Roles of lineage sorting and phylogenetic relationship in the genetic diversity at the self-incompatibility locus of Solanaceae

YINGQING LU

Department of Biology, Duke University, Box 90338, Durham, NC 27708-0338, U.S.A.

Allelic polymorphism at the *S* locus that determines the gametophytic self-incompatibility (GSI) system in the pistil predates speciation. Understanding the evolution of a GSI system therefore requires knowledge of how lineage sorting and interspecific phylogenetic relationship affect *S* allele polymorphism. In searching for patterns of lineage sorting among species of various phylogenetic relationships, 22 *S*-alleles from 34 genets randomly taken at three Tennessee sites from a newly known GSI species *Physalis longifolia* were sequenced. Analyses of these data along with the previous sequences of three solanaceous species indicate that much of the combined allelic genealogy may be explained by lineage sorting and phylogenetic relationship. Using the mean terminal branch lengths of *trans*-specific alleles on the allelic genealogy to infer phylogenetic relationship among species, *P. longifolia* was found to be more closely related to *P. cinerascens* than to *P. crassifolia*. Nonetheless, the distribution of terminal branch lengths of *P. longifolia* was more similar to that of *P. crassifolia* than to that of *P. cinerascens*, suggesting phylogenetic relationship may have little effect on species-specific polymorphism. Similar habitat and growth characters, yet contrasting *S*-polymorphism, between *P. longifolia* and *P. cinerascens* also reject previous hypotheses that habitat and growth characters are the major factors responsible for interspecific differences in *S*-polymorphism. A likely scenario is that species-specific *S*-polymorphism is based on lineage sorting whose effect is further modified by species age and historical changes in population parameters.

Keywords: inferring phylogeny, lineage sorting, *Physalis longifolia*, *S* locus, *trans* alleles, *trans*-specific polymorphism.

Introduction

Extremely polymorphic loci often display *trans*-specific polymorphism, where alleles between species may be more similar to each other than alleles within species, such as the *S* locus governing pollination of many plant species (Ioerger *et al.*, 1990) and the major histocompatibility (MHC) loci in mammals (Figueroa *et al.*, 1988). *Trans*-specific polymorphism has been explained by lineage sorting (Avice *et al.*, 1990) referring to the random sorting of ancestral alleles into descendent species during speciation. Intuitively, lineage sorting should be the primary factor determining patterns of allelic distribution, especially for recent species, although thorough investigations on such effects have not been seen. The current work intends to fill this gap by focusing on lineage sorting at the *S* locus, with

implications for other multiallelic loci where allelic divergence predates speciation.

Alleles at the *S* locus are known to be responsible for the gametophytic self-incompatibility (GSI) reaction in the family Solanaceae. The expression of *S*-alleles in the pistil aborts self-pollen and selectively rejects cross-pollen carrying identical alleles to the pistil. *S*-alleles are under strong balancing selection because of frequency-dependent advantage and lack of homozygotes (Wright, 1939, 1960). Although the effect of balancing selection has been cited for *trans*-specific polymorphism (Ioerger *et al.*, 1990), nucleotide polymorphism among *S*-alleles (Clark, 1993; Vekeman & Slatkin, 1994) and fast approaching equal allele frequency at equilibrium (Wright, 1969; Nagylaki, 1975), surveys at the *S* locus conducted for natural populations of the Solanaceae (Richman *et al.*, 1995, 1996a; Richman & Kohn, 1999) revealed significant variations among species in allele number and average allelic diversity, which are not easily explained by balancing selection. For instance,

Correspondence. E-mail: verena@duke.edu

28 *S*-alleles from a natural population of *Physalis crassifolia* are on average more similar to each other than 13 alleles from *P. cinerascens*. This contrasting pattern of allelic genealogies was examined statistically under the null hypothesis of balancing selection (Uyenoyama, 1997), and significant deviations were identified for the allelic genealogies of *Solanum carolinense* (Uyenoyama, 1997) and *P. cinerascens* (Richman & Kohn, 1999). These results prompted the search for additional forces responsible for the interspecific differences in allelic polymorphism. Proposed candidates include population bottlenecks (Richman *et al.*, 1996b), rates of allele turnover (Uyenoyama, 1997), nonrandom assortment of *S*-alleles, life history characteristics and ephemeral habitat (Richman & Kohn, 1999). Here, two more factors, lineage sorting and phylogenetic relationships, are considered in relation to the interspecific differences.

Recent investigations of *S*-polymorphism at the population level in a third species of *Physalis*, *P. longifolia*, enables the assessment of the effects of phylogenetic relationships on *S*-polymorphism. Previous hypotheses concerning the roles of life history characters and habitat in shaping *S*-polymorphism may also be evaluated with the new data obtained in this study. When the phylogeny among *Physalis* congeners is known, observed *trans*-specific polymorphism induced by lineage sorting may be temporally examined, and lead to critical evaluation of the effect of lineage sorting on *S*-polymorphism. Analyses of this work depend on information provided by the alleles participating in *trans*-specific polymorphism, hence these alleles are referred to as *trans* alleles for simplicity. It will be shown that: (1) effects of lineage sorting on species-specific polymorphism decrease with species age; (2) species phylogeny may be inferred from *trans* alleles; (3) lineage sorting and species phylogeny may explain much of the distribution of *S*-alleles on an allele genealogy.

Materials and methods

Field collections

Physalis longifolia is a weedy perennial growing in disturbed habitats in the central and eastern United States. It is partially sympatric with *P. cinerascens* and possibly allopatric to *P. crassifolia*. The species reproduces both vegetatively via underground rhizomes and sexually through its GSI system. The GSI system was revealed by the microscopic observations made on pollen germination *in vivo* (unpublished data). Ninety-two collections were made in June 1997 from three sites, FH (Frosty's house), BF (Beefalo Farm), and VA (VA Hospital). All of those sites were in Johnson City area,

Washington county of Tennessee. Twelve additional rhizomes were collected from the BF site in November 1997. Plants were transplanted and maintained in a greenhouse of Duke University.

