*Genet. Res., Camb.* (2008), **90**, *pp.* 17–26. © 2008 Cambridge University Press
doi:1017/S0016672307009044 Printed in the United Kingdom


# The number, age, sharing and relatedness of *S*-locus specificities in *Prunus*

JORGE VIEIRA[1], NUNO A. FONSECA[1], RAQUEL A. M. SANTOS[1], TSUYOSHI HABU[2], RYUTARO TAO[3] AND CRISTINA P. VIEIRA[1]*
[1] *Instituto de Biologia Molecular e Celular (IBMC), University of Porto, Rua do Campo Alegre 823, 4150-180 Porto, Portugal*
[2] *Experimental Farm, Graduate School of Agriculture, Kyoto University, Takatsuki 569-0096, Japan*
[3] *Laboratory of Pomology, Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan*

(*Received 19 September 2007 and in revised form 12 November 2007*)

**Summary**

In gametophytic self-incompatibility systems, many specificities (different 'lock-and-key' combinations) are maintained by frequency-dependent selection for very long evolutionary times. In Solanaceae, trans-specific evolution (the observation that an allele from one species may be more closely related to an allele from another species than to others from the same species) has been taken as an argument for the very old age of specificities. In this work, by determining, for the first time, the age of extant *Prunus* species, we show that this reasoning cannot be applied to Prunoideae. Furthermore, since our sample size is large (all *S-RNase* encoding the female component and *SFB* encoding the male component GenBank sequences), we were able to estimate the age of the oldest *Prunus* specificities. By doing so, we show that the lower variability levels at the *Prunus S*-locus, in comparison with Solanaceae, is due to the younger age of *Prunus* alleles, and not to a difference in silent mutation rates. We show that the ancestor to extant *Prunus* species harboured at least 102 specificities, in contrast to the maximum of 33 observed in extant *Prunus* species. Since the number of specificities that can be maintained in a population depends on the effective population size, this observation suggests a bottleneck in *Prunus* evolutionary history. Loss of specificities may have occurred during this event. Using only information on amino acid sites that determine specificity differences, and a simulation approach, we show that a model that assumes closely related specificities are not preferentially lost during evolution, fails to predict the observed degree of specificity relatedness.

## 1. Introduction

Gametophytic self-incompatibility (GSI) is a genetic mechanism present in flowering plants that is controlled by the multi-allelic *S*-locus and that prevents self-fertilization, by enabling the pistil to reject pollen from genetically related individuals (de Nettancourt, 1977). Recognition is a 'lock-and-key' system, and each 'lock-and-key' is defined as a specificity (Charlesworth *et al*., 2005). The *S*-locus contains two separate genes: one that determines pistil specificity and another that determines pollen specificity. GSI has been extensively studied in Solanaceae, Plantaginaceae and Rosaceae species. In these families, the pistil component of GSI has been shown to be an *S-RNase* (see review by Wang *et al*., 2003). *S-RNase*-based GSI has been proposed to have evolved only once, before the separation of Asteridae and Rosidae (Igic & Kohn, 2001; Steinbachs & Holsinger, 2002; Vieira *et al*., 2007*c*).

In GSI, many allele specificities are maintained by frequency-dependent selection for very long evolutionary times (Wright, 1939; Takahata, 1990). In Solanaceae, for instance, alleles from species that diverged 30 million years ago are found mingled in the phylogenetic tree (Richman *et al*., 1996; Charlesworth

* Corresponding author. IBMC, Molecular Evolution Group, Rua do Campo Alegre 823, 4150-180 Porto, Portugal. Telephone: +351 226074900. Fax: +351 226099157. e-mail address: cgvieira@ibmc.up.pt

& Guttman, 1997). Therefore, an allele from one species may be more closely related to an allele from another species than to other alleles from the same species, a pattern named trans-specific evolution (Richman *et al*., 1996). Trans-specific evolution has been described also in *Prunus* (see for instance, Ortega *et al*., 2006). Nevertheless, it is unclear whether trans-specific evolution can be taken as evidence for the very old age of *S*-alleles, as argued by Richman *et al*. (1996), in genera where no estimation of the species age is available. For the genus *Prunus*, the relationship of the different species has been determined (Bortiri *et al*., 2002), but no estimate is given for the speciation times. In this work, the first age estimate for the genus *Prunus* is provided, using the published chloroplast sequence data and the estimated divergence time between the Maloideae and Prunoideae (32 million years; Wikstrom *et al*., 2001).

Levels of synonymous variability are, on average, 3·7 times lower in the *Prunus S-RNases* than in Solanaceae *S-RNases* (Vieira *et al*., 2007*b*). This observation suggests that *Prunus S-RNase* alleles are considerably younger than the Solanaceae *S-RNase* alleles, or that the neutral mutation rate is different in *Prunus* and Solanaceae. Furthermore, *S*-alleles from *Prunus* species never cluster with Maloideae *S*-alleles (Ushijima *et al*., 1998; Ma & Oliveira, 2002). Therefore, *Prunus S*-alleles must be younger than 32 million years. In this work, for the first time, the age of the oldest *Prunus S*-alleles is estimated. This is feasible since more than 150 different *Prunus S-RNase* sequences have been reported (although most are partial); thus, it is likely that at least some of the oldest *S*-alleles are present in the sample. Furthermore, the pollen component of *Prunus* GSI has been identified as being an *F*-box gene (named *SFB*; Entani *et al*., 2003; Ushijima *et al*., 2003). Levels of variability at the *S-RNase* and *SFB* genes are similar, the evolutionary histories of the two genes are correlated (although not completely), and amino acid sites under frequency-dependent selection (those likely to be responsible for specificity differences) are found in both genes (Nunes *et al*., 2006; Vieira *et al*., 2007*a*, *b*). More than 75 *SFB* allele sequences have been reported; thus, it is likely that at least some of the oldest *S*-alleles are present in the sample. Therefore, it is feasible to calculate the age of the oldest *Prunus S*-alleles based on the *SFB* data, as well.

