Gene and Allelic Genealogies at a Gametophytic Self-Incompatibility Locus

Xavier Vekemans[1] **and Montgomery Slatkin**

Department of Integrative Biology, University of California, Berkeley, California 94720

Manuscript received January 6, 1994
Accepted for publication May 4, 1994

ABSTRACT

The properties of gene and allelic genealogies at a gametophytic self-incompatibility locus in plants have been investigated analytically and checked against extensive numerical simulations. It is found that, as with overdominant loci, there are two genealogical processes with markedly different time scales. First, functionally distinct allelic lines diverge on an extremely long time scale which is inversely related to the mutation rate to new alleles. These alleles show a genealogical structure which is similar, after an appropriate rescaling of time, to that described by the coalescent process for genes at a neutral locus. Second, gene copies sampled within the same functional allelic line show genealogical relationships similar to neutral gene genealogies but on a much shorter time scale, which is on the same order of magnitude as the harmonic mean of the number of gene copies within an allelic line. These results are discussed in relation to data showing *trans*-specific polymorphisms for alleles at the gametophytic self-incompatibility locus in the Solanaceae. It is shown that population sizes on the order of 4×10^5 and a mutation rate per locus per generation as high as 10^{-6} could account for estimated allelic divergence times in this family.

GAMETOPHYTIC self-incompatibility is a self-recognition mechanism that has been found in many families of flowering plants (DE NETTANCOURT 1977; CHARLESWORTH 1985). It is thought to have evolved as a means for regulating outcrossing distances in populations of hermaphroditic plants (UYENOYAMA 1988). In most systems studied, the incompatibility reaction is governed by a single gene with multiple alleles (called, respectively, *S* locus and *S* alleles). The rejection mechanism involves a recognition between the allele carried by the haploid pollen and either allele expressed in the diploid style. Numerous theoretical studies on gametophytic self-incompatibility systems have been undertaken to predict the number of alleles maintained at equilibrium in finite populations (for a review see CLARK 1993). These studies show that, because of the frequency-dependent selection induced by the incompatibility system, a larger number of alleles are maintained at an *S* locus than at a neutral locus with the same mutation rate (WRIGHT 1939). These predictions are consistent with the large numbers of *S* alleles typically found at *S* loci. For example, in natural populations between 30 and 45 different alleles have been found (ATWOOD 1944; EMERSON 1939; LAWRENCE and O'DONNELL 1981).

Recent molecular studies have led to the isolation of style glycoproteins that cosegregate with individual *S* alleles (KAMBOJ and JACKSON 1986). Clones of cDNA encoding the *S* allele-associated glycoproteins have been isolated and sequenced in several species of the family Solanaceae (see references in SIMS 1993). Comparisons of amino acid sequences of *S* alleles within and among species have led to the following generalizations: (1) there is a wide range of variation among alleles (between 38.7 and 93.5% identity); (2) levels of polymorphism within and among species are of the same order of magnitude; (3) there are numerous examples of pairs of alleles from different species that are more similar to each other than are functionally different alleles from the same species (IOERGER *et al.* 1990; CLARK and KAO 1991; CLARK 1993).

These observations suggest that *S* alleles exhibit shared polymorphisms (IOERGER *et al.* 1990), i.e. that some of the polymorphisms are more ancient than the times of species divergence. Shared polymorphisms have also been found in the mammalian major histocompatibility complex (MHC) and have been used to support the hypothesis that the high degree of polymorphism at those loci is maintained by overdominant selection (HUGHES and NEI 1988). This hypothesis is reinforced by the theory of TAKAHATA (1990) and TAKAHATA and NEI (1990) who showed that loci under balancing selection would have alleles with very long residence times and coalescence times that would far exceed those of neutral alleles. Similarly, the frequency-dependent selection induced by the incompatibility system has been used to explain the observations of long divergence times among *S* alleles (IOERGER *et al.* 1990; CLARK and KAO 1991; CLARK 1993). Moreover CLARK (1993) suggested that a one-locus gametophytic self-incompatibility system is formally similar to a locus under symmetric overdominant selection with complete homozygote lethality.

