SCIENTIFIC REPORTS
nature research

OPEN

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

Shashi N. Goonetilleke[1], Adam E. Croxford[1], Timothy J. March[1], Michelle G. Wirthensohn[1], Maria Hrmova[1,2] & Diane E. Mather[1*]

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*-locus[1–5]. Sequencing of a 71,953 bp region of one haplotype (S_7, also known as S_c) 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 S_7 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 sequenced[1–5]. Among haplotypes that have been physically mapped, the order and orientations of *S*-locus features are conserved, but the distances between these features vary[6].

In *Prunus*, including almond, the specificity of SI is mainly determined by the *S-RNase* and *SFB* genes[6,7], which are expressed in pistils and pollen tubes, respectively. S-RNases act as cytotoxins in self pollen tubes[8] 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' proteins[9,10]. This allows

[1]School of Agriculture, Food and Wine, Waite Research Institute, The University of Adelaide, PMB 1, Glen Osmond, SA, 5064, Australia. [2]School of Life Sciences, Huaiyin Normal University, Huai'an, 223300, China. *email: diane.mather@adelaide.edu.au

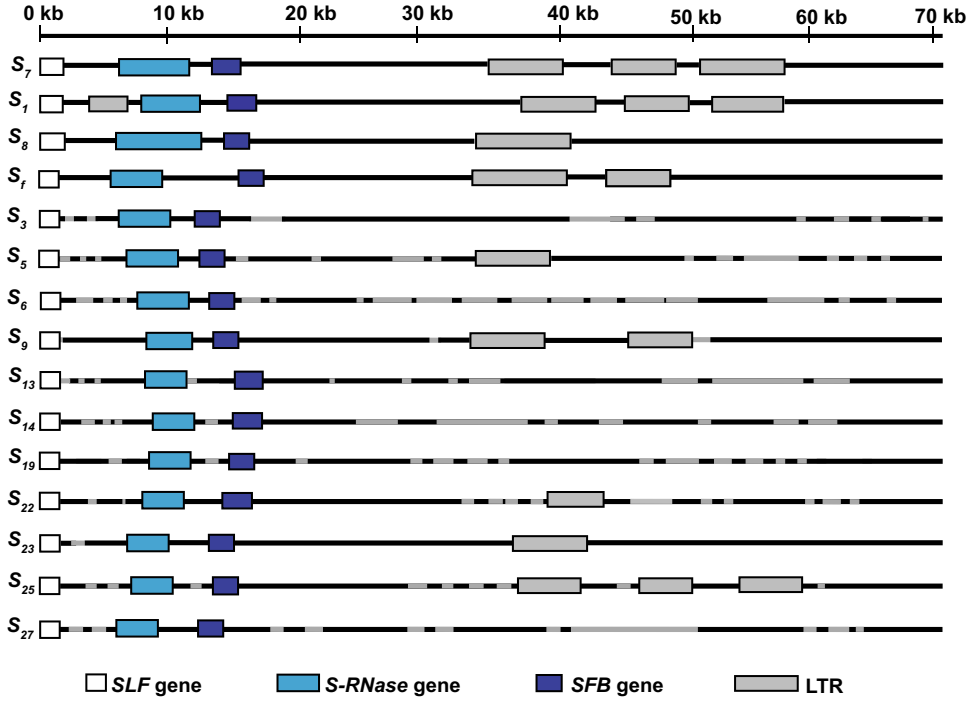

Figure 1. *S*-locus structure. Structure of the almond *S* locus showing the positions of the *SLF*, *S-RNase* and *SFB* genes and long terminal repeat retrotransposons (LTRs). Black lines indicate regions for which sequences were obtained and grey lines indicate gaps in the sequence.

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 S_f[11,12] or S_{fi} (S_f-*inactive*)[13]. Plants carrying this allele do not express an active S_f-RNase, possibly due to poor transcription[14], and are not able to block the growth of S_f pollen tubes. A similar allele, S_{fa} (S_f-*active*) expresses an active S-RNase and confers SI[13]. Despite their contrasting phenotypes, S_{fi}- and S_{fa}-*RNase* alleles have identical nucleotide sequences and are linked with identical *SFB* alleles[14]. This apparent paradox was resolved by the discovery that S_{fi} and S_{fa} are epialleles[15], differing by the methylation of a single nucleotide upstream of the coding sequence. This epigenetic difference may determine whether the S_f-*RNase* allele is expressed. Consistent with this interpretation, $S_{fi}S_{fa}$ heterozygotes have been found to be fully self-incompatible[16], with their S_{fa}-RNase able to block the growth of both S_{fi} and S_{fa} 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_7 haplotype were extracted and aligned with the previously available S_7 haplotype sequence (AB081587). An S_7 haplotype sequence was assembled for each clone (Supplementary Table S2). 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 S_7. It includes the identity of 21 nucleotides that were missing from AB081587. It differs from AB081587 at just five positions. At all five of those positions, the same nucleotide was called for all eight clones.

Using the consensus S_7 sequence as a new reference for S_7, haplotype sequences were obtained for: S_1, from Brown Nonpareil (S_1S_7); S_8, from Nonpareil and McKinlays (both S_7S_8); and S_f, from the self-fertile clones Carina, Mira, Capella, T5 and T7 (all S_7S_f). There is just 2% inconsistency between the two S_8 sequences and 3% inconsistency among the five S_f sequences. Among the four fully-sequenced haplotypes (S_1, S_7, S_8 and S_f), positional sequence identity ranges from 51 to 84% (Supplementary Table S3).

Using the four complete haplotype sequences as references, partial sequences were derived for other haplotypes (S_3, S_5, S_6, S_9, S_{13}, S_{14}, S_{19}, S_{22}, S_{23}, S_{25} and S_{27}). The total lengths of the sequences obtained for these haplotypes range from 47% (S_6) to 99% (S_{23}) of the length of the S_7 haplotype sequence (Fig. 1).

In each of the completely sequenced haplotypes, 12 or more open reading frames (ORFs) were detected: 12 in S_7, 12 in S_8, 14 in S_1 and 18 in S_f. (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_f haplotype and the last ORF in each haplotype), no similarity to known-function genes was detected. For others (five for S_7 and S_f, six for S_1 and eight for S_8), homology with transposases from other *Prunus* species was detected. Most of these transposases belong to the *Ty*1-*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_7, S_8, S_9, S_{22}, S_{23}, S_{25}, S_{27} and S_f). These sequences, which are AT-rich (65–70%), range in length from 1.2 (S_9) to 6.6 kb (S_f). The sequence identity among them ranges from 21% (between S_1 and S_f) to 98% (between S_7 and S_{23}) (Supplementary Table S5).