In finite populations, the number of specificities that can be maintained at equilibrium is dependent on selection, mutation and drift (Wright, 1939). For the same strength of selection and mutation rate, large populations will harbour more specificities than small populations (Wright, 1939). For most of the species exhibiting GSI (16 of 19 species; Lawrence, 2000) the estimated number of alleles in natural populations is below 45. In *Physalis crassifolia* (Solanaceae), which is estimated to have 44 specificities, the implied effective population size is 6000 to 10 000 individuals (Richman *et al*., 1996). In *Prunus*, most species sampled exhibit fewer than 34 specificities (Vieira *et al*., 2007*a*). The screening of a large number of individuals never revealed more than 33 specificities. By screening 145 *P. lannesiana* individuals Kato & Mukai (2004) found 22 specificities; in 65 *P. avium* individuals studied by de Cuyper *et al*. (2005) 18 specificities were found, and later by screening 164 *P. avium* individuals Schueler *et al*. (2006) found 15 specificities; Ortega *et al*. (2006) found 25 specificities to be present in 104 *P. dulcis* cultivars. Thus, the effective population size of most *Prunus* species exhibiting GSI is thought to be below 10 000 individuals. Historical changes in population size may lead to *S*-allele loss. Therefore, another goal of this work was to determine the number of specificities found in the common ancestor to all *Prunus* species. This number can be compared with that observed in extant *Prunus* species in order to infer historical changes in population size. Such changes in population size also have been inferred in Solanaceae species (Richman *et al*., 1996; Richman & Kohn, 2000; Lu, 2006).

Theoretical models predict that a novel specificity is expected to replace the one that gave origin to it in the local population where it arose, but migration from another population may prevent the loss of the original specificity (Uyenoyama *et al*., 2001). Nevertheless, when two closely related specificities are present in one population, it is possible that in a fraction of all attempted fertilizations one of the specificities is misrecognized as being the closely related one (Newbigin & Uyenoyama, 2005). Therefore, closely related specificities should be rare in natural populations, but could be frequent when comparing closely related species. *Prunus S-RNase* and *SFB* amino acid sites under frequency-dependent selection have been identified (Nunes *et al*., 2006; Vieira *et al*., 2007*a*, *b*). Therefore, it is possible to test whether, in *Prunus*, closely related specificities are found at the expected frequency, under the assumptions that they are not preferentially lost during evolution, and that a single amino acid change at a site identified as being under frequency-dependent selection results in a different specificity. Although the latter assumption is debatable, in two *P. spinosa* populations two alleles differing at a single such site have been found in the same individual, implying that they probably represent two different specificities (Vieira *et al*., 2007*a*). Here, we use for the first time a simulation approach to investigate the expected shape of the relationship between specificities, under the above assumptions, using only information at sites identified as under frequency-dependent selection.

## 2. Materials and methods

### (i) *Data sets used and computational methods*

To estimate the age of the different *Prunus* subgenera, neutral genes are needed, since the *S*-locus is under frequency-dependent selection. Three genes for which nucleotide sequence data are available for at least one species from each of the *Prunus* subgenera (*Prunus*, *Amygdalus* and *Cerasus*) and one species of Maloideae were used (the non-coding chloroplast regions *trnL-trnF*, *trnS-trnG* intergenic spacers, and the *rpl16* intron; accession numbers are shown in Supplementary Table 1 at http://evolution.ibmc.up.pt/data/prunus_GR). The nucleotide sequences were aligned using ClustalX (Thompson *et al*., 1997). The number of silent changes per silent site (*Ks* values with Jukes–Cantor correction) was computed using the DNasp software (Rozas *et al*., 2003). Based on two plastid genes (*rbcL* and *atpB*) and one nuclear gene (*18S rDNA*), and using as a calibration point two fossils that imply the split between the Fagales and Cucurbitales occurred 84 million years ago, Wikstrom *et al*. (2001) provide a mean estimate of 32 million years ago for the split between Prunoideae and Maloideae species.

*Prunus S-RNase* and *SFB* protein sequences were retrieved from GenBank by searching blastp with one complete *S-RNase* (BAC65203) and *SFB* (BAC65204) protein sequence, respectively. Sequences labelled as non-functional and those from *Prunus* self-compatible species were included in the analyses, since they are closely related to functional *S*-alleles (see, for instance, Hauck *et al*., 2006; Tsukamoto *et al*., 2006). Sequences labelled as non *S-RNases* and non-*SFB* that were retrieved in this way were discarded from the data set except for two *S*-like *RNases* from *P. dulcis* (*PD*1 and *PD*2) and one *SFB*-like sequence from *P. armeniaca*. These sequences were used to root the linearized minimum evolution trees. Since, using this procedure, *S-RNase* and *SFB* paralogous genes were also retrieved, it is likely that all *Prunus S-RNase* and *SFB* sequences available in GenBank, at that time, were retrieved. For the same species, when two or more identical protein sequences were retrieved only the longest one was used. Furthermore, when similar proteins (less than three amino acid differences) from the same species with identical specificity names were found only the longest one was used. *P. dulcis S*-allele synonyms (i.e. identical sequences with different specificity names) reported in Ortega *et al*. (2006) were also considered. The final data set contains 158 and 78 *S-RNase* and *SFB* sequences, respectively (see Supplementary Tables 2 and 3 for accession numbers at http://evolution.ibmc.up.pt/data/prunus_GR). Protein sequences were aligned using ClustalX (Thompson *et al*., 1997). The minimum evolution trees were obtained using pairwise deletion as implemented in the MEGA software (Kumar *et al*., 2004), since most sequences are partial. Maximum likelihood and Bayesian methods for phylogeny reconstruction, using the corresponding nucleotide sequences, could not be used, since they use a complete-deletion approach. Since most sequences are partial, all positions have alignment gaps (data not shown). Furthermore, there is nucleotide substitution saturation when sequences from the Maloideae and Prunoideae are compared (data not shown). In order to estimate the age of the oldest *Prunus S*-allele, as a calibration point we use the age of the estimated split between Prunoideae and Maloideae species (32 million years; Wikstrom *et al*., 2001).

