number of sequences in the dataset that show 100% identity with the reported sequences.

giving rise to short contigs. Nevertheless, although only 78% of the *SFBB* alleles reported in the literature are represented in the two transcriptome assemblies, we could find at least one allele for all *SFBB* genes (Table 1).

Since large size contigs were obtained with Edena assembly (50% of the sequences are larger than 290 bp), we use this to address how many contigs can represent *SFBB* genes. 825 contigs were retrieved from the tblastn of *SFBB3beta* protein (AB270796) and the combined Edena filtered assemblies (identical sequences included within longer sequences have been removed) of the nine anthers transcriptomes. 75 of these contigs present identities higher than 97% with *SLFL*-like genes (not determining GSI specificity)[10], and thus were also removed. The presence of *SLFL*-like genes in the blast results implies that no other *SFBB* genes are present in these transcriptomes. The remaining 750 sequences could represent *SFBB* genes. The number of contigs per cultivar varied from 57 ('Northern Spy') to 99 ('Red Delicious' and 'Honeycrisp'). It should be noted that more than one contig can represent the same *SFBB* allele since the preliminary blast searches revealed that most assembled transcripts are incomplete (see Material and Methods). Indeed 87% of these sequences had a size smaller than 500 bp, and the coding region of the *SFBB* genes is larger than 1Kb. Moreover, if two sequences overlapped but covered different regions, they were both retained at this point. Therefore, to help the assembly and confirm the identified sequences, we characterized *SFBB* sequences from genomic DNA of these individuals, using the primers SFBBgenF and SFBBgenR, described in Aguiar *et al.*[22], that amplify a region of about 900 bp. Although these primers do not amplify all *SFBB*s[22], with this additional information most of these sequences will be assembled into larger fragments.

***M. domestica SFBB* sequences obtained with primers SFBBgenF and SFBBgenR.** For each of the nine cultivars an amplification product of about 900 bp was obtained and cloned from genomic DNA with primers SFBBgenF and SFBBgenR[22]. Due to sequence variation within the primer binding sites, these primers are expected to support the amplification of only 65.5% of *Malus* and *Pyrus SFBB* GenBank sequences (n = 165)[22]. Of the 32 *SFBB*s described for *S3*, *S9* and *S10*- haplotypes, 14 of the *Malus* sequences described in the literature[23,27,44] (Fig. 1 - boxes with vertical lines) could not be amplified for this reason. Sequencing of the insert of more than 30 colonies exhibiting different RFLP patterns for each cultivar (see Material and Methods), revealed 188 coding sequences, plus seven putative pseudogenes (Supplementary Table S5). The presence of identical sequences in two cultivars having a common *S*-haplotype, that are not present in the other cultivars, implies that the sequence comes from the shared *S*-haplotype. It should be noted that, no (or little) diversity is observed at the alleles of the

S-genes within the same specificity[51–55]. Thus, for all *SFBB* sequences, except those of *SFBB15* (identical sequences are found in cultivars not sharing a *S*-haplotype; boxes with horizontal lines in Fig. 1; Supplementary Table S5), we could assign the sequences into a *S*-haplotype. The putative pseudogene sequences belong to the *S2*-, *S10*-, and *S28*-haplotypes (Supplementary Table S5). One of these sequences corresponds to *S2-SFBB2* gene that presents a nucleotide insertion that is absent in all other *SFBB2* sequences from the other *S*-haplotypes (the star in Fig. 1), that creates in-frame stop codons. This insertion is not a sequencing error since an identical sequence has been obtained from 'Golden Delicious' and 'Honeycrisp' cultivars, and is also present in the *Malus* genome (NW_007545880.1–1139053… 1137851). All the remaining sequences appeared to be functional *SFBB* alleles. Since most of the *S*-haplotypes are common between cultivars, the number of different coding sequences was 127. Phylogenetic analyses of the coding sequences defined 19 *SFBB* genes (white boxes in Fig. 1; Supplementary Table S5; Fig. 2 (sequences in bold)). Nevertheless, the presence of two sequences for *S10*-haplotype clustering with sequences from other *S*-haplotypes assigned as *SFBB7*, and two sequences for *S25*-haplotype clustering with sequences from other *S*-haplotypes assigned as *MdSFBB1*/*SorbusSFBB13*, implies the presence of 21 *SFBB* genes (Fig. 1; Supplementary Table S5; Fig. 2). It should be noted that 14 of the 32 *SFBB*s described for *S3*-, *S9*-, and *S10*-haplotypes (boxes with vertical lines in Fig. 1; those underlined in Supplementary Table S5)[23,27,44] were not characterized using the PCR approach, with primers SFBBgenF and SFBBgenR. The 141 different coding sequences, that include the previously reported sequences (using local Blastn, 100% identity and a minimum size for the alignment of 100 bp; see Material and Methods) cover 555 contigs from the Edena assemblies. These were used to enlarge the size of the region sequenced when showing 100% identity in an overlapping region larger than 100 bp.

Amplification of *MdSFBB1*/*SorbusSFBB13- SFBB4*, *SFBB7-SFBB14*, *SFBB16*, and *SFBB18* genes using specific primers. Since the 10 *S*-haplotypes have not been characterized for all *SFBB* genes using SFBBgenF and SFBBgenR primers (Table 2), the 195 Edena contigs that did not show a 100% match to known *SFBB* sequences, may represent uncharacterized alleles and/or new genes. Since polymorphism levels at *SFBB* genes are below 10%[22], we have used the longest sequences of each gene to identify (using local blast and an overlap of at least 50 bp) putative allelic sequences for each *SFBB* gene. Thus, we inferred that 145 Edena contigs can represent allele sequences of the known genes. Therefore, we used specific primers for *MdSFBB1*/*SorbusSFBB13-SFBB4*, *SFBB7-SFBB14*, *SFBB16*, and *SFBB18* genes (Supplementary Table S6), to amplify the uncharacterized alleles from genomic DNA of the cultivars having that *S*-haplotype. All expected amplification products were cloned and sequenced, as described in Material and Methods. A total of 31 sequences were obtained that included alleles missing for *S2-, S24-MdSFBB1*/*Sorbus SFBB13*, *S10-SFBB2*, *S3-, S7-, S28-SFBB3*, *S2-, S7-SFBB4*, *S2-, S5-, S28-SFBB7*, *S2-SFBB8, S5- MdSFBB9*/*SorbusSFBB11*,

S5-, S7-, S10-MdSFBB10/*SorbusSFBB12*, *S7- MdSFBB11*/*SorbusSFBB9, S5-, S7-, S24-MdSFBB12*/*SorbusSFBB10, S9-, S24-, S25-*Md*SFBB13*/*SorbusSFBB1*, *S5-, S7-, S10-, S28-SFBB14*, *S10-SFBB16*, and *S5-, S7-, S25-SFBB18* genes (grey blocks in Fig. 1). These sequences were only present in the transcriptome of the cultivars presenting those *S*-haplotypes, when blastn was performed. There were still seven alleles (*S25-SFBB7*, *S25-SFBB8, S28-MdSFBB9*/*SorbusSFBB11*, *S5-MdSFBB11*/*SorbusSFBB9*, *S3-MdSFBB12*/*SorbusSFBB10*, *S10-MdSFBB12*/*SorbusSFBB10*, and *S9-SFBB16*; Fig. 1) that were not amplified using specific primers. They may represent divergent alleles or missing genes in these *S*-haplotypes.

The 157 sequences obtained by PCR, plus the 13 from *S3*-, *S9*-, and *S10*-haplotypes (Fig. 1) show 100% match to 728 Edena contigs. Manual inspection of the 22 remaining Edena contigs revealed eight (from 'Empire' and 'Red Delicious' transcriptomes) that were assembled into a single larger sequence that shared less than 92% identity with sequences in our dataset. This sequence was present in the 'Empire' and 'Red Delicious' transcriptome, and was named *S28-SFBB22* (black boxes in Fig. 1). Three other contigs from 'McIntosh' were also assembled into a larger sequence that shared 98% identity with *MdSFBB1*/*SorbusSFBB13* sequences. This sequence was only present in the 'McIntosh' transcriptome and was called *S25-SFBB23* (black boxes in Fig. 1). Four other sequences from the 'Fuji' and 'Northern Spy' anthers transcriptomes show overlap and, thus they can be assembled into a larger sequence that shows less than 90% homology with sequences in our dataset. This sequence is only present in transcriptomes of these two cultivars and was called *S1-SFBB24* (black boxes in Fig. 1). These sequences have been confirmed using specific primers, in PCR reactions using genomic DNA of these cultivars. The remaining seven sequences represent almost exclusively 5′and 3′ regions of alleles for which data has been obtained. In conclusion, 173 sequences were obtained that covered more than 80% of the *SFBB* coding region, and 98% of these sequences include the F-box region (60% have the start codon).

Number of *SFBB* genes in *M. domestica*. The phylogenetic relationship of the 173 sequences obtained in this work support the existence of, at least, 24 *SFBBs* (Fig. 2). *S9-SFBB19* clusters within *SFBB8* and *SFBB16* sequences. Although this sequence could represent a very divergent *SFBB16* allele for the *S9*-haplotype, diversity levels (0.228, after Jukes and Cantor correction) support that this sequence represents a different gene. When primers were designed for this gene sequence (*SFBB19*, Supplementary Table S6), an amplification product with expected size (810 bp) was observed only in 'Fuji' and 'Red Delicious', the cultivars with the *S9*-haplotype. A similar result was obtained when SFBB19F primer was combined with the SFBBgenR, and SFBB19R primer with SFBBgenF (Supplementary Table S6).