Pairs of LTRs were detected within each haplotype, mostly in positions that correspond to the LTR-containing region of the S_7 haplotype (Fig. 1). No LTRs were detected within the *SLF*, *S-RNase* or *SFB* genes or between the *S-RNase* and *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 *Ty*1-*copia*-like retrotransposons with TG/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[10] showed that the predicted products of all 15 almond *SLF* alleles products are most similar to PavSLFL1 (sequence identity between 82 and 95%, compared to between 55 and 61% for PavSLFL4/5 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 almond *S*-locus: Prudul26A008798 (97% sequence identity with PavSLFL2), Prudul26A009208 (95% sequence identity with PavSLFL3), Prudul26A016695 (95% sequence identity with PavSLFL6) and Prudul26A015917 (97% identity with PavSLFL8).

The *S-RNase* gene, which was completely sequenced for 11 haplotypes (S_1, S_5, S_7, S_8, S_{13}, S_{14}, S_{23}, S_{25}, S_{27} and S_f) and partially sequenced for four haplotypes (S_3, S_6, S_9 and S_{19}), is much more variable. The completely sequenced *S-RNase* alleles range in length from 1.0 kb (S_1) to 4.5 kb (S_8). Their pairwise nucleotide sequence identities range from 19% (between S_1 and S_8) to 51% (between S_7 and S_{27}) (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_5, S_7, S_8, S_9, S_{22}, S_{23}, S_{25}, S_{27} and S_f) and partially sequenced for four haplotypes (SFB_6, SFB_{13}, SFB_{14} and SFB_{19}). The completely sequenced alleles range in length from 1.1 kb (S_9) to 1.5 kb (S_1). Their pairwise sequence identities range from 35% (between SFB_7 and SFB_f) to 86% (between SFB_1 and SFB_{23}) (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 (HVa 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 SFB_6, SFB_{13}, SFB_{14} and SFB_{19} proteins each have an insertion of a single arginine, while SFB_{23} and SFB_{27} 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_f-*RNase* allele, a primer pair (WriPdSf-1; Supplementary Table S10) was designed for an S_f-specific site within intron 2 of the *S-RNase* gene (Supplementary Fig. S4). With this assay, fluorescence is detected for the S_f allele and no signal is detected for any of the other alleles. For example, when this assay was applied to Nonpareil (S_7S_8) × Vairo (S_9S_f) F_1 progeny, HEX fluorescence was detected for S_7S_f and S_8S_f progeny and little or no fluorescence was detected for S_7S_9 or S_8S_9 progeny (Fig. 3a).

Four additional primer sets (Supplementary Table S10) were designed to query an A/C SNP that distinguishes S_f (A) from each of the SI 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 S_f sequence throughout the annealing site. The HEX-tailed primer is exactly complementary to the S_3, S_9, S_{23} and S_{25} sequences throughout the annealing site but not to the S_1, S_5, S_7 and S_8 sequences. For S_5, the first mismatch is too far from the target SNP to interfere with annealing and amplification. For S_1, S_7 and S_8, the mismatches are close enough to the target SNP to prevent annealing and amplification. When the WriPdSf-2 primer set was applied to Chellaston (S_7S_{23}) × Lauranne (S_3S_f) F_1 progeny, HEX fluorescence was detected from the HEX-HEX genotype S_3S_{23} (half from each allele) and from the null-HEX genotype S_3S_7 (all from S_3), FAM fluorescence was detected for the