We used our estimate of the age of extant *Prunus* species, and our calibrated amino acid minimum evolution tree, to estimate the number of *S*-alleles in the ancestor to extant *Prunus* species. For pairs of *S-RNase* and *SFB* amino acid sequences, which could represent instances of two copies of the same ancestral specificity inherited by two different species (see Section 3), the corresponding nucleotide sequences were retrieved from GenBank (see Supplementary Tables 4 and 5 at http://evolution.ibmc.up.pt/data/prunus_GR for accession numbers). The nucleotide sequences were aligned using ClustalX (Thompson *et al*., 1997), and the number of silent changes per silent site (*Ks* values with Jukes–Cantor correction) was computed using the DNasp software (Rozas *et al*., 2003). For the *SFB* gene, it is not possible to estimate the per site silent site nucleotide substitution rate, since the Maloideae orthologous gene has not been identified yet (Sassa *et al*., 2007). For the *S-RNase*, the average *Ks* for comparisons involving Maloideae and Prunoideae sequences is larger than 4·93 (data not shown), and it can be shown that there is nucleotide substitution saturation (data not shown). Thus, it is not possible to obtain an accurate estimate for the silent site nucleotide substitution rate, as well.

In order to estimate the degree of ancestral specificities shared between *Prunus* species A and B (where species B is always the one for which more specificities have been described) we calculated the percentage of ancestral specificities in species A that are represented in species B.

For the purpose of evaluating whether there is a lack of closely related *Prunus* specificities, only amino acids located at sites identified as being under frequency-dependent selection (those likely to be responsible for defining specificities) at the *S-RNase* (Vieira *et al*., 2007*b*) and *SFB* (Vieira *et al*., 2007*a*) genes were used to construct two independent sets of words, named S-RNase-w and SFB-w. Only sequences encompassing all sites identified as being under frequency-dependent selection were used. We have used only amino acid sites that were identified with a high degree of confidence (those amino acid

sites that have a posterior probability of selection higher than 50% when using two different methodologies; Nunes *et al*., 2006; Vieira *et al*., 2007*a*, *b*).

Random word networks were then generated using the following computational approach (program available on request from the authors): for the S-RNase-w and the SFB-w data sets, containing *n* different words (92 and 68 for the S-RNase-w and SFB-w, respectively), we first calculate the number and set of amino acid variants at each position; then a set of words is generated by performing the following steps: (i) Randomly select one word from the data set being considered (either the S-RNase-w or the SFB-w data set), and add it to a bag of words. (ii) Choose the distance *x* (number of differences) between the selected word and a new word to be created. This value is randomly taken from a Poisson distribution with a mean equal to the mean of the mean distance between a given word and all other words present in the data set being considered (either the S-RNase-w or the SFB-w data set). (iii) Create a new word by incrementally making *x* mutations; thus, by doing so *x* new words will be created and added to the bag of words. In order to perform this step, a random position in the word is first selected and then one amino acid is randomly selected from the set of amino acids observed in the data set being used, at that particular position. Repeat this process *x* times using the most recently created word as a template. (iv) For every word position check whether it uses all amino acids observed at that same position in the data set being considered; if not, then randomly select one word from the simulated bag of words and repeat steps (ii) and (iii). This procedure will generate a bag of simulated words much larger than the number of words in the data set being considered, thus suggesting that the empirical data set (either the S-RNase-w or the SFB-w) contains only a fraction of all words that could have arisen during evolution. The procedure described above is repeated 100 times, to generate 100 independent simulated bags of words for the S-RNase-w and the SFB-w data sets. From each bag of words, we take 100 random samples with size *n*. Thus, in the end we have 10 000 simulated samples from which we compute the appropriate summary statistics. The graphics and summary statistics were calculated using standard functions in the *R* language and environment (http://www.r-project.org/).

## 3. Results

### (i) *Age of the subgenera* Amygdalus, Prunus *and* Cerasus

Bortiri *et al*. (2002) determined the relationship of the three *Prunus* subgenera (*Amygdalus*, *Prunus* and *Cerasus*). Nevertheless, there is no estimate for the age of the subgenera or speciation times. In *Prunus*, trans-specific evolution has been described (see, for instance, Ortega *et al*., 2006). Nevertheless, without an estimation of *Prunus* speciation times it is not possible to know whether this observation can be used as evidence for the very old age of *S*-alleles, as in Solanaceae (Richman *et al*., 1996).

The split between the Prunoideae and Maloideae lineages has been estimated by Wikstrom *et al*. (2001) to have occurred between 29 and 35 million years ago (the average of the three estimates given by these authors is 32 million years). The relative synonymous divergences, based on published chloroplast data, of species from two different *Prunus* subgenera to the synonymous divergence of those species and Maloideae species are shown in Table 1. In this table, estimated ages (in millions of years) for the separation of the different *Prunus* subgenera are also shown. The common ancestor to the living species of the subgenera *Amygdalus*, *Prunus* and *Cerasus*, based on neutral genes, lived no more than 5 million years ago (Table 1).

### (ii) *Age of the oldest* Prunus S-*alleles*

Patterns of variability at *Prunus S-RNase* and *SFB* genes suggest that *Prunus S*-alleles are, on average, much younger than *S*-alleles in Solanaceae (Vieira *et al*., 2007*b*). At the *S-RNase* gene, the average amino acid divergence between Prunoideae and Maloideae sequences is 32% (data not shown). This value corresponds to 32 million years (Wikstrom *et al*., 2001). A linearized minimum evolution tree, using all available *Prunus S-RNase* amino acid sequences, is shown in Fig. 1*a*. In this tree, the oldest specificities show an estimated 15–20% amino acid divergence relative to most other specificities. Therefore, the oldest specificities are about 15–20 million years old.

A linearized minimum evolution tree using all available *SFB* amino acid sequences is presented in Fig. 1*b*. It is not possible to calibrate this tree in the same way as the *S-RNase* tree since the Maloideae orthologous gene has not been identified yet (Sassa *et al*., 2007). If we use the rate of change estimated for the *S-RNase* gene, then the oldest specificity also seems to be 15–20 million years old.

### (iii) *Specificity number in the ancestor to extant* Prunus *species*

Historical changes in population size may lead to *S*-allele loss, since the number of alleles that can be maintained in a population depends on the effective population size (Wright, 1939). Therefore, by comparing the number of specificities in the ancestor to extant *Prunus* species with that observed in extant *Prunus* species (fewer than 34 specificities), historical changes in population size can be inferred.