In this report we re-examine the problem of coalescence times at *S* loci, considering both allelic genealogies and gene genealogies within each allelic class. We develop analytic theory that describes the relationship

[1] Present address to which correspondence should be addressed: Université Libre de Bruxelles, Laboratoire de Génétique et d'Ecologie Végétales, Chaussée de Wavre 1850, B-1160 Bruxelles, Belgium.

between allelic genealogies at *S* loci and those for overdominant loci. Then we test our analytical predictions using extensive computer simulations. Finally, we discuss the model with reference to the observed properties of the *S* locus in the Solanaceae.

THEORY

Allelic genealogies at an S-locus: TAKAHATA (1990) developed the theory of allelic genealogies, *i.e.*, genealogies of functionally distinct alleles sampled from a finite population, for an overdominant locus under relatively strong selection. He showed that these genealogies are similar to neutral gene genealogies, *i.e.*, genealogies of randomly sampled genes at a neutral locus irrespective of their allelic type, but with a different time scale (see also TAKAHATA *et al.* 1992). Hence some analytic predictions obtained from the theory of neutral alleles can be applied to allelic genealogies for overdominant alleles after an appropriate rescaling of time, which is equivalent to a change in the effective population size. CLARK (1993) pointed out that TAKAHATA's approach could also apply to allelic genealogies at an *S* locus, because there are similarities in the diffusion approximations for overdominant models and for frequency-dependent models such as for gametophytic self-incompatibility systems (YOKOYAMA and NEI 1979). In both cases the stationary distribution of allelic frequencies shows a set of common alleles with similar frequencies, in contrast to the frequency spectrum for a neutral locus for which there is usually only one common allele and a number of rare alleles (MARUYAMA and NEI 1981; WRIGHT 1939).

The problem for the *S* locus is to find the appropriate rescaling factor for time. According to TAKAHATA's (1990) theory for overdominant alleles, time should be rescaled by a factor f_s, where

$$f_s = \tfrac{1}{2} n_c t(1/n_c) \tag{1}$$

$t(1/n_c)$ is the expected time until a particular common allele, starting at frequency $1/n_c$, becomes lost and n_c is the number of common allelic lines at equilibrium. We can find expressions for these two quantities for an *S* locus as follows.

We assume a population of *N* diploid individuals with a one-locus gametophytic self-incompatibility system. WRIGHT (1939) and FISHER (1958) used a diffusion approximation to study the conditions of equilibrium for this model under mutation, selection and random genetic drift. Assuming the infinite-allele model of mutation, the drift and diffusion coefficients for the model are as follows (WRIGHT, 1960, p. 70)

$$E(\delta x) = -ax(x - F) - ux \tag{2}$$

$$\mathrm{Var}(\delta x) = x(1 - 2x)/2N \tag{3}$$

where *u* is the mutation rate per gene per generation, $F = \sum_i x_i^2$ is the homozygosity, and $a = 1/[(1 - F)(1 - 2F)]$.

The stationary distribution of allelic frequencies is given by

$$\Phi(x) = 4Nue^{2Nax}(1 - 2x)^{2Nb-1}x^{-1} \tag{4}$$

where $b = 1/[2(1 - F)] + u$ (YOKOYAMA and NEI 1979).

The homozygosity, *F*, and, consequently, the coefficients *a* and *b* can be obtained by using the relationship $\int_0^{1/2} x\Phi(\mathrm{x})\,\mathrm{dx} = 1$, hence by solving for *F* in

$$u\sqrt{8\pi N}\exp\left(\frac{2NF}{(1 - F)(1 - 2F)}\right) = (1 - F)^{-1/2}(1 - 2F)^{-M[1/(1-F)+2u]} \tag{5}$$

(YOKOYAMA and HETHERINGTON 1982).

