Molecular Ecology (2007) **16**, 1315–1325 doi: 10.1111/j.1365-294X.2007.03233.x

Population structure at the S-locus of *Sorbus aucuparia* L. (Rosaceae: Maloideae)

OLIVIER RASPÉ* and JOSHUA R. KOHN†
National Botanic Garden of Belgium, Domein van Bouchout, B-1860 Meise, Belgium, †Section of Ecology, Behaviour and Evolution, Division of Biological Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0116, USA

Abstract

Low sequence divergence within functional alleles is predicted for the self-incompatibility locus because of strong negative frequency-dependent selection. Nevertheless, sequence variation within functional alleles is essential for current models of the evolution of new mating types. We genotyped the stylar self-incompatibility RNase of 20 *Sorbus aucuparia* from a population in the Pyrenees mountains of France in order to compare alleles found there to those previously sampled in a Belgian population. Both populations returned 20 different alleles from samples of 20 individuals, providing maximum-likelihood estimates of 24.4 (95% CI 20–34) alleles in each. Ten alleles occurred in both samples. The maximum likelihood (ML) estimate of the overlap in the alleles present in both populations was 16, meaning that an estimated eight alleles are private to each population, and a total of 32 alleles occur across the two populations examined. We used Fisher's (1961) missing plot method to estimate that 40 alleles occur in the species. In accord with population genetics theory, we observed at most one synonymous sequence difference between copies of alleles sampled from the different populations and no variation within populations. Phylogenetic analysis shows that nearly every allele in *S. aucuparia* arose prior to divergence of this species from members of three different genera of the Rosaceae subfamily, Maloideae. Lack of observable sequence variation within alleles, coupled with the slow pace of allelic relative to taxonomic diversification, implies that finding intermediate stages in the process of new allele creation will be difficult in this group.

Keywords: balancing selection, population structure, Rosaceae, self-incompatibility, S-locus, S-RNase

Received 14 July 2006; revision accepted 13 November 2006

Introduction

Strong negative frequency-dependent selection acting on the gametophytic self-incompatibility locus of flowering plants leads to both very large numbers of alleles in populations (reviewed in Lawrence 2000; Castric & Vekemans 2004) and large sequence divergence among them (Ioerger *et al*. 1990; Richman & Kohn 2000). Under single-locus gametophytic self-incompatibility, haploid pollen is rejected before fertilization if the allele it carries at the self-incompability locus (S-locus) matches either allele in the diploid pollen recipient. Rare alleles have a mating advantage while common alleles are more frequently rejected (Clark & Kao 1994). Novel alleles accumulate in populations until the effects of negative frequency-dependent selection are balanced by drift (Wright 1939).

Single-locus gametophytic self-incompatibility is found in many families of flowering plants (Weller *et al*. 1995), with 12–45 alleles commonly found within single populations (Lawrence 2000; Castric & Vekemans 2004). In addition to their large numbers, alleles in populations show extreme sequence divergence (Ioerger *et al*. 1990), both because selection favours sequence changes that result in novel mating types and because, whenever alleles drift towards extinction, selection acts to increase their frequency. This greatly increases the coalescence time of functionally distinct alleles (Takahata 1990, 1993; Vekemans & Slatkin 1994), allowing for the accumulation of large amounts of sequence variation. The predicted long coalescence time of functional polymorphism at the S-locus is confirmed by the common observation that an allele drawn from a certain

Correspondence: Joshua R. Kohn, Fax: 858-534-7108; E-mail: jkohn@ucsd.edu

species is often more closely related to an allele from a species in another genus than to other alleles in its own species (Ioerger *et al*. 1990). Observation of multiple *trans*-generic alleles implies that S-locus polymorphism present in the common ancestor of sampled species persisted down to the present day (Igic *et al*. 2004, 2006).

In contrast to extreme divergence among functionally distinct alleles, the coalescence time of neutral variation (variation that does not alter the mating specificity) at the S-locus is expected to be short relative to neutral variation at loci that are not subject to diversifying selection. This is because the effective population size of an S-allele is only $1/n$ times that of the population as a whole, where n is the number of S-alleles maintained (Clark & Kao 1994). With such shortened coalescence time, little neutral variation is expected within a deme (Clark & Kao 1994; Vekemans & Slatkin 1994).

While much is known about S-allele number and sequence variation from single populations, multiple populations of the same species have been sampled less frequently (Richman *et al*. 1995; Charlesworth *et al*. 2003; Lu 2006). Such sampling is important in order to examine several predictions regarding evolution at the S-locus. The extent of subdivision at a locus under negative frequency-dependent selection is expected to be lower than for neutral loci, because a migrant allele entering a deme in which that allele was absent will be strongly favoured. This increases the effective migration rate at the S-locus in comparison with neutral variation (Schierup 1998; Schierup *et al*. 2000; Muirhead 2001), reducing the ability of geographical separation to promote among-population divergence.

Finally, the evolution of a new specificity at the S-locus is a difficult issue and the extent of population subdivision at the S-locus bears critically upon it. Evolution of new mating types is difficult because in all known cases, including the system studied here, the S-locus contains separate, linked, style and pollen genes that interact to confer specificity. Mutations that alter specificity in one gene are expected to result in a self-compatible haplotype, whose persistence under conditions favouring self-incompatibility is uncertain. Only concerted changes in both genes result in the evolution of a new functional allele (Charlesworth 2000).

Several solutions to the puzzle of the origin of new S-alleles have been suggested. The observation that a single amino acid change can cause expression of a stylar S-allele with dual rejection specificity led Matton *et al*. (1999) to propose that new functional alleles might arise through a transitional dual-specificity state, avoiding self-compatibility. Uyenoyama & Newbigin (2000) showed that such a process was likely to lead to the evolution of a new functional allele only if the pollen gene evolved in advance of the style gene, rather than the converse. A second scenario (Uyenoyama *et al*. 2001) involves an intermediate self-compatible phase, which may persist in populations given certain levels of inbreeding depression and selfing rates of the self-compatible form. A key feature of both models, however, is that the new allele replaces its progenitor in the population in which it evolves. However, if the species were sufficiently subdivided, evolution of a new specificity in one deme, followed by subsequent migration to other demes, might lead to the presence of both the progenitor and descendant alleles in populations (Uyenoyama & Newbigin 2000; Uyenoyama *et al*. 2001). Using a verbal model, Chookajorn *et al*. (2004) suggested that a series of small functional changes in both style and pollen genes could lead to the evolution of new alleles within demes, without the loss of the progenitor allele. For such a model to work, however, incremental intrapopulation variation within specificities must be maintained.

