Plant Molecular Biology **42:** 169–179, 2000.

Evolutionary genetics of self-incompatibility in the Solanaceae

Adam D. Richman[1,*] and Joshua R. Kohn[2]
[1]*Plant Sciences Department, Montana State University, Bozeman, MT 59717-0346, USA (*author for correspondence; e-mail arichman@montana.edu);* [2]*Biology Department, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0116, USA (e-mail jkohn@ucsd.edu)*

Key words: gene genealogy, genetic variation, population bottlenecks, RT-PCR, *S* gene

Abstract

The self-incompatibility (*S*) gene in flowering plants has long been appreciated as an example of extreme allelic polymorphism maintained by frequency-dependent selection. Recent studies of population samples of *S*-allele sequences obtained by RT-PCR from five species of Solanaceae now reveal a picture of conspicuous inter-specific variation in both *S*-allele number and age. Explanations for this variation are examined with reference to current theory. We propose that changes in species' effective population size, particularly those associated with the evolution of different life histories, best account for interspecific differences in both the number and average age of *S* alleles.

Introduction

Self-incompatibility (SI) in flowering plants is of interest to biologists working at all levels of organization, from those concerned with the molecular basis of self recognition and rejection, to population biologists investigating the evolution of genetic polymorphism. Here we review empirical work which takes advantage of recent progress in understanding the molecular genetics of gametophytic self-incompatibility (e.g. [2, 4, 8, 38] to study population genetic variation at the self-incompatibility gene in the Solanaceae.

In the Solanaceae, self-incompatibility is determined by a single (*S*) gene with multiple alleles. A pollen tube is rejected in the style if it carries either allele expressed in the stylar tissue of the pollen recipient. An important consequence of this form of self-incompatibility for the evolution of genetic polymorphism is that the fitness of an *S* allele is inversely related to its frequency in the population. Alleles which become rare due to chance will on average more frequently encounter compatible mates compared to other (more common) alleles, and therefore will tend to increase in frequency in successive generations [41, 42]. Rare allele advantage resulting from negative frequency-dependent selection is responsible for two of the most conspicuous features of *S*-gene polymorphism: (1) the maintenance of large numbers of different *S* alleles within populations and (2) very long persistence of these alleles in populations and species compared to the persistence of selectively neutral allelic polymorphism.

The expectation that frequency-dependent selection can maintain large numbers of *S* alleles within populations is born out by empirical studies. For example, 10–50 different alleles have been estimated to occur within single natural populations in the Solanaceae [23]. In addition to maintaining many alleles in populations, frequency-dependent selection also tends to preserve allelic variation over time, because rare allele advantage will tend to rescue from extinction *S* alleles which become rare due to chance. Thus allelic polymorphism at the *S* gene can be extremely old. A key observation concerning the age of *S*-allelic polymorphism in the Solanaceae is trans-specific and often trans-generic evolution, where an allele found in one species is more closely related to an allele in another species or genus than to other

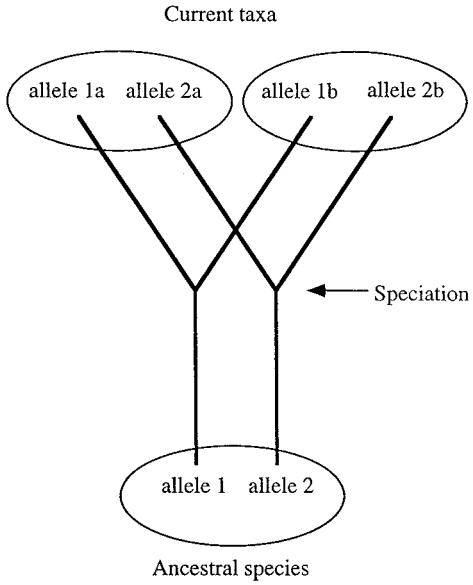

Figure 1. Trans-specific evolution as a consequence of long persistence of allelic lineages. Two allelic lineages present in an ancestral species are each inherited by two daughter species. In this case the closest relative of an allele in one daughter species is found not in that species but in the other daughter species.

conspecific alleles [3]. This observation indicates that the origin of the allelic lineage predates the origin of the species in which it is currently found (Figure 1). Trans-specific evolution is also observed in the sporophytic self-incompatibility system of the Brassicaceae [9], and in other genes under balancing selection, of which frequency-dependent selection at the *S* gene is a special case [36], including the MHC class II genes of vertebrates [11], and the self-recognition genes of some fungi [43].

In this chapter we review recent empirical studies of *S*-allele variation in the Solanaceae which find marked inter-specific differences in both *S*-allele number and age. We evaluate recent models of the evolution of *S*-gene polymorphism which attempt to account for this variation, and conclude that the evolution of ecological differences among species is the most important factor affecting inter-specific variation in allele number and persistence.

Background

In the Solanaceae, the stylar *S* gene product is an extracellular RNase which is both necessary and sufficient to determine the specificity of rejection of incompatible pollen [13]. This molecule contains relatively conserved domains known to be involved in RNA catalysis [16] as well as hydrophilic and highly variable regions which may play a role in determining allelic specificity [10, 15, 20]. Extreme sequence divergence is commonly observed among *S* alleles, with alleles from the same species often differing at 50% or more of amino acid residues [31]. Such extreme sequence divergence among alleles at this locus reflects the joint contributions of (1) the great age of many alleles, (2) diversifying selection for novel allelic specifities and (3) the lack of the homogenizing effect of recombination. While points 1 and 2 are readily understood as consequences of frequency-dependent selection acting to preserve allelic variation, the absence of recombination is also important in maintaining allelic variation, because repeated exchanges of homologous regions among different alleles would be expected to homogenize differences among alleles, obscuring our ability to reconstruct the history of allelic diversification using phylogenetic methods. However, recombination in the region of the *S* gene appears to be suppressed, greatly simplifying phylogenetic and molecular sequence analyses. Evidence for absence of recombination comes from molecular sequence analyses of the *S* gene [3] and nearby regions [1, 5]. For example, RFLP patterns and DNA sequences of regions flanking the *S* gene differ markedly among *S* alleles in the same species, indicating the absence of recombination which would be expected to homogenize these sequences.