The sequences assigned as alleles of a *SFBB* gene cluster together with strong support. The exceptions are the *SFBB5* gene (*S1*- and *S24-SFBB5* are divergent alleles) and *MdSFBB1*/*Sorbus SFBB13* gene (*S5- MdSFBB1*/*Sorbus SFBB13* is a divergent allele). According to the levels of synonymous diversity (0.1 and 0.09 after Jukes and Cantor correction, respectively), there is no support for these sequences representing new genes.

Differences in number and order of *SFBBs* between *S*-haplotypes has been previously observed[23,44]. The number of *SFBB* genes varied from 17 (*S3*-, *S5*-, and *S25*-haplotypes) to 19 (*S1*-, and *S28*-haplotypes) (Fig. 1). When

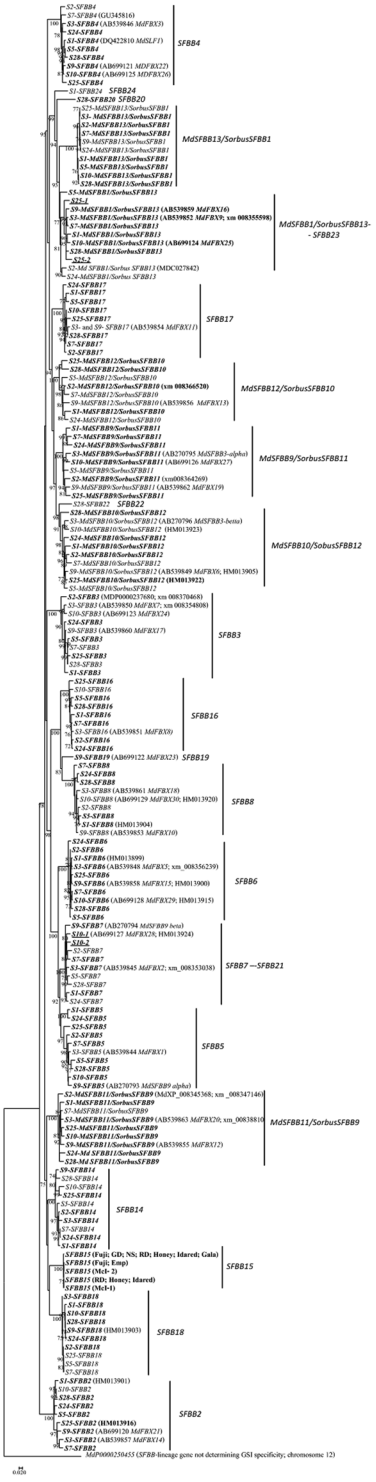

Figure 2. Maximum-likelihood phylogenetic tree showing the relationship of the 173*M. domestica SFBB* sequences obtained for 10 *S*-haplotypes. The tree was rooted with *SFBB* -lineage gene MDP0000250455 (not located in the *S*-locus region, and not involved in GSI)[10]. In brackets are the GenBank acc. numbers for the sequences previously described. In bold are the sequences obtained from the PCR reaction using primers SFBBgenF and SFBBgenR. Numbers below the branches represent bootstrap values above 70.

the 24 genes were used as query in a blast search against reads form style, stigma, ovary, filaments, receptacle, petals, sepals, receptacle and young leaves from 'Golden Delicious' cultivar transcriptomes, no reads supported the existence of these sequences. In contrast, all *SFBB* genes are expressed in anthers (Supplementary Fig. S3). Therefore, all these genes are expressed in anthers and pollen only, as those involved in GSI (Aguiar *et al.*[10], and

Gene	Alleles not amplified with SFBBgen primers	Alleles characterized with specific primers
MdSFBB1/SorbusSFBB13	*S2; S24*	*S2; S24*
SFBB2	*S10*	*S10*
SFBB3	*S5; S7; S28*	*S5; S7; S28*
SFBB4	*S2; S7*	*S2; S7*
SFBB7	*S2; S5; S25; S28*	*S2; S5, S28*
SFBB8	*S2; S25*	*S2*
MdSFBB9/SorbusSFBB11	*S5; S28*	*S5*
MdSFBB10/SorbusSFBB12	*S5, S7, S10*	*S5, S7, S10*
MdSFBB11/SorbusSFBB9	*S5; S7*	*S7*
MdSFBB12/SorbusSFBB10	*S3; S5; S7; S24*	*S5; S7; S24*
MdSFBB13/SorbusSFBB1	*S9; S24; S25*	*S9; S24; S25*
SFBB14	*S5; S7; S10; S28*	*S5; S7; S10; S28*
SFBB16	*S9; S10*	*S10*
SFBB18	*S5; S7; S25*	*S5; S7; S25*

Table 2. Alleles for 14 *SFBB* genes that were not identified with SFBBgen primers, but were identified with specific primers. The alleles were named with the *SFBB* and the haplotype from which it was identified.

references therein). Except for *SFBB15*, in every case each allele could be associated to a *S*-haplotype, thus indicating linkage to the *S-RNase* gene.

Associations between *SFBB* genes and the *S-RNase* gene. Progeny segregation from three crosses were analysed to test the linkage between the *S-RNase* and each of the *SFBB*- like genes: 'Golden Delicious' (*S2, S3*) × 'Red Delicious' (*S9, S28*) - 27 individuals analyzed, 'Gala' (*S2, S5*) × 'McIntosh' (*S10, S25*) - 48 individuals analyzed, and 'Fuji' (*S1, S9*) × 'Honeycrisp' (*S2, S24*) - 34 individuals analyzed (Supplementary Table S7; Supplementary Table S8). We used specific primers for conserved regions of each *SFBB* gene (Supplementary Table S6) and the amplification products for each individual was digested with selected enzymes that distinguished the alleles present in each individual, according to the sequences previously obtained (Supplementary Table S7; Supplementary Table S8). This methodology differentiated 17 genes (Supplementary Table S7; Supplementary Table S8) but not *MdSFBB13/SorbusSFBB1*, *SFBB15*, and *SFBB18*. These three *SFBB* genes have levels of synonymous diversity bellow 0.013, and thus there were no polymorphic restriction enzyme cut sites that could be used as allele specific markers. The *S-RNase* alleles were also genotyped for the 109 individuals (Supplementary Table S7; Supplementary Table S8), using specific primers (Supplementary Table S9). All 17 *SFBB* genes analyzed were linked with the *S-RNase* gene (Supplementary Table S7; Supplementary Table S8). This result supports the role of these *SFBB*s as *S*-pollen genes.

Inferring recombination, mutation and diversifying selection at *SFBB* genes. Levels of polymorphism for the *SFBB* genes are, on average, 4.2 times lower than those observed for the *S-RNase* (Table 3)[22,23,26,34,48,49], despite the evidence for specific associations between *SFBB*s and the *S-RNase*. To address the effect of the *S*-locus on the polymorphism levels at *SFBB* genes, the levels of diversity were determined for 126 single copy genes expressed in *M. domestica* anthers transcriptomes (Material and Methods), for which a sequence fragment larger than 100 bp had been obtained in, at least, four cultivars. Both synonymous and non-synonymous diversity levels at the *SFBB*s were higher than those of the 126 *M. domestica* single copy genes expressed in anthers (Fig. 3; Mann-Whitney, $P < 0.001$). Therefore, *SFBB* diversity is being affected by recombination, and/or diversifying selection. We found evidence for recombination for all *SFBB*s present in more than one *S*-haplotype, except *SFBB15* using different methodologies (Table 3), although for *MdSFBB13/SorbusSFBB1*, *SFBB16*, and *SFBB18* genes not all tests support evidence for recombination. It should be noted that RDP uses phylogenetic incongruence, and thus depends on the amount of diversity in the data, and thus is less powerful[56]. Nevertheless, we find evidence for recombination at the *S-RNase* gene. Therefore, the differences observed seem not to be due to recombination alone. On the other hand, we found evidence for diversifying selection only at one, two, and three amino acid positions at *SFBB7*, *SFBB6*, and *SFBB8*, respectively (Table 3). Thus, there is little evidence for diversifying selection at the *SFBB* genes, in contrast with the *S-RNase*. The different selection regimes at the *S*-pollen and *S-RNase* genes seem to be the major cause for the differences on levels of diversity.

