Characterization of 25 full-length S-RNase alleles, including flanking regions, from a pool of resequenced apple cultivars

Characterization of 25 full-length *S-RNase* alleles, including flanking regions, from a pool of resequenced apple cultivars

Paolo De Franceschi[1] (corresponding author), Luca Bianco[2], Alessandro Cestaro[2], Luca Dondini[1], Riccardo Velasco[2,3]

[1] Dipartimento di Scienze e Tecnologie Agroalimentari (DISTAL), Università degli Studi di Bologna, Bologna, Italy

[2] Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy

[3] Present address: Centro di Ricerca in Viticoltura ed Enologia del Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA-VE), Conegliano, Treviso, Italy

e-mail: paolo.defranceschi2@unibo.it - Telephone number: +39-0512096422

Key message

Data obtained from Illumina resequencing of 63 apple cultivars were used to obtain full-length *S-RNase* sequences using a strategy based on both alignment and *de novo* assembly of reads.

Abstract

The reproductive biology of apple is regulated by the *S*-RNase-based gametophytic self-incompatibility system, that is genetically controlled by the single, multi-genic and multi-allelic S locus. Resequencing of apple cultivars provided a huge amount of genetic data, that can be aligned to the reference genome in order to characterize variation to a genome-wide level. However, this approach is not immediately adaptable to the *S*-locus, due to some peculiar features such as the high degree of polymorphism, lack of colinearity between haplotypes and extensive presence of repetitive elements. In this study we describe a dedicated procedure aimed at characterizing *S-RNase* alleles from resequenced cultivars. The *S*-genotype of 63 apple accessions is reported; the full length coding sequence was determined for the 25 *S-RNase* alleles present in the 63 resequenced cultivars; these included 10 previously incomplete sequences (S_5, S_{6a}, S_{6b}, S_8, S_{11}, S_{23}, S_{39}, S_{46}, S_{50} and S_{58}). Moreover, sequence divergence clearly suggests that alleles S_{6a} and S_{6b}, proposed to be neutral variants of the same alleles, should be instead considered different specificities. The promoter sequences have also been analyzed, highlighting regions of homology conserved among all the alleles.

Keywords: self-incompatibility, *S*-locus, *Malus*, *S*-genotyping, apple genome

Acknowledgements: This work has been partly funded under the EU Seventh Framework Programme by FruitBreedomics project no. 265582: Integrated Approach for Increasing Breeding Efficiency in Fruit Tree Crop, and the Italian Ministry of Education, Universities and Research project PRIN 2015BPM9H3: Investigating Self Incompatibility DEterminants in fruit trees (ISIDE). We would like to thank Dr. Massimo Pindo and Dr. Erika Stefani (Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy) for technical support on sequencing reactions.

Introduction

The domesticated apple (*Malus × domestica* Borkh.) is one of the most important crops in temperate regions; it belongs to the family Rosaceae, tribe Pyreae and subtribe Pyrinae (Potter et al. 2007), together with pears (gen. *Pyrus*) and other cultivated tree species such as quince, loquat and medlar. All the Pyrinae have a distinctive base chromosome number x=17, which is considered to have arisen from an ancestor with x=9 through an event of autopolyploidization and subsequent aneuploidization (Evans and Campbell 2002). The center of origin of domesticated apple is thought to be in Kazakhstan and Central Asia (Vavilov 1930), but thanks to its great genetic variability it was adapted to cultivation in different environments ranging from subtropical to subarctic climates. A wide genetic base is maintained also due to self-incompatibility, which forces allogamy by preventing self-fertilization (de Nettancourt 2001).

Apple and several other Rosaceae possess the *S*-RNase-based gametophytic self-incompatibility system, which is considered the most widespread and ancient self-incompatibility system in eudicots (Steinbachs and Holsinger 2002). In this system, self-pollen tubes are recognized and rejected by the *S*-locus encoded, pistil-specific ribonuclease (*S*-RNase); compatible pollen tubes, on the other hand, have the ability to inactivate the *S-RNase*s through a pool of *S*-locus *F-box Brother* genes (*SFBB*s, Sassa et al. 2007) which are thought to bind it in an allele-specific way and mark it for proteolytic degradation by the ubiquitine-proteasome pathway (reviewed by De Franceschi et al. 2012). The recognition of all non-self *S*-RNase alleles is provided by a pool of different *SFBB*s acting collaboratively (Kubo et al. 2010) and being all encoded within the *S*-locus. The suppression of recombination allows the multi-genic *S*-haplotypes to be inherited as single segregating units, maintaining their integrity and functionality across generations (Wang et al. 2012). On a practical point of view, knowledge of *S*-genotypes of cultivars is fundamental both to establish productive cultivar combinations in orchards and to plan compatible crosses in breeding; therefore, PCR-based *S*-genotyping assays have been developed since the early characterization of the *S-RNase* gene (Janssens et al. 1995).

The genome of domesticated apple was sequenced for the first time from 'Golden Delicious' (Velasco et al. 2010) and subsequently from a doubled haploid derived from the same cultivar (Daccord et al. 2017). The initial polyploidization, the hybridization with wild relatives and the forced allogamy of apple resulted in a complex genome, in which syntenic portions of chromosomes reflect the rearrangements of the two ancestral genomes and a huge variability is maintained (Velasco et al. 2010). The availability of a genome sequence and modern high-throughput sequencing technologies make it possible to characterize the genetic variation of apples by resequencing cultivars and aligning them to the reference genome, also enabling the development of dense SNP-genotyping arrays which efficiently cover the whole genome (Chagné et al. 2012; Bianco et al. 2014; Bianco et al. 2016).

While this approach proved to be extremely effective in other genomic regions, the *S*-locus shows a series of features that greatly hamper this process. First of all, different *S*-haplotypes proved to be very variable in terms of number, position and orientation of genes (Minamikawa et al. 2010; Okada et al. 2011; Wang et al. 2012; Okada et al. 2013); this particular structure is the product of a complex evolutionary dynamics of *SFBB*s, in which gene duplication and horizontal transfer are expected to play a major role (De Franceschi et al. 2012; Kubo et al. 2015). The lack of colinearity between haplotypes, together with the massive presence of transposable elements, are the major hurdles when aligning short reads to a reference genome, as the haplotypes borne by the reference cultivar are not likely to show a common structure with accessions

having different *S*-alleles. The *S-RNase* gene, moreover, shows a degree of allelic polymorphism much higher than other genes; evidence of positive selection, favoring amino acid change and the rise of new alleles, has been reported by several studies (Ishimizu et al. 1998a; Vieira et al. 2007; Vieira et al. 2010; De Franceschi et al. 2011). Frequency-dependent balancing selection, on the other hand, preserved *S-RNase* alleles from being lost due to frequency fluctuations and genetic drift, maintaining allelic polymorphism generated before speciation in the Pyrinae (Sassa et al. 1996; Ishimizu et al. 1998a; Raspé and Kohn 2002; Dreesen et al. 2010; De Franceschi et al. 2011). As a result, many alleles are present in extant species: around 50 different *S-RNase*s have been reported in *Malus* (Kim et al. 2016) and the highest sequence homology can be found among alleles of different species or genera rather than within the same species. The single intron of the *S-RNase* has a length varying from around 100bp to more than 3kb (Takasaki-Yasuda et al. 2013). For all these reasons, a reference sequence does not always provide a good template for the alignment of resequencing reads, either at haplotype or single-gene level.

In the present study, we take advantage of the availability of genetic data from 63 previously resequenced apple cultivars (Bianco et al. 2016) to characterize the *S-RNase* gene and its flanking regions, using a dedicated procedure specifically developed to identify and assemble *S-RNase* allele sequences. The full-length sequence of 25 alleles is described, enabling efficient comparisons and resolving previous doubts about *S* specificities of grown apple cultivars.

Materials and methods

Plant material and resequencing

In a previous study, 65 apple accessions were resequenced using an Illumina HiSeq2000 platform with the aim of developing an Axiom®Apple480K SNP genotyping array (Bianco et al. 2016). The resequencing panel included 63 diploid apple cultivars chosen to cover the highest possible genetic diversity of apple germplasm and two doubled haploids; in the present study only the sequence reads obtained from the 63 diploid cultivars were used. Freeze-dried leaves used for DNA extraction were provided by various institutions as described by Bianco et al. (2016); the cultivar 'Mela Rozza' used for sequencing the allele S6 (see later in the text) is housed at the experimental station of Cadriano of the Department of Agricultural Sciences of the University of Bologna (Italy).

Synthetic Sequences

We developed a two-step approach based on the construction of synthetic sequences to determine the *S*-genotype of each cultivar. A first synthetic sequence was built using the information available in literature and on Genbank (https://www.ncbi.nlm.nih.gov/genbank/) and used to gain insights on the genotypes available through alignment of the resequencing data. The second synthetic sequence was built to extend the information by means of a local assembly phase.

Alignment of reads to known apple *S-RNases* and determination of S-genotypes

The first synthetic sequence contains 32 apple *S-RNase* alleles (see Table 1 for accession numbers) separated from each other by a spacer of 1000 'Ns'. Full-length genomic sequences were used when available, but for 13 out of 32 alleles only a portion of the coding sequence has been characterized (Table

1). For the S_8 allele, moreover, besides an incomplete coding sequence the intron was also lacking, given that the available sequence was obtained from mRNA. In this case an additional 1000 Ns spacer was manually inserted to separate the two exons. Alleles with multiple neutral variants, such as S_{16} (Matsumoto and Furusawa 2005), were included selecting a single variant to avoid redundancy in the alignment of short reads.

S-genotypes were determined by aligning resequencing data against the synthetic sequence. Reads were aligned with BFAST (Homer et al. 2009) keeping all best scoring matches and filtering results allowing 2 mismatches; this parameter was set after testing the strategy on the cultivar 'Cox's Orange Pippin', with known *S*-genotype, allowing 2 to 7 mismatches per read in different runs to determine the best setting to ensure allele-specificity.

After aligning and filtering, a coverage analysis was performed to determine the hypothetical alleles. The mean coverage was calculated separately for each exon using SAMtools (Li et al. 2009). *S*-genotypes were attributed assuming the presence of the two alleles with the highest coverage. We imposed, for calling the second allele, a coverage threshold of 30% with respect to the first one; when this threshold was not reached the presence of an unknown allele was assumed for that cultivar, and it was further analyzed with a different strategy (see "Identification of alleles not included in the synthetic sequence").