Amplifying, cloning and sequencing *S*-alleles

Floral buds were collected and frozen immediately. Total RNAs were extracted from styles using the Purescript RNA isolation kits (Gentra Systems, Inc.) for cDNA synthesis. cDNAs were synthesized as one strand with the cDNA cycle kit (Invitrogen) and oligo dT primer, and used as templates for RT-PCR reactions with two degenerate primers. The degenerate primers targeted C2 and C5 regions (*sensu* Ioerger *et al.*, 1991), as described in Richman *et al.* (1995). A single band of ~420 bp from the PCR amplification was cloned using the TA cloning kit (Invitrogen). Inserts were individually sequenced from 8 to 25 clones per plant on automated sequencers ABI 373 and ABI 377. More than four sequences per allele were obtained to eliminate PCR errors. Each allelic sequence was further verified using partial genomic DNAs and allele-specific primers. These sequences have been deposited in the GenBank (accession numbers AF281180–AF281201).

Greenhouse crosses

To confirm that the sequences identified are indeed responsible for mating behaviour in *P. longifolia*, four sets of crosses (~20 replicates per cross) were carried out in a greenhouse using parents of known *S*-genotypes. Harvested seeds from compatible crosses were randomly chosen and germinated for genotypic assays. The genotypes of the seedlings were determined by screening genomic DNAs using allele-specific primers. All single bands amplified by the primers were further sequenced.

Analyses

Allelic genealogies

Satisfactory alignment of *S*-sequences was obtained by visual inspection based on known conserved motifs (Ioerger *et al.*, 1991) and references on the secondary structure of *S*-alleles (Ishimizu *et al.*, 1998). The alignment was then translated into pairwise distances using Dayhoff's PAM matrix (Dayhoff *et al.*, 1978) by PHYLIP 3.572 (Felsenstein, 1998). The allele genealogy, with a *Malus* *S*-allele *S*7 (Broothaerts *et al.*, 1995) as an outgroup, was constructed using the Neighbour-Joining (NJ) algorithm, based on a principle of minimum evolution (Rzhetsky & Nei, 1992). The first genealogy was based on all 75 amino acid (aa) sequences available

so far from *Solanum carolinense* and three *Physalis* species, most of which are 135-aa long though several are only 57-aa long. The 76 *S*-genealogy is skewed by the long branches of nine *S*-alleles of *P. crassifolia* and two of *S. carolinense*. The sequences available for these 11 alleles are shorter than for the rest of the alleles but span the two hypervariable regions, thus severely inflating the pairwise distances involving these sequences. To circumvent the problem, a genealogy was constructed using 65 *S*-alleles of the same length (~135-aa). The resulting 65 *S*-genealogy provides the basis for the subsequent analyses.

Testing the molecular clock among lineages of different species

Alleles distributed among species may have experienced different rates of evolution due to species-specific changes. In order to search for evidence of recent lineage-specific rate heterogeneity, Tajima's (1993) chi-squared test for the molecular clock was applied to the data sets, which consisted of the three most closely related sequences. These data sets did not overlap on the allelic genealogy, and the tests were considered independent.

Models for understanding allelic number and trans-specific polymorphism

The current number of *S*-alleles for a given species (n_t) is the result of inherited *S*-alleles from the most recent common ancestor (MRCA) population (n_0), plus newly emerged *S*-alleles after speciation, minus loss of *S*-alleles since speciation after the t -th generation. Let the rate of allele loss at the *S* locus be ρ per allele per generation, and the rate of allele origination be μ per allele per generation, then,

$$n_t = n_0(1 + (\mu - \rho)t). \quad (1)$$

Obviously, the relative magnitude of μ and ρ determines whether the allele number in a species decreases or increases, and both μ and ρ are subject to changes of population parameters.

While allele numbers change with time within each species, the number of *trans* allele pairs between two species (k_0) generally decreases after speciation. This is because k_0 reduces each time when a new allele is derived from a *trans* allele. Random losses of *trans* alleles in both species further change the number of *trans* allelic pairs after t generations. Taking these factors into account, the current number of *trans* allelic pairs (k_t) becomes

$$k_t = k_0(1 - (\mu_A + \mu_B + \rho_A + \rho_B - \mu_A\mu_B - \rho_A\rho_B - \mu_A\mu_B\rho_A\rho_B)t), \quad (2)$$

where ρ_A and ρ_B are the rates of allele loss in species A and B, respectively, and μ_A and μ_B are rates of allele gain. Since rates of allele loss and allele origination are both expected to be small, eqn (2) can be approximated by

$$k_t = k_0(1 - (\mu_A + \mu_B + \rho_A + \rho_B)t). \quad (3)$$

If $\rho_A = \rho_B = \rho$, and $\mu_A = \mu_B = \mu$, then

$$k_t = k_0(1 - 2(\mu + \rho)t). \quad (4)$$

Since t is the generation time after speciation, eqns (2)–(4) indicate that the longer the time since speciation, the smaller the total number of *trans*-specific allele pairs between species. If ρ and μ are known, the time to observe zero *trans*-specific pairs (t^*) can be estimated by letting eqn 4 be zero. That is,

$$t^* = 1/(2\mu + 2\rho) \quad (5)$$

If $\rho \ll \mu$, about $1/(2\mu)$ generations are needed to observe no *trans* alleles in a genealogy, regardless of the initial number of alleles shared between species at speciation. Because of balancing selection operating at the *S* locus, the rate of allelic loss is extremely low, making t^* exceedingly long. Hence extensive *trans*-specific polymorphism among species of recent origin is observed. Lineage sorting into subsequent species described in eqns (1) and (3) shows that the time since speciation is a major factor determining the current degree of *trans*-specific polymorphism.