Positively selected amino acid sites in 17–19 *M. domestica SFBB* genes at each of the 10 *S*-haplotypes. In the collaborative non-self -recognition model each *S*-pollen gene recognises a sub-set of non-self- S-RNases, but not the S-RNase of its *S*-haplotype[19–24]. Recently, Kubo and co-authors[20] proposed for *Petunia* species a more detailed model that falls under the general collaborative non-self -recognition model. Under Kubo and co-authors[20] model, having either a diverged or deleted allele at a *SLF* gene, whose product usually recognizes a given Sx-RNase, is the way by which recognition avoidance of the own Sx-RNase is achieved. All non-divergent alleles would recognize the Sx-RNase. In agreement with this model, in *Petunia*, phylogenetic analyses show divergent and non-divergent alleles as two distinct allele groups. The phylogenetic inferences led to the identification of *SLF* genes that recognize seven S-RNases, among eight *S*-haplotypes analysed, and in five cases their predictions have been confirmed with experimental evidence. It should, however, be noted, that in *Petunia*, different

Gene	N	K_s	K_a	Number of sites analysed	Rm	4GT	RDP	Number of synonymous mutations inferred in the phylogeny	Model	Recombination events per synonymous mutation
MdSFBB1/SorbusSFBB13	9	0.08361	0.03399	898	7	103/6670	3	70.70163	M0	0.042432
SFBB2	9	0.05807	0.01914	750	4	12/2926	1	40.32852	M0	0.024796
SFBB3	10	0.04639	0.02032	812	3	96/2346	3	33.565	M0	0.089379
SFBB4	10	0.06119	0.02836	1156	13	193/10585	1	84.20016	M0	0.011876
SFBB5	8	0.08712	0.02078	879	6	82/4371	3	63.89838	M0	0.04695
SFBB6	10	0.03746	0.00726	935	7	73/946	2	39.35508	M2 (100; 271)	0.050819
SFBB7	8	0.07138	0.02701	770	7	56/3486	1	49.29736	M2 (144)	0.020285
SFBB8	9	0.02204	0.0183	809	8	84/1176	1	16.94118	M2 (117;182;304)	0.059028
MdSFBB9/SorbusSFBB11	9	0.10253	0.03253	837	9	70/8385	3	82.37295	M0	0.03642
MdSFBB10/SorbusSFBB12	9	0.10285	0.02224	851	8	50/4753	1	79.27522	M0	0.012614
MdSFBB11/SorbusSFBB9	9	0.07466	0.02444	761	6	31/4186	1	54.92304	M0	0.018207
MdSFBB12/SorbusSFBB10	8	0.15013	0.03066	465	6	65/2415	2	60.88446	M0	0.032849
MdSFBB13/SorbusSFBB1	10	0.01162	0.00251	765	1	4/55	0	7.31024	M0	0
SFBB14	10	0.07694	0.01797	712	4	19/2485	1	46.15336	M0	0.021667
SFBB15	5	0.00212	0.00204	876	0	0/6	0	1.06967	M0	0
SFBB16	9	0.05735	0.01659	803	5	34/2556	0	38.2542	M0	0
SFBB17	9	0.03279	0.01341	812	1	13/1378	4	25.50332	M0	0.156842
SFBB18	10	0.01231	0.0022	543	1	1/10	0	4.15359	M0	0
S-RNase	10	0.25541	0.20965	420	26	610/10731	5	108.769	M2 (17)	0.045969
	19*	0.22702	0.19136	660	51	1580/20706	14	220.86	M2 (26)	0.0634

Table 3. DNA sequence variation summary for sequences of 18 *SFBBs* and the *S-RNase* from *M. domestica*. N- number of sequences used. K_s - ratio of synonymous substitutions per synonymous site. K_a - ratio of non-synonymous substitutions per non-synonymous site. Rm- minimum number of recombination events[87]. 4GT - number of pairwise comparisons presenting the four gametic types over the total number of all pairwise comparisons. RDP- number of independent recombination events[85]. Model- Yang's[57] model used to infer the total number of synonymous mutations implied by the data. In brackets- amino acid sites identified as positively selected, using the method of Yang[57] implemented in ADOPS[86] with a probability higher than 90% in both NEB (naive empirical Bayes) and BEB (Bayes empirical Bayes). *only complete sequences were used.

non-divergent alleles of the same *SLF* gene (see Fig. 4 in [34]for *SLF1* gene) can recognize different S-RNases. For instance, *Petunia S7*- and *S5-SLF1* alleles can recognize S17- and S9-RNases, but S11-SLF1 only recognizes the S17-RNase, and not the S9-RNase.

In *Malus*, for six *SFBB* genes, alleles could be found for only 8 or 9 of the 10 *S*-haplotypes analysed (missing boxes in Fig. 1). This finding could suggest that the *Malus* system may work in a way similar to that proposed by Kubo and co-authors[20] for *Petunia*, since it is conceivable that if more *S*-haplotypes are analysed missing alleles will be found at all *SFBB* genes. In *Malus*, no divergent alleles were, however, found at *SFBB* genes. It should be noted that in *Petunia*, divergent alleles show less than 90% identity with non-divergent alleles, and in *Malus* such value is only observed when different genes are being compared. This could be due to the effect of recombination, since in *Petunia* intragenic recombination is only inferred for two *SLF* genes[20] and in *Malus* recombination is inferred for 94% of the *SFBB* genes (Table 3). In the presence of recombination it is not possible to have clearly defined allele groups. Nevertheless, we can still make some inferences, by assuming that for any *SFBB* gene, alleles that are identical at the amino acid sites responsible for specificity recognition, are targeting the same set of S-RNases. Moreover, those alleles cannot recognize the S-RNases to which they are linked. Indeed, polymorphism levels at the *SFBB* genes are low (see above), and thus, natural selection must favour diversification of *SFBB* genes within a *S*-haplotype[22,34]. Evidence for adaptive evolution at the *SFBB* paralogous genes has been found using codeML[57] and 11 *SFBB* genes of two *Sorbus S*-haplotypes[22]. 12 amino acid sites were identified as being positively selected, and these amino acid positions were found to be polymorphic when comparing the alleles of different *S*-haplotypes for each *SFBB* gene, thus supporting the involvement of these amino acids in specificity determination[22]. Using the same methodology and the 17 to 19*M. domestica SFBB* genes of each *S*-haplotype here characterized, we identified 21amino acid sites under positive selection (Table 4; B + database[58] (bpositive.i3s.up.pt; see the *Malus* SFBB BP2017000011 dataset)). Supplementary Fig. S4 shows these amino acid sites on top of a reference alignment of the *SFBB4* gene. Since most of the sequences here used do not cover the region where positively selected amino acid position 377 is located, this position has not been considered in the remaining analyses. Assuming that the positively selected amino acid sites are those determining *S*-pollen specificity, the *MdSFBB13/SorbusSFBB1* and *SFBB18* genes, which do not present polymorphism at these positions, are not involved in the recognition of any of the 10 *S-RNases* here studied. Assuming that within a *SFBB* gene an allele showing one difference at the positively selected amino acid positions (those that are involved in specificity determination) is sufficient to prevent the recognition of a given *S-RNase* specificity, the number of sequences that can be distinguished based on these amino acid positions only, gives insight into the maximum number of different

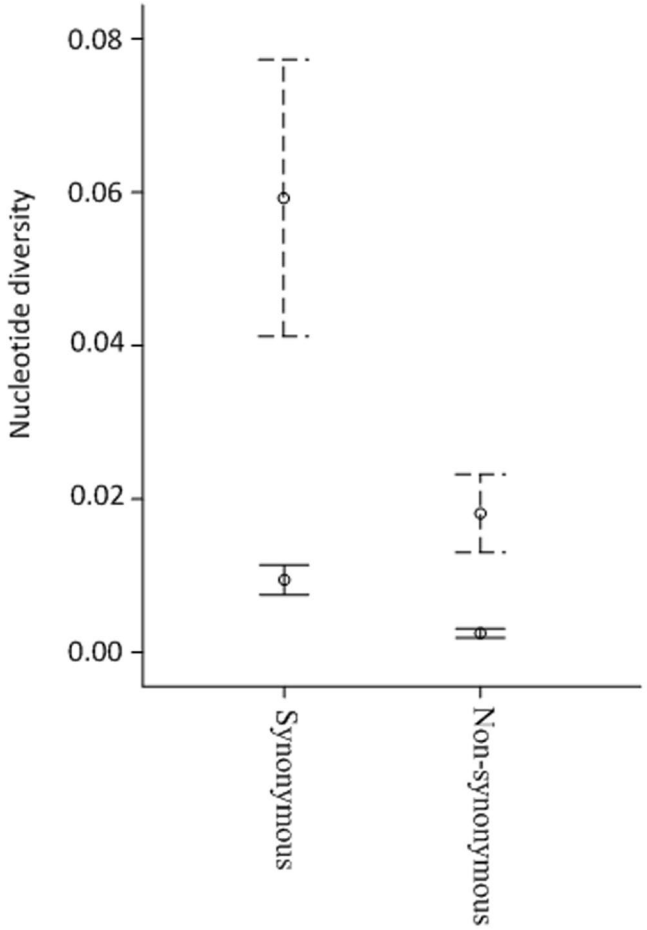

Figure 3. Box plot of synonymous and non-synonymous nucleotide diversity at genes expressed in anthers that are not located at the *S*-locus and *SFBBs* (dotted lines).

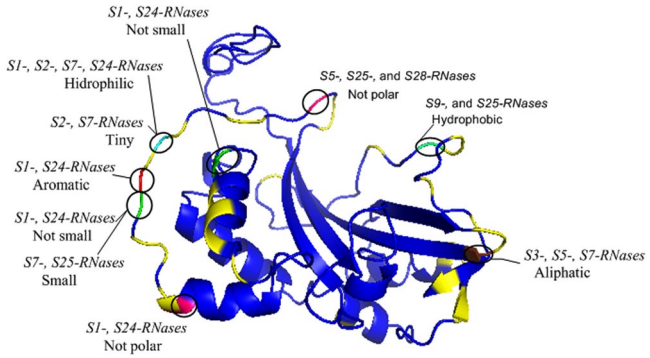

Figure 4. Positively selected amino acid sites mapped onto the S-RNase crystal structure of *M. domestica* S7-RNase, obtained as in Vieira *et al.*[61]. Positively selected amino acid positions that are putatively involved in SFBB specificities recognition are highlighted. The features of those amino acid positions that have been inferred to be important for discriminating different SFBBs are shown. Green- size, brown- aliphatic, red- aromatic, pink-polarity, light green- hidrophobicity, and light blue- hidrophobicity and size.

