Variation among S-locus haplotypes and among stylar RNases in almond

Shashi N. Goonetilleke¹, Adam E. Croxford³, Timothy J. March³, Michelle G. Wirthensohn¹, Maria Hrmova¹,² & Diane E. Mather¹*

In many plant species, self-incompatibility systems limit self-pollination and mating among relatives. This helps maintain genetic diversity in natural populations but imposes constraints in agriculture and plant breeding. In almond [Prunus dulcis (Mill.) D.A. Webb], the specificity of self-incompatibility is mainly determined by stylar ribonuclease (S-RNase) and S-haplotype-specific F-box (SFB) proteins, both encoded within a complex locus, S. Prior to this research, a nearly complete sequence was available for one S-locus haplotype. Here, we report complete sequences for four haplotypes and partial sequences for 11 haplotypes. Haplotypes vary in sequences of genes (particularly S-RNase and SFB), distances between genes and numbers and positions of long terminal repeat transposons. Haplotype variation outside of the S-RNase and SFB genes may help maintain functionally important associations between S-RNase and SFB alleles. Fluorescence-based assays were developed to distinguish among some S-RNase alleles. With three-dimensional modelling of five S-RNase proteins, conserved active sites were identified and variation was observed in electrostatic potential and in the numbers, characteristics and positions of secondary structural elements, loop anchoring points and glycosylation sites. A hypervariable region on the protein surface and differences in the number, location and types of glycosylation sites may contribute to determining S-RNase specificity.

Many plant species, including almond [Prunus dulcis (Mill.) D.A. Webb] and some other important tree crops, exhibit self-incompatibility (SI); they are unable to set seed from self-pollination or from pollination by genetically identical or genetically similar plants. While biologically important as a means of maintaining population diversity and avoiding inbreeding, self-incompatibility imposes constraints on agricultural and horticultural practices (requiring polliniser varieties) and in plant breeding (restricting the choice of cross combinations). Self-incompatibility can be sporophytic, involving recognition of the genotype of the pollen parent, or gametophytic, involving recognition of the pollen genotype. In sporophytic SI, incompatibility reactions prevent the germination of incompatible pollen grains on the stigma. In gametophytic SI, incompatibility reactions impede the growth of incompatible pollen tubes through the style.

In Prunus, including almond, SI is gametophytic and under the genetic control of complex and highly variable S loci. Based on the results of experimental crosses, there are thought to be at least 50 variants at the almond S-locus⁴⁻⁵. Sequencing of a 71,953 bp region of one haplotype (S₇, also known as S₇) showed that the almond S locus includes S-locus F-box (SLF), stylar RNase (S-RNase) and S-haplotype-specific F-box (SFB) genes, other open reading frames and pairs of long-terminal-repeat retrotransposons (LTRs)⁶. For this complex locus, variant forms of individual genes are referred to as alleles, while variant forms of the entire locus are referred to as haplotypes. Although the S₇ haplotype is the only one for which a nearly complete sequence has been published, the SLF, S-RNase and/or SFB alleles of some other haplotypes have been fully or partially sequenced⁴⁻⁷. Among haplotypes that have been physically mapped, the order and orientations of S-locus features are conserved, but the distances between these features vary⁶.

In Prunus, including almond, the specificity of SI is mainly determined by the S-RNase and SFB genes⁶⁻⁷, which are expressed in pistils and pollen tubes, respectively. S-RNases act as cytotoxins in self pollen tubes⁶ but the role of SFB proteins is not completely understood. In sweet cherry (P. avium), SFB proteins act as ‘blockers’ that protect self S-RNases from detoxification by SFB-like and SLF-like ‘general inhibitor’ proteins⁹⁻¹⁰. This allows
self S-RNases to remain active and capable of arresting pollen tube growth. It is not known which, if any, F-box proteins play general inhibitor roles in the SI system of almond.

A few almond cultivars are self-fertile. This phenotype has been attributed to a dominant S-RNase allele, designated Sf inactive). Plants carrying this allele do not express an active Sf-RNase, possibly due to poor transcription and are not able to block the growth of Sf-pollen tubes. A similar allele, Sf (Sf-active) expresses an active Sf-RNase and confers SI. Despite their contrasting phenotypes, Sf- and Sf-RNase alleles have identical nucleotide sequences and are linked with identical SFB alleles. This apparent paradox was resolved by the discovery that Sf and Sf are epialleles, differing by the methylation of a single nucleotide upstream of the coding sequence. This epigenetic difference may determine whether the Sf-RNase allele is expressed. Consistent with this interpretation, Sf,Sf heterozygotes have been found to be fully self-incompatible, with their Sf-RNase able to block the growth of both Sf and Sf pollen tubes.

To further investigate variation among S-locus haplotypes, we amplified and sequenced PCR products from 48 diverse almond clones. To improve the efficacy of S allele detection, we developed simple fluorescence-based marker assays to distinguish among S-RNase alleles. To investigate how structural features might affect S-RNase function and specificity, we conducted three-dimensional (3D) protein modelling for the predicted products of five S-RNase alleles and investigated how sequence variation in a highly variable region could affect domain structure, glycosylation and physical interacting forces that might influence the specificity of SI.

Results

S-locus haplotype sequences. From Illumina paired-end sequence data generated for a pooled library of products amplified from the S locus of diverse almond clones (Supplementary Table S1), sequences from each of eight clones known to carry the S-locus haplotypic sequence. Among these haplotypes, there were sequence discrepancies at just eight of 71,953 nucleotide positions. Haplotype sequences from the cultivars Keanes (MH029539) and Capella (MH029540) were complete and were identical to each other. Their sequence was selected as the consensus sequence for S7 haplotype sequences. Among the four fully-sequenced haplotypes (S1, S4, S7, S8), positional sequence identity ranges from 51 to 84% (Supplementary Table S3).

Using the consensus S7 sequence as a new reference for S7, haplotype sequences were obtained for: S7, from Brown Nonpareil (S7,S7); S7, from Nonpareil and McKinlay’s (both S7,S7); and S7, from the self-fertile clones Carina, Mira, Capella, T5 and T7 (all S7,S7). There is just 2% inconsistency between the two S7 sequences and 3% inconsistency among the five S7 sequences. Among the four fully-sequenced haplotypes (S7, S7, S7 and S7), positional sequence identity ranges from 47% (S7) to 99% (S7) of the length of the S7 haplotype sequence (Fig. 1).

Figure 1. S-locus structure. Structure of the almond S locus showing the positions of the SLF, S-RNase and SFB genes and long terminal repeat retrotransposons (LTRs). Black lines indicate regions for which sequences were obtained and grey lines indicate gaps in the sequence.
In each of the completely sequenced haplotypes, 12 or more open reading frames (ORFs) were detected: 12 in \( S_1 \), 12 in \( S_6 \), 14 in \( S_3 \) and 18 in \( S_5 \) (Supplementary Table S4; Supplementary Fig. S1). In each case, these include ORFs that correspond with the SLF, S-RNase and SFB genes. For some ORFs (including ten in the \( S_6 \) haplotype and the last ORF in each haplotype), no similarity to known function genes was detected. For others (five for \( S_6 \) and \( S_5 \) six for \( S_1 \) and eight for \( S_5 \)), homology with transposases from other \( P. \) \( truncatus \) species was detected. Most of these transposases belong to the \( Ty1-copia \) RNase family and contain a DDE motif.

For the interval between the S-RNase and SFB genes, complete sequences were obtained for 11 haplotypes (\( S_1 \), \( S_3 \), \( S_5 \), \( S_6 \), \( S_{22} \), \( S_{23} \), \( S_7 \) and \( S_9 \)). These sequences, which are AT-rich (65–70%), range in length from 1.2 \( (S_6) \) to 6.6 \( (S_1) \). The sequence identity among them ranges from 21% (between \( S_1 \) and \( S_3 \)) to 98% (between \( S_6 \) and \( S_9 \)) (Supplementary Table S5).