Estimating interspecific distances using trans-specific terminal branch lengths

Typical divergence among *S*-alleles implies that most of the alleles sorted between species during speciation are also divergent due to balancing selection. It follows that the *trans* alleles presently viewed within each species are the consequences of independent accumulations of mutations in the common ancestral alleles during the time separating the two species. Therefore, the terminal branches of a pair of *trans* alleles may be taken as a measure for the time distance between the two species (Fig. 1). If k pairs of *trans* alleles are observed between species A and B, and *trans* allele j in species A has a terminal branch of length A_j and the corresponding *trans* allele in species B has a terminal branch length of B_j , the phylogenetic distance between A and B ($d_{A,B}$) may be estimated as

$$\hat{d}_{A,B} = \sum_{j=1}^k \frac{A_j + B_j}{2k} \quad (6)$$

Fig. 1 A graphical illustration of the concept that *trans* alleles are mostly derived from the ancestral alleles at the time of speciation. The circles represent the ancestral *S*-alleles with lines showing their genealogy. X refers to allele extinction.

Two possible biases may inflate the estimation. One is at lineage sorting when two closely related ancestral alleles were both sampled into the species A and B, respectively. The other is under-sampling of alleles in current species creating *pseudo-trans* alleles. Ways of reducing the biases include increasing sample sizes and excluding outliers showing excessive terminal branch lengths.

Results

S alleles in *P. longifolia*

A total of 21 distinct *S*-genotypes containing 22 alleles were found among 34 genets from the Tennessee sites.

The genets were identified by 6-phosphogluconate dehydrogenase (unpublished data) and *S*-genotypes. The segregation study on eight randomly chosen alleles indicates that the alleles behave consistently with the GSI mode (Table 1). Sequenced *P. longifolia* *S*-alleles were about 380-nt long, covering nearly 60% of the coding regions of typical *S*-genes (Fig. 2). Moreover, these sequences span from C2 to C5 and include the previously identified hypervariable regions (Ioerger *et al.*, 1991).

Allele genealogies of the *S*-alleles from *Physalis* and *Solanum*

The 76-*S* allele genealogy represents all available *S*-alleles (75) from natural populations of four solanaceous species (Fig. 3). Two *P. longifolia* alleles (*S*20 and *S*28) and one (*S*1) *P. cinerascens* allele form a clade basal to previously defined *Physalis* *S*-allele clades. All other *P. longifolia* *S*-alleles congregate in the clades that have been previously identified for the genus *Physalis*. Most interestingly, *S*26 of *P. longifolia* and *S*3 of *P. cinerascens* are identical for the 255-nucleotide sites sequenced. Excluding the shorter sequences in the 76-*S* allele genealogy, the resulting 65 *S*-genealogy (Fig. 4) appears to have less skewed branches. In addition to the features described in the 76 *S*-tree, four *trans*-specific pairs (marked with *) were identified. The pattern of *trans* alleles initiated by lineage sorting is most pronounced in *P. cinerascens*. On the 65 *S*-tree, 10 of the 12 sampled *P. cinerascens* *S*-alleles (83%) show more similarity with *S*-alleles of different species than with its conspecific alleles. Twelve

Table 1 Segregation ratios of *S*-alleles among offspring of four crosses in *Physalis longifolia*. Each cross was carried out with the first genotype as maternal plant and the second as paternal plant

Cross	Offspring genotype	Offspring number	Haplotype	Observed	Expected	Chi-test statistic	<i>P</i> -value
<i>S</i> 13 <i>S</i> 25 × <i>S</i> 01 <i>S</i> 02	<i>S</i> 01 <i>S</i> 13	8	<i>S</i> 01	20	21	0.992	> 0.5
	<i>S</i> 02 <i>S</i> 13	13	<i>S</i> 02	22	21		
	<i>S</i> 01 <i>S</i> 25	12	<i>S</i> 13	21	21		
	<i>S</i> 02 <i>S</i> 25	9	<i>S</i> 25	21	21		
<i>S</i> 01 <i>S</i> 02 × <i>S</i> 01 <i>S</i> 28	<i>S</i> 01 <i>S</i> 28	9	<i>S</i> 01	9	9	1	> 0.5
	<i>S</i> 02 <i>S</i> 28	9	<i>S</i> 02	9	9		
			<i>S</i> 28	18	18		
<i>S</i> 02 <i>S</i> 03 × <i>S</i> 01 <i>S</i> 28	<i>S</i> 01 <i>S</i> 02	6	<i>S</i> 01	20	19.5	0.881	> 0.5
	<i>S</i> 01 <i>S</i> 03	14	<i>S</i> 02	17	19.5		
	<i>S</i> 02 <i>S</i> 28	11	<i>S</i> 03	22	19.5		
	<i>S</i> 03 <i>S</i> 28	8	<i>S</i> 28	19	19.5		
<i>S</i> 01 <i>S</i> 28 × <i>S</i> 04 <i>S</i> 15	<i>S</i> 01 <i>S</i> 04	5	<i>S</i> 01	12	11	0.948	> 0.5
	<i>S</i> 01 <i>S</i> 15	7	<i>S</i> 04	10	11		
	<i>S</i> 04 <i>S</i> 28	5	<i>S</i> 15	12	11		
	<i>S</i> 15 <i>S</i> 28	5	<i>S</i> 28	10	11		