The expression for $t(1/n_c)$, the expected time until a particular common allele starting at frequency $1/n_c$ becomes lost, has been derived for overdominant alleles by TAKAHATA (1990) by using the diffusion equation

$$t(x) = 2\left[\int_0^x \frac{dy}{\mathrm{Var}(\delta y)\Psi(y)} \int_0^y \Psi(z)\,dz + \int_x^1 \frac{dy}{\mathrm{Var}(\delta y)\Psi(y)} \int_0^x \Psi(z)\,dz\right]$$

with $\Psi(y) = \exp[-2\int^y \{E(\delta z)/\mathrm{Var}(\delta z)\}\delta z]$ (EWENS 1979, p. 119).

Using (2) and (3), we have, for gametophytic self-incompatibility,

$$\Psi(y) \approx \exp[2Na\{x - (F - u/a)\}^2] \tag{6}$$

This expression is the same as the one obtained by TAKAHATA (1990) if we replace his coefficients *S* and *m* by $2Na$ and $F - u/a$, respectively. Following his derivation but using the coefficients appropriate for *S* loci we get

$$t\left(\frac{1}{n_c}\right) = \frac{\sqrt{2}}{8N^2ua(F - u/a)^2} \tag{7}$$

In (7), *F* is itself a function of *N* and *u* but we will treat it as a separate parameter for now and later find its value by solving (5).

To complete the derivation for *S* loci, we need an expression for the average number of common alleles at equilibrium, n_c. This quantity is approximately equal to the effective number of alleles in the population (KIMURA and CROW 1964) and hence $n_c \approx 1/F$, with *F* computed by solving (5).

Finally, using (1), the formula for the rescaling factor becomes

$$f_s = \frac{\sqrt{2}}{16N^2uaF(F - u/a)^2} \tag{8}$$

We should note that, as discussed by TAKAHATA (1990),

TABLE 1

Values of the homozygosity, F, the coefficient a, and the rescaling factor, f_s, for a gametophytic self-incompatibility locus and an overdominant locus with lethal homozygotes when $u = 10^{-6}$

N	F_{Slocus}[a]	$F_{\text{Overd.}}$[b]	a[c]	$f_{s_{\text{S locus}}}$[d]	$f_{s_{\text{Overd.}}}$[e]
10^2	0.159	0.222	1.746	1252	802
10^3	0.059	0.066	1.205	355	305
10^4	0.0188	0.0194	1.059	125	119
10^5	0.0056	0.0057	1.017	48.9	48.3
10^6	0.0016	0.0016	1.005	20.6	20.5

[a] Computed by solving for F in (5).
[b] Computed according to (11).
[c] Computed as $1/[(1 - F)(1 - 2F)]$.
[d] Computed according to (8).
[e] Computed according to (12).

this approach provides valid approximations only when $\sqrt{2Na/Nu} \geq 100$, *i.e.* only when mutation rates are low.

The formula derived for a neutral gene genealogy described by the coalescence process (KINGMAN 1982) can now be applied to the allelic genealogy at an S locus after multiplication by f_s. As an example, following the notation of TAKAHATA (1990), the average pairwise divergence times between alleles can be computed as

$$E\{T_d\} \approx 2Nf_s \tag{9}$$

with the time expressed in generations. Similarly the mean time to coalescence of all alleles in the population is given by

$$E\{T_c\} \approx 4Nf_s(1 - 1/n_e) \approx 4Nf_s(1 - F). \tag{10}$$

Comparison of allelic genealogies under gametophytic self-incompatibility and overdominance: The scaling factor we find for an S locus is somewhat different from that suggested by CLARK (1993) because we have taken explicit account of the dynamics of S alleles. CLARK (1993) based his suggestion on the analysis of YOKOYAMA and NEI (1979) and YOKOYAMA and HETHERINGTON (1982), who showed that when $4Nu \gg 1$, the equilibrium value of F for an S locus is given by

$$F = [-\ln(u\sqrt{8\pi N})/2N]^{1/2} \tag{11}$$

which is the same as for an overdominant locus with lethal homozygotes (formula 5 from TAKAHATA 1990 with $s = 1.0$). Under these conditions the formula for f_s becomes

$$f_s = \sqrt{2N}/2Nu[\ln(1/8\pi Nu^2)]^{-3/2} \tag{12}$$

(CLARK 1993). Table 1 compares values for both F and f_s for our approximation and that of CLARK (1993). We confirm that his approximation is valid for sufficiently large N ($N > 10{,}000$) but that for small population sizes there is a substantial difference. In effect, selection at S loci in small populations is stronger than can possibly be obtained with overdominant loci affecting viability because selection at S loci is discriminating among heterozygotes, in favour of those heterozygotes with low frequency alleles. In larger populations, however, many alleles are maintained at equilibrium so that most alleles are found at low frequencies.