Pairs of LTRs were detected in each haplotype, mostly in positions that correspond to the LTR-containing region of the \( S_1 \) haplotype (Fig. 1). No LTRs were detected within the SLF, S-RNase or SFB genes but in the \( S_1 \) haplotype, an LTR pair was detected between the SLF and S-RNase genes. All of the LTRs detected here are \( Ty1-copia \)-like retrotransposons with TGG/CA boxes in their 5’ and 3’ ends. Their protein-binding sites are TyrGTA, IleAAT, MetCAT and AlaTGC.

**SLF, S-RNase and SFB allele sequences.** The SLF gene, which is about 1.2 kb long and has no introns, was sequenced for all 15 haplotypes. Pairwise sequence identities among SLF alleles are high, ranging from 70 to 98% (Supplementary Table S6). Among the predicted products of the 15 SLF alleles, 210 of 325 amino acid residues are absolutely conserved across all 15 alleles (Supplementary Fig. S2). Sequence comparisons with SLF-like proteins from sweet cherry showed that the predicted products of all 15 almond SLF alleles are most similar to PavSLFL1 (sequence identity between 82 and 99%, compared to between 55 and 61% for PavSLFL5 and no more than 36% for any other PavSLFL) (Supplementary Table S7). Comparisons of PavSLFL sequences with a pseudomolecule sequence for almond chromosome 6 revealed possible homologs of PavSLFLs near the pseudomolecule sequence for almond chromosome 6 revealed possible homologs of PavSLFLs near the S-F-box motif, ranging from 0.5 to 0.9.

The S-RNase gene, which was completely sequenced for 11 haplotypes (\( S_1 \), \( S_3 \), \( S_6 \), \( S_9 \), \( S_{13} \), \( S_{14} \), \( S_{15} \), \( S_{22} \), \( S_{23} \), \( S_7 \) and \( S_9 \)) and partially sequenced for four haplotypes (\( S_3 \), \( S_6 \) and \( S_9 \)), is much more variable. The completely sequenced S-RNase alleles range in length from 1.0 kb (\( S_6 \)) to 4.5 kb (\( S_3 \)). Their pairwise nucleotide sequence identities range from 19% (between \( S_1 \) and \( S_3 \)) to 51% (between \( S_1 \) and \( S_7 \)) (Supplementary Table S8). Differences among alleles include both sequence differences within exons and length polymorphisms within introns (especially intron 2). The deduced protein sequences of the completely sequenced S-RNase alleles contain previously reported conserved regions (C1, C2, C3, RC4 and C5) and variable regions (RHV, V1 and V2) (Fig. 2a). Three additional variable regions were identified: V3 between C1 and C2; V4 between C2 and RHV; and V5 between C3 and RC4 (Fig. 2a). The nonsynonymous-to-synonymous ratio (Ka/Ks) for these alleles is 0.60 (Ka = 0.15, Ks = 0.25), with most codon differences occurring within V1, V2 and RHV (Fig. 2b).

The SFB gene was completely sequenced for 11 haplotypes (\( S_1 \), \( S_3 \), \( S_6 \), \( S_9 \), \( S_{13} \), \( S_{14} \), \( S_{15} \), \( S_{22} \), \( S_{23} \), \( S_7 \) and \( S_9 \)) and partially sequenced for four haplotypes (\( S_1 \), \( S_3 \), \( S_6 \) and \( S_9 \)). The completely sequenced alleles range in length from 1.0 kb to 6.6 kb (\( S_6 \)). Their pairwise sequence identities range from 35% (between SFB2 and SFB3) to 86% (between SFB2 and SFB5) (Supplementary Table S9). The SFB gene has one intron, which is within its 5’ untranslated region and is less polymorphic than either of the S-RNase introns. In protein sequences deduced from complete SFB allele sequences, several previously reported features of the protein are evident: an F-box motif, two variable regions (V1 and V2) and two hypervariable regions (HVs and HVb) (Supplementary Fig. S3). Two additional short highly variable regions (V3 and V4) were detected, both between V1 and V2. Within the F-box motif, the SFB2, SFB15, SFB14 and SFB10 proteins each have an insertion of a single arginine, while SFB13 and SFB27 each have a deletion of eight amino acid residues. Within V1, many amino acid residues are conserved among the alleles examined here. An overall Ka/Ks ratio of 0.50 (Ka = 0.11, Ks = 0.22) was computed using the complete SFB gene sequences. Most of the non-synonymous changes are in the hypervariable regions HVa and HVb, within which Ka/Ks values range from 0.9 to 1.5 (Supplementary Fig. S3), but there is also considerable variation in Ka/Ks values in the F-box motif, ranging from 0.5 to 0.9.

**Marker assays to distinguish among S-RNase alleles.** To provide a presence-absence assay for the S-RNase allele, a primer pair (WriPdSf-1; Supplementary Table S10) was designed for an S5 specific site within intron 2 of the S-RNase gene (Supplementary Fig. S4). With this assay, fluorescence is detected for the S5 allele and no signal is detected for any of the other alleles. For example, when this assay was applied to Nonpareil (\( S_5S_5 \)) × Vairo (\( S_3S_3 \)) F1 progeny, HEX fluorescence was detected for \( S_5 \) and \( S_5 \) progeny and little or no fluorescence was detected for \( S_3 \) and \( S_3 \) progeny (Fig. 3a).

Four additional primer sets (Supplementary Table S10) were designed to query an A/C SNP that distinguishes \( S_5 (A) \) from each of the Sl alleles considered here (all C) (Supplementary Fig. S5). Each of these sets includes a primer in the conserved region C1 and two allele-specific primers that overlap with part of the conserved region C2.

In primer set WriPdSf-2, the FAM-tailed primer is exactly complementary to the S5 sequence throughout the annealing site. The HEX-tailed primer is exactly complementary to the \( S_3 \), \( S_6 \), \( S_9 \) and \( S_6 \) sequences throughout the annealing site but not to the \( S_5 \), \( S_5 \) and \( S_5 \) sequences. For \( S_9 \), the first mismatch is too far from the target SNP to prevent annealing and amplification. When the WriPdSf-2 primer set was applied to Chellaston (\( S_6S_6 \)) × Lauranne (\( S_3S_3 \)) F1 progeny, HEX fluorescence was detected from the HEX-HEX genotype \( S_6S_6 \) (half from each allele) and from the null-HEX genotype \( S_3S_3 \) (all from \( S_3 \)), FAM fluorescence was detected for the
null-FAM genotype Sf-Sf (all from Sf) and both HEX and FAM fluorescence were detected for the HEX-FAM genotype Sf-Sf (HEX from Sf and FAM from Sf) (Fig. 3b). As is normally expected for KASP markers, the total amount of HEX fluorescence for HEX-HEX genotypes was about the same as for null-HEX genotypes and about twice that for HEX-FAM genotypes. In summary, this primer set successfully discriminated Sf genotypes from all Sf genotypes and among some non-Sf genotypes.

The primer sets WriPdSf-3, WriPdSf-4 and WriPdSf-5 are similar to WriPdSf-2 but include degenerate allele specific primers, to accommodate polymorphisms other than the target SNP. For each of these primer sets, Sf fluorescence is detected and at least one SI allele for which little or no fluorescence is detected. Application of WriPdSf-2, WriPdSf-3, WriPdSf-4 and WriPdSf-5 to synthesised DNA representing the alleles Sf, Sf, Sf, Sf.
Figure 3. Marker assay results. Results obtained for (a) Nonpareil (S8S8) × Vairo (S5S5) F1 progeny using primer set WriPdSf-1, (b) Chellaston (S5S7) × Lauranne (S5S5) F1 progeny using primer set WriPdSf-2, (c) Carmel (S5S7) × 12–350 (S5S5) F1 progeny using primer set WriPdS1 and (d) Johnston’s Prolific (S8S8) × Lauranne (S5S5) F1 progeny using primer set WriPdS3. Data shown are intensities of FAM and HEX fluorescence, each normalised against fluorescence from an internal ROX reference.