S-RNase specificities recognized by a single *SFBB*. In our dataset, the maximum number of different *S-RNase* specificities recognized by a single *SFBB* gene is eight (see *MdSFBB1*/Sorbus*SFBB13* gene; Table 4). Positively selected amino acid sites that are invariant within a given *SFBB* gene can be devoted to the recognition of the same S-RNase as proposed by Kubo and co-authors[20]. In the case of *SFBB* genes presenting missing alleles, the S-RNase that is recognised is likely the one linked to the *S*-haplotype presenting the missing allele, as proposed by Kubo and co-authors[20]. Within a *S*-haplotype there is always a minimum set of genes that can, in principle, recognize all *S-RNase* specificities here considered, but the self *S-RNase* (for instance for the *S1*-haplotype, genes

Gene	71	77	81	112	117	119	132	160	162	169	170	188	217	232	235	251	253	281	303	304	S-haplotype
MdSFBB1/SorbusSFBB13	A	N	P	L	F	Q	A	P	K	I	G	Q	M	E	H	N	G	E	D	E	S1
	.	.	.	.	.	E	.	.	.	.	.	.	.	.	.	T	.	.	.	A	S2
	.	.	.	.	.	.	.	.	Q	.	.	.	.	.	.	I	.	.	.	.	S3
	.	.	.	.	.	E	.	.	.	.	.	.	.	D	.	T	.	.	.	A	S5
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	S7, S25
	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	I	.	.	.	A	S9
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	S10, S28
	.	.	.	.	.	E	.	.	.	.	.	.	T	.	.	T	.	.	.	.	S24
SFBB2	M	R	R	L	H	S	V	P	E	T	Q	K	T	D	N	S	D	L	D	N	S1, S9, S24, S28
	.	.	.	.	L	.	.	.	K	S	.	.	.	.	.	.	.	.	.	.	S3
	.	.	.	.	N	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	S5
	.	.	.	.	.	.	.	Q	G	.	.	.	.	.	.	.	.	.	.	.	S7
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	S10
	.	.	.	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	S25
SFBB3	F	Q	R	R	H	Q	E	P	E	T	H	Q	T	S	P	T	G	N	E	D	S1
	.	.	H	.	.	H	.	.	.	.	.	.	.	T	.	R	.	K	.	.	S2
	.	.	.	.	P	.	.	.	.	.	.	.	.	T	.	.	.	E	.	.	S3
	.	.	.	.	P	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	S5, S7, S9, S24, S25, S28
	.	.	.	.	P	E	.	.	.	.	.	.	.	T	.	I	.	E	.	.	S10
SFBB4	V	K	H	R	H	L	S	L	G	D	G	R	M	K	P	R	G	E	Q	D	S1, S24
	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.	.	.	E	.	S2
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	E	.	S3, S5, S9, S10
	.	.	Q	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	E	.	S7
	.	.	.	.	.	.	.	.	S	.	.	.	.	.	.	.	.	.	E	.	S25
	.	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	E	.	S28
SFBB5	V	R	Q	M	N	E	V	—	K	I	K	Q	M	D	P	N	N	K	—	—	S1, S24
	.	.	.	.	.	.	.	.	.	V	.	R	.	.	.	M	.	.	—	—	S2, S3, S5, S7, S28
	.	.	.	.	.	.	.	.	.	V	.	R	.	.	.	C	.	.	—	—	S9
	.	.	.	.	.	.	.	.	.	.	.	R	.	.	.	.	.	.	—	—	S10
	.	K	.	.	.	D	.	.	.	.	.	R	.	.	.	M	.	.	—	—	S25
SFBB6	V	R	R	I	N	Q	V	—	M	L	K	R	T	D	P	N	N	K	—	—	S1, S25
	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	S2, S3, S5, S7, S9, S10, S24, S28
SFBB7	V	R	Q	I	N	E	.	—	K	T	K	R	T	E	P	Y	N	N	.	.	S1
	.	.	.	M	.	.	V	.	.	I	.	.	.	.	.	.	.	K	.	.	S2
	.	.	.	M	.	.	V	.	.	.	.	.	.	.	.	.	.	K	.	.	S3, S5, S7
	.	.	.	M	.	.	V	.	.	.	.	.	M	.	.	.	.	K	.	.	S9
	.	.	.	M	.	.	.		.	.	.	.	.	.	.	.	.	K	—	.	S10—1
	.	.	.	R	.	.	V		.	.	.	.	.	.	.	.	.	K	—	.	S10—2
	.	.	.	.	.	.	V		.	.	.	.	.	.	.	.	.	K	.	.	S24
	.	.	.	M	.	D	.		.	.	K	.	.	.	.	.	.	K	.	.	S28
SFBB8	A	E	Q	R	E	E	V	G	K	T	K	R	M	K	P	C	V	K	—	—	S1, S5
	.	.	.	K	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	S2, S9, S10, S24, S28
	.	Q	.	K	.	.	.	.	.	.	.	.	.	.	.	.	L	.	.	.	S3
	.	Q	.	K	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	S7
MdSFBB9/SorbusSFBB11	A	Q	Q	L	F	L	A	P	E	S	Q	R	T	T	S	T	G	R	E	D	S1, S2, S7, S24
	.	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	.	.	.	S3
	P	.	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	S5
	.	.	.		.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	S9
	.	.	.	.	.	.	.	.	.	N			M	.	.	.	.	.	.	.	S10
	.	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	.	.	.	.	S25

Continued

Gene	71	77	81	112	117	119	132	160	162	169	170	188	217	232	235	251	253	281	303	304	S-haplotype
MdSFBB10/SorbusSFBB12	A	Q	Q	L	F	L	A	P	K	T	Q	R	T	Q	H	S	S	T	E	D	*S1, S5, S25*
	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	.	N	.	.	.	*S2, S7*
	.	.	.	.	.	.	.	.	.	.	.	.	.	K	.	.	N	.	.	.	*S3*
	.	.	.	.	.	.	.	.	.	.	.	Q	.	.	.	.	.	.	.	.	*S9*
	.	.	.	.	.	.	.	.	.	N	.	.	.	S	.	.	N	.	.	.	*S10*
	.	.	.	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	*S24*
	.	.	.	.	.	.	.	.	.	.	.	.	.	E	.	.	.	.	.	.	*S28*
MdSFBB11/SorbusSFBB9	M	Q	Y	T	P	Q	I	P	E	I	E	Q	T	K	Q	N	G	K	E	D	*S1*
	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	S	.	.	.	.	*S2*
	.	.	.	.	.	.	T	.	.	.	.	.	A	.	.	.	.	.	.	.	*S3*
	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	*S7, S10, S24*
	.	.	.	M	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	*S9, S28*
	.	.	.	.	.	.	T	.	Q	.	.	.	.	.	.	.	.	.	.	.	*S25*
MdSFBB12/SorbusSFBB10	P	K	Q	L	F	Q	V	E	G	T	E	Q	T	T	S	T	D	T	G	D	*S1, S9*
	—	—	—	—	—	—	A	.	.	.	.	.	.	.	.	.	.	.	.	.	*S5*
	.	.	.	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	*S2, S7*
	—	—	—	—	—	.	.		.	.	.	.	.	.	.	.	.	—	—	—	*S24*
	.	Q	.	.	.	.	.	.	.	.	.	.	.	K	.	.	.	.	.	.	*S25*
	.	Q	.	.	.	.	L	.	.	.	.	.	.	N	.	.	.	.	.	.	*S28*
MdSFBB13/SorbusSFBB1	N	R	P	L	F	E	A	S	R	I	T	Q	T	E	C	T	E	K	D	E	*S1, S2, S3, S5, S7, S9, S10, S24, S25, S28*
SFBB14	M	K	Y	L	P	Q	A	P	E	I	G	Q	I	K	P	S	G	K	E	D	*S1, S2, S3, S5, S7, S24*
	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	D	.	*S9, S25*
	.	.	.	.	.	.	.	.	.	S	.	.	.	E	.	.	.	.	.	.	*S10*
	.	.	.	.	.	.	.	.	.	.	R	.	.	.	.	.	.	.	.	—	*S28*
SFBB16	T	D	R	Q	E	I	L	G	K	T	K	R	T	K	P	S	D	K	—	—	*S1, S2, S3, S7, S24*
	M	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	*S5, S25, S28*
	M	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.	*S10*
SFBB17	T	N	Q	L	Y	L	A	P	K	V	R	Q	T	K	S	T	A	K	D	K	*S1, S2, S3, S7, S9, S10, S25, S28*
	.	.	.	.	.	.	.	.	.	.	.	.	.	R	.	.	.	.	.	.	*S5*
	.	.	.	.	.	.	.	.	.	.	.	.	.	E	H	.	G	.	.	.	*S24*
SFBB18	M	D	Y	M	P	L	T	P	E	T	R	R	T	K	P	T	G	K	E	D	*S1, S2, S3, S5, S7, S9, S10, S24, S25, S28*

Table 4. Amino acid composition for each SFBB in the 10 *S*-haplotypes for the amino acid sites identified as positively selected in intra-haplotypic analyses of *M. domestica* SFBBs using 10 *S*-haplotypes (see B + database[58] (bpositive.i3s.up.pt; see the *Malus* SFBB BP2017000011 dataset, for analyses). Sites were identified using the method of Yang[57] implemented in ADOPS[86] with a probability higher than 95% in NEB (naive empirical Bayes) or BEB (Bayes empirical Bayes) in at least one *S*—haplotype. The positions are according to the alignment of *SFBB4* gene presented in Supplementary Fig. S4.

MdSFBB1/SorbusSFBB13, *SFBB3*, *SFBB7*, *MdSFBB11/SorbusSFBB9*, could alone recognize the *S2-*, *S3-*, *S5-*, *S7-*, *S9-*, *S10-*, *S24, S25*, and *S28-RNases*; Table 4). For the *S*-haplotypes here considered, on average, there are five (3 to 7) *SFBB* genes that alone can recognise all the *S-RNase* specificities here considered but the self *S-RNase* (Table 4). Thus, there are multiple *SFBB* genes recognizing the same *S-RNase* specificity.