The strategy was repeated on all the 63 cultivars. Alignments were individually visualized and inspected using Tablet (Milne et al. 2013) to confirm the presence of each allele and to check for possible mismatches with the sequences used as input, as it may occur for neutral variants of the same allele.

De novo assembly of flanking regions

To obtain sequence information about flanking regions, sequencing reads had to be assembled *de novo* starting from the ends of known sequences. Targeted *de novo* assemblies were performed using GapFiller version 1.10 (available at https://www.baseclear.com/services/bioinformatics/basetools/gapfiller/, Boetzer and Pirovano 2012), making the target regions figure as a gap to fill. The second synthetic sequence was therefore built, in which each *S-RNase* allele identified through the alignment described above was flanked by two artificial 1000 N gaps, as follows:

A50 – N1000 – Allele 1 – N1000 – Allele 2 – N1000 – (…) – Allele n – N1000 – C50

The initial (A50) and final (C50) strings were inserted to provide an end for the terminal gaps flanking the first and last alleles. This synthetic sequence was used as template for GapFiller, so that each region flanking an allele in the template could be considered by the program as a gap to fill with Illumina reads, under the assumption that only the gaps flanking the two alleles borne by an accession could be filled with reads coming from that accession. Two changes were made with respect to the sequences used in the alignment step: the allele S_{16a} was substituted with S_{16b} (Acc. Number: AF327222 mRNA, full length cds, Van Nerum et al. 2001; AB428430 genomic, partial cds, Kim et al. 2009), as this is the variant actually detected in all the six cultivars in which S16 was detected; and S_{6a} was substituted by S_{6b} (Acc. Number AB094493 genomic, partial cds, Matsumoto et al. 2003) in the template to be used for the two cultivars which proved to possess this variant. Again, an artificial gap was also placed to separate the two S_8 exons as the intron sequence was unknown. Initially, the S_9-*RNase* allele was chosen to test the strategy as its flanking regions were sequenced from a BAC library from the cultivar 'Florina' and available at NCBI GeneBank database

under the accession number AB270792 (Sassa et al. 2007). GapFiller was run separately on each accession possessing the *S_9-RNase* allele, setting the minimum overlap length to 20 and 3 iterations per individual. The *S_9* region was selected from the GapFiller output for each genotype individually, and the obtained sequences were aligned and compared to the known *S_9* genomic region of 'Florina' using ClustalOmega (Sievers et al. 2011). The same approach was then applied to all the 63 cultivars, increasing the number of iterations per individual when necessary to increase the length of the output sequence. The output sequence for each allele carried by each single cultivar was extracted and all the obtained sequences of the same allele were aligned together using ClustalOmega.

Identification of alleles not included in the synthetic sequence

For those cultivars showing only one clearly identifiable *S-RNase* allele, the alignment of reads to the initial synthetic sequence was filtered increasing the number of allowed mismatches per read to 4; the output was then visually inspected to find reads aligning to alleles other than the first one. The reads were isolated, aligned to known alleles using BlastN (Altschul et al. 1990) and manually assembled in short contigs, which were then used to build a separate template for GapFiller as described above, running 10 iterations for each individual plus 10 additional iterations when the output sequence was not including the full-length coding sequence of the *S-RNase* allele. The sequences obtained were aligned to known allele using BlastN to find homologies and to determine the putative coding sequence.

Experimental validation of *in silico* assemblies

Genomic DNA was extracted from freeze-dried leaves of selected cultivars using a standard CTAB (cetyltrimethylammonium bromide) protocol (Maguire et al. 1994) and diluted to a final concentration of 50ng/µl.

RNA extraction was performed from the styles of 10 flowers collected from the cultivar 'Mela Rozza' at full bloom, using the protocol described by Zamboni et al. (2008). cDNA was synthesized from 100 ng of total RNA using SuperScript II Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA) following the manufacturer's instructions. To enable 3' RACE (rapid amplification of cDNA ends) a tailed oligo(dT) primer was used (5'-gactcgagtcgacatcga-t17-3') according to the protocol described in Sambrook et al. (1989). The 3' end of cDNA was then amplified using a gene-specific forward primer and a reverse one designed to anneal on the tail generated by reverse transcription.

Primer pairs (Online Resource 1) were designed on the obtained sequences using Primer3 (Untergasser et al. 2012) and PCR was performed in 20µl mixtures containing 1X PCR buffer, 2.5 mM MgCl2, 0.2 mM of each dNTP, 0.3 mM of each primer, 0.03 U/µl Amplibiotherm DNA polymerase (Fisher Molecular Biology, Rome, Italy) and 50ng genomic DNA or 1 µl of the reverse transcription mixture diluted 1:10 as template. The thermal profile included an initial denaturation at 95°C for 3' followed by 30 cycles of denaturation at 95° C for 20'', annealing at 58°C (54°C when the primer MdProm-box1f was used, Online Resource 1) for 30'', extension at 72°C for 1'; and final extension at 72°C for 8'. For cloning and sequencing, primer pairs were designed to amplify the largest possible portion of the sequence including flanking regions; amplicons sized 1810 and 1348bp were obtained for *S_{23}* and *S_{58}*, respectively; for the allele S8, however, given the large size of the obtained sequence (4098 bp) we decided to divide it in three partially overlapping chunks (MdS8_5UTR/MdS8_Ex1rev, 1297bp; MdS8_Ex1for/PycomC5r1, 1642bp; and MdS8_Ex2for/MdS8_3UTR, 1417bp) covering a total of 3983bp.

PCR products were cloned into pGEM-T Easy vectors (Promega, Madison, WI, USA) following manufacturer's instructions and three independent colonies were isolated for each fragment. Plasmids were extracted using GeneJet Plasmid Miniprep Kit (Thermo Scientific, Waltham, MA, USA) and sequenced using universal primers T7 and M13rev, plus consensus primers PycomC1f1 and PycomC5r1 (Sanzol 2009a) as internal oligos when necessary to sequence very large inserts.

Sequence analysis

The 5' and 3' flanking regions obtained for each *S-RNase* allele were aligned using Kalign (Lassmann and Sonnhammer 2005). The promoter regions were analyzed using TSSPlant (Shahmuradov et al. 2017) to determine the putative transcript start site (TSS) and TATA box. Coding regions and deduced proteins were aligned using ClustalW (Larkin et al. 2007) and a phylogenetic tree was built using the Neighbor-Joining (NJ) algorithm as implemented in MEGA6 (Tamura et al. 2013), using complete deletions and the maximum composite likelihood calculated using the three codon positions. Statistical support for the topologies was obtained by bootstrap analysis with 1000 replications.

Results and discussion

Determination of the S-genotype of resequenced cultivars

Illumina reads of 63 apple cultivars (Bianco et al. 2016) were aligned to a synthetic sequence containing 32 known *S-RNase* alleles with the purpose of identifying the *S-RNase* alleles specific of each cultivar. This pool of 32 known alleles included all the alleles reported by Matsumoto (2014) with the exclusion of S_{18}, for which no sequence is available; the alleles S_{41}, S_{42}, S_{44}, S_{45}, S_{46}, S_{50} and S_{53} were added because previously characterized in apple cultivars (De Franceschi et al. 2016) or known to be present in the Italian apple germplasm from previous analyses carried out in our laboratory (unpublished data). This pool is expected to cover most of the allelic variability of domesticated apple at this locus based on recent reports (Matsumoto 2014; Kim et al. 2016; Larsen et al. 2016); however, *S-RNase*s from related *Malus* species could also be present in the analyzed apple germplasm (Dreesen et al. 2010; Halász et al. 2011), as well as possible new, still uncharacterized alleles.

We assumed that under the appropriate filtering parameters, reads from each cultivar would selectively cover the positions on the synthetic sequence corresponding to the alleles present in its genome; as all cultivars are diploid and forcedly heterozygous at the *S*-locus, we expected to find two matches for each genotype, both with a similar coverage given that the two alleles are present in a 1:1 ratio. We tested our system using a cultivar of known *S*-genotype, 'Cox's Orange Pippin' (S_5S_9; Janssens et al. 1995; Bošković and Tobutt 1999) filtering with a maximum number of allowed mismatches ranging from 2 to 7 (i.e. from 2% to 7% of the read). Importantly, the filtering of reads had to be strict enough to preserve allele-specificity, but relaxed enough to detect possible variants or point mismatches in the sequence of an allele. A possible source of error in this approach is the presence of repetitive elements in *S-RNase* introns, which could be highly represented between reads and result in a high coverage even when the rest of the allele sequence is not covered; this was the case, in the first runs from 'Cox's Orange Pippin', of alleles S_3, S_{10}, S_{16}, S_{25}, S_{42} and S_{45}, which showed a very high, non-allele specific coverage within introns (data not shown). On the other hand, only exons from S_5 and S_9 yielded a good coverage under all the tested conditions. These

preliminary results highlighted that non-specific matching of reads on intron sequences could lead to alleles mis-identification; all the subsequent coverage analyses, therefore, were carried out on the two exons separately not taking into consideration reads aligning to introns.

The best results from the test cultivar 'Cox's Orange Pippin' were obtained setting the maximum allowed mismatches per read to 2; the hits were distributed only on the two expected alleles and the mean coverage for each of their exons ranged from 10.2 to 12.5, consistently with a >20-fold genome coverage and a 1:1 ratio of the two alleles (Online Resource 2). The same settings were extended to the alignment of reads from all the other cultivars. The coverage obtained for the exons of all the 42 template *S-RNase*s is reported on Online Resource 2. Ideally, a 10X coverage on each allele would be expected for a 20X genome coverage; however, the number of reads obtained from each accession and therefore the coverage varied quite significantly, as reported by Bianco et al. (2016).

The designed synthetic sequence was suitable for defining 2 alleles for 59 of 63 cultivars (Online Resource 2), while only one allele was detected for the remaining 4 cultivars, which indicated that the second one was not included in the synthetic sequence. In total 22 different *S-RNase* alleles were detected in the analyzed accessions. All the six cultivars possessing allele S_{16} featured the S_{16b} variant, while among the three matching allele S_6, one ('Mela Rozza') possessed S_{6a} and two ('Heta' and 'Maikki') S_{6b}. The lack of significant coverage on more than two alleles for diploid cultivars reflects the accuracy of this alignment strategy for defining the S-genotypes from resequencing data; we could not test this approach with triploid accessions, as none was present in the resequenced pool, but these results suggest that it could be equally effective even with higher ploidy levels, provided that a sufficient genome coverage is obtained.