Richman *et al.* [22–24, 27, 28] used RT-PCR to amplify *S* alleles from species in the Solanaceae using RNA extracted from styles. Amplification resulted in a single band of expected size. RFLP analysis of RT-PCR products from single individuals revealed the presence of two different sequences, as expected for the obligately heterozygous *S* gene. PCR products were cloned to separate the two partial *S* alleles prior to sequencing. In single donor matings, transmission of PCR-amplified sequences was consistent with expectation for a gametophytic *S* gene. For example, in crosses where donor and recipient share one *S* allele in common, only the compatible paternal allele was transmitted to offspring [22, 27]. While it is possible to determine the number of mating type alleles in a population with gametophytic self-incompatibility from large diallele crossing experiments [6, 12], the use of a molecular sampling technique allows work on species not amenable to greenhouse studies (e.g. [25]) while at the same time providing sequence information otherwise unavailable.

Two distinct aspects of genetic variation at the *S* gene are accessible through molecular sequence analysis at the population level. The first is the number of *S* alleles found in different populations or species. As the number of alleles increases under frequency-dependent selection, the strength of selection favoring rare alleles becomes attenuated. At equilibrium, selection favoring new alleles is balanced by the force of drift which removes variation. The number of alleles at equilibrium is expected to increase with increasing population size, because the effect of drift is weaker in large populations than in small ones. The number of alleles is also expected to evolve relatively rapidly in response to changes in population size, due to increased strength of selection relative to drift following population expansion, and conversely the increased importance of drift relative to selection upon reduction in population size.

A second kind of genetic variation, the age of alleles as inferred from either the number of trans-generic lineages represented in the sample or the amount of sequence divergence among alleles, is expected to evolve much more slowly than allele number. For example, the number of trans-generic lineages in the population or species is expected to evolve more slowly than the number of alleles, because loss of a trans-generic lineage requires the loss of all alleles represented within a lineage [35]. Allele age may also be inferred from molecular sequence divergence, because apparently very few changes may be required to alter allelic specificity [15, 29]. The accumulation of molecular sequence differences at the *S* gene can occur over millions of years, as inferred from levels of sequence divergence among alleles within species showing trans-generic evolution [28].

Variation in *S* allele number and age afford different opportunities for historical inference. Because the number of alleles is expected to respond relatively rapidly to changes in population size, it is expected to reflect more recent history (changes in population size) of these taxa. In contrast, the slower evolution of the age of *S* alleles reflects historical changes in population size which may have occurred in the distant past [23, 28], before the origin of extant species.

In analyzing molecular genetic data from population samples of *S* alleles, we take advantage of current theory on maintenance of variation at the *S* gene. For example, frequency-dependent selection is expected to maintain alleles at equal frequency, and this permits efficient estimation of the number of alleles in the population [21] and a corresponding asymmetric likelihood interval [18], based on the number of different alleles recovered from a population sample of a given size. The assumption of a uniform frequency distribution is evaluated using Mantel's test [14]. Formulae for estimating the expected number of alleles in a sample from a finite population are given by Yokoyama and Hetherington [45].

In analyzing variation in allele age among samples, we focus on the number of trans-generic lineages recovered in a population sample, as inferred from phylogenetic analysis. Because samples differ in size, we use a coalescent model to adjust for differences in sample size when estimating the amount of allelic extinction and origination (turnover) consistent with observed differences in the number of trans-generic lineages among samples [35]. Differing amounts of allelic turnover in population samples are then attributable to differences in population size, which determines the rate of loss of balanced lineages due to drift. The method is a specific application of coalescent analysis of balanced genealogies, an approach pioneered by Takahata who showed that the genealogical structure of balanced polymorphism is expected to be mathematically analogous to that of neutral polymorphism, differing by a constant scaling factor [33]. A coalescent model specific to gametophytic self-incompatibility was investigated by Vekemans and Slatkin [40].

The sophistication of coalescent models is a potential drawback if assumptions underlying them are not met. Uyenoyama [39] developed tree-shape statistics including R_{sd}, the sample size-independent ratio of the sum of all terminal branch lengths in the genealogy relative to the depth (the length from the base to the tip) of the genealogy. We use this statistic to test whether sequences recovered from a sample have a genealogical structure consistent with assumptions of the coalescent model. We compare the R_{sd} value derived from sampled alleles to the expected mean and variance of this statistic estimated by computer simulation of populations under gametophytic self-incompatibility.