The extent of variation within gametophytic S-alleles of the same specificity needs to be evaluated empirically. If variation is common within populations, then populations must be quite large. If functionally neutral variation is found among populations then gene flow must be very rare, despite the expected increased effective migration rate at the S-locus. On the other hand, if neutral variation is rare or absent, models for allele formation may still be correct, but the development of new alleles by these processes may occur over very long periods, or only under unusual demographical circumstances. One way to infer the pace of new allele formation is to examine the phylogeny of alleles drawn from different taxa. If close pairs of alleles usually come from different species or genera, then alleles are rarely forming over the life spans of individual species. On the other hand, if species harbour monophyletic clusters of alleles, then considerable diversification has occurred since species divergence.

Population subdivision can affect not only the amount of sequence variation within S-alleles, but also the number of functionally distinct S-alleles found in a species (Schierup 1998; Schierup *et al*. 2000; Muirhead 2001). As for neutral variation, subdivided species can harbour more allelic variation than panmictic species, but under balancing selection the effect of subdivision is quite dynamic. Under island models with strong balancing selection, both Schierup (1998) and Muirhead (2001) showed that when migration rates are very low (less than or equal to the origination rate of new alleles) the number of S-alleles maintained is higher than for a panmictic population of the same total size. However, unlike the neutral case, a counterintuitive result is that at intermediate rates of migration (higher than the origination rate but lower than a critical value; see Muirhead 2001), the total number of alleles maintained is slightly lower than if the population were panmictic. When many demes are examined, the average fraction of unique alleles they contain (alleles not found in other demes) can provide information on the migration rate among them (Muirhead 2001).

In this paper we examine S-alleles present in a population of *Sorbus aucuparia* from the Pyrenees mountains of southern France in order to compare them to a previously censused population from Belgium (Raspé & Kohn 2002). We compare the number of S-alleles and ask how much overlap there is in the alleles present in each population as a measure of population subdivision at the S-locus. We further estimate the total number of alleles in the two populations combined, and in the species. In addition, we determine the range of sequence variation observed within vs. between populations. In particular, we ask whether small sequence differences are present between populations that are absent within them. Finally, we put the 30 alleles recovered from *S. aucuparia* in this and our previous study (Raspé & Kohn 2002) onto a phylogeny of alleles from the Rosaceae subfamily, Maloideae, in order to infer the pace of allelic diversification relative to species diversification in the group.

Materials and methods

RNA isolation, S-allele amplification and sequencing

Twigs bearing inflorescences were sampled from 20 individuals in a population located around the Somport Pass in the Pyrenees. Approximately 30 styles of freshly opened flowers, or flower buds 1–2 days prior to anthesis, were frozen in liquid nitrogen and stored at –80 °C. Styles were ground on dry ice and total RNA was isolated using the Purescript® RNA isolation kit (Gentra Systems) according to the manufacturer's instructions, and resuspended in 25 μL of RNA hydration solution. *S*-alleles were amplified by PCR using-3′RACE. First, cDNA synthesis was performed with 2–3 μL of RNA solution, using the Sensiscript™ RT kit (Qiagen). The cDNA was primed with an oligo composed of 17 T-residues at the 3′ end plus with sequence 5′-GACTCGAGTCGACATCGATCC-3′ at the 5′ end. A 2 μL aliquot of the cDNA synthesis reaction was then used in a 25 μL amplification reaction, using a proofreading polymerase (Platinum® Pfx, Invitrogen). The amplification was performed with the adaptor given above and an internal primer anchored in either the conserved region C1 (5′-CAATTTACGCAGCAATATCAGC-3′) or, if the latter failed to amplify two alleles, the conserved region C2 (5′-GTTYACBGTTCACGGWTTGTGGC-3′; Raspé & Kohn 2002). The PCR mix contained 2× enzyme buffer, 1 mM $MgSO_4$, 250 μM of each of the four dNTPs, 300 nM of each primer and 0.625 U of Pfx polymerase. Reactions were run using one step of 4 min at 94 °C, 35 cycles of 15 s at 94 °C, 30 s at 55 °C and 1 min at 68 °C, followed by one step of 4 min at 68 °C. The PCR products were then electrophoresed on 1.6% agarose gels at 4 V/cm for 3 h, with the aim of separating the two alleles. Plugs were taken from either of the two bands when the two alleles differed in size, or from the single band when they did not. The DNA from plugs was used as a template for reamplification using either Pfx polymerase as above or a nonproofreading polymerase (Taq DNA polymerase, Roche) when subsequent cloning was necessary because the two alleles could not be separated by electrophoresis. PCR conditions used with the Taq DNA polymerase were the same as with the Pfx polymerase except that the PCR contained 1× Taq buffer, 1.5 mM $MgCl_2$ in place of $MgSO_4$ and 200 μg/mL BSA, and that the elongation steps were performed at 72 °C. Cloning was performed with the TOPO TA Cloning® kit for sequencing (Invitrogen) and the cloned alleles were amplified by PCR and identified by size differences or RFLP analysis. PCR products were then purified with the NucleoSpin® PCR purification kit (Macherey-Nagel), and sequenced using BigDye terminators and an ABI377XL automatic sequencer (Applied Biosystems). We used the RT-PCR approach because PCR from genomic DNA using universal primers amplifies both S-alleles and related non-S RNases (personal observation; Sassa *et al*. 1996). RT-PCR on total RNA isolated from styles improved both the reliability and exclusivity of S-RNase amplification. The disadvantage of this approach is that sequences of the single intron of the S-RNase gene, a potentially useful source of information on silent variation, were not obtained.

Analysis

The number of S-alleles present in the population was estimated from the sample using the method of Paxman (1963). The 95% likelihood interval of this estimator was computed as in O'Donnell & Lawrence (1984). These estimators assume isoplethy of S-alleles, an assumption that was tested using the Mantel (1974) statistic (see Campbell & Lawrence 1981). Overlap in the alleles present in the Pyrenean and Belgian populations was estimated using the method of O'Donnell *et al*. (1993). The number of S-alleles in the species was then estimated using the 'missing plot' method of Fisher (1961; see also O'Donnell *et al*. 1993). This method assumes that each population sample represents an independent random draw from the same total species' pool of alleles. This assumption cannot be evaluated when fewer than three populations are sampled, so we interpret the estimate with caution.