	1	65
PlgS01	DNSEKQGLLVGCEPPPNYTNFKD-KMFDDLDKHWIQLLLKDKVGKIEQRIWKYQYKRHGSCCRNL	
PlgS02	EKKGID-QLTFCAQSNYTLFQDK-KMLDELDKHWIQLQYSQEIQLQRQESWKRQYKHKGSCCCLNR	
PlgS03	DNSEQVGLLVDCSTSTNYTNLKD-KMLDDLDTHWTQLQDKETGLAQQRIRWKYQYKRHGSCCLEL	
PlgS04	EKRQGN-KMVSCLEPVKYTLFEDTKMLDVLDKHWIQLKVKKDEIPVKQELWKNQYKHKGACQEV	
PlgS06	DNISK--PINFCHWNDYDRKLLNDYKKNKLYISWPDLFADAEADCKNAQKFWFQYKHKGTCSLDS	
PlgS07	DKNNS--LLMDCKPSPKYTKFPRNKMFADLDKHWIQLKIEDDAETNQSTWSNQYIKHGSCCRNL	
PlgS08	DKNNS--LLMDCKPTPNYTNFPRNKMFDDLDKHWIQLKIEGAAETNQSTWSYQYKHKGSCCCLNR	
PlgS09	DNIST--TLNKCIPVYAKNMTD-DKKNKLYIRWPDLLVGEANCKDKQKFWKNEYKHKGSCEKS	
PlgS12	EKRGRK-MLVSCFPQVNYTLFEDMKMLDILDKHWIQLKVLKNESLVQQLWKNQYKHKGACQEV	
PlgS13	DKNDS--LLMDCKPPPNYTNFQD-KMFDELDKHWIQLKIFENQSKINQSTWSYQYKHKGSCCRNL	
PlgS14	EKKGVD-KLTFCAQPNYTLFKDKMLDDLDKHWIQLMYSKEKGLQKQEFWKRQYKHKGTCCCLNR	
PlgS15	DKNNS--LLMDCKPPPNYTNFHN-KMFDDLDKHWIQLKIPKQRAKIDQSTWSYQYKHKGSCCRNL	
PlgS16	EKRKK-MLVSCLDHSYTLFEDR-KMLDDLDKHWIQLKIFENQSKINQSTWSYQYKHKGSCCRNL	
PlgS17	DKAGM--TLKDCSTAPNYTKFKD-KMLDDLDTHWTQLTLTKTGLAEQSTWNYQFTKHKSCCHEL	
PlgS18	DNISK--TINFCRRNDYDSSLMDKYKKNKLYISWPDLFAGEAECKNDQKFWLQYKHKGTCCVDR	
PlgS20	DNKYV--RLDCKPNAVINTE-GDKAILDKRWPQLDITVEYGLTKQFPFWKQYKHKGTCCCLNR	
PlgS22	DNFNT--TLNPCSERGYEKNKKNYTKDMLYIHWPDLLVVEAECKNYQEFWIRQYKHKGTCCCTS	
PlgS23	DKAGV--TLMDCLPIPNYTNFKD-KMLDDLDTHWTQLRLTKEIGLAEQRIWNYQFTKHKGSCCQEL	
PlgS25	????K-IIVSCLPKVNYTLFQDRKMLDDLDKHWIQLKVSKNELGQQQEAQWFKYKHKGTCCCKNS	
PlgS26	EEKGRN-KLKFCPLPNYTRFEDKKMLDDLDKHWIQLKQKDSQDGLQKQDFWKRQYKHKGSCCRNL	
PlgS27	DNFSA--MLNFCANKYDKAILTDYKKNKLYIHWPDLVVDEAKCKNDQEFWFKYKHKGTCCCKNS	
PlgS28	DTGR---RLYICIPNVYFLKIEQGEKSDILDKSWPQLEYNKTDGLDKQAFWKRQYKHKGTCCCKNS	
	66	130
PlgS01	-YNQNMYSFSLALHLKDKVDLKRNLGKQNIIPGGNYTFAQIIKAVKTVSK-SEPNIKCIIKFKGPLE	
PlgS02	-YNQTPYFDLALRLKDRIGLLSTLHTSGIDPGENYTFKQIVKARTATN-ADFLFKCVK-GTTQE	
PlgS03	-YNQRMYSFSLALHLKDKVDLKRNLGKQNIIPGGNYTFAEIIKAVKIVSK-SEPNIKCIIK--GPQE	
PlgS04	-YDQNMYSFSLALRLYERFDLLSLALQNHISVPGENYTIQEISKAIKNVTK-AGSDIKCVK--GKPE	
PlgS06	-YKQEYFDLAMDLDKDFLLQSFKSHGIIIRGGDYTLKSINDTVRAIT-T-GVPSLWCTA---QMQ	
PlgS07	-YNQNMYSFSLALHLKNRVNLTLNLITQKIFPPGRHYTFDEIADAVKTAQ-SVPNIKCMQ--RPQE	
PlgS08	-YNQNMYSFSLALHLKNRVNLTLNLITQKIFPPGGQYTLDEIAKAVKAALK-SVPNIKCIKVKGPQE	
PlgS09	-YNQEYFDLAMDGLDKDFLLSLSLGRYGIIPGKSYPVQNTINNTVKAITQ-GFPRLACTD---LME	
PlgS12	-YNQNMYSFSLALRLYERFDLLSLKNSHISVPGKYTIQELSKAIKTVTI-SDSDIRCDE--RTHE	
PlgS13	-YNQNMYSFSLALRLKDRFDLLRLNLNTEKIFPPGGQYTLDEIAEAVKTVLK-SVPNIKCVQ--RPQE	
PlgS14	-YNQTAYFSLALHLKEKIDLLSLHNSGIDPGENYTFQEIKAIKTVTN-ADSLFKCVK-RTTQE	
PlgS15	YNQNMYSFSLALHLKNRVNLTLNLITKIIYPPGGQYTLDEIAKAVKAALK-SVPNIKCIKVKGPQE	
PlgS16	-YNQSTYFSLALRLYERFDLLSLKNSHISVPGENYTIQEIAKAIKTVINADSDIKCSI--RTRE	
PlgS17	-YDQSMYSFSLALGLAKNLTLTLKSKHIFPGRKYTVDEISTAVKTVST-SQPKIKCIK--GPQE	
PlgS18	-NNQELYFDLAMDGLDKDFLLQSFKSYGIIRGGDYTLKSINDTVRAITR-GVPSLWCAA---QME	
PlgS20	-YTQEAYFNLAIKLDRFDILKALGDQGIIPGNKYVVKVKNIIISRVTL-QPPKLMCTEASQVLE	
PlgS22	-YNQEYFDLAMDVLKDKFNLLQSFNYGIIIRGRYTLKEINDTIRAITR-GYPSLSCTQ--RME	
PlgS23	-YNQSMYSFSLALDLKAKVNLTLTLKSKHIFPGRKYTVDEISTAVKTVST-SQPKIKCIK--GPQE	
PlgS25	-YNQNMYSFSLALRLYERFDLLSLKNSHISVPGENYTIQEISKAIKTVTI-SDSDIRCAB--GTQE	
PlgS26	-YNQTAYFNLAIRLKEKIDLLKTLKQGIIVPGENYTFYFIASAVKTVTL-ADSFVKCIE--GTQE	
PlgS27	-YSQEYFDLAMDVLKDKFLLQSFKRYGIIPGTSHTVQTINNTVKAITH-GFPNLLCTE---RME	
PlgS28	-YTQEAYFNLAIKLDRFDVLKILGDQGIIPGNKYAFQKVHDAIKKVIIR-RPPKLMCTEASQVLE	