To our knowledge, the *S*-genotype of 14 of the 63 accessions was analyzed in previous studies; 11 of these are confirmed by our results, namely 'Braeburn' (S_9S_{24}), 'Cox's Orange Pippin' (S_5S_9), 'Delicious' (S_9S_{28}), 'Filippa' (S_7S_{24}), 'Fuji' (S_1S_9), 'Jonathan' (S_7S_9), 'Lady Williams' (S_7S_{23}), 'Macoun' (S_3S_{25}), 'McIntosh' ($S_{10}S_{25}$), 'Spässerud' (S_1S_7) and 'Åkerö' ($S_1S?$) for which a single *S*-allele was reported (Janssens et al. 1995; Sassa et al. 1996; Matsumoto et al. 1999; Van Nerum et al. 2001; Kitahara and Matsumoto 2002a; Kitahara and Matsumoto 2002b; Broothaerts et al. 2004; Garkava-Gustavsson et al. 2008; Nybom et al. 2008; Matsumoto 2014). In three cases, however, the alleles detected did not match the previously reported *S*-genotype : 'Borowitsky' (syn. 'Oldenburg') was reported as S_3S_7 (Long et al. 2010a) while we detected alleles S_3S_{28}; 'Priscilla' was reported as S_3S_9 (Broothaerts et al. 2004) or S_9S_{20} (Morita et al. 2009), but our results indicate its genotype to be S_7S_{10}; and finally 'Worcester Pearmain', previously reported as S_2S_{24} (Kitahara et al. 2000), resulted $S_{24}S_{25}$. In all these cases the coverage on the detected alleles was good (>12 for 'Borowitsky'; >11 for 'Priscilla'; >10 for 'Worchester Pearmain', Online Resource 2); the reads aligning to each allele were isolated from each genotype and assembled to a contig, which confirmed to correspond to the detected *S-RNase* allele after the alignment to the corresponding reference sequence (Online Resource 3). We therefore propose to correct the *S*-gentypes of these three cultivars accordingly.

De novo assembly of allele sequences and flanking regions

PCR-based *S*-genotyping assays, in apple as well as in other species, require *a priori* sequence information; most methods currently rely on the use of allele-specific primers, restriction enzymes or a combination of both methods for the identification of *S-RNase* alleles (De Franceschi et al. 2012 and references therein; Larsen et al. 2016). However, not all the known alleles have been sequenced completely. Full-length sequences are available for 19 of the 32 alleles used in our synthetic sequence (Table 1); among the

remaining alleles, 2 are incomplete in the first exon (S_5 and S_{44}), 9 in both exons (S_6, S_{11}, S_{21}, S_{23}, S_{29}, S_{42}, S_{45}, S_{46}, S_{50} and S_{53}) and one (S_8) lacks the intron and part of the first exon. Extending the sequence information is important especially for those alleles, like S_6, for which only a small portion of the coding sequence is known. The characterization of flanking regions, the promoter and terminator, is also desirable to improve our knowledge on the *S-RNase* gene and open new possibilities for molecular assays.

The high-coverage whole-genome resequencing of 63 cultivars yielded a number reads in which virtually all positions in the genome are expected to be represented; these reads have been aligned against the apple reference genome (Velasco et al. 2010) enabling the study of genetic variation within this species to a level previously unexplored (Bianco et al. 2016). When it comes to the *S-RNase*, however, this approach is not immediately effective due to the extremely high degree of polymorphism of this gene (Ma and Oliveira 2002; Vieira et al. 2007; Vieira et al. 2010; De Franceschi et al. 2011), which hampers the process of aligning reads to a reference sequence especially outside the coding region (Ushijima et al. 1998). Therefore, *S-RNase* coding and flanking regions are indeed present in the resequencing output, but cannot immediately be obtained by simply aligning to a reference sequence.

To exploit this amount of information fragmented into 100bp reads we envisaged a strategy based on a targeted *de novo* assembly. This operation is routinely applied in genomic studies to fill sequence gaps in genome assemblies, using programs such as GapFiller (Boetzer and Pirovano 2012); we reasoned that the same program could help in reconstructing unknown *S-RNase* regions, if they are treated as gaps in an artificial sequence. To test this possibility, we took advantage of the availability of the genomic sequence surrounding the S_9-*RNase*, which was obtained by a BAC library and deposited on GenBank (Sassa et al. 2007). An initial testing template was built by using the S_9-*RNase* sequence from the start to the stop codon (including intron) and creating two artificial gaps of 1000 nucleotides at both sides. Gapfiller was then run using this template from all the cultivars which proved to possess the allele S_9; 10 initial iterations resulted in the elongation of the sequence by 140 to 633bp at the 5' side in the different cultivars (Fig. 1a); this variation is likely due to the different coverage obtained from the accessions. On the opposite side, a much shorter sequence was obtained at the 3' end: in this case, assembly of reads was stopped by the presence of a >70bp microsatellite starting 16 bp downstream the stop codon (Fig. 1b). The newly assembled portions were aligned to the known S_9 region, and proved to be perfectly matching (Fig. 1).

This initial test showed that the approach with GapFiller could be effective in assembling reads to extend known allele sequences, but also highlighted that the low coverage obtained from some cultivars and the presence of repetitive DNA close to the coding region are two major problems of this strategy. A template containing all the 22 *S-RNase* alleles detected in the pool of cultivars was then used with GapFiller from all the accessions. After the GapFiller run, all the extended sequences were isolated and the output for the same allele were aligned to check their consistency; importantly, we did not observe any aspecific allele extension, i.e. only the two alleles present in a cultivars were extended by the program (data not shown). The procedure was repeated for those cultivars in which the newly generated portion of sequence was <300 bp upstream the start or downstream the stop codon.

The results of this approach are summarized in Table 2. In most cases the program successfully assembled several hundred up to more than one thousand bp on both the 5' and 3' flanking regions. The worst results were observed on the S_6 (*a* and *b*) alleles, for which the full-length coding sequence could not be determined despite the several additional iterations tested. This poor result was likely a consequence of

the low coverage obtained from the cultivars 'Mela Rozza', 'Heta' and 'Maikki', also observed on the first alignment of reads (Online Resource 2). Five alleles (S_4, S_9, S_{32}, S_{33}, S_{46}) were extended on the 3' side for less than 100bp (Table 1), possibly due to the presence of microsatellites downstream the stop codon, as observed in the initial test with S9 and at least in 3 of these alleles (S_4, S_9, S_{46}). Finally, for the allele S50 the assembly was very satisfactory at the 3' end (1028bp downstream the stop codon) but only yielded a 14bp fragment of the 5' flanking region (Table 1); of the three cultivars possessing S_{50}, two had a medium-to-low coverage ('Aport Kuba' and 'Durello di Forlì', ranging from 3.4X to 7.1X on S50 exons on the initial alignment) while the third, 'Ag Alma', resulted covered >10X on this allele (Online Resource 2). Therefore, also considering the good results obtained on the 3'side, the inability to assemble the 5' flanking region was unlikely to be due to coverage problems and the reason for this result remains unclear.

Identification of new alleles

For 4 cultivars, namely for 'Åkerö', 'Antonovka Pamtorutka', 'Ijunskoe Ranee' and 'Young America', no second alleles were clearly found. An attempt to identify them was therefore made using reads that aligned with the relaxed parameters (4 mismatches).

After manual inspection, candidate reads aligning on the S_{50} allele were found in 'Åkerö' and chosen for a step of manual assembly that yielded a template for the gapfiller phase. A sequence of 1549 bp was obtained, with 2 exons of 252 and 432 bp separated by an intron of 164 bp. Unfortunately, for 'Antonovka Pamtorutka', the limited number of reads did not allow for the assembly phase; however, the perfect match of Antonovka reads against the contig assembled from 'Åkerö' (data not shown) suggested that the two cultivars share the same allele; this finding was later confirmed by cloning and sequencing the allele from 'Antonovka Pamtorutka' (see next section).

The assembled sequence shows a considerable similarity with the S_{50}-*RNase* (93.1 % identity at nucleotide level in the coding region) but nevertheless it can be considered a different allele with a good confidence, as its deduced protein differs from S_{50} for 30 amino acids, showing a 86,8% sequence identity. A blast search of the nucleotide collection (nt) database highlighted a 96% nucleotide identity with S_{37} of *M. sylvestris* (NCBI GenBank Acc. Number EU419864, Dreesen et al. 2010); for this allele only a 200bp portion of the coding region is known, encoding a 66 amino acids sequence which shows 5 mismatches (92,4% identity) with the allele from 'Åkerö' and only 2 (97,0% identity) with S_{50}. On the other hand, a 98,7% identity both at the nucleotide and protein level was found with allele S_{121} from *Pyrus communis* (NCBI GenBank Acc. Number EU477839, Sanzol 2009b) suggesting the common origin of this allele before the divergence between the genera *Malus* and *Pyrus*, as frequently observed for *S-RNase*s (Ishimizu et al. 1998a; Raspé and Kohn 2002). While this manuscript was in preparation, an allele from the crabapple 'Mt Blanc' named S_{58} (Acc. Number MG262529) was released, that shows a high sequence identity with the allele identified in 'Åkerö' and 'Antonovka Pamtorutka'. The two sequences differ for 4 nucleotides, one in the first exon causing a conservative amino acid change (leucine to phenylalanine), one in the intron, and two silent substitutions in the second exon (Online Resource 4). We therefore named the new allele *M.* × *domestica* S_{58} and defined the *S*-genotypes of 'Åkerö' and 'Antonovka Pamtorutka' as S_1S_{58} and S_8S_{58}, respectively.

