# Genealogical Structure Among Alleles Regulating Self-Incompatibility in Natural Populations of Flowering Plants

**Marcy K. Uyenoyama**

*Department of Zoology, Duke University, Durham, North Carolina 27708-0325*

Manuscript received April 5, 1997
Accepted for publication August 5, 1997

ABSTRACT

A method is proposed for characterizing the structure of genealogies among alleles that regulate self-incompatibility in flowering plants. Expected distributions of ratios of divergence times among alleles, scaled by functions of allele number, were generated by numerical simulation. These distributions appeared relatively insensitive to the particular parameter values assigned in the simulations over a fourfold range in effective population size and a 100-fold range in mutation rate. Generalized least-squares estimates of the scaled indices were obtained from genealogies reconstructed from nucleotide sequences of self-incompatibility alleles from natural populations of two solanaceous species. Comparison of the observed indices to the expected distributions generated by numerical simulation indicated that the allelic genealogy of one species appeared consistent with the symmetric balancing selection generated by self-incompatibility. However, the allelic genealogy of the second species showed unusually long terminal branches, suggesting the operation of additional evolutionary processes.

RICHMAN *et al.* (1996) examined phylogenetic relationships among nucleotide sequences encoding self-incompatibility (*S*)-alleles in natural populations of *Solanum carolinense* and *Physalis crassifolia.* These solanaceous species express gametophytic self-incompatibility (GSI), under which the specificity expressed by a pollen tube is determined by the *S*-allele in its haploid genome and seed parents reject fertilization by pollen tubes that express specificities encoded by either of their own *S*-alleles (see DE NETTANCOURT 1977). In the Solanaceae, the *S*-locus derives from a multigene family of ribonucleases (MCCLURE *et al.* 1990), with RNase activity directly mediating rejection of incompatible pollen tubes (HUANG *et al.* 1994; LEE *et al.* 1994; MURFETT *et al.* 1994). GSI in apple (*Malus domestica;* BROOTHAERTS *et al.* 1995; SASSA *et al.* 1996) and snapdragon (*Antirrhinum hispanicum;* XUE *et al.* 1996) appears to be homologous. By promoting fertilization by pollen that express rare specificities, GSI imposes intense balancing selection on the *S*-locus (see CLARK and KAO 1994).

*S*-allele genealogies estimated from nucleotide sequences derived from natural populations of the two solanaceous species showed strikingly different patterns: few *S*-alleles of ancient divergence in *S. carolinense* and many *S*-alleles of relatively recent divergence in *P. crassifolia* (RICHMAN *et al.* 1996). To explore the mode of evolution, a method originally developed by TAKAHATA (1993) for the analysis of balanced polymorphisms at loci within the vertebrate major histocompatibility complex (MHC) was modified for the *S*-locus. Comparison between the species of the number of *S*-alleles maintained and the number of *S*-allele lineages shared across solanaceous genera suggested that the long-term effective population size of *S. carolinense* may have exceeded that of *P. crassifolia* by at least an order of magnitude. A possible scenario is that much of the *S*-allele variation presently segregating in the *P. crassifolia* population was generated after a bottleneck in population size during which many ancient *S*-lineages were lost (RICHMAN *et al.* 1996; RICHMAN and KOHN 1996).

This approach assumed knowledge of key parameters, particularly rate of mutation to new *S*-alleles. Because such information is unavailable, RICHMAN *et al.* (1996) arbitrarily assigned a range of values for the mutation rate. An apparent discrepancy exists between the low number of *S*-alleles maintained in *S. carolinense* and the estimate of effective population size under the assigned mutation rates. Further, implicit in the estimation procedure was the assumption that genealogical aspects have converged to their steady-state distributions. This equilibrium assumption may be inconsistent with the interpretation developed on the basis of the estimates that the *P. crassifolia* population has experienced a bottleneck within the period since divergence among its *S*-allele lineages.

In the present study, I explore prospects for the development of descriptive indices of genealogies that are relatively insensitive to parameter values that are in general unknown for natural populations, particularly effective population size and rate of mutation. An ultimate goal is to design a tool for the diagnosis of demo-

*Corresponding author:* M. K. Uyenoyama, Department of Zoology, Box 90325, Duke University, Durham, NC 27708-0325.
E-mail: marcy@amida.zoo.duke.edu

graphic history. Toward this end, I adopt a numerical approach to detecting departures from the null hypothesis of constant mutation rate in a panmictic population of constant size. Numerical simulations generated frequency distributions for descriptors of genealogical structure among *S*-alleles. These indices, ratios of divergence times scaled by functions of allele number, show substantial variation, but less than the variation shown by the raw divergence times. The scaled ratios appear relatively insensitive to assignments of effective population size over a fourfold range and of rate of mutation to new *S*-allele specificities over a 100-fold range. Examination of observed values of these indices in samples from the natural populations surveyed suggests that it is the *S*-allele genealogy of *S. carolinense* rather than *P. crassifolia* that deviates more sharply from expectation. In particular, the *S. carolinense* genealogy exhibits unusually long terminal branches relative to maximum divergence time and total tree length.

## METHODS

**Gametophytic self-incompatibility:** I conducted numerical simulations of the evolution of *S*-alleles under GSI in a population of hermaphroditic diploid individuals. Because the specificity expressed by pollen is determined by the gametophyte itself, the simulations incorporated immediate expression in pollen of mutations to new specificities. In contrast, the specificities rejected by the seed parent were limited to those determined by its own genotype at conception, even if a mutation had occurred subsequently in an egg cell of that individual. A mutation transmitted through the egg was first expressed in the zygote it formed rather than in the parent in which it arose. Because all mutations were assumed to generate novel specificities (infinite alleles mutation), a pollen grain bearing a new mutation could fertilize all individuals, including the individual that produced it.

Zygote formation in the numerical simulations reflected these assumptions. A diploid maternal genotype and a haploid pollen gamete were drawn, without exclusion of the maternal parent as a pollen donor. After mutation in the pollen gamete, its *S*-allele class was compared to the unmutated maternal genotype to determine compatibility. If incompatible, the pollen gamete was discarded and the process of sampling of pollen and mutation repeated until a compatible pairing occurred.

**Genealogical structure:** *Genealogies of genes and alleles:* Methods for recording genealogical structure among genes and *S*-allele classes differed in some respects from those of TAKAHATA and NEI (1990) and VEKEMANS and SLATKIN (1994). Divergence times among all pairs of genes and the *S*-allele type of each gene were recorded. Pairwise divergence times unambiguously determine (possibly multifurcating) genealogies among genes in the population. The allelic genealogies analyzed are in fact genealogies among genes chosen to represent their allelic class. For example, the node joining two allelic classes corresponds to the last common ancestor of the genes that represent the two classes and not the origin of the first specificity-determining mutation that distinguishes the alleles, as is the case for the genealogies described by TAKAHATA and NEI (1990).

Newly arisen *S*-allele lineages may well be nested within their parental lineages. Upon the origin of a new *S*-allele specificity, the particular gene that gave rise to the new *S*-allele class will have diverged more recently from the offspring class than from other members of its own class. Variation in divergence time among pairs of genes expressing the offspring and parental *S*-allele specificities will persist until coalescence occurs within the parental class. If coalescence within the parental class traces back to the particular gene that gave rise to the offspring class, then the generation of divergence between the classes will correspond to the generation in which the offspring class arose. If coalescence traces back to a different member of the parental class, then the divergence time between the classes will exceed the time since the origin of the new class, corresponding instead to the divergence time between the parental gene and the coalescent gene.

