

## ORIGINAL ARTICLE

# Studies of self-incompatibility in wild tomatoes: I. S-allele diversity in *Solanum chilense* Dun. (Solanaceae)

B Igic<sup>1,4</sup>, WA Smith<sup>2,4</sup>, KA Robertson<sup>1</sup>, BA Schaal<sup>2</sup> and JR Kohn<sup>3</sup>

<sup>1</sup>Department of Biological Sciences, University of Illinois-Chicago, Chicago, IL, USA; <sup>2</sup>Department of Biology, Washington University, St Louis, MO, USA and <sup>3</sup>Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA

We characterized the molecular allelic variation of RNases at the self-incompatibility (SI) locus of *Solanum chilense* Dun. We recovered 30 S-RNase allele sequences from 34 plants representing a broad geographic sample. This yielded a species-wide estimate of 35 (95% likelihood interval 31–40) S-alleles. We performed crosses to confirm the association with SI function of 10 of the putative S-RNase allele sequences. Results in all cases were consistent with the expectation that

these sequences represent functional alleles under single-locus gametophytic SI. We used the allele sequences to conduct an analysis of selection, as measured by the excess of nonsynonymous changes per site, and found evidence for adaptive changes both within the traditionally defined hypervariable regions and downstream, near the 3'-end of the molecule.

*Heredity* (2007) **99**, 553–561; doi:10.1038/sj.hdy.6801035; published online 15 August 2007

**Keywords:** self-incompatibility; *Solanum chilense*; tomato; S-alleles; molecular evolution; positive selection

## Introduction

Self-incompatibility (SI) is the ability of simultaneous hermaphrodites to avoid self-fertilization, mediated by a genetic mechanism for self-pollen recognition and rejection. In individuals with a gametophytic SI system, pollen tube growth is arrested in the style when the haploid pollen S-allele matches either of the S-alleles in a diploid pistil. Self-pollinations are disabled, as are some among close relatives and occasionally those among unrelated individuals bearing identical alleles. Population genetic analyses of SI, primarily concerned with the number and frequency of S-alleles, were pioneered by Wright (1939) whose interest was sparked by Emerson's (1938) discovery of the astonishing allelic diversity at the S-locus in the narrow endemic species *Oenothera organensis*. Wright (1939) noted that an active system ensures that there are no homozygotes, and that at least three alleles must be present in populations to ensure mate availability for individual plants. In addition, plants bearing rare S-alleles have more available mates than those with common ones. This negative frequency-dependent selection is expected to result in equal frequencies of all alleles (isoplethy) at equilibrium. Wright also noted that the equilibrium number of alleles is dependent on population size, and that the strength of negative frequency-dependent selection would be

attenuated when populations contained large numbers of alleles.

The discovery of the molecular basis of SI in the past two decades enabled studies of the molecular diversity at the S-locus, which for the first time introduced analyses of the temporal aspects of S-locus evolution. In all taxa examined, many S-alleles or lineages of S-alleles predate the origin of species that contain them. This pattern of extensive *trans*-specific polymorphism is another consequence of negative frequency-dependent selection (Ioerger *et al.*, 1990), which increases the frequencies of alleles that drift towards rarity. Negative frequency-dependent selection greatly increases the persistence of polymorphism at the S-locus, extending the expected time to coalescence.

In three plant families, Solanaceae, Plantaginaceae and Rosaceae, the specificity-encoding component of SI in the pistil is a functional S-RNase (McClure *et al.*, 1989; Sassa *et al.*, 1992; Xue *et al.*, 1996). Phylogenetic studies of the style component suggest that RNase-based gametophytic SI has a single origin in the early stages of eudicot evolution (Igic and Kohn, 2001; Steinbachs and Holsinger, 2002). To date, nine population surveys of S-RNases of Solanaceae have been published (Richman *et al.*, 1995, 1996b; Richman, 2000; Richman and Kohn, 2000; Wang *et al.*, 2001; Lu, 2002; Stone and Pierce, 2005; Savage and Miller, 2006). These analyses found considerable variation in the number of deep lineages and average diversity at the S-loci of different species (for example, Richman *et al.*, 1996a; Stone and Pierce, 2005). While most species show a pattern of broad *trans*-specific sharing of ancient lineages, species of the genera *Physalis* and *Witheringia* have S-alleles representing only a small subset of old S-allele lineages. Richman *et al.*

Correspondence: Professor B Igic, Department of Biological Sciences, University of Illinois-Chicago, M/C 067, 840 West Taylor Street, Chicago, IL 60607, USA.

E-mail: boris@uic.edu

<sup>4</sup>These authors contributed equally to this work.

Received 8 January 2007; revised 15 June 2007; accepted 20 June 2007; published online 15 August 2007

(1996a) attributed the finding of only three deep S-allele lineages in *Physalis crassifolia* to a historical reduction in the population, although other factors may contribute (Uyenoyama, 1997).

Another extrapolation of the simple genetic principles that govern the evolution of SI suggests that the S-locus should experience positive selection (selection for change). The traditional estimates of selection at the molecular level at the S-locus were based on whole-gene or sliding window nonsynonymous/synonymous rate ratios (for example, Clark and Kao, 1991; Richman *et al.*, 1996a), and generally had low power to detect positive selection (Yang and Bielawski, 2000). Analyses based on these ('approximate') methods were seminal for the development of this field of study (Hughes and Nei, 1988; Clark and Kao, 1991). However, appropriate statistical tests to detect selection at a particular amino-acid site lacked power, and the available methods often made highly simplifying and unrealistic assumptions (Yang and Bielawski, 2000). The advent of the likelihood-based models of codon substitution (Goldman and Yang, 1994; Muse and Gaut, 1994), and the application of the Empirical Bayes approach (Nielsen and Yang, 1998), led to the development and widespread use of these methods to assess adaptive molecular evolution at the level of individual codons. Such methods can also be used in conjunction with molecular genetic experiments to elucidate the mechanisms that create novel specificities (phenotypes) from genetic changes (for example, Matton *et al.*, 1997, 1999).