In comparative analysis of samples from different species, we will often be interested in determining whether samples differ statistically in various aspects of their genealogy, correcting for inevitable differences in sample size. The use of Uyenoyama's statistics is intended to test the fit of genealogy to theoretical expectation, and is not necessarily the most efficient way of examining whether two genealogies differ significantly in shape. In particular, the method corrects

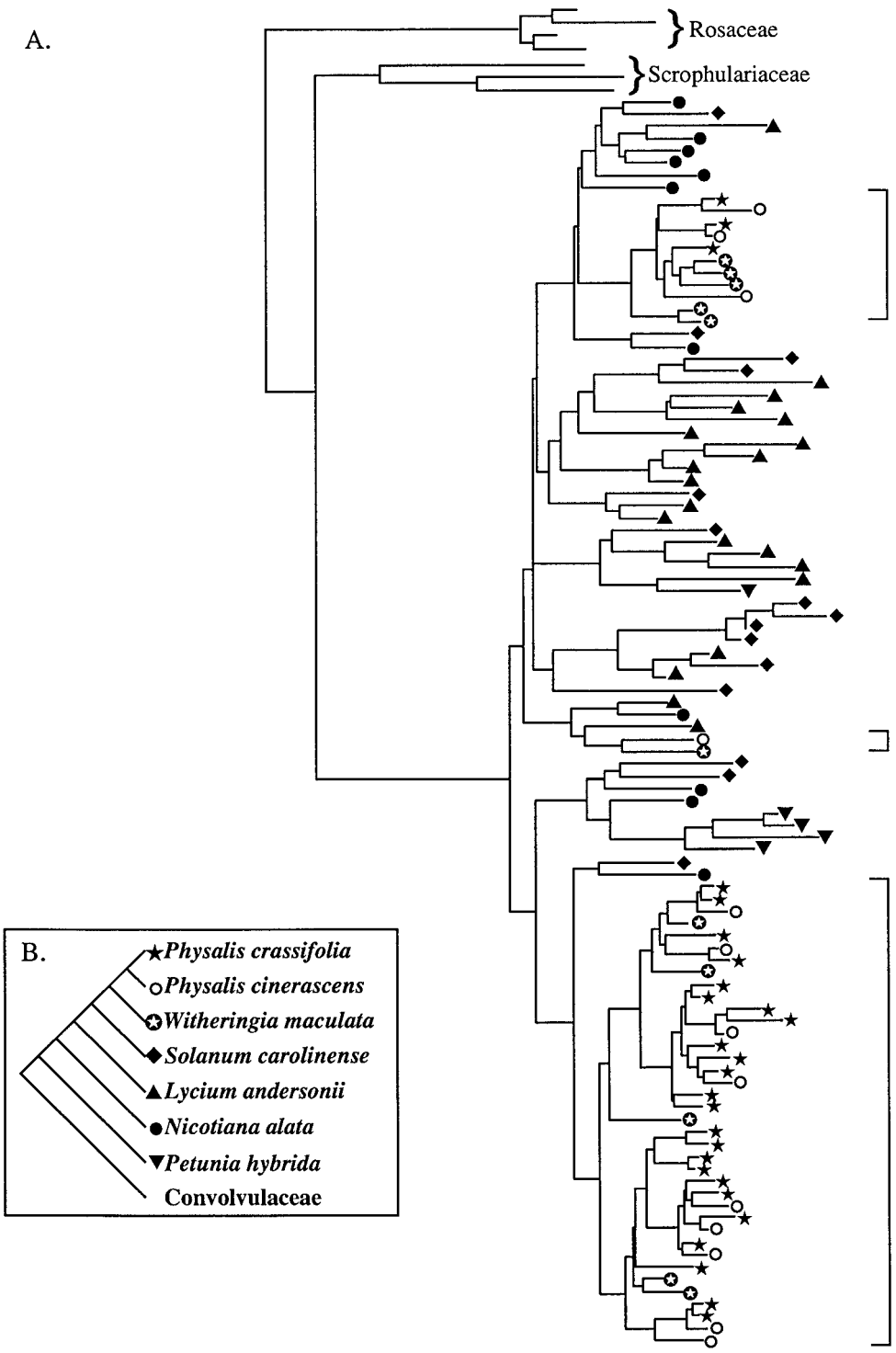

Figure 2. A. Neighbor joining topology for *S*-allele DNA sequences in the Solanaceae. Citations for published *S* sequences and/or their GenBank accession numbers are given in [24–26] with the exception of the Rosaceae [2, 44]. Estimation of pairwise distances using the Kimura 2-parameter model and phylogeny construction were carried out in PAUP* [32]. DNA sequences used correspond to amino acid positions 1–129 in Figure 1 of Richman *et al.* [28]. B. Phylogeny of selected genera in the Solanaceae [19].

for differences in sample size based on the assumption of equilibrium conditions which may not hold. As an alternative, we use resampling statistics to correct for differences in sample size when comparing genealogies of different population samples. We then take advantage of theoretical models in investigating the cause of observed differences among species in allele age.

Results of empirical studies of *S* gene polymorphism

Variation among species in both the number and age of S *alleles*

Studies in five species of Solanaceae reveal marked differences in allele number recovered from population samples (Table 1). Three species (*Solanum carolinense*, *Physalis cinerascens*, and *Witheringia maculata*) have estimated allele numbers of 15 or fewer while the remaining two (*Physalis crassifolia* and *Lycium andersonii*) have estimated allele numbers in excess of 35. The 95% likelihood intervals on the estimates from these two groups do not overlap (Table 1).

Phylogenetic analysis of *S*-allele sequences sampled from multiple genera of Solanaceae finds a striking difference among species in the relative age of alleles as well. We estimated the number of trans-generic lineages as the number of allelic lineages in a sample that predate the divergence of the four genera under study (*Lycium*, *Physalis*, *Solanum*, and *Witheringia*). For instance, an allele or group of alleles from *Solanum carolinense* is considered trans-generic if it joins the genealogy of *S* alleles from Solanaceae (Figure 2A) at or above a node which includes an allele from *Lycium*, *Nicotiana*, or *Petunia*. Such a node must predate the divergence of *Lycium* from the other three genera from which population samples are drawn (Figure 2B). Confidence intervals on the number of transgeneric lineages (Table 1) were estimated by bootstrap resampling of the data and counting the number of trans-generic lineages for each species in each bootstrap replicate. Whereas alleles sampled from *S. carolinense* and *L. andersonii* show extensive trans-generic evolution, estimates for *Physalis* and *Witheringia* spp. are much lower, indicating more extensive lineage turnover (Table 1). Further, all alleles sampled from *Physalis* and *Witheringia* arise from the same limited number of trans-generic lineages,