S0, S9, S23 and S25 and to mixtures of synthetic DNA representing heterozygous combinations of these alleles confirmed that all four primer sets yield FAM fluorescence when S1 is present and HEX fluorescence when any of S0, S5, S8, S23 or S25 are present (Supplementary Fig. S6). In each case the amount of HEX fluorescence obtained for one allele with a mismatch (Sf) was very similar to that detected for the alleles with no mismatches (S1, S9, S23 and S25). In addition, WriPdSf-3 yields HEX fluorescence when S1 or Sf is present and WriPdSf-4 yields HEX fluorescence when S1 is present. When these markers were tested on synthetic DNA representing alleles for which neither HEX nor FAM fluorescence was expected, there was some HEX fluorescence detected (Supplementary Fig. F6), indicating that with an abundance of template DNA, some annealing of the HEX-tailed primer occurred despite mismatches in the annealing site. Nevertheless, these data points were well separated from those for which HEX fluorescence was expected. When these markers were tested on mapping populations segregating for null alleles, the results were exactly as expected (Supplementary Fig. S7): HEX fluorescence for null-HEX heterozygotes (SfS1 and SfS2; for WriPdSf-3; SfS5 for WriPdSf-4). FAM fluorescence for null-FAM heterozygotes (SfS5 for WriPdSf-2 and WriPdSf-3; SfS3 for WriPdSf-2, WriPdSf-4 and WriPdSf-5), both HEX and FAM fluorescence for HEX-FAM heterozygotes (SfS1 and SfS5 for WriPdSf-3 and SfS5 for WriPdSf-4) and a low levels of HEX fluorescence for null-null heterozygotes (SfS1 for WriPdSf-2 and WriPdSf-6; SfS5 for WriPdSf-2, WriPdSf-4 and WriPdSf-5).

To provide assays to distinguish among S-RNase alleles that confer SI, ten additional primer sets were designed. Seven of these (WriPdS1, WriPdS5, WriPdS7-2, WriPdS8, WriPdS9, WriPdS23 and WriPdS25-2) consist of just two primers each (Supplementary Fig. S8). They provide presence-absence assays with which FAM fluorescence is detected for the target allele (e.g. S1 in Fig. 3c) and little or no fluorescence is detected when the target allele is not present. The other three assays (WriPdS3, WriPdS7-1 and WriPdS25-1) consist of three primers each (Supplementary Fig. S9) and can distinguish FAM target alleles from HEX target allele(s) (e.g. S1 vs S25 in Fig. 3d).

Results from application of the 15 primer sets to a variety panel with known S genotypes and to the progeny of appropriate crosses are shown in Supplementary Fig. S10. In each set of results, there was significant variation (p < 0.001) among genotypic clusters defined based on FAM and HEX fluorescence intensities (Supplementary Tables S11 and S12). For members of the variety panel, there were no inconsistencies between prior genotypic information and the clusters to which they were assigned. Across a total of 3,417 progeny that were assigned to genotypic classes and for which observed genotypic ratios were compared to expected ratios using a chi-square test (χ², α = 0.05), there were no statistically significant deviations from the expectation that half of the progeny would carry the target S allele (Supplementary Table S13). In populations that were analysed with more than one marker, there were no inconsistencies in results among markers. On trees that had been genotyped as self-fertile and for which branches were bagged to exclude foreign pollen, fruits were consistently set on the bagged branches (Supplementary Table S14).
**3D models of S-RNase proteins.** PSI-BLAST searches yielded eight candidate templates (1J1G, 1J1F, 1BK7, 1UCA, 1UCC, 1UCD, 1UCG and 1V9H) for almond S-RNases, all with protein sequence identities above 35% and similarities of 47% or 48%. Among these candidate templates, 1J1G (for the MC1 RNase isolated from seeds of bitter gourd (Momordica charantia L.) was selected for generating 3D models for the S5-, S7-, S8-, S23- and Sf-RNases. The MC1 protein sequence has 38% identity and 48% similarity to the S-RNase and Sf-RNase protein sequences (E-value = 1.71e-27) and 36% identity and 47% similarity to the S-RNase (E-value = 1.72e-27), S-RNase (E-value = 1.69e-27) and Sf-RNase (E-value = 1.71e-27) protein sequences. The first 22 residues at the N-termini of the almond S-RNases could not be modelled because no suitable structural template was identified for this region. Modelling of the remaining protein sequences indicated that the folding topologies of the almond S-RNases are similar to those of the template protein and other T2 RNases. All of these RNases consist of α-helices and β-strands that are inter-connected by loops (Fig. 4) and can be classified in family d.124.1.1 of the SCOPe 2.06 database.

Among models generated based on alternative alignments, no differences were found in the locations of α-helices and β-sheets, DOPE values or MOF values. Based on Ramachandran plots, none of the residues of any 3D models are positioned in disallowed regions. All models can therefore be considered to have satisfactory stereo-chemical quality. The G factor values of the models range from 0.25 (S3-RNase) to 0.33 (S8-RNase), compared to 0.35 for the template structure (Supplementary Table S15). Analysis with ProSa 2003 indicated that the conformational energies of residues are in negative regions in all models. For these reasons, all structural models can be deemed to be correct.

The 1J1G template structure has eight α-helices and eight β-sheets, while each of the protein structures generated for almond S-RNases has seven α-helices and between five and seven β-sheets (Fig. 4; Supplementary Fig. S11). In all five S-RNases, the α-helices are located in approximately the same positions. They range in length from six to 15 residues, with α5 being the longest in all cases. Within the S5-, S7-, S8- and S23-RNases, four β-sheets (β3, β2, β1 and β0) are located in approximately the same positions. Three of these (β3, β2 and β1) form an antiparallel β-sheet that packs well with α-helices in the interior of the molecule. The overall molecular dimensions of each S-RNase are approximately 50 Å × 40 Å × 30 Å. The estimated solvent-accessible surface areas of the predicted almond S-RNases are 78% for the S5-, S7-, S8- and S23-RNases and 81% for the Sf-RNase. The percentages of the exposed surface occupied by positively charged residues are 22%, 25%, 25%, 23% and 19% for the S5-, S7-, S8-, S23- and Sf-RNases, respectively. Electropositive regions (RHV, V1, V2 and V4) of the S-RNases have high Ka/Ks ratios (ranging from 1.2 to 1.8). Those regions have higher average exposed surface (25%) than neutral and negatively charged regions (10%).

Each of the five S-RNase proteins modelled here has eight conserved cysteine residues. These residues are predicted to form four disulphide bridges (Fig. 2). These connect a region upstream of C1 with a region between C1 and V3; V4 with C3, C5 with a region downstream of V2 and two regions downstream of C5 with each other.

Comparisons with the S-RNase of Japanese pear enabled identification of putative active sites within almond S-RNases. These sites include conserved cysteine, histidine, glutamic acid, lysine and tryptophan residues, which are separated by distances ranging from 3.2 Å to 5.0 Å (dashed lines in Fig. 5).

In all five S-RNases, the variable regions V3, V4 and V5 and the hypervariable region RHV are exposed on the protein surface. In the S5-, S7-, and S3-RNases, V4 is highly positively charged (Fig. 4c). Considerable structural variation was detected in RHV, which consists of 18 residues in the S5-, S7-, S8- and S23-RNases but only 14 residues in the Sf-RNase. In the S5-, S7-, and S3-RNases, RHV consists of two α-helices (α0 and α1) and a short loop, while in the S5- and Sf-RNases, it has only one α-helix (α0) and a short loop. The residues in the RHV loop-anchoring points vary in size, charge, polarity and hydrophobicity. Notably, the last residue of each RHV loop-anchoring point is polar: glutamine in the S5-RNase, asparagine in the S7-RNase, tyrosine in the S8-RNase and serine in the S3- and Sf-RNases (Supplementary Table S16).