Here, we present the results of the first part of analyses of natural variation at the S-locus of *Solanum chilense*. We report the outcome of a survey of S-alleles from a broad geographic sample of natural populations and estimate the number of S-alleles in the species. We use phylogenetic analysis of S-alleles to determine the extent of polymorphism that arose before the divergence of this species from its congener *S. carolinense* and from a well-sampled member of the genus *Lycium* that possesses S-alleles representing many deep lineages (Savage and Miller, 2006). We then conduct an analysis of adaptive molecular evolution at the S-locus using the recovered samples. This provides a powerful analysis of selection because of the relatively large number of sequences analyzed and because all sequences shared their entire selective history continually, until present. Previous analyses of the sites under selection often analyzed sets of alleles from multiple taxa (Takebayashi *et al.*, 2003), or fewer and shorter sequences than those analyzed here (Savage and Miller, 2006).

## Methods

### The study species

The Chilean wild tomato, *S. chilense* Dun., is a self-incompatible herbaceous perennial found along the western slopes of the Andean cordillera from southern Peru to northern Chile and also in a disjunct region to the southwest, along the ranges that are flanked by the Atacama desert and the Pacific Ocean (Figure 1; Rick and Lamm, 1955). Although it was historically included in the highly polymorphic *S. peruvianum* complex, its status as a separate species was well established in biosystematic (Rick and Lamm, 1955) and phylogenetic studies



**Figure 1** Approximate geographic distribution of *S. chilense* Dun. (shaded area). See Table 1 for sampling location coordinates marked on the map.

(Peralta and Spooner, 2001). Populations are found from the Pacific coast to ~3800 m in the Andes and are frequently highly visible in the barren basin areas and seasonal washes. We collected seeds from individuals growing in natural populations (Table 1) of *S. chilense* in May 1995. Seeds were germinated, grown in a glasshouse and three to four styles were collected from each plant and either used fresh or stored at  $-80^{\circ}\text{C}$ .

### S-allele sequences

We extracted stylar RNA and performed cDNA synthesis using the protocols originated by Richman *et al.* (1995), and further described by Raspé and Kohn (2002), with a few modifications. Instead of using the poly-T primer for cDNA synthesis, we performed 3'-rapid amplification of cDNA ends (RACE) (Frohmann *et al.*, 1988), initiating cDNA synthesis with primer RaceA 5'-GCGCACGCGTC GACTAGTACTTTTTTTTTTTTTTTT-3'. PCR recipe for cDNA was the same as in Raspé and Kohn (2002). Primers PR1 5'-GAATTCAYGGNYTNTGGCCNGA-3' (Richman *et al.*, 1995) and RaceB 5'-GCGCACGCGTCTAGTAC-3' were used for amplification. As each individual is heterozygous at the S-locus, and S-alleles are highly divergent, we used the Invitrogen T-A Cloning Kit to separate the alleles. The cloned alleles, spanning the region from the conserved region 2 (C2) to the 3'-end of the transcript (~750 bp long), were identified by restriction digests and sequenced with standard M13 primers. Each allele was sequenced from two or more plants and independent bacterial clones, except for allele 13, which was amplified and sequenced directly from DNA with allele-specific primers designed for a separate study. In addition, we used a C1 forward primer (PR0 5'-CARCTNGHTHTVMVWTGGCC-3') along with allele-specific primers to amplify the region upstream of C2 for some sequences. One or two neutral nucleotide polymorphisms were occasionally detected

**Table 1** Location names, approximate geographic position (latitude south, longitude west, and altitude), and sampling information for natural populations of *S. chilense* collected in Chile and included in this study

Population	Lat	Long	Alt (m)	No. of plants	Genotypes
Socoroma	18.16	69.57	3200	8	(1,10); (10,14); (9,17); (10,17); (7,26); (6,10); (3,10); (5,25)
Livilcar	18.30	69.43	2400	1	(9,18)
Timar	18.45	69.41	2800	1	(6,30)
Mocha	19.79	69.29	2580	1	(3,23)
Pachica	19.89	69.34	2380	2	(8,18); (16,27)
Toconce	22.25	68.42	2770	1	(2,20)
Lincacabur	22.92	68.04	2600	1	(2,30)
Talabre	23.33	68.00	3100	9	(19,24); (17,18); (4,18); (4,21); (17,28); (7,18); (4,19); (15,18); (8,24)
Taltal	25.00	70.40	20	10	(5,30); (11,24); (6,19); (11,19); (22,30); (24,27); (27,30); (11,29); (12,13); (15,27)

within putative functional alleles. The sequence analyses presented here were conducted on consensus sequences. Analysis of variation within allele types will be presented elsewhere. We genotyped one offspring from each wild individual for analysis of allele number. We found four additional alleles in sibs grown at the same time. These additional sequences were included in our phylogenetic analyses as were four sequences from a previously published study of *S. chilense* (Kondo et al., 2002).

#### Allele number and identity

The estimate of the total number of alleles ( $N$ ) was made using the method of Paxman (1963). The 95% likelihood interval was computed according to O'Donnell and Lawrence (1984). Both assume isoplethy, as expected in a population at equilibrium (Wright, 1939). The assumption of isoplethy was tested using Mantel (1974) statistic (see Campbell and Lawrence, 1981a, b).

Numerous attempts were made to generate homozygous lines through self-pollinations, including bud pollinations, but all failed. Consequently, we performed controlled crosses between individuals that shared one putative S-allele in common to test whether the sequenced S-alleles are associated with the SI phenotype. If pairs of available genotyped plants (for example, dam BC × sire AB) contain one matching allele (B), only two progeny genotypes are possible (AB and AC). Such crosses are termed 'half-compatible' and all of the resulting progeny bear the single compatible paternal allele (A). Otherwise, if the putative alleles are not associated with the same incompatibility phenotype, because they are either nonfunctional or function as a different specificity (for example, dam B'C × sire AB), the progeny will be fully compatible and segregate into four genotypic groups (AC, AB', BB' and BC). Association of genotype and phenotype can be demonstrated if half-compatible crosses follow the expected patterns and each seedling contains the compatible paternal allele. Consequently, we performed PCR amplifications of the unmatched paternal allele on the cross progeny (sequences of primers employed for allele-specific PCR can be obtained from the authors upon request). Because the appropriate material was not consistently available, the crosses were performed on a subset of alleles. We performed these tests for 10 alleles, on 5–27 seedlings per cross, depending on the seed set from crosses, germination rate and time to germination. Every seedling from a half-compatible cross should contain the nonmatching paternal allele (for example, A;  $P=1$ ). In the fully compatible case,  $P=0.5$ . This expectation was

used to calculate the probabilities of observing all progeny bearing the paternal allele fortuitously, if the cross was fully compatible instead of half-compatible ( $P=(0.5)^n$ , where  $n$  = the number of progeny screened).

