Molecular Evolution within and between Self-Incompatibility Specificities

V. Castric,*[1] J. S. Bechsgaard,[2] S. Grenier,[1,3] R. Noureddine,[1,4] M. H. Schierup,[2] and X. Vekemans[1]

[1]Université des Sciences et Technologies de Lille 1, Laboratoire Génétique et Evolution des Populations Végétales, CNRS UMR, France

[2]Ecology and Genetics, Institute of Biological Sciences, University of Aarhus, Denmark

[3]Unité de Recherche Pluridisciplinaire Prairies et Plantes Fourragères (URP3F), INRA Lusignan, France

[4]INRA, ENVN, UMR 1300 Bio-agression, Epidémiologie et Analyse de Risques (BioEpAR), Ecole Nationale Vétérinaire de Nantes, France

***Corresponding author:** E-mail: Vincent.Castric@univ-lille1.fr.

Associate editor: Marcy Uyenoyama

Research article

Abstract

Genes under multiallelic balancing selection have sharply contrasted evolutionary dynamics across timescales, with much longer coalescence time among functionally distinct allelic lines but much shorter coalescence time among gene copies within allelic lines as compared with the genomic background. In this paper, we combine theoretical and empirical approaches to investigate patterns of molecular evolution within and between self-incompatibility (SI) specificities. We first use numerical simulations to investigate coalescence times within allelic lines in a subdivided population for a sporophytic SI system. We then report on a comprehensive analysis of nucleotide polymorphism among gene copies within five distinct allelic lines in the closely related *Arabidopsis halleri* and *Arabidopsis lyrata*. In line with our model predictions, we find that the observed level of polymorphism among gene copies was generally low but differed among allelic lines. The data provide compelling direct evidence for recombination and/or gene conversion not only within the two most recessive allelic lines but also between two closely related but distinct allelic lines, suggesting that recombination at the Arabidopsis SI locus is possible in the absence of large sequence divergence among haplotypes. We observed shared polymorphic sites between the two species in one allelic line and strikingly similar haplotypes in another allelic line. We discuss whether convergent evolution may have led to this pattern and suggest that these observations are consistent with ongoing or very recent introgression, as previously documented.

Key words: multiallelic balancing selection, intrahaplotype polymorphism, recombination, introgression, sporophytic self-incompatibility, gene conversion.

Introduction

Self-incompatibility (SI) is a widespread genetic system preventing selfing and inbreeding with close relatives in flowering plants (de Nettancourt 2001) and is a classical example of multiallelic balancing selection (Wright 1939). A distinctive feature of the molecular evolution of genes under multiallelic balancing selection is their sharply contrasted evolutionary dynamics across two different timescales: between and within allelic lines (fig. 1). Although phenotypically distinct allelic lines may be difficult to delineate in many genetic systems such as magor histocompatibility complex or plant disease resistance genes, different SI specificities can easily be determined by crossing, making SI genes particularly relevant models to learn about general properties of balancing selection. Theory predicts that functionally distinct allelic lines should be actively maintained within species or populations within species for extended periods of time, such that they diverge on a much longer timescale as compared with neutral genes. Takahata (1990) and Vekemans and Slatkin (1994) showed that the overall genealogical structure for these genes is expected to be similar to that for genes at a neutral locus but with the total length of the genealogy inflated by a factor inversely related to the mutation rate to new functional alleles. A large body of empirical work in a number of self-incompatible species has been devoted to characterizing evolutionary processes involving distinct specificities, focusing on the level of divergence among allelic lines (Ioerger et al. 1990), the molecular targets of balancing selection (Saainudin et al. 2005; Castric and Vekemans 2007), recombination among allelic lines (Charlesworth et al. 2006; Takuno et al. 2007), divergence between species (Dwyer et al. 1991; Castric et al. 2008), spatial distribution (Schierup et al. 2008), or on relative branch length (Uyenoyama 1997; Richman and Kohn 2000). In contrast, patterns of molecular evolution within functional allelic lines have received much less attention.

Yet patterns of polymorphism within SI specificities could indeed shed light on three major processes of molecular evolution. First, balancing selection maintains over long periods many allelic lines that can each be considered as a distinct subpopulation, hence proportionally reducing the depth of gene genealogies within each of the allelic lines. Thus, because of short expected coalescence times, molecular polymorphism within specificities at the SI locus (S-locus) should be orders of magnitude lower than among

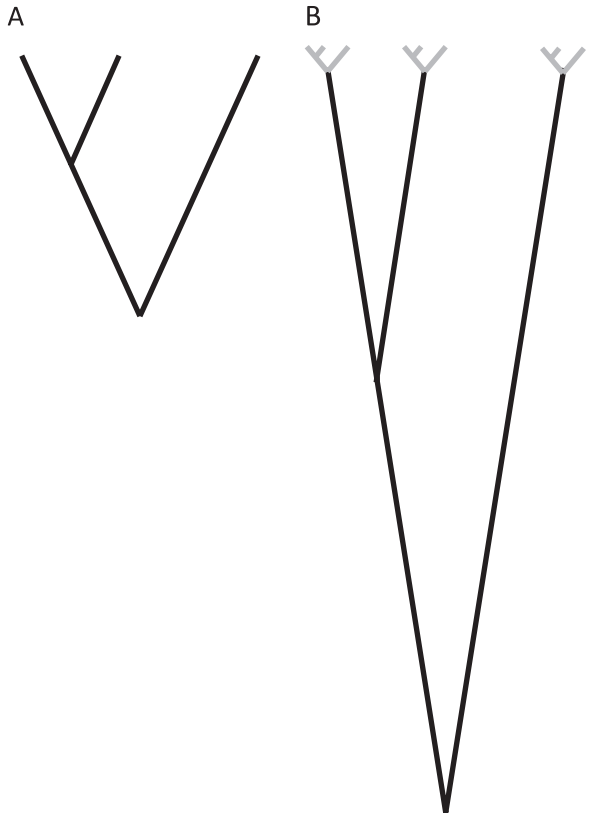

Fig. 1. Comparison of gene genealogies at the genomic background (*A*) and S-alleles (*B*). The genealogy of S-alleles can be separated into two nested genealogies: among allelic lines (black lines) and within allelic lines (gray lines).

specificities, the magnitude of this reduction being essentially determined by the number of functionally distinct allelic lines (May et al. 1999). Moreover, although in gametophytic SI the evolutionary dynamics of all alleles at the S-locus (S-alleles) should be identical in expectation (Wright 1939), S-alleles in sporophytic SI typically show dominance relationships that entail different selective pressures among them (Schierup et al. 1997; Billiard et al. 2007). Hence, intuition suggests that diversity may differ among alleles at different levels of the dominance hierarchy. However, the direction of the effect is currently not known, especially in the presence of population structure. On the one hand, recessive alleles are expected to reach higher frequencies than dominant alleles in a panmictic population, such that the number of gene copies and thus the timescale of the genealogy of gene copies sampled within recessive allelic lines may be expected to be larger than that within dominant allelic lines. On the other hand, population subdivision also has an important impact on coalescence times (Schierup et al. 2000) and may potentially offset this effect if dominant and recessive alleles differ with respect to their effective migration rate (i.e., migration rate taking into account the probability of successful introduction, which depends on the allele's dominance level). More specifically, dominant S-alleles could be present in a larger number of demes than recessive S-alleles if their effective migration rate is higher, a feature that would increase coalescence times within dominant alleles. The relative magnitude of these two effects remains to be determined, both theoretically and empirically.