Using the same approach, hits partially matching S_5 were found for 'Ijunskoe Ranee' and 'Young America', isolated and assembled into sequences to be used as template for GapFiller; the two cultivars showed the presence of the same allele. The obtained sequence consists of two 252 and 438bp exons, a long 1446bp

intron, a 360bp 5' flanking region and 267bp downstream the stop codon. The alignment with known sequences using Blast revealed to be almost identical to the sequence of allele S_{39} from *Malus sylvestris* (NCBI GenBank Acc. Number EU419871, Dreesen et al. 2010); even in this case the full-length sequence is not available, but the 204bp portion of the coding sequence showed 3 single nucleotide mismatches with respect to the allele obtained from 'Ijunskoe Ranee' and 'Young America', 2 of which being synonymous substitutions and the third one causing a conservative amino acid change (methionine to isoleucine) (Online Resource 4). The two sequences thus showed a 98,5% identity both at the nucleotide and protein level; although the existence of more relevant differences in other portions of the protein cannot be ruled out, we considered the new allele as the *M. × domestica* variant of *M. sylvestris* S_{39}*-RNase* and defined the *S*-genotypes of for 'Ijunskoe Ranee' and 'Young America' as $S_{39}S_{46}$ and $S_{28}S_{39}$, respectively.

Validation and sequencing of uncharacterized regions

The initial analysis carried out on the S_9*-RNase* supported the accuracy of sequence assembly by the program GapFiller; however, we decided to validate the *in silico* sequence assembly by amplifying, cloning and sequencing a set of alleles including S_8 from 'Antonovka', S_{23} from 'Lady Williams' and S_{58} from 'Antonovka Pamtorutka'. These alleles were chosen among those for which leaf material was available, to include: an allele with a very large flanking region and intron portions assembled *in silico* (S_8); a previously unknown allele, whose sequence was entirely assembled *in silico* (S_{58}); and an additional control (S_{23}). The obtained sequences were aligned with those generated *in silico* by GapFiller, highlighting a 100% identity (Online Resource 5).

Additionally, we further investigated those alleles for which the assembly yielded shorter or incomplete portions, i.e. S_{50} and S_{6a}/S_{6b}. The sequence of S_{50} was extended by GapFiller by 1028bp on the 3' side, but only 14bp upstream the start codon. Interestingly, it resulted quite similar to that of S_{58}, as also highlighted by the partial coverage obtained from 'Åkerö' on S_{50} in the initial alignment (Online Resource 2). Therefore we combined the primer designed on the 5' region of S_{58} (MdS58_5UTR) with MdS50_3UTR designed on the 3' extreme of S_{50} (Online Resource 1), arguing that the high sequence similarity could allow the amplification of both alleles with the same primers. The amplification was tested on 'Durello di Forlì' (S_3S_{50}, Online Resource 2) and yielded a fragment sized 1745bp, which was cloned and sequenced and proved to include a 343bp fragment of the 5' flanking region. Again, the alignment with the *in silico* obtained sequence highlighted a 100% identity in the overlapping portion (Online Resource 5), confirming once again the accuracy of sequence assembly with GapFiller.

For the couple of alleles S_{6a}/S_{6b}, the low coverage obtained for the cultivars 'Mela Rozza', 'Heta' and 'Maikki' greatly hampered the assembly process, making it impossible to obtain the complete exon sequences (Table 2). We therefore defined a different strategy to amplify the full-length gene; on the 5' side, we used the consensus (non-allele specific) primer MdProm-box1f designed on a conserved motif found after the alignment of the promoters of all the other alleles (see the next section "Analysis of promoter sequences"), coupled with PycomC5r1 (Online Resource 1). Although none of these primers was specific for S_6, the amplification from 'Mela Rozza' and 'Heta' yielded two clearly distinct bands from each genotype, as in both cases the S_6 allele was coupled to a much larger one: S_{6a} with S_3 in 'Mela Rozza', S_{6b} with S_{16b} in 'Heta' (S_{6a} and S_{6b} amplified fragments sized about 1000bp, while amplicons from S_3 and S_{16b} were sized >2100bp and >3000bp respectively). After cloning the amplicons, colonies bearing the fragment of the correct allele could be easily distinguished by PCR and selected for sequencing, yielding in both cases

the entire 252bp first exon plus a 274bp and 273bp flanking region for S_{6a} and S_{6b}, respectively. The 3' side was amplified from 'Mela Rozza' cDNA through 3' Rapid Amplification of cDNA Ends (RACE); we designed a primer (MdS6a_Ex1for, Online Resource 1) which could selectively amplify S_{6a} excluding S_3, and coupled it to the primer designed on the oligo(dT) RACE adapter. This allowed us to amplify, clone and sequence a 994bp fragment including part of the first exon, the whole intron and second exon (167 and 435bp respectively) plus a 344bp 3' untranslated region (UTR). Interestingly, this latter portion highlighted a very large (230bp) microsatellite region of mixed TA and CA repeats. Three different reverse primers (MdS6a_3UTR1-3, Online Resource 1) were designed on the S_{6a} 3'UTR of 'Mela Rozza', one upstream and two downstream the microsatellite repeats, and tested on 'Heta' coupled with S6a_Ex1for in order to obtain amplification of the corresponding region of S_{6b}. Of these, however, only the upstream one (MdS6a_UTR1), annealing 10bp after the stop codon, proved to work on S_{6b} allowing the characterization of only 9bp of the 3'UTR of this latter allele, but including the complete second exon which made it possible to obtain the full-length coding sequence of the S_{6b} allele from 'Heta'.

Analysis of coding sequences and difference between S_6 (=S_{6a}) and S_{17} (=S_{6b})

Overall, the full length coding sequence was determined for all the 25 *S-RNase* alleles present in the 63 resequenced cultivars; these included 10 previously incomplete sequences (S_5, S_{6a}, S_{6b}, S_8, S_{11}, S_{23}, S_{39}, S_{46}, S_{50} and S_{58}). The nucleotide and deduced amino acid sequences were aligned and a Neighbor-Joining tree was built based on the protein alignment (Fig. 2, 3).

Surprisingly, the distance (number of amino acid substitutions per site/per sequence) between S_{6a} and S_{6b}-*RNase*s, which are proposed to be neutral variants of the same allele, was comparable to that between clearly distinct alleles (Table 3). The complete protein sequences of S_{6a} and S_{6b} differ for 14 amino acids (Fig. 2) out of 228 and their calculated distance is 0.063; two other couple of alleles differ for 14 amino acids, S_3/S_{10} and S_1/S_{20}, whose distances were 0.063 and 0.064 respectively. Similar but slightly higher differences were calculated for S_{20}/S_{24} (15 amino acids, 0.069), S_9/S_{46} (16 amino acids, 0.073) and S_1/S_{24} (17 amino acids, 0.078).

Before the present study, only partial genomic sequences were known for S_{6a} (previously also reported as S_{12}) and S_{6b} (S_{17} or S_{19}), limiting the possibility to compare their protein sequences to a portion of just 54 amino acid residues, showing a single difference; nevertheless, the two alleles were thought to encode the same *S*-specificity as they showed an identical hyper-variable region (Fig. 2), which is thought to be mainly responsible for *S*-specific protein interaction (Matsumoto et al. 2003; Morita et al. 2009; Matsumoto 2014). Later, however, it was proved that different *S-RNase* alleles can share identical hyper-variable regions, highlighting that other residues in the protein are indeed also involved in the determination of specificities (Zisovich et al. 2004). As no conclusive evidence supporting their functional identity has been provided so far, some authors considered separately the alleles S_{6a} and S_{6b}, maintaining the designations S_6 and S_{17} (Dreesen et al. 2010; Kim et al. 2016).

It is known that neutral variants of the same *S-RNase* allele coexist within a species; it is the case, for example, of apple S_{16a}/ S_{16b}/S_{16c} (Matsumoto and Furusawa 2005; Morita et al. 2009) or European pear $S_{104\text{-}1}$/$S_{104\text{-}2}$ (Sanzol 2010). In all these cases, however, sequence identity is much higher than within the pair S_{6a}/S_{6b}: S_{16a}, S_{16b} and S_{16c} encode identical proteins, while $S_{104\text{-}1}$ and $S_{104\text{-}2}$ differ for 2 amino acids; moreover , in both cases the functional identity of their *S*-specificities was proved after pollination tests, which to our knowledge have not been performed between S_{6a} and S_{6b}; only the identity between alleles previously

reported as S_{17} and S_{19}, which both correspond to S_{6b}, has been functionally proved (Matsumoto et al. 2006). On the other hand, alleles S_3 and S_5 from *Pyrus pyrifolia* differ for only 11 amino acids, yet encoding clearly distinct specificities (Ishimizu et al. 1998b).

The distribution of the 14 amino acid differences between S_{6a} and S_{6b} further supports a difference in their encoded specificities; although the first three substitutions (residues 3, 9 and 24; Fig. 2) fall in the signal peptide and can therefore be excluded from playing a role in the allele-specific interaction with pollen determinants, at least 5 of the remaining 11 replacements (residues 160 between C4 and C5; 212, 215, 216 and 224 after C5, Fig. 2) fall in sites previously reported to be under positive selection (Vieira et al. 2010), and therefore likely involved in determining *S*-specificity.

All these evidences strongly suggest that S_{6a} and S_{6b} encode distinct *S*-specificities, although a pollination test would be needed to clear any doubt. For this reason, we propose to rename S_{6a} as simply S_6, and S_{6b} as S_{17}; accordingly, the *S*-genotypes of 'Mela Rozza', 'Heta' and 'Maikki' are designated as S_3S_6, $S_{16b}S_{17}$ and $S_{10}S_{17}$, respectively. Among the sequences available in GenBank, AB094495 from 'Oetwiler Reinette', AB105061 and EU427461 from 'Citron d'Hiver' correspond to S_6, while AB094493 from 'Bohnapfel' and AB105062 from 'Blenheim Orange' match with S_{17} (sequences from Matsumoto et al. 2003; Dreesen et al. 2010).

The final *S*-genotype determined for all 63 cultivars is reported on table 4. Among the 25 different alleles detected, S_3 and S_1 were the most frequent, being detected in 17 and 13 cultivars respectively; S_6, S_{32} and S_{46} on the contrary appeared just in one individual (Fig. 4). All the sequences characterized in this study have been deposited in GenBank and are available under the accession numbers MG598487 (S_1), MG598488 (S_2), MG598489 (S_3), MG598490 (S_4), MG598491 (S_5), MG598492 (S_6), MG598493 (S_7), MG598494 (S_8), MG598495 (S_9), MG598496 (S_{10}), MG598497 (S_{11}), MG598498 (S_{16b}), MG598499 (S_{17}), MG598500 (S_{20}), MG598501 (S_{23}), MG598502 (S_{24}), MG598503 (S_{25}), MG598504 (S_{26}), MG598505 (S_{28}), MG598506 (S_{32}), MG598507 (S_{33}), MG598508 (S_{39}), MG598509 (S_{46}), MG598510 (S_{50}) and MG598511 (S_{58}).