Phylogenetic relationships among *S*-allele lineages were determined from divergence times among genes by choosing the first gene encountered of a given *S*-allele class to represent the class. Different choices of genes to represent the *S*-allele classes might have given rise to different genealogies if at the time of the census both parental and offspring classes were segregating in the population and a genetic turnover had not yet occurred within the parental class. Functionally distinct *S*-allele classes persist over time scales orders of magnitude greater than coalescence among functionally equivalent genes within *S*-allele class [see discussions of "effective gene number" in TAKAHATA (1990) and "coalescence time of all gene copies" in VEKEMANS and SLATKIN (1994)]. Consequently, the ambiguity in *S*-allele genealogies introduced by the arbitrary choice of genes to represent the allelic classes is insignificant in practice.

*Independent epochs:* All genes segregating in the population at a given point in time descend from a single gene in a previous generation, and a single gene in the present population will eventually give rise to all genes in a future generation. That future generation marks a complete turnover of genetic lineages. All genes in the initial generation of the simulations were regarded as distinct (not identical by descent) and the end of the first epoch defined as the generation of the first complete turnover. At the end of each epoch, all lineages were again considered distinct and genealogies were

traced at intervals until the next complete turnover. As a consequence of the extreme balancing selection imposed by the expression of self-incompatibility, S-allele lineages persist over very long periods of time. Sharing of lineages causes correlations in genealogical structure across generations. Genealogies are considered independent only between epochs, with genealogies within epochs treated as observations of a single phylogenetic structure.

*Records of divergence:* A population of 1000 diploid individuals was initiated with three S-alleles and run through the first epoch (at which time all genes descended from a single gene in the initial generation) to the second epoch (at which time all genes descended from a single member of the first turnover generation). At the second epoch, the population was considered to be independent of initial conditions. This seed population was used to initiate all simulation runs. To reduce the influence of the seed population, records of genealogical structure during the first epoch after initialization at the seed population were discarded. Further, records for the generation terminating each epoch were also discarded because genealogies at the moment of coalescence show unusual structure (Figure 1; see TAJIMA 1990a).

**Genealogical structure under symmetric balancing selection:** Theoretical expectations for divergence times among neutral genes are well known. Derivation of these expressions relies only on the exponential distribution of successive coalescence times (see lucid reviews by TAVARÉ 1984 and HUDSON 1990). TAKAHATA (1990) showed that coalescence times among alleles evolving under symmetric overdominance in viability also follow an exponential distribution, and VEKEMANS and SLATKIN (1994) showed that the coalescence process among alleles subject to gametophytic self-incompatibility shares this property as well. TAKAHATA's (1990) proposal that neutral theory may provide a qualitative guide to genealogical structure under symmetric balancing selection provides a key motivation for the present study.

I conducted numerical simulations to determine genealogical structure among S-alleles evolving under constant population size and mutation rate. Under pure neutrality, expectations of the five time intervals recorded are given by

$$\begin{aligned} E[T] &= 4Na_n \\ E[D] &= 4N(1 - 1/n) \\ E[P] &= 2N \\ E[S] &= 4N \\ E[B] &= 4Nb_n \end{aligned} \qquad (1)$$

in which $N$ represents effective population size; $T$ total time in the genealogy of a sample of $n$ genes and $a_n = \Sigma_1^{n-1} 1/i$ (WATTERSON 1975); $D$ maximum divergence time (coalescence time of all genes) and $P$ average time since divergence between pairs of the $n$ genes sampled (HUDSON 1982; KINGMAN 1982; TAJIMA 1983); $S$ the expected sum of the terminal branch lengths (FU and LI 1993); and $B$ the average length of the base branches, which emanate from the root, with $b_n = 1/n + \Sigma_2^{n-1} 1/i^2$ (see APPENDIX).

To establish a basis for comparison between empirical observation and simulation results and among simulation results under different parameter assignments, an attempt was made to remove the influence of population size ($N$) and rate of mutation to new S-allele specificities ($\mu$) by considering ratios of the lengths in (1), scaled by simple functions of the number of alleles in the sample ($n$). Of the 10 possible ratios of pairs of five time intervals, up to four are independent. Values for four scaled ratios were obtained by numerical simulation and also estimated from empirical observations:

$$\begin{aligned} R_{PT} &= \frac{2Pa_n}{T} \\ R_{ST} &= \frac{Sa_n}{T} \\ R_{SD} &= \frac{S(1 - 1/n)}{D} \\ R_{BD} &= \frac{B(1 - 1/n)}{Db_n} \end{aligned} \qquad (2)$$

The coefficients used to scale the ratios were determined by requiring the expressions in (2) to reduce to unity upon replacement of the lengths by their neutral expectations, given in (1).

I conducted numerical simulations to explore the distribution of the scaled indices (2) in S-allele genealogies evolving under the form of symmetric balancing selection induced by the expression of GSI in a population of constant size and rate of mutation to new S-alleles. Because genealogies among all segregating S-alleles were examined, the number of alleles in the sample ($n$) in (1) and (2) was replaced by the total number of S-alleles in the population. Under symmetric balancing selection, the scaled indices may well be expected to deviate from unity, not only because the ratio of expectations departs from the expectation of ratios, but also because the correspondence between gene genealogies under neutrality and allelic genealogies under balancing selection is only approximate. My objective is to examine not the quantitative values of these indices but rather to explore whether they can be used to detect departures in genealogical structure from that expected for populations of constant size and rate of mutation to new allelic classes. Most of the simulation results were generated under a single assignment of population size ($N = 1000$ diploid individuals) and mu-

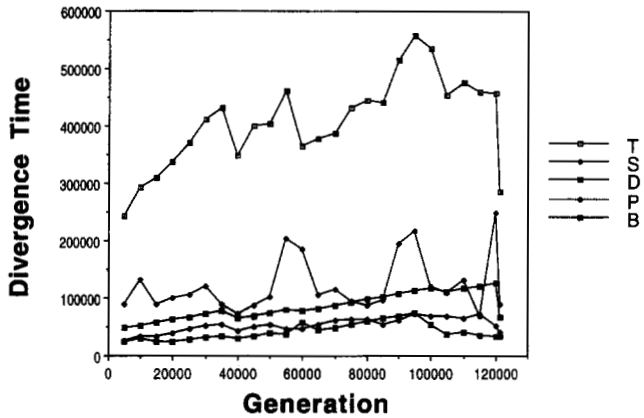


FIGURE 1.—Divergence times, including total time in the genealogy ($T$), sum of the terminal branch lengths ($S$), maximum divergence time ($D$), average pairwise divergence time ($P$), and average base branch length ($B$), in genealogies of $S$-alleles sampled at 5000-generation intervals during a course of a typical epoch.

tation rate ($\mu = 2.5 \times 10^{-6}$ per gamete). Additional simulations explored the effects of changing the parameter values.

To examine the distribution of lengths of the terminal branches, I scaled each terminal branch as in (2), replacing $S$ by $S/n$. This scaling reflects that the length of each terminal branch corresponds in expectation to $S/n$.

In the numerical simulations, genealogical relationships among representatives of all $S$-allele classes were determined at intervals within epochs. Average standard deviations of the genealogical measures within each of 151 epochs describe the magnitude of variation observed over the course of a complete turnover. Records compiled over several intervals within an epoch generated a frequency distribution for the scaled indices. These frequency distributions for different epochs were then averaged to generate the summary distributions shown in Figure 2, representing the stochastic variation among independent genealogies generated by the same evolutionary process. I also report means among different epochs and standard errors of the means, which are indicative of the magnitude of stochastic variation.

