Learning from rejection: the evolutionary biology of single-locus incompatibility

Adam D. Richman and Joshua R. Kohn

Species in a wide array of flowering plant families prevent self-fertilization by physiological rejection of self-pollen. The genetic basis of rejection of self pollen is best understood in single-locus homomorphic gametophytic and sporophytic self-incompatibility systems. Under gametophytic self-incompatibility (GSI), if the S-allele carried by the pollen grain matches either allele of the ovule parent, the pollen is rejected. Thus, matings fall into three categories: (1) compatible (no alleles shared between maternal and paternal parents); (2) semi-compatible (one allele shared, half of the pollen rejected); and (3) incompatible (two alleles shared, all pollen rejected). In sporophytic self-incompatibility (SSI), the genotype of the pollen parent (rather than that of the haploid pollen grain) determines whether or not pollen will be rejected, and dominance relationships among alleles play a role in determining mating type. Alleles at the self-incompatibility (S-) locus are less frequently rejected when rare, leading to diversifying selection and some of the highest levels of allelic polymorphism known for any locus. Many other self-incompatibility (SI) mechanisms are found in flowering plants, including heteromorphic incompatibility (distyly and tristyly) and homomorphic systems controlled by two or more loci. However, the conceptual simplicity of single locus SI, coupled with recent advances in understanding its molecular basis, make the S-locus particularly amenable to population and evolutionary genetic study.

Polymorphism at the S-locus

Two empirical findings have been pivotal in generating theoretical inquiry into the dynamics of the S-locus. The first was the discovery that natural populations contain very large numbers of S-alleles. Emerson1 found 45 different S-alleles in GSI Oenothera organensis, an endemic species that he estimated to comprise only 500 individuals, raising the question of how this remarkable diversity is maintained in a small, isolated population. Wright2 modeled frequency dependent selection (FDS) at a gametophytic S-locus and determined the number of S-alleles maintained in a finite population under the balance of drift, mutation and selection. For Oe. organensis, Wright and later workers3 were forced to conclude that population numbers had recently declined from sizes between 2000 and 10 000. More recent population estimates of Oe. organensis place the number of individuals near 5000 (Ref. 4), a remarkable validation of theoretical prediction. Subsequent studies in several taxa have recovered similarly large estimates of the

Reference:
Box 1. The number of S-alleles and effective population size

Given the number of gametophytic self-incompatibility (GSI) alleles, \( n \), in a sample of diploid individuals, the number of alleles in the population, \( N \), can be found by iteratively solving Pawitan’s \( n^2 \)

\[
N = \frac{n(1 - (1 - 2/n)^{n-1})}{1 - (1 - 2/n)^{n-1}}
\]

This estimator assumes that genotype frequencies are equal, as expected under GSI. The assumption that the frequency distribution of alleles in the sample is drawn from a uniform distribution can be tested using Mantel’s \( n^2 \) test:

\[
\chi^2 = \frac{\sum (O - E)^2}{E}
\]

where \( O \) is the number of times an allele occurs, \( E \) is the number of alleles found, and \( r \) is the number of genotypes sampled. Alternative estimators for the number of alleles in a population are available if allele frequencies are not uniform. Substantial deviations have been detected only in Papaver rhaes37.

The population size required to maintain a specified number of S-alleles is a solution of:

\[
N_s = \int_{1/2N_0}^{1} \frac{dN}{dx} dx
\]

where \( dN/dx \) is the number of alleles whose frequencies fall in interval from \( x \) to \( y \), and where:

\[
\delta N = 4N_0\mu N_s(1 - q)\delta V x \times 1
\]

and \( V \) is the effective population size, \( N_0 \), for values of the origin rate \( V \) of new alleles from 10\(^{-4}\) to 10\(^{-9}\) gene\(^{-1}\) generation\(^{-1}\), is shown below. Within the expected range of origin rates \( V = 10^{-10} \) (see text) estimates of \( N_0 \) are relatively insensitive to the value of \( V \).

\[
N_s = \frac{1}{2N_0} \left[ \frac{1}{(1 - 2N_0)} \right]^{1/2}
\]

The relationship between the number of alleles maintained in a population \( N \) and the effective population size \( N_s \), for values of the origin rate \( V \) of new alleles from 10\(^{-4}\) to 10\(^{-9}\) gene\(^{-1}\) generation\(^{-1}\), is shown below. Within the expected range of origin rates \( V = 10^{-10} \) (see text) estimates of \( N_0 \) are relatively insensitive to the value of \( V \).