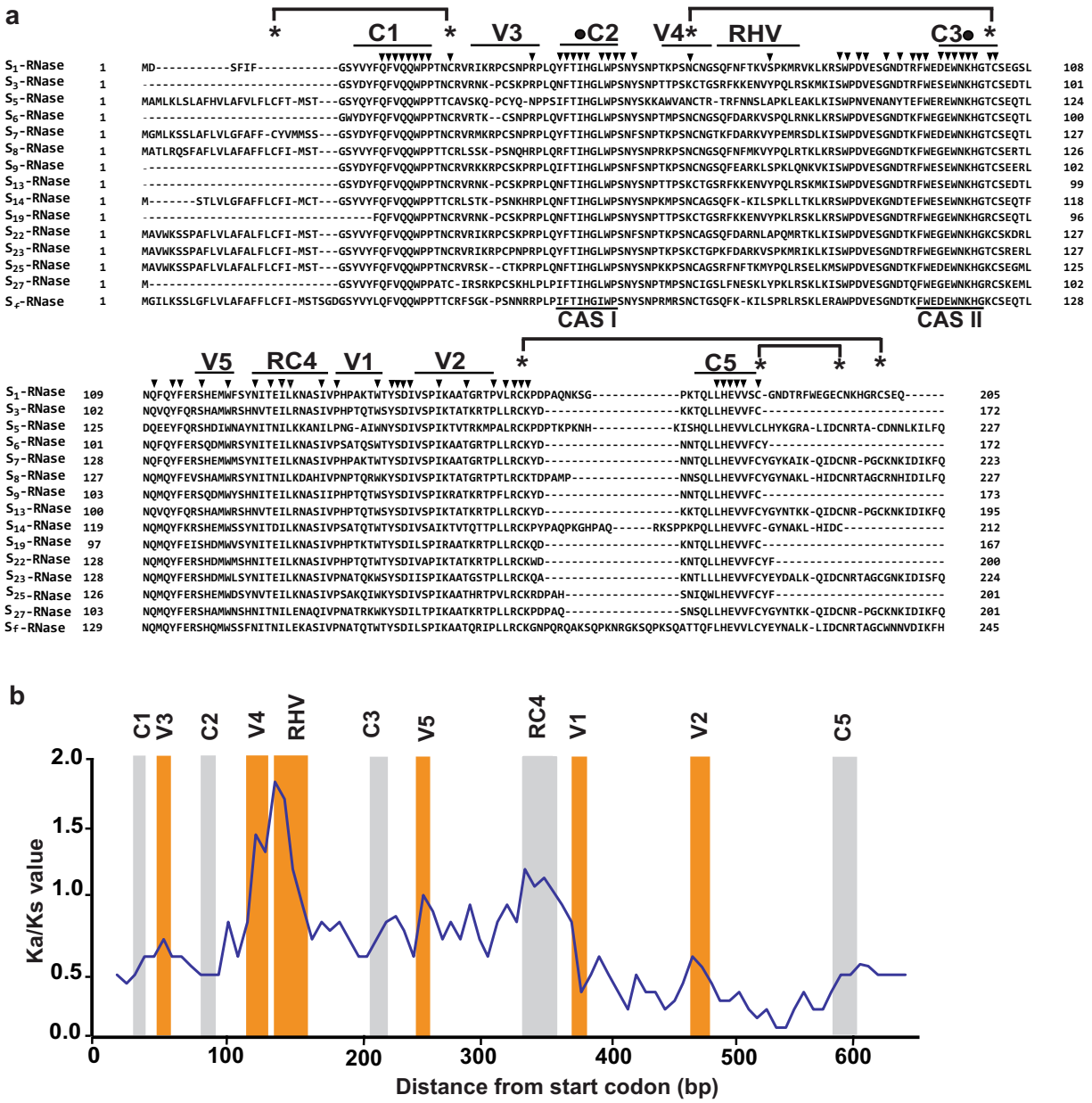

Figure 2. S-RNase sequence alignment and Ka/Ks ratios. (**a**) Sequence alignment of 15 almond S-RNases, showing conserved regions (C1, C2, C3, RC4 and C5), variable regions (V1–V5) a hypervariable region (RHV) and conserved active segments (CAS I and CAS II). Positions of absolutely conserved residues, conserved histidine residues and conserved cysteine residues are indicated by arrowheads, circles and asterisks, respectively. Cysteine residues that are predicted to be linked by disulphide bridges are connected by lines. (**b**) Mean nonsynonymous/synonymous (Ka/Ks) ratios for coding regions of the *S-RNase* gene of almond, average Ka/Ks values calculated for 100 bp sliding windows with a 20 bp step size. Conserved regions (C1, C2, C3, RC4 and C5) are highlighted in grey and variable (V1–V5) or hypervariable (RHV) regions in the S-RNase are highlighted in orange.

null-FAM genotype S_7S_f (all from S_f) and both HEX and FAM fluorescence were detected for the HEX-FAM genotype $S_{23}S_f$ (HEX from S_{23} and FAM from S_f) (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 S_f genotypes from all tested non-S_f genotypes, while also discriminating among some S_f genotypes and among some non-S_f 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, S_f is the only allele for which FAM fluorescence is detected. In each case, there are some SI alleles for which HEX 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 S_f, S_1, S_3, S_5, S_7,

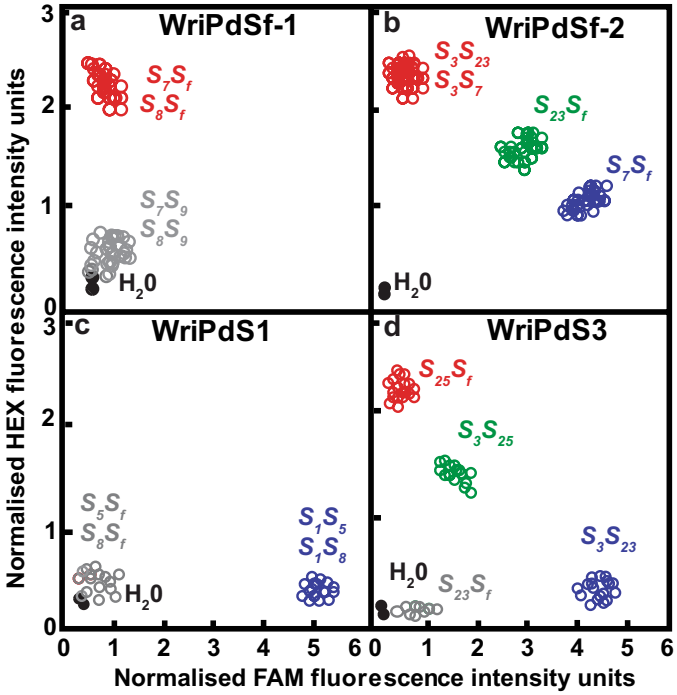

Figure 3. Marker assay results. Results obtained for (**a**) Nonpareil (S_7S_8) × Vairo (S_9S_f) F_1 progeny using primer set WriPdSf-1, (**b**) Chellaston (S_7S_{23}) × Lauranne (S_3S_f) F_1 progeny using primer set WriPdSf-2, (**c**) Carmel (S_5S_8) × 12–350 (S_1S_f) F_1 progeny using primer set WriPdS1 and (**d**) Johnston's Prolific ($S_{23}S_{25}$) × Lauranne (S_3S_f) F_1 progeny using primer set WriPdS3. Data shown are intensities of FAM and HEX fluorescence, each normalised against fluorescence from an internal ROX reference.

S_8, S_9, S_{23} and S_{25} and to mixtures of synthetic DNA representing heterozygous combinations of these alleles confirmed that all four primer sets yield FAM fluorescence when S_f is present and HEX fluorescence when any of S_3, S_5, S_9, S_{23} or S_{25} are present (Supplementary Fig. S6). In each case the amount of HEX fluorescence obtained for one allele with a mismatch (S_5) was very similar to that detected for the alleles with no mismatches (S_3, S_9, S_{23} and S_{25}). In addition, WriPdSf-3 yields HEX fluorescence when S_1 or S_8 is present and WriPdSf-4 yields HEX fluorescence when S_7 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 (S_1S_7 and S_7S_{23} for WriPdSf-3; S_1S_5 for WriPdSf-4). FAM fluorescence for null-FAM heterozygotes (S_7S_f for WriPdSf-2 and WriPdSf-5; S_8S_f for WriPdSf-2, WriPdSf-4 and WriPdSf-5), both HEX and FAM fluorescence for HEX-FAM heterozygotes (S_1S_f and $S_{23}S_f$ for WriPdSf-3 and S_5S_f for WriPdSf-4) and a low levels of HEX fluorescence for null-null heterozygotes (S_1S_7 for WriPdSf-2 and WriPdSf-6; S_1S_8 for WriPdSf-2, WrPdSf-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. S_1 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. S_3 vs S_{25} 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 (χ^2, $\alpha = 0.05$), 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, IUCG 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 S_5-, S_7-, S_8-, S_{23}- and S_f-RNases. The MC1 protein sequence has 38% identity and 48% similarity to the S_7-RNase and S_{23}-RNase protein sequences (E-value = $1.71e^{-27}$) and 36% identity and 47% similarity to the S_f-RNase (E-value = $1.72e^{-27}$), S_5-RNase (E-value = $1.69e^{-27}$) and S_8-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[17]. 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[18].