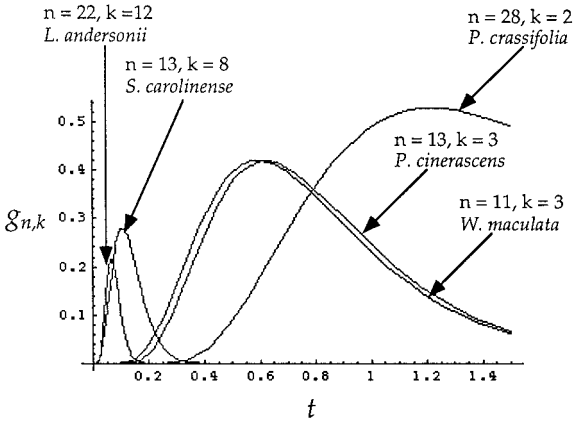

Figure 3. Likelihood estimates ($g_{n,k}$) of a given amount of allelic origination and extinction (t) given the number of trans-specific lineages (k) detected in a population sample of size n [35]. A value of $t = 1$ indicates complete turnover (the loss of all lineages present at time 0). The definition of a trans-specific lineage used to obtain k (the number of lineages sampled which were present at time $t = 0$) is any lineage pre-dating the divergence of the four genera *Solanum*, *Lycium*, *Physalis* and *Witheringia*, as inferred from Figure 2. Because the time that has elapsed for the observed n and k is the same for all taxa, differences in the likelihood estimate of t may be attributed to differences in effective population size (and/or generation time).

indicating reduction of the number of trans-generic lineages occurred in a common ancestor of these taxa (Figure 2).

Because the number of alleles sampled varied among species, we estimated the amount of allele turnover (origination and extinction) consistent with observed number of trans-generic lineages in a sample using the likelihood approach of Takahata [34, 35]. The number of trans-generic lineages in a population sample was used to estimate lineage turnover since the most recent common ancestor of these genera. Likelihood estimates of the amount of allelic turnover broadly overlap for *S. carolinense* and *L. andersonii* as do the estimates for *Witheringia maculata* and *Physalis* spp., while there is little overlap between the estimates from these two groups (Figure 3). This analysis indicates differences in the amount of turnover between these groups persists when differences in sample size are corrected for, supporting the result that a common explanation is required for the observation of extensive trans-generic evolution in *Solanum* and *Lycium* relative to *Physalis* and *Witheringia*. Richman *et al.* [28] interpreted recent *S*-allele diversification in *P. crassifolia* as evidence for a severe population restriction in the history of this taxon which caused the loss of most trans-generic lineages,

Table 1. Allele number and age in five species of Solanaceae. Sample sizes, numbers of alleles recovered, and the maximum likelihood estimate of population allele number and corresponding 95% likelihood interval are given for each species. For *S. carolinense*, two populations were assayed allowing the estimate of overlap between the two samples and the species-wide estimate of allele number. The number of trans-generic lineages is the number of alleles or clades of alleles from each species that join the genealogy in Figure 2 at a node as old as or older than the most recent common ancestor of these four genera. Confidence intervals on the number of trans-generic lineages were estimated using 100 bootstrap replicates of the data in Figure 2. R_{sd} (see text) for the five species was estimated for species-specific gene genealogies determined using the program KITSCH in PHYLIP [7].

Species	*Solanum carolinense*		*Physalis crassifolia*
Population	NC	TN	Deep Canyon, CA
Plants sampled	12	14	22
Alleles recovered	12	11	28
ML estimate of allele number	12 (11–15)	14 (13–18)	44 (33–60)
ML estimate of overlap	12		
Species allele number	15		
Trans-generic lineages (95% CI)	8 (8−12)		2 (2−3)
R_{sd}	6.27***		2.64‡
Species	*Physalis cinerascens*		*Lycium andersonii*
Population	TX		Granite Mtns., CA
Plants sampled	12		16
Alleles recovered	12		22
ML estimate of allele number	14 (12−20)		38 (28−55)
Trans-generic lineages (95% CI)	3 (3−4)		12 (9−14)
R_{sd}	2.59‡		5.42***
Species	*Witheringia maculata*		
Population	Monte Verde, Costa Rica		
Plants sampled	12		
Alleles recovered	10		
ML estimate of allele number	14−15 (12−23)		
Trans-generic lineages (95% CI)	3 (3−4)		
R_{sd}	1.48		

*** $P<0.001$; ‡ $0.1>P>0.05$

followed by rediversification within the few surviving lineages. Subsequent work summarized here (Figure 2A; and see [24, 26]) shows that the bottleneck event occurred in a common ancestor of *Witheringia* and *Physalis*.

Testing the fit of allelic genealogies to theoretical expectation

We examined whether species-specific allelic genealogies conformed to expectation for a coalescent model specific to balancing selection at the *S* gene using the tree-shape statistic $R_{sd} = S(1 - 1/n)/D$ [39] where S is the sum of terminal branch lengths, n is the number of sequences in the sample and D is the depth of the genealogy (Table 1). We find that the genealogies of *S. carolinense* and *L. andersonii* deviate significantly from expectation, in that terminal branches are too long relative to the depth of the genealogy. Allelic genealogies for *Physalis* and *Witheringia* species tend in the same direction (towards long tips), although in these cases the deviation is not significant. Although application of Uyenoyama's [39] method fails to indicate that allelic genealogies in *Physalis* and *Witheringia* deviate significantly from expectation, there is reason to be skeptical about this result. These taxa experienced a dramatic loss of trans-generic lineages, followed by rediversification of *S* alleles within surviving lineages (Figure 2). It is therefore unlikely these genealogies meet equi-

librium assumptions, which include the assumption of constant population size. Failure of Uyenoyama's method to detect a deviation from equilibrium upon diversification following a relatively recent bottleneck event suggests that the method is insensitive to important aspects of variation in the *S*-gene genealogy. As a consequence, we have used resampling methods which control for differences in sample size in order to compare the shape of allelic genealogies from post-bottleneck species.