Table 1. *Average silent site divergence and estimated age in million of years* (*within parentheses*) *between pairs of species from the three* Prunus *subgenera and from the Maloideae estimated using three chloroplast gene regions*

| | | *Amygdalus* | *Prunus* | *Cerasus* | Average |
|---|---|---|---|---|---|
| *Prunus* | *trnL-trnF* spacer | 0·00661 (2·2[a]) | | | |
| | *trnS-trnG* spacer | 0·00774 (2·8[a]) | | | |
| | *rpl16* intron | 0·00510 (2·5[a]) | | | |
| | Average | (2·5) | | | |
| *Cerasus* | *trnL-trnF* spacer | 0·00872 (2·8[a]) | 0·01267 (4·1[a]) | | |
| | *trnS-trnG* spacer | 0·01215 (4·4[a]) | 0·01557 (5·6[a]) | | |
| | *rpl16* intron | 0·00477 (2·3[a]) | 0·00440 (2·2[a]) | | |
| | Average | (3·2) | (4·0) | | |
| Maloideae | *trnL-trnF* spacer | 0·09610 | 0·09610 | 0·10170 | 0·09797 (32[b]) |
| | *trnS-trnG* spacer | 0·08097 | 0·08607 | 0·09876 | 0·08860 (32[b]) |
| | *rpl16* intron | 0·06581 | 0·06503 | 0·06503 | 0·06529 (32[b]) |

[a] The average values between *Prunus* subgenera and Maloideae species were used in the calculations to obtain divergence ages in million of years.
[b] The split between the Prunoideae and Maloideae lineages has been estimated to have occurred between 29 to 35 million years ago (Wikstrom *et al*., 2001); thus we use the average of these values (32).

The common ancestor to the living species of the subgenera *Amygdalus*, *Prunus* and *Cerasus* lived no more than 5 million years ago, and, in this time period, *S-RNase* amino acid sequences are expected to have accumulated about 5% amino acid divergence. Therefore, pairs of *Prunus S-RNase* alleles that represent the capturing of the same ancestral specificity by two different *Prunus* species are expected to show less than 5% amino acid divergence. The same approach cannot be performed for the *SFB* gene, since a clear orthologue has not yet been identified in Maloideae species (Sassa *et al*., 2007). Nevertheless, the depth of the *Prunus S-RNase* and *SFB* trees is similar (see Fig. 1). Thus, it seems reasonable to use the 5% value for the *SFB* gene as well.

In Fig. 1*a* (the *S-RNase* tree), there are 100 groups of alleles that differ by more than 5% amino acid divergence from any others. Therefore, the most recent common ancestor of extant *Prunus* species is assumed to have harboured at least 100 specificities. Of the five short amino acid sequences that could not be incorporated in the tree (see Fig. 1*a* legend; *P. armeniaca* $S_{13}$, *P. salicina* $S_{12}$, *P. salicina* $S_{13}$, *P. armeniaca* $S_h$ and *P. webbii* $S_5$), two (*P. armeniaca* $S_{13}$ and *P. salicina* $S_{13}$) differ by more than 5% amino acid divergence from any others, thus increasing to 102 the number of specificities in the most recent common ancestor of extant *Prunus* species. In Fig. 1*b* (the *SFB* tree), there are 64 groups of alleles that differ by more than 5% amino acid divergence from any others. Therefore, the most recent common ancestor of extant *Prunus* species is assumed to have harboured at least 64 specificities.

A pairwise deletion approach was used to build the minimum evolution trees shown in Fig. 1 due to the different lengths of the amino acid sequences used. Thus, we are assuming that the different regions of these proteins are evolving at the same rate, which is known not to be true since there are conserved and hypervariable regions along these proteins (Nunes *et al*., 2006; Vieira *et al*., 2007*b*). Therefore, these trees should be interpreted with caution and used only as a rough guide for the relationship of the different sequences. Most confidence can be placed on the relationship of the closely related amino acid sequences that are supported by high bootstrap values. Only two pairs of alleles (X1 and X2, Fig. 1*a*) that seem to have been diverging for less than 5 million years do not present high bootstrap values. Silent (synonymous plus intron sites; *Ks* values with Jukes–Cantor correction) divergence values were thus obtained for all pairs of *S-RNase* and *SFB* alleles with less than 5% amino acid differences (see Supplementary Tables 6*a* and 6*b* at http://evolution.ibmc.up.pt/data/prunus_GR). Only four of 69 such allele pairs produce *Ks* values above 10%. For the *S-RNase* and *SFB* gene 62% and 57%, respectively, of the closely related alleles from two different species differ in at least one

https://doi.org/10.1017/S0016672307009044 Published online by Cambridge University Press