Phylogenetic reconstruction of S-allele diversification was performed to assess the degree of *trans*-specific evolution in the sample of *S. aucuparia* alleles. By doing so, one can assess how much diversification of the S-alleles in *S. aucuparia* occurred prior to divergence of this species from the other Maloideae in the sample. A phylogenetic tree of all 30 S-alleles from *S. aucuparia*, together with S-alleles sequenced from other species in the Rosaceae subfamily Maloideae (the genera *Malus*, *Pyrus*, and *Crataegus*), was constructed using four alleles from *Prunus avium* (Rosaceae, subfamily Amygdaloideae) as an outgroup. In the Rosaceae, alleles from *Prunus* and those from the subfamily Maloideae

form well supported, reciprocally monophyletic groups (Igic & Kohn 2001; Stienbachs & Holsinger 2002). The alignment of Maloid and *Prunus* S-alleles is 611 bp in length, which represents 94% of the mature S-RNase, with the region under conserved region 1 (our primer site) plus the first 4–6 amino acids at the 5′ end of the protein absent. Also missing from the 5′ end of our sequences is coding region for the leader peptide, usually 78 bp in length, thought to be cleaved before the protein is mature (Ishizumi *et al*. 1998). However, because the primers used to amplify the alleles varied within and between this study and that of Raspé & Kohn (2002), there is some variation in sequence length among alleles. Most *Sorbus* and *Crataegus* alleles have a length of between 447 and 611. All include the hypervariable regions a and b, except the shortest sequence (Sauc 20), which is only 249 bp long.

Amino acid sequences of all alleles were aligned using CLUSTALX (Thompson *et al*. 1994), adjusted manually, and then converted back to DNA sequences for analysis. The alignment is available from the authors. We used PAUP* version 4.0b10 (Swofford 2002) to find the best tree using a maximum-likelihood heuristic search. For this search we first used MODELTEST version 3.0 (Posada & Crandall 1998) to find the optimal model of evolution (HKY + I + G; Hasegawa *et al*. 1985), selected using the Akaike Information Criterion. Five-hundred bootstraps replicates were run to evaluate the support of nodes. This analysis failed to find a well supported root position for the Maloideae S-alleles (see Results). Because of this we ran a second analysis with only Maloideae S-alleles using mid-point rooting. For this second analysis, MODELTEST version 3.0 found the best-fit molecular model to be TrNef + I + G (Kimura 1980; Posada & Crandall 1998). Other phylogenetic procedures were the same as in the first analysis.

A weakness of this study is that we do not demonstrate whether divergent S-RNases encode the same or different specificities, a problem not uncommon in studies of the S-locus, particularly in trees or other species not amenable to greenhouse culture and crossing (Raspé & Kohn 2002; Nunes *et al*. 2006). To partially compensate for this, we compared sequence divergences among *S. aucuparia* alleles to those found among 16 alleles from apple (*Malus domestica*) another member of the Maloideae in which the correspondence between sequence and mating type has been demonstrated for many alleles (reviewed in Broothaerts 2003). Using MEGA we computed pairwise amino acid divergences, rates of nonsynonymous substitutions per nonsynonymous site (K_a), and synonymous substitutions per synonymous site (K_s) for sets of alleles from each species. We assume any allele pairs in *Sorbus* that are as divergent, or more divergent than the closest demonstrated alleles in *Malus* represent different specificities. If any allele pairs in *Sorbus* are less divergent than any pair of *Malus* alleles, we ask additionally whether these allele pairs appear to be *trans*-generic in the phylogenetic analysis on the assumption that neutral variation within an S-allele would have a shorter coalescent time than genera. If such pairs are not *trans*-generic, then they may represent divergent copies of the same specificity.

Results

We recovered 20 different alleles from 20 diploid individuals sampled in the Pyrenees, precisely the same recovery rate and sample size as in the previously censused Belgian population (Raspé & Kohn 2002). This yields an estimate 24.4 alleles with a 95% likelihood interval of 20–34 alleles. Allele frequencies did not differ from the expectation of isoplethy ($\chi^2 = 10.55$, NS), which was also the case for the Belgian population (Raspé & Kohn 2002).

Of the 20 alleles sampled in the Pyrenees population, eight were found only once, six twice, four three times and two four times (Table 1). No sequence differences were found between copies of the same allele amplified from different individuals. Six alleles were found to be identical to copies previously sampled from Belgium (alleles Sauc 2, 3, 6, 13, 15, and 19). Another four alleles (Sauc 4, 7, 10, 18) were found to differ by one to three nucleotides from sequences first recorded in the Belgian population. We then used allele-specific primers to amplify these alleles from the Belgian population using genomic DNA stored from previous studies in order to confirm whether between-population differences were real rather than errors introduced by cDNA

Table 1 S-locus genotypes of 20 *Sorbus aucuparia* individuals from the Pyrenees and Belgian populations. Genotypes are given by the allele numbers. Alleles 1–20 were previously sequenced from the population sampled in Belgium (Raspé & Kohn 2002)

plant	Pyrenees	Belgium
1	21,22	10,5
2	21,23	5,12
3	15,25	7,15
4	24,27	10,5
5	3,26	17,20
6	22,29	2,10
7	7,10	3,11
8	24,25	10,4
9	23,29	10,16
10	19,26	8,10
11	6,28	14,19
12	6,19	6,7
13	23,25	7,17
14	19,21	2,17
15	4,25	1,19
16	7,18	9,13
17	2,27	1,16
18	24,30	17,18
19	13,23	2,12
20	13,26	2,11

synthesis, amplification, cloning, and sequencing procedures. For three of the four alleles, Belgian sequences amplified from DNA using allele-specific primers matched exactly sequences from the Pyrenees population, indicating errors in the original Belgian RT-PCR sequences. For the fourth allele (Sauc 7) we were unable to amplify the original Belgian sequence so we cannot confirm whether the single synonymous change differentiating the copy from the Belgian and Pyrenees populations is real. In sum, for the 10 alleles amplified from the two populations, we find at most a single nucleotide difference, and no amino acid differences.

We used the method of O'Donnell *et al*. (1993) to find the maximum likelihood (ML) estimate of the overlap in the alleles found in the two populations, which was 16 with 95% confidence interval of 11–18. This means that if each population contains the predicted 24 alleles, then eight are expected to be unique to each population and 16 common to both, for a total of 32 alleles in the two populations. This estimate is very close to the ML estimate for the total number of alleles if the samples from the two populations are pooled. In that case, discovery of 30 alleles from 40 individuals provides an estimate of 32.6 (95% CI 30–38) alleles. One consequence of structure at the S-locus is that there should be more alleles in samples of the same size drawn from among, rather than within, populations. To test for this we repeatedly sampled 20 individuals drawn randomly from among the 40 *S aucuparia* genotypes in our two studies. Of 500 randomizations, only 12 (2.4%) contained only 20 alleles, as observed in samples from both the Belgian and Pyrenean populations, while all others contained more than 20 (Fig. 1). The probability of two such observations if the samples came from a panmictic population is $P < 0.001$. The mean number of alleles in these random samples of 20 genotypes was 23.4, providing a ML estimate of 32.3 alleles, virtually the same estimate as by the other methods.