A second distinctive feature of the molecular evolution of SI genes is the role played by recombination and/or gene conversion. Indeed, whether recombination occurs at all in the genomic region of the S-locus is still a debated topic. Although Wang et al. (2001) found recombinant haplotypes in *Petunia inflata*, and so direct evidence that intragenic recombination contributed to the generation of new specificities in this species, Vieira et al. (2003) found no evidence for a decline of linkage disequilibrium between linked polymorphic sites in the S-locus in Solanaceae, Rosaceae, and Scrophulariaceae. Similarly, Charlesworth, Barolome, et al. (2003), Charlesworth, Mable, et al. (2003), and Castric and Vekemans (2007) found no clear decline of synonymous diversity across exons in the S-locus pistil specificity gene *SRK* (S-receptor kinase) in *A. lyrata* and *Arabidopsis halleri*. In the surrounding genomic region, Kawabe et al. (2006) found evidence for reduced recombination across ~600 kb from each side of the S-locus compared with more distant flanking regions in *A. lyrata*. However, in the closely related *A. halleri*, Ruggiero et al. (2008) found no difference in estimated recombination rates between a 4-Mb region centered on the S-locus compared with the average genomic rate in noncentromeric regions. Thus, whether recombination actually occurs at SI loci is still unclear.

Rates of recombination are known to vary widely across regions within individual genomes, and three general mechanisms are believed to modulate the level of recombination across a genome. First, recombination may actually occur but be selected against. Given the dual molecular structure of the S-locus (typically, pollen + pistil component genes), any recombinant haplotype between the component genes would result in self-compatible combinations, which could be selected against in the presence of strong inbreeding depression. Empirical evidence suggests that the level of inbreeding depression in self-incompatible species is indeed high (Charlesworth and Charlesworth 1987; Stone 2004; Glémin et al. 2006; Bechsgaard 2007). Second, sequence-specific features as well as genomic localization (centromeric vs. telomeric) and GC content have been shown to affect recombination and may be responsible for recombination hot and coldspots (De Massy 2003). Third, the local rate of recombination decreases with increasing sequence divergence between homologous chromatides. Opperman et al. (2004) observed in *Arabidopsis thaliana* a rapid drop-off in somatic recombination rates with sequence divergence. In their experiment, the introduction of a single mismatch (0.16% nucleotide divergence) lowered the recombination rate by threefold, whereas additional mismatches reduced the recombination rate to a lesser extent (see also Datta et al. 1997; Chen and Jinks-Robertson 1999 in yeast). Because of the typically very high level of sequence divergence and physical rearrangement among allelic lines at the S-locus (Kusaba et al. 2001; Tang et al. 2007; Boggs, Dwyer, Shah, et al. 2009), this

process is very likely to affect recombination at the S-locus. Formally testing this prediction would require comparing recombination among allelic lines with low versus High sequence divergence, whereas only cases of the latter are typically found when comparing functionally distinct SI specificities. In contrast to gametophytic SI, sporophytic SI allows the spontaneous generation of homozygote genotypes for a given allelic line, thereby potentially providing opportunity for recombination in a context of low sequence divergence, unless another mechanism than overall sequence divergence acts to suppress recombination. Hence, the analysis of molecular polymorphism among gene copies within allelic lines in sporophytic SI provides a unique opportunity to test whether recombination does indeed occur at the gene controlling sporophytic SI in the absence of strong sequence divergence. Because homozygosity is expected to be higher for recessive than for dominant alleles, we predict that recombination should occur more readily at recessive than at dominant alleles.

A third feature of the molecular evolution of SI genes is the preponderant role of introgression. Balancing selection is expected to favor introgression at the genes controlling SI because any novel S-allele introduced by introgression would be locally rare in the recipient species and thus enter the species with a higher probability than alleles at neutral genes (Schierup et al. 2000). *Arabidopsis halleri* and *A. lyrata* are two Brassicaceae species with functional sporophytic SI (Schierup et al. 2001; Llaurens et al. 2008) that diverged approximately 2 Ma (Koch and Matschinger 2007; Castric et al. 2008). Polymorphism of the genomic background has been characterized in a number of genes across the species range and was found to be partially shared between the two species (Ramos-Onsins et al. 2004), leaving open the possibility that introgression is still ongoing. We have recently documented that the gene controlling SI specificity (*SRK*) was undergoing introgression at a ~5-fold higher rate than the genomic background (Castric et al. 2008). Whether introgression also had such a profound influence on molecular polymorphism within allelic lines remains open to investigation. In particular, because coalescence of gene copies within allelic lines is expected to occur over a much shorter timescale than that of neutral genes, incomplete lineage sorting would be an excellent indication that introgression between closely related species is either ongoing or at least very recent.

In spite of its potential interest to document processes of molecular evolution at the S-locus, molecular polymorphism within allelic lines has been investigated in a few species only. Although all studies thus far confirmed the expected absolute low levels of molecular variation (Walker et al. 1996 in *Papaver rhoeas*, Wang et al. 2001 in *P. inflata*, Raspé and Kohn 2007 in *Sorbus aucuparia*, Nunes et al. 2006 in *Prunus spinosus*, Kusaba et al. 2000; Miege et al. 2001 in *Brassica oleraceae*, and Charlesworth, Barolome, et al. 2003; Charlesworth, Mable, et al. 2003 in *A. lyrata*), none of them included "reference" neutral polymorphism data. Hence, it remains difficult to evaluate whether the observed level of polymorphism is indeed lower than that of the genomic background. Moreover, only pairs of sequences were typically compared (but see Miege et al. 2001), and in any case, the number of sequences and/or the number of allelic lines compared was always very low, preventing any strong conclusion at this point.

In this paper, we first use numerical simulations to investigate coalescence times within allelic lines in a subdivided population for a sporophytic SI system, focusing on whether polymorphism is expected to vary according to dominance. We then perform a comprehensive analysis of polymorphism among gene copies within allelic lines in *A. halleri* and *A. lyrata* at each of the four levels of the dominance hierarchy described for *A. lyrata* in Prigoda et al. (2005). More specifically, we aim at 1) comparing polymorphism within allelic lines with that of the genomic background, 2) testing whether recombination and/or gene conversion occurs at S genes in the absence of large sequence divergence, and 3) testing for shared polymorphism within allelic lines between the two closely related *A. halleri* and *A. lyrata*.