Among the five S-RNase proteins that were modelled here, the most obvious structural difference involves a loop located between the variable region V2 and the conserved region C5. In the S-RNase, this loop is much longer (30 residues) than in other S-RNase proteins (10 to 20 residues) (Fig. 4), with eight Sf-specific residues: a loop located between the variable region V2 and the conserved region C5. In the Sf-RNase, this loop is much longer (30 residues) than in other S-RNase proteins (10 to 20 residues) (Fig. 4), with eight Sf-specific residues:

- **Electropositive regions (RHV, V1, V2 and V4)** of the S-RNases have high Ka/Ks ratios (ranging from 1.2 to 1.8). Those regions have higher average exposed surface (25%) than neutral and negatively charged regions (10%).
- **Each of the five S-RNase proteins modelled here has eight conserved cysteine residues.** These residues are predicted to form four disulphide bridges (Fig. 2). These connect a region upstream of C1 with a region between C1 and V3; V4 with C3, C5 with a region downstream of V2 and two regions downstream of C5 with each other.
- **Comparisons with the S-RNase of Japanese pear enabled identification of putative active sites within almond S-RNases.** These sites include conserved cysteine, histidine, glutamic acid, lysine and tryptophan residues, which are separated by distances ranging from 3.2 Å to 5.0 Å (dashed lines in Fig. 5).

In all five S-RNases, the variable regions V3, V4 and V5 and the hypervariable region RHV are exposed on the protein surface. In the S5-, S7-, and S3-RNases, V4 is highly positively charged (Fig. 4c). Considerable structural variation was detected in RHV, which consists of 18 residues in the S5-, S7-, S8- and S23-RNases but only 14 residues in the Sf-RNase. In the S5-, S7-, and S3-RNases, RHV consists of two α-helices (α0 and α1) and a short loop, while in the S5- and Sf-RNases, it has only one α-helix (α0) and a short loop. The residues in the RHV loop-anchoring points vary in size, charge, polarity and hydrophobicity. Notably, the last residue of each RHV loop-anchoring point is polar: glutamine in the S5-RNase, asparagine in the S7-RNase, tyrosine in the S8-RNase and serine in the S3- and Sf-RNases (Supplementary Table S16).

Among the five S-RNase proteins that were modelled here, the most obvious structural difference involves a loop located between the variable region V2 and the conserved region C5. In the S-RNase, this loop is much longer (30 residues) than in other S-RNase proteins (10 to 20 residues) (Fig. 4), with eight Sf-specific residues: an isoleucine, an asparagine, a glycine, a phenylalanine, an alanine and three glutamates. The average exposed surface of the extended loop region of the Sf-RNase is 16%.

With in silico prediction of individual amino acid residues of the Sf-RNase by the corresponding residues present at the same positions in other almond S-RNases (Supplementary Fig. S12), it was possible to estimate relative changes in unfolding enthalpy values (∆∆G) due to specific mutations (Fig. 6; Supplementary Fig. S12). Most of the mutations were destabilising but a few were classified as highly stabilising (∆∆G < -1.84 kcal/mol). Most of the highly destabilising mutations (∆∆G > 1.84 kcal/mol) were at positions near the N- or C-termini.

Between four and seven N-glycosylation sites were identified in each of the S5-, S7-, S8- and Sf-RNases (Supplementary Fig. S11). Many of these sites were within loops, and none were within β-sheets. Two types of sequon were detected: Asn-Xaa-Thr and Asn-Xaa-Ser, where Xaa is any amino acid residue except proline. Only one of these (an Asn-Ile-Thr sequon in the RC4 region) is conserved among all five S-RNases. C-terminal regions had higher Asn-Xaa-Thr to Asn-Xaa-Ser ratios than N-terminal regions. Substitution of Asn residues within S-RNase sequons did not substantially affect electrostatic potential (Fig. 6), but substitution of Asn residues within some S-RNase sequons affected ∆∆G values. For example, replacement of Asn79 by Thr79 caused an energy loss of 0.18 kcal/mol and replacement of Asn151 by Asp151 caused an energy gain of 0.12 kcal/mol. In each of the S5-, S7- and Sf-RNases, one O-glycosylation site was detected in the V2 region (Supplementary Fig. S12). No O-glycosylation sites were detected in either the S5- or Sf-RNases.
Figure 4. Molecular properties of almond S-RNase structural models. (a) Stereo representation of the superposition of 3D structures of S-RNases, whereby the template crystal structure (the MC1 RNase from seeds of bitter gourd) is in pink and the 3D models of almond Sf-RNase, S5-RNase, S7-RNase, S8-RNase and S23-RNases are blue, green, grey, yellow and tint blue, respectively. Almond structural models were superposed on the template structure with RMSD values in the range of 0.15 Å to 0.19 Å for 179 Cα atoms. (b) Dispositions of secondary structure elements in the template, with the Sf-, S5-, S7-, S8- and S23-RNases indicated in pink, blue, green, grey, yellow and tint blue, respectively. (c) Molecular surface morphologies of the template structure, and almond Sf-, S5-, S7-, S8- and S23-RNase models coloured by electrostatic potentials display electroneutral (white), electropositive (blue, contoured at +5 kilotesla einstein⁻¹) and electronegative (red, contoured at −5 kilotesla einstein⁻¹) regions, and presented in the same orientations as the cartoons in panel.
Prunus, breeding, and horticulture, the can be applied at high throughput to large numbers of samples. Given the considerable advantages of self-fertility for some alleles, the fluorescence-based KASP assays developed here do not require gel electrophoresis. They distinguish among SI alleles can be used to design compatible crosses. Unlike assays that were previously available

![Image](https://example.com/image.png)

**Figure 5.** Stereo representation of the active site residues of the almond S7-RNase. Active site residues (His40, Trp43 in C2 region and Glu93, His97 in C3 region) are shown in cpk magenta on the background of the cartoon model (in grey). The positions of disulphide bridges (Cys20-Cys28, Cys56-Cys100, Cys161-Cys192, Cys176-Cys187) are shown in yellow. Distances among the active site residues are shown with dashed lines.

### Discussion

Given the complexity of the S locus, the extent of variation among haplotypes, the heterozygosity of the clones used here and the availability of just one reference sequence, it was difficult to obtain uniform DNA amplification across the entire S locus from all haplotypes and samples. Given that a high level of heterozygosity was expected, some primers were designed with degenerated 3’-end sequences. These primers tended to have low PCR sensitivity and a high degree of non-specific binding, and most of them were not selected for use to obtain amplicons for sequencing. Although SI enforces S-locus heterozygosity, some primer pairs seemed to yield only one product from some clones. This could be due to lack of length polymorphism between alleles (generating two products of equal length), sequence polymorphism at primer annealing sites (generating just one product) and/or preferential amplification of some products (generating predominantly one product). It was particularly difficult to obtain useful amplicons for the region in which the previously sequenced haplotype was known to contain LTRs.

The complete consensus S7 sequence obtained here based on data from eight clones provided minor improvements over the AB081587 sequence and provided new reference upon which an iterative process could be undertaken to assemble complete or partial sequences for other haplotypes. With these sequences, we were able to investigate sequence diversity throughout the locus.

Consistent with difficulties that were experienced in obtaining amplicons, the LTR-containing region was the least completely sequenced. In our S7 consensus sequence, one of four previously reported LTRs (LTR0) was not detected. This may have been due to the use of different reference sequences (Arabidopsis here, but rice in the previous work). In other haplotypes, between one and four LTR pairs were detected, almost all in approximately the same region of the locus in which LTRs were detected in the S7 haplotype. No LTRs were detected within the S-RNase or SFB genes. This is in contrast to Japanese apricot (Prunus mume Siebold & Zucc.), for which LTR insertions in the SFB gene have been reported to lead to breakdown of SI. While there is no evidence that S-locus LTRs are functionally relevant, they may be genetically relevant, with variation in their numbers, lengths and positions contributing to maintaining tight associations between S-RNase and SFB alleles.