Analysis of promoter sequences

The *S-RNase* gene is known to be specifically expressed in pistil tissues during flower development (Sassa et al. 1993; Ishimizu et al. 1996). The fine regulation of its expression is thought to depend on the presence of cis-acting elements in the promoter, which despite the strong allelic polymorphism must be present in the promoter of all *S-RNase*s (Norioka et al. 2000; Dissanayake et al. 2002). In the present study, 5' flanking regions containing at least part of the *S-RNase* promoter have been obtained for 25 different alleles; the complete alignment is provided as Online Resource 6, while the portion including the 800 positions upstream of the start codon is reported in Fig. 5. The sequences have been individually analyzed with TSSPlant (Shahmuradov et al. 2017), the results are reported in Table 5. For 20 out of 25 alleles the most likely transcript start site was found at a position ranging from -90 to -67, with a conserved TATA box at -123 to -100 (Fig. 5); a more distant TSS and TATA box was predicted for the allele S_{23}, while putative TATA-less promoters were identified for S_6, S_{17}, S_{20} and S_{24} (Table 5). Based on the sequence alignment (Fig. 5), however, the conserved TATA box identified in most alleles seems to be also present in S_6, S_{17}, S_{20} and S_{24} while a single substitution occurs in the same region of S_{23}. Despite the different predictions by TSSP for these alleles, all the analyzed *S-RNase* promoters might as well share a conserved structure and a common TSS; experimental evidences are needed to test this hypothesis.

Compared to solanaceous *S-RNase*s, the Pyrinae promoters show a higher sequence conservation; Solanaceae and Rosaceae however share a single conserved motif named IA-like (Ushijima et al. 1998). This motif falls within an approximately 200bp conserved region designated box1, that was described just upstream the TATA box in the promoters of apple *S-RNase* alleles S_1 and S_9, Japanese pear S_2, S_3, S_4 and S_5, and Chinese pear S_{12}, S_{13} and S_{21}, but not identified in apple S_2 (Ushijima et al. 1998; Norioka et al. 2000; Liu et al. 2012); our results show however that the S2-*RNase* promoter also contains the conserved box1 sequence, but it is more distant from the TATA box due to a 350bp insertion (Fig. 5). Box1 is thought to contain cis-acting regulatory elements driving the *S-RNase* expression; consistently with this hypothesis, transformation experiments performed in *Arabidopsis* with a reporter gene (GUS) under control of the Chinese pear S_{12}-*RNase* promoter carrying sequential deletions showed that a truncated box1 region results in the lack of expression in pistils (Liu et al. 2012). Other regulatory elements, however, must be present upstream box1, as sequential deletions resulted in a gradually decreasing reporter gene expression. A second conserved region named box2 has also been reported, but the degree of conservation is lower than box1; moreover, it was not detected in all promoters (Ushijima et al. 1998). Accordingly, sequence identity seems to fall quickly upstream box1 as visible on the alignment in Online Resource 6.

The presence of conserved motifs in the 5' flaking region can be exploited to design consensus PCR primers that would allow the amplification of *S-RNase* alleles including portion of the promoter and, most importantly, the entire first exon. Obtaining full-length coding sequences is fundamental not only to better characterize alleles, but also to make reliable comparisons, as highlighted by the case of S_{6a}/S_6 and S_{6b}/S_{17} which were believed to encode the same specificity because only a small portion of the protein-coding sequence was determined. We designed a primer on a short, highly conserved motif within box1, placed a few bases upstream the IA-like motif (Fig. 5); this primer, MdProm-box1f (5'-arggcabtgcacatga-3', Online Resource 1), contains two degenerations in order to include all possible variants within this 16bp region. Importantly, its sequence was found not only in all the apple promoters hereby obtained, but also in the Japanese and Chinese pear alleles so far characterized (Ushijima et al. 1998; Norioka et al. 2000; Liu et al. 2012). This primer coupled with a consensus reverse designed on exon 2 (PycomC5r) allowed us to amplify and clone alleles S_6 and S_{17} (see section "Validation and sequencing of uncharacterized regions" above). The same primer pair was also successfully tested on several genotypes of European, Japanese and Chinese pear (Online Resource 7) and wild species of *Malus* and *Pyrus* (data not shown), suggesting that it anneals efficiently on the majority, if not all, the promoters of apple and pear *S-RNase*s; the possibility to obtain amplification from more distant genera such as *Sorbus*, *Crataegus* and *Eriobotrya* has to be tested. We suggest the use of MdProm-box1f as a quick and cheap alternative to 5' RACE to characterize the 5' portion of *S-RNase*s. Unfortunately, the alignment of 3' flanking regions did not display similar clearly conserved motifs (Online Resource 8).

Conclusions

Using a combined alignment and assembly procedure, specifically devised for the analysis of the *S-RNase* gene from resequencing datasets, we were able to determine the *S*-genotype of 63 apple cultivars and to obtain the full-length sequence, including portions of the flanking regions, for 23 *S-RNase* alleles; traditional cloning and sequencing tests confirmed the accuracy of sequence assemblies. Two additional alleles, S_{6a} and S_{6b}, for which only a short portion of the coding sequence was available, were fully sequenced and proved to encode most likely different *S*-specificities; we therefore propose to rename S_{6a} as S_6 and S_{6b} as S_{17}.

The availability of full-length *S-RNase* sequences marks an important step forward in the study of Pyrinae *S-RNase*s and provides essential information for the development of *S*-genotyping assays. Conserved regions were found in all the promoters, making it possible to amplify and clone 5' flanking regions directly from genomic DNA of apple and pear species using consensus primers.

Literature cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410 . doi: 10.1016/S0022-2836(05)80360-2

Bianco L, Cestaro A, Linsmith G, Muranty H, Denancé C, Théron A, Poncet C, Micheletti D, Kerschbamer E, Di Pierro EA, Larger S, Pindo M, Van De Weg E, Davassi A, Laurens F, Velasco R, Durel CE, Troggio M (2016) Development and validation of the Axiom®Apple480K SNP genotyping array. Plant J 86:62–74 . doi: 10.1111/tpj.13145

Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S, Di Guardo M, Salvi S, Jansen J, Viola R, Gut I, Laurens F, Chagné D, Velasco R, Van De Weg E, Troggio M (2014) Development and validation of a 20K Single Nucleotide Polymorphism (SNP) whole genome genotyping array for apple (Malus × domestica Borkh). PLoS One 9:e110377 . doi: 10.1371/journal.pone.0110377

Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. Genome Biol 13:R56 . doi: 10.1186/gb-2012-13-6-r56

Bošković R, Tobutt KR (1999) Correlation of stylar ribonuclease isoenzymes with incompatibility alleles in apple. Euphytica 107:29–43 . doi: 10.1023/A:1003516902123

Broothaerts W, Van Neram I, Keulemans J (2004) Update on and review of the incompatibility (S-) genotypes of apple cultivars. HortScience 39:943–947

Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A, Velasco R, Main D, Rees JD, Iezzoni A, Mockler T, Wilhelm L, van de Weg E, Gardiner SE, Bassil N, Peace C (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS One 7:e31745 . doi: 10.1371/journal.pone.0031745

Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, Van De Geest H, Bianco L, Micheletti D, Velasco R, Di Pierro EA, Gouzy J, Rees DJG, Guérif P, Muranty H, Durel CE, Laurens F, Lespinasse Y, Gaillard S, Aubourg S, Quesneville H, Weigel D, Van De Weg E, Troggio M, Bucher E (2017) High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nat Genet 49:1099–1106 . doi: 10.1038/ng.3886

De Franceschi P, Cova V, Tartarini S, Dondini L (2016) Characterization of a new apple S-RNase allele and its linkage with the Rvi5 gene for scab resistance. Mol Breed 36:1–11 . doi: 10.1007/s11032-015-0427-x

De Franceschi P, Dondini L, Sanzol J (2012) Molecular bases and evolutionary dynamics of self-incompatibility in the Pyrinae (Rosaceae). J Exp Bot 63:4015–4032 . doi: 10.1093/jxb/ers108

De Franceschi P, Pierantoni L, Dondini L, Grandi M, Sansavini S, Sanzol J (2011) Evaluation of candidate F-box genes for the pollen S of gametophytic self-incompatibility in the Pyrinae (Rosaceae) on the basis of their phylogenomic context. Tree Genet Genomes 7:663–683 . doi: 10.1007/s11295-011-0365-7

de Nettancourt D (2001) Incompatibility and Incongruity in Wild and Cultivated Plants. Springer-Verlag,

560 Berlin, Germany

561 Dissanayake DMRKK, Norioka S, Norioka N, Takasaki T, Nakanishi T (2002) Cis-regulatory elements for pistil
562 specific expression in S-RNase promoter region of Japanese pear (Pyrus pyrifolia Nakai). Acta Hortic
563 587:459–465 . doi: 10.17660/ActaHortic.2002.587.60

564 Dreesen RSG, Vanholme BTM, Luyten K, van Wynsberghe L, Fazio G, Roldán-Ruiz I, Keulemans J (2010)
565 Analysis of Malus S-RNase gene diversity based on a comparative study of old and modern apple
566 cultivars and European wild apple. Mol Breed 26:693–709 . doi: 10.1007/s11032-010-9405-5

567 Evans RC, Campbell CS (2002) The origin of the apple subfamily (Maloideae; Rosaceae) is clarified by DNA
568 sequence data from duplicated GBSSI genes. Am J Bot 89:1478–1484 . doi: 10.3732/ajb.89.9.1478

569 Garkava-Gustavsson L, Kolodinska Brantestam A, Sehic J, Nybom H (2008) Molecular characterisation of
570 indigenous Swedish apple cultivars based on SSR and S-allele analysis. Hereditas 145:99–112 . doi:
571 10.1111/j.0018-0661.2008.02042.x

572 Halász J, Hegedus A, György Z, Pállinger É, Tóth M (2011) S-genotyping of old apple cultivars from the
573 Carpathian basin: Methodological, breeding and evolutionary aspects. Tree Genet Genomes 7:1135–
574 1145 . doi: 10.1007/s11295-011-0401-7

575 Homer N, Merriman B, Nelson SF (2009) BFAST: An alignment tool for large scale genome resequencing.
576 PLoS One 4:e7767 . doi: 10.1371/journal.pone.0007767