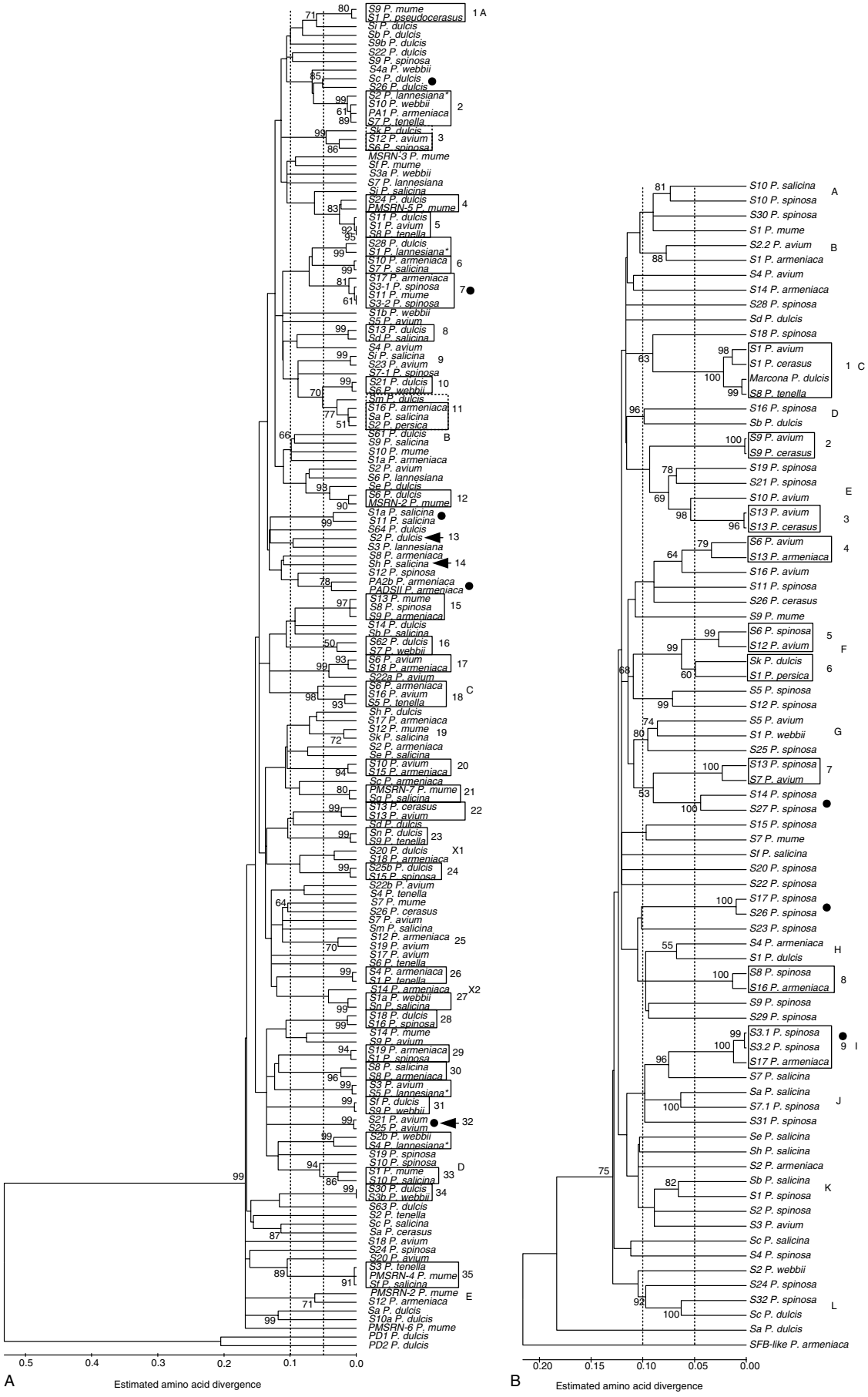


Fig. 1. For legend see opposite page.

https://doi.org/10.1017/S0016672307009044 Published online by Cambridge University Press

amino acid site identified as being under positive selection by Vieira *et al*. (2007*a*, *b*; see Supplementary Tables 6*a* and 6*b* at http://evolution.ibmc.up.pt/data/prunus_GR).

The percentage of ancestral specificities shared between *Prunus* species pairs ranges from 0 to 0·45, and from 0 to 0·23 for comparisons involving *Prunus* species from the same subgenus or from two different subgenera, respectively (see Supplementary Table 7 at http://evolution.ibmc.up.pt/data/prunus_GR). Despite the large variances associated with these estimates, *Prunus* species from the same subgenus share on average a higher percentage of ancestral specificities (the average is 0·18) than *Prunus* species from different subgenera (the average is 0·07; non-parametric Mann–Whitney test; $P<0·025$).

### (iv) *Relatedness of putative* Prunus *specificities*

Theoretical models predict that closely related specificities could be rare in natural populations, but could be frequent when comparing closely related species (Uyenoyama *et al*., 2001; Newbigin & Uyenoyama, 2005). It is conceivable that distantly related *S-RNase* or *SFB* sequences could represent closely related specificities. Thus, it is better to use information on amino acid sites shown to be under frequency-dependent selection than the full *S-RNase* or *SFB* sequences, as was done here, for the first time. Identifying all amino sites under diversifying selection (those likely to be responsible for specificity differences) at the *S-RNase* and *SFB* genes is, nevertheless, a difficult task since the different methodologies (a phylogenetic one (Yang, 1997), and a population genetics method that uses an approximation to the coalescent with recombination (Wilson & McVean, 2006)), typically yield overlapping but not fully compatible amino acid sites (Nunes *et al*., 2006; Vieira *et al*., 2007*a*, *b*). Furthermore, it is unknown whether the methodologies used are able to detect all amino acid sites under diversifying selection. It is assumed that a single amino acid change at an amino acid site identified as being under frequency-dependent selection results in a different specificity. Although this assumption is debatable, there is some evidence to support it. For the larger data set (the *S-RNase* data set), the frequency of closely related specificities (those that are less than 5 million years old) is not statistically different ($P>0·05$; two-sided Fisher exact test) within species (0·26% of all 1948 possible comparisons) and between species (0·20% of all 14 165 possible comparisons).

A simulation approach was, nevertheless, used, for the first time, to investigate the expected shape of the relationship between specificities under a simple model that assumes that closely related specificities are not preferentially lost during evolution. In order to do so, words must be assembled into word clusters. If a word can be connected by fewer than five changes to another word then we consider that both words belong to the same cluster. Five changes is an arbitrary choice, and using a different cluster definition could affect the results for a given summary statistic (the factors considered here are: number of clusters, average cluster size, maximum cluster size and average distance within clusters). Nevertheless, the simulations we performed using different cluster definitions show that there is always at least one feature in the empirical data that is unexpected (data not shown). Thus, the conclusion that the empirical data are incompatible with the simple model being tested is not dependent on the cluster definition used.