All 12 ORFs that had previously been reported in the S7 haplotype were detected in our S7 consensus sequence, and up to 18 ORFs were detected in other haplotypes. These included the ORFs for the SLF, S-RNase and SFB genes and ORFs with high homology with known DDE RNase transposes from other Prunus species. The catalytic domains of DDE transposases are known to exhibit considerable sequence variability, possibly reflecting different ways of recognizing transposon DNA and leading to non-specific and/or weak DNA binding activity. The variable number of ORFs among S-locus haplotypes and differences in transposon DNA recognition and/or DNA binding ability may contribute to maintaining the specificity of SI in almond. In other Prunus species, insertion of transposable elements into S-RNase and SFB genes has been reported to lead to breakdown of SI, but no such insertions were observed here.

This work increased the number of sequenced SLF alleles from just two (SLF1 and SLF3) to fifteen. Consistent with the expectation that SLF does not affect SI specificity in Prunus spp., pairwise sequence identities among SLF alleles are high and the predicted SLF protein sequences are highly conserved. Based on its sequence similarity with the PavSLFL1 protein, SLF might be considered among the candidates for the general inhibitor role in SI interactions of almond. Further, there are candidate homologs for other PavSLFL-encoding genes near the almond S-locus; their roles are also worthy of investigation. As expected, we observed considerable sequence variation among alleles of the S-RNase and SFB genes, which are known to encode the determinants of pistil-pollen specificity. The S-RNase alleles also vary considerably in length, mainly because of polymorphisms in the second intron.

With analysis of S-RNase allele sequences, it was possible to design new marker assays for use in almond breeding. Assays that distinguish S7 alleles from SI alleles can be used to select self-fertile progeny, while those that distinguish among SI alleles can be used to design compatible crosses. Unlike assays that were previously available for some alleles, the fluorescence-based KASP assays developed here do not require gel electrophoresis. They can be applied at high throughput to large numbers of samples. Given the considerable advantages of self-fertility in breeding and horticulture, the S7 assays are likely to be particularly useful. Among the S7 assays, the two-primer presence-absence assay WriPdSf-1 should be sufficient for most applications. The others (WriPdSf-2 through WriPdSf-5) could be advantageous in cases where it is useful to know what other allele is present in combination...
with \( S_f \). As all of these assays are based on genomic sequence polymorphisms, none of them can be expected to distinguish the active \( S_f \) allele (\( S_{fa} \)) from its inactive epi-allele (\( S_{fi} \)).

Due to the high level of sequence variation among \( S\)-RNase alleles, the approaches used for assay design had to go beyond the methods that are routinely used to design KASP assays for individual SNPs within otherwise conserved regions. Some differences could only be detected as presence-absence polymorphisms. Some assays were designed using degenerate primers. Several assays have the same common primer but different sets of alternative allele-specific primers. The approaches used here could be useful for designing markers for other \( S \)-alleles in almond or other species, or for other multi-allelic loci.

Prior to this research, some variation had been noted in the length and sequence of the interval between the \( S\)-RNase and \( SFB \) genes and it had been suggested that this variation could contribute to \( S \)-haplotype specificity by limiting recombination within the interval. Among 11 haplotypes for which this region was completely sequenced, we observed over five-fold variation in length and substantial sequence variation. Consistent with the idea that this could be a region of low recombination, this region is AT-rich. Recombination-enriched sites are often in regions with high GC content.

The protein sequences deduced from \( S\)-RNase allele sequences contain five conserved regions (C1, C2, C3, RC4 and C5) and the hypervariable region (RHV) that are considered characteristic of Rosaceae \( S\)-RNases, two variable regions (V1 and V2) that had previously been reported between RC4 and C5, and three highly variable regions that had not previously been reported. Similarly, the protein sequences deduced from \( SFB \) allele

---

**Figure 6.** Examples of effects of the mutation of individual \( S\)-RNase residues. (a) Wild type \( S\)-RNase showing the positions of Val45 and Asn53 residues. (b) Mutant \( S\)-RNase with Val45 replaced by Leu45, resulting in an energy loss of 1.58 kcal/mol. (c) Mutant \( S\)-RNase with Asn53 replaced by Gln53, resulting in an energy gain of 5.1 kcal/mol. (d) Wild-type \( S\)-RNase showing the positions of Asn79 and Asn151 residues. (e) Mutant \( S\)-RNase with Asn79 replaced by Tyr79, resulting in an energy loss of 0.18 kcal/mol. (f) Mutant \( S\)-RNase with Asn151 replaced by Asp151, resulting in an energy gain of 0.12 kcal/mol. The lower part of each panel presents protein surfaces coloured by electrostatic potentials: blue = + 5 kT·e\(^{-1}\); white = neutral; red = −5 kT·e\(^{-1}\). Structures are shown in the same orientations as in panels a and b of Fig. 4.
sequences contain previously reported features (the F-box motif, two variable regions (V1 and V2) and two
hypervariable regions (HVa and HVb) and two additional short highly variable regions (V3 and V4). The Ka/Ks
ratio obtained for the complete S-RNase allele sequences (0.60) is similar to values that were previously reported
for almond58, while that obtained for the complete SFB allele sequences (0.50) is similar to what has been reported
for sweet cherry (Prunus avium L.)59.

To generate three-dimensional models for almond S-RNases, we needed to select a template from among
RNases for which crystal structures had been determined. After thorough evaluation of the sequences of eight
candidate RNases in comparison with almond S-RNase sequences, we selected the 1J1G (MC1 RNase) template.
This template had previously been used for 3D modelling of three almond S-RNases: the Sγ-, Sη-, and Sρ-RNases.
We constructed models for those three proteins and for two others (the Sε- and Sτ-RNases). In agreement with
what was previously reported56, these models consist of α-helices and β-strands that are inter-connected by loops.
The variation that we observed in the numbers, lengths and positions of secondary structure elements was similar
to that reported for S-RNase proteins of Japanese pear and apple19,37,38 but greater than what had been reported
for almond56. This difference is likely due to the more comprehensive modelling processes used here, with several
tools used to align protein sequences and identify mismatches, many models generated from slightly different
alignments and final models selected based on optimisation and evaluation of binding energy for multiple stable
low-energy models.

Within the selected models, the positions of conserved cysteine residues and the predicted positions of
disulphide bridges between them are similar to what has been observed for other members of the T2 RNase
enzyme family, including the Sγ- and Sτ-RNases of Japanese pear19,39. Disulphide bridges may help stabilise
the secondary and tertiary structure of S-RNases, contributing to maintaining the proteins in a flexible yet active
conformation40.

The active sites of RNases include amino acid residues that temporarily bind with RNA and residues that cat-
alyse cleavage of RNA. Cysteine, histidine, glutamic acid, lysine and tryptophan residues have been proposed to
be particularly important for these roles19,41. Such residues were observed within the putative active sites that we
identified for almond S-RNases, all at positions that correspond with those of similar residues in the active site of
the Sγ-RNase of Japanese pear19. A conserved cysteine residue in the active site may influence the binding affinity
of the protein by enhancing the interaction between the enzyme and its substrate. Conserved histidine residues
in the active site may be catalytically important. Histidine residues in the active sites of RNases of the fungus
Rhizopus niveus M. Yamaz., have been shown to act as the key residues that mediate catalysis19,42. Consistent with
this, it has been shown that the loss of a histidine residue from the C2 region of an S-RNase leads to self-fertility
in Peruvian tomato (Solanum peruvianum L.)43 and that carboxymethylation of histidine residues inactivates
S-RNases in jasmine tobacco (Nicottiana alata Link & Otto)44. Conserved glutamic acid and lysine residues in the
α3 element within the active site may be important in stabilising a penta-covalently associated RNA substrate
intermediate42,45. Conserved tryptophan residues within the active site may be important for fixation of catalyt-
ically important histidine and glutamic acid residues45,46, via formation of hydrogen bonds between tryptophan
residues and the γ-carboxyl groups of glutamic acid residues and/or stacking interactions between the indole ring
of tryptophan residues and the imidazole ring of histidine residues. For other RNases, tryptophan residues have
been reported to contribute to energy transfer with bound substrates46,47. Among the almond S-RNases examined
here, two lysine residues are completely conserved: one in C3 and the other between V2 and C5. Another lysine
residue, in V2, is conserved among all of the almond S-RNases except Sη-RNase. The conserved lysine residues
in V2 and between V2 and C5 correspond with lysine residues that have been detected in other species of the
Rosaceae48.