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 (S_{23}-RNase) to 0.33 (S_8-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 (five for S_7-, S_8-, and S_{23}-RNases, six for S_f-RNase and seven for S_5-RNase) (Fig. 4a,b; 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 α_5 being the longest in all cases. Within the S_5-, S_7-, S_8- and S_{23}-RNases, four β-sheets (β_1, β_2, β_5 and β_6) are located in approximately the same positions. Three of these (β_1, β_2, and β_6) 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 S_5-, S_7-, S_8- and S_{23}-RNases and 81% for the S_f-RNase. The percentages of the exposed surface occupied by positively charged residues are 22%, 25%, 25%, 23% and 19% for the S_5-, S_7-, S_8-, S_{23}- and S_f-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_3-RNase of Japanese pear enabled identification of putative active sites within almond S-RNases[19]. 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 S_7-, S_8-, and S_{23}-RNases, V4 is highly positively charged (Fig. 4c). Considerable structural variation was detected in RHV, which consists of 18 residues in the S_5-, S_7-, S_8- and S_{23}-RNases but only 14 residues in the S_f-RNase. In the S_7-, S_8-, and S_{23}-RNases, RHV consists of two α-helices (α_3 and α_4) and a short loop, while in the S_f- and S_5-RNases, it has only one α-helix (α_4) 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 S_5-RNase, asparagine in the S_7-RNase, tyrosine in the S_8-RNase and serine in the S_{23}- and S_f-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_f-RNase, this loop is much longer (30 residues) than in other S-RNase proteins (10 to 20 residues) (Fig. 4), with eight S_f-specific residues: an isoleucine, an asparagine, a glycine, a phenylalanine, an alanine and three glutamines. The average exposed surface of the extended loop region of the S_f-RNase is 16%.

With *in silico* mutation of individual amino acid residues of the S_7-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 (ΔΔG) due to specific mutations (Fig. 6; Supplementary Fig. S12). Most of the mutations were destabilising but a few were classified as highly stabilising ($\Delta\Delta G < -1.84$ kcal/mol). Most of the highly destabilising mutations ($\Delta\Delta G > 1.84$ kcal/mol) were at positions near the N- or C-termini.

Between four and seven *N*-glycosylation sites were identified in each of the S_5-, S_7-, S_8-, S_{23}- and S_f-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_7-RNase sequons did not substantially affect electrostatic potential (Fig. 6), but substitution of Asn residues within some S_7-RNase sequons affected ΔΔG values. For example, replacement of Asn79 by Try79 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 S_5-, S_7- and S_8-RNases, one *O*-glycosylation site was detected in the V2 region (Supplementary Fig. S12). No *O*-glycosylation sites were detected in either the S_f- or S_{23}-RNases.

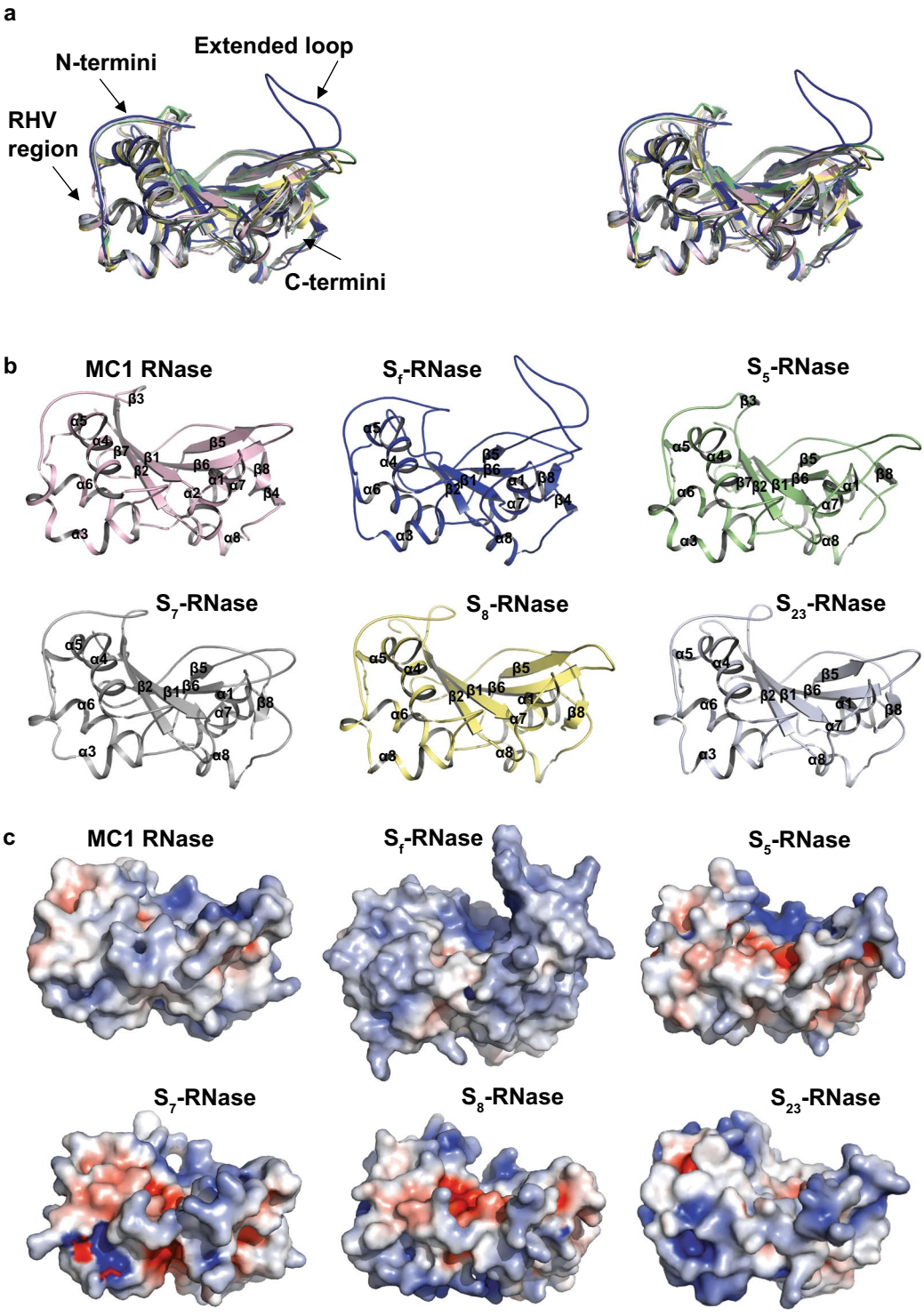

Figure 4. Molecular properties of almond S-RNase structural models. (**a**) Stereo representation of the superposition of 3D structures of S-RNases, whereby the template crystal structure (the MC1 RNase from seeds of bitter gourd) is in pink and the 3D models of almond S_f-RNase, S_5-RNase, S_7-RNase, S_8-RNase and S_{23}-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 C^{α} atoms. (**b**) Dispositions of secondary structure elements in the template, with the S_f-, S_5-, S_7-, S_8- and S_{23}-RNases indicated in pink, blue, green, grey, yellow and tint blue, respectively. (**c**) Molecular surface morphologies of the template structure, and almond S_f-, S_5-, S_7- S_8- and S_{23}-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.