## RESULTS

**Frequency distributions generated by numerical simulation:** *Large variation in divergence times:* Figure 1 presents for a typical run aspects of the genealogy of all segregating $S$-alleles, including the total time in the genealogy ($T$), the time since the most recent common ancestor ($D$), the average time since divergence between pairs of $S$-alleles ($P$), the sum of the terminal branch lengths ($S$), and the average of the base branch lengths ($B$). All genes in the initial generation diverged from their most recent common ancestor 42,390 generations prior to the start of the run. The coalescence time increased by the time interval separating observations (5000 generations) if the lineages that separated at that deepest divergence persisted and decreased if those lineages were lost. After 105,000 generations, only two lineages that diverged prior to the initial generation remained. At generation 120,000, a single $S$-allele out of 22 segregating classes represented the sole descendant of one of these lineages. This lineage was lost in generation 121,275, with $D$ falling from 126,189 to 66,896 generations, the coalescence time among the remaining $S$-alleles. This event, representing a complete turnover of $S$-allele lineages in the population, ended the epoch.

Figure 1 illustrates the considerable variation exhibited by the divergence times over the course of an epoch. Table 1 presents for 151 independent epochs means, standard errors, and average standard deviations within epochs of the actual number of $S$-alleles ($n$), effective number of alleles ($n_e$), and divergence times. Effective allele number corresponds to the inverse of homozygosity, defined as the sum of the squared frequencies of alleles (KIMURA and CROW 1964). Divergence times show variances on the order of the square of the means, as expected for exponentially distributed variables.

*Relative stability of scaled ratios:* Table 2 provides for the same 151 epochs means, standard errors, and average standard deviations of the scaled ratios given in (2). Although these indices show appreciable variation, they show considerably smaller coefficients of variation than the raw divergence times. Further, the means for none depart significantly from the approximate expectation of unity. This apparent agreement supports TAKAHATA's (1990) proposal that neutral theory provides an approximate basis for expectations concerning the genealogical structure of symmetrically balanced polymorphisms.

Figure 2 presents frequency distributions for the scaled ratios, averaged over epochs, together with cumulative frequency distributions. All distributions appear to have a single mode in the vicinity of unity, and $R_{PT}$ appears to approximate a Gaussian variable. I use these frequency distributions, generated by numerical simulation, to examine genealogical structure among $S$-alleles observed in natural populations.

**Comparison to natural populations:** *Long terminal branches:* Values of the indices defined in (2) were computed from generalized least-squares estimates of branch lengths in genealogies reconstructed using nucleotide sequences sampled from natural populations of the two solanaceous species (see RICHMAN *et al.* 1996). Comparison to the frequency distributions generated by numerical simulation reveals significant departures in *S. carolinense* (Table 2). Large values of $R_{ST}$ and $R_{SD}$ for *S. carolinense* indicate unusually long termi-

**TABLE 1**

**Variation in allele number and divergence times**

| | $n$ | $n_e$ | $T$ | $D$ | $P$ | $S$ | $B$ |
|---|---|---|---|---|---|---|---|
| Full | | | | | | | |
| Mean | 23.3 | 20.5 | 391,946 | 95,439 | 50,589 | 114,802 | 61,281 |
| SE | 0.7 | 0.5 | 110,710 | 43,110 | 19,999 | 21,662 | 38,319 |
| MSD | 1.6 | 1.2 | 75,089 | 28,353 | 13,420 | 41,771 | 28,431 |
| Subsample | | | | | | | |
| Seed | 23 | 20.0 | 502,739 | 131,311 | 84,258 | 87,067 | 69,188 |
| Mean | 13.2 | 10.8 | 451,804 | 131,311 | 84,212 | 126,390 | 72,590 |
| SE | 1.5 | 1.6 | 29,554 | 0 | 6,267 | 52,794 | 11,971 |

MSD, mean standard deviation.

nal branches relative to the total length of the tree and maximum divergence time. Estimates for these indices in *P. crassifolia* show similar trends but do not depart significantly from expectation.

Approximate significance levels were inferred from the frequency distributions in Figure 2, in which arrows indicate the observed values and bars approximate 95% confidence intervals. Treating the indices as Gaussian and applying *t*-tests gave comparable results. Unusually long terminal branches in *S. carolinense* (and to a lesser extent in *P. crassifolia*) emerge as the most striking characteristic of the genealogies estimated for the natural populations.

An expected distribution of terminal branch lengths was formed by averaging across the 151 independent epochs generated by numerical simulation distributions of terminal branch lengths within epochs. Figure 3 shows the expected distribution (Full) of terminal branch lengths scaled to total tree length ($T$). Distributions scaled to the other time lengths ($D$, $P$, and $B$) showed similar forms.

Figures 4 and 5 contrast the expected terminal branch distribution with those observed for *S. carolinense* and *P. crassifolia*. To ensure a minimum of five branches expected given the numbers of alleles observed, I compared the expected and observed numbers of scaled terminal branches in the first category ($\leq 0.5$). Chi-square tests indicated that both *S. carolinense* ($\chi^2_{[1]} = 9.33$) and *P. crassifolia* ($\chi^2_{[1]} = 7.14$) show significant ($P < 0.01$) deficiencies of terminal branches in this smallest relative length category.

*Star phylogenies:* Long terminal branches are also characteristic of a star phylogeny. In a star phylogeny, the time since divergence between all pairs of alleles are identical ($P = D$), and the sum of the terminal branches corresponds to the total time in the genealogy ($S = T = nD$). Table 2 shows values of the scaled ratios for a star phylogeny, given the numbers of alleles examined for *S. carolinense* ($n = 13$) and *P. crassifolia* ($n = 17$). Values of the scaled ratios observed for both species lie between the values obtained by simulation and for star phylogenies, suggesting a tendency toward a star-like structure.

**Effects of undersampling:** One factor that might generate unusually long terminal branches in a genealogy is undersampling of the population: the failure to observe an allele would cause an apparent extension of the terminal branch of its sister allele. Under neutrality,

**TABLE 2**

**Variation in scaled ratios**

| | $R_{PT}$ | $R_{ST}$ | $R_{SD}$ | $R_{BD}$ |
|---|---|---|---|---|
| Full | | | | |
| Mean ± SE | 0.93 ± 0.11 | 1.14 ± 0.22 | 1.40 ± 0.46 | 0.87 ± 0.16 |
| MSD | 0.16 | 0.34 | 0.52 | 0.21 |
| Empirical | | | | |
| *S. carolinense* | 0.63 ± 0.02 | 2.40 ± 0.08*** | 6.27 ± 0.41*** | 0.71 ± 0.05 |
| *P. crassifolia* | 0.72 ± 0.02 | 1.88 ± 0.09 | 2.64 ± 0.24 | 0.92 ± 0.05 |
| Star phylogeny | | | | |
| $n = 13$ | 0.48 | 3.10 | 12 | |
| $n = 17$ | 0.40 | 3.38 | 16 | |
| Subsamples | | | | |
| Seed | 1.25 | 0.65 | 0.64 | 0.78 |
| Mean ± SE | 1.16 ± 0.09 | 0.86 ± 0.34 | 0.89 ± 0.37 | 0.80 ± 0.13 |

MSD, mean standard deviation. *** $P < 0.001$.