It is known that protein-protein interactions depend on properties such as residue interface propensities, hydrophobicity and conformational changes[59,60]. In *P. hybrida* it has been shown that one alteration of an amino acid under positive selection at the C-terminal SLF protein, was sufficient to change *S*- pollen specificity, because it causes a change in the surface electrostatic potential[36]. To identify features at the pistil amino acid sites under positive selection[61](see Supplementary Fig. S5 for those amino acid sites for the 10 *S-RNases* here analysed), such as hydrophobicity, polarity, aliphatic, charge, size, and aromatic, that can determine pistil-pollen interactions, we determined whether these amino acid proprieties are exclusively found in a group of *S-RNases* that share in their *S*-haplotype, for a particular *SFBB* gene, *SFBB* alleles with identical sequences at the amino acids under positive selection (*SFBB* alleles that recognize the same *S-RNase* specificity; Table 4). For instance, *S1-* and

S24- haplotypes have identical amino acids at sites under positive selection at two genes, namely *SFBB4* and *SFBB5*. Therefore, none of these *SFBB* genes is able to recognize either the *S1-* and *S24-RNases*. Comparing the above features of the amino acids under positive selection for the *S1-* and *S24-RNases* to the remaining *S-RNases* present in the cultivars analysed, we observed that *S1-* and *S24-RNases* have proprieties that are unique in four of these sites (the two *S-RNases* are the only ones that at amino acid position 80 present an aromatic amino acid, at positions 81 and 125 amino acids are not small, and at position 88 the two *S-RNases* are the only ones that are not polar; Fig. 4). This suggests that the amino acid composition of the *S-RNase* at these sites prevents the interaction with the protein encoded by the *SFBB4* gene at amino acid position 304 (that is unique in the *SFBB4* alleles analysed, Table 4) or/and with *SFBB5* gene at position 188 (Table 4). For *S2- S7-MdSFBB10*/*SorbusSFBB12*/*S2- S7-RNases* and also *S2- S7-MdSFBB12*/*SorbusSFBB10*/*S2- S7-RNases*, at amino acid position 78, *S2-, S7-RNases* are the only ones presenting a tiny amino acid. This position prevents the interaction with the protein encoded by the *MdSFBB10*/*SorbusSFBB12* gene at amino acid position 132 (Table 4), and with *MdSFBB12*/*SorbusSFBB10* gene at position 217 (Table 4). Results pointing to one amino acid site preventing the protein-protein interaction between *S-RNase* and a *SFBB* are also obtained for *S3-, S5-, S7- SFBB7*/*S3-, S5-, S7-RNases* (at amino acid position 200 *S3-, S5-, S7-RNases* are the only ones presenting an aliphatic amino acid; Fig. 4), *S1-, S2-, S7-, S24- MdSFBB9*/*SorbusSFBB11*/*S1-, S2-, S7-, S24-RNases* (at amino acid position 78 *S1-, S2-, S7-, S24-RNases* are the only ones presenting a non hydrophobic amino acid; Fig. 4), *S7-, S25-MdSFBB1*/*SorbusSFBB13*/*S7-, S25-RNases* (at amino acid position 81 *S7-, S25-RNases* are the only ones presenting a small amino acid; Fig. 4), *S9-S25-SFBB14*/*S9- S25-RNases* (at amino acid position 227, *S9-, S25-RNases* are the only ones presenting a hydrophobic amino acid), and *S5-, S25-, S28-SFBB16*/*S5-, S25-, S28-RNases* (at amino acid position 70 *S5-, S25-, S28-RNases* are the only ones presenting a non polar amino acid; Fig. 4).

Discussion

The number of *M. domestica SFBB* genes present in a given *S-* haplotype varies from 17 to 19 (Fig. 1). A similar number of genes is observed in *Petunia*[20,21], although the two systems may have evolved independently[10]. Under the assumption that each *S*-pollen can recognize a different proportion of target *S-RNases* (according to *Petunia* transformation experiments a *S*-pollen gene can recognize 18.6% of S-RNases on average), Monte Carlo simulations revealed that between 16 to 20 *S*-pollen genes are sufficient to recognize 40 *S-RNases* specificities[20]. Since the number of *Malus SFBB* genes is lower than the number of S-RNase specificities described in *M. domestica*[61], each *S*-pollen must recognize a different proportion of target *S-RNases*, like in *Petunia*. In *M. domestica* there are 59 *S-RNase* unique sequences in GenBank, that according to the sites under positive selection[61], define 34 *S-RNase* specificities. Moreover, our results show that within a *S*-haplotype, 20% of the genes can recognise all the *S-RNase* specificities studied but the self *S-RNase*. Therefore, it is not surprising that the two systems have a similar number of *S*-pollen genes, independently of their evolution.

In *Petunia*, it has been observed that either divergent or absent alleles at a particular *S*-pollen gene are those determining *S-RNase* specificity recognition[20]. In *M. domestica* we find six genes that are absent in five *S*-haplotypes (*S3-*, *S5-*, *S9-*, *S25-*, and *S28-* haplotypes; Fig. 1; Table 2) and six genes that were detected in a single *S*-haplotypes (*S1-*, *S9-*, *S10-*, *S25-*, and *S28-* haplotypes; Fig. 1 and Fig. 2). Therefore, it seems that absent alleles are also important in *M. domestica* specificity determination, although these observations are not sufficient to account for the 10 *S-RNase* specificities here analysed. When amino acids under positive selection are considered, we can account for all specificities in the data set. Furthermore, the data supports the prediction that different *SFBB* genes are involved in the recognition of the same non-self *S-RNase* specificity.

Although we do not know how *S*-pistil and *S*-pollen proteins interact to allow self/non-self recognition and discrimination, the chemical characteristics of amino acids under positive selection at both proteins must be determinant for such interactions. Under the assumption that two *SFBB* alleles, from two different *S*-haplotypes, showing identical amino acids at sites under positive selection cannot recognize any of the two *S-RNase* specificities of the two *S*-haplotypes, we find at the corresponding *S-RNase* chemical characteristics at amino acids under positive selection such as hydrophobicity, polarity, aromatic, aliphatic, and size, that are exclusively found in these, and thus must be involved in the self/non-self recognition (Fig. 4). The assumption that one amino acid under positive selection is sufficient for self/non-self recognition seems to be realistic since in *Petunia* the alteration of a single C-terminal amino acid under positive selection at one *S*-pollen gene is sufficient to change *S*-pollen specificity[36]. Here we identified putative interactions for the amino acid positions unique in the *SFBB* alleles that could recognize as self a particular set of *S-RNase* specificities (Fig. 4), but further interactions can be predicted by considering amino acid sites under positive selection that are shared with other *SFBB* alleles. Such inferences are essential for guided experimental validation.

Having multiple *SFBB*s to detoxify a given non-self S-RNase will reduce the loss of cross-compatibility caused by mutations and/or recombination[62]. In *M. domestica* we found evidence for duplications within a *S*-haplotype for two genes (within *S10*, for the *SFBB7* vs. *SFBB21* genes, and within *S25*, for the *MdSFBB1*/*SorbusSFBB13* vs. *SFBB23* genes). For the *MdSFBB1*/*SorbusSFBB13* vs. *SFBB23* gene pair, the observed sequence relationships are not those expected under a model of gene duplication without intragenic recombination. Nevertheless, it is compatible with a model where there is intragenic recombination and where the duplicated gene no longer recombines with the gene that gave origin to it, making most alleles of the *MdSFBB1*/*SorbusSFBB13* similar among them, but not the *MdSFBB1*/*SorbusSFBB13* and the *SFBB23* gene. Recombination can also contribute to the gene number variation observed in *S*-haplotypes, as well as in the development of chimeric *SFBB*-genes that can encode novel specificities. Intragenic recombination is detected in all *SFBB* genes showing more than two different sequences, except for *SFBB15*. In *Petunia* evidence for *S*-pollen genes intragenic recombination has been reported[20]. Therefore, duplication and recombination are essential for functional diversification, and thus for generation of *S*-pollen specificities. Nevertheless, it is possible that the number of *SFBB* genes per *S*-haplotype

is constrained by the fitness costs of having more genes, as observed for genes involved in the recognition of pathogen avirulence[63–73].

Material and Methods

Plant material and RNA-DNA extractions. In Pyrinae there are no homozygous lines and thus, in this work, we selected a set of nine cultivars ['Fuji' (*S1*, *S9*), 'Northern Spy' (*S1*, *S3*), 'Golden Delicious' (*S2*, *S3*), 'Gala' (*S2, S5*), 'Honeycrisp' (*S2*, *S24*), 'Idared' (*S3*, *S7*), 'Red Delicious' (*S9*, *S28*), 'McIntosh' (*S10*, *S25*), and 'Empire' (*S10*, *S28*)], where two to three cultivars share six specificities (*S1, S2, S3, S9, S10*, and *S28*). Since no or little diversity is expected for alleles of the *S*-genes within the same specificity[51–55], these are, in principle, equivalent to the use of two biological replicates for the *S*-locus genes. Three of these *S*-haplotypes were used as controls since the *SFBB* genes have already been characterized[23,27,44]. Anthers from flower buds 1–3 days prior to opening were collected from trees of the above nine *M. domestica* cultivars, growing at Michigan State University campus, East Lansing, Michigan. The anthers were immediately frozen in liquid nitrogen and stored at −80 °C for RNA extraction. Anthers were used since *SFBBs* show higher expression levels at this tissue (Fig. 4 in Aguiar *et al.*[10]). Flower buds were also collected from these individuals for DNA extraction. Additional tissues were collected for 'Golden Delicious': petals, sepals, filaments, receptacle, styles, stigmas, and ovaries from open flowers, and immature leaves from new shoot growth for RNA extraction.