Materials and Methods

Simulation Study

We used the forward coalescent simulation program of sporophytic SI from Schierup et al. (2000) to compare the genealogy of gene copies sampled within dominant versus recessive allelic lines. Briefly, we simulated the DOM model from Schierup et al. (1997), assuming an identical linear hierarchy of dominance levels in pollen and pistil, in $N = 1{,}000$ diploid individuals either in a panmictic population or in a subdivided population with 10 demes ($N_{deme} = 100$) connected by gene flow according to an *n*-island model of migration ($Nm = 0.6$, corresponding to the level of population structure in *A. halleri*, i.e., $F_{ST} = 0.3$, Pauwels M, personal communication). Each simulation was started with $2N$ functionally different alleles in the population and allowed to evolve for 316,000 generations (= 5 times the expected coalescence time of allelic lines in a gametophytic SI system as computed according to Vekemans and Slatkin 1994) to reach mutation–selection–drift equilibrium. We then computed for each allelic line in the dominance hierarchy the coalescence time of all gene copies sampled within the allelic line from the whole population, corresponding to the depth of the coalescent tree within the allelic line. The mutation rate to new S-alleles was $u = 5 \times 10^{-5}$ per gene per generation. Five hundred independent replicates were performed for both panmictic and subdivided populations.

Empirical Study

In the Brassicaceae, SI specificity is encoded by two physically linked genes, *SRK* and *SCR*, expressed in papilla cells at the surface of the stigma and in anthers during pollen maturation, respectively. *SRK* contains three hypervariable regions HV1, HV2, and HV3 (Nishio and Kusaba 2000). These regions are enriched in positively selected sites (Takebayashi et al. 2003; Castric and Vekemans 2007); they show low divergence between trans-specific pairs from *A.*

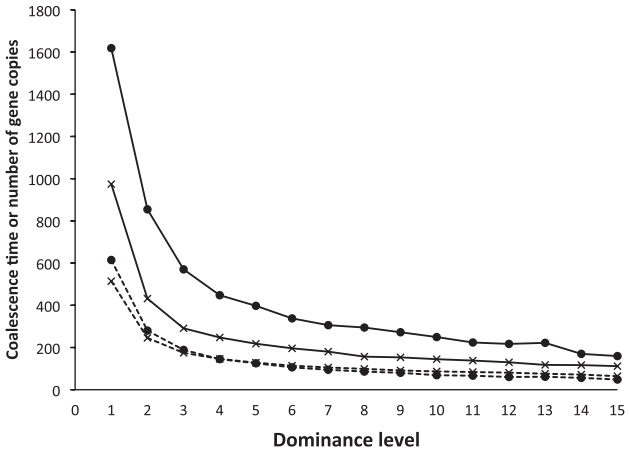

Fig. 2. Expected number of gene copies within allelic lines (dotted lines) and total coalescence times of gene copies sampled within allelic lines (solid lines) as a function of dominance level in a panmictic population (X) or a subdivided population of 10 demes with $Nm = 0.6$ (•). Allelic lines are ranked by increasing dominance level. Simulations were performed with $N = 1{,}000$ and mutation rate to new S-alleles as 5×10^{-5} per gene per generation.

halleri and *A. lyrata* (Castric et al. 2008); and functional analyses confirmed that they are involved in specificity determination (Boggs, Dwyer, Nasrallah, et al. 2009). In *A. lyrata*, S-alleles belong to at least four different levels along a dominance hierarchy, coined A1, B, A3, and A2 in ascending order by Prigoda et al. (2005). Llaurens et al. (in press) confirmed that this hierarchy was roughly preserved in *A. halleri*. A set of 39 European *A. halleri* populations was screened for S-alleles using allele-specific targeted primers as described in Llaurens et al. (in press). Multiple copies from this sampling collection were then obtained by direct sequencing for five different alleles, including at least one allele from each level of the dominance hierarchy described in Prigoda et al. (2005). Sequencing primers for each allele were designed so as to amplify the longest possible fragment based on the available sequence published. AhSRK01 or AlSRK01 is the most recessive allele known in *A. halleri* or *A. lyrata*, respectively (class A1 according to Prigoda et al. 2005) and is also the most frequent allele in each species. AhSRK01 and AlSRK01 sequences were obtained for a 1,112- and a 1,081-bp polymerase chain reaction (PCR) fragment, respectively, covering the nearly complete S-domain in 29 *A. halleri* and 19 *A. lyrata* individuals. The currently known nucleotide sequence of AhSRK01 includes the whole S-domain (exon 1) but neither introns nor exons from the kinase domain. AhSRK03/AlSRK03 and AhSRK28 belong to the second most recessive class of alleles in *A. halleri* and *A. lyrata* (class B), and AhSRK03 is the second most frequent allele in *A. halleri*. AhSRK03 and AhSRK28 sequences were obtained by direct sequencing of three overlapping PCR fragments (730, 790, and 1,240 bp) covering the S-domain up to the third exon of the kinase domain in 14 *A. halleri* and 7 *A. lyrata* individuals. In *A. lyrata*, only an incomplete portion of the S-domain is known for SRK28 (Charlesworth, Barolome, et al. 2003; Charlesworth, Mable, et al. 2003), so we did not include it in the analysis. AhSRK04 and AlSRK37 belong to an intermediate class in the dominance hierarchy (class A3) and are believed to represent the same functional specificity in *A. halleri* and *A. lyrata*, respectively (Bechsgaard et al. 2006). Seven AhSRK04 sequences were obtained by direct sequencing of two overlapping fragments (1,839 and 1,777 bp) spanning from the 5′UTR to the last exon of the kinase domain. Three AlSRK37 sequences were obtained using five primer pairs to tile a fragment from the S-domain to the last exon of the kinase domain. AhSRK05 belongs to class A2 and ranks among the most dominant alleles in *A. halleri*. It occurs at low frequency in natural populations. Based on the available nucleotide sequence of this allele (only the S-domain is known for this allele, Castric and Vekemans 2007), 10 AhSRK05 sequences were obtained by direct sequencing of a 483-bp portion of the 3′ end of the S-domain. All primer sequences are reported in supplementary table S1, Supplementary Material online. Sequences were obtained on an ABI capillary sequencer (Applied Biosystems) using the BigDye Terminator kit 3.1 chemistry (Applied Biosystems) and run on an ABI-3100 capillary sequencer. Sequences were assembled and aligned in BioEdit 7.0.9 (Hall 1999), and polymorphism was analyzed in DNAsp 4.20.2 (Rozas et al. 2003). Hypervariable regions were assigned with comparison to the BoSRK60 sequence (Genbank accession AB032474, Suzuki et al. 2000) according to Nishio and Kusaba (2000), and intron–exon boundaries were assigned according to Bechsgaard et al. (2006) and Castric and Vekemans (2007). Evidence for recombination or gene conversion among gene copies within each allelic line was sought using visual inspection and by applying the four-gamete test of Hudson and Kaplan (1985) implemented in DNAsp.