Fig. 2 Alignment of 130 amino acids of 22 *P. longifolia* *S*-alleles. Conserved residues are indicated in bold, gaps by '-', and unknowns by '?'.
PlgS01
PlgS02
PlgS03
PlgS04
PlgS06
PlgS07
PlgS08
PlgS09
PlgS12
PlgS13
PlgS14
PlgS15
PlgS16
PlgS17
PlgS18
PlgS20
PlgS22
PlgS23
PlgS25
PlgS26
PlgS27
PlgS28

out of 22 *P. longifolia* *S*-alleles (55%) exhibit the same pattern, and nine out of 19 *P. crassifolia* *S*-alleles (47%) display *trans*-specific polymorphisms.

Species-specific polymorphism

Species-specific NJ trees were also constructed for each species (not shown) using *S*-allele amino acid sequences. The average length of terminal branches of *P. longifolia* is significantly different from those of previously surveyed species, but is closer to that of *P. crassifolia* than to those of *P. cinerascens* and *S. carolinense* (Table 2).

No statistical difference is found when the terminal branch length distribution of *P. longifolia* is compared to that of *P. crassifolia* (Kolmogorov–Smirnov two-sample test, $P > 0.1$, two-tailed), indicating that the different mean terminal branch lengths (Table 2) are possibly caused by the shift of means rather than the overall distributions. However, the distribution of terminal branch lengths in *P. longifolia* differs significantly

from that of *P. cinerascens* ($P = 0.025$), as well as from that of *S. carolinense* ($P = 0.01$), using the Kolmogorov–Smirnov two-sample tests (two-tailed).

Because *S26* of *P. longifolia* is the same as *S3* of *P. cinerascens* among all the sites sequenced, introgression between the two species was considered, however, it seems unlikely, since all reciprocal crosses on 26 flowers (involving two *P. cinerascens* individuals and four *P. longifolia* genets) failed in the greenhouse.

Tests of the molecular clock among lineages

Tajima's chi-squared statistics detected significant departures from a molecular clock for the lineages leading to *S4* and *S9* in *P. cinerascens* (Table 3). Excessive (~7-fold) substitutions were found in sequences *S4* and *S9* in comparison with their sister alleles in *P. crassifolia*, suggesting higher rates of evolution along the two terminal lineages of *P. cinerascens*. No significant departure from a molecular clock was found among

0.1000

Fig. 3 The Neighbour-Joining tree based on the amino acids of 75 *Physalis* *S*-sequences, including the 22 new *S*-alleles from *P. longifolia* and 53 previously published sequences (accession numbers L40539–L40551: Richman *et al.*, 1995; accession numbers L46653–L46680: Richman *et al.*, 1996a; accession numbers AF058930–AF058941: Richman & Kohn, 1999). The outgroup is *Malus* S7 (accession number U19792: Broothaerts *et al.*, 1995). Amino acid distances were calculated using PHYLIP 3.572's PROTDIST.

other terminal branches of *P. longifolia* and *P. crassifolia*.

Interspecific relationship based on *trans* alleles

The total substitutions between *trans* alleles for estimating within-species evolution at the *S* locus are listed (Table 4). Excluding the marked outliers, an average per 100 bases of about 1.5 substitutions separates *P. cinerascens* and *P. longifolia*, and about 5–6 substitutions separate *P. crassifolia* from *P. cinerascens* and *P. longifolia*, respectively. Nearly 70–80% of all these changes are due to replacement substitutions. Some of the changes were incorporated in the terminal branch lengths of *trans* alleles identified by the 65 *S*-genealogy, and included in the interspecific distances (Table 5) computed by eqn (6). The distances from *P. crassifolia* to *P. cinerascens* and *P. longifolia* are about twice the distance between *P. longifolia* and *P. cinerascens* (Fig. 5), and the differences are significant with Wilcoxon rank sum tests (Table 5).