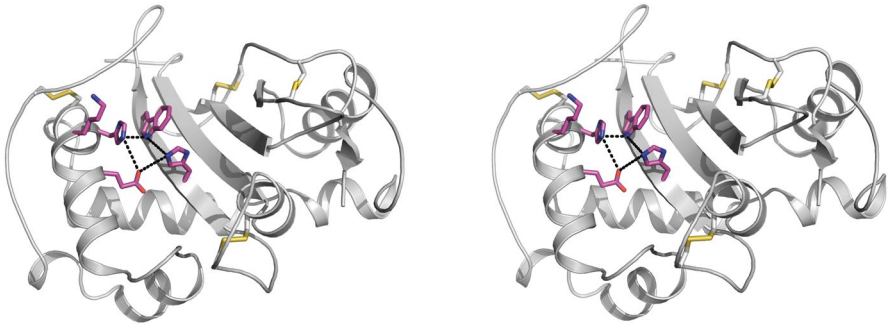

Figure 5. Stereo representation of the active site residues of the almond S_7-RNase. Active site residues (His40, Trp43 in C2 region and Glu93, Lys96, His97 in C3 region) are shown in cpk magenta on the background of the cartoon model (in grey). The positions of disulphide bridges (Cys20-Cys28, Cys56-Cys100, 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 S_7 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 S_7 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[6]). 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 S_7 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[20]. 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 S_7 haplotype[6] were detected in our S_7 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[21,22], possibly reflecting different ways of recognising 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[20,23,24], but no such insertions were observed here.

This work increased the number of sequenced *SLF* alleles from just two (SLF_7 and SLF_8) to fifteen. Consistent with the expectation that SLF does not affect SI specificity in *Prunus* spp.[6], 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 S_f 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[25–28], 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 S_f assays are likely to be particularly useful. Among the S_f 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

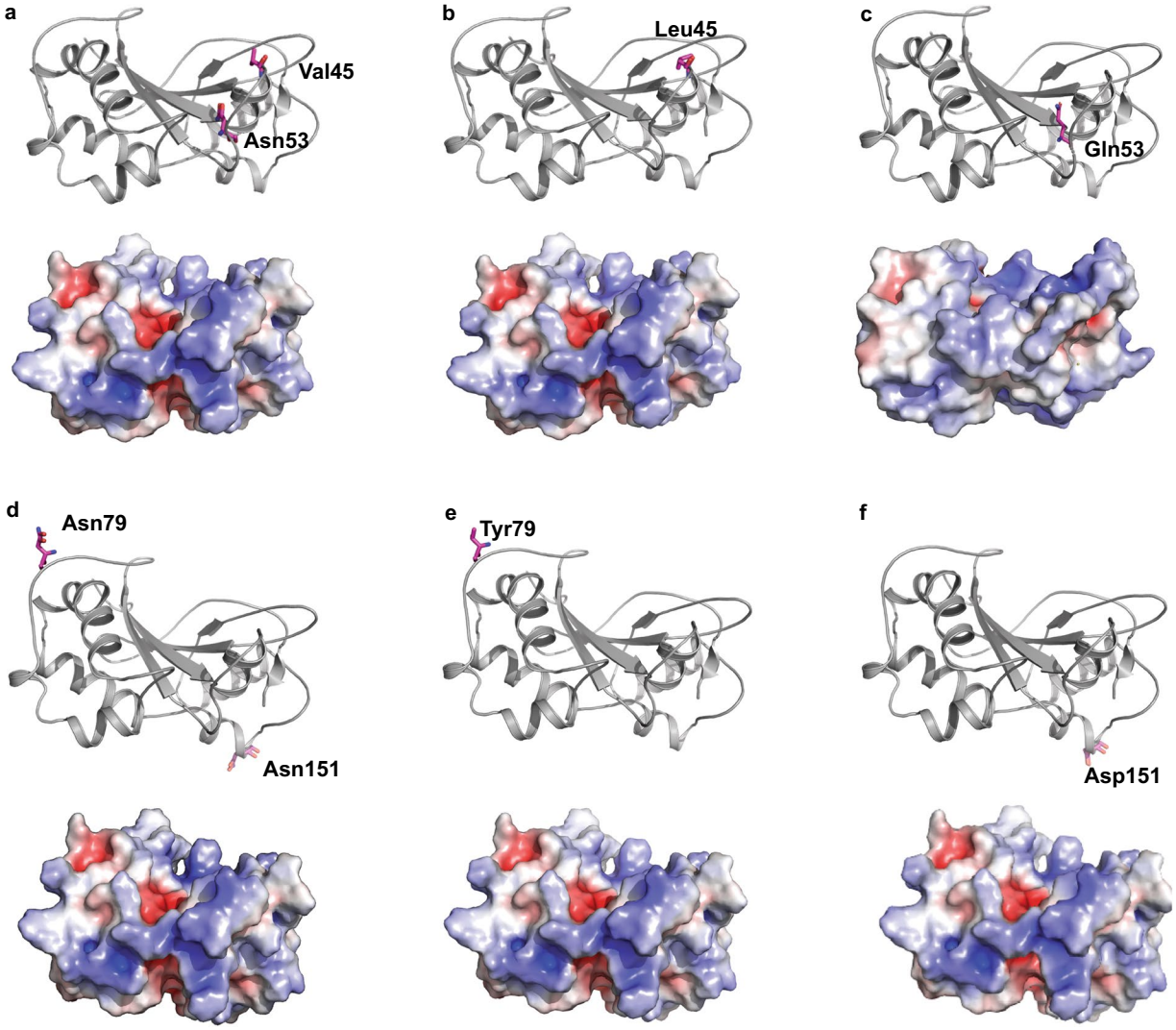

Figure 6. Examples of effects of the mutation of individual S_7-RNase residues. (**a**) Wild type S_7-RNase showing the positions of Val45 and Asn53 residues. (**b**) Mutant S_7-RNase with Val45 replaced by Leu45, resulting in an energy loss of 1.58 kcal/mol. (**c**) Mutant S_7-RNase with Asn53 replaced by Gln53, resulting in an energy gain of 5.1 kcal/mol. (**d**) Wild-type S_7-RNAse showing the positions of Asn79 and Asn151 residues. (**e**) Mutant S_7-RNase with Asn79 replaced by Tyr79, resulting in an energy loss of 0.18 kcal/mol (**f**) Mutant S_7-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.

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[6,29,30] and it had been suggested that this variation could contribute to *S*-haplotype specificity by limiting recombination within the interval[26]. 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[31,32].