Variation in the numbers, lengths and positions of α-helices, β-sheets and loops may contribute to functional
differences among almond S-RNases. Residues at loop-anchoring points could be particularly important in
influencing protein folding topologies, depending on their sizes and whether they have hydrophobic/hydrophilic
or polar/non-polar characteristics. Among the five proteins that were modelled here, the most obvious struc-
tural difference involves an extended loop in the Sf-RNase. This loop was previously reported, with discussion
of how it might contribute to self-fertility46. Now that it has been demonstrated that the S-RNase encoded by
the S9 epiallele can function in self-incompatible interactions15, it seems unlikely that the long loop determines
self-fertility. Nevertheless, it is intriguing that this S-RNase has such a distinct structural feature.

Although the main-chain backbones of all five almond S-RNases superposed very well with each other in
other parts of the molecules, there are noticeable differences in the RHV region, which is exposed on the protein
surface. This supports the idea that the RHV region is important in determining specificity. Variation in the
length and charge of the loop within RHV could provide flexibility for mediation of intermolecular interactions
on the protein surface. The V4 region, which is also on the protein surface may contribute to the regulation of
protein–protein interactions by affecting the conformation of secondary structural elements such as α-helices and
loops. The hydrophobic and electronegative nature of the α-helices may stabilise the conformation of both the
α-helices and the loop.

In the S-RNase proteins examined here, predicted N-glycosylation sites were more abundant in internal
regions than in C termini and were often within loops and loop anchoring points. N-glycosylation of these
proteins may be predominantly post-translational rather than co-translational and may contribute to ensuring
proper protein folding and stabilising secondary structures such as α-helices and β-sheets. None of these proteins
had more than one predicted O-glycosylation site, in accordance with previous reports that O-glycosylation is not
common in plants48.

In Solanum chacoense, site-directed mutagenesis of specific sites within the RHV-encoding region of an
S-RNase allele led to the acquisition of dual SI49,50, in which one S-RNase can recognise two SFB alleles. Here, we
applied in silico mutagenesis to investigate effects of changing specific residues in the almond Sγ-RNase. High
destabilisation energies were observed at N- and C- termini of S-RNase proteins, indicating that these termini are
flexible. Stabilising or destabilising effects in the RHV region and differences in the number, types and positions of N-glycosylation sites could also contribute to maintaining substrate specificity and function.

This is the first report on high-throughput sequencing of haplotypes of the complex S-locus of almond. With this approach, we completed the DNA sequence of the $S_7$ haplotype and generated complete sequences for three other haplotypes and partial sequences for 11 haplotypes. This provided new information on structural variation within the S locus and on allelic variation in the genes that determine SI specificity and made it possible to design high-throughput marker assays for application in almond breeding. The 3D protein modelling, surface morphology assessments and assessments of sites with N-glycosylation and/or O-glycosylation potential conducted here broaden knowledge on the structure and possible mechanisms of S-RNase-based SI, indicating how numbers, lengths, sequences and positions of secondary structural elements, electrostatic potential and surface conformation of the RHV region and post-translational modification could affect S-RNase function and specificity.

Methods

Plant materials, library preparation and DNA sequencing. To enable design of primer pairs that would provide overlapping amplicons from the S-locus, the $S_7$ haplotype sequence (AB081587) was aligned with available sequences for almond Sf, S-RNase and SFB alleles and for pseudomolecule 6 of peach (Prunus persica (L.) Batsch) (Supplementary Table S17), using ten iterations of the Map to Reference alignment algorithm in Geneious software version 9.0.2. Primer pairs (Supplementary Tables S18 and S19) were designed using Primer3 software (http://primer3.ut.ee). Genomic DNA was extracted from young leaves of 48 almond clones (Supplementary Table S1), using an Isolate II Plant DNA Extraction Kit (Bioline, NSW, Australia). DNA quality and quantity were assessed on 1% (w/v) agarose gels using an Agilent 2100 Bioanalyzer (Agilent Technologies, VIC, Australia), diluted to 12 pM and mixed with 1% (w/v) Illumina PhiX library. Paired-end sequencing was performed in a total volume of 20 μL using 20 ng of DNA with 1x Phusion® HF, 1.25 mM dNTPs, 1 μM primer mix, and 0.2 U of Phusion High-Fidelity DNA polymerase (New England Biolabs, Ipswich, MA, USA). The PCR conditions used were 98 °C for 30 s, 34 cycles of 98 °C for 10 s, annealing temperature for 30 s and 72 °C for 10 min followed by a final extension at 72 °C for 15 min. Samples (5 μL) of the amplified products and the HyperLadder I DNA ladder (Bioline, NSW, Australia) were run on 1% (w/v) agarose gels at 100 V for 30 min. Gels were stained with SYBR® Safe (Invitrogen, NSW, Australia).

With each of seven primer pairs (Supplementary Table S18), two products differing in length were amplified from each of eight clones that carry the $S_7$ haplotype (Supplementary Table S19). In each case, one of these products was of the length expected for the $S_7$ haplotype. When the same primer pairs were applied to clones that do not carry the $S_7$ haplotype, only four of the seven pairs amplified products. With additional primer pairs (Supplementary Table S20), additional products were amplified (Supplementary Table S21).

Each PCR product was classified as strong or weak based on the intensity of electrophoretic bands. For each almond clone, strong and weak products were pooled in separate tubes. Each pooled sample was purified using AMPure® XP beads (Agencourt Bioscience, MA, USA). For each almond clone, the strong and weak pools were mixed together at a ratio that should provide approximately uniform coverage across the S locus.

A sequencing library was prepared using an IlluminaNextera DNA Library Prep Kit (V3) (Illumina, VIC, Australia) and 50 ng of DNA from each of the resulting samples. The Tn5 transposase from the kit was used to digest DNA samples to generate segments of about 300 bp containing read 1 (5'–TGTCGGGCTCGG–3') and read 2 (5'–GTCTCGTGGGCGGCCTC–3') sequences. Index primers i5 and i7 and paired-end primers P5 and P7 were annealed to each sample using reduced-cycle PCR amplification. Amplified products were purified using AMPure® XP, quantified by qPCR using Kapa SYBR® FAST Master Mix (Kapa Biosystems, MA, USA) on a Rotor-Gene Q instrument (Qiagen, VIC, Australia) and assayed for quality using a TapeStation 2002 instrument (Agilent Technologies, VIC, Australia). Each sample was normalised to 4 nM and the samples were pooled. The resulting library was quality for analysis in a Bioanalyzer 2001 instrument (Agilent Technologies, VIC, Australia), diluted to 12 pM and mixed with 1% (w/v) Illumina PhiX library. Paired-end sequencing was conducted on an Illumina MiSeq instrument, using an Illumina 600-cycles reagent kit.