Results

Patterns of Polymorphism within Allelic Lines: Simulations

Dominant versus Recessive Allelic Lines. Simulations confirmed our prediction that dominance should strongly influence the timescale of gene genealogies within allelic lines, regardless of the presence of population structure. This effect is mainly caused by a higher mean allelic frequency for recessive than for dominant allelic lines (fig. 2). Indeed, gene copies sampled within recessive allelic lines showed consistently higher coalescence times (e.g., 974 generations on average for the most recessive allelic line) than gene copies sampled within dominant allelic lines (e.g., 112 generations for the most dominant allelic line). Overall, coalescence times within allelic lines were found to be very close to the neutral coalescent expectations (slightly under twice the number of gene copies per allelic line according to Vekemans and Slatkin 1994, fig. 2). Moreover, the contrast in coalescence times between dominant and recessive alleles was maintained under population subdivision ($Nm = 0.6$), although coalescence times within allelic lines were consistently higher. The effect of dominance on coalescence times was robust to

Table 1. Nucleotide Polymorphism within the S-Domain of Four SI Specificities.

		N	*L*	*h*	*S*	S_N[a]	π_S
SRK alleles	**AhSRK01 (very recessive)**	**29**	**1,112**	**13**	**9**	**5**	**0.00374**
	AhSRK03 (recessive)	**14**	**845**	**3**	**2**	**0**	**0.00571 (0.00138)[a]**
	AhSRK04 (dominant)	**7**	**1,311**	**4**	**7**	**2**	**0.00581**
	AhSRK05 (very dominant)	**10**	**483**	**2**	**1**	**1**	**0**
	Average						**0.00441**
Reference genes[b]	**HAT4**	**34**	**884**				**0.00710**
	CAUL	**36**	**824**				**0.02420**
	ScADH	**34**	**1,196**				**0.02880**
	Aly9	**30**	**497**				**0.00820**
	CHS	**28**	**860**				**0.03990**
	Average						**0.02164**

NOTE.—*N* refers to the number of sequences analyzed. *L* is the length of the sequence in base pairs. *h* is the number of different haplotypes. *S* is the number of segregating sites. S_N is the number of nonsynonymous segregating sites. π_S is the synonymous diversity estimate.

[a] π_S given in parentheses is obtained excluding haplotypes that are most likely derived from recombination between AhSRK03 and AhSRK28 (i.e., including AhSRK03a, b, and c only).

[b] From Ruggiero et al. (2008).

variations in population size (checked for $N = 200$, number of S-alleles maintained $n_a \approx 5$ and $N = 5{,}000$, $n_a \approx 38$, data not shown).

Polymorphism within and across Demes. Under population subdivision ($Nm = 0.6$), we compared coalescence times for genealogies of genes sampled within allelic lines either at the total population level or at the deme level (supplementary fig. S1, Supplementary Material online). The results showed that coalescence times are indeed reduced but still remain substantial at the deme level for most dominance levels ($T_{within(deme)}/T_{within(total)} = 0.36$–0.63), in particular for the most recessive allelic line, which showed a slight reduction only ($T_{within(deme)}/T_{within(total)} = 0.83$).

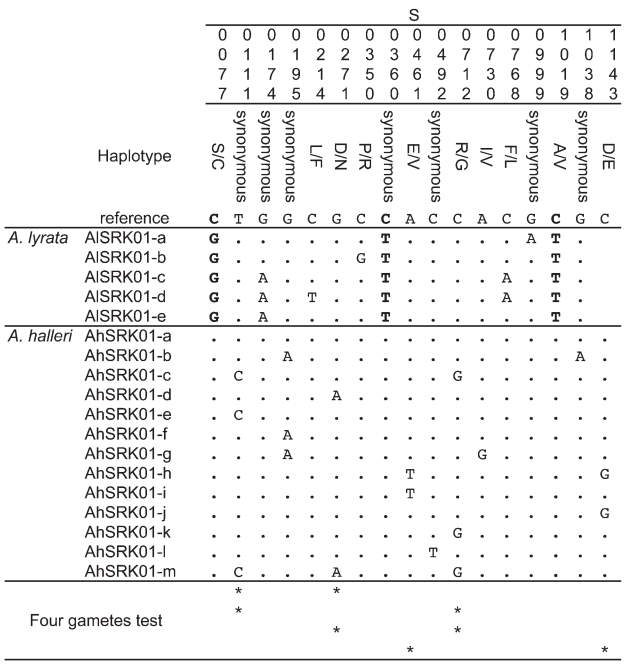

FIG. 3. Polymorphism within and between AhSRK01 and AlSRK01. Fixed differences between *Arabidopsis halleri* and *Arabidopsis lyrata* are figured in bold. Pairs of sites in *A. halleri* that show evidence for recombination using the four-gamete test are flagged with stars. Amino acids corresponding to nonsynonymous SNPs in exons are reported. S: extracellular S-domain. Nucleotide positions are given with reference to the start codon of BoSRK60.

Patterns of Polymorphism within Allelic Lines: Empirical Study

The empirical analysis revealed polymorphism within each of the allelic lines (table 1, figs. 3–5), with as many as 13 different haplotypes for a given allelic line (AhSRK01).

Synonymous specieswide diversity at the S-domain was consistently low within AhSRK allelic lines (average $\pi_S = 0.00441$) and varied from zero for AhSRK05 to 0.00581 for AhSRK04 (table 1). Taken at face value (but see Discussion), synonymous diversity was clearly not correlated with the level of dominance, because the two intermediate-dominance alleles (AhSRK03 and AhSRK04) had the highest π_S, and the two extreme-dominance alleles (AhSRK01, most recessive and AhSRK05, most dominant) had the lowest π_S.

Contrasting the average observed synonymous nucleotide polymorphism among gene copies within allelic lines ($\pi_{within} = 0.00441$) with that estimated for the genomic background ($\pi_{genomic} = 0.0216$, table 1, Ruggiero et al. 2008) produces a ratio $\pi_{within}/\pi_{genomic} = 0.204$, hence confirming the expected reduction in polymorphism.