Controlled crosses between 'McIntosh' × 'Gala', 'Fuji' × 'Honeycrisp', and 'Golden Delicious' × 'Red Delicious' were performed and 48, 34, and 27 seeds, respectively, were obtained. The seeds where germinated and leaves were collected from the seedlings and stored at −20 °C for DNA extraction. No permits were required for the field collection, since the plant location is part of Michigan State University and *M. domestica* is not an endangered or protected species.

RNA and DNA extraction, RNA Library Construction, and Sequencing. Total RNA was extracted using the mirVana™ miRNA Isolation Kit (Ambion), using the manufacturer's guidelines for recovery of total RNA. RNA quantity was assessed using a NanoDrop v.1.0 (Thermo Scientific) and RNA quality was evaluated by BioRad's Experion System. cDNA libraries construction and sequencing was performed using the Illumina TruSeq protocol and reagents with 100-bp, paired-end sequencing. A total of 138380723 read pairs were obtained for the anther transcriptomes (Supplementary Table S1). Genomic DNA was extracted using the method of Ingram *et al.*[74] or the Puregene® DNA purification system (Gentra Systems, Minneapolis, USA).

Transcriptome Assembly and Coverage. The Transcriptome Shotgun reads have been deposited at Sequence Read Archive (SRA) under BioProject PRJNA419119. Only high quality reads were used. Before assembly, adaptor sequences were removed from raw reads. FASTQC reports were then generated and based on this information the resulting reads were trimmed at both ends. Nucleotide positions with a score lower than 20 were also masked (replaced by an N). These analyses were performed using the FASTQ tools implemented in the Galaxy platform[75,76]. The total number of reads for each transcriptome is presented in Supplementary Table S1. To assess the changing rate of new gene detection as a function of sequencing sampling for the nine anthers transcriptomes here obtained (Supplementary Fig. S1A), plus the nine Golden delicious tissues analysed (Supplementary Fig. S1B), we have obtained an accumulation curve by dividing the reads in sets of one million paired reads and looking for the number of *M. domestica* CDS, retrieved from the *M. domestica* RefSeq at NCBI, that show evidence for expression. Blastn search using as query the 33 *SFBB* sequences previously identified for *S3*-, *S9*- and *S10*- haplotypes[23,27,44] and identities higher than 90% revealed 10 *SFBBs* in the *M. domestica* RefSeq (Supplementary Table S2). FPKM values in these 18 transcriptomes were estimated using Express as implemented in Trinity (default parameters)[77]. The reads were then used in the transcriptome assembly using Trinity (default parameters)[77] and also using Edena[78] with the following K-mer values 20, 25, 30, 35, 40, 45, 50, 55 and 60. Assembly statistics for both assemblies were obtained with ABySS 2.0[79] (Supplementary Table S3 and Supplementary Table S4). The resulting files were merged and contigs that have a 100% match along the full sequence with larger contigs were eliminated. All contigs were used as subject for tblastn searches using local blast[80], and the *SFBB9* (*MdSFBB3-Beta*; AB270796) sequence as query, and an expect value of 0.05. It should be noted that when using such parameters we also obtained sequences that show high identity (more than 97% identity over more than 100 bp) with previously reported *SFBB*-like genes. Therefore, it is unlikely that using this methodology we missed any *SFBB* gene. Nevertheless, not all alleles of each *SFBB* gene were obtained here,when the selected contigs were used as the query to perform a local blastn search[80], against a database of 33 *SFBB* sequences from *S3*-, *S9*-, and *S10*-haplotypes[23,27,44] (see Results).

Amplification of *SFBB* genes. Genomic DNA of each of the nine *M. domestica* cultivars was used as template in PCRs using primers SFBBgenF and SFBBgenR[22]. Standard amplification conditions were 35 cycles of denaturation at 94 °C for 30 seconds, primer annealing at 48 °C for 30 s, and primer extension at 72 °C for 2 min. The amplification products were cloned using the TA cloning kit (Invitrogen, Carlsbad, CA). For each cultivar and amplification product, the insert of an average of 60 colonies was cut separately with *Rsa*I, *Alu*I, *Ava*II and *Sau*3AI restriction enzymes. For each cultivar and restriction pattern two colonies were sequenced. The ABI PRISM BigDye cycle-sequencing kit (Perkin Elmer, Foster City, CA), and specific primers, or the primers for the M13 forward and reverse priming sites of the pCR2.1 vector, were used to prepare the sequencing reactions. Sequencing runs were performed by STABVIDA (Lisboa, Portugal). Local blastn was performed using theses sequences as query and *Sorbus SFBB* genes[22] as subject. Sequences with homology higher than 95% were grouped as alleles of a particular gene.

Sequences were then aligned, using clustalW as implemented in Mega7[81] to identify conserved regions in all sequences for a given gene but that are different in other *SFBB* sequences. These regions were used to design specific primers for *SFBB* genes (Supplementary Table S6). These primers were used to amplify *SFBB* alleles from genomic DNA of *M. domestica* cultivars for which allele sequences were not obtained with SFBBgenF and SFBBgenR primers. Amplification conditions are described in Supplementary Table S6. The amplification products were cloned as described above. For each cultivar and amplification product, the insert of an average of 20 colonies was cut separately with *Rsa*I, *Alu*I, *Ava*II and *Sau*3AI restriction enzymes. The colonies that show a different restriction pattern from that of the alleles obtained with SFBBgenF and SFBBgenR primers were selected for sequencing. For each pattern three colonies were sequenced, as described above.

Genotyping and linkage analyses between 15 *SFBB* genes and nine *S-RNases*. For 17 out of the 18 *SFBB* genes present in more than one *M. domestica* cultivar, we were able to infer the allele that goes with a particular *S-RNase* (see Results). To show that these 17 *SFBB* genes are located in the *S*-locus region, we used 48, 34, and 27 individuals of the crosses 'McIntosh' (*S10, S25*) × 'Gala' (*S2, S5*), 'Fuji' (*S1, S9*) × 'Honeycrisp' (*S2, S24*), and 'Golden Delicious' (*S2, S3*) × 'Red Delicious' (*S9, S28*), respectively, that were genotyped for *S-RNase* alleles, using specific primers (Supplementary Table S9). For *SFBB* genes that present synonymous nucleotide diversity higher than 0.01 (all except *SFBB1*, *SFBB15*, and *SFBB18*; see Results) specific primers (Supplementary Table S6) were used to amplify these genes from genomic DNA of the 108 individuals analyzed from the three controlled crosses. For each *SFBB* gene, alleles present in the parents were used to select RFLPs that could be used to identify each of the *SFBB* alleles (Supplementary Table S8). It should be noted that, it is often not possible to develop a diagnostic marker for all four alleles segregating in a particular cross, since alleles of these genes have low levels of diversity.

Phylogenetic analyses, summary statistics, recombination, and testing for positive selection at the *M. domestica SFBB* genes. *SFBB* sequences were deposited in GenBank (accession numbers MG458438-MG458668). These *SFBB* sequences together with those reported for *S3*-, and *S9*- and *S10*-haplotypes[23,27,44], and the *SFBB* -lineage gene MDP0000250455 (not located in the *S*-locus, and not involved in GSI)[10], used to root the phylogenetic tree, were aligned with Clustal Omega[82]. Maximum-likelihood trees were obtained with FastTree2[83], using the general time reversible model with a proportion of invariant sites. A "CAT" rate for each site from among 20 fixed possibilities is first computed and then the lengths rescaled to optimize the gamma20 likelihood.

Analyses of DNA polymorphism, and minimum number of recombination events were performed using DnaSP v5[84]. The number of independent recombination events was inferred by RDP[85] using the RDP, Chimaera, BootScan, 3Seq, GeneConv, MaxChi and SiScan methods (default options). A sequence is taken as recombinant if at least one of the methods identifies a recombination tract in that sequence with a probability smaller than 0.05. For each *SFBB* gene, the total number of synonymous mutations implied by the data was inferred using Yang's[57] methodology, under the appropriate model (M0 or M2; see Results), in ADOPS[86].

For the identification of sites under positive selection we have used ADOPS[86] and 10 datasets corresponding to *SFBB*s in each of the *S*-haplotypes here analyzed. Sequences were first aligned with the ClustalW2, and Muscle alignment algorithms as implemented in ADOPS[86]. Only codons with a support value above two are used for phylogenetic reconstruction. Bayesian trees were obtained using MrBayes, as implemented in the ADOPS pipeline[86], using the GTR model of sequence evolution, allowing for among-site rate variation and a proportion of invariable sites. Third codon positions were allowed to have a gamma distribution shape parameter different from that of first and second codon positions. Two independent runs of 1,000,000 generations with four chains each (one cold and three heated chains) were set up. The average standard deviation of split frequencies was always about 0.01 and the potential scale reduction factor for every parameter about 1.00 showing that convergence has been achieved. Trees were sampled every 100th generation and the first 2500 samples were discarded (burn-in). The remaining trees were used to compute the Bayesian posterior probabilities of each clade of the consensus tree (see the B + database (bpositive.i3s.up.pt; see the Malus SFBB BP2017000011 dataset).We compare M2-M1 and M8-M7 models using codeML as implemented in ADOPS[86]. We consider as positively selected those amino acid sites that show a probability higher than 95% for both naive empirical Bayes (NEB) or Bayes empirical Bayes (BEB) methods in at least one of the analyses.