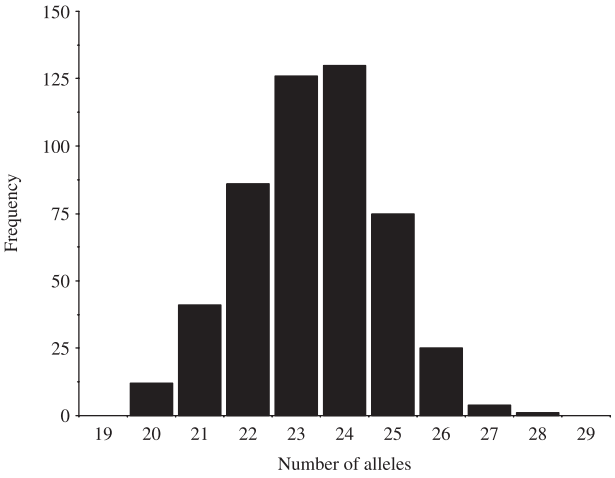

Fig. 1 The number of alleles recovered when sampling 20 genotypes randomly drawn from the two populations combined. Results of 500 randomizations are shown. The value of 20 alleles, as found in both the Pyrenean and Belgian populations, was observed 12 times (2.4%). The probability of drawing only 20 alleles twice if genotypes are randomly distributed over the two populations is $P < 0.001$. The mean value is 23.3 alleles.

Fisher's (1961) method for estimating the number of alleles missing from a set of samples yields an estimate of 10 meaning that, in addition to the 30 alleles so far recovered from this species, an estimated 10 more exist, bringing the species-wide estimate to 40 alleles. We must assume no geographical structure at the S-locus in order to apply Fisher's (1961) method, an assumption that cannot be directly tested when fewer than three populations are sampled (Fisher 1961). The preceding analysis of overlap suggests some structure, though its extent, while significant, is not enough to alter the discovery rate of alleles very much (Fig. 1). The general lack of even silent sequence variation in S-alleles drawn from different populations, and the relatively low frequency of private alleles (Muirhead 2001) estimated even when just two populations were sampled, are indications that the rate of gene flow is not extremely low. Therefore, our estimate of 40 alleles is not expected to differ greatly from the actual value in the species.

Phylogenetic analysis of S-alleles from *S. aucuparia* together with alleles from other Maloideae is shown in Fig. 2. The ML method was unable to find a reliable root position, apparently because of the great genetic distances between alleles from *Prunus* and the Maloideae. We therefore ran a second analysis using only alleles from the Maloideae and mid-point rooting (Fig. 3). In nearly every case, an allele from *S. aucuparia* is at least as closely related to an allele from another genus, than to any other allele from its own species. Of the 30 alleles from *S. aucuparia*, the only monospecific sister pair are alleles 26 and 27. The same can be said of alleles from the other Maloideae represented; they too are nearly always at least as closely related to alleles from other species as to other alleles from their own. Notably absent are any substantial monophyletic clades of alleles restricted to a single genus.

Average divergences among putative alleles in *S. aucuparia* are quite comparable to those observed in *Malus domestica* (Table 2). However, six allele pairs in *S. aucuparia* show

Table 2 Mean (SD) pairwise rates of nonsynonymous (K_a) and synonymous (K_s) nucleotide substitution, and mean pairwise proportion of amino acid differences (pAA) for alleles from *Malus domestica* and *Sorbus aucuparia*. Allele Sauc20 was not included in this analysis because of the short length of sequence recovered for this allele (Raspé & Kohn 2002)

	Malus $n = 16$	*Sorbus* $n = 29$
K_a	0.211 (0.04)	0.209 (0.05)
K_s	0.169 (0.05)	0.127 (0.05)
pAA	0.343 (0.06)	0.337 (0.08)