577 Ishimizu T, Endo T, Yamaguchi-Kabata Y, Nakamura KT, Sakiyama F, Norioka S (1998a) Identification of
578 regions in which positive selection may operate in S-RNase of Rosaceae: Implication for S-allele-
579 specific recognition sites in S-RNase. FEBS Lett 440:337–342 . doi: 10.1016/S0014-5793(98)01470-7

580 Ishimizu T, Sato Y, Saito T, Yoshimura Y, Norioka S, Nakanishi T, Sakiyama F (1996) Identification and partial
581 amino acid sequences of seven S-RNases associated with self-incompatibility of Japanese pear, Pyrus
582 pyrifolia Nakai. J Biochem 120:326–334 . doi: 10.1093/oxfordjournals.jbchem.a021417

583 Ishimizu T, Shinkawa T, Sakiyama F, Norioka S (1998b) Primary structural features of rosaceous S-RNases
584 associated with gametophytic self-incompatibility. Plant Mol Biol 37:931–941 . doi:
585 10.1023/A:1006078500664

586 Janssens GA, Goderis IJ, Broekaert WF, Broothaerts W (1995) A molecular method for S-allele identification
587 in apple based on allele-specific PCR. Theor Appl Genet 91:691–698 . doi: 10.1007/BF00223298

588 Kim H, Kakui H, Kotoda N, Hirata Y, Koba T, Sassa H (2009) Determination of partial genomic sequences and
589 development of a CAPS system of the S-RNase gene for the identification of 22 S haplotypes of apple
590 (Malus × domestica Borkh.). Mol Breed 23:463–472 . doi: 10.1007/s11032-008-9249-4

591 Kim HT, Moriya S, Okada K, Abe K, Park JI, Yamamoto T, Nou S (2016) Identification and characterization of
592 S-RNase genes in apple rootstock and the diversity of S-RNases in Malus species. J Plant Biotechnol
593 43:49–57 . doi: 10.5010/JPB.2016.43.1.49

594 Kitahara K, Matsumoto S (2002a) Cloning of the S25cDNA from “McIntosh” apple and an S25-allele
595 identification method. J Hortic Sci Biotechnol 77:724–728 . doi: 10.1080/14620316.2002.11511563

596 Kitahara K, Matsumoto S (2002b) Sequence of the S10 cDNA from “McIntosh” apple and a PCR-digestion
597 identification method. HortScience 37:187–190

598 Kitahara K, Soejima J, Komatsu H, Fukui H, Matsumoto S (2000) Complete sequences of the S-genes, Sd-

599 and Sh-RNase cDNA in apple. HortScience 35:712–715

600 Kubo KI, Entani T, Takara A, Wang N, Fields AM, Hua Z, Toyoda M, Kawashima SI, Ando T, Isogai A, Kao TH,
601 Takayama S (2010) Collaborative non-self recognition system in S-RNase-based self-incompatibility.
602 Science (80-) 330:796–799 . doi: 10.1126/science.1195243

603 Kubo KI, Paape T, Hatakeyama M, Entani T, Takara A, Kajihara K, Tsukahara M, Shimizu-Inatsugi R, Shimizu
604 KK, Takayama S (2015) Gene duplication and genetic exchange drive the evolution of S-RNase-based
605 self-incompatibility in Petunia. Nat Plants 1:140005 . doi: 10.1038/nplants.2014.5

606 Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm
607 A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0.
608 Bioinformatics 23:2947–2948 . doi: 10.1093/bioinformatics/btm404

609 Larsen B, Ørgaard M, Toldam-Andersen TB, Pedersen C (2016) A high-throughput method for genotyping S-
610 RNase alleles in apple. Mol Breed 36:1–10 . doi: 10.1007/s11032-016-0448-0

611 Lassmann T, Sonnhammer ELL (2005) Kalign - An accurate and fast multiple sequence alignment algorithm.
612 BMC Bioinformatics 6:298 . doi: 10.1186/1471-2105-6-298

613 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The
614 Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079 . doi:
615 10.1093/bioinformatics/btp352

616 Li T, Long S, Li M, Bai S, Zhang W (2012) Determination S-Genotypes and Identification of Five Novel S-
617 RNase Alleles in Wild Malus Species. Plant Mol Biol Report 30:453–461 . doi: 10.1007/s11105-011-
618 0345-y

619 Liu X ying, Wuyun T na, Zeng H yan (2012) Cloning, characterization and promoter analysis of S-RNase gene
620 promoter from Chinese pear (Pyrus pyrifolia). Gene 505:246–253 . doi: 10.1016/j.gene.2012.06.017

621 Long S, Li M, Han Z, Wang K, Li T (2010a) Characterization of three new S-alleles and development of an S-
622 allele-specific PCR system for rapidly identifying the S-genotype in apple cultivars. Tree Genet
623 Genomes 6:161–168 . doi: 10.1007/s11295-009-0237-6

624 Long S, Li M, Han Z, Zhang B, Wang K, Li T (2010b) Characterization of two novel S-RNase genes and PCR
625 analyzing of S-genotypes of 46 cultivars in Malus domestica Borkh. J Agric Biotechnol 18:265–271

626 Ma RC, Oliveira M (2002) Evolutionary analysis of S-RNase genes from Rosaceae species. Mol Genet
627 Genomics 267:71–78 . doi: 10.1007/s00438-002-0637-x

628 Maguire TL, Collins GG, Sedgley M (1994) A modified CTAB DNA extraction procedure for plants belonging
629 to the family proteaceae. Plant Mol Biol Report 12:106–109 . doi: 10.1007/BF02668371

630 Matsumoto S (2014) Apple pollination biology for stable and novel fruit production: Search system for
631 apple cultivar combination showing incompatibility, semicompatibility, and full-compatibility based on
632 the S-RNase allele database. Int J Agron 2014:1–9 . doi: 10.1155/2014/138271

633 Matsumoto S, Furusawa Y (2005) Genomic DNA sequence of S16c (= 16)-RNase in apple: re-numbering of
634 S16 (= 27a)-and S22 (= 27b)-allele to S16a and S16b. Sci Rep Rac Educ Gifu Univ (Nat Sci) 29:7–12

635 Matsumoto S, Furusawa Y, Kitahara K, Soejima J (2003) Partial Genomic Sequences of S6-, S12-, S13-, S14-,
636 S17-, S19-, and S21-RNases of Apple and Their Allele Designations. Plant Biotechnol 20:323–329 . doi:
637 10.5511/plantbiotechnology.20.323

638 Matsumoto S, Kitahara K, Komatsu H, Abe K (2006) Cross-compatibility of apple cultivars possessing S-
639 RNase alleles of similar sequence. J Hortic Sci Biotechnol 81:934–936 . doi:
640 10.1080/14620316.2006.11512178

641 Matsumoto S, Komori S, Kitahara K, Imazu S, Soejima J (1999) S genotypes of 15 apple cultivars and self-
642 compatibility of “Megumi.” J Japan Soc Hort Sci 68:236–241

643 Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shawand PD, Marshall D (2013) Using tablet for
644 visual exploration of second-generation sequencing data. Brief Bioinform 14:193–202 . doi:
645 10.1093/bib/bbs012

646 Minamikawa M, Kakui H, Wang S, Kotoda N, Kikuchi S, Koba T, Sassa H (2010) Apple S locus region
647 represents a large cluster of related, polymorphic and pollen-specific F-box genes. Plant Mol Biol
648 74:143–154 . doi: 10.1007/s11103-010-9662-z

649 Morita J, Abe K, Matsumoto S (2009) S-RNase genotypes of apple cultivars grown in Japan and
650 development of A PCR-RFLP method to identify the S 6-and S 21-RNase alleles. J Hortic Sci Biotechnol
651 84:29–34 . doi: 10.1080/14620316.2009.11512475

652 Norioka N, Katayama H, Matsuki T, Ishimizu T, Takasaki T, Nakanishi T, Norioka S (2000) Sequence
653 comparison of the 5’ flanking regions of Japanese pear (Pyrus pyrifolia) S-RNases associated with
654 gametophytic self-incompatibility. Sex Plant Reprod 13:289–291 . doi: 10.1007/s004970100067

655 Nybom H, Sehic J, Garkava-Gustavsson L (2008) Self-incompatibility alleles of 104 apple cultivars grown in
656 northern Europe. J Hortic Sci Biotechnol 83:339–344 . doi: 10.1080/14620316.2008.11512389

657 Okada K, Moriya S, Haji T, Abe K (2013) Isolation and characterization of multiple F-box genes linked to the
658 S9- and S10-RNase in apple (Malus × domestica Borkh.). Plant Reprod 26:101–111 . doi:
659 10.1007/s00497-013-0212-0

660 Okada K, Tonaka N, Taguchi T, Ichikawa T, Sawamura Y, Nakanishi T, Takasaki-Yasuda T (2011) Related
661 polymorphic F-box protein genes between haplotypes clustering in the BAC contig sequences around
662 the S-RNase of Japanese pear. J Exp Bot 62:1887–1902 . doi: 10.1093/jxb/erq381

663 Potter D, Eriksson T, Evans RC, Oh S, Smedmark JEE, Morgan DR, Kerr M, Robertson KR, Arsenault M,
664 Dickinson TA, Campbell CS (2007) Phylogeny and classification of Rosaceae. Plant Syst Evol 266:5–43 .
665 doi: 10.1007/s00606-007-0539-9

666 Raspé O, Kohn JR (2002) S-allele diversity in Sorbus aucuparia and Crataegus monogyna (Rosaceae:
667 Maloideae). Heredity (Edinb) 88:458–65 . doi: 10.1038/sj/hdy/6800079

668 Sambrook J, Fristch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd Ed. Cold Spring Harb
669 Lab Press Cold Spring Harb New York

670 Sanzol J (2010) Two neutral variants segregating at the gametophytic self-incompatibility locus of European
671 pear (Pyrus communis L.) (Rosaceae, Pyrinae). Plant Biol 12:800–805 . doi: 10.1111/j.1438-
672 8677.2009.00277.x

673 Sanzol J (2009a) Genomic characterization of self-incompatibility ribonucleases (S-RNases) in European
674 pear cultivars and development of PCR detection for 20 alleles. Tree Genet Genomes 5:393–405 . doi:
675 10.1007/s11295-008-0194-5