#### Genealogy of S-alleles

Amino-acid sequences of 34 S-allele sequences from *S. chilense* were aligned with *ClustalX* (Thompson et al., 1997), and adjusted manually to ensure the proper alignment of conserved cysteines and the two intron-exon junctions. In addition, we added 22 alleles from *S. carolinense* (Richman et al., 1995; Lu, 2006) and 24 from *Lycium parishii* (Savage and Miller, 2006). Nearly all alleles from *S. carolinense* have been shown in previous analyses to predate the genus *Solanum* and the alleles found in *L. parishii* also represent a broad sample of the ancient S-allele lineages found among Solanaceae (Richman et al., 1996a; Richman and Kohn, 2000; Savage and Miller, 2006). After alignment of amino acids, it was converted back to the original nucleotide sequences.

To generate a phylogenetic hypothesis for these sequences, we implemented Bayesian analyses using Markov chain Monte Carlo sampling with the Metropolis-Hastings-Green algorithm running four chains (three heated, one cold) for 5 000 000 generations in Mr Bayes v3.0 (Huelsenbeck and Ronquist, 2001). We sampled every 2500th tree in this analysis. The burn-in procedure was used to discard the initial 1001 trees. The posterior probabilities of individual clades were calculated using the remaining 1000 trees. We also used *PAUP\*4.0b10* (Swofford, 2002) to heuristically find the best tree in a maximum likelihood (ML) search. For this search, we used Modeltest 3.0 (Posada and Crandall, 1998) to find the optimal model of evolution (TVM+I+G), selected using the Akaike Information Criterion (Akaike, 1974).

#### Selection analysis

We used *PAML 3.14* (Yang, 1997) for estimates of nonsynonymous/synonymous rate ratio ( $\omega$ ) for the *S. chilense* alleles. Sites with values of  $\omega < 1$  indicate an excess of synonymous substitutions, and therefore can be inferred to evolve under purifying selection that constrains certain nucleotide changes at those sites. Alternatively, those sites for which a significant excess of nonsynonymous substitutions is recorded, or  $\omega > 1$ , suggest molecular evolution under positive selection pressure. Values of  $\omega = 1$ , are inferred as consistent with neutrality. We calculated posterior probabilities that codon sites are under positive selection using the *codeml* package in *PAML v3.14*. We compared values of

likelihood ratios for a number of nested models. Our methodology and nomenclature follow those of Yang and co-workers (Yang and Bielawski, 2000; Yang and Swanson, 2002; Wong *et al.*, 2004). We compared the likelihoods of more complex models to null models  $M_0$  (which assumes neutral codon evolution) and  $M_{1a}$  (nearly neutral codon evolution with two codon classes allowed to take on values from  $0 \leq \omega_0 \leq 1$  or  $\omega_0 = 1$ ) with those of a more complex model  $M_{2a}$ , which incorporates an additional positively selected site class ( $\omega_2 > 1$ ). We also examined the results of analyses using models  $M_7$  and  $M_8$ , which both assume a  $\beta$ -distribution for  $0 \leq \omega \leq 1$ , with the latter model allowing for an extra class of sites with  $\omega > 1$ . We test for positive selection by comparing two times the log-likelihood differences of  $M_{1a}$  vs  $M_{2a}$  and  $M_7$  vs  $M_8$ . The posterior probabilities that each codon belongs to one of the selection classes were calculated. Codons with a significant posterior probability ( $pp > 0.95$ ) of being in a positively selected class ( $\omega > 1$ ) are considered likely to have experienced positive selection.

## Results

### Allele number and identity

In the initial sample, a total of 30 different S-alleles were sequenced from the 34 *S. chilense* individuals. Subsequent genotyping of sibships revealed four additional alleles. All sequences were deposited in GenBank (accessions EF680077–EF680110). A previously published sequence (Kondo *et al.*, 2002; Figure 2) from plants without provenance information was identical to one in our data set. Three sequences from that study contained potentially nontrivial differences (more than two amino acids) when compared to ours. An alignment of sequences is available from the authors upon request. Pairwise amino-acid sequence identities ranged from 30–96%. The most closely related alleles, designated 13 and 18 (see Figure 2), differed by four amino-acids out of the total 113 available for comparison. Allele frequencies did not deviate from isoplethy ( $\chi^2 = 30.76$ ,  $P > 0.33$ ). The ML estimate of the number of S-alleles maintained in this species is 35 (95% CI range, 31–40).

We tested 10 alleles using half-compatible crosses and allele-specific PCR (Table 2). Allele determination based on sequence data is perfectly correlated with functionality. In all 10 cases, the null hypothesis of full compatibility is rejected. For a comparison, a cross was performed involving alleles expected to have different specificities (5 and 14), with a relatively high (86%) amino-acid identity (Table 2). The crossing result in this case confirms that they are allelic and encode different specificities, as the null hypothesis of full compatibility cannot be rejected.

### Genealogy of S-alleles

The sample from *S. chilense* represents many ancient S-allele lineages. Seventeen clades of S-alleles from *S. chilense* join the phylogeny at nodes ancestral to one or more alleles from *L. parishii*, indicating that much extant polymorphism predates the divergence of these genera (Figure 2). This represents a minimum estimate of the amount of polymorphism at the S-locus of *S. chilense* that predates the genus *Solanum* because additional

sampling from *S. chilense*, or from species in other genera, could only increase this number.