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[7,33], two variable regions (V1 and V2) that had previously been reported between RC4 and C5[26,34] and three highly variable regions that had not previously been reported. Similarly, the protein sequences deduced from *SFB* allele

sequences contain previously reported features (the F-box motif, two variable regions (V1 and V2) and two hypervariable regions (HVa and HVb)[6] 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 almond[26], while that obtained for the complete *SFB* allele sequences (0.50) is similar to what has been reported for sweet cherry (*Prunus avium* L.)[35].

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 S_8-, S_{23}-, and S_f-RNases[36]. We constructed models for those three proteins and for two others (the S_5- and S_7-RNases). In agreement with what was previously reported[36], 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 apple[19,37,38] but greater than what had been reported for almond[36]. 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 S_3- and S_4-RNases of Japanese pear[19,39]. Disulphide bridges may help stabilise the secondary and tertiary structure of S-RNases, contributing to maintaining the proteins in a flexible yet active conformation[40].

The active sites of RNases include amino acid residues that temporarily bind with RNA and residues that catalyse cleavage of RNA. Cysteine, histidine, glutamic acid, lysine and tryptophan residues have been proposed to be particularly important for these roles[19,41]. 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 S_3-RNase of Japanese pear[19]. 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 catalysis[19,42]. 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 (*Nicotiana alata* Link & Otto)[44]. Conserved glutamic acid and lysine residues in the α_5 element within the active site may be important in stabilising a penta-covalently associated RNA substrate intermediate[42,45]. Conserved tryptophan residues within the active site may be important for fixation of catalytically important histidine and glutamic acid residues[19,42] 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 substrates[46,47]. 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 S_{19}-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 Rosaceae[38].

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 structural difference involves an extended loop in the S_f-RNase. This loop was previously reported, with discussion of how it might contribute to self-fertility[36]. Now that it has been demonstrated that the S-RNase encoded by the S_{fa} epi-allele can function in self-incompatible interactions[15], 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*-glycoslyation is not common in plants[48].

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 SI[49,50], 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 S_7-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 *SLF, S-RNase* and *SFB* alleles and for pseudomolecule 6 of peach (*Prunus persica* (L.) Batsch)[51] (Supplementary Table S17), using ten iterations of the Map to Reference alignment algorithm in Geneious software version 9.0.2[52]. Primer pairs (Supplementary Tables S18 and S19) were designed using Primer3 software (http:/primer3.ut.ee)[53,54].

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 a HyperLadder I DNA ladder (Bioline, NSW, Australia). PCR amplification was performed in a total volume of 20 uL 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 Illumina Nextera 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′-TCGTCGGCAGCGT-3′) and read 2 (5′-GTCTCGTGGGCTCGG-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 assessed for quality 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-cycle Version 3 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 ILLUMINACLIP option in Trimmomatic V0.32[55]. 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 ten clones known to carry the S_7 haplotype were aligned to the AB081587 sequence using the BWA-mem algorithm in the Burrows-Wheeler alignment (BWA-0.6) tool[56]. The resulting binary alignment/map (BAM) files were visualised using Tablet graphical viewer version 1.16.09.06[57].

Trimmed reads from each clone were assembled using Mimicking Intelligent Read Assembler (MIRA) version 4.0.2[58]. The resulting contig sequences were mapped to the AB081587 sequence and visualised using CONTIGuator software[59]. 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[60]. Polymorphisms were graphically visualised using Integrated Genomic Viewer version 2.3[61].

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[62] 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_1S_7, S_7S_8 and S_7S_f clones were then compared to the overall consensus S_7 sequence using the variant call format in VCFtools v0.1.13[63]. This provided haplotype sequences for S_1, S_8, and S_f, 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[64] tool on the Phylogeny.fr online server (www.phylogeny.fr) with the 'less stringent' data selection setting.

Protein-coding sequences of the S_1, S_7, S_8, and S_f haplotypes were predicted and analysed with BLASTX 2.8.[65] using the refseq protein database, and with GENSCAN[66] using Arabidopsis (*Arabidopsis thaliana* L.) as the reference. LTR retrotransposons were detected with LTR_Finder version 1.0.5[67] using an Arabidopsis tRNA database (http://lowelab.ucsc.edu/GtRNAdb/) 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)[10]. 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 $1e^{-5}$) with the query sequences of *P. avium* SLFLs using the genome databases of *P. dulcis* Texas Genome v2.0[68] 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[69] with a 100 bp sliding window and 20 bp 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[70] and using Primer 3 software version 4.0[54]. Tail sequences complementary to the FRET cassettes 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 *Prunus dulcis* and S referring to the *S* locus, followed by a number or letter designating a target *S* allele (e.g. f for S_f). 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, Antoñeta, Carmel, Francolí, Johnston's Prolific, Lauranne, Mandaline, 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-[16] 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® Gene Fragments (Integrated DNA Technologies, Iowa, USA) were synthesised to represent segments of the *S-RNase* alleles S_f, S_1, S_3, S_5, S_7, S_8, S_9, S_{23} and S_{25} (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 with more than two clusters, *post hoc* comparisons among clusters were conducted using pairwise Tukey contrasts, as implemented in the R package MANOVA.RM[71]. 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, IUCG, 1V9H) were downloaded from the Protein Data Bank[72]. The sequences were aligned with the deduced protein sequences for almond *S-RNase* alleles using PSI-BLAST[73], PSIPRED v3.3[74] and RaptorX[75].

Five almond S-RNases (S_f, S_5, S_7, S_8, and S_{23}) were subjected to comparative protein modelling using Modeller V9.19[76]. 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[77] 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)[78] and discrete optimised protein energy (DOPE)[79] parameters were selected. Each model was optimised using FoldX4[80] and evaluated using ProSa 2003[81] and PROCHECK[82]. 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_3-RNase of Japanese pear (*Pyrus pyrifolia* (Burm.) Nak) (BAA93052.1)[83]. The positions of disulphide bonds in the S-RNases were identified by using the SSBOND