\[
N_s = \frac{1}{2N_0} \left[ \frac{1}{(1 - 2N_0)} \right]^{1/2}
\]

\[
N_s = \frac{1}{2N_0} \left[ \frac{1}{(1 - 2N_0)} \right]^{1/2}
\]
In the Solanaceae, assays of large pollen populations estimated the rate of spontaneous mutations to be 0.2–0.4 $\times 10^{-6}$ gene-sequence exchanges per generation$^4$. However, all such mutations caused self-compatibility instead of conversion to a new mating type, so the rate of origination of new specificities is expected to be lower than this value. In order for coalescence times of alleles to reach 40 My as implied by the allele allelogeny, an origination rate as high as $10^{-6}$ would require a long-term $N_e$ of $4 \times 10^{10}$ (Ref. 19). At this population size, 375 alleles are expected to be maintained, a value in excess of all observed allele numbers, even for species-wide estimates$^{21}$. At an origination rate of $10^{-7}$, a long-term population size of $2 \times 10^4$ is required for $T_e$ to equal 40 My and 66 alleles are maintained at equilibrium, a value more in keeping with observed allele numbers. Origination rates $\leq 10^{-8}$ gene$^{-1}$ generation$^{-1}$ increase $T_e$ but make it difficult to maintain observed allele numbers in relatively small, isolated populations$^{19}$.

Molecular tools for assaying S-diversity

Until recently, data on allele numbers and sequence relationships came from different sources. Labor-intensive diallel crossing studies were used to estimate the number of alleles in populations, their frequencies and the degree of overlap in the alleles found among populations of a species. Sequence data, on the other hand, were limited to relatively few alleles cloned from various species without reference to natural population structure. PCR (polymerase chain reaction)-based methods (Box 4) can now be used to amplify S-alleles from individuals, thus providing rapid population surveys of S-allele numbers as well as sequence information for gene-genealogical studies. Application of models to data on allele numbers and sequence diversity from natural populations provides a tool for examining recent and long-term population dynamics.

Box 2. Trans-specific evolution

Trans-specific evolution occurs when a polymorphism present in an ancestral taxon is transmitted to descendants. In the figure below, alleles A and B were transmitted to extant Species 1 and 2 from their ancestor. Both alleles have accumulated changes through time but a gene-genealogy will group A' with A" and B' with B", so that sister alleles will occur in different species. Trans-specific evolution is a common feature of loci under balancing selection, such as the S-locus in plants (Fig. 1) and the MHC loci of vertebrates$^{31,34,36}$. In contrast, neutral polymorphism usually coalesces within extant species.

Box 3. Estimating long-term effective population size

Gene-genealogies from loci under balancing selection can be used to estimate long-term effective population size ($N_e$). One method uses the number of trans-specific lineages in a sample assuming a stochastic model for lineage turnover. Specifically, we calculate the probability $g_{2N_e}$ that, given a sample of n extant alleles, they coalesce into k lineages $t$ generations ago. It is convenient to replace $t$ by $t'$, defined as:

$$t' = \frac{t}{2N_e}$$

where $t'$ is a scaling factor specific to frequency dependent selection at the S-locus:

$$t'_s = \frac{\sqrt{2}}{16N_e^2(1-J\nu)}$$

where $\nu$ is the origination rate, and $J$ and $a$ are defined in Box 1.

The probability, $g_{2N_e}(t')$, is then given by:

$$g_{2N_e}(t') = \sum_{n=1}^{\infty} (2m-1)(1-\frac{1}{n})^{n-1} k(n)\exp \left\{ -\frac{m(m-1)t'}{2} \right\}$$

for $2 < k < n$, where $n_{m}$ = $n_{m-1} - 2(n-1)(n-2)...[(n-m+1), n_{m} = n(n-1)(n-2)...(n-m-1)$, and $m$ is a variable changing from $k$ to $n$ (Refs. 17, 19). For a given $t'$, we then calculate the probability $g_{2N_e}(t')$ over the range of $t'$ to find the value of $t'$ that maximizes the probability of observing $n$ and $k$.

The phylogeny in Fig. 1 was used to estimate $k$ for Solanum carolinense and Physalis crassifolia. An S-lineage was considered to be as old or older than the divergence of Nicotiana from the clade containing both Solanum and Physalis, if it inserted into the phylogeny at a node ancestral to an allele from Nicotiana or Petunia$^{36}$. Nine such ancient lineages are found in S. carolinense, while only two occur in the sample of alleles from P. crassifolia (see Fig. 1). Likelihood curves for both species (S. carolinense, dotted line, $k = 9$, $n = 15$; P. crassifolia, solid line, $k = 2$, $n = 28$) are shown in the figure below.