Discussion

Inference of species phylogeny from allelic genealogy

Knowing the phylogeny of the species whose *S*-allelic polymorphisms have been documented helps to put the *S*-polymorphism in a historical perspective. While an inclusive phylogeny of *Physalis* has yet to be achieved (Sudhakaran & Ganapathi, 1993), deep coalescence at the *S* locus and the existence of *trans* alleles on an allelic genealogy enable the inference of phylogenetic relationships among species. This mean-based inference method can be applied to many loci characterized by *trans* alleles and balancing selection.

Adopting the average distance measured in both species (eqn 6) in inferring phylogeny helps reducing bias when a molecular clock does not hold. In addition, inspecting the immediate ancestors of *trans* alleles on the genealogy may reduce overestimation of interspecific distances. Outliners among *trans* alleles should be

excluded as in the cases of plg22–pci11 and plg3–pc15 (Table 4), not only because two closely related ancestral alleles could be occasionally sampled but also because pseudo-*trans* alleles could be generated by undersampling. In this study, none of the ancestral lineages among the three *Physalis* congeners was closely related on the genealogy (Fig. 4). Despite *P. cinerascens* S4 and S9 alleles exhibiting higher substitution rates relative to their corresponding *trans* alleles in *P. crassifolia* (Table 3), they imposed a negligible effect on the inference of phylogenetic relationship (Table 4).

Effects of lineage sorting and species phylogeny on allelic polymorphism

The simple models in this analyses (eqns 1 and 2) appear to have captured the major features of lineage sorting. The parameters of a lineage-sorting event are the number of alleles sampled in a species and the relationship among these alleles. After the initial lineage sorting that provided the basis of within-species polymorphism at the *S* locus, species age and rates of allele loss and gain are critical factors in determining allele number within species and the extent of shared polymorphism between species. An incipient species is expected to show inherited alleles from its most recent ancestor, since new alleles are yet to emerge within the species. Clearly, with the increasing age of a species, the signature left by lineage sorting in the species, in terms of total ancestral alleles and their relationship, may be modified substantially by frequent lineage losses or allele turnover. Fortunately, balancing selection at the *S* locus has prevented both events from happening frequently, for at least closely related species.

Since a phylogenetic relationship defines the allele space where lineage sorting operates, it is not surprising to see that lineage sorting and phylogenetic relationship jointly influence the current distribution of *S*-alleles among species. In this study, all *P. longifolia* *S*-alleles fall nicely into the clades that are characteristic of the genus (Fig. 4). However, sharing clades in genealogy does not necessarily lead to sharing ancestral alleles among species. Due to the more or less random nature of lineage sorting during speciation, most species are expected to show somewhat different patterns of allelic polymorphism, as have been seen among congeners of *Physalis*. The lack of *trans* alleles observed between *Solanum* and *Physalis* indicates that the time separating the two taxa is at least $0.5\mu^{-1}$ generations as suggested by eqn 5.

More detailed analyses of lineage sorting require reliable methods to be developed to estimate species age. Phylogenetic relationship among sampled species is often not a good measure of species age, because species

Fig. 4 The Neighbour-Joining tree of 65 *S*-alleles whose sequences are of similar lengths (~135 amino acids). *Trans* alleles with * are those not found in the 76 *S*-genealogy.

Table 2 Comparison of terminal branch lengths among species-specific Neighbour-Joining trees of four solanaceous species. Number in parentheses is the sample size. The second mean was obtained without short sequences. Significant mean differences were found with the Wilcoxon two-sample rank sum tests on samples with the same superscripts (^aone-tailed, $P = 0.02$; ^bone-tailed, $P = 0.009$)

Length of terminal branch	<i>P. crassifolia</i>	<i>P. longifolia</i>	<i>P. cinerascens</i>	<i>S. carolinense</i>
Mean	0.101 ^a (28) 0.108 (19)	0.161 ^{a,b} (22)	0.270 ^b (12)	0.263 (13) 0.348 (11)
SE	0.042 0.017	0.018	0.044	0.061 0.056

Table 3 Results of Tajima's molecular clock tests on interspecific lineages of *P. longifolia* (Plg), *P. cinerascens* (Pci) and *P. crassifolia* (Pcr). Test statistics follow the chi-squared distribution with 1 degree of freedom

Outgroup	Sister alleles	Chi-squared	P-value
PcrS20	<i>PciS11–PlgS22</i>	1.000	NS
PcrS17	<i>PciS06–PlgS13</i>	0.333	NS
PcrS11	<i>PciS07–PlgS07</i>	0.818	NS
PcrS01	<i>PciS03–PlgS26</i>	0.000	NS
PlgS28	<i>PlgS20–PciS01</i>	0.000	NS
PlgS03	<i>PlgS01–PcrS16</i>	1.250	NS
PlgS02	<i>PlgS14–PcrS03</i>	1.143	NS
PciS02	<i>PcrS01–PcrS13</i>	0.818	NS
PlgS07	<i>PcrS07–PciS04</i>	9.000	**
PcrS28	<i>PcrS04–PciS09</i>	9.000	**

age is the time from the most recent speciation while a phylogeny only reflects relative time distances among sampled species which may or may not include the most recently differentiated species-pair(s).

Impact of speciation on S-allelic polymorphism

One striking result of this study is that allele S26 in *P. longifolia* is identical to allele S3 in *P. cinerascens* as reported by Richman & Kohn (1999) along a 255-nucleotide long sequence. Although this is consistent with the close phylogenetic relationship between the two species detected in this study, additional explanatory factors may lie in lack of selection pressure on both alleles to differentiate within each species in the presumably short period of time that has separated the two species. Modern gene introgression between *P. longifolia* and *P. cinerascens* appears unlikely since the two species are well defined (Sullivan, 1984, 1986) and the reciprocal crosses between the two species yielded no hybrids in this study.