Although most haplotypes tended to be restricted to one or a few geographically close populations only, some haplotypes had wide geographic distributions (supplementary table S2, fig. S2, Supplementary Material online). For instance, haplotype AhSRK01a was found in populations A8 (Austria), SLO7 (Slovenia), B1 (Belgium), and D13 (Germany), and haplotype AhSRK03d was found in populations PL1 (Poland), I1, I2, and I9 (Italy). Moreover, of 12 populations where diversity within population could be estimated in at least one allelic line (more than a single individual sampled), two or more haplotypes different in sequence were found in seven populations. This polymorphism included replacement differences, whose functional significance remains to be determined. The most spectacular example of this pattern was allele AhSRK03 in population PL01, where four different haplotypes (b–e) were identified

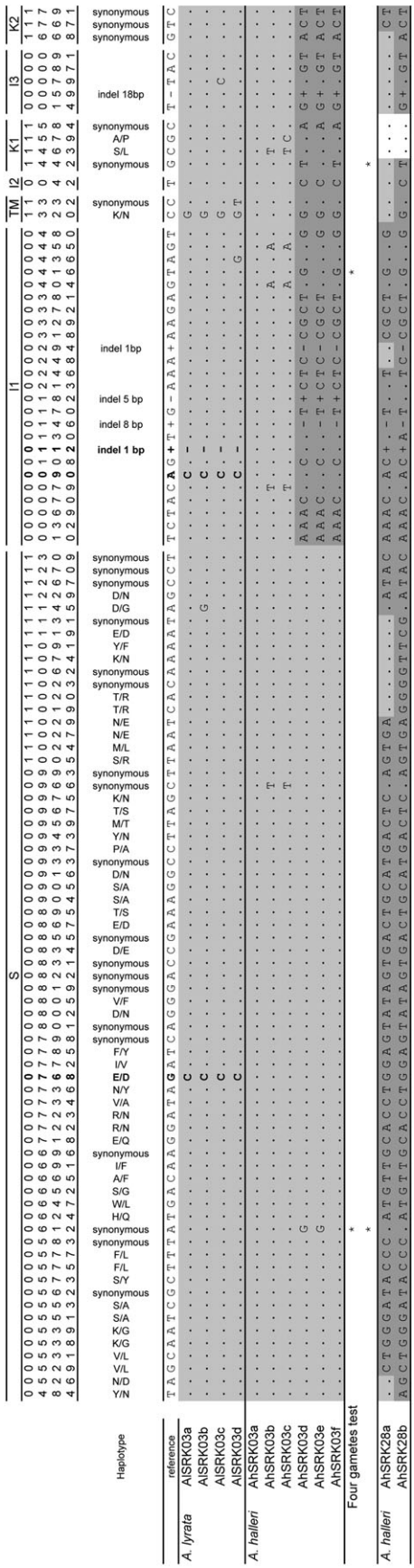

Fig. 4. Polymorphism within AhSRK03, AlSRK03, and AhSRK28. Fixed differences between *Arabidopsis halleri* and *Arabidopsis lyrata* are figured in bold. A possible scenario of recombination and gene conversion is figured in shades of gray (dark gray: typical SRK03 nucleotides positions, light gray: typical SRK28 positions). Portions where no scenario could be reliably inferred in AhSRK28a were left uncolored. Pairs of sites in *A. halleri* that show evidence for recombination using the four-gamete test are flagged with stars. Amino acids corresponding to nonsynonymous SNPs in exons are reported. S: extracellular S-domain, TM: transmembrane domain, I1–3: introns 1–3, and K1–2: exons of the kinase domain. Nucleotide positions are given with reference to the start codon of BoSRK60. Because introns could not be aligned reliably due to large sequence divergence among alleles, nucleotide positions for introns are given with reference to the start of each individual intron.

among the six sequences obtained. The fact that the two allelic lines within which intrahaplotype polymorphism was found even within populations (AhSRK01 and AhSRK03) are the two most recessive ones is consistent with the results from our simulation study comparing intrademe versus total population coalescence times that showed that the most recessive allele is expected to have a higher within-deme diversity than any other allele (supplementary fig. S1, Supplementary Material online).

Evidence for Recombination and/or Gene Conversion

Recombinant haplotypes were detected for AhSRK01 and AhSRK03, the two most recessive allelic lines. For AhSRK01, four pairs of sites (111–271, 111–712, 271–712, and 461–1143) showed all four gametic combinations (fig. 3), providing unambiguous evidence that at least two recombination events must have occurred, one between sites 111 and 271 and one between sites 271 and 712. The four-gamete test revealed all four gametic combinations for AhSRK03 at sites 612 in the S-domain and site 404 in the first intron, and at sites 612 in the S-domain and 1442 in the first exon of the kinase domain, thus implying that at least one recombination must have occurred between sites 612 and 404 (fig. 4). Strikingly, *A. halleri* showed two distinct groups of AhSRK03 haplotypes. Haplotypes a–c closely resembled the *A. lyrata* AlSRK03 haplotypes, whereas in the second half of the sequence (from the first intron to the kinase domain), haplotypes d–f resembled more *A. halleri* AhSRK28b haplotypes. This observation strongly suggests that some *A. halleri* haplotypes (d–f) result from interallelic recombination, combining the SRK03 S-domain with the SRK28 kinase domain. Interestingly, the second SRK28 haplotype (AhSRK28a) showed signs of gene conversion from AhSRK03 on at least four scattered fragments (fig. 4), suggesting that recombination/gene conversion must be very intense. The individual analyzed does not carry AhSRK03, thus excluding the possibility of artificial PCR chimera. No evidence for recombination was found for AhSRK04, and the test could not be performed for AhSRK05 because of insufficient polymorphism (a single polymorphic site). Note however that recombination with other alleles can still be excluded for this allelic line, because that would have introduced some variants.

Rare Shared Polymorphisms and Evidence for Introgression

Shared polymorphisms between *A. lyrata* and *A. halleri* were found for only one of the three allelic lines tested. For AhSRK01/AlSRK01, of the 16 polymorphic sites overall identified across the S-domain in the two species combined (fig. 3), three sites showed fixed differences, and none showed shared polymorphisms. No shared polymorphism was found between SRK03 sequences of *A. halleri* and *A. lyrata*. Yet, SRK03 haplotypes of *A. halleri* that showed no sign of recombination and/or gene conversion with SRK28 (i.e., haplotypes AhSRK03a–c) were indeed extremely similar to *A. lyrata* haplotypes with only three fixed nucleotide