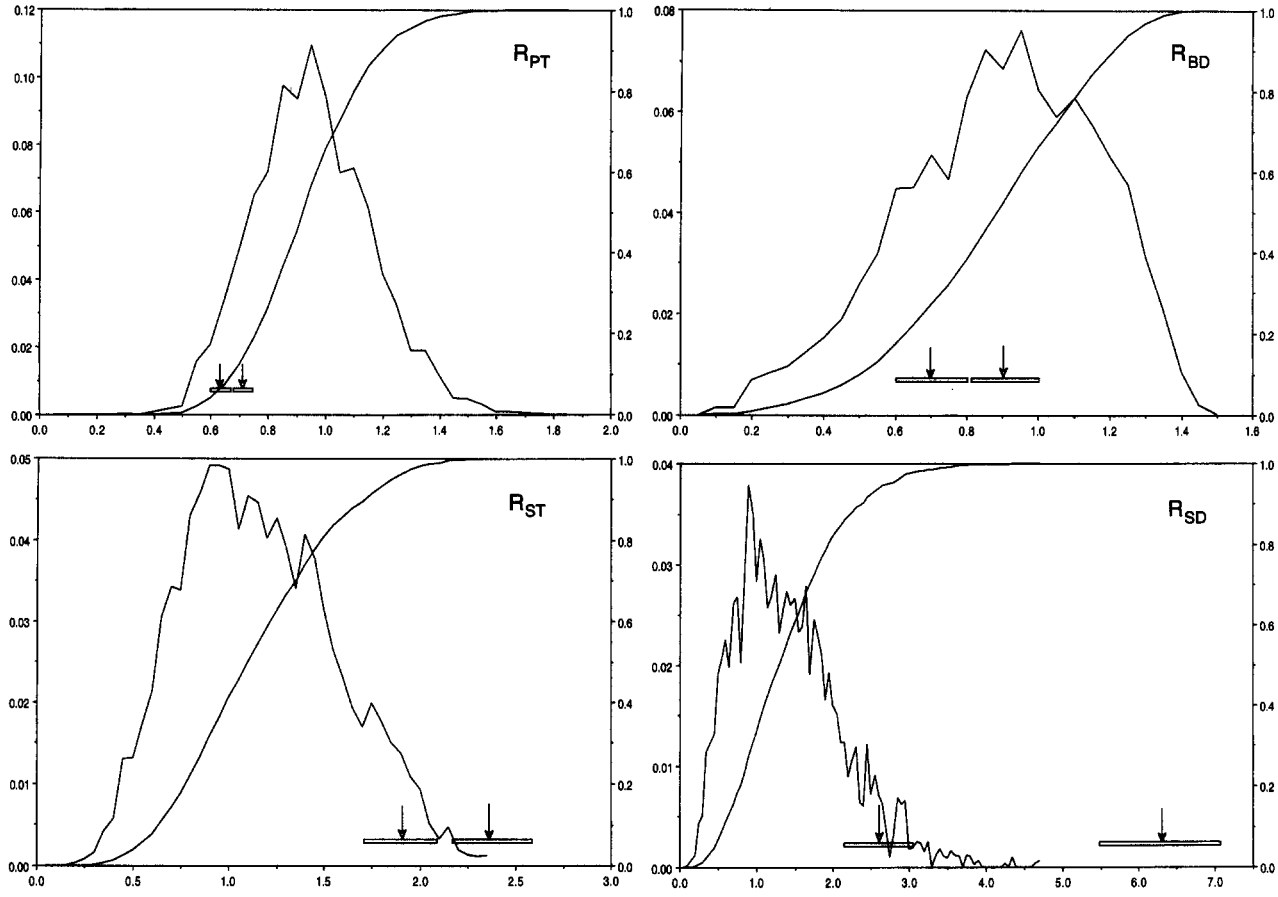


FIGURE 2.—Frequency and cumulative frequency distributions for the scaled ratios of divergence times, obtained by numerical simulation over 151 independent epochs. Closed arrows indicate values estimated from a sample of *S*-alleles from *S. carolinense* and open arrows *P. crassifolia;* bars indicate approximate 95% confidence intervals.

undersampling would not cause this effect because the expectations given in (1) refer to genealogical structure among genes within the sample, rather than among all genes segregating in the population (EWENS 1972; WATTERSON 1975). However, little analytical theory has been developed that addresses the properties of samples of alleles involved in the expression of self-incompatibility. To examine whether the unusually long terminal branches in the genealogy among *S*-alleles ob-

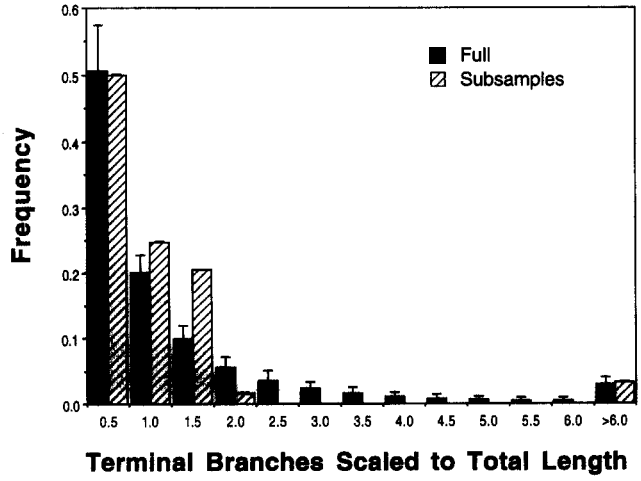


FIGURE 3.—Frequencies of terminal branch lengths, scaled to total length of the genealogy, expected and observed (Subsamples) in genealogies of *S*-alleles in 10,000 subsamples of 10 individuals from a simulated population of 1000 individuals.

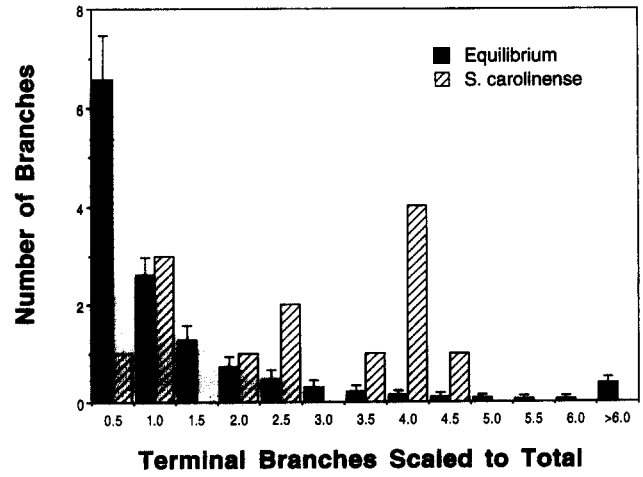


FIGURE 4.—Numbers of terminal branch lengths, scaled to total length of the genealogy, expected and observed in a genealogy of *S*-alleles sampled from *S. carolinense.*

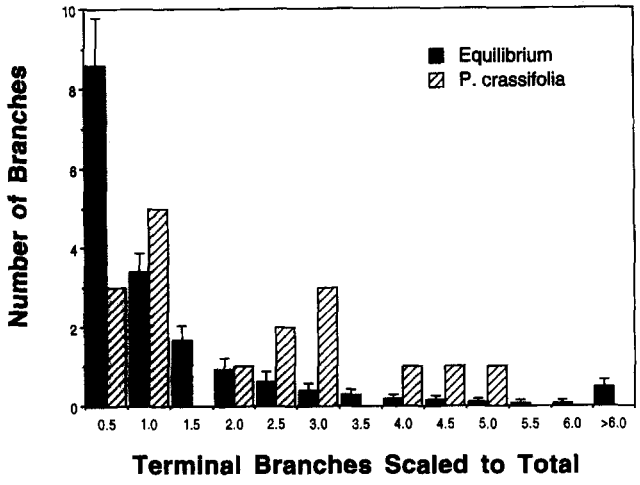


FIGURE 5.—Numbers of terminal branch lengths, scaled to total length of the genealogy, expected and observed in a genealogy of S-alleles sampled from *P. crassifolia*.