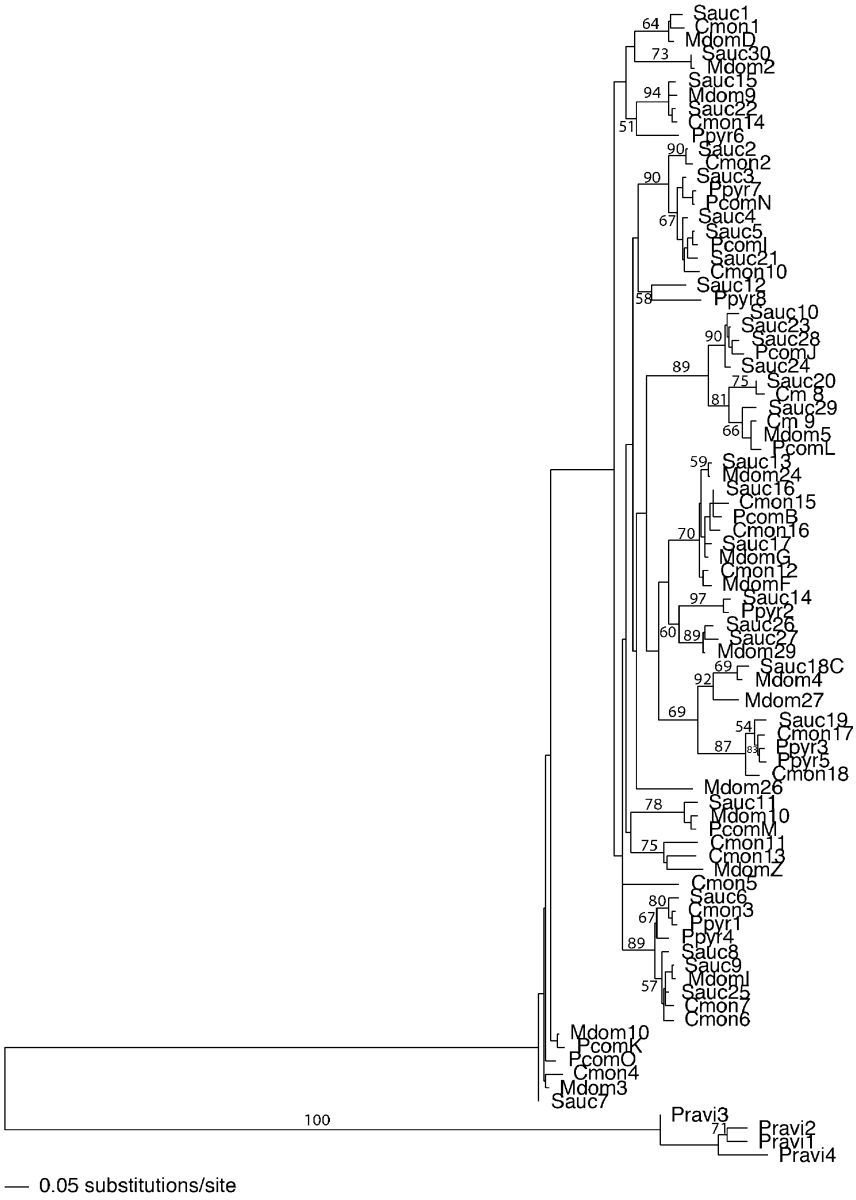

Fig. 2 Maximum-likelihood phylogeny of 80 S-alleles from the Rosaceae subfamily Maloideae represented by *Sorbus aucuparia* (Sauc, 30 alleles, GenBank accessions AF504253–72, and pending), *Crataegus monogyna* (Cmon, 18 alleles, AF504273–90), *Malus domestica* (Mdom, 16 alleles, D50837, AB017636, AB019184, AB032246, AB052683, AB062100, AF016918, AF016920, AF239809, AF327222–3, AY039702, U12199–200, U19791, U19793), *Pyrus pyrifolia* (Ppyr, 8 alleles, AB002139, AB002141–3, AB009385, AB014073, AB025421, AB104908) and *Pyrus communis* (Pcom, 8 alleles, AF457594–5, AF518319, AY103408–9, AY159323, AY195840, AY261994). The tree is rooted with S-alleles from *Prunus avium* (Pravi, 4 alleles, AB028154, AJ298310–2) from the subfamily Amygdaloideae, No root position within the Maloideae was supported by the 50% bootstrap consensus tree, suggesting that the genetic distance between *Prunus* and Maloideae alleles is too great to reliably infer the placement of the root.

K_a values lower than the minimum value observed among *M. domestica* alleles (K_a = 0.027 for alleles MdomeF and MdomeG). These pairs are Sauc8 and 9 (K_a = 0.023); 10 and 23 (K_a = 0.016); 9 and 25 (K_a = 0.015); 15 and 22 (K_a = 0.024); 23 and 24 (K_a = 0.019); 23 and 28 (K_a = 0.022). Because 30 alleles have now been recovered from *S. aucuparia* while 16 appear in GenBank for *M. domestica*, finding some less-divergent allele pairs in *Sorbus* may be expected. We cannot disprove the possibility that close pairs might represent divergent copies of the same specificity. However, in each case, the terminal branch leading to each allele of a close pair attaches to the phylogenetic tree at a node that is ancestral to an allele in another genus of Maloideae. The *trans*-generic divergence of each of these pairs suggests they represent different functional alleles. In addition, each allele pair differs by a minimum of five amino acids over the portion of the gene compared. Pairwise amino acid divergences are: Sauc8 and 9, 7; 10 and 23, 5; 9 and 25, 5; 15 and 22, 9; 23 and 24, 8; 23 and 28, 8. Nine amino acid differences separate S3 and S5 alleles in *Pyrus pyriflolia* (Fig. 3, bottom; Ishizumi *et al*. 1998) a species from which only seven alleles have been characterized. It would be odd to find several cases of *trans*-generic copies of the same functional alleles, each with multiple amino acid differences, in the absence of smaller divergences.

Discussion

Consistent with the population genetics theory for the S-locus, we find little evidence of sequence variation within

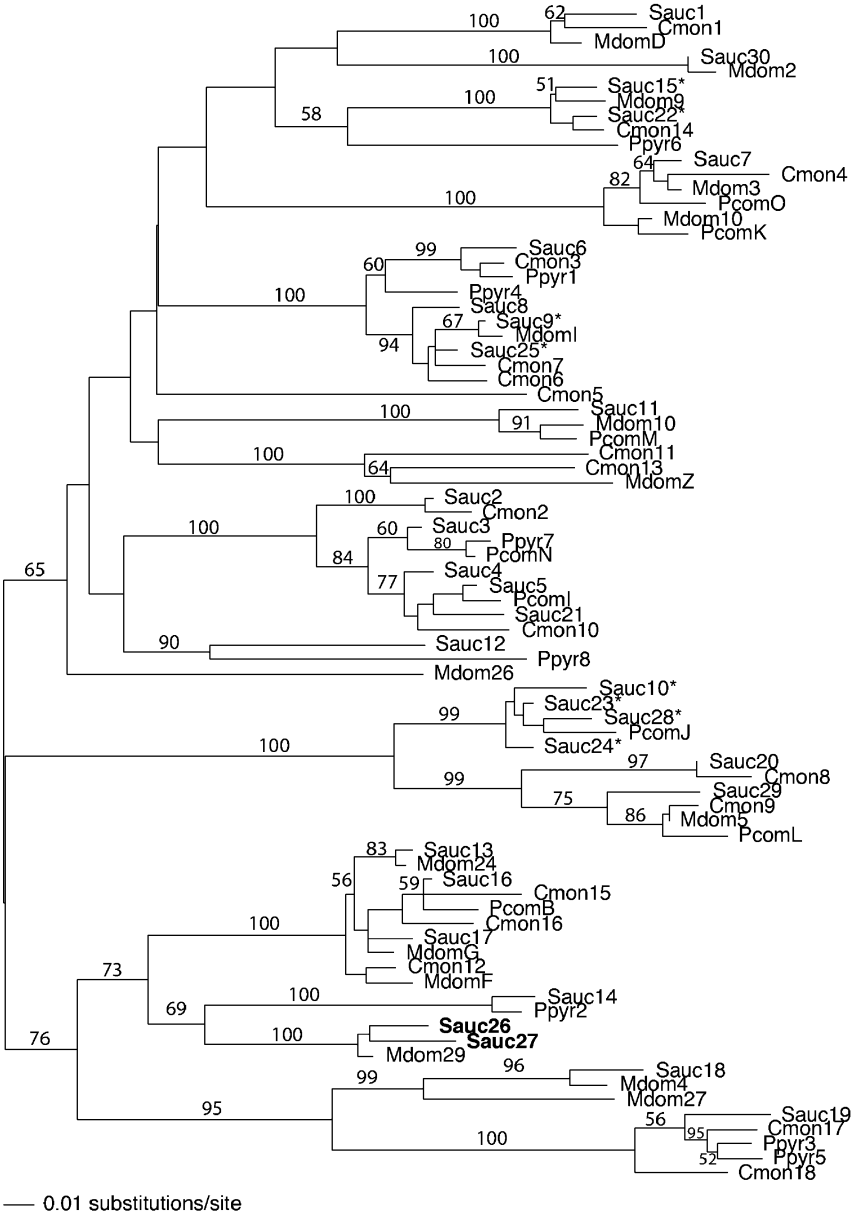

Fig. 3 Midpoint-rooted maximum-likelihood phylogeny of 80 S-alleles from the Maloideae. Boldface, sister alleles within *S. aucuparia*. Asterisks, alleles that are members of pairs within *S. aucuparia* that are less divergent than any pair of alleles from *Malus domestica* (see Results).