A derived observation is that once an allele is sorted into two descendent species, it becomes two alleles, each within its own species boundary. From the view of the overall S-genealogy in the Solanaceae, the process of speciation itself helps generating new alleles and creating coalescences among *trans*-alleles. If speciation is more frequent in one taxon than in another, disproportionate amounts of alleles may be observed between the two groups.

Species-specific polymorphism at the S locus

P. longifolia shares many characteristics of life history with those of *P. cinerascens* and *S. carolinense*, such as underground rhizomes, a perennial growth habit, and even the same disturbed habitat. However, a richer allele number and substantially shorter terminal branches are found in *P. longifolia* than in *S. carolinense* and *P. cinerascens* (Table 2). The overall distribution of terminal branches of *P. longifolia* is indistinguishable from that of *P. crassifolia*, which has a shrubby growth habit and lives in a relatively undisturbed desert. These comparisons argue against a recent proposition (Richman & Kohn, 1999) that growth habit and habitat are among the major factors that determine species-specific allele number and terminal branch length.

Although species phylogeny and lineage sorting do not satisfactorily explain why alleles in *P. longifolia* are closer to *P. crassifolia* than to *P. cinerascens*, faster rates of substitution detected in *P. cinerascens* (Table 3) hint at a strong influence of species-specific event(s) which may have significantly modified the species-specific patterns of allelic polymorphism. The rate variation itself may be due to intensified selection for advantageous substitutions (Kimura & Ohta, 1973) or rapid expansion(s) in population size (Tajima, 1989) that these alleles have experienced. In these scenarios, species age and drastic changes in population parameters would profoundly alter the S-polymorphism at the species level.

Table 4 Substitutions (synonymous (syn.) and nonsynonymous (nonsyn.)) between *trans* alleles of three *Physalis* species. The *trans*-specific relationships were identified from *S*-allelic genealogies constructed by amino acid sequences via the neighbour-joining method. The pairs found specific to an allelic genealogy are marked with superscripts: a for 76 *S*-genealogy using sequences of varying lengths, b when using 65 sequences 135 amino-acid long; c when using 76 *S*-sequences 57 amino-acid long. Allelic pairs showing excessively high numbers of substitutions (defined as $> \text{mean} + 3 \text{ SE}$) are indicated with *, and were excluded from the calculation of means and standard errors

Species-pair	Allelic pair	Nonsyn.	Syn.	Bases compared	Substitutions/100 bases
<i>P. longifolia</i> – <i>P. cinerascens</i>	<i>Plg20</i> – <i>Pci01</i>	3	1	255	1.57
	<i>Plg22</i> – <i>Pci11</i>	12	3	255	5.88*
	<i>Plg07</i> – <i>Pci07</i>	7	2	255	3.53
	<i>Plg13</i> – <i>Pci06</i>	2	0	255	0.78
	<i>Plg26</i> – <i>Pci03</i>	0	0	255	0.00
	Mean	3	0.75		1.47
	SE	1.5	0.5		0.76
<i>P. longifolia</i> – <i>P. crassifolia</i>	<i>Plg06</i> – <i>Pc23</i> ^a	7	1	168	4.76
	<i>Plg01</i> – <i>Pc16</i>	17	6	384	5.99
	<i>Plg03</i> – <i>Pc15</i> ^a	15	4	171	11.11*
	<i>Plg23</i> – <i>Pc12</i> ^a	11	2	165	7.88
	<i>Plg14</i> – <i>Pc03</i>	10	5	381	3.94
	<i>Plg04</i> – <i>Pc10</i> ^b	18	6	378	6.35
	<i>Plg27</i> – <i>Pc20</i> ^b	12	6	324	5.56
	<i>Plg09</i> – <i>Pc21</i> ^b	29	2	369	8.40
	<i>Plg15</i> – <i>Pc26</i> ^c	6	4	157	6.37
	Mean	13.8	4.0		6.16
	SE	2.6	0.6		0.52
<i>P. cinerascens</i> – <i>P. crassifolia</i>	<i>Pci10</i> – <i>Pc21</i>	17	5	255	8.63
	<i>Pci09</i> – <i>Pc04</i>	6	3	255	3.53
	<i>Pci04</i> – <i>Pc07</i> ^b	8	1	255	3.53
	<i>Pci05</i> – <i>Pc09</i> ^b	17	4	255	8.24
	<i>Pci03</i> – <i>Pc24</i> ^c	4	0	162	2.47
	<i>Pci12</i> – <i>Pc22</i> ^c	3	0	161	1.86
	Mean	9.2	2.2		4.71
	SE	2.6	0.9		1.21

Table 5 Estimation of interspecific distances using mean terminal branch lengths of *trans*-specific sister *S*-alleles among three *Physalis* species based on the 65S tree. *Trans* alleles used in the inferences are *Plg20*–*Pci1*, *Plg22*–*Pci7*, *Plg7*–*Pci7*, *Plg13*–*Pci6* and *Plg26*–*Pci3* between *P. longifolia* and *P. cinerascens*; *Pci9*–*Pcr4*, *Pci4*–*Pcr7*, *Pci5*–*Pcr9* between *P. crassifolia* and *P. cinerascens*; *Plg4*–*Pcr10*, *Plg14*–*Pcr3*, *Plg1*–*Pcr16*, *Plg27*–*Pcr20* and *Plg9*–*Pcr21* between *P. longifolia* and *P. crassifolia*. Numbers in parentheses are the standard errors for the means. Wilcoxon two-sample tests were performed for samples of the same superscript: superscript a, one-tailed, $P = 0.016$; superscript b, one-tailed, $P = 0.018$

From/to	<i>P. crassifolia</i>	<i>P. cinerascens</i>	<i>P. longifolia</i>
<i>P. crassifolia</i>	—	0.052 (0.017) ^b	0.063 (0.012) ^a
<i>P. cinerascens</i>	0.076 (0.010)	—	0.037 (0.008) ^a
<i>P. longifolia</i>	0.077 (0.015)	0.032 (0.011) ^b	—
Interspecific distance			
	<i>P. crassifolia</i>	<i>P. cinerascens</i>	
<i>P. cinerascens</i>	0.064		
<i>P. longifolia</i>	0.070	0.035

Fig. 5 Estimated species phylogeny of three *Physalis* taxa using *Solanum carolinense* as an outgroup. The branch lengths within *Physalis* were estimated using *trans*-alleles.