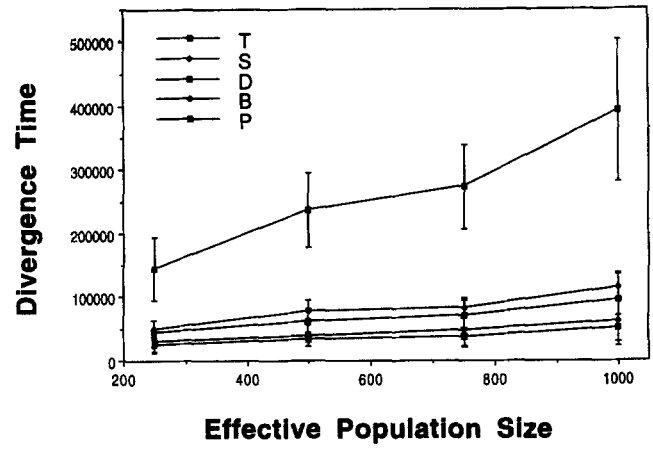


FIGURE 6.—Divergence times (in generations) in genealogies of S-alleles segregating in simulated populations over a fourfold range in population size under a rate of mutation to new S-alleles of $2.5 \times 10^{-5}$ per gamete. Divergence times include total time in the genealogy ($T$), sum of the terminal branch lengths ($S$), maximum divergence time ($D$), average pairwise divergence time ($P$), and average base branch length ($B$).

served in *S. carolinense* (and to a lesser extent in *P. crassifolia*) may merely represent undersampling of natural populations, I generated subsamples from a population obtained in the numerical simulations.

One of the simulated populations was arbitrarily selected and all aspects in a single generation recorded. The population was initiated at the standard seed population and iterated to the third complete turnover. In the chosen simulation, this event occurred 205,864 generations after the second complete turnover. A generation (81,659) between the second and third turnovers was arbitrarily chosen using a uniform random number generator. A total of 24 S-alleles were segregating in that generation. From the 1000 individuals constituting this population, 10,000 subsamples of 10 individuals were generated.

Tables 1 and 2 report measures for the seed population (Seed) and for the subsamples. As expected, the subsamples contained fewer distinct S-alleles, related through genealogies with shorter total length ($T$) and longer terminal branches ($S$; Table 1). In contrast, the scaled indices for the subsamples depart significantly from neither the population from which they were drawn (Seed) nor the full set of simulated populations (Full; Table 2). Genealogies for the subsamples do not reflect the large $R_{ST}$ or $R_{SD}$ values observed for *S. carolinense;* while the mean values of these indices in the subsamples exceed the values for the Seed population from which they were derived, the departures are not significant. Figure 3 indicates that the distribution of terminal branch lengths scaled to total length in the subsamples (Subsamples) does not deviate significantly from the full set of simulated populations (Full).

These comparisons indicate that undersampling does not in itself generate deviations from the expected distributions of the nature exhibited by the natural populations. Further, they suggest that the scalings in (2) adequately correct for twofold differences in allele number.

**Robustness to mutation rate and population size:** To explore the sensitivity of the results to the particular parameter values assigned in the simulations used to generate the expected distributions, I conducted additional numerical simulations, varying population size over a fourfold range and rate of mutation to new S-alleles over a 100-fold range (Table 3). The very close agreement between the observed ($n_e$) and expected ($E[n_e]$) effective numbers of alleles (Table 3) confirms both the accuracy of YOKOYAMA and HETHERINGTON's (1982) formula and the reliability of the numerical results.

Decreasing population size reduces both S-allele number and divergence time (Table 3 and Figure 6). In contrast, comparison of the *S. carolinense* and *P. crassifolia* S-allele genealogies indicates a negative relationship between number of S-allele lineages and divergence time among lineages. A difference in population size alone does not appear to account for differences between the S-allele genealogies observed in the two species.

S-allele number and divergence time show a strong dependence on rate of mutation to new S-alleles over a 100-fold range for $\mu$ (Table 3 and Figure 7). Interestingly, the log of the divergence times shows a nearly linear relationship to the log of the mutation rate (Figure 7; *cf.* Figure 4A of VEKEMANS and SLATKIN 1994). Unlike population size, variation in mutation rate induces a negative association between S-allele number and divergence time. This relationship supports the interpretation that the negative correlation observed in comparing the *S. carolinense* and *P. crassifolia* S-allele

**TABLE 3**

**Influence of population size and mutation rate**

| $\mu^a$ | $N$ | $n_e$ | $E[n_e]$ | $R_{PT}$ | $R_{ST}$ | $R_{SD}$ | $R_{BD}$ |
|---|---|---|---|---|---|---|---|
| 1 | 250 | 10.4 ± 0.3 | 10.7 | 0.98 ± 0.08 | 1.09 ± 0.20 | 1.27 ± 0.36 | 0.91 ± 0.15 |
| 1 | 500 | 14.5 ± 0.5 | 14.8 | 0.95 ± 0.08 | 1.12 ± 0.15 | 1.33 ± 0.29 | 0.87 ± 0.11 |
| 1 | 750 | 17.8 ± 0.3 | 18.0 | 0.93 ± 0.12 | 1.13 ± 0.21 | 1.35 ± 0.39 | 0.91 ± 0.16 |
| 1 | 1000 | 20.5 ± 0.5 | 20.1 | 0.93 ± 0.11 | 1.14 ± 0.22 | 1.40 ± 0.46 | 0.87 ± 0.16 |
| 5 | 1000 | 24.4 ± 0.7 | 24.3 | 0.97 ± 0.14 | 1.05 ± 0.19 | 1.23 ± 0.36 | 0.85 ± 0.16 |
| 10 | 1000 | 26.6 ± 0.7 | 26.5 | 0.93 ± 0.12 | 1.04 ± 0.18 | 1.21 ± 0.31 | 0.91 ± 0.13 |
| 20 | 1000 | 29.6 ± 1.1 | 29.4 | 0.90 ± 0.10 | 1.05 ± 0.23 | 1.35 ± 0.44 | 0.78 ± 0.14 |
| 50 | 1000 | 35.8 ± 1.0 | 35.4 | 1.00 ± 0.12 | 0.99 ± 0.18 | 1.20 ± 0.40 | 0.84 ± 0.14 |
| 100 | 1000 | 41.6 ± 2.3 | 43.1 | 0.90 ± 0.17 | 1.07 ± 0.22 | 1.32 ± 0.42 | 0.90 ± 0.14 |

[a] Relative to $2.5 \times 10^{-5}$ mutations/gamete.

genealogies suggests greater differences between the species in mutation rate than population size (UYENOYAMA 1997).

In contrast with allele number and divergence times, the means of the scaled ratios appear insensitive to the parameter assignments over the ranges examined (Table 3).

## DISCUSSION

**Analysis of genealogical structure:** *Scaled ratios of divergence times:* Divergence times among segregating genetic lineages reflect selective regime, effective population size, and other aspects of the context in which they evolve. By enforcing strict heterozygosity at the *S*-locus, gametophytic self-incompatibility engenders an extreme form of balancing selection that promotes the maintenance of *S*-allele lineages for exceedingly long periods. Variation in divergence times increases as the square of their means, reflecting the exponential distribution of coalescence times among lineages.