Statistical comparison of allelic genealogies by resampling analysis

We use a resampling procedure to show that there are significant differences in the shape of allelic genealogies of post-bottleneck species. We argue that inter-specific differences in the shape of the allelic genealogy are due to differences in the rate of *S*-allelic diversification in different taxa.

The genealogy for *W. maculata* shows diversification from the same, limited number of lineages found in *Physalis* (Figure 2). However, the estimate of the number of alleles from the population sample indicates that it has significantly fewer alleles than *P. crassifolia* (Table 1), suggesting that it has diversified at a slower rate since the bottleneck event. Limited diversification is also suggested by the observation that the limited number of alleles detected in this species nevertheless tend to group together in phylogenetic analysis (Figure 2). To evaluate the significance of this observation, we tested whether an equivalent size sample of alleles from *P. crassifolia* would be expected to contain similarly close pairs of alleles using a resampling analysis. Tree measures (total genealogy length, genealogy depth, total terminal branch length) of the *W. maculata* sample were compared to random samples of alleles from *P. crassifolia* of equivalent size. Total length of the genealogy is the sum of all branch lengths, total terminal branch length is the sum of the lengths of all terminal branches, and depth of the genealogy is the sum of the branches from the base to any terminal tip.

The *W. maculata* genealogy differs significantly from the resampled genealogies [26] in both total length of the genealogy and total terminal branch length but not in the depth of the genealogy, indicating that the cause of the difference is significantly shorter terminal branches, and not differences in branch lengths deeper in the genealogy. This occurs because several close pairs of alleles were recovered

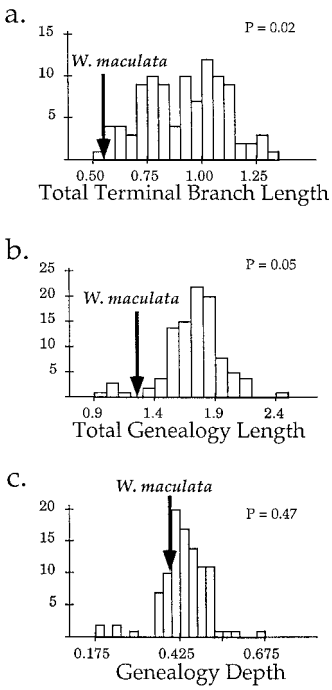

Figure 4. Frequency distributions for statistics of *S*-allele genealogies generated by randomly resampling subsets of *P. crassifolia* sequences. Arrows indicate the value of these statistics for *W. maculata*. a. Total terminal branch length, the sum of the lengths of all terminal branches. b. Total length of the genealogy, the sum of all branch lengths. c. Depth of the genealogy, the sum of the branches from the base to any terminal tip. Some *P. crassifolia* alleles (P12, 15, 18, 19, 23, and 24; see [28]) were omitted from resampling because the available partial sequence information would result in biased estimates of branch lengths [26]. Trees for resampled *P. crassifolia* sequences were estimated using the KITSCH algorithm in PHYLIP. The method assumes a molecular clock, and this assumption was examined using the computer package LINTRE [37]. The assumption of a molecular clock is not rejected for data sets considered here [24, 28].

in the *W. maculata* sample, whereas resampling statistics indicate that an equal-size sample of *P. crassifolia* would not be expected to contain such close pairs (Figure 4). Because random sampling tends to recover the deeper branches of the genealogy preferentially [39], recovery of close pairs of alleles indicates that we have sampled a greater fraction of the alleles in *W. maculata* than in *P. crassifolia*, implying that there are fewer *S* alleles species-wide in *W. maculata* compared to *P. crassifolia*. Because the time since the bottleneck event is the same for both taxa, lower *S*-allele number in *W. maculata* relative to *P. crassifolia* also indicates diversification of *S* alleles in *W. maculata* has occurred at a slower rate than in *P. crassifolia*.

The sample of *S* alleles from *P. cinerascens* is similar to that for *W. maculata* in having significantly

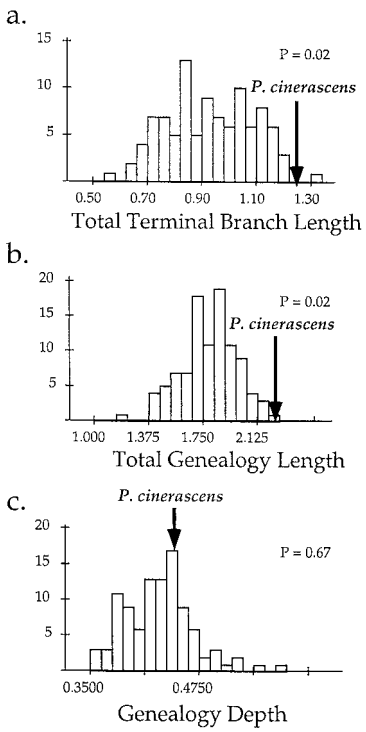

Figure 5. Frequency distributions for statistics of *S*-allele genealogies generated by randomly resampling subsets of *P. crassifolia* sequences. Arrows indicate the value of these statistics for *P. cinerascens*. a. Total terminal branch length, the sum of the lengths of all terminal branches. b. Total length of the genealogy, the sum of all branch lengths. c. Depth of the genealogy, the sum of the branches from the base to any terminal tip. Statistics were calculated for *P. cinerascens* sequences excluding PCIN1, which does not fall into either of the main clades of *Physalis* alleles (see Figure 2). Because PCIN1 is relatively divergent from all other *Physalis* alleles, its exclusion is conservative with respect to testing the hypothesis that *P. cinerascens* has long terminal branch lengths relative to *P. crassifolia*. Some *P. crassifolia* alleles (P12, 15, 18, 19, 23, and 24; see [23]) were omitted from resampling because the available partial sequence information would result in biased estimates of branch lengths [24].