For instance, a PCR-based approach was used to survey S-allele numbers and sequence diversity in natural populations of two solanaceous species$^{37,38}$. In Solanum

<table>
<thead>
<tr>
<th>Effective population size ($N_e$)</th>
<th>S. carolinense</th>
<th>P. crassifolia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin rate ($\nu$)</td>
<td>$0.5 \times 10^4$</td>
<td>$0.8 \times 10^4$</td>
</tr>
<tr>
<td>$10^{-7}$</td>
<td>$2.2 \times 10^4$</td>
<td>$0.9 \times 10^4$</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>$3.0 \times 10^4$</td>
<td>$0.8 \times 10^4$</td>
</tr>
</tbody>
</table>

The value of $N_e$ associated with the maximum-likelihood estimate of $t'$ is obtained from eqn (1) above.

Assuming the time since divergence of Nicotiana from the clade containing both Physalis and Solanum is 30 My (Ref. 9), and assuming an average generation time of 2 years, the $N_e$ for origination rates ($\nu$) from $10^{-2}$ to $10^{-4}$ are presented below. For all origination rates, the long-term $N_e$ for P. crassifolia is substantially smaller than for S. carolinense in contrast to estimates of current $N_e$ based on the number of alleles (Box 1).
carolinense, a weedy rhizomatous plant of the southeastern USA and northern Mexico, sampling of a total of 24 individuals from two populations separated by approximately 250 km revealed a total of (only) 13 alleles, 11 of which were present in both populations providing a species-wide estimate of 13–14 S-alleles. By contrast, allelic diversity in Physalis crassifolia, a perennial subshrub of southwestern deserts, was similar to that found in previous studies of GSI plants. Twenty-eight alleles were found in the first 22 individuals sampled, providing a single population estimate of 44 alleles. The striking difference in the numbers of alleles found in these species may be attributable to differences in ecology. The rhizomatous S. carolinense occurs in small, short-lived patches on disturbed sites suggesting small N_0. By contrast, P. crassifolia lacks the weedy habit occurring in undisturbed habitats of the Mojave and Colorado deserts and does not reproduce vegetatively. It is interesting to note that species of SSI Asteraceae thought to suffer reduced reproduction owing to low mating-type diversity reproduce clonally, a trait shared in common with S. carolinense.24,25

Sequences of S-alleles from S. carolinense and P. crassifolia indicate surprising differences in the tempo and mode of diversification. Alleles from S. carolinense show extensive trans-specific evolution (Fig. 1). The 13 alleles represent nine lineages that pre-date the divergence of Solanum and Nicotiana (Fig. 1, Box 3). In P. crassifolia, on the other hand, there is extensive, relatively recent allelic diversification, with evidence for only two old (transgeneric) lineages (Fig. 1, Box 3). Recent diversification is indicated also by the relatively low level of sequence divergence among alleles within each of the old lineages, resulting in significant differences in average allele age between the samples from the two species.26 The clustering of Physalis alleles is even more striking given molecular phylogenetic information that places Physalis and Solanum as sister genera relative to the more basal Nicotiana and Petunia.27,28 The sequence diversity of alleles found in Physalis implies that most extant alleles were derived relatively recently, as, for example, following a population bottleneck, rather than as the result of continuous turnover in the absence of fluctuation in population size (Box 3). Following a bottleneck, the rate of successful invasion by mutant alleles with new specificities substantially increases, leading to rapid diversification within the extant lineages.21 The striking differences in both allele number and age among the first species surveyed indicates that the S-locus may provide a wealth of information pertaining to current and to historical population processes.

Origins of the S-locus

Molecular analysis of S-gene products in GSI Solanaceae and Papaveraceae, as well as SSI Brassicaceae, have revealed three different molecular mechanisms of rejection and, by extension, independent origins of SI in these families or their ancestors.8 In contrast, the stylar S-gene products in GSI Rosaceae and Scrophulariaceae have recently been cloned and found to be RNases as in the Solanaceae.29–31 Whether this represents homology of the SI systems of these dicot families is unclear. Some phylogenetically intervening families (e.g. Brassicaceae) have evolved independent SI mechanisms but the possibility remains that ga-metophytic SI could be homologous in a number of higher dicot families. Equally intriguing is the possibility that the use of RNases as incompatibility genes in different plant families represents convergence. It has been suggested that RNases functioning in defense against pathogen invasion through floral organs may have been converted to function in self-pollen rejection.22 Identification of RNases involved in defense may lead to discovery of the antecedents of the S-locus. Whether or not S-RNases in different angiosperm families are homologous, the PCR methods used to survey S-alleles in the Solanaceae will probably be useful in any plant family with an RNase-based incompatibility system.