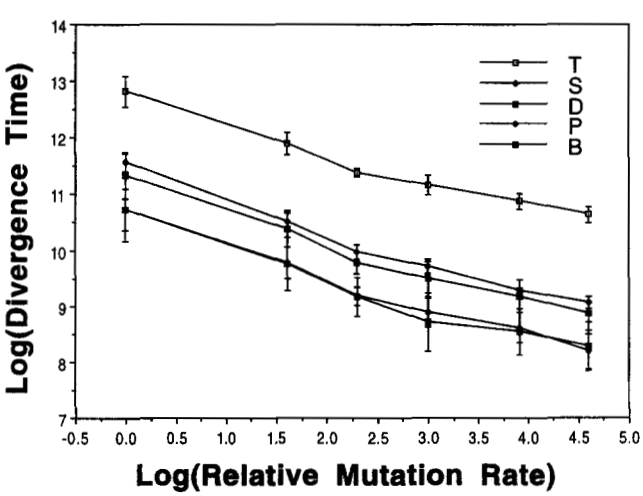


FIGURE 7.—Divergence times in genealogies of *S*-alleles segregating in a simulated population of 1000 individuals over a 100-fold range in rate of mutation to new *S*-alleles. Divergence times (in generations) include total time in the genealogy (*T*), sum of the terminal branch lengths (*S*), maximum divergence time (*D*), average pairwise divergence time (*P*), and average base branch length (*B*). Mutation rate is scaled relative to a base rate of $2.5 \times 10^{-5}$ per gamete.

This study explores whether structural aspects of gene genealogies reveal the signatures of evolutionary and demographic processes, apart from the circumstances of particular populations. Other approaches have examined the distribution of internodal lengths (see HEY 1992; NEE *et al.* 1995; KUBO and IWASA 1995). Internodal lengths in genealogies estimated from sequences of alleles regulating sporophytic self-incompatibility in Brassica appeared unusual, but a crude analysis indicated nonsignificant departures (UYENOYAMA 1995). Because stochastic variation in internodal lengths is large and the number of internodal lengths estimated was small, whether genealogies of Brassica *S*-alleles in fact conform to expectation remains undetermined.

Ratios of divergence times, scaled to minimize dependence on allele number, may provide useful descriptors of genealogical structure. Although the scaled indices were proposed on the basis of theoretical expectations under selective neutrality, I studied their behavior under the expression of gametophytic self-incompatibility. Motivation for this approach derives from TAKAHATA's (1990) finding that the process of coalescence under symmetric overdominant viability selection resembles that under selective neutrality, a property shared by the symmetric balancing selection imposed by GSI (VEKEMANS and SLATKIN 1994). Results from numerical simulations of evolving *S*-locus variation indicated that the proposed scaled ratios of divergence times show relatively low sensitivity to effective population size, mutation rate, and allele number.

Coalescence times observed in the numerical simulations exhibited the great variability expected for exponentially distributed events (Table 1; see also HEY 1992). Divergence times showed a strong dependence on the parameters assigned in the simulations: approximately linear with effective population size and a power function of rate of mutation to new *S*-alleles (Figures 6

and 7). In contrast, the scaled ratios of divergence times appeared less variable (Table 2) and showed no discernible dependence on effective population size over a fourfold range or mutation rate over a 100-fold range (Table 3). Results of numerical simulations of over 151 independent coalescent episodes provided frequency distributions for the scaled indices of genealogical structure (Figure 4). While the means of all indices corresponded roughly to the approximate expected value of unity, two ($R_{ST}$ and $R_{SD}$) appeared to give consistently high values (Tables 2 and 3). These departures fell short of significance relative to the substantial variances of the distributions. To examine whether empirical observations are consistent with the null hypothesis of stable populations of constant size, I compared empirical estimates of the indices, not to the theoretical expectation of unity, but to frequency distributions generated by numerical simulation. The analysis appeared to be sufficiently sensitive to detect unusually long terminal branches in the *S*-allele genealogy constructed from nucleotide sequences observed in natural populations of a solanaceous plant (*S. carolinense*).

*Diagnosis of demographic history:* An ultimate goal is to develop a framework for using the structure of genealogies of neutral genes or of functionally distinct alleles maintained by symmetric balancing selection as a basis for the diagnosis of evolutionary process apart from the particular circumstances in which that process occurred. This study explores the distribution of the scaled ratios of divergence times under the null hypothesis of stochastic equilibrium under stable conditions. I hope to develop profiles of descriptors for various historical events, including changes in population size or mutation rate, that may generate significant deviations from the null hypothesis.

Scaled ratio $R_{PT}$ is related to TAJIMA's (1989a) *D*-statistic, which compares the average pairwise nucleotide differences ($\pi$) to the number of segregating sites ($K$). Under the assumption that all mutations occur at different sites (infinite sites model), the expectation of $\pi$ is proportional to twice the average pairwise divergence time ($P$) and $K$ proportional to total time in the genealogy ($T$). Values of $R_{PT}$ greater than unity correspond approximately to positive values of the *D*-statistic and less than unity to negative values. TAJIMA (1989a,b; 1990b; 1993) described effects of change in population size or of purifying or balancing selection on the *D*-statistic. To the extent that demographic changes induce discernible patterns in the scaled ratios or other descriptive indices, the approach explored in the present study may provide a means of recognizing the hallmarks of such events.

*Application to other symmetric evolutionary processes:* For the analysis of selectively neutral variation, the case to which the theoretical expectations in (1) in fact correspond, the method would likely require modification to generate expected distributions of divergence times among alleles (distinct sequences), conditional on the occurrence of the mutations that distinguish them. Further, because the processes of genetic turnover within and between allelic classes would occur on the same time scale, divergence time between two given allelic classes may show variation among sampled pairs of genes.

In addition, the method may be useful for the analysis of genealogical structure among class I and class II MHC alleles and other systems subject to strong symmetric balancing selection. Under the symmetric balancing selection generated by the expression of gametophytic self-incompatibility with codominant stylar expression of the rejection reaction, divergence time among *S*-allele classes depends on effective population size ($N$) and the rate of mutation to functionally different alleles ($\mu$). Under symmetric overdominance in viability, divergence time among functionally distinct alleles depends on the intensity of selection as well as $N$ and $\mu$ (TAKAHATA 1990; SASAKI 1992). Whether indices that show relatively low sensitivity to all three parameters can be developed remains unexplored.

**Application to gametophytic self-incompatibility:** *Contrasting genealogical structures:* RICHMAN *et al.* (1996) observed striking differences in genealogical structure among *S*-alleles sampled from natural populations of two self-incompatible solanaceous species. Their application of a modified form of TAKAHATA's (1993) method for inferring historical population size assumed that the two genealogies were representative of steady-state distributions under identical rates of mutation to new *S*-alleles. Comparison between the species of the number of segregating *S*-alleles and the number of *S*-allele lineages shared across genera indicated much larger long-term effective population size in *S. carolinense*, while the higher number of alleles maintained in *P. crassifolia* indicated larger short-term effective population size in *P. crassifolia*. They suggested that the *P. crassifolia* population had suffered a bottleneck in population size, during which many *S*-allele lineages were lost, with much of the presently segregating *S*-allele variation having been generated since that event.