fewer alleles than is found in *P. crassifolia* (Table 1). However, a parallel resampling analysis comparing the genealogy of *P. cinerascens* to that obtained for equivalent size resamples of *P. crassifolia* alleles finds that terminal branches are significantly longer in *P. cinerascens* than in *P. crassifolia*, opposite to the result obtained for the *W. maculata* genealogy (Figure 5, see also [24]). Whereas the low *S*-allele number in *W. maculata* appears to have been due to limited diversification over an extended period, the low *S*-allele number in *P. cinerascens* appears to be more recently derived, due to loss of alleles from an ancestor with high *S*-allele number similar to its congener *P. crassifolia*. All alleles recovered from *P. cinerascens* are most closely related to alleles or lineages found in *P. crassifolia* rather than to conspecific alleles (Figure 2) indicating that none of the alleles sampled from *P. cinerascens* arose after divergence from the common ancestor of both species. Conversely, many alleles in the more diverse *P. crassifolia* are also quite closely related to alleles from *P. cinerascens* (Figure 2), indicating that most alleles in both taxa arose prior to species divergence and therefore that low allele number in *P. cinerascens* is recently derived from a common ancestor with allele number similar to that found currently in *P. crassifolia*. Importantly, the observation of significantly longer terminal branches for *P. cinerascens* relative to its congener suggests that the loss of alleles did not occur at random. Instead, more divergent alleles were preferentially maintained during reduction in allele number resulting in significantly longer terminal branches than found in resampled data for *P. crassifolia* [24].

Discussion

There is marked variation among species in both the number and age of S *alleles*

Samples from different species vary significantly in both the number and age of *S* alleles (Table 1), whether the latter is inferred from the degree of trans-generic evolution or from variation among species in terminal branch lengths of species-specific genealogies (Figures 4 and 5). While significance tests take account of variation in sample size, most estimates of the number of alleles in Table 1 are based on samples from single populations, raising the possibility that other populations may contain additional alleles. For the inference of significant inter-specific differences in the number of alleles this issue is important only with respect to samples with low *S*-allele number (*S. carolinense*, *P. cinerascens*, *W. maculata*), because additional samples from species with high number (*P. crassifolia*, *L. andersonii*) can only increase current estimates. The estimate of the number of alleles in *S. carolinense* is based on two widely separated population samples [22], and is therefore not subject to this concern. Estimates for *P. cinerascens* and *W. maculata* are based on single population samples, but genealogical analyses indicate that *P. cinerascens* has undergone a recent and significant reduction in allele number relative to its congener *P. crassifolia* whereas a parallel analysis

in *W. maculata* indicates that it has maintained low *S*-allele number for a longer time [26]. Because the genealogy is in large measure older than current population structure, this constitutes evidence for lower diversity at the global scale in these taxa.

There is no association between S*-allele number and age*

The shapes of *S*-allele genealogies for *S. carolinense* and *L. andersonii* deviate from expectation under balancing selection in that terminal branch lengths are too long. To explain the observation of long terminal branches relative to expectation, Uyenoyama [39] suggested that the accumulation of deleterious recessive mutations arising in close linkage with the obligately heterozygous *S* gene could inhibit allelic origination. A new allele arising in the population would share a similar genetic background to the allele from which it is descended, resulting in the expression of linked genetic load in individuals carrying both ancestor and descendant alleles. This would result in a relative fitness disadvantage of both alleles compared to other alleles in the population, leading to the preferential elimination of one or the other allele and leaving no evidence of diversification on the genealogy. Under this view, the importance of a bottleneck event, as was inferred for the common ancestor of *Witheringia* and *Physalis* spp., is to purge genetic load which has accumulated over time, allowing rapid *S*-allele diversification and an increase in the number of alleles maintained at equilibrium. Uyenoyama's model therefore predicts a negative association between the age and number of alleles. This prediction was consistent with inferences based on samples then available for *S. carolinense* and *P. crassifolia* [28, 39]. However, further population samples from additional taxa now indicate that variation in *S*-allele number is decoupled from variation in average age of *S* alleles (Table 1). Species which share a historical bottleneck in *S* lineages show the same range of variation in *S*-allele number as those that do not. Moreover, the genetic load hypothesis cannot explain significant differences in diversification rate for *W. maculata* and *P. crassifolia*, since species arising after the bottleneck event share the same history of purging. These results suggest that postulated differences in origination rate are not the primary determinants of variation in the number of *S* alleles maintained in these species.

Evidence that more divergent alleles are preferentially maintained

We propose that long terminal branches relative to expectation may result from violations of assumptions of the coalescent model other than change in the origination rate. In particular, there is evidence from *P. cinerascens* that more divergent alleles are selectively retained during a reduction in *S*-allele number. This may occur given linked genetic load [39], or if the ability to discriminate self from non-self is a function of the degree of pairwise sequence divergence and matings between close alleles are sometimes falsely rejected. The latter mechanism differs from a mechanism of genetic load in that the inhibiting effect on origination of new alleles cannot be purged, so no association between allele number and age is expected. In this case, species' *S*-allele genealogies with long terminal branches may be derived secondarily by the selective maintenance of more divergent alleles. In support of this possibility, analyses presented here indicate that low *S*-allele diversity in *P. cinerascens* was apparently derived from a more diverse ancestor similar to the congener *P. crassifolia*, and alleles which were retained were significantly more divergent on average, generating a genealogy with significantly longer terminal branches. The evolution of a weedy habit and consequent reduction in population size may have triggered the loss of *S*-allele diversity in *P. cinerascens* [24] (and see below). Thus perturbations of the genealogy driven by the evolution of life history characters affecting population structure and persistence may best account for the observation of long terminal branches and high average age of *S* alleles in some population samples.