Sequence analysis. Raw sequence reads were assessed for quality, adapter sequences and barcode contamination using FASTQC v0.11.5 (http://bioinformatics.babraham.ac.uk/projects/fastqc). Adapter sequences were removed using the ULLUMINACLIP option in Trimmomatic V0.32. This was followed by another run of FASTQC. Sequence data were deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) as study SRP133723. Trimmed reads from 10 clones known to carry the $S_7$ haplotype were aligned with the AB081587 sequence using the BWA-mem algorithm in the Burrows-Wheeler alignment (BWA-0.6) tool. The resulting binary alignment/map (BAM) files were visualised using Tableau graphical viewer version 1.16.09.06. Trimmed reads from each clone were assembled using Mimicking Intelligent Read Assembler (MIRA) version 4.0.2. The resulting contig sequences were mapped to the AB081587 sequence and visualised using CONTIGuator software. Further, large contigs (size $\geq$ 500 bp; N $\geq$ 50) were aligned to the AB081587 sequence using the Map to Reference function in Geneious software version 9.0.2. Sequences were obtained from output files of the pileup command in SAMtools version 1.2. Polymorphisms were graphically visualised using Integrated Genomic Viewer version 2.3.

Sequences from each of ten clones that carry the $S_7$ haplotype were aligned with the AB081587 sequence using the Clustal W multiple sequence alignment algorithm in Geneious software version 9.0.2. An $S_7$ sequence was established for each clone and an overall consensus sequence was established for $S_7$. Sequences from $S_1, S_8$ and $S_7$ were then compared to the overall consensus $S_7$ sequence using the variant call format in VCFtools v0.1.13. This provided haplotype sequences for $S_1, S_8$ and $S_7$, which were then used as references to obtain partial
sequences for other haplotypes. Sequences of S haplotypes and SLF, S-RNase and SFB alleles were deposited in GenBank; their accession numbers are listed in Table S2.

Pairwise sequence differences among haplotypes and among SLF, S-RNase, and SFB alleles were determined using Clustal W. Conserved blocks were identified using the Gblocks version 0.91b tool on the Phylogeny.fr online server (www.phylogeny.fr) with the ‘less stringent’ data selection setting.

Protein-coding sequences of the S<sub>2</sub>, S<sub>3</sub>, S<sub>9</sub>, and S<sub>19</sub> haplotypes were predicted and analysed with BLASTX 2.8.56 using the refseq protein database, and with GENSCAN<sup>66</sup> using Arabidopsis (Arabidopsis thaliana L.) as the reference. LTR retrotransposons were detected with LTR_Finder version 1.0.5<sup>67</sup> using an Arabidopsis tRNA database (http://lowelab.ucsc.edu/GtRNAadb/) to predict protein-binding sites. The deduced amino acid sequences of almond SLF proteins were compared to those of sweet cherry SLF-like proteins (XP_021802052.1 (PavSLFL1), XP_021803309.1 (PavSLFL2), XP_021816935 (PavSLFL3), X_P021800841 (PavSLFL4/5), XP_021821224.1 (PavSLFL6), XP_021802446.1 (PavSLFL7) and XP021816963.1 (PavSLFL8)<sup>10</sup>. Presence of SLF-like genes in the vicinity of the almond S-locus were identified by conducting a homology search using Blastp (word size 6 and an E-value of 1e<sup>-5</sup>) with the query sequences of <i>P. avium</i> SLFLs using the genome databases of <i>P. dulcis</i> Texas Genome v2.0<sup>68</sup> in the GDR database (https://www.rosaceae.org/). Coding sequences of S-RNase and SFB genes were analysed for non-synonymous (Ka) and synonymous (Ks) variation using DnaSP v6.10<sup>69</sup> with a 100 bp sliding window and 20bp steps. Regions with sequence identity below 35% were considered to be variable.

### Design and application of S-allele markers.

Primer sets consisting of two primers (an allele-specific primer for a target allele and a second primer) or three primers (two allele-specific primers and a common primer) were designed using the S-RNase allele sequences following KASP™ (LGC Ltd, Teddington, UK) primer design guidelines<sup>70</sup> and using Primer 3 software version 4.0<sup>54</sup>. Tail sequence complementary to the FRET casette in the KASP Master Mix were added to the 5′ ends of the allele-specific primers. The resulting primer sets were named with the prefix WriPdS with Wri referring to the Waite Research Institute, Pd referring to <i>Prunus dulcis</i> and S referring to the S locus, followed by a number or letter designating a target S allele (e.g. f for S<sub>1</sub>). In cases where more than one primer set was developed to detect the same target allele, a number was appended to distinguish the primer sets (e.g., WriPdS7-1 and WriPdS7-2). Two DNA samples of each of Nonpareil, Antonéta, Carmel, Francolí, Johnston's Prolific, Lauranne, Mandalone, Somerton, Vairo, 12–350, Capella, Mira, Carina and Maxima and two water samples (negative controls) were assayed with all primer sets. DNA samples of 10 ng (5 μL of 2 ng/μL) were dried at 55°C for 1 h. A mixture of 0.028 μL (containing 12 μM of each allele specific forward primer and 30 μM of the common primer) and 1.972 μL of 1 × KASP Master Mix was added to each sample. Amplification was conducted using the standard KASP PCR protocol in a Hydrocycler-14 thermocycler (LGC Ltd, Teddington, UK). Fluorescence detection was performed in a Pherastar Plus plate reader (BMG LABTECH, Germany). Each primer set that was shown to be informative based on results from this panel was assayed on progeny of relevant crosses (Supplementary Table S22). For further evaluation of four primer sets, gBlocks<sup>60</sup> Gene Fragments (Integrated DNA Technologies, Iowa, USA) were synthesised to represent segments of the S-RNase alleles S<sub>2</sub>, S<sub>3</sub>, S<sub>9</sub>, and S<sub>23</sub> (Supplementary Fig. S13). Primer sets WriPdSf-2 through WriPdSf-5 were applied to samples of these synthetic DNA fragments and to 1:1 mixtures of each possible pairwise combination of these fragments (representing heterozygous genotypes) using the KASP screening procedure as described above.

### Analysis of marker data.

Fluorescence data were analysed using Kraken™ software (LGC Ltd, Teddington, UK), which normalises FAM and HEX fluorescence intensities relative to an internal ROX control, identifies clusters of data points and assigns individual data points to clusters. Statistical analyses were conducted using R (https://www.R-project.org/). Normalised FAM and HEX fluorescence intensities were subjected to one-way multivariate analysis of variance (MANOVA) with cluster (genotype call) as the independent variable. In cases where more than two clusters, post hoc comparisons among clusters were conducted using pairwise Tukey contrasts, as implemented in the R package MANOVA.RM<sup>71</sup>. For molecular marker data collected from the progeny of crosses, chi-square tests were used to assess the deviation of observed genotypic ratios from expected ratios.

### 3D modelling of S-RNase proteins.

Protein sequences and crystal structures of eight RNases (1J1G, 1J1F, 1BK7, 1UCA, 1UCC, 1UCD, 1UCCG, 1V9H) were downloaded from the Protein Data Bank<sup>72</sup>. The sequences were aligned with the deduced protein sequences for almond S-RNase alleles using PSI-BLAST<sup>73</sup>, PSIPRED v3.3<sup>74</sup> and RaptorX<sup>75</sup>. Five almond S-RNases (S<sub>2</sub>, S<sub>3</sub>, S<sub>9</sub>, S<sub>19</sub>, and S<sub>23</sub>) were subjected to comparative protein modelling using Modeller V9.19<sup>76</sup>. From among known 3D structures with more than 35% sequence identity to the target sequences, the one with the highest positional sequence identity and the lowest E-value was selected for each target S-RNase. Each target S-RNase was aligned with its selected template using MUSCLE alignment<sup>77</sup> in Geneious version 9.0.2. Alignments were checked using PSIPRED v3.3. For each S-RNase, four structurally aligned sequences were used to construct 100 3D models using Modeller V9.19. From among these models, five models with favourable modeller objective function (MOF)<sup>78</sup> and discrete optimised protein energy (DOPE)<sup>79</sup> parameters were selected. Each model was optimised using FoldX4<sup>80</sup> and evaluated using ProSa 2003<sup>81</sup> and PROCHECK<sup>82</sup>. Energy and stability were calculated using FoldX4. Based on all evaluations, the best-scoring models were selected for each S-RNase. Selected models were superposed on the template structures, yielding root mean square deviation (RMSD) values as indicators of structural folds. Images were generated with PyMOL Molecular Graphics V1.8.2.0 (Schrödinger LLC, NY, USA).