### Selection analysis

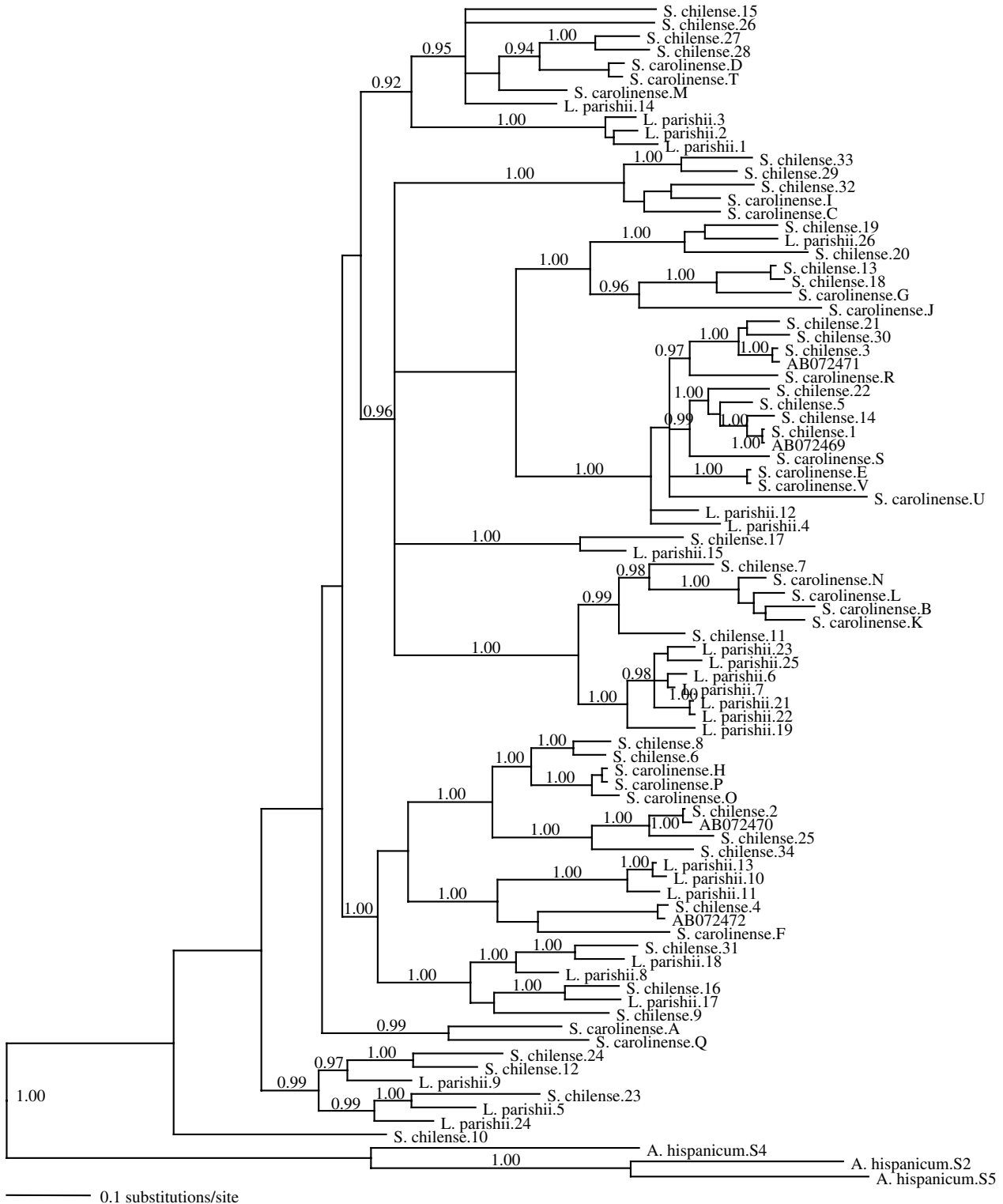
To conduct analyses of selection, we compared several models of codon substitution. Generally, the models that allow a positive selection category provided a significantly better fit than those that did not. In particular, model  $M_{2a}$  ( $\ln L = -10543.02$ ) is superior to  $M_{1a}$  ( $\ln L = -10571.61$ ,  $LRT = 57.18$ ,  $P \ll 0.0001$ ). The discrete three-category model ( $M_{2a}$ ;  $\ln L = -10518.10$ ) shows an additional improvement over  $M_{2a}$  ( $LRT = 49.8$ ,  $P \ll 0.0001$ ). Calculations under the best directly comparable model,  $M_{2a}$ , suggest three site classes with the following proportions (or prior probabilities,  $P$ ) and  $\omega$  values:  $P_0 = 0.35$ ,  $\omega_0 = 0.10$ ;  $P_1 = 0.39$ ,  $\omega_1 = 0.53$ ; and  $P_2 = 0.26$ ,  $\omega_2 = 1.40$ . Thus, for example, the mean value of the  $\omega$  ratio in the positive selection class is 1.40, indicating the presence of selective incorporation of new variants, and the prior probability that any site belongs in this class is 0.26. Similarly, under the second set of analyses, model  $M_8$  ( $\ln L = -10502.60$ ) provides a significantly better fit to the data than  $M_7$  ( $\ln L = -10526.48$ ;  $LRT = 47.75$ ,  $P \ll 0.0001$ ).  $M_8$  analysis assigns a prior probability  $P = 0.16$  that a site is under positive selection with  $\omega = 1.60$ . The prior probabilities and sequence data at each site are used to generate the posterior probabilities that a particular site is in each selection class ( $pp$ ; Figure 3). Sites inferred in the positive selection class, as well as the corresponding posterior probabilities, are listed in Table 3.

## Discussion

### Allele number and identity

The estimated number of alleles in *S. chilense* (35) falls well within the range estimated for other species of Solanaceae. Given that we recovered 34 alleles in total, after a broader, nonrandom search, it is likely that the species-wide diversity may be slightly higher, although we have no reason to believe that it is higher than the upper end of the estimated 95% CI (41). In addition, sequences from independently sampled plants (Kondo *et al.*, 2002), if they are indeed conspecifics, can be used to reinforce a higher species-wide estimate.

Estimates from the available molecular genetic surveys of natural populations in this family range from 15 to 44 (Lawrence, 2000; Lu, 2002; Stone and Pierce, 2005; Wang *et al.*, 2001; Savage and Miller, 2006). Most of these estimates come from analyses of variation present in a single population. Paradoxically, the lowest estimate for Solanaceae came from the only species-wide survey available to date, that from *S. carolinense* (Richman *et al.*, 1995). Estimates derived from sequence data from other species with RNase-based SI are very similar. Single populations of *Crataegus monogyna* and *Sorbus aucuparia* (Rosaceae) are estimated to harbor 27 and 24 alleles, respectively (Raspé and Kohn, 2002) and the species-wide estimate for *S. aucuparia* based on sampling from two populations is 40 (Raspé and Kohn, 2007). The only estimates that substantially differ from these come from crossing studies of clover species, *Trifolium repens* and *T. pratense* (Fabaceae), in which roughly 100–200 alleles are estimated to segregate in natural populations (cf. Lawrence, 2000). Any molecular verification of these



**Figure 2** A phylogenetic tree of S-alleles from the Solanaceae (*S. chilense*, *S. carolinense*, *L. parishii*), rooted with outgroup alleles from the Plantaginaceae (*Antirrhinum hispanicum*). Four previously published *S. chilense* alleles (Kondo *et al.*, 2002) are shown with their GenBank accession numbers. The best tree obtained in a heuristic ML search in PAUP\* (Swafford, 2002) is shown. Posterior probabilities in excess of 0.9 derived from a Bayesian analysis (Huelsenbeck and Ronquist, 2001) are shown above each branch of the tree. *S. chilense* alleles cover the entire depth of the tree and are contained within nearly all major allele lineages.