record of PDB files of ribonucleases and using protein sequences as input in the Disulfind online server (http://disulfind.dsi.unif.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://wolf.bms.umist.ac.uk/naccess/). Electrostatic potentials were calculated with the Adaptive Poisson-Boltzmann Solver[84] using the PyMOL plug-in APBS Tool2, with AMBER force field parameters[85] 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 investigated 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 (S_5, S_7, S_8, S_{23} and S_f) were aligned using PROMALS3D[86] (Supplementary Fig. S12) to identify positions at which the S_7-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 FoldX4[80] was used to mutate the S_7-RNase residue to the alternative residue (or residues), considering both single and multiple mutation options. A similar process was applied to the S_8-RNase, with residues replaced by those in the S_7-RNase. The energies for the wild-type (ΔG wild-type) and mutant (ΔG mutant) proteins were computed using FoldX4) 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 ($\Delta\Delta G < -1.84$ kcal/mol), (ii) stabilising (-1.84 kcal/mol $\leq \Delta\Delta G < -0.92$ kcal/mol), (iii) slightly stabilising (-0.92 kcal/mol $\leq \Delta\Delta G < -0.46$ kcal/mol), (iv) neutral (-0.46 kcal/mol $< \Delta\Delta G \leq +0.46$ kcal/mol), (v) slightly destabilising ($+0.46$ kcal/mol $< \Delta\Delta G \leq +0.92$ kcal/mol), (vi) destabilising ($+0.92$ kcal/mol $< \Delta\Delta G \leq +1.84$ kcal/mol), and (vii) highly destabilising ($\Delta\Delta G > +1.84$ kcal/mol).

The sequences of five S-RNases (S_5, S_7, S_8, S_{23} and S_f) were analysed using the NetNGlyc 1.0 server (http://www.dtu.dk/services/NetNGlyc/) and sequons 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_7-RNase, Asn and Xaa residues were replaced by the corresponding residues from the S_8-RNase, using the BuildModel command of FoldX4[80]. *O*-glycosylation sites in the same five S-RNases were identified using the DictyOGlyc 1.1 Server (http://www.cbs.dtu.dk/services/DictyOGlyc/), 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*-haplotype 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 *P. dulcis* and *P. webbii*: detection and cloning of 'wildtype S_f' and new self-compatibility alleles encoding inactive *S*-RNases. *Mol. Gen. Genomics* **278**, 665–676 (2007).
3. Channuntapipat, C., Sedgley, M. & Collins, G. Sequences of the cDNAs and genomic DNAs encoding the *S1*, *S7*, *S8*, and *Sf* alleles from almond, *Prunus dulcis*. *Theor. Appl. Genet.* **103**, 1115–1122 (2001).
4. Halász, J., Fodor, Á., Pedryc, A. & Hegedüs, A. *S*-genotyping of Eastern European almond cultivars: identification and characterization of new (*S36 – S39*) selfincompatibility ribonuclease alleles. *Plant Breed* **129**, 227–232 (2010).
5. Hafizi, A., Shiran, B., Maleki, B., Imani, A. & Banović, 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., Batlle, 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-RNasebased 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 selfincompatibility system. *Plant Mol. Biol.* **100**, 367–378 (2019).
11. Grasselly, C. & Olivier, G. Mise en évidence de quelques types autocompatibles parmi les cultivars d'amandier (*P. amygdalus* Batsch) de la population des Pouilles. *Ann. Amélio. Plantes* **26**, 107–113 (1976).
12. Socias 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., Socias 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_f allele. *J. Am. Soc. Hortic. Sci.* **134**, 221–227 (2009).
14. Fernández i Martí, À. *et al.* The almond S_f haplotype shows a double expression despite its comprehensive genetic identity. *Sci. Hort.* **125**, 685–691 (2010).
15. Fernández i Martí, A., Gradziel, T. & Socias i Company, R. Methylation of the S_f allele in almond is associated with S-RNase loss of function. *Plant. Mol. Biol.* **86**, 681–689 (2014).
16. Kodad, O., Socias i Company, R. & Alonso, J. M. Unilateral recognition of the S_f allele in almond. *Sci. Hort.* **185**, 29–33 (2015).