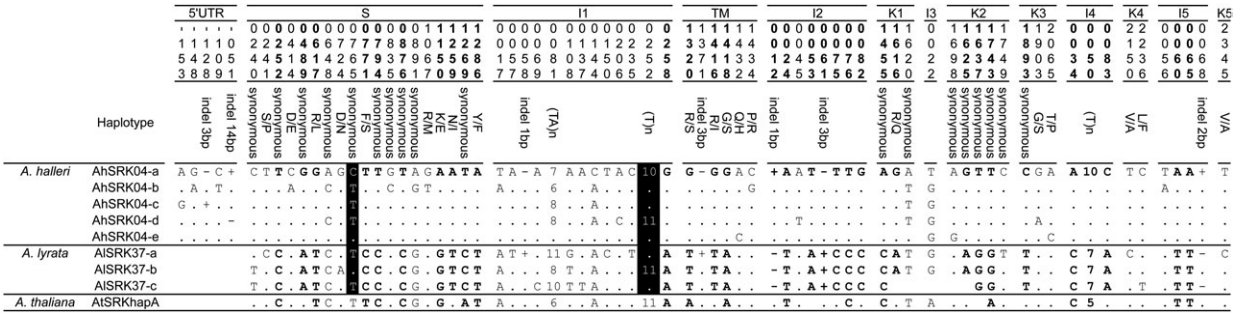

Fig. 5. Polymorphism within AhSRK04 and AlSRK37 and ancestral states inferred from *Arabidopsis thaliana* haplogroup A. Fixed differences between *Arabidopsis halleri* and *Arabidopsis lyrata* are figured in bold, and shared polymorphisms between *A. halleri* and *A. lyrata* are figured in black. Amino acids corresponding to nonsynonymous SNPs in exons are reported. The sequence spans from the 5′UTR to the fifth exon of the kinase domain for *A. halleri* and from the S-domain to the fifth exon of the kinase domain for *A. lyrata*. S: extracellular S-domain, I1–5: introns 1–5, TM: transmembrane domain, and K1–K5: exons 1–5 of the kinase domain. Nucleotide positions in the exons are given with reference to the start codon of BoSRK60. Because introns could not be aligned reliably due to large sequence divergence among alleles, nucleotide positions for introns are given with reference to the start of each individual intron.

differences along the whole sequence (~1,780 bp, fig. 4), possibly suggesting recent introgression from *A. lyrata*. For AhSRK04/AlSRK37, two sites showed shared polymorphism between the two species: site 765 in the S-domain and site 252 in the first intron, also possibly suggesting ongoing introgression. However, the former is a 4-fold degenerate site and the latter is a polyT length polymorphism that may both be prone to homoplasy. Thus, convergent evolution of these two single nucleotide polymorphisms (SNPs) seems possible, particularly as there are 32 fixed differences between the *A. halleri* and *A. lyrata* sequences (fig. 5).

Sequences from this study have been deposited in the GenBank database under accession numbers GQ915332–GQ915371.

Discussion

Overall, our combined theoretical and empirical investigations confirm that the observed level of polymorphism among gene copies is generally low (in line with previous results of Miege et al. 2001; Charlesworth, Barolome, et al. 2003; Charlesworth, Mable, et al. 2003) but differs among allelic lines and may have been influenced by dominance and population subdivision. Our results provide unambiguous evidence for recombination and/or gene conversion not only within but also between closely related allelic lines, especially for the most recessive allelic lines tested. We observed shared polymorphic sites between *A. halleri* and *A. lyrata* in one allelic line and strikingly similar haplotypes in another line. We discuss whether convergent evolution may have led to this pattern and suggest that these observations are consistent with ongoing or very recent introgression, as previously documented.

Interspecific Introgression and Interallelic Recombination Inflate Polymorphism within Allelic Lines

Our analysis confirmed that polymorphism was indeed reduced within allelic lines as compared with the genomic background in *A. halleri*. The model also predicts that recessive S-alleles should have higher polymorphism than dominant allelic lines, but our empirical study revealed no such clear trend. At least two factors that were not taken into account in our simulation study may explain this apparent discrepancy. First, as shown by Castric et al. (2008), introgression is expected to be favored at this gene as compared with the genomic background. Indeed, haplotypes AhSRK03a–c are strikingly similar to AlSRK03 haplotypes, with only 3-nt differences across ~1,780 bp. This level of divergence between *A. halleri* and *A. lyrata* is about an order of magnitude lower than that of other trans-specifically shared allelic lines (27 and 46 differences across sequences of similar extent for AhSRK04/AlSRK37 and AhSRK10/AlSRK16, respectively, Bechsgaard et al. 2006). These pairs are themselves less divergent than the genomic background ($K_{4\text{-fold}} = 0.0287$ and 0, respectively, vs. $K_{4\text{-fold}} = 0.0743$ for the genomic background, Castric et al. 2008), suggesting that AhSRK03 may have been transferred from *A. lyrata* into *A. halleri* very recently. Hence, our data suggest that polymorphism within AhSRK03 has been inflated by recent introgression, leading to the two divergent clades AhSRK03a–c and AhSRK03d–e. The elevated level of polymorphism for the AhSRK04 allelic line also suggests that some important process of molecular evolution (possibly introgression) was not taken into account in our simple model. Second, polymorphism within AhSRK03 was indeed lower than that of AhSRK01 after excluding from the analysis recombinant haplotypes AhSRK03d–f whose 3′ half, including part of the first intron, the transmembrane domain, and part of the kinase domain, is most likely derived from AhSRK28 (π_S=0.00138). This result would be more in line with theoretical predictions, possibly suggesting that interallelic recombination has inflated the observed level of polymorphism within AhSRK03. Before the current study, Miege et al. (2001) reported a previous comprehensive analysis of polymorphism within an allelic line for the Brassica BoSRK2 allele. Interestingly, the 10 BoSRK2 sequences obtained fell into two phylogenetically distinct groups, similarly to our observation for AhSRK03, possibly suggesting that either introgression and/or gene conversion from a related

allelic line may also play an important role in intrahaplotype polymorphism in Brassica species.

The weak effect of population structure on intrahaplotype polymorphism is exemplified by the observation that the two most recessive allelic lines (SRK01 and SRK03) show polymorphism even within demes. This observation is consistent with our simulation results that show that for recessive allelic lines, coalescence times within demes are generally very close to coalescence times at the total population level, thus providing ample time for mutations to accumulate even within demes.

Recombination at the Arabidopsis SI Locus Is Possible in the Absence of Large Sequence Divergence

Although we cannot tell from our data whether natural selection or sequence divergence prevents recombination among allelic types, our data provide compelling evidence for pervasive recombination within and even between closely related allelic types and are therefore clearly not consistent with complete recombination suppression, suggesting that S-allele sequences per se do not constrain recombination. This observation is consistent with the poor congruence between phylogenies based on S- and kinase domains in *A. lyrata* (Charlesworth, Barolome, et al. 2003; Charlesworth, Mable, et al. 2003). Although gene conversion between the *SRK* and *SLG* genes has long been suspected based on phylogenetic reconstructions (Sato et al. 2002; Charlesworth, Barolome, et al. 2003; Charlesworth, Mable, et al. 2003), this is to our knowledge the first direct evidence for gene conversion between *SRK* alleles in the Brassicaceae. Interestingly, we found evidence for recombination and/or gene conversion for the two most recessive alleles only. Although not a direct demonstration that recombination does not occur at higher levels of the dominance hierarchy, recombination among gene copies within allelic lines requires the formation of homozygotes for a given allelic line but heterozygotes for different gene copies of this allelic line. As predicted by theoretical models, homozygosity is expected to be much higher for recessive alleles than for dominant alleles (Billiard et al. 2007). Moreover, the second most recessive allele (AhSRK03) may locally reach high frequency in populations where the most recessive allele AhSRK01 is absent (ca. 5% of all populations, Castric V, unpublished data), thus specifically increasing opportunities for recombination for this allele. For AhSRK03, the recombination event seems to have swapped kinase domains (along with intron 1 and transmembrane domain) with its most closely related allelic line, AhSRK28. It should be noted that among all allelic lines that were compared in *A. halleri* for the S-domain nucleotide sequence, AhSRK28 and AhSRK03 were found to be the two most similar (Castric et al. 2008).