References

1. Igic, B., Lande, R. & Kohn, J. R. Loss of self-incompatibility and its evolutionary consequences. *Int J Plant Sci* **169**, 93–104 (2008).
2. De Nettancourt, D. *Incompatibility in angiosperms*. (Springer-Verlag, Berlin, 1977).
3. Roalson, E. H. & McCubbin, A. G. S-RNases and sexual incompatibility: structure, functions, and evolutionary perspectives. *Mol Phylogenet Evol* **29**, 490–506, https://doi.org/10.1016/S1055-7903(03)00195-7 (2003).
4. McClure, B. Darwin's foundation for investigating self-incompatibility and the progress toward a physiological model for S-RNase-based SI. *J Exp Bot* **60**, 1069–1081, https://doi.org/10.1093/jxb/erp024 (2009).
5. Nowak, M. D., Davis, A. P., Anthony, F. & Yoder, A. D. Expression and trans-specific polymorphism of self-incompatibility RNases in *Coffea* (Rubiaceae). *PLoS One* **6**, e21019, https://doi.org/10.1371/journal.pone.002101910.1371/journal.pone.0021019 (2011).
6. Huang, S., Lee, H. S., Karunanandaa, B. & Kao, T. H. Ribonuclease activity of *Petunia inflata S* proteins is essential for rejection of self-pollen. *The Plant Cell* **6**, 1021–1028 (1994).
7. Igic, B. & Kohn, J. R. Evolutionary relationships among self-incompatibility RNases. *Proc Natl Acad Sci USA* **98**, 13167–13171 (2001).
8. Steinbachs, J. E. & Holsinger, K. E. *S*-RNase-mediated gametophytic self-incompatibility is ancestral in eudicots. *Mol Biol Evol* **19**, 825–829 (2002).
9. Vieira, J., Fonseca, N. A. & Vieira, C. P. An *S-RNase*-based gametophytic self-incompatibility system evolved only once in eudicots. *J Mol Evol* **67**, 179–190, https://doi.org/10.1007/s00239-008-9137-x (2008).
10. Aguiar, B. *et al*. Convergent evolution at the gametophytic self-incompatibility system in *Malus* and *Prunus*. *PloS one* **10**, e0126138, https://doi.org/10.1371/journal.pone.0126138 (2015).

11. Vieira, J., Santos, R. A., Ferreira, S. M. & Vieira, C. P. Inferences on the number and frequency of S-pollen gene (SFB) specificities in the polyploid *Prunus spinosa*. *Heredity* **101**, 351–358, https://doi.org/10.1038/hdy.2008.60 (2008).
12. Nunes, M. D., Santos, R. A., Ferreira, S. M., Vieira, J. & Vieira, C. P. Variability patterns and positively selected sites at the gametophytic self-incompatibility pollen SFB gene in a wild self-incompatible *Prunus spinosa* (Rosaceae) population. *New Phytol* **172**, 577–587, https://doi.org/10.1111/j.1469-8137.2006.01838.x (2006).
13. Sonneveld, T., Tobutt, K. R., Vaughan, S. P. & Robbins, T. P. Loss of pollen-*S* function in two self-compatible selections of *Prunus avium* is associated with deletion/mutation of an *S* haplotype-specific F-box gene. *Plant Cell* **17**, 37–51, https://doi.org/10.1105/tpc.104.026963 (2005).
14. Ikeda, K. *et al*. Primary structural features of the *S* haplotype-specific F-box protein, SFB, in *Prunus*. *Sex Plant Reprod* **16**, 235–243, https://doi.org/10.1007/s00497-003-0200-x (2004).
15. Entani, T. *et al*. Comparative analysis of the self-incompatibility (*S*-) locus region of *Prunus mume*: identification of a pollen-expressed F-box gene with allelic diversity. *Genes Cells* **8**, 203–213 (2003).
16. Ushijima, K. *et al*. Characterization of the S-locus region of almond (*Prunus dulcis*): analysis of a somaclonal mutant and a cosmid contig for an S haplotype. *Genetics* **158**, 379–386 (2001).
17. Ushijima, K. *et al*. Structural and transcriptional analysis of the self-incompatibility locus of almond: identification of a pollen-expressed F-box gene with haplotype-specific polymorphism. *Plant Cell* **15**, 771–781 (2003).
18. Romero, C. *et al*. Analysis of the *S*-locus structure in *Prunus armeniaca* L. Identification of S-haplotype specific S-RNase and F-box genes. *Plant Mol Biol* **56**, 145–157, https://doi.org/10.1007/s11103-004-2651-3 (2004).
19. 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. *Frontiers in plant science* **6**, 41, https://doi.org/10.3389/fpls.2015.00041 (2015).
20. Kubo, K. *et al*. Gene duplication and genetic exchange drive the evolution of *S*-RNase-based self-incompatibility in *Petunia*. *Nature Plants* **1**, 14005, https://doi.org/10.1038/nplants.2014.5 (2015).
21. Williams, J. S., Der, J. P. & Kao, T. H. Transcriptome analysis reveals the same 17 *S-Locus F-Box* genes in two haplotypes of the self-incompatibility locus of *Petunia inflata*. *The Plant Cell Online* **26**, 2873–2888, https://doi.org/10.1105/tpc.114.126920 (2014).
22. Aguiar, B. *et al*. Patterns of evolution at the gametophytic self-incompatibility *Sorbus aucuparia* (Pyrinae) *S* pollen genes support the non-self recognition by multiple factors model. *J Exp Bot* **64**, 2423–2434, https://doi.org/10.1093/jxb/ert098 (2013).
23. 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, https://doi.org/10.1007/s11103-010-9662-z (2010).
24. Kubo, K. *et al*. Collaborative non-self recognition system in S-RNase-based self-incompatibility. *Science* **330**, 796–799, https://doi.org/10.1126/science.1195243 (2010).
25. Cheng, J., Han, Z., Xu, X. & Li, T. Isolation and identification of the pollen-expressed polymorphic F-box genes linked to the *S*-locus in apple (*Malus* × *domestica*). *Sex Plant Reprod* **19**, 175–183, https://doi.org/10.1007/s00497-006-0034-4 (2006).
26. Kakui, H., Tsuzuki, T., Koba, T. & Sassa, H. Polymorphism of SFBB-gamma and its use for *S* genotyping in Japanese pear (*Pyrus pyrifolia*). *Plant Cell Rep* **26**, 1619–1625, https://doi.org/10.1007/s00299-007-0386-8 (2007).
27. 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, https://doi.org/10.1534/genetics.106.068858 (2007).
28. Wheeler, D. & Newbigin, E. Expression of 10 S-class *SLF-like* genes in *Nicotiana alata* pollen and its implications for understanding the pollen factor of the S locus. *Genetics* **177**, 2171–2180, https://doi.org/10.1534/genetics.107.076885 (2007).
29. Ushijima, K. *et al*. The *S* haplotype-specific F-box protein gene, *SFB*, is defective in self-compatible haplotypes of *Prunus avium* and *P. mume*. *Plant J* **39**, 573–586, https://doi.org/10.1111/j.1365-313X.2004.02154.x (2004).
30. Wang, L. *et al*. Genome-wide analysis of *S*-Locus F-box-like genes in *Arabidopsis thaliana*. *Plant Mol Biol* **56**, 929–945, https://doi.org/10.1007/s11103-004-6236-y (2004).
31. Vieira, J., Fonseca, N. A. & Vieira, C. P. RNase-based gametophytic self-incompatibility evolution: Questioning the hypothesis of multiple independent recruitments of the S-pollen gene. *J Mol Evol* **69**, 32–41, https://doi.org/10.1007/s00239-009-9249-y (2009).
32. Luu, D.-T. *et al*. Rejection of *S*-heteroallelic pollen by a dual-specific S-RNase in *Solanum chacoense* predicts a multimeric SI pollen xomponent. *Genetics* **159**, 329–335 (2001).
33. Matsumoto, D. & Tao, R. Distinct self-recognition in the *Prunus S-RNase*-based gametophytic self-incompatibility system. *The Horticulture Journal*, https://doi.org/10.2503/hortj.MI-IR06 (2016).
34. Kakui, H. *et al*. Sequence divergence and loss-of-function phenotypes of *S locus F-box* brothers genes are consistent with non-self recognition by multiple pollen determinants in self-incompatibility of Japanese pear (*Pyrus pyrifolia*). *Plant J* **68**, 1028–1038, https://doi.org/10.1111/j.1365-313X.2011.04752.x (2011).
35. Sun, P., Li, S., Lu, D., Williams, J. S. & Kao, T. H. Pollen *S*–locus F–box proteins of *Petunia* involved in S–RNase based self-incompatibility are themselves subject to ubiquitin- mediated degradation. *The Plant J* **83**, 213–223, https://doi.org/10.1111/tpj.12880 (2015).
36. Li, J. *et al*. Electrostatic potentials of the *S*-locus F-box proteins contribute to the pollen *S* specificity in self-incompatibility in *Petunia hybrida*. *The Plant J* **89**, 45–57, https://doi.org/10.1111/tpj.13318 (2016).
37. Sijacic, P. *et al*. Identification of the pollen determinant of S-RNase-mediated self-incompatibility. *Nature* **429**, 302–305, https://doi.org/10.1038/nature02523 (2004).
38. Brewbaker, J. t. & Natarajan, A. Centric fragments and pollen-part mutation of incompatibility alleles in Petunia. *Genetics* **45**, 699 (1960).
39. Entani, T. *et al*. Relationship between polyploidy and pollen self-incompatibility phenotype in Petunia hybrida Vilm. *Biosci Biotechnol Biochem* **63**, 1882–1888 (1999).
40. Golz, J. F., Oh, H. Y., Su, V., Kusaba, M. & Newbigin, E. Genetic analysis of *Nicotiana* pollen-part mutants is consistent with the presence of an S-ribonuclease inhibitor at the S locus. *Proc Natl Acad Sci USA* **98**, 15372–15376, https://doi.org/10.1073/pnas.261571598 (2001).
41. Yuan, H. *et al*. A novel gene, MdSSK1, as a component of the SCF complex rather than MdSBP1 can mediate the ubiquitination of S-RNase in apple. *J Exp Bot* **65**, 3121–3131, https://doi.org/10.1093/jxb/eru164 (2014).
42. Ashkani, J. & Rees, D. A comprehensive study of molecular evolution at the self-incompatibility locus of rosaceae. *J Mol Evol* **82**, 128–145, https://doi.org/10.1007/s00239-015-9726-4 (2016).
43. Tsukamoto, T. *et al*. Genetic and molecular characterization of three novel S-haplotypes in sour cherry (*Prunus cerasus* L.). *J Exp Bot* **59**, 3169–3185, https://doi.org/10.1093/jxb/ern172 (2008).
44. 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 reproduction* **26**, 101–111, https://doi.org/10.1007/s00497-013-0212-0 (2013).
45. McCubbin, A. G., Wang, X. & Kao, T. H. Identification of self-incompatibility (S-) locus linked pollen cDNA markers in *Petunia inflata*. *Genome* **43**, 619–627 (2000).
46. Wang, Y., Wang, X., McCubbin, A. G. & Kao, T. H. Genetic mapping and molecular characterization of the self-incompatibility (S) locus in *Petunia inflata*. *Plant Mol Biol* **53**, 565–580 (2003).
47. Hua, Z., Meng, X. & Kao, T. H. Comparison of *Petunia inflata* S-Locus F-box protein (Pi SLF) with Pi SLF like proteins reveals its unique function in S-RNase based self-incompatibility. *Plant Cell* **19**, 3593–3609, https://doi.org/10.1105/tpc.107.055426 (2007).