17. Ramanauskas, K. & Igić, B. The evolutionary history of plant T2/S-type ribonucleases. *PeerJ* **5**, e3790, https://doi.org/10.7717/peerj.3790 (2017).
18. Fox, N. K., Brenner, S. E. & Chandonia, J. M. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–309 (2014).
19. Matsuura, T. *et al*. Crystal structure at 1.5-Å resolution of *Pyrus pyrifolia* pistil ribonuclease responsible for gametophytic self-incompatibility. *J. Biol. Chem.* **276**, 45261–45269 (2001).
20. 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 (2004).
21. Jaskolski, M., Alexandratos, J. N., Bujacz, G. & Wlodawer, A. Piecing together the structure of retroviral integrase, an important target in AIDS therapy. *FEBS J* **276**, 2926–2946 (2009).
22. Nesmelova, I. V. & Hackett, P. B. DDE transposases: Structural similarity and diversity. *Adv. Drug Deliv. Rev.* **62**, 1187–1195 (2010).
23. Halász, J., Kodad, O. & Hegedűs, A. Identification of a recently active *Prunus*-specific non-autonomous Mutator element with considerable genome shaping force. *Plant J.* **79**, 220–231 (2014).
24. Tao, R. *et al*. Self-compatible peach (*Prunus persica*) has mutant versions of the *S* haplotypes found in self-incompatible *Prunus* species. *Plant Mol. Biol.* **63**, 109–123 (2007).
25. Channuntapipat, C. *et al*. Identification of incompatibility genotypes in almond (*Prunus dulcis* Mill.) using specific primers based on the introns of the Salleles. *Plant Breed.* **122**, 164–168 (2003).
26. Ortega, E., Bošković, R., Sargent, D. & Tobutt, K. Analysis of *S*-RNase alleles of almond (*Prunus dulcis*): characterization of new sequences, resolution of synonyms and evidence of intragenic recombination. *Mol. Genet. Genomics* **276**, 413–426 (2006).
27. Ortega, E., Sutherland, B. G., Dicenta, F., Boskovic, R. & Tobutt, K. R. Determination of incompatibility genotypes in almond using first and second intron consensus primers: detection of new *S* alleles and correction of reported *S* genotypes. *Plant Breed.* **124**, 188–196 (2005).
28. Sánchez-Pérez, R., Dicenta, F. & Martínez-Gómez, P. Identification of *S*-alleles in almond using multiplex PCR. *Euphytica* **138**, 263–269 (2004).
29. Gómez, E. M., Prudencio, A. S., Dicenta, F. & Ortega, E. Characterization of S_f-RNase and SFB_f adjacent regions in the *S*-locus of almond. *Acta Hortic* **1231**, 105–108 (2019).
30. Hanada, T. *et al*. Cloning and characterization of a self-compatible *S* haplotype in almond [*Prunus dulcis* (Mill.) D.A. Webb. syn. *P. amygdalus* Batsch] to resolve previous confusion in its *S -RNase* sequence. *HortScience* **44**, 609–613 (2009).
31. Mercier, R., Mezard, C., Jenczewski, E., Macaisne, N. & Grelon, M. The molecular biology of meiosis in plants. *Annu. Rev. Plant Biol.* **66**, 297–327 (2015).
32. Lambing, C., Franklin, F. C. H. & Wang, C.-J. R. Understanding and manipulating meiotic recombination in plants. *Plant Physiol.* **173**, 1530–1542 (2017).
33. Ishimizu, T., Shinkawa, T., Sakiyama, F. & Norioka, S. Primary structural features of rosaceous S-RNases associated with gametophytic self-incompatibility. *Plant Mol. Biol* **37**, 931–941 (1998).
34. Gu, C. *et al*. Characterization of the *S-RNase* genomic DNA allele sequence in *Prunus speciosa* and *P. pseudocerasus*. *Sci. Hort.* **144**, 93–101 (2012).
35. Newbigin, E., Paape, T. & Kohn, J. R. RNase-based self-incompatibility: puzzled by *pollen S*. *Plant Cell* **20**, 2286–2292 (2008).
36. Fernández i Marti, Á., Wirthensohn, M., Alonso, J., Socias i Company, R. & Hrmova, M. Molecular modelling of S-RNases involved in almond selfincompatibility. *Front. Plant Sci.* 3 (2012).
37. Ashkani, J. & Rees, D. J. G. A comprehensive study of molecular evolution at the self-incompatibility locus of Rosaceae. *J. Mol. Evol.* **82**, 128–145 (2016).
38. Vieira, J., Ferreira, P., Aguiar, B., Fonseca, N. & Vieira, C. 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).
39. Ishimizu, T., Norioka, S., Kanai, M., Clarke, A. E. & Sakiyama, F. Location of cysteine and cystine residues in *S*-ribonucleases associated with gametophytic self-incompatibility. *Eur J Biochem* **242**, 627–635 (1996).
40. Arolas, J. L., Aviles, F. X., Chang, J. Y. & Ventura, S. Folding of small disulfide-rich proteins: clarifying the puzzle. *Trends Biochem. Sci.* **31**, 292–301 (2006).
41. Kurihara, H. *et al*. The crystal structure of ribonuclease Rh from *Rhizopus niveus* at 2.0 Å resolution. *J. Mol. Biol.* **255**, 310–320 (1996).
42. Nakagawa, A. *et al*. Crystal structure of a ribonuclease from the seeds of bitter gourd (*Momordica charantia*) at 1.75 Å resolution. *Biochim. Biophys. Acta* **1433**, 253–260 (1999).
43. Royo, J. *et al*. Loss of a histidine residue at the active site of *S*-locus ribonuclease is associated with self-compatibility in *Lycopersicon peruvianum*. *Proc. Natl. Acad. Sci. USA* **91**, 6511–6514 (1994).
44. Ishimizu, T. *et al*. Identification of histidine 31 and cysteine 95 in the active site of self-incompatibility associated *S6* -RNase in *Nicotiana alata*. *J. Biochem.* **118**, 1007–1013 (1995).
45. Tanaka, N. *et al*. Crystal structure of a plant ribonuclease, RNase LE. *J. Mol. Biol.* **298**, 859–873 (2000).
46. Irie, M., Harada, M. & Sawada, F. Studies on the state of tryptophan residues in ribonuclease from *Aspergillus saitoi*. *J. Biochem* **72**, 1351–1359 (1972).
47. Lima, W. F. *et al*. Human RNase H1 uses one tryptophan and two lysines to position the enzyme at the 3′-DNA/5′-RNA terminus of the heteroduplex substrate. *J. Biol. Chem.* **278**, 49860–49867 (2003).
48. Strasser, R. Plant protein glycosylation. *Glycobiology* **26**, 926–939 (2006).
49. Matton, D. P. *et al*. Production of an S RNase with dual specificity suggests a novel hypothesis for the generation of new *S* alleles. *Plant Cell* **11**, 2087–2097 (1999).
50. Matton, D. P. *et al*. Hypervariable domains of self-incompatibility RNases mediate allele-specific pollen recognition. *Plant Cell* **9**, 1757–1766 (1997).
51. Verde, I. *et al*. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet* **45**, 487–494 (2013).
52. Kearse, M. *et al*. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
53. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–1291 (2007).
54. Untergasser, A. *et al*. Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
55. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
57. Milne, I. *et al*. Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
58. Chevreux, B. *et al*. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147–1159 (2004).
59. Galardini, M., Biondi, E. G., Bazzicalupo, M. & Mengoni, A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol. Med* **6**, 11, https://doi.org/10.1186/1751-0473-6-11 (2011).
60. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Robinson, J. T. *et al*. Integrative genomics viewer. *Nat. Biotech* **29**, 24–26 (2011).

62. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680 (1994).
63. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
64. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol* **17**, 540–552 (2000).
65. Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
66. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
67. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
68. Alioto, T. et al. Transposons played a major role in the diversification between the closely related almond and peach genomes; results from the almond genome sequence. *Plant J*., https://doi.org/10.1111/tpj.14538 (2019).
69. Rozas, J. *et al*. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol* **34**, 3299–3302 (2017).
70. He, C., Holmes, J. & Anthony, J. SNP genotyping: the KASP assay. *Methods. Mol. Biol* **1145**, 75–86 (2014).
71. Friedrich, K., Konietschke, F. & Pauly, M. MANOVA.RM: Analysis of multivariate data and repeated measures designs. *R Package Version 0.3.2* (2019).
72. Berman, H. M. *et al*. The protein data bank. *Nucleic Acids Res* **28**, 235–242 (2000).
73. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
74. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
75. Källberg, M. *et al*. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **7**, 1511–1522 (2012).
76. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
77. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).
78. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).
79. Eswar, N., Eramian, D., Webb, B., Shen, M.-Y. & Sali, A. Protein structure modelling with MODELLER in *Structural proteomics: high-throughput methods* (eds Kobe B., Guss, M. & Huber, T.) 145–159 (Humana Press, 2008).
80. Schymkowitz, J. W. H. *et al*. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl. Acad. Sci. USA* **102**, 10147–10152 (2005).
81. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct., Funct., Bioinf* **17**, 355–362 (1993).
82. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. App. Crystall* **26**, 283–291 (1993).
83. Norioka, N. *et al*. Sequence comparison of the 5' flanking regions of Japanese pear (*Pyrus pyrifolia*) S-RNases associated with gametophytic selfincompatibility. *Sex. Plant Reprod.* **13**, 289–291 (2001).
84. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W665–667 (2004).
85. Weiner, P. K. & Kollman, P. A. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Comput. Chem.* **2**, 287–303 (1981).
86. Pei, J., Kim, B.-H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, 2295–2300 (2008).

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 www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020