For instance, using the above cluster definition, the most striking feature of the S-RNase-w data set is the presence of a word cluster of size 50. Obtaining such a large cluster is unexpected ($P<0·001$; see Supplementary Fig. 1*a* at http://evolution.ibmc.up.pt/data/prunus_GR). Other aspects of the S-RNase-w data set, such as the average size of the clusters, and the total number of clusters, occur with a probability of 0·05 to 0·1 (see Supplementary Fig. 1*a* at http://evolution.ibmc.up.pt/data/prunus_GR). The within-cluster average word distance is slightly higher than in the simulated data (2·57 vs 2·38, respectively; $P>0·05$; see Supplementary Fig. 1*a* at http://evolution.ibmc.up. pt/data/prunus_GR). For the SFB-w data set the within-cluster average word distance is significantly lower than the average of the simulated data (1·46 vs 2·10, respectively; $P<0·005$; Supplementary Fig. 1*b* at http://evolution.ibmc.up.pt/data/prunus_GR). The average cluster size is also significantly lower in the SFB-w data set than in the simulated data (1·33 vs 1·94, respectively; $P<0·0005$; Supplementary Fig. 1*b* at http://evolution.ibmc.up.pt/data/prunus_GR). More clusters are found in the SFB-w data set than in the simulated data and this difference

Fig. 1. Linearized rooted minimum evolution tree showing the relationship of *S-RNase* (*a*) and *SFB* (*b*) sequences from *Prunus* species. The tree shown in (*a*) has been rooted by using the *PD1* and *PD2* sequences from *P. dulcis*, two *S-RNase*-like sequences. The tree shown in (*b*) has been rooted by using an *SFB*-like sequence from *P. armeniaca*. Squares delineate sets of alleles that are believed to be derived from the same ancestral specificity (see text for details). Dotted squares indicate uncertainty. Five short amino acid sequences could not be incorporated (*P. armeniaca S13*, *P. salicina S12*, *P. salicina S13*, *P. armeniaca Sh* and *P. webbii S5*) since they did not overlap other short sequences already in the tree. Arrowheads point to sequences that show similarity with one of those sequences. Black dots point to pairs of closely related alleles from the same species that are unlikely to represent the same specificity because they show amino acid differences at sites previously identified as being positively selected.

is significant (51 vs 37 respectively; $P<0{\cdot}0001$; see Supplementary Fig. 2*b* at http://evolution.ibmc.up.pt/data/prunus_GR). Nevertheless, the observed maximum cluster size is compatible with that obtained in the simulations ($P>0{\cdot}05$; see Supplementary Fig. 1*b* at http://evolution.ibmc.up.pt/data/prunus_GR).

## 4. Discussion

In Solanaceae, trans-specific evolution has been taken as evidence for the very old age of specificities, since alleles from species that diverged 30 million years ago cluster in the phylogenetic tree (Richman *et al*., 1996; Charlesworth & Guttman, 1997). Here we show, for the first time, that extant *Prunus* are a group of closely related species, since all speciation events happened in the last 5 million years. There is a Prunoideae fossil that is estimated to be 44 million years old (Magallón *et al*., 1999). The observation that a fossil shares similarities with living species of a given family does not indicate that the fossil taxon is part of the crown group of living species (Wikstrom *et al*., 2001). It is thus concluded that, in *Prunus*, trans-specific evolution cannot be taken as evidence for the very old age of alleles, as in Solanaceae.

Given the large sample size, it is likely that very old *Prunus* specificities are present in our sample. Based on the available *S-RNase* and *SFB* amino acid sequence data, the oldest specificities have an estimated age that ranges from 15 to 20 million years. Therefore, a much younger age for *Prunus* than for Solanaceae specificities is the likely cause of the difference in synonymous variability levels at the *S-RNase*, when the two groups are compared (Vieira *et al*., 2007*b*).

Present-day *Prunus* species harbour fewer than 34 specificities (Vieira *et al*., 2007*a*). Although it could be argued that the relatively small number of specificities found in *Prunus* could reflect in many cases the continuing process of domestication, a similar number of specificities has been found by Raspé and Kohn (2002) for two wild Rosaceae (Maloideae) species (*Sorbus aucuparia* and *Crataegus monogyna* presenting 24 and 27 inferred specificities, respectively). Recently Kato *et al*. (2007) reported 75 *P. lannesiana* putative specificities. However, *P. lannesiana* forms interspecific hybrids with *P. jamasakura* and *P. incisa* and it is very difficult to distinguish precisely between these hybrids and *P. lannesiana* (Kato *et al*., 2007). Thus, depending on the number of hybrids sampled, the number of specificities in *P. lannesiana* could be inflated. In contrast, it is inferred that the ancestral *Prunus* population harboured at least 102 specificities, implying a larger effective population size. The genus *Prunus* is geographically widely distributed, and therefore it is conceivable that the common ancestor to extant *Prunus* also was geographically widely distributed (Bortiri *et al*., 2002).

The relationship of *Prunus S-RNase* and *SFB* allele pairs older than 5 million years is poorly resolved (see Fig. 1). This lack of resolution can be due to the pairwise deletion option used since it implies that different regions of the protein are evolving at the same rate and this is not a realistic assumption. It could also simply reflect a problem of amino acid saturation that it is difficult to correct for. A maximum-likelihood approach using long *Prunus S-RNase* and *SFB* nucleotide sequences also shows poor resolution for old specificities (see, for instance, Nunes *et al*., 2006). Therefore, the lack of phylogenetic resolution observed in this work is not entirely attributable to using amino acid sequences, a pairwise deletion approach and minimum evolution for tree reconstruction.

Features compatible with recombination have been observed at the *S-RNase* and *SFB* genes, although the evidence is still not unequivocal (Vieira *et al*., 2003, 2007*a*; Nunes *et al*., 2006; Ortega *et al*., 2006). In principle, recombination could also significantly affect the shape and resolution of the phylogenies, although it remains to be demonstrated that the amount of recombination implied by the *SFB* and *S-RNase* data sets is enough to cause the observed pattern. It should be noted that rare recombination does not, in principle, greatly affect our estimate of the number of specificities present in the common ancestor. Only recombination events occurring after *Prunus* speciation may create the illusion that the recombinant allele already existed in the common ancestor to all living *Prunus*, thus inflating our estimate. Since the common ancestor to present-day *Prunus* species lived about 5 million years ago, and recombination is assumed to be rare at the *S*-locus, it is unlikely that our estimate is greatly inflated.