alleles at the S-locus in these populations separated by 950 kilometers. Alleles sequenced from the two populations are either identical, or nearly so, differing by at most one synonymous substitution, or else they differ substantially, by a minimum of five amino acids, usually more, suggesting that they represent different functional alleles. Low sequence variation within functionally distinct S-alleles, and high divergence between them, are the predicted results of negative frequency-dependent selection acting on the S-locus. Such selection preserves functional polymorphism over very long periods of time, promoting divergence among functional allele classes. At the same time, frequency-dependent selection restricts neutral polymorphism owing to low effective population sizes of functionally distinct alleles within populations, and higher effective migration rates among them.

Many population studies recover multiple copies of at least some alleles (e.g. Wang *et al*. 2001; Raspé & Kohn 2002; Vieira *et al*. 2003; see also studies reviewed by Lawrence (2000) and Castric & Vekemans 2004). Only Nunes *et al*. (2006) have reported intrapopulation variation in S-alleles. Two of 21 alleles of the S-locus gene *SFB*, thought to encode pollen specificity in the allotetraploid species *Prunus spinosa*, were found to be variable within the population studied. Variant pairs differed by one and two amino acids. The stylar RNases associated with both variants of one allele were sequenced and found to differ by five nucleotides and two amino acids. It is not yet clear whether the divergent

copies encode the same or different specificities or whether both copies encode functional alleles (Nunes *et al*. 2006). It is also unclear whether the observed variation arose within *P. spinosa*. This variation could have been inherited from *P. spinosa's* two diploid ancestors or by gene flow from pentaploid *P. insistitia,* which apparently hybridizes with *P. spinosa* in the population studied (Nunes *et al*. 2006).

Few studies have looked for sequence variation among populations. Several studies have sequenced alleles from more than one population and not reported any differences (Richman *et al*. 1995; Richman *et al*. 1995; Lu 2006) though finding polymorphism within S-alleles was not a primary goal of any of these studies. Charlesworth *et al*. (2003) found a maximum of two nonsynonymous changes within stylar alleles of the sporophytic S-locus of *Arabidopsis lyrata* when the same S-alleles from populations as geographically disparate as Iceland, Scotland, and North Carolina were compared. Most S-alleles from these disparate populations were identical. Walker *et al*. (1996) cloned the S_3 allele of *Papaver rhoeas* cv *Shirley* and from a wild population from Spain. In the coding region, just two synonymous substitutions were found. In a related study, May *et al*. (1999) found very low sequence divergence within fungal mating type alleles from a worldwide sample. The major exception to the finding of little or no polymorphism within functional allele classes across populations is the study of (Miege *et al*. 2001) who found, among different cultivars of *Brassica oleracea*, two different types of S_2 alleles whose stylar proteins at the S-related kinase (*SRK*) locus differ by as many as 12 amino acids. Evolution of the sporophytic S-locus in *Brassica* appears to involve concerted changes in *SRK* and the associated S-locus glycoprotein (*SLG*) loci, perhaps through gene conversion Miege *et al*. (2001). It is not clear whether the rather large degree of polymorphism observed within the S_2 allele in this species represents a general facet of S-locus evolution in natural populations, but it is worth noting that *Arabidopsis lyrata,* also a member of the Brassicaceae, lacks *SLG* and shows limited intrahaplotype S-locus polymorphism (Charlesworth *et al*. 2003).

While the populations studied here have substantially overlapping sets of S-alleles, each population also contains at least several alleles not found in the other population. Ten alleles were sampled in both populations while 10 were unique to each population sample. Since the ML estimate of the number of alleles occurring in each population is 24, with 16 being common to both, each population is estimated to contain eight alleles that are not found in the other. The existence of 'private' alleles might be taken as evidence of structure at the S-locus, but structure can mean several things. Private alleles can result from the fact that local populations are too small to support the number of alleles found in the entire species, even if there is no spatial structure to where alleles occur across the geographical range of the species. In that case, each population contains a random draw of the alleles found in the species. Conversely, if gene flow is sufficiently restricted, drift may lead to the differentiation of allele sets in the geographically separated regions. Under such a scenario, variation at silent sites is expected to accumulate and provide a geographical pattern. We have little or no evidence of neutral variation.

It is entirely possible that the evidence for private alleles in this study arises solely from the limited spatial scale of sampling in each locality. Because allele distributions must be affected by accidents of recent dispersal and establishment in the secondary succession habitats occupied by these small trees, local populations cannot be expected to be at evolutionary equilibrium. More thorough sampling over a somewhat broader spatial scale in both regions might reduce or eliminate the estimate of private alleles.

The ecological and geographical history of *Sorbus aucuparia* may contribute to the lack of sequence variation within alleles recorded in this study. Allozyme and cpDNA studies found significant differentiation between Belgian and Pyrenean populations, but G_{ST} values were low in comparison to values for other tree species sampled at similar spatial scales (Raspé & Jacquemart 1998; Raspé *et al*. 2000). The generally low levels of differentiation were attributed to the ecology of this bird-dispersed, insect-pollinated, self-incompatible species. In addition, there may have been little time for differentiation to arise given that Belgian populations may represent postglaciation re-invasion of northern Europe from a Pleistocene refugium in SW Europe. A separate refugium in SE Europe has been implicated in studies of genetic variation in other European trees and shrubs (e.g. Comps *et al*. 1990; Breitenbach-Dorfer *et al*. 1992; Dumolin-Lapègue *et al*. 1997; King & Ferris 1998; Hampe *et al*. 2003; Rendell & Ennos 2003; Heuertz *et al*. 2004), but the number and localization of refugia remain unknown in *S. aucuparia*. Examination of the S-locus from eastern European populations could reveal some structure. Nevertheless, whatever structure might be present at the S-locus within *S. aucuparia*, the relationships among alleles from different Maloideae genera suggests it has contributed little, if anything, to allelic diversification.