Polymorphism within S-Alleles: First Step Toward Diversification?

Our analysis revealed molecular polymorphism within each of the five allelic lines tested, some of the differences involved being nonsynonymous differences in the S-domain. Chookajorn et al. (2004) showed that sequence variations of the pollen protein SCR might alter binding affinity to the pistil protein SRK. They proposed that these slight functional variations among gene copies within a given allelic line occurring jointly in pollen and pistil genes may be the first step toward gradual functional diversification of SI systems. This model requires that standing variation for mutations altering specificity does segregate within species, and here we demonstrate in natural populations of two species with functional SI that nonsynonymous polymorphism indeed segregates at least at the pistil gene (see also Charlesworth, Barolome, et al. 2003; Charlesworth, Mable, et al. 2003).

However, whether the low level of nonsynonymous polymorphism observed here within and across *A. halleri* populations has functional consequences remains to be determined. Although functionally distinct S-alleles found within species are typically highly divergent, a number of exceptions have been reported in several species. The phenotypically distinct S_1/S_4 and S_3/S_5 pairs of *Pyrus pyrifolia* show 90% and 95.5% amino acid identity, respectively (Ishimizu et al. 1998), and the S_{11}/S_{13} pair of *Solanum chacoense* (Saba El Leil et al. 1994) shows only 10 amino acid substitutions, four of which are sufficient to switch SI specificity between the two alleles (Matton et al. 1997). In *P. inflata*, the functionally distinct S_6 and S_9 alleles differ only by 4 nt, all in the second hypervariable region, resulting in two amino acid differences (Wang et al. 2001). Using a different approach based on the comparison of S-alleles in two closely related species, Sato et al. (2003, 2006) found that seven of nine pairs of alleles had retained identical specificity between the closely related *B. oleraceae* and *Brassica rapa* despite an average of 4.7% amino acid divergence within the S-domain. The two exceptions were BoSRK2b/BrSRK44 with 14 amino acid differences at *SRK* and completely different recognition specificities and BoSRK5/BrSRK40 with 20 amino acid differences at *SRK* and slightly different recognition specificities. Finally, Miege et al. (2001) reported the only phenotypic comparison of gene copies within allelic lines within species and showed that even the most divergent pairs of sequences for BoSRK02 (belonging to clades 2a and 2b and showing 12 amino acid differences, including one in HV3) have retained identical specificity. Overall, these results suggest that a low number of amino acid differences may in some cases be sufficient to switch specificity, although the location of these differences along the gene also largely determines their phenotypic consequences. Therefore, precisely assessing the phenotypic consequences of the naturally occurring variants that we identified will now require controlled crosses. In particular, whether the chimeric AhSRK28a haplotype has retained identical specificity as AhSRK28b in spite of the four-gene conversion events we inferred from AhSRK03a, b, or c will be central.

Supplementary Material

Supplementary tables S1 and S2 and supplementary figures S1 and S2 are available at *Molecular Biology*

and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Pierre Saumitou-Laprade for access to *A. halleri* samples and three anonymous reviewers for many relevant comments. This work was supported by an ATIP and ATIP-plus research grant from the life science department of the CNRS, and research grants ANR-06-BLAN-0128 and ANR-06-BDIV-003 from the French National Research Agency to X.V. Grants from the Danish National Research Council to M.H.S. are gratefully acknowledged. Much of it was done while V.C. was on sabbatical leave at CNRS.

References

Bechsgaard JS. 2007. Cause and consequence of self-incompatibility, ecology and genetics [PhD thesis]. Denmark, University of Aarhus.

Bechsgaard JS, Castric V, Charlesworth D, Vekemans X, Schierup MH. 2006. The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 million years. *Mol Biol Evol*. 23:1741–1750.

Billiard S, Castric V, Vekemans X. 2007. A general model to explore complex dominance patterns in plant sporophytic self-incompatibility systems. *Genetics*. 175:1351–1369.

Boggs NA, Dwyer KG, Nasrallah ME, Nasrallah JB. 2009. In vivo detection of residues required for ligand-selective activation of the S-locus receptor in Arabidopsis. *Curr Biol*. 19:1–6.

Boggs NA, Dwyer KG, Shah P, McCulloch AA, Bechsgaard J, Schierup MH, Nasrallah ME, Nasrallah JB. 2009. Expression of distinct self-incompatibility specificities in *Arabidopsis thaliana*. *Genetics*. [Epub ahead of print] doi:10.1534/genetics.109.102442.

Castric V, Bechsgaard J, Schierup MH, Vekemans X. 2008. Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet*. 4(8):e1000168.

Castric V, Vekemans X. 2007. Evolution under strong balancing selection: how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae? *BMC Evol Biol*. 7:132.

Charlesworth D, Bartolome C, Schierup MH, Mable BK. 2003. Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. *Mol Biol Evol*. 20:1741–1753.

Charlesworth D, Charlesworth B. 1987. Inbreeding depression and its evolutionary consequences. *Annu Rev Ecol System*. 18:237–268.

Charlesworth D, Kamau E, Hagenblad J, Tang C. 2006. Trans-specificity at loci near the self-incompatibility loci in Arabidopsis. *Genetics*. 172(4):2699–2704.

Charlesworth D, Mable BK, Schierup MH, Bartolome C, Awadalla P. 2003. Diversity and linkage of genes in the self-incompatibility gene family in *Arabidopsis lyrata*. *Genetics*. 164:1519–1535.

Chen W, Jinks-Robertson S. 1999. The role of the mismatch repair machinery in regulating mitotic and meiotic recombination between diverged sequences in yeast. *Genetics*. 151:1299–1313.

Chookajorn T, Kachroo A, Ripoll DR, Clark AG, Nasrallah JB. 2004. Specificity determinants and diversification of the Brassica self-incompatibility pollen ligand. *Proc Natl Acad Sci U S A*. 101:911–917.

Datta A, Hendrix M, Lipsitch M, Jinks-Robertson S. 1997. Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. *Proc Natl Acad Sci U S A*. 94:9757–9762.

De Massy B. 2003. Distribution of meiotic recombination sites. *Trends Genet*. 19:514–522.

de Nettancourt D. 2001. Incompatibility and incongruity in wild and cultivated plants. Springer-Verlag, Berlin.

Dwyer KG, Balent MA, Nasrallah JB, Nasrallah ME. 1991. DNA sequences of self-incompatibility genes from *Brassica campestris* and *B. oleracea*: polymorphism predating speciation. *Plant Mol Biol*. 16:481–486.