Our conclusion on the number of specificities in the ancestor to extant *Prunus* is also dependent on the assumption that at that time there was a single ancestral species. Alternatively, there could have been multiple species (all except one of the lineages going extinct), with restricted gene flow among them. Extant *Prunus* species from the same subgenus hybridize easily (Nunes *et al*., 2006; Kato *et al*., 2007; http://www.rjb.csic.es/floraiberica/PHP/cientificos.php), but not *Prunus* species from different subgenera (Surbanovski *et al*., 2007). Under the scenario of multiple species with frequent gene flow among them, the entire nuclear genome is expected to be affected. Therefore, this hypothesis can be tested by looking at variability patterns of nuclear genes in extant *Prunus* species, but such data are not yet available. Such data for multiple extant *Prunus* species are needed to achieve accurate estimates of the effective population size in the current and ancestral species, and thus to

estimate the magnitude of the reduction in population size. A reduction in population size can lead to loss of alleles and, thus, to a variability loss at the *S*-locus.

Even closely related *Prunus* species do not share more than 50% of the ancestral specificities (see Supplementary Table 7 at http://evolution.ibmc.up.pt/data/prunus_GR) and, on average, species from the same *Prunus* subgenus share more ancestral specificities than species from different subgenera. Therefore, it is possible that in *Prunus* speciation is often associated with population bottlenecks or that the process of domestication of *Prunus* species resulted in the loss of specificities.

Closely related specificities are expected to be rare in natural populations, but could be frequent when comparing closely related species (Uyenoyama *et al*., 2001; Newbigin & Uyenoyama, 2005). Many closely related *S*-alleles (those that are less than 5 million years old) from different *Prunus* species present differences at amino acid sites previously identified as being positively selected (see Supplementary Tables 6*a* and 6*b* at http://evolution.ibmc.up.pt/data/prunus_GR). Therefore, many of the closely related *S*-alleles could represent different specificities, although the lack of information on the minimum number of changes needed to create a new specificity precludes a firm conclusion. Ideally, crosses should be made between individuals harbouring these closely related alleles, but often this is not possible since individuals are not marked in the field, and/or because it is not possible to cross different species (Surbanovski *et al*., 2007). Using the *S-RNase* data set (the larger one), the frequency of closely related putative specificities (those that are less than 5 million years old) is not statistically different ($P > 0{\cdot}05$) within species (0·26% of all possible comparisons) and between species (0·20% of all possible comparisons). Therefore, theoretical expectations seem not to be fulfilled.

The simulation approach shows, on the other hand, that the empirical data are incompatible with a scenario where closely related specificities are not preferentially lost during evolution, and this conclusion is not dependent on the cluster definition used. The conclusion that closely related specificities are preferentially retained (not predicted by current theoretical models) or lost (as predicted by Uyenoyama *et al*., 2001; Newbigin & Uyenoyama, 2005) is, however, dependent on the cluster definition used, and it is not obvious which one should be used. Under the assumption that two words that can be connected by fewer than five changes belong to the same cluster, the *S-RNase* data set seems to be biased towards an excess of closely related putative specificities, and this is unexpected according to current theoretical models. An apparent excess of putative specificities may be caused by wrongly assuming that the alleles considered represent different specificities. Although little polymorphism exists within specificities (Nunes *et al*., 2006; Vieira *et al*., 2007*a*), the *S-RNase* sequences used come from different *Prunus* species, and thus it is likely that the same specificity in two different species is not represented by the same amino acid sequence. Nevertheless, this feature is not observed in the *SFB* data set, where sequences also come from closely related *Prunus* species. The *SFB* data set seems to be biased towards an excess of distantly related putative specificities, in agreement with theoretical expectations. It should be noted, however, that this is a smaller data set than the *S-RNase*. More *SFB* sequences from many *Prunus* species are needed in order to solve this issue.

Barbara Mable made helpful comments on an earlier version of this work. This work has been partially funded by Fundação para a Ciência e Tecnologia (FCT; research projects POCTI/AGG/44800/2002 and POCI/BIA-BDE/59887/2004 funded by POCI 2010, co-funded by FEDER funds). N.A.F. is the recipient of a Postdoctoral grant SFRH/BPD/26737/2006 from FCT.

## References

Bortiri, E., Oh, S.-H., Gao, F.-Y. & Potter, D. (2002). The phylogenetic utility of nucleotide sequences of sorbitol 6-phosphate dehydrogenase in *Prunus* (Rosaceae). *American Journal of Botany* **89**, 1697–1708.

Charlesworth, D. & Guttman, D. S. (1997). Seeing selection in *S* allele sequences. *Current Biology* **7**, R34–37.

Charlesworth, D., Vekemans, X., Castric, V. & Glemin, S. (2005). Plant self-incompatibility systems: a molecular evolutionary perspective. *New Phytologist* **168**, 61–69.

de Cuyper, B., Sonneveld, T. & Tobutt, K. R. (2005). Determining self-incompatibility genotypes in Belgian wild cherries. *Molecular Ecology* **14**, 945–955.

de Nettancourt, D. (1997). *Incompatibility in Angiosperms*. Berlin: Springer.

Entani, T., Iwano, M., Shiba, H., Che, F. S., Isogai, A. & Takayama, S. (2003). Comparative analysis of the self-incompatibility (*S*-) locus region of *Prunus mume*: identification of a pollen-expressed F-box gene with allelic diversity. *Genes to Cells* **8**, 203–213.

Hauck, N. R., Ikeda, K., Tao, R. & Iezzoni, A. F. (2006). The mutated *S*1-haplotype in sour cherry has an altered *S*-haplotype-specific F-box protein gene. *Journal of Heredity* **97**, 514–520.

Kato, S. & Mukai, Y. (2004). Allelic diversity of *S-RNase* at the self-incompatibility locus in natural flowering cherry populations (*Prunus lannesiana* var. *speciosa*). *Heredity* **92**, 249–256.

Kato, S., Iwata, H., Tsumura, Y. & Mukai, Y. (2007). Distribution of *S*-alleles in island populations of flowering cherry, *Prunus lannesiana* var. *speciosa*. *Genes & Genetic Systems* **82**, 65–75.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in Bioinformatics* **5**, 150–163.

Lawrence, M. J. (2000). Population genetics of the homomorphic self-incompatibility polymorphisms in flowering plants. *Annals of Botany* **85**, 221–226.