Association between the ecology of species and allele number

We propose that differences in species' life histories affecting population size and persistence are the primary contributors to inter-specific variation in *S*-allele number. In particular, a conspicuous aspect of life history variation among species in the Solanaceae is the frequent evolutionary transition between a weedy and a non-weedy habit, where weediness is defined by the exploitation of transient disturbed habitats. Weedy species typically share a number of life history characteristics suggesting small population size compared to non-weedy taxa, including small and short-lived populations, and (partial) clonal reproduction. *S. carolinense*, *P. cinerascens* and *W. maculata*, species

with low allele number, are weedy herbaceous species with small and short-lived populations which suggests reduced species effective population size relative to *P. crassifolia* and *L. andersonii*, which are common perennial species of undisturbed habitats.

Low S*-allele number has been achieved by different evolutionary pathways*

Our analyses indicate that species with low allele number have arisen by different evolutionary processes, with the implication that the common thread uniting the observation of low *S*-allele number in different taxa is the evolution of a weedy life history and associated population structure. Genealogical analysis suggests that low *S*-allele number in *P. cinerascens* is due to a relatively recent reduction in diversity from a more diverse ancestor similar to its congener *P. crassifolia* at present. Perhaps the evolution of a weedy habit in *P. cinerascens* triggered a reduction in effective size and corresponding reduction in the number of alleles maintained at equilibrium. Low *S*-allele number in *W. maculata* has an origin different from that inferred for *P. cinerascens*. *W. maculata* has apparently maintained low *S*-allele number for a longer time period. The observation of relatively few but closely related *S* alleles in *W. maculata* is consistent with the proposition that small effective size has limited the extent of *S*-allele diversification in this weedy taxon. The genus *Witheringia* is composed of herbaceous species inhabiting light gaps in tropical cloud forest, suggesting that limited population size may be a common and possibly long-standing feature of species in this genus.

Conclusions

A hypothesis that population size affects the number of *S* alleles in the species assumes that the number of alleles are at an equilibrium. In contrast to the evolution of the *S*-gene genealogy, which shows the effects of historical perturbations for millions of generations [28], the number of *S* alleles is expected to approach equilibrium more quickly [35], making this equilibrium interpretation more reasonable. The connection between life history characters and species effective size also presumes tight linkage between local *S*-allele number and global (species) population size. Recent theoretical results show that population subdivision for a given level of gene flow is far lower for a gametophytic *S* gene than for a neutral marker [30], as had already been shown for genes under overdominant balancing selection [34]. This is not surprising, because a migrant allele not already represented in the population has an immediate advantage in finding compatible matings, increasing the effective migration rate at the *S* locus relative to neutral markers. Thus the estimate of population size is necessarily averaging over a larger area than neutral markers. An important future goal is to determine the scale at which population size is being measured in a local sample by determining the spatial distribution of *S*-allele diversity in natural populations [12, 17]. It is possible, given the high effective migration at the *S* gene, that the local sample reflects diversity at the species level, as suggested by the observation that weedy species appear to show low allele number not only at the local but the global scale as well, due to the combination of high effective migration and high population turnover [22]. High effective migration at the *S* gene provides a theoretical justification for the interpretation of local estimates of *S*-allele number as functions of inter-specific differences in life history characters. Just as the long persistence of balanced genetic polymorphism has been recognized as providing a unique opportunity for historical inference, *S*-allele number offers a similar opportunity with respect to estimating population size over a larger area.

Acknowledgement

This study was supported by NSF grants DEB95-27835 (to J.R.K. and A.D.R.) and DEB98-70766 (to A.D.R.).

References

1. Anderson MA, McFadden GI, Bernatzky R, Atkinson A, Orpin T, Dedman H, Tregear G, Fernley R, Clarke AE: Sequence variability of three alleles of the self-incompatibility gene of *Nicotiana alata*. Plant Cell 1: 483–491 (1989).
2. Broothaerts W, Janssens GA, Proost P, Broekaert WF: cDNA cloning and molecular analysis of two self-incompatibility alleles from apple. Plant Mol Biol 27: 499–511 (1995).
3. Clark AG, Kao T-h: Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc Natl Acad Sci USA 88: 9823–9827 (1991).
4. Clarke AE, Newbigin E: Molecular aspects of self-incompatibility in flowering plants. Annu Rev Genet 27: 257–279 (1993).
5. Coleman CE, Kao T-h: The flanking regions of two *Petunia inflata S*-alleles are heterogeneous and contain repetitive sequences. Plant Mol Biol 18: 725–737 (1992).
6. Emerson S: A preliminary survey of the *Oenothera organensis* population. Evolution 24: 524–537 (1939).
7. Felsenstein J: PHYLIP (Phylogeny Inference Package). Distributed by the author. Department of Genetics, University of Washington, Seattle, WA (1996).