Some aspects revealed by this analysis warrant further examination. Differences in effective population size alone would be expected to induce positive associations between *S*-allele number and divergence times (Table 3 and Figure 6), unlike the pattern observed between the species. Further, the low number of *S*-alleles maintained in *S. carolinense* appeared inconsistent with the large estimate of effective population size under the assigned rates of mutation. While the differences observed between the species appear to exceed the levels that might plausibly be attributed to estimation error, whether neither, either, or both *S*-allele genealogies deviate from expectation under the null hypothesis of

demographic stability with constant rates of mutation to new *S*-alleles remained equivocal.

Numerical simulations described here provide some indication of the extent of variation in genealogy structure consistent with the null hypothesis. This analysis suggests that it is the *S*-allele genealogy of *S. carolinense* rather than *P. crassifolia* that exhibits unusual structure: very long terminal branches relative to the rest of the tree (Table 2 and Figure 2), reminiscent of a star phylogeny. Undersampling of *S*-allele variation in natural populations does not appear to account for this characteristic.

*Differences in population size:* Population expansion promotes the generation of phylogenies with star-like structure (SLATKIN and HUDSON 1991). Genealogies retain this signature of population expansion until the next genetic turnover. Under the extreme balancing selection imposed by GSI, evidence of such events may persist for exceedingly long periods.

In view of the extensive sharing of *S. carolinense S*-allele lineages across genera, attributing the star-like structure of the *S. carolinense* genealogy to population expansion would entail hypothesizing a very ancient episode of diversification, prior to divergence between *S. carolinense* and *P. crassifolia.* That the *P. crassifolia S*-allele genealogy does not show a comparable pattern might reflect more rapid genetic turnover of *S*-allele lineages in *P. crassifolia.*

Lower rates of genetic turnover in *S. carolinense* may reflect higher long-term effective population size, as suggested by the analysis of RICHMAN *et al.* (1996). However, the low number of *S*-allele lineages maintained in *S. carolinense* appears inconsistent with this hypothesis. In addition, differences in population size alone would be expected to induce a positive association between allele number and divergence time (Table 3), contrary to the negative association observed between the species.

A survey of two sites, in Tennessee and North Carolina, provided the estimate of the number of *S*-alleles in *S. carolinense* (RICHMAN *et al.* 1995). If the *S*-alleles segregating at the two sites constitute independent samples of the *S*-allele variation in the species, then the nearly complete overlap between the samples indicates a low total number of *S*-alleles. If, however, the sites function as a single demographic unit in spite of their 250-km separation, then the species may maintain higher numbers of *S*-alleles. Further sampling of *S*-allele variation in *S. carolinense* populations over a more extensive geographical scale is planned. In any event, revision of the estimate of the total number of *S*-alleles would not affect the significance of the finding of unusually long terminal branches in the genealogy of the present sample of *S*-alleles.

*Differences in mutation rate:* While the analysis of RICHMAN *et al.* (1996) attributed differences in the pattern of *S*-allele variation between *S. carolinense* and *P. crassifolia* to differences in effective population size under a common rate of mutation to new *S*-alleles, a joint consideration of mutation rate and population size suggested a modest increase in population size relative to a reduction in mutation rate in *S. carolinense* (UYENOYAMA 1997). A difference in mutation rate is consistent with the negative association between allele number and divergence times observed between the species (Figures 7 and 8). Because GSI in the two solanaceous species studied appears to derive from the same RNase-based mechanism, a large difference in rate of mutation to new *S*-alleles may appear implausible. A difference in the rate of incorporation of new *S*-allele lineages, rather than mutation *per se,* might account for the observations.

At stochastic steady-state, incorporation of a new *S*-allele class into the population entails the loss on average of an existing *S*-allele class. If the new class replaces the parental class from which it was derived, the specificity of the *S*-allele lineage will have changed without bifurcation (see TAKAHATA *et al.* 1992). Processes that promote antagonistic interactions between parental and offspring *S*-allele classes would both reduce the number of distinct *S*-alleles segregating in the population and extend divergence times between *S*-allele lineages. Such antagonism may reflect lineage-specific mutational load: *S*-alleles associated with distinct arrays of recessive deleterious mutations. The extent to which *S*-alleles differ with respect to the number, position, and effects of associated mutations determines the level of variation in viability among *S*-allele genotypes.

Enforced heterozygosity at the *S*-locus may shelter recessive deleterious mutations that arise in closely linked regions from expression and purging. Upon the derivation by point mutation of a new *S*-allele class from an existing class, flanking deleterious mutations may be expressed in genotypes carrying both parental and offspring *S*-alleles. Viability selection would oppose the maintenance of both the parental and offspring *S*-allele classes. Elimination of the offspring class would preserve the *S*-allele lineage without a change in specificity, while elimination of the parental class would permit a change in specificity but without bifurcation of the lineage. This process would tend to promote two characteristics that distinguish the *S*-allele genealogy of *S. carolinense:* exclusion between parental and offspring classes would reduce allele number and a progressive increase in sheltering would cause a progressive retardation of branching among segregating *S*-allele lineages.

Sheltering of recessive deleterious factors may require tight linkage between the *S*-locus and loci affecting fitness. COLEMAN and KAO (1992) reported that the sequences of regions flanking *S*-alleles in Petunia (Solanaceae) are unalignable, suggesting large divergence times among regions both internal and external

to the coding region of the *S*-locus, perhaps reflecting some level of recombination suppression. Large tracts of DNA, up to hundreds of kilobases, appear to cosegregate with the *S*-locus complex in Brassica (BOYES *et al.* 1991, 1997; BOYES and NASRALLAH 1993).

Demographic events may influence the evolutionary consequences of sheltering. For example, a reduction in the number of segregating *S*-alleles during a population bottleneck or founder event would permit the fixation of deleterious mutations held in common by the surviving lineages. As a consequence, both variation in divergence time among *S*-alleles and the relative depression in viability suffered by zygotes that bear more recently diverged *S*-alleles may decline. These effects together with a strong selective pressure to restore *S*-allele variation after population expansion would favor incorporation of new *S*-alleles, even with their parental alleles. Such an interaction between lineage-specific mutational load and historical factors might generate the genealogical pattern observed for *P. crassifolia* *S*-alleles. Separate studies will address this hypothesis.

I am indebted to N. TAKAHATA, Y. SATTA, Y. ICHIKAWA, and all members of the academic and administrative staff at The Graduate University for generous hospitality, innumerable acts of kindness and grace, and essential assistance in all aspects of life and to M. SEKI for lifting the veil. X. VEKEMANS, A. G. CLARK, an anonymous reviewer, Y. LU, and J. STONE contributed many insightful comments and suggestions. A fellowship from the Japan Society for the Promotion of Science (Japan) and the John E. Fogarty International Center of the National Institutes of Health supported my visit to The Graduate University for Advanced Studies to develop this collaborative project with N. TAKAHATA and Y. SATTA. Public Health Service grant GM-37841 provided additional support.

## LITERATURE CITED

BOYES, D. C., and J. B. NASRALLAH, 1993 Physical linkage of the *SLG* and *SRK* genes at the self-incompatibility locus of *Brassica oleracea*. Mol. Gen. Genet. **236:** 369–373.

BOYES, D. C., C.-H. CHEN, T. TANTIKANJANA, J. J. ESCH and J. B. NASRALLAH, 1991 Isolation of a second *S*-locus-related cDNA from *Brassica oleracea:* genetic relationships between the *S* locus and two related loci. Genetics **127:** 221–228.