676 Sanzol J (2009b) Pistil-function breakdown in a new S-allele of European pear, S 21 ??, confers self-
677 compatibility. Plant Cell Rep 28:457–467 . doi: 10.1007/s00299-008-0645-3

678 Sassa H, Hirano H, Ikehashi H (1993) Identification and characterization of stylar glycoproteins associated
679 with self-incompatibility genes of Japanese pear, Pyrus serotina Rehd. MGG Mol Gen Genet 241:17–
680 25 . doi: 10.1007/BF00280196

681 Sassa H, Kakui H, Miyamoto M, Suzuki Y, Hanada T, Ushijima K, Kusaba M, Hirano H, Koba T (2007) S locus
682 F-box brothers: Multiple and pollen-specific F-box genes with S haplotype-specific polymorphisms in
683 apple and Japanese pear. Genetics 175:1869–1881 . doi: 10.1534/genetics.106.068858

684 Sassa H, Nishio T, Kowyama Y, Hirano H, Koba T, Ikehashi H (1996) Self-incompatibility (S) alleles of the
685 Rosaceae encode members of a distinct class of the T2/S ribonuclease superfamily. Mol Gen Genet
686 250:547–557 . doi: 10.1007/s004380050108

687 Shahmuradov IA, Umarov RK, Solovyev V V. (2017) TSSPlant: A new tool for prediction of plant Pol II
688 promoters. Nucleic Acids Res 45: . doi: 10.1093/nar/gkw1353

689 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J,
690 Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence
691 alignments using Clustal Omega. Mol Syst Biol 7:539 . doi: 10.1038/msb.2011.75

692 Steinbachs JE, Holsinger KE (2002) S-RNase-mediated gametophytic self-incompatibility is ancestral in
693 eudicots. Mol Biol Evol 19:825–829 . doi: 10.1093/oxfordjournals.molbev.a004139

694 Takasaki-Yasuda T, Nomura N, Moriya-Tanaka Y, Okada K, Iwanami H, Bessho H (2013) Cloning an S-RNase
695 allele, including the longest intron, from cultivars of European pear (Pyrus communis L.). J Hortic Sci
696 Biotechnol 88:427–432 . doi: 10.1080/14620316.2013.11512987

697 Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics
698 analysis version 6.0. Mol Biol Evol 30:2725–2729 . doi: 10.1093/molbev/mst197

699 Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3-new
700 capabilities and interfaces. Nucleic Acids Res 40:e115 . doi: 10.1093/nar/gks596

701 Ushijima K, Sassa H, Hirano H (1998) Characterization of the flanking regions of the S-RNase genes of
702 Japanese pear (Pyrus serotina) and apple (Malus x domestica). Gene 211:159–167 . doi:
703 10.1016/S0378-1119(98)00105-X

704 Van Nerum I, Geerts M, Van Haute a., Keulemans J, Broothaerts W, Nerum I van, Geerts M, Haute A van,
705 Keulemans J, Broothaerts W (2001) Re-examination of the self-incompatibility genotype of apple
706 cultivars containing putative “new” S-alleles. Theor Appl Genet 103:584–591 . doi:
707 10.1007/PL00002913

708 Vavilov NI (1930) Wild Progenitors of the fruit trees of Turkistan and Cacausus and the problem of the
709 origin of fruit trees. Report Proc 9th Int Hort Congr 271–286

710 Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio
711 M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A,
712 Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M,
713 Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma
714 T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D,
715 Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R,
716 Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagné
717 D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett J a, Proost S, Rouzé P, Sterck L, Toppo S, Lazzari B,
718 Hellens RP, Durel C-E, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y,
719 Salamini F, Viola R (2010) The genome of the domesticated apple (Malus × domestica Borkh.). Nat

720 Genet 42:833–839 . doi: 10.1038/ng.654

721 Vieira J, Ferreira PG, Aguiar B, Fonseca NA, Vieira CP (2010) Evolutionary patterns at the RNase based
722 gametophytic self - incompatibility system in two divergent Rosaceae groups (Maloideae and Prunus).
723 BMC Evol Biol 10:200 . doi: 10.1186/1471-2148-10-200

724 Vieira J, Morales-Hojas R, Santos RAMM, Vieira CP (2007) Different positively selected sites at the
725 gametophytic self-incompatibility pistil S-RNase gene in the Solanaceae and Rosaceae (Prunus, Pyrus,
726 and Malus). J Mol Evol 65:175–185 . doi: 10.1007/s00239-006-0285-6

727 Wang S, Kakui H, Kikuchi S, Koba T, Sassa H (2012) Interhaplotypic heterogeneity and heterochromatic
728 features may contribute to recombination suppression at the S locus in apple (Malus x domestica). J
729 Exp Bot 63:4983–4990 . doi: 10.1093/jxb/ers176

730 Zamboni A, Pierantoni L, De Franceschi P (2008) Total RNA extraction from strawberry tree (Arbutus unedo)
731 and several other woody-plants. iForest - Biogeosciences For 1:122–125 . doi: 10.3832/ifor0465-
732 0010122

733 Zisovich AH, Stern RA, Sapir G, Shafir S, Goldway M (2004) The RHV region of S-RNase in the European pear
734 (Pyrus communis) is not required for the determination of specific pollen rejection. Sex Plant Reprod
735 17:151–156 . doi: 10.1007/s00497-004-0225-9

736

737 **Tables**

738 **Table 1.** List and accession numbers of the *S-RNase* sequences used for building the synthetic alignment
739 template; alleles from 1 to 34 follow the numeration reported by Matsumoto (2014), while other alleles
740 were described by Long et al. 2010b; Long et al. 2010a; Dreesen et al. 2010; Li et al. 2012; and De
741 Franceschi et al. 2016

Allele	Acc. #	Allele	Acc. #	Allele	Acc. #	Allele	Acc. #
1	D50837[b] EU427454[c]	9	AB270792	25	AB062100[b] AB428431[c]	34	AB540122
2	U12199[b] HQ693077[c]	10	AF327221[b] AB428428[c]	26	AF016918[b] AB428432[c]	41	KT724706
3	U12200[b] EU427455[c]	11[a]	FJ008669[c]	28	AB035273[b] AF201748[c]	42[a]	EU427453[c]
4	AF327223[b] EU427456[c]	16	AF016919[b] AB428429[c]	29[a]	AY039702[c]	44[a]	EU443101[c] FJ008673[c]
5[a]	U19791[b] AB428427[c]	20	AB019184[b] HQ689397[c]	30	AB035928[b] AB052268[c]	45[a]	FJ008671[c]
6[a]	EU427461[c]	21[a]	FJ008670[c]	31	DQ135990	46[a]	FJ008672[c]
7	AB032246[b] AB050634[c]	23[a]	AF239809[c]	32	DQ135991	50[a]	FJ535241[c]
8[a]	AY744080[b]	24	AF016920[b] HQ693064[c]	33	AB540121	53[a]	FJ602074[c]

[a] full-length coding sequence is not available
[b] mRNA
[c] partial genomic sequence

742

743

744 **Table 2.** Summary of the sequences obtained by the assembly of reads; the size (bp) of flanking regions,
745 exons and the intron are reported. The number between parentheses indicates the portion of sequence
746 known before the assembly.

Allele	5' flanking	Exon 1	Intron	Exon 2	3' flanking
S_1	1167	243	344	438	1346
S_2	877	252	147	435	1053
S_3	747	252	1290	435	1108
S_4	912	244	140	440	44
S_5	536	252 (159)	1158	438	754
S_{6a}	0	246 (138)	167	392 (26)	0
S_{6b}	0	242 (138)	167	34 (26)	0
S_7	571	252	118	435	866
S_8	1077	252 (171)	1232 (0)	435	1102
S_9	1084	252	144	435	85
S_{10}	766	252	1717	435	1152
S_{11}	363	246 (150)	175	438 (384)	506
S_{16b}	1275	246	2130	438	1559
S_{20}	750	243	318	438	1437
S_{23}	882	252 (132)	147	435 (324)	198
S_{24}	1013	243	339	438	1361
S_{25}	1323	252	2359	435	251
S_{26}	1165	252	159	432	1057
S_{28}	813	252	169	432	1011
S_{32}	901	252	151	435	12
S_{33}	990	261	726	438	68
S_{46}	541	252 (156)	143	435 (381)	42
S_{50}	14	252 (156)	162	429 (381)	1028

747

748

Table 3. Estimates of evolutionary divergence between *S-RNase* sequences. The distances calculated as number of amino acid substitutions per site using the Poisson correction model are reported in the lower-left part, while the upper-right shows the number of amino acid differences per sequence.