Acknowledgements

I thank two anonymous reviewers, the designated editor, M. K. Uyenoyama for helpful comments and criticism, M. K. Uyenoyama for help in logistics, F. Levy, E. Walker for kindly locating field sites and providing field help, J. Stone, Z. Yang for the field collections, D. Liu for help in screening the offspring of the greenhouse crosses, and the greenhouse staff of the Duke University for assisting in plant maintenance. This work was supported by a NSF/Sloan postdoctoral fellowship in Molecular Evolution.

References

- AVISE, J. C., ANKNEY, C. D. AND NELSON, W. S. 1990. Mitochondrial gene trees and the evolutionary relationship of mallard and black ducks. *Evolution*, **44**, 1109–1119.
- BROOThAERTS, W., JANSSENS, G. A., PROOST, P., BROEKAERT, W. F. 1995. cDNA cloning and molecular analysis of two self-incompatibility alleles from apple. *Plant Mol. Biol.*, **27**, 499–511.
- CLARK, A. G. 1993. Evolutionary inferences from molecular characterization of self-incompatibility alleles. In: Takahata, N. and Clark, A. G. (eds) *Mechanisms of Molecular Evolution*, pp. 79–108. Sinauer, Sunderland, MA.
- DAYHOFF, M. O., SCHWARTZ, R. M. AND ORCUTT, B. C. 1978. A model of evolutionary change in proteins. In: Dayhoff, M. O. (ed.) *Atlas of Protein Sequence and Structure*, vol. 5, (suppl. 3), pp. 345–352. National Biomedical Research Foundation, Washington DC.
- FELSENSTEIN, J. 1998. PHYLIP 3.572 (*Phylogeny inference package*). Department of Genetics, University of Washington, Seattle.
- FIGUEROA, F., GUNTHER, E. AND KLEIN, J. 1988. MHC polymorphism predating speciation. *Nature*, **335**, 265–267.
- IOERGER, T. R., CLARK, A. G. AND KAO, T.-H. 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 9732–9735.
- IOERGER, T. R., GOHLKE, J. R., XU, B. AND KAO, T.-H. 1991. Primary structural features of the self-incompatibility protein in Solanaceae. *Sex. Plant Reprod.*, **4**, 81–87.
- ISHIMIZU, T., ENDO, T., YAMAGUCHI, K. Y., NAKAMURA *ET AL.* 1998. Identification of regions in which positive selection may operate in S-RNase of Rosaceae: Implication for S-allele-specific recognition sites in S-RNase. *FEBS Letters*, **440**, 337–342.
- KIMURA, M. AND OHTA, T. 1973. Mutation and evolution at the molecular level. *Genetics*, **73** (supplement), 19–35.
- NAGYLAKI, T. 1975. The deterministic behavior of self-incompatibility alleles. *Genetics*, **79**, 545–550.
- RICHMAN, A. D. AND KOHN, J. R. 1999. Self-incompatibility alleles from *Physalis*: Implication for historical inference from balanced genetic polymorphisms. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 168–172.
- RICHMAN, A. D., KAO, T.-H., SCHAEFFER, S. W. AND UYENOYAMA, M. K. 1995. S-allele sequence diversity in natural populations of *Solanum carolinense* Horsenettle. *Heredity*, **75**, 405–415.
- RICHMAN, A. D., UYENOYAMA, M. K. AND KOHN, J. R. 1996a. S-allele diversity in a natural population of ground cherry *Physalis crassifolia* (Solanaceae) assessed by RT-PCR. *Heredity*, **76**, 497–505.
- RICHMAN, A. D., UYENOYAMA, M. K. AND KOHN, J. R. 1996b. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. *Science*, **273**, 1212–1216.
- RZHETSKY, A. AND NEI, M. 1992. A simple method of estimating and testing minimum-evolution. *Mol. Biol. Evol.*, **9**, 945–967.
- SUDHAKARAN, S. AND GANAPATHI, A. 1993. Structure and distribution of plant trichomes in relation to taxonomy. *Feddes Repertorium*, **104**, 469–474.
- SULLIVAN, J. R. 1984. Pollination biology of *Physalis viscosa* var. *cinerascens* (Solanaceae). *Am. J. Bot.*, **71**, 815–820.
- SULLIVAN, J. R. 1986. Systematics of the *Physalis viscosa* complex (Solanaceae). *Syst. Bot.*, **10**, 426–444.
- TAJIMA, F. 1989. The effect of change in population size on DNA polymorphism. *Genetics*, **123**, 597–601.
- TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, **135**, 599–607.
- UYENOYAMA, M. K. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics*, **147**, 1389–1400.
- VEKEMAN, X. AND SLATKIN, M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics*, **137**, 1157–1165.
- WRIGHT, S. 1939. The distribution of self-fertility alleles in populations. *Genetics*, **24**, 538–552.
- WRIGHT, S. 1960. On the number of self-incompatibility alleles maintained in equilibrium by a given mutation rate in a population of a given size: a re-examination. *Biometrics*, **16**, 61–85.
- WRIGHT, S. 1969. *Evolution and the Genetics of Populations*, vol. 2, *The Theory of Gene Frequencies*. University of Chicago Press.