BOYES, D. C., M. E. NASRALLAH, J. VREBALOV and J. B. NASRALLAH, 1997 The self-incompatibility (*S*) haplotypes of Brassica contain highly divergent and rearranged sequences of ancient origin. Plant Cell **9:** 237–247.

BROOTHAERTS, W., G. A. JANSSENS, P. PROOST and W. F. BROEKAERT, 1995 cDNA cloning and molecular analysis of two self-incompatibility alleles from apple. Plant Mol. Biol. **27:** 499–511.

CLARK, A. G., and T.-H. KAO, 1994 Self-incompatibility: theoretical concepts and evolution, pp. 220–241 in *Genetic Control of Self-Incompatibility and Reproductive Development in Flowering Plants*, edited by E. G. WILLIAMS, A. E. CLARKE and R. B. KNOX. Kluwer Academic Publishers, Boston.

COLEMAN, C. E., and T.-H. KAO, 1992 The flanking regions of two *Petunia inflata* *S*-alleles are heterogeneous and contain repetitive sequences. Plant Mol. Biol. **18:** 725–737.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

FU, Y.-X., and W.-H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

HEY, J., 1992 Using phylogenetic trees to study speciation and extinction. Evolution **46:** 627–640.

HUANG, S., H.-S. LEE, B. KARUNANADAA and T.-H. KAO, 1994 Ribonuclease activity of *Petunia inflata* S proteins is essential for rejection of self-pollen. Plant Cell **6:** 1021–1028.

HUDSON, R. R., 1982 Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. J. FUTUYMA and J. ANTONOVICS. Oxford University Press, New York.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KINGMAN, J. F. C., 1982 On the genealogy of large populations. J. Appl. Prob. **19A:** 27–43.

KUBO, T., and Y. IWASA, 1995 Inferring the rates of branching and extinction from molecular phylogenies. Evolution **49:** 694–704.

LEE, H.-S., S. HUANG and T.-H. KAO, 1994 *S* proteins control rejection of incompatible pollen in *Petunia inflata*. Nature **367:** 560–563.

MCCLURE, B. A., J. E. GRAY, M. A. ANDERSON and A. E. CLARKE, 1990 Self-incompatibility in *Nicotiana alata* involves degradation of pollen rRNA. Nature **347:** 757–760.

MURFETT, J., T. L. ATHERTON, B. MOU, C. S. GASSER and B. A. MCCLURE, 1994 *S*-RNase expressed in transgenic Nicotiana causes *S*-allele-specific pollen rejection. Nature **367:** 563–566.

NEE, S., E. C. HOLMES, A. RAMBAUT and P. H. HARVEY, 1995 Inferring population history from molecular phylogenies. Phil. Trans. R. Soc. Lond. B **349:** 25–31.

DE NETTANCOURT, D., 1977 *Incompatibility in Angiosperms*. Springer-Verlag, Berlin.

RICHMAN, A. D., and J. R. KOHN, 1996 Learning from rejection: the evolutionary biology of single-locus self-incompatibility. Trends Ecol. Evol. **11:** 497–502.

RICHMAN, A. D., T.-H. KAO, S. W. SCHAEFFER and M. K. UYENOYAMA, 1995 *S*-allele sequence diversity in natural populations of *Solanum carolinense* (Horsenettle). Heredity **75:** 405–415.

RICHMAN, A. D., M. K. UYENOYAMA and J. R. KOHN, 1996 Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science **273:** 1212–1216.

SASAKI, A., 1992 The evolution of host and pathogen genes under epidemiological interaction, pp. 247–263 in *Population Paleo-Genetics*, edited by N. TAKAHATA. Japan Scientific Society Press, Tokyo.

SASSA, H., T. NISHIO, Y. KOWYAMA, H. HIRANO, T. KOBA *et al.* 1996 Self-incompatibility (S) alleles of the Rosaceae encode members of a distinct class of the T2/S ribonuclease superfamily. Mol. Gen. Genet. **250:** 547–557.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

TAJIMA, F., 1989a Statistical method for testing the neutral mutation hypothesis. Genetics **123:** 585–595.

TAJIMA, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

TAJIMA, F., 1990a Relationship between DNA polymorphism and fixation time. Genetics **125:** 447–454.

TAJIMA, F., 1990b Relationship between migration and DNA polymorphism in a local population. Genetics **126:** 231–234.

TAJIMA, F., 1993 Measurement of DNA polymorphism, pp. 37–59 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer, Sunderland, MA.

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

TAKAHATA, N., 1993 Evolutionary genetics of human paleo-populations, pp. 1–21 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer, Sunderland, MA.

TAKAHATA, N., and M. NEI, 1990 Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124:** 967–978.

TAKAHATA, N., Y. SATTA and J. KLEIN, 1992 Polymorphism and balancing selection at major histocompatibility complex loci. Genetics **130:** 925–938.

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and

their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

UYENOYAMA, M. K., 1995 A generalized least-squares estimate for the origin of sporophytic self-incompatibility. Genetics **139:** 975–992.

UYENOYAMA, M. K., 1997 The evolution of breeding systems, to appear in *Evolutionary Genetics from Molecules to Morphology,* edited by R. S. SINGH and C. KRIMBAS. Cambridge University Press, New York.

VEKEMANS, X., and M. SLATKIN, 1994 Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics **137:** 1157–1165.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

XUE, Y., R. CARPENTER, H. G. DICKINSON and E. S. COEN, 1996 Origin of allelic diversity in Antirrhinum S locus RNases. Plant Cell **8:** 805–814.

YOKOYAMA, S., and L. E. HETHERINGTON, 1982 The expected number of self-incompatibility alleles in finite plant populations. Heredity **48:** 299–303.

Communicating editor: A. G. CLARK

## APPENDIX

**Base branches:** Base branches bifurcate from the root of a genealogy. Expectations for the lengths of these branches rely only on the expectation of $t_i$, the time for coalescence from $i$ lineages to $i - 1$ lineages:

$$E[t_i] = \frac{4N}{i(i-1)}$$

for $N$ the effective population size (HUDSON 1982; KINGMAN 1982; TAJIMA 1983). The length of the shorter base branch ($B_1$) is $t_2$, the time for coalescence of two lineages into the root (first node). The length of the longer base branch ($B_2$) is determined by the time at which that branch first bifurcates.

Beyond the second node, $t_2$ generations after the root, three branches exist. If the third node occurs on the longer base branch (probability 1/3), $B_2$ is $t_2 + t_3$. Similarly, with probability $(1/j)\prod_3^{j-1}(1 - 1/k)$, the $j$th node ($j = 3, 4, \ldots, n-1$) in the genealogy of a sample of $n$ genes is the first to occur on the longer base branch; in this case, $B_2$ is $\Sigma_2^j t_i$. With probability $\prod_3^{n-1}(1 - 1/k)$, no nodes occur on one of the base branches and $B_2$ is $\Sigma_2^n t_i$. Using $E[t_i]$, the expectations of $B_1$, $B_2$, and the average base branch length [$B = (B_1 + B_2)/2$] are

$$E[B_1] = 2N$$

$$E[B_2] = E\left[\sum_{j=3}^{n-1}\sum_{i=2}^{j}(t_i/j)\prod_{k=3}^{j-1}(1-1/k) + \sum_{i=2}^{n} t_i \prod_{k=3}^{n-1}(1-1/k)\right]$$

$$= 8N\left[1/n + \sum_{3}^{n-1} 1/j^2\right]$$

$$E[B] = 4N\left[1/n + \sum_{2}^{n-1} 1/j^2\right].$$