	S_1	S_2	S_3	S_4	S_5	S_{6a} (S_6)	S_{6b} (S_{17})	S_7	S_8	S_9	S_{10}	S_{11}	S_{16b}	S_{20}	S_{23}	S_{24}	S_{25}	S_{26}	S_{28}	S_{32}	S_{33}	S_{39}	S_{46}	S_{50}	S_{58}
S_1	-	82	68	72	81	68	69	78	85	70	67	79	77	14	83	17	77	71	67	79	77	89	73	71	73
S_2	0.455	-	76	77	87	76	78	66	87	67	76	76	82	83	86	83	86	70	77	81	78	92	66	71	78
S_3	0.361	0.405	-	76	74	68	68	70	82	66	14	77	76	66	74	69	71	78	71	76	80	79	69	71	72
S_4	0.389	0.414	0.407	-	84	75	74	80	81	75	76	62	36	75	81	74	74	66	70	76	78	83	76	66	69
S_5	0.451	0.483	0.394	0.462	-	83	84	73	89	82	73	85	83	82	89	84	84	90	83	84	50	45	85	85	85
S_{6a} (S_6)	0.359	0.407	0.356	0.403	0.457	-	14	74	81	67	68	83	80	70	81	72	73	71	30	74	74	80	68	62	69
S_{6b} (S_{17})	0.366	0.421	0.356	0.396	0.464	0.063	-	74	78	69	68	85	80	71	84	75	74	71	28	72	76	81	71	66	71
S_7	0.430	0.343	0.368	0.434	0.385	0.396	0.396	-	86	58	70	81	82	76	79	79	74	69	71	77	66	79	61	70	75
S_8	0.477	0.480	0.445	0.441	0.497	0.441	0.421	0.476	-	76	78	87	85	83	79	85	32	88	77	66	84	89	78	78	87
S_9	0.374	0.347	0.341	0.401	0.448	0.349	0.362	0.295	0.405	-	69	79	70	67	78	70	67	61	65	64	74	86	16	62	65
S_{10}	0.355	0.405	0.063	0.407	0.387	0.356	0.356	0.368	0.418	0.360	-	79	75	65	73	67	66	76	67	76	75	75	70	67	70
S_{11}	0.437	0.407	0.414	0.319	0.469	0.457	0.471	0.441	0.483	0.427	0.427	-	57	81	85	82	80	75	83	75	80	88	79	73	75
S_{16b}	0.423	0.448	0.407	0.172	0.455	0.436	0.436	0.448	0.469	0.368	0.401	0.289	-	79	84	77	76	70	74	77	82	84	73	67	71
S_{20}	0.063	0.462	0.349	0.409	0.458	0.372	0.379	0.416	0.462	0.355	0.342	0.451	0.437	-	80	15	76	72	67	78	77	86	71	73	74
S_{23}	0.462	0.473	0.392	0.441	0.497	0.441	0.462	0.427	0.425	0.418	0.385	0.469	0.462	0.441	-	84	75	79	78	84	78	89	83	77	85
S_{24}	0.078	0.462	0.368	0.403	0.472	0.385	0.405	0.437	0.477	0.374	0.355	0.458	0.423	0.068	0.470	-	78	73	69	80	78	86	73	74	74
S_{25}	0.421	0.473	0.373	0.394	0.462	0.387	0.394	0.394	0.151	0.347	0.341	0.434	0.407	0.414	0.398	0.428	-	77	70	55	77	82	70	69	76
S_{26}	0.381	0.368	0.421	0.345	0.507	0.375	0.375	0.364	0.490	0.312	0.407	0.403	0.370	0.387	0.427	0.394	0.414	-	65	65	79	90	62	60	69
S_{28}	0.355	0.416	0.377	0.372	0.460	0.141	0.131	0.379	0.416	0.339	0.351	0.460	0.398	0.355	0.423	0.368	0.370	0.339	-	70	73	83	69	60	68
S_{32}	0.434	0.438	0.405	0.407	0.462	0.394	0.381	0.414	0.341	0.329	0.405	0.401	0.414	0.428	0.459	0.441	0.276	0.337	0.370	-	75	82	64	72	81
S_{33}	0.421	0.418	0.432	0.421	0.246	0.394	0.407	0.341	0.459	0.392	0.398	0.434	0.448	0.421	0.418	0.428	0.412	0.427	0.390	0.398	-	47	74	79	85
S_{39}	0.509	0.519	0.427	0.455	0.218	0.436	0.443	0.425	0.497	0.476	0.401	0.490	0.462	0.487	0.497	0.487	0.448	0.507	0.460	0.448	0.229	-	89	90	93
S_{46}	0.394	0.341	0.360	0.407	0.469	0.356	0.375	0.312	0.418	0.072	0.366	0.427	0.387	0.381	0.452	0.394	0.366	0.319	0.364	0.329	0.392	0.497	-	65	68
S_{50}	0.383	0.377	0.377	0.347	0.474	0.320	0.345	0.372	0.423	0.320	0.351	0.392	0.353	0.396	0.416	0.403	0.364	0.308	0.310	0.383	0.430	0.510	0.339	-	30

S_{58}	0.394	0.421	0.381	0.364	0.471	0.362	0.375	0.403	0.483	0.337	0.368	0.403	0.377	0.401	0.469	0.401	0.407	0.362	0.357	0.441	0.469	0.530	0.356	0.142	-

Table 4. Summary of the *S*-genotypes determined for the 63 apple accessions analyzed in this study.

Cultivar	S genotype	Cultivar	S genotype	Cultivar	S genotype
Abbondanza	S_3S_5	Delicious[a]	S_9S_{28}	Malinové Holovouské	S_3S_{20}
Ag Alma	$S_{23}S_{50}$	Dr. Oldenburg	S_3S_{16b}	Mcintosh[a]	$S_{10}S_{25}$
Aivaniya	S_1S_{25}	Durello Di Forlì	S_3S_{50}	Mela Rosa (Pd)	S_3S_{25}
Ajmi	S_8S_{23}	F22682922	S_5S_{24}	Mela Rozza	S_3S_6
Åkerö[a]	S_1S_{58}	Filippa[a]	S_7S_{24}	Ovčí Hubička	S_3S_{28}
Alfred Jolibois	S_1S_{28}	Fuji[a]	S_1S_9	Panenské České	S_7S_{10}
Amadou	$S_{20}S_{33}$	Fyriki	$S_{16b}S_{26}$	Papirovka	S_1S_5
Annurca	S_7S_{26}	Gelata	S_7S_{16b}	Patte De Loup	S_1S_2
Antonovka	S_8S_{32}	Godelieve Hegmans	S_2S_7	Pepino Jaune	S_1S_3
Antonovka Pamtorutka	S_8S_{58}	Heta	$S_{16}bS_{17}$	Precoce De Karage	S_4S_{26}
Aport Kuba	S_1S_{50}	Hetlina	S_1S_{16b}	President Roulin	S_5S_{24}
Belle Et Bonne	S_3S_{33}	Ijunskoe Ranee	$S_{39}S_{46}$	Priscilla[b]	S_7S_{10}
Borowitsky[b]	S_3S_{28}	Jantarnoe	S_8S_{26}	Reinette Clochard	S_3S_4
Braeburn[a]	S_9S_{24}	Jonathan[a]	S_7S_9	Reinette Dubois	S_3S_{11}
Budimka	S_1S_{16b}	Keswick Codlin	S_4S_{20}	Renetta Grigia Torriana	S_7S_{28}
Busiard	S_1S_5	Kmenotvorná	S_3S_7	Rosa (Fi)	$S_{20}S_{28}$
Cabarette	S_3S_{24}	Košíkové	S_3S_8	Skry	$S_{24}S_{28}$
Chodské	S_2S_{10}	Kronprins	$S_{11}S_{28}$	Sonderskow	$S_{10}S_{33}$
Court-Pendu Henry	S_3S_5	Lady Williams[a]	S_7S_{23}	Spässerud[a]	S_1S_7
Cox's Orange Pippin[a]	S_5S_9	Macoun[a]	S_3S_{25}	Worchester Pearmain[b]	$S_{24}S_{25}$
De L'Estre	S_1S_2	Maikki	$S_{10}S_{17}$	Young America	$S_{28}S_{39}$

[a] *S*-genotypes in agreement with previous reports (Janssens et al. 1995; Sassa et al. 1996; Matsumoto et al. 1999; Van Nerum et al. 2001; Kitahara and Matsumoto 2002a; Kitahara and Matsumoto 2002b; Broothaerts et al. 2004; Garkava-Gustavsson et al. 2008; Nybom et al. 2008; Matsumoto 2014)

[b] *S*-genotypes in disagreement with previous reports (Kitahara et al. 2000; Broothaerts et al. 2004; Morita et al. 2009; Long et al. 2010a)

Table 5. Prediction of transcript start site (TSS) and TATA box position in the promoter region of *S-RNase*s.

Allele	TSS position	TSS score	TATA position	TATA score
S_1	-78	1.988	-111	8.193
S_2	-83	1.988	-116	8.193
S_3	-78	1.988	-111	5.506
S_4	-78	1.988	-111	7.781
S_5	-77	1.989	-111	8.193
S_6	-50	1.888	TATA-	
S_7	-68	1.988	-102	8.193
S_8	-67	1.986	-100	6.901
S_9	-73	1.988	-105	8.193
S_{10}	-78	1.988	-111	5.506
S_{11}	-72	1.989	-105	8.193
S_{16b}	-78	1.989	-111	8.193
S_{17}	-50	1.887	TATA-	
S_{20}	-322	1.991	TATA-	
S_{23}	-199	1.968	-230	4.240
S_{24}	-333	1.993	TATA-	
S_{25}	-77	1.988	-111	8.193
S_{26}	-74	1.989	-108	8.193
S_{28}	-90	1.989	-123	8.193
S_{32}	-78	1.986	-111	6.972
S_{33}	-71	1.989	-104	8.193
S_{39}	-77	1.989	-111	8.193
S_{46}	-79	1.988	-113	8.193
S_{50}	-78	1.988	-111	8.193
S_{58}	-76	1.988	-109	8.193

Figure captions

Fig. 1 Alignment of the newly assembled flanking regions of the *S_9-RNase* to the genomic sequence. The varying length obtained at the 5' side (a) after 10 iterations depended essentially on the different coverage obtained from the tested cultivars, while at the 3' side (b) assembly of reads was hindered by the presence of a simple sequence repeat

Fig. 2 Alignment of deduced protein sequences of the 25 *S-RNase* alleles detected in the pool of resequenced cultivars. Black and grey backgrounds mark conserved sites and conservative substitutions, respectively; the five conserved (C1-C5) and one hyper-variable (RHV) regions of Rosaceous *S-RNase*s are underlined

Fig. 3 Phylogenetic tree built on the alignment of S-RNase protein sequences using the Neighbor-Joining method; the percentage of replicate trees in which the associated sequences clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The evolutionary distances were computed using the Poisson correction method and are expressed as the number of amino acid substitutions per site

Fig. 4 Bar plot of allele frequencies among the 63 cultivars

Fig. 5 Aligment of *S-RNase* promoters. Transcript start site (TSS) and TATA box were predicted using TSSPlant (see text); the conserved region box1, the motif IA-like and the annealing site of primer MdProm-box1f are indicated

Electronic Supplementary Material

Online Resource 1. List and sequence of primers used in this study.

Online Resource 2. Coverage obtained from the alignment of Illumina reads on the synthetic sequence.

Online Resource 3. Verification of *S-RNase* alleles detected in 'Borowitsky', 'Priscilla' and 'Worchester Pearmain'.

Online Resource 4. Alignment of *Malus* hybrid 'Mt Blanc' *S_{58}* with *M. × domestica S_{58}-RNase* and *M. sylvestris S_{39}* with *M. × domestica S_{39}-RNase*.

Online Resource 5. Alignment of sequences derived from assembly of reads and traditional cloning and sequencing.

Online Resource 6. Alignment of *S-RNase* 5' flanking regions.

Online Resource 7. Test amplification with primer MdProm-box1f from a set of cultivars of different pear species.

Online Resource 8. Alignment of *S-RNase* 3' flanking regions.