**Table 2** Tests of allelic association for ten sequences and corresponding incompatibility phenotypes

Genotypes		Allele tested	Allele for PCR	Cross results	Probability
Dam	Sire				
(15,3)	(15,20)	15	20	17/17	$P < 0.0001$
(4,18)	(4,21)	4	21,18 <sup>a</sup>	29/29	$P < 0.0001$
(17,9)	(17,10)	17	10	20/20	$P < 0.0001$
(18,9)	(18,15)	18	15	5/5	$P = 0.03125$
(20,6)	(20,2)	20	2	12/12	$P < 0.001$
(23,4)	(23,3)	23	3	27/27	$P < 0.0001$
(2,30)	(2,20)	2	20	5/5	$P = 0.03125$
(3,10)	(3,17)	3	17	18/18	$P < 0.0001$
(17,9)	(9,18)	9	18	17/17	$P < 0.0001$
(11,21)	(11,15)	11	15	12/12	$P < 0.001$
(5,25)	(10,14)	5 vs 14 <sup>b</sup>	10	6/13 <sup>b</sup>	$P > 0.7094$ NS

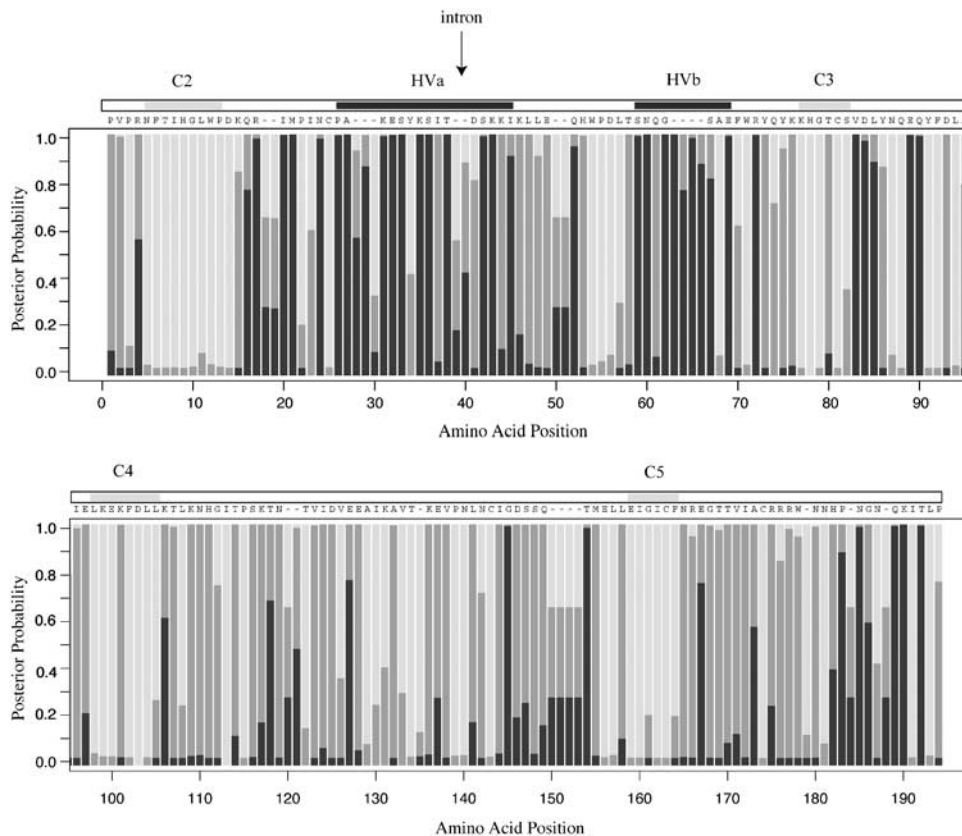
Plants bearing the genotypes indicated in the Dam/Sire columns were used in crosses. The second column lists the allele whose identity is tested for association between sequence and phenotype. It is followed by the sire allele to be transmitted to progeny, if the cross is half-compatible. This allele was used in PCR genotyping of the progeny. Results of the crosses are scored as (no. of progeny bearing the expected sire allele)/(no. of progeny tested). The final column indicates cumulative binomial probability of observing a particular progeny screening result if the sequence is, in fact, not associated with an allele phenotype and all four possible genotypes are present at expected frequencies (see text for details).

<sup>a</sup>Combined results of reciprocal crosses.

<sup>b</sup>Test involves two closely related alleles (5 and 14).

estimates is precluded by the lack of characterization of the S-locus in Fabaceae.

Occasionally, studies of S-allele variation in SI plants uncover sequences of S-like RNases that resemble S-alleles closely, as measured by sequence similarity, but do not function as S-alleles (Lee *et al.*, 1992; Liang *et al.*, 2003). For example, Lee *et al.* (1992) studied the S-like RNase X2 in *Petunia integrifolia* ssp. *inflata*, which is not linked to the S-locus. Although phylogenetic analyses of X2 show that it groups with functional S-RNases, and is expressed in the style, it is also monomorphic, and not involved in SI. Such genes likely originate as paralogous copies of S-alleles. Many more-distantly related genes in the same RNase superfamily exist, but with much lower amino-acid identities, and they are expressed in other plant tissues (for example, Hugot *et al.*, 2002). We found that the putative S-alleles we sequenced co-segregated with S-allele phenotypes by performing a series of half-compatible crosses to test the association between sequence and functional phenotype. The results confirm that all tested putative alleles, inferred from sequencing, are associated with a functional phenotype (Table 2). Unfortunately, we were unable to test the functional association of the two most closely related putative alleles (13 and 18), because the original plant with allele 13 was short lived (the first plant to die), and none of the plant's sibs genotyped subsequently carried this allele.



**Figure 3** Selection on S-allele sequences under model  $M_{2x}$ . Posterior probabilities ( $pp$ ) that any listed amino-acid site belongs to the positive selection site class with  $\omega > 1$  ( $\omega_2$ ) are given in black. The remainder of the total posterior probabilities is split between the two site classes with  $\omega < 1$  ( $\omega_0$ , light grey;  $\omega_1$ , dark grey). A reference sequence (S15) is given along the top of the figure. The arrow points to the location of the ubiquitous intron site.