8. Franklin FCH, Franklin-Tong VE, Thorlby GJ, Howell EC, Atwal K, Lawrence MJ: Molecular basis of the incompatibility mechanism in *Papaver rhoeas* L. Plant Growth Regul 11: 5–12 (1992).
9. Hinata K, Watanabe M, Yamakawa S, Satta Y, Isogai A: Evolutionary aspects of the S-related genes of the *Brassica* self-incompatibility system: synonymous and nonsynonymous base substitutions. Genetics 140: 1099–1104 (1995).
10. Ioerger TR, Gohlke JR, Xu B, Kao T-h: Primary structural features of the self-incompatibility protein in Solanaceae. Sex Plant Reprod 4: 81–87 (1991).
11. Klein J, Satta Y, Takahata N, O'hUigin C: Trans-specific *Mhc* polymorphism and the origin of species in primates. J Med Primatol 22: 57–64 (1993).
12. Lawrence MJ, Lane MD, O'Donnell S, Franklin-Tong VE: The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. V. Cross-classification of the S-alleles of samples from three natural populations. Heredity 71: 581–590 (1993).
13. Lee HS, Huang SS, Kao T-H: S-proteins control rejection of self-incompatible pollen in *Petunia inflata*. Nature 367: 560–563 (1994).
14. Mantel N: Approaches to a health research occupancy problem. Biometrics 30: 355–362 (1974).
15. Matton DP, Maes O, Laublin G, Xike Q, Bertrand C, Morse D, Cappadocia M: Hypervariable domains of self-incompatibility RNases mediated allele-specific pollen recognition. Plant Cell 9: 1757–1766 (1997).
16. McClure BA, Haring V, Ebert PR, Anderson MA, Simpson RJ, Sakiyama F, Clarke AE: Style self-incompatibility gene products of *Nicotiana alata* are ribonucleases. Nature 21: 955–957 (1989).
17. O'Donnell S, Lane MD, Lawrence MJ: The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*.VI. Estimation of the overlap between the allelic complements of a pair of populations. Heredity 71: 591–595 (1993).
18. O'Donnell S, Lawrence MJ: 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 (1984).
19. Olmstead RG, Sweere JA: Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst Biol 43: 467–481 (1994).
20. Parry S, Newbigin E, Craik D, Nakamura KT, Bacic A, Oxley D: Structural analysis and molecular model of a self-incompatibility RNase from wild tomato. Plant Physiol 116: 463–469 (1998).
21. Paxman GJ: The maximum likelihood estimation of the number of self-sterility alleles in a population. Genetics 48: 1029–1032 (1963).
22. Richman AD, Kao T-h, Schaeffer SW, Uyenoyama MK: *S*-allele sequence diversity in natural populations of *Solanum carolinense* Horsenettle. Heredity 75: 405–415 (1995).
23. Richman AD, Kohn JR: Learning from rejection: the evolutionary biology of single-locus incompatibility. Trends Ecol Evol 11: 497–502 (1996).
24. Richman AD, Kohn JR: Self-incompatibility alleles in *Physalis*: implications for historical inference from balanced polymorphisms. Proc Natl Acad Sci USA 96: 168–172 (1999).
25. Richman AD: *S*-allele diversity in *Lycium andersonii*: implications for inter-specific variation in *S*-allele age in the Solanaceae. Ann Bot (in press).
26. Richman AD, Kohn JR: Significant differences in *S*-allele diversification in taxa arising after a bottleneck event. Heredity (submitted).
27. Richman AD, Uyenoyama MK, Kohn J: *S*-allele diversity in a natural population of ground cherry *Physalis crassifolia* (Solanaceae) assessed by RT-PCR. Heredity 76: 497–505 (1996).
28. Richman AD, Uyenoyama MK, Kohn JR: Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science 273: 1212–1216 (1996).
29. Saba-El-Leil MK, Rivard S, Morse D, Cappadocia M: The S11 and S13 self-incompatibility alleles in *Solanum chacoense* Bitt. are remarkably similar. Plant Mol Biol 24: 571–583 (1994).
30. Schierup M: The number of self-incompatibility alleles in a finite, subdivided population. Genetics 149: 1153–1162 (1998).
31. Singh A, Kao T-h: Gametophytic self-incompatibility: biochemical, molecular genetic, and evolutionary aspects. Int Rev Cytol 140: 449–483 (1992).
32. Swofford DL: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, MA (1998).
33. Takahata N: A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc Natl Acad Sci USA 87: 2419–2423 (1990).
34. Takahata N: Allelic genealogy and human evolution. Mol Biol Evol 10: 2–22 (1993).
35. Takahata N: Evolutionary genetics of human paleopopulations. In: Takahata, N, Clarke AG (eds), Mechanisms of Molecular Evolution, pp. 1–21. Sinauer Associates, Sunderland MA (1993).
36. Takahata N, Nei M: Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124: 967–978 (1990).
37. Takezaki N, Rzhetsky A, Nei M: Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12: 823–833 (1995).
38. Thompson RD, Kirsch H-H: The *S* locus of flowering plants: when self-rejection is self-interest. Trends Genet 8: 381–387 (1992).
39. Uyenoyama MK: Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics 147: 1389–1400 (1997).
40. Vekemans X, Slatkin M: Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics 137: 1157–1165 (1994).
41. Wright S: On the number of self-incompatibility alleles maintained in equilibrium by a given mutation rate in the population of a given size: a re-examination. Biometrics 16: 61–85 (1960).
42. Wright S: The distribution of self-incompatibility alleles in populations. Evolution 18: 609–619 (1965).
43. Wu J, Saupe SJ, Glass NL: Evidence for balancing selection operating at the *het-c* heterokaryon incompatibility locus in filamentous fungi. Proc Natl Acad Sci USA 95: 12398–12403 (1998).
44. Xue Y, Carpenter R, Dickinson HG, Coen ES: Origin of allelic diversity in *Antirrhinum* S locus RNases. Plant Cell 8: 805–814 (1996).
45. Yokoyama S, Hetherington LE: The expected number of self-incompatibility alleles in finite plant populations. Heredity 48: 299–303 (1982).