For gametophytic self-incompatibility, studies of multiple populations of the same species have generally revealed substantial overlap in the alleles present (Emerson 1939; O'Donnell *et al*. 1993; Richman *et al*. 1995; Kato & Mukai 2004). Lu (2006) found lower overlap in the alleles present in samples from multiple populations of *Physalis longifolia* than *Solanum carolinense* and concluded that migration rates were lower in the former species. However, *P. longifolia* populations appear to harbour considerably more alleles (Lu 2001) than do populations of *S. carolinense* (Richman *et al*. 1995). Since sample sizes were limited and equal for the two species, lower overlap among population samples of *P. longifolia* might be expected. For sporophytic self-incompatibility, the picture is less clear. The genetics of this

system make theoretical predictions more difficult (Schierup 1998), but Glémin *et al*. 2005) report a high frequency of private alleles among Corsican populations of *Brassica insularis*.

The phylogenetic analysis of alleles from *S. aucuparia* and other members of the Rosaceae subfamily, Maloideae, provides evidence concerning the pace of allelic relative to taxonomic diversification in the group. Nearly every allele sequenced from *S. aucuparia* appears older than the genus *Sorbus*. With only one exception, each allele from *S. aucuparia* joins the phylogenetic tree at a node that is ancestral to an allele from another genus. The only exception is the sister relationship of *S. aucuparia* alleles 26 and 27. Apparently, very few if any alleles have evolved since the most recent common ancestor *S. aucuparia* and the other sampled species. This finding is not likely to be greatly affected by the level of sampling in *S. aucuparia* or the other species. In *S. aucuparia*, we estimate that the 30 alleles so far sequenced represent the majority of the alleles found in the species. While some unsampled alleles could be younger than the genus *Sorbus*, rates of allelic relative to phyletic diversification must be low because there is virtually no evidence of monophyletic groups of alleles unique to *S. aucuparia* in the phylogeny (Figs 2 and 3). Undersampling of alleles from the other taxa on the phylogeny could only cause an overestimate of the number of sister relationships among alleles from *S. aucuparia*. The lack of sequence variation within alleles of *S. aucuparia*, coupled with the lack of phylogenetic evidence for allelic diversification within species makes it appear unlikely that transitional states in the creation of new alleles will be easily found in this group.

One possibility, which the phylogenetic study cannot rule out, is that new alleles are arising within species but descendent alleles replace their immediate progenitor leaving no apparent diversification (Uyenoyama & Newbigin 2000; Uyenoyama *et al*. 2001). This would result in no change in allele number and no development of monophyletic clades private to different taxa. However, the lack of variation within alleles either within or between populations means that this process, if it is occurring, is difficult to document in *S. aucuparia*.

Close sister relationships among alleles from different taxa might suggest that hybridization has played a role in distributing alleles among Maloideae. This appears unlikely because minimum divergences between alleles from different species (Fig. 3) are generally substantial, indicating that any hybridization events cannot have been recent. That is, while the same allelic lineages appear to be represented in different genera, the alleles within each lineage but drawn from different genera show considerable divergence (Fig. 3). These divergences likely represent the build-up of change since species divergence.

In summary, our findings are consistent with the predictions of population genetics theory for loci under balancing selection. No sequence variation within putative alleles was seen within populations and very little if any appears to occur between populations. The pace of allelic diversification has been slow relative to taxonomic diversification. All, or nearly all of the alleles present in *S. aucuparia* appear to be at least as old as the common ancestor of the sampled genera of Maloideae. Because of the slow rate of diversification of alleles relative to species, it appears unlikely that surveys of natural populations will succeed in finding intermediate stages of the process of novel allele formation.

Acknowledgements

We wish to thank the authorities of the Parc National des Pyrénées for granting the permission to sample *S. aucuparia* flowers within the park. We also wish to thank Jean-Jacques Lazare for valuable information on the flowering time. We are grateful to Bénédicte Purnelle (Université catholique de Louvain, Unité de Biochime Physiologique) for her sequencing work. This work was partially supported by a grant from the National Science Foundation (DEB 01–08173) to JRK.

References

Breitenbach-Dorfer M, Pensker W, Hacker R, Müller F (1992) Clone identification and clinal allozyme variation in populations of *Abies alba* from the Eastern Alps. *Plant Systematics and Evolution*, **181**, 109–120.

Broothaerts W (2003) New findings in apple S-genotype analysis resolve previous confusion and request the re-numbering of some S-alleles. *Theoretical and Applied Genetics*, **106**, 703–714.

Campbell JM, Lawrence MJ (1981) The population genetics of the self-incompatibility polymorphism in Papaver rhoeas. II. The number and frequency of S-alleles in a natural population (R106). *Heredity*, **46**, 81–90.

Castric V, Vekemans X (2004) Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. *Molecular Ecology*, **13**, 2873–2889.

Charlesworth D (2000) How can two-gene models of self-incompatibility generate new specificities? *Plant Cell*, **12**, 309–310.

Charlesworth D, Bartolomé C, Schierup MH, Mable BK (2003) Haplotype structure of the stigmatic self-incompatibility gene in natural populations of Arabidopsis lyrata. *Molecular Biology and Evolution*, **20**, 1741–1753.

Chookajorn T, Kachroo A, Ripoll DR, Clark AG, Nasrallah JB (2004) Specificity determinants and diversification of the *Brassica* self-incompatibility pollen ligand. *Proceedings of the National Academy of Sciences of the USA*, **101**, 911–917.

Clark AG, Kao T-H (1994) Self-incompatibility: theoretical concepts and evolution. In: *Genetic Control of Self-Incompatibility and Reproductive Development in Flowering Plants* (eds Williams EG, Clarke AE, Knox RB), pp. 220–242. Kluwer Academic Publishers, Boston.

Comps B, Thiébaut B, Paule L, Merzeau D, Letouzey J (1990) Allozymic variability in beechwoods (*Fagus sylvatica* L.) over central Europe: spatial differentiation among and within populations. *Heredity*, **65**, 407–417.

Dumolin-Lapègue S, Demesure B, Fineschi S, Le Corre V, Petit RJ (1997) Phylogeographic structure of white oaks throughout the European continent. *Genetics*, **146**, 1475–1487.

Emerson S (1939) A preliminary survey of the *Oenothera organensis* population. *Genetics*, **24**, 524–537.

Fisher RA (1961) The possible differentiation in wild populations of *Oenothera organensis*. *Australian Journal of Biological Sciences*, **14**, 76–78.

Glémin S, Gaude T, Guilleman M-L, Lourmas M, Olivieri I, Mignot A (2005) Balancing selection in the wild: Testing population genetics theory of self-incompatibility in the rare species *Brassica insularis*. *Genetics*, **171**, 279–289.

Hampe A, Arroyo J, Jordano P, Petit RJ (2003) Rangewide phylogeography of a bird-dispersed Eurasian shrub: contrasting Mediterranean and temperate glacial refugia. *Molecular Ecology*, **12**, 3415–3426.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Jounal of Molecular Evolution*, **22**, 160–174.