**Table 3** Sites under positive selection ( $\omega > 1$ ), and the corresponding posterior probabilities

Site	$M_{2a}$ pp ( $\omega$ )	$M_8$ pp ( $\omega$ )	Site	$M_{2a}$ pp ( $\omega$ )	$M_8$ pp ( $\omega$ )
4	0.551		64	<b>0.762</b>	
16	0.763		65	<b>0.984*</b>	<b>0.583</b>
17	0.982*	0.572	66	<b>0.874</b>	
20	0.998**	0.841	67	<b>0.812</b>	
21	1.000**	0.951*	69	<b>0.993**</b>	<b>0.543</b>
24	0.982*	0.527	72	0.998**	0.780
26	<b>0.996**</b>	<b>0.787</b>	83	1.000**	0.995**
27	<b>0.999**</b>	<b>0.935</b>	84	0.973*	0.531
28	<b>0.558</b>		85	0.883	
29	<b>0.863</b>		89	0.997**	0.675
31	<b>0.992**</b>	<b>0.685</b>	90	0.993**	0.675
32	<b>0.997**</b>	<b>0.752</b>	106	0.600	
33	1.000**	1.000**	118	0.674	
35	1.000**	<b>0.956*</b>	127	0.763	
36	1.000**	<b>0.997**</b>	145	0.993**	0.706
38	1.000**	<b>0.990**</b>	154	0.986*	0.565
42	<b>0.992**</b>	<b>0.689</b>	167	0.750	
43	1.000**	<b>0.993**</b>	173	0.561	
45	<b>0.909</b>		183	0.881	
52	0.949		185	0.989*	0.713
59	<b>0.994**</b>	<b>0.799</b>	186	0.580	
60	1.000**	<b>0.988*</b>	189	0.994**	0.772
62	1.000**	<b>0.999**</b>	190	1.000**	0.964*
63	1.000**	<b>0.974*</b>	192	0.995**	0.753

Sites with posterior probabilities in excess of 95% ( $pp > 95\%$ ) are followed by\*, those with  $pp > 99\%$  are followed by\*\*. Boldface cells identify sites located in HVa or HVb. Columns listed as  $M_{2a}$  or  $M_8$  indicate the results of analyses under those models.

### Genealogy of S-alleles

*S. chilense* harbors a broad sample of the allelic variation at the S-locus found among Solanaceae. Like its congener *S. carolinense*, much of the variation at the S-locus appears to predate the genus *Solanum*, thought to have differentiated more than 5 million years ago, based on fossil seed from the mid- to upper miocene (Benton, 1993). Species restricted to the genera *Physalis* and *Witheringia* remain the only Solanaceae known whose S-allele complement contains only a restricted number (3–4; Richman and Kohn, 2000; Lu, 2002; Stone and Pierce, 2005) of ancient S-lineages.

### Selection analysis

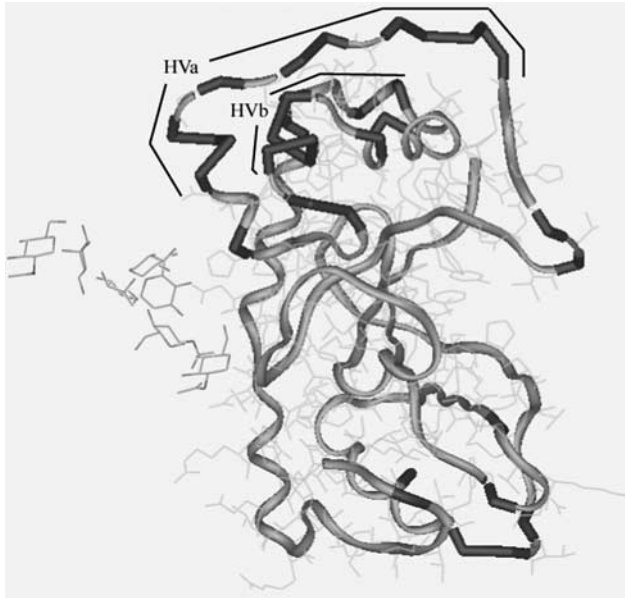
The detection of positive selection at the S-locus is unsurprising, but confirms and extends earlier studies. Previous analyses that employed the approximate methods (for example, Clark and Kao, 1991), as well as analyses using the codon-based methods (Takebayashi *et al.*, 2003; Savage and Miller, 2006) found that nonsynonymous mutations are preferentially fixed at some regions of the S-locus. However, our analyses of *S. chilense* sequences yielded two interesting results. Although we find that the traditionally defined hypervariable regions (HVa and HVb) are replete with selected sites (Figure 2), many other sites have experienced selection. Of 31 sites inferred under positive selection with  $pp > 95\%$ , 16 are in HVa and HVb, and 15 are outside. When all 48 sites inferred under positive selection ( $pp > 50\%$ ) are considered, 22 are found in HVa and HVb, and 26 are outside. These results provide further evidence (Charlesworth *et al.*, 2000; Takebayashi

*et al.*, 2003; Savage and Miller, 2006) that parts of the molecule other than HVa and HVb are almost certainly involved in encoding self-recognition specificity. It should be noted here that  $M_{2a}$  is known to falsely detect positive selection when there are a large number of neutral sites (Anisimova *et al.*, 2002). We include it here as a liberal estimator of selected sites.

Verica *et al.* (1998) have previously argued that the single successful experiment, in which the HV domains were swapped between two closely related alleles (Matton *et al.*, 1997), may not be indicative of the general pattern of evolution. Their arguments are based on studies that fail to recover a change in specificity in larger experimental domain swaps (cf. Kao and McCubbin, 1996). We note here that a disproportional amount of molecular adaptation in this gene still appears to take place in HVa and HVb. However, it remains possible that different groups of S-allele lineages may rely on alternate sites for changes in specificity. In addition, the sites involved in specificity may change throughout the history of allelic lineages. Finally, several selected sites are located far outside of HVa and HVb on the 3'-end of the primary structure (see Figure 3, residues 145, 154, 185, 189, 190 and 192). Extrapolated from the solved crystal structure of *Nicotiana glauca* S11 (Ida *et al.*, 2001), these sites are clustered at the external portion of the folded RNase, on the pole opposite to HVa and HVb (Figure 4, bottom). These findings are likely to hold for other species in this family, because S-alleles from different species share long periods of evolutionary history. The role of these sites in the SI reaction will remain equivocal until the appropriate experiments are conducted.