Lu, Y. (2006). Historical events and allelic polymorphism at the gametophytic self-incompatibility locus in Solanaceae. *Heredity* **96**, 22–28.

Igic, B. & Kohn, J. R. (2001). Evolutionary relationships among self-incompatibility *RNases*. *Proceedings of the National Academy of Sciences of the USA* **98**, 13167–13171.

Ma, R. C. & Oliveira, M. M. (2002). Evolutionary analysis of *S-RNase* genes from Rosaceae species. *Molecular Genetics and Genomics* **267**, 71–78.

Magallón, S., Crane, P. R. & Herendeen, P. S. (1999). Phylogenetic pattern, diversity, and diversification of eudicots. *Annals of the Missouri Botanic Garden* **86**, 297–372.

Newbigin, E. & Uyenoyama, M. K. (2005). The evolutionary dynamics of self-incompatibility systems. *Trends in Genetics* **21**, 500–505.

Nunes, M. D. S., Santos, R. A. M., Ferreira, S. M., Vieira, J. & Vieira, C. P. (2006). Variability patterns and positively selected sites at the gametophytic self-incompatibility pollen *SFB* gene in a wild self-incompatible *Prunus spinosa* (Rosaceae) population. *New Phytologist* **172**, 577–587.

Ortega, E., Boskovic, R. I., Sargent, D. J. & Tobutt, K. R. (2006). Analysis of *S-RNase* alleles of almond (*Prunus dulcis*): characterization of new sequences, resolution of synonyms and evidence of intragenic recombination. *Molecular Genetics and Genomics* **276**, 413–426.

Raspé, O. & Kohn, J. R. (2002). S-allele diversity in *Sorbus aucuparia* and *Crataegus monogyna* (Rosaceae: Maloideae). *Heredity* **88**, 458–465.

Richman, A. D. & Kohn, J. R. (2000). Evolutionary genetics of self-incompatibility in the Solanaceae. *Plant Molecular Biology* **42**, 169–179.

Richman, A. D., Uyenoyama, M. K. & Kohn, J. R. (1996). Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. *Science* **273**, 1212–1216.

Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497.

Sassa, H., Kakui, H., Miyamoto, M., Suzuki, Y., Hanada, T., Ushijima, K., Kusaba, M., Hirano, H. & Koba, T. (2007). *S* locus F-box brothers: multiple and pollen-specific F-box genes with *S* haplotype-specific polymorphisms in apple and Japanese pear. *Genetics* **175**, 1869–1881.

Schueler, S., Tusch, A. & Scholz, F. (2006). Comparative analysis of the within-population genetic structure in wild cherry (*Prunus avium* L.) at the self-incompatibility locus and nuclear microsatellites. *Molecular Ecology* **15**, 3231–3243.

Steinbachs, J. E. & Holsinger, K. E. (2002). *S-RNase*-mediated gametophytic self-incompatibility is ancestral in eudicots. *Molecular Biology and Evolution* **19**, 825–829.

Surbanovski, N., Tobutt, K. R., Konstantinović, M., Maksimović, V., Sargent, D. J., Stevanović, V. & Bosković, R. I. (2007). Self-incompatibility of *Prunus tenella* and evidence that reproductively isolated species of *Prunus* have different *SFB* alleles coupled with an identical *S-RNase* allele. *Plant Journal* **50**, 723–734.

Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proceedings of the National Academy of Sciences of the USA* **87**, 2419–2423.

Thompson, J., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The ClustalX window interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**, 4876–4882.

Tsukamoto, T., Hauck, N. R., Tão, R., Jiang, N. & Iezzoni, A. F. (2006). Molecular characterization of three non-functional *S*-haplotypes in sour cherry (*Prunus cerasus*). *Plant Molecular Biology* **62**, 371–383.

Ushijima, K., Sassa, H., Tao, R., Yamane, H., Dandekar, A. M., Gradziel, T. M. & Hirano, H. (1998). Cloning and characterization of cDNAs encoding *S-RNases* from almond (*Prunus dulcis*): primary structural features and sequence diversity of the *S-RNases* in Rosaceae. *Molecular & General Genetics* **260**, 261–268.

Ushijima, K., Sassa, H., Dandekar, A. M., Gradziel, T. M., Tao, R. & Hirano, H. (2003). 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.

Uyenoyama, M. K., Zhang, Y. & Newbigin, E. (2001). On the origin of self-incompatibility haplotypes: transition through self-compatible intermediates. *Genetics* **157**, 1805–1817.

Vieira, C. P., Charlesworth, D. & Vieira, J. (2003). Evidence for rare recombination at the gametophytic self-incompatibility locus. *Heredity* **91**, 262–267.

Vieira, J., Santos, R. A. M., Ferreira, S. M. & Vieira, C. P. (2007*a*). Molecular evolution at the *Prunus spinosa SFB*: allele diversity, population structure and amino acid sites under positive selection. *Heredity*, submitted.

Vieira, J., Morales-Hojas, R., Santos, R. A. M. & Vieira, C. P. (2007*b*). Different positively selected sites at the gametophytic self-incompatibility pistil *S-RNase* gene in the Solanaceae and Rosaceae (*Prunus*, *Pyrus* and *Malus*). *Journal of Molecular Evolution* **65**, 175–185.

Vieira, J., Fonseca, F. A. & Vieira, C. P. (2007*c*). Further support for the hypothesis that S-RNase based gametophytic self-incompatibility system evolved only once in eudicots. *Journal of Molecular Evolution*, submitted.

Wang, Y., Wang, X., Skirpan, A. L. & Kao, T. H. (2003). *S-RNase*-mediated self-incompatibility. *Journal of Experimental Botany* **54**, 115–122.

Wikstrom, N., Savolainen, V. & Chase, M. W. (2001). Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London, Series B* **268**, 2211–2220.

Wilson, D. J. & McVean, G. (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* **172**, 1411–1425.

Wright, S. (1939). The distribution of self-sterility alleles in Populations. *Genetics* **24**, 538–552.

Yang, Z. (1997). PAML a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences: CABIOS* **13**, 555–556.