Glémin S, Vimond L, Ronfort J, Bataillon T, Mignot T. 2006. Marker-based investigation of inbreeding depression in the endangered species Brassica insularis. *Heredity*. 97:304–311.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser*. 41:95–98.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*. 111:147–164.

Ioerger TR, Clark AG, Kao TH. 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proc Natl Acad Sci U S A*. 87(24):9732–9735.

Ishimizu T, Shinkawa T, Sakiyama F, Norioka S. 1998. Primary structural features of rosaceous S-RNases associated with gametophytic self-incompatibility. *Plant Mol Biol*. 37:931–941.

Kawabe A, Hansson B, Forrest A, Hagenblad J, Charlesworth D. 2006. Comparative gene mapping in *Arabidopsis lyrata* chromosomes 6 and 7 and *A. thaliana* chromosome IV: evolutionary history, rearrangements and local recombination rates. *Genet Res*. 88:45–56.

Koch MA, Matschinger M. 2007. Evolution and genetic differentiation among relatives of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*. 104:6272–6277.

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME. 2001. Self-incompatibility in the genus *Arabidopsis*: characterization of the S locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. *Plant Cell*. 13:627–643.

Kusaba M, Matsushita M, Okazaki K, Satta Y, Nishio T. 2000. Sequence and structural diversity of the S-Locus genes from different lines with the same self-recognition specificities in *Brassica oleracea*. *Genetics*. 154:413–420.

Llaurens V, Billiard S, Leducq JB, Castric V, Klein EK, Vekemans X. Does frequency-dependent selection with complex dominance interactions accurately predict allelic frequencies at the self-incompatibility locus in *Arabidopsis halleri*? *Evolution*. 62: 2545–2557.

Llaurens V, Castric V, Austerlitz F, Vekemans X. 2008. High paternal diversity in the herbaceous perennial herb *Arabidopsis halleri* despite functional self-incompatibility, clonal reproduction and spatially-restricted pollen dispersal. *Mol Ecol*. 17:1577–1588.

Matton DP, Maes O, Laublin G, Xike Q, Bertrand C, et al. 1997. Hypervariable domains of self-incompatibility RNases mediate allele-specific pollen recognition. *Plant Cell*. 9:1757–1766.

May G, Shaw F, Badrane H, Vekemans X. 1999. The signature of balancing selection: fungal mating compatibility gene evolution. *Proc Natl Acad Sci USA*. 96:9172–9177.

Miege C, Ruffio-Châble V, Schierup MH, Cabrillac D, Dumas C, et al. 2001. Intrahaplotype polymorphism at the Brassica S-Locus. *Genetics*. 159:811–822.

Nishio T, Kusaba M. 2000. Sequence diversity of *SLG* and *SRK* in *Brassica oleracea* L. *Ann Bot*. 85:141–146.

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 Phytolog*. 172:577–587.

Opperman R, Emmanuel E, Levy AA. 2004. The effect of sequence divergence on recombination between direct repeats in Arabidopsis. *Genetics*. 168:2207–2215.

Prigoda NL, Nassuth A, Mable BK. 2005. Phenotypic and genotypic expression of self-incompatibility haplotypes in *Arabidopsis lyrata* suggests unique origin of alleles in different dominance classes. *Mol Biol Evol*. 22:1609–1620.

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aguade M. 2004. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. *Genetics*. 166:373–388.

Raspé O, Kohn JR. 2007. Population structure at the S-locus of *Sorbus aucuparia* L. (Rosaceae: Maloideae). *Mol Ecol*. 16: 1315–1325.

Richman AD, Kohn JR. 2000. Evolutionary genetics of self-incompatibility in the Solanaceae. *Plant Mol Biol*. 42:169–179.

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

Ruggiero MV, Jacquemin B, Castric V, Vekemans X. 2008. Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in *Arabidopsis halleri*. *Genet Res*. 90:37–46.

Saba El Leil M, Rivard S, Morse D, Cappadocia M. 1994. The S11 and S13 self-incompatibility alleles in *Solanum chacoense* are remarkably similar. *Plant Mol Biol*. 24:571–583.

Sainudiin R, Wong WSW, Yogeeswaran K, Nasrallah JB, Yang Z, Nielsen R. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the Human Major Histocompatibility Complex and the SRK of the plant sporophytic self-incompatibility system. *J. Mol. Evol*. 60:315–326.

Sato Y, Fujimoto R, Toriyama K, Nishio T. 2003. Commonality of self-recognition specificity of S haplotypes between *Brassica oleracea* and *Brassica rapa*. *Plant Mol Biol*. 52:617–626.

Sato Y, Sato K, Nishio T. 2006. Interspecific pairs of class II S haplotypes having different recognition specificities between *Brassica oleracea* and *Brassica rapa*. *Plant Cell Physiol*. 47: 340–345.

Schierup MH, Bechsgaard JS, Christiansen FB. 2008. Selection at work in self-incompatible *Arabidopsis lyrata*. II. Spatial distribution of S haplotypes in Iceland. *Genetics*. 180:1051–1059.

Schierup MH, Mable BK, Awadalla P, Charlesworth D. 2001. Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of *Arabidopsis lyrata*. *Genetics*. 158:387–399.

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

Schierup MH, Vekemans X, Christiansen FB. 1997. Evolutionary dynamics of sporophytic self-incompatibility alleles in plants. *Genetics*. 147:835–846.

Stone JL. 2004. Sheltered load associated with S-alleles in *Solanum carolinense*. *Heredity*. 92:335–342.

Suzuki T, Kusaba M, Matsushita M, Okazaki K, Nishio T. 2000. Characterization of Brassica S-haplotypes lacking S-locus glycoprotein. *FEBS Lett*. 482:102–108.

Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proc Natl Acad Sci USA*. 87:2419–2423.

Takebayashi N, Brewer PB, Newbigin E, Uyenoyama MK. 2003. Patterns of variation within self-incompatibility loci. *Mol Biol Evol*. 20:1778–1794.

Takuno S, Fujimoto R, Sugimura T, Sato K, Okamoto S, Zhang SL, Nishio T. 2007. Effects of recombination on hitchhiking diversity in the Brassica self-incompatibility locus complex. *Genetics*. 177(2):949–958.

Tang C, Toomajian C, Sherman-Broyles S, Plagnol V, Guo YL, Hu TT, Clark RM, Nasrallah JB, Weigel D, Nordborg M. 2007. The evolution of selfing in *Arabidopsis thaliana*. *Science*. 317: 1070–1072.

Uyenoyama MK. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics*. 147:1389–1400.

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, et al. 1996. Molecular analysis of two functional homologs of the S3 allele of the *Papaver rhoeas* self-incompatibility gene isolated from different populations. *Plant Mol Biol*. 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 Physiol*. 125:1012–1022.

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