Heuertz M, Fineschi S, Anzidei M *et al.* (2004) Chloroplast DNA variation and postglacial recolonization of common ash (*Fraxinus excelsior* L.) in Europe. *Molecular Ecology*, **13**, 3437–3452.

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

Igic B, Bohs L, Kohn JR (2004) Historical inferences from the self-incompatibility locus. *New Phytologist*, **161**, 97–105.

Igic B, Bohs L, Kohn JR (2006) Ancient polymorphism reveals unidirectional breeding system shifts. *Proceedings of the National Academy of Sciences of the USA*, **103**, 1359–1363.

Ioerger TR, Clark AG, Kao T-H (1990) Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proceedings of the National Academy of Sciences of the USA*, **87**, 9732–9735.

Ishizumi T, Shinkawa T, Sakiyama F, Norioka S (1998) Primary structural features of rosaceous S-RNases associated with gametophytic self-incompatibility. *Plant Molecular Biology*, **37**, 931–941.

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

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

King RA, Ferris C (1998) Chloroplast DNA phylogeography of *Alnus glutinosa* (L.) Gaertn. *Molecular Ecology*, **7**, 1151–1161.

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

Lu Y (2001) Roles of lineage sorting and phylogenetic relationship in the genetic diversity at the self-incompatibility locus. *Heredity*, **86**, 195–205.

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

Mantel N (1974) Approaches to a health research occupancy problem. *Biometrics*, **30**, 355–362.

Matton DP, Luu DT, Xike Q, Laublin G, O'Brien M, Maes O, Morse D, Cappadocia M (1999) Production of an S RNase with dual specificity suggests a novel hypothesis for the generation of new S alleles. *Plant Cell*, **11**, 2087–2098.

May G, Shaw F, Badrane H, Vekemans X (1999) The signature of balancing selection: fungal mating compatibility gene evolution. *Proceedings of the National Academy of Sciences of the USA*, **96**, 9172–9177.

Miege C, Ruffino-Chable V, Schierup MH, Cabrillac D, Dumas C, Gaude T, Cock JM (2001) Intrahaplotype polymorphism at the Brassica *S* locus. *Genetics*, **159**, 811–822.

Muirhead CA (2001) Consequences of population structure on genes under balancing selection. *Evolution*, **55**, 1532–1541.

Nunes MDS, Santos RAM, Ferreira SM, Vieira J, Vieira CP (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.

O'Donnell S, Lawrence MJ (1984) The population genetics of the self-incompatibility polymorphism in Papaver rhoeas. IV. The estimation of the number of alleles in a population. *Heredity*, **53**, 495–507.

O'Donnell S, Lane MD, Lawrence MJ (1993) The population genetics of self-incompatibility polymorphism in *Papaver rhoeas*. VI. Estimation of the overlap between the allelic complements of a pair of populations. *Heredity*, **71**, 591–595.

Paxman GJ (1963) The maximum likelihood estimation of the number of self-sterility alleles in a population. *Genetics*, **43**, 1029–1042.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Raspé O, Jacquemart A-L (1998) Allozyme diversity and genetic structure of European populations of *Sorbus aucuparia* L. (Rosaceae: Maloideae). *Heredity*, **81**, 537–545.

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

Raspé O, Saumitou-Laprade P, Cuguen J, Jacquemart A-L (2000) Chloroplast DNA haplotype variation and populations differentiation in *Sorbus aucuparia* L. (Rosaceae: Maloideae). *Molecular Ecology*, **9**, 1113–1122.

Rendell S, Ennos RA (2003) Chloroplast DNA diversity of the dioecious European tree *Ilex aquifolium* L. (English holly). *Molecular Ecology*, **12**, 2681–2688.

Richman AD, Kao T-H, Schaeffer SW, Uyenoyama MK (1995) *S*-allele sequence diversity in natural populations of *Solanum carolinense* (Horsenettle). *Heredity*, **75**, 405–415.

Richman AD, Kohn JR (2000) Evolutionary genetics of self-incompatibility in the Solanaceae. *Plant Molecular Biology*, **42**, 169–179.

Sassa H, Nishio T, Kowyama Y, Hirano H, Koba T, Ikehashi H (1996) Self-incompatibility (S) alleles of the *Rosaceae* encode members of a distinct class of the T2/S ribonuclease superfamily. *Molecular and General Genetics*, **250**, 547–557.

Schierup MH (1998) The number of self-incompatibility alleles in a finite, subdivided population. *Genetics*, **149**, 1153–1162.

Schierup MH, Vekemans X, Charlesworth D (2000) The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetical Research*, **76**, 51–62.

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

Swofford DL (2002) PAUP*: *Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Takahata N (1993) Evolutionary genetics of human paleopopulations. In: *Mechanisms of Molecular Evolution* (eds Takahata N, Clark AG), pp. 1–21. Sinauer, Sunderland, Massachusetts.

Takahata N (1990) A simple genealogical structure of strongly balanced allelic lines and trans-specific evolution of polymorphism.

Proceedings of the National Academy of Sciences of the USA, **87**, 2419–2425.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Uyenoyama MK, Newbigin E (2000) Evolutionary dynamics of dual-specificity self-incompatibility alleles. *Plant Cell*, **12**, 310–312.

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

Vekemans X, Slatkin M (1994) Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics*, **137**, 1157–1165.

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

Walker EA, Ride JP, Kurup S, Franklin-Tong VE, Lawrence MJ, Lawrence FCH (1996) Molecular analysis of two functional homologs of the S_3 allele of the Papaver rhoeas self-incompatibility gene isolated from different populations. *Plant Molecular Biology*, **30**, 983–994.

Wang X, Hughes AL, Tsukamoto T, Ando T, Kao T (2001) Evidence that intragenic recombination contributes to allelic diversity of the *S-RNase* gene at the self-incompatibility (*S*) locus in *Petunia inflata*. *Plant Physiology*, **125**, 1012–1022.

Weller SG, Donoghue MJ, Charlesworth D (1995) The evolution of self-incompatibility in angiosperms: a phylogenetic approach. In: *Experimental and Molecular Approaches to Plant Biosystematics* (eds Hoch PC, Stephenson AG), pp. 355–382. Missouri Botanical Garden, St. Louis.

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

Olivier Raspé uses molecular methods to address questions in plant population genetics, evolution, and systematics. This paper is part of an ongoing programme on the population genetics of *Sorbus aucuparia*. Joshua Kohn studies the evolution and consequences of plant mating system variation with a focus on self-incompatibility.