Takebayashi *et al.* (2003) and Savage and Miller (2006) find similar results to those presented here, but infer fewer strongly selected sites. Generally, these discrepancies are minor and may be, in part, explained by differences in alignment, lengths of sequences compared, and the sample of sequences considered. Takebayashi *et al.* (2003) used sequences from six Solanaceae species representing three genera. While broad *trans*-generic pattern of evolution ensures that most of the evolutionary history is captured in both studies, ours may have slightly increased power to detect selection, because all allelic lineages within a species share their entire evolutionary history. Savage and Miller (2006) analyzed selection at the S-locus in two species of *Lycium* separately. Their analysis used somewhat fewer sequences per species (24 in *L. parishii*, 16 in *L. andersonii*) and the sequences did not include the 3'-end of the molecule, but also found evidence for substantial selection outside of the hypervariable regions.

The recent discovery of the pollen component, an F-box-containing gene (Ushijima *et al.*, 2003; Qiao *et al.*, 2004; Sijacic *et al.*, 2004), in the Solanaceae and other families with RNase-based SI is expected to spur new studies of genetic variation at the S-locus. Its finding confirms that the SI recognition reaction involves two distinct components (Golz *et al.*, 2001), and supports the common origin of this form of gametophytic SI (Igic and Kohn, 2001; Steinbachs and Holsinger, 2002), but the exact mechanism of recognition and rejection remains clouded (Goldraij *et al.*, 2006; McClure, 2006). It is surprising that, and presently unclear why, the pollen F-box-containing gene fails to show the pattern of



**Figure 4** Model of the S-RNase molecule showing residues inferred to be under selection (black). The tertiary structure, solved for *Nicotiana alata* SF11, is used here as a generalized approximation for all aligned sequences. HVa and HVb are the commonly recognized large hypervariable regions, here found along the top of the molecule (approximately indicated by lines). For a detailed discussion of S-allele structures, see Parry *et al.* (1998); Ida *et al.* (2001).

reciprocal monophyly (Ushijima *et al.*, 2004), exhibited by S-RNase alleles from the three well-studied families. Therefore, it is clear that experimental focus of research on SI will shift in favor of patterns at the pollen gene and its interactions with S-RNases. Still, some lingering questions about the general patterns from the style part remain. For example, the role of gene conversion and recombination in the evolution of both pollen and style parts remains cloudy, at best, as does the number of changes required for shifts to new S-allele recognition phenotypes, given that the pollen and style S-alleles are tightly linked and have to coevolve. There is also surprisingly little research into the fascinating question regarding the quantity of neutral variation within functional S-alleles. Our future work in this series will use *S. chilense* and related species as models for studies of these and related questions.

## Acknowledgements

Support for this work was provided by NSF DEB-0108173 to JRK and DEB-0309184 to BI and JRK.

## References

- Akaike H (1974). A new look at the statistical model identification. *IEEE Trans Aut Contr* **AC-19**: 716.
- Anisimova M, Bielawski JP, Yang Z (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* **19**: 950–958.
- Benton MJ (1993). *The Fossil Record 2*. Chapman & Hall: London, 845p.
- Campbell JM, Lawrence MJ (1981a). The population genetics of the self-incompatibility polymorphism *Papaver rhoeas*. I. The