The fact that related non-S-genes are present in various plant families suggests an approach to dating the origin of SI by estimating the time taken since the change of the non-S-genes from their ancestral function to their use as an SI mechanism. What is needed is a measure of the rate of substitution along branches of the gene-family tree, and a divergence time from the fossil record, or another source, with which to calibrate the clock. Uyenoyama's26,27,34,35 used sequence data on the SLG/SRK/SNR multigene family in the Brassicaceae to estimate the time since the origin of SI. This estimate was 4–5 times as long as the time since divergence of the Brassica oleracea and B. campestris, an event assumed to have occurred 10 My ago.

In addition to estimating the age of the SI system found in the Brassicaceae, Uyenoyama's method estimated rates of substitution in S-alleles relative to those of related genes not subject to frequency-dependent selection. Rates of substitution at the S-locus of Brassicaceae were found to be significantly lower than those of related genes. This result was important since hypermutability had been suggested as a factor in the maintenance of extreme polymorphism.36 Instead, the large numbers of alleles and their high levels of sequence divergence appeared to be brought about through long persistence of allelic lineages. In fact, it appeared that most S-lineages appeared soon after the origin of the system, and subsequently, the origination rate of allelic lineages slowed down. To account for this...
pattern, Uyenoyama speculates that descendant alleles may preferentially replace their parents. Alternatively, the slowdown in the appearance of new S-alleles might reflect constraints on the nature of changes that can generate new specificities. Given the recent data from Physalis, it would appear that some lineages, at least, can produce more than 20 different specificities. Further analysis of the limits of within-lineage diversification will require the results of molecular studies of additional species and populations.

**Conclusion**

Empirical data from the S-locus have provided impetus for the development of population genetic theory several times during the 20th century. Emerson's discovery of extreme polymorphism prompted Wright and others to model allele numbers and sampling distributions under frequency dependent selection. The discovery of extensive trans-specific evolution at the MHC loci of vertebrates and the S-locus of plants has helped to promote extensions of coalescence theory to loci under overdominant and frequency dependent selection. Identification of the genes responsible for rejection specificity permits the application of molecular methods to the study of natural populations. We have only begun to accumulate samples from nature. The striking differences in both allele number and age among samples from *S. carolinense* and *P. crassifolia* suggest that a wealth of information concerning ecology and history may be stored at the S-locus. In particular, the S-locus will provide a useful tool for investigation of the relationship of species ecology to effective population size. The prevalence of historical bottlenecks at the S-locus provides a means for evaluating the frequency of founder-event speciation.

**Fig. 1.** The phylogeny of 63 (partial) S-allele sequences from the Solanaceae. The neighbor-joining tree was derived from amino acid distances using the PAM 200. Dayhoff similarity matrix. Numbers indicate consensus bootstrap values that exceed 50%. Brackets indicate alleles from *Physalis crassifolia* (Pc), and arrows indicate alleles from *Solanum carolinense* (Sc). Other solanaceous taxa are *S. chacoense* (Scs), *S. tuberosa* (Stub), *Lycopersicon peruvianum* (Lperu), *Petunia hybrida* (Phybrida), *P. inflata* (Pinflata) and *Nicotiana alata* (Nalata). The tree is rooted with non-S RNases from *Lycopersicon esculentum* (LE) and *Momordica charantia* (MC; see Ref. 26 for sequence sources). Taxa in bold represent basal genera.
Acknowledgements

The support of NSF to both authors made this work possible. We are grateful to M.K. Uyenoyama and two anonymous reviewers for comments on the manuscript.

References

2 Wright, S. (1960) 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
16 Takahata, N. and Nei, M. (1980) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci, Genetics 124, 677-678
22 Richman, A.D. et al. (1995) S-allele sequence diversity in natural populations of Solanum carolinense (Hornemianic), Heredity 75, 403-415
29 Broohtans, W. et al. (1995) cDNA cloning and molecular analysis of two self-incompatibility alleles from apple, Plant Mol. Biol. 27, 499-511
35 Paxman, G.J. (1965) The maximum likelihood estimation of the number of self-sterility alleles in a population, Genetics 48, 1029-1032
41 Williams, R.D. and Williams, W. (1947) Genetics of red clover (Trifolium pratense L.) compatibility. III. The frequency of incompatibility S alleles in two non-pedigree populations of red clover, J. Genet. 48, 69-19
43 Bateman, A.J. (1954) Self-incompatibility systems in angiosperms. II. Iberis amara, Heredity 8, 365-332