48. De Franceschi, P. *et al*. Evaluation of candidate F-box genes for the pollen *S* of gametophytic self-incompatibility in the Pyrinae (Rosaceae) on the basis of their phylogenomic context. *Tree Genet Genomes* **7**, 663–683, https://doi.org/10.1007/s11295-011-0365-7 (2011).
49. Okada, K. *et al*. Related polymorphic F-box protein genes between haplotypes clustering in the BAC contig sequences around the *S-RNase* of Japanese pear. *J Exp Bot* **62**, 1887–1902, https://doi.org/10.1093/jxb/erq381 (2011).
50. Velasco, R. *et al*. The genome of the domesticated apple (*Malus* x *domestica* Borkh.). *Nat Genet* **42**, 833–839, https://doi.org/10.1038/ng.654 (2010).
51. Vieira, C. & Charlesworth, D. Molecular variation at the self-incompatibility locus in natural populations of the genera *Antirrhinum* and *Misopates*. *Heredity* **88**, 172–181 (2002).
52. Wright, S. The Distribution of Self-Sterility Alleles in Populations. *Genetics* **24**, 538–552 (1939).
53. Clark, A. Evolutionary inferences from molecular characterization of self-incompatibility alleles. *Mechanisms of Molecular Evolution*, 79–108 (1993).
54. Richman, A. D., Uyenoyama, M. K. & Kohn, J. R. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. *Science* **273**, 1212 (1996).
55. Richman, A. D. & Kohn, J. R. In *Plant Molecular Evolution* 169–179 (Springer, 2000).
56. Posada, D. & Crandall, K. A. Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution* **16**, 37–45 (2001).
57. Yang, Z. H. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS* **13**, 555–556 (1997).
58. Vázquez, N. *et al*. *in 11th International Conference on Practical Applications of Computational Biology & Bioinformatics*. 18 (Springer) (2017).
59. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc Natl Acad Sci* **93**, 13–20 (1996).
60. Sudha, G., Nussinov, R. & Srinivasan, N. An overview of recent advances in structural bioinformatics of protein–protein interactions and a guide to their principles. *Progress in biophysics and molecular biology* **116**, 141–150 (2014).
61. 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, https://doi.org/10.1186/1471-2148-10-200 (2010).
62. Sun, P. & Kao, T. H. Self-incompatibility in *Petunia inflata*: the relationship between a self-incompatibility locus F-box protein and its non-self S-RNases. *The Plant Cell* **25**, 470–485 (2013).
63. Boccara, M. *et al*. The *Arabidopsis* miR472-RDR6 silencing pathway modulates PAMP-and effector-triggered immunity through the post-transcriptional control of disease resistance genes. *PLoS Pathog* **10**, e1003883 (2014).
64. Kato, H., Shida, T., Komeda, Y., Saito, T. & Kato, A. Overexpression of the activated disease resistance 1-like1 (*ADR1-L1*) gene results in a dwarf phenotype and activation of defense-related gene expression in *Arabidopsis thaliana*. *Journal of Plant Biology* **54**, 172–179 (2011).
65. Aboul-Soud, M. A. *et al*. Activation tagging of ADR2 conveys a spreading lesion phenotype and resistance to biotrophic pathogens. *New Phytol* **183**, 1163–1175, https://doi.org/10.1111/j.1469-8137.2009.02902.x (2009).
66. Grant, J. J., Chini, A., Basu, D. & Loake, G. J. Targeted activation tagging of the *Arabidopsis* NBS-LRR gene, *ADR1*, conveys resistance to virulent pathogens. *Molecular Plant-Microbe Interactions* **16**, 669–680 (2003).
67. Mindrinos, M., Katagiri, F., Yu, G.-L. & Ausubel, F. M. The *A. thaliana* disease resistance gene *RPS2* encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell* **78**, 1089–1099 (1994).
68. Shah, J., Kachroo, P. & Klessig, D. F. The *Arabidopsisssi1* mutation restores pathogenesis-related gene expression in *npr*1 plants and renders defensin gene expression salicylic acid dependent. *The Plant Cell* **11**, 191–206 (1999).
69. Shirano, Y., Kachroo, P., Shah, J. & Klessig, D. F. A gain-of-function mutation in an *Arabidopsis* Toll Interleukin1 Receptor–Nucleotide Binding Site–Leucine-Rich Repeat type R gene triggers defense responses and results in enhanced disease resistance. *The Plant Cell* **14**, 3149–3162 (2002).
70. Stokes, T. L., Kunkel, B. N. & Richards, E. J. Epigenetic variation in *Arabidopsis disease resistance*. *Genes & Development* **16**, 171–182 (2002).
71. Xiao, S., Brown, S., Patrick, E., Brearley, C. & Turner, J. G. Enhanced transcription of the *Arabidopsis* disease resistance genes*rpw8. 1* and *rpw8. 2* via a salicylic acid–dependent amplification circuit is required for hypersensitive cell death. *The Plant Cell* **15**, 33–45 (2003).
72. Igari, K. *et al*. Constitutive activation of a CC-NB-LRR protein alters morphogenesis through the cytokinin pathway in *Arabidopsis*. *The Plant J* **55**, 14–27 (2008).
73. Palma, K. *et al*. Autoimmunity in *Arabidopsisacd11* is mediated by epigenetic regulation of an immune receptor. *PLoS Pathog* **6**, e1001137, https://doi.org/10.1371/journal.ppat.1001137 (2010).
74. Ingram, G. C. *et al*. Dual role for fimbriata in regulating floral homeotic genes and cell division in *Antirrhinum*. *Embo J* **16**, 6521–6534 (1997).
75. Blankenberg, D. *et al*. Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, 19.10.01-19.10.21, https://doi.org/10.1002/0471142727.mb1910s89 (2010).
76. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86, https://doi.org/10.1186/gb-2010-11-8-r86 (2010).
77. Haas, B. J. *et al*. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494–1512, https://doi.org/10.1038/nprot.2013.084 (2013).
78. Hernandez, D., François, P., Farinelli, L., Østerås, M. & Schrenzel, J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* **18**, 802–809, https://doi.org/10.1101/gr.072033.107 (2008).
79. Jackman, S. D. *et al*. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res* **27**, 768–777, https://doi.org/10.1101/gr.214346.116 (2017).
80. Johnson, M. *et al*. NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**, W5–W9 (2008).
81. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**, 1870–1874, https://doi.org/10.1093/molbev/msw054 (2016).
82. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, https://doi.org/10.1038/msb.2011.75 (2011).
83. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490, https://doi.org/10.1371/journal.pone.0009490 (2010).
84. Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497, https://doi.org/10.1093/bioinformatics/btg359 (2003).
85. Martin, D. P., Williamson, C. & Posada, D. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**, 260–262 (2005).
86. Reboiro-Jato, D. *et al*. ADOPS - Automatic Detection Of Positively Selected Sites. *J Integr Bioinform* **9**, 200, https://doi.org/10.2390/biecoll-jib-2012-200 (2012).
87. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).

Acknowledgements
This work was financed by the project Norte-01–0145-FEDER-000008 -Porto Neurosciences and Neurologic Disease Research Initiative at I3S, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (FEDER).

Author Contributions
J.V., A.I., S.van N., N.A.F., and C.P.V. designed the research. A.I., and S.van N., collected the plant material. S.vanN. obtained the transcriptomes. A.I. performed the controlled crosses. M.I.P., B.A., V.N., V.T. and C.P.V. performed the molecular work. J.V., and C.P.V. processed the data and performed the analyses. All authors reviewed and approved the manuscript.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-19820-1.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.