- number and distribution of S-alleles in families from three localities. *Heredity* **46**: 69–80.
- Campbell JM, Lawrence MJ (1981b). 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.
- Charlesworth D, Awadalla P, Mable BK, Schierup MH (2000). Population-level studies of multiallelic self-incompatibility loci, with particular reference to Brassicaceae. *Ann Bot* **85** (Supplement A): 227–239.
- Clark AG, Kao T-h (1991). Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. *Proc Natl Acad Sci USA* **88**: 9823–9827.
- Emerson S (1938). The genetics of self-incompatibility in *Oenothera organensis*. *Genetics* **23**: 190–202.
- Frohmann MA, Dush MK, Martin GR (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci USA* **85**: 8998–9002.
- Goldman N, Yang ZH (1994). Codon-based model of nucleotide substitution for protein-coding DNA-sequences. *Mol Biol Evol* **11**: 725–736.
- Goldraj A, Kondo K, Lee CB, Hancock CN, Sivaguru M, Vazquez-Santana S, *et al.* (2006). Compartmentalization of S-RNase and HT-B degradation in self-incompatible *Nicotiana*. *Nature* **439**: 805–810.
- Golz JF, Oh HY, Su V, Kusaba M, Newbigin E (2001). Genetic analysis of *Nicotiana* pollen-part mutants is consistent with the presence of an S-ribonuclease inhibitor at the S locus. *Proc Natl Acad Sci USA* **98**: 15372–15376.
- Huelsenbeck JP, Ronquist F (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.
- Hughes AL, Nei M (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**: 167–170.
- Hugot K, Ponchet M, Marais A, Ricci P, Galiana E (2002). A tobacco S-like RNase inhibits hyphal elongation of plant pathogens. *Mol Plant-Microbe Inter* **15**: 243–250.
- Ida K, Norioka S, Yamamoto M, Kumasaka T, Yamashita E, Newbigin E *et al.* (2001). The 1.55 Å resolution structure of *Nicotiana alata* SF11-RNase associated with gametophytic self-incompatibility. *J Mol Biol* **314**: 103–112.
- Igić B, Kohn JR (2001). Evolutionary relationships among self-incompatibility RNases. *Proc Natl Acad Sci USA* **98**: 13167–13171.
- Ioerger TR, Clark AG, Kao T-h (1990). Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proc Natl Acad Sci USA* **87**: 9732–9735.
- Kao T-h, McCubbin AG (1996). How flowering plants discriminate between self- and non-self pollen to prevent inbreeding. *Proc Natl Acad Sci USA* **93**: 12059–12065.
- Kondo K, Yamamoto M, Itahashi R, Sato T, Egashira H, Hattori T *et al.* (2002). Insights into the evolution of self-compatibility in *Lycopersicon* from a study of stelar factors. *Plant J* **30**: 143–153.
- Lawrence MJ (2000). Population genetics of the homomorphic self-incompatibility polymorphisms in flowering plants. *Ann Bot* **85**: 221–226.
- Lee H-S, Singh A, Kao T-h (1992). RNase X2, a pistil-specific ribonuclease from *Petunia inflata*, shares sequence similarity with solanaceous S-proteins. *Plant Mol Biol* **20**: 1131–1141.
- Liang LZ, Huang JA, Xue YB (2003). Identification and evolutionary analysis of a relic S-RNase in *Antirrhinum*. *Sex Plant Reprod* **16**: 17–22.
- Lu YQ (2002). Molecular evolution at the self-incompatibility locus of *Physalis longifolia* (Solanaceae). *J Mol Evol* **54**: 784–793.
- Lu YQ (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 et al. (1999). Production of an S-RNase with dual specificity suggests a novel hypothesis for the generation of new S-alleles. *Plant Cell* **11**: 2087–2097.
- Matton DP, Maes C, Laublin G, Qin XK, Bertrand C, Morse D et al. (1997). Hypervariable domains of self-incompatibility RNases mediate allele-specific pollen recognition. *Plant Cell* **9**: 1757–1766.
- McClure BA (2006). New views of S-RNase-based self-incompatibility. *Curr Opin Plant Biol* **9**: 639–646.
- McClure BA, Haring V, Ebert PR, Anderson MA, Simpson RJ, Sakiyama F et al. (1989). Style self-incompatibility gene products of *Nicotiana glauca* are ribonucleases. *Nature* **342**: 955–957.
- Muse SV, Gaut BS (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* **11**: 715–724.
- Nielsen R, Yang Z (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929–936.
- 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.
- Parry SE, Newbiggin D, Craik KT, Nakamura KT, Bacic A, Oxley D (1998). Structural analysis and molecular model of a self-incompatibility RNase from wild tomato. *Plant Physiol* **116**: 463–469.
- Paxman GJ (1963). The maximum likelihood estimation of the number of self-sterility alleles in a population. *Genetics* **48**: 1029–1032.
- Peralta IE, Spooner DM (2001). Granule-bound starch synthase (GBSSI) gene phylogeny of wild tomatoes (*Solanum* L. section *Lycopersicon* (Mill.) Wettst. subsection *Lycopersicon*). *Am J Bot* **88**: 1888–1902.
- Posada D, Crandall KA (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.
- Qiao H, Wang F, Zhao L, Zhou JL, Lai Z, Zhang YS et al. (2004). The F-Box protein AhSLF-S-2 controls the pollen function of S-RNase-based self-incompatibility. *Plant Cell* **16**: 2307–2322.
- Raspé O, Kohn JR (2002). S-allele diversity in *Sorbus aucuparia* and *Crataegus monogyna* (Rosaceae: Maloideae). *Heredity* **88**: 458–465.
- Raspé O, Kohn JR (2007). Population structure at the S-locus of *Sorbus aucuparia* L. (Rosaceae: Maloideae). *Mol Ecol* **16**: 1315–1325.
- Richman AD (2000). S-allele diversity in *Lycium andersonii*: implications for the evolution of S-allele age in the Solanaceae. *Ann Bot* **85**: 241–245.
- Richman AD, Kao TH, 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 Mol Biol* **42**: 169–179.
- Richman AD, Uyenoyama MK, Kohn JR (1996a). Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. *Science* **273**: 1212–1216.
- Richman AD, Uyenoyama MK, Kohn JR (1996b). S-allele diversity in a natural population of *Physalis crassifolia* (Solanaceae) assessed by RT-PCR. *Heredity* **76**: 497–505.
- Rick CM, Lamm R (1955). Biosystematic studies on the status of *Lycopersicon chilense*. *Am J Bot* **42**: 663–675.
- Sassa H, Hirano H, Ikehashi H (1992). Self-incompatibility-related RNases in styles of Japanese pear (*Pyrus serotina* Rehd. *Plant Cell Physiol* **33**: 811–814.
- Savage AE, Miller JS (2006). Gametophytic self-incompatibility in *Lycium parishii* (Solanaceae): allelic diversity, genealogical structure, and patterns of molecular evolution. *Heredity* **96**: 434–444.
- Sijacic P, Wang X, Skirpan AL, Wang Y, Dowd PE, McCubbin AG et al. (2004). Identification of the pollen determinant of S-RNase-mediated self-incompatibility. *Nature* **429**: 302–305.
- Steinbachs JE, Holsinger KE (2002). S-RNase-mediated gametophytic self-incompatibility is ancestral in eudicots. *Mol Biol Evol* **19**: 825–829.
- Stone JL, Pierce SE (2005). Rapid recent radiation of S-RNase lineages in *Witheringia solanacea* (Solanaceae). *Heredity* **94**: 547–555.
- Swafford DL (2002). PAUP\*: phylogenetic analysis using parsimony (\*and other methods), version 4.0. Sinauer, Sunderland, Mass: USA.
- Takebayashi N, Brewer PB, Newbiggin E, Uyenoyama MK (2003). Patterns of variation within self-incompatibility loci. *Mol Biol Evol* **20**: 1778–1794.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **24**: 4876–4882.
- Ushijima K, Sassa H, Dandekar AM, Gradziel TM, Tao R, Hirano H (2003). Structural and transcriptional analysis of the self-incompatibility locus of almond: identification of a pollen-expressed F-box gene with haplotype-specific polymorphism. *Plant Cell* **15**: 771–781.
- Ushijima K, Yamane H, Watari A, Kakehi E, Ikeda K, Hauck NR et al. (2004). The S haplotype-specific F-box protein gene, *SFB*, is defective in self-compatible haplotypes of *Prunus avium* and *P. mume*. *Plant J* **39**: 573–586.
- Uyenoyama MK (1997). Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics* **137**: 1389–1400.
- Verica JA, McCubbin AG, Kao T-h (1998). Letter to the Editor. *Plant Cell* **10**: 311–314.
- Wang X, Hughes AL, Tsukamoto T, Ando T, Kao TH (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.
- Wong WSW, Yang Z, Goldman N, Nielsen R (2004). Accuracy and power of statistical tests for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**: 1041–1051.
- Wright S (1939). The distribution of self-sterility alleles in populations. *Genetics* **24**: 538–552.
- Xue Y, Carpenter R, Dickinson HG, Coen ES (1996). Origin of allelic diversity in *Antirrhinum* S locus RNases. *Plant Cell* **8**: 805–814.
- Yang Z (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555–556.
- Yang Z (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**: 568–573.
- Yang Z, Bielawski JP (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**: 496–503.
- Yang Z, Swanson WJ (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among sites. *Mol Biol Evol* **19**: 49–57.