To identify putative catalytic sites, PSI-BLAST, PSIPRED and PROMALS3D were used to align almond S-RNase sequences with the sequence of the S<sub>1</sub>-RNase of Japanese pear (<i>Pyrus pyrifolia</i> (Burm.) Nak) (BAA93052.1)<sup>83</sup>. The positions of disulphide bonds in the S-RNases were identified using the SSBOND
record of PDB files of ribonucleases and using protein sequences as input in the Disulfind online server (http://disulfind.dsi.unifi.it/process.php). Molecular surfaces of S-RNases were generated in PyMOL, using a probe radius of 1.80 Å. Solvent-accessible areas were estimated with Naccess V2.1.1 (http://wulf.bms.umist.ac.uk/naccess/). Electrostatic potentials were calculated with the Adaptive Poisson-Boltzmann Solver48 using the PyMOL plug-in APBS Tool2, with AMBER force field parameters48 and dielectric constants of 78 (solvent) and 2 (solute). The values of electrostatic potentials were expressed using Boltzmann constant (k) and the temperature (T) per Einstein (e) (kT/e).

Amino-acid residue positions were expressed for their potential to create new S-RNase specificities by examining Ka/Ks ratios, electrostatic potentials and the cumulative charges of the residues in positively charged regions.

The sequences of five S-RNases (S5, S7, S8, S23 and Sf) were aligned using PROMALS3D46 (Supplementary Fig. S12) to identify positions at which the S7-RNase sequence differed from the sequence(s) of one or more of the other four S-RNases. At each of these positions, the BuildModel command in FoldX48 was used to mutate the S7-RNase residue to the alternative residue (or residues), considering both single and multiple mutation options. A similar process was applied to the Sf-RNase, with residues replaced by those in the Sf–RNase. The energies for the wild-type (ΔG wild-type) and mutant (ΔG mutant) proteins were computed using FoldX) to find differences in stability (ΔΔG = ΔG mutant − ΔG wild-type) of proteins. The ΔΔG values were classified into seven bins based on the standard deviation in FoldX: (i) highly stabilising (ΔΔG < −1.84 kcal/mol), (ii) stabilising (−1.84 kcal/mol ≤ ΔΔG < −0.92 kcal/mol), (iii) slightly stabilising (−0.92 kcal/mol ≤ ΔΔG < −0.46 kcal/mol), (iv) neutral (−0.46 kcal/mol < ΔΔG ≤ 0.46 kcal/mol), (v) slightly destabilising (0.46 kcal/mol < ΔΔG ≤ 0.92 kcal/mol), (vi) destabilising (+0.92 kcal/mol < ΔΔG ≤ +1.84 kcal/mol), and (vii) highly destabilising (ΔΔG > +1.84 kcal/mol).

The sequences of five S-RNases (S5, S7, S8, S23 and Sf) were analysed using the NetNGlyc 1.0 server (http://www.dtu.dk/services/NetNGlyc/) and sequences with N-glycosylation potential greater than 0.5 were selected as potential N-glycosylation sites. Within the sequons identified as potential N-glycosylation sites in the S-RNase, Asn and Xaa residues were replaced by the corresponding residues from the S7-RNase, using the BuildModel command of FoldX48. O-glycosylation sites in the same five S-RNases were identified using the DictryOGlyc 1.1 Server (http://www.cbs.dtu.dk/services/DictryOGlyc/), with sites with O-glycosylation potential greater than 0.5 considered as potential O-glycosylation sites.

Data availability
Sequence data have been deposited in the National Center for Biotechnology information (NCBI) Short Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra/) as study SRP133723. S-haploype sequences and allele sequences for SFB, S-RNase and SFB genes have been deposited in the National Center for Biotechnology information (NCBI) GenBank database (https://www.ncbi.nlm.nih.gov/genbank/). GenBank accession numbers are listed in Supplementary Table S2.

Received: 22 April 2019; Accepted: 31 December 2019;
Published online: 17 January 2020

References
1. Ballester, J. et al. Location of the self-incompatibility gene on the almond linkage map. Plant Breed 117, 69–72 (1998).
2. Bošković, R., Tobutt, K., Ortega, E., Sutherland, B. & Godini, A. Self-(in)compatibility of the almonds from almond, Prunus dulcis. Theor. Appl. Genet. 103, 1115–1122 (2001).
3. Channuntapipat, C., Sedgley, M. & Collins, G. Sequences of the cDNAs and genomic DNAs encoding the S-RNases. Plant Mol. Gen. Genomics 278, 665–676 (2007).
4. Halász, J., Fodor, Á., Pedryc, A. & Hegedüs, A. Characterization of natural mutations in Iranian almond cultivars. Trees 27, 497–510 (2013).
5. Hafizi, A., Shiran, B., Maleki, B., Imani, A. & Banovic, B. Identification of new S-RNase self-incompatibility alleles and characterization of natural mutations in Iranian almond cultivars. Trees 27, 497–510 (2013).
6. 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).
7. 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. 260, 261–268 (1998).
8. Bošković, R., Tobutt, K. R., Baffle, I. & Duval, H. Correlation of ribonuclease zymograms and incompatibility genotypes in almond. Euphytica 97, 167–176 (1997).
9. Matsumoto, D. & Tao, R. Recognition of a wide range of S-RNases by S locus F-box like 2, a general-inhibitor candidate in the Prunus-specific S-RNase-based self-incompatibility system. Plant Mol. Biol. 91, 459–469 (2016).
10. Matsumoto, D. & Tao, R. Recognition of S-RNases by an S locus F-box like protein and an S haplotype-specific F-box protein in the Prunus-specific self-incompatibility system. Plant Mol. Biol. 100, 367–378 (2019).
11. Grasso, C. & Olivier, G. Mise en évidence de quelques types autocompatibles parmi les cultivars d’amandier (P. amygdalus Batsch) de la population des Pouilles. Ann. Amél. Plantes 26, 107–113 (1976).
12. Socías i Company, R. Breeding self-compatible almonds in Plant Breeding Reviews Vol. 8 (ed. Janick, J.) 313–338 (John Wiley & Sons, Inc., 1990).
13. Kodad, O., Socías i Company, R., Sánchez, A. & Oliveira, M. M. The expression of self-compatibility in almond may not only be due to the presence of the S allele. J. Am. Soc. Hortic. Sci. 134, 221–227 (2009).
14. Fernández i Martí, A. et al. The almond S allele shows a double expression despite its comprehensive genetic identity. Sci. Hortic. 125, 685–691 (2010).
15. Fernández i Martí, A., Gradziel, T. & Socías i Company, R. Methylation of the S allele in almond is associated with S-RNase loss of function. Plant. Mol. Biol. 86, 681–689 (2014).
16. Kodad, O., Socías i Company, R. & Alonso, J. M. Unilateral recognition of the S allele in almond. Sci. Hortic. 185, 29–33 (2015).
Acknowledgements
This research was funded by Hort Innovation using the almond industry and development levy and contributions from the Australian Government (Project AL12015), Illumina, Inc. (pilot project grant for sequencing); the Australian Research Council (Project DP120100900) and the University of Adelaide (Australian Postgraduate Award to SNG).

Author contributions
S.N.G. and A.E.C. prepared the Illumina sequencing library. S.N.G., A.E.C. and T.J.M. conducted DNA sequence analysis, S.N.G. designed and tested marker assays, S.N.G. and M.H. conducted protein sequence analysis and modelling. M.G.W. provided almond materials and S-genotype information. S.N.G., M.H. and D.E.M. wrote the manuscript. All authors contributed to the design of the study, interpretation of the results and revision of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-57498-6.
Correspondence and requests for materials should be addressed to D.E.M.

Reprints and permissions information is available at http://www.nature.com/reprints.

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

© The Author(s) 2020