SIMULATIONS

Methods of computer simulation: We simulated reproduction in a diploid plant population of size N with non-overlapping generations. In each generation progeny were produced by randomly choosing one of the $2N$ genes as the female gamete, and one as the male gamete (pollen), and then checking for compatibility, *i.e.*, testing if the allele carried by the pollen is different than each allele carried by the female parent. If the pollen is compatible then a new zygote is formed. If the pollen is incompatible then the female gamete is retained and new pollen randomly chosen until a compatible one is found, as described by MAYO (1966). The process was repeated until N new zygotes were produced. A given number of mutations, drawn from a Poisson distribution with mean $2Nu$, were then applied to randomly selected genes. Each mutation was assumed to produce a new functional allelic type, according to the infinite-allele model of mutation. Each run was started with $2N$ different alleles and, in order to reach the equilibrium distribution of allelic frequencies, was continued until the effective number of alleles in the population had crossed the expected value ($1/F$) a given number of times (20 in our runs). Beginning at that time genealogical information was recorded.

The genealogy was tracked by building a genealogical tree, generation after generation in a forward manner, with each node representing a gene with its generation number, allelic type and pointers to ancestral and descendant nodes. After each generation, lineages which were not represented in the last generation (extinct lineages) were removed from the tree. The time to fixation was recorded when only one gene from the first generation remained present. Then simulations were continued for a time randomly chosen between 2 and 6 times the observed time to fixation. At this stage, the tree represents a random gene genealogy with $2N$ genes at its tips. The corresponding allelic genealogy is constructed by counting the number of distinct allelic lines in the last generation, *i.e.*, the number of tips in the allelic genealogy, and by removing from the tree all but one gene from each allelic line.

The average divergence time between different alleles for all pairwise comparisons (T_d) and the coalescence time of all alleles in the population (T_c) are then computed according to TAKAHATA and NEI (1990). The whole process was repeated for 1000 replicates for each set of parameter values. The program was tested by simulating a neutral locus and checking the results against the expectations for neutral gene genealogies.

Numerical results for allelic genealogies: The results of some simulations performed with population sizes of

TABLE 2

Numerical results and analytical expectations (in parentheses) for the mean actual number of alleles, n_a, the mean heterozygosity, H, the mean coalescence time of all alleles, T_c, and the average pairwise divergence time between alleles, T_d at a gametophytic self-incompatibility locus

$4Nu$	N	u	n_a	H	T_c^a	T_d^a
0.004	50	2×10^{-5}	5.23 ± 0.52 (5.86)	0.796 ± 0.018 (0.818)	475.8 ± 284.4 (496.4)	303.6 ± 164.8 (303.4)
	100	1×10^{-5}	6.75 ± 0.69 (7.39)	0.841 ± 0.014 (0.855)	559.7 ± 284.6 (606.5)	338.1 ± 155.7 (354.7)
	200	5×10^{-6}	8.83 ± 0.78 (9.49)	0.879 ± 0.009 (0.887)	669.2 ± 379.6 (754.8)	386.3 ± 194.6 (425.5)
0.04	50	2×10^{-4}	6.29 ± 0.84 (6.99)	0.821 ± 0.020 (0.840)	86.1 ± 50.7 (83.2)	51.9 ± 27.6 (49.5)
	100	1×10^{-4}	8.16 ± 0.94 (8.85)	0.863 ± 0.014 (0.874)	95.4 ± 50.2 (100.0)	55.8 ± 26.1 (57.2)
	200	5×10^{-5}	10.66 ± 1.09 (11.37)	0.895 ± 0.009 (0.902)	120.2 ± 61.3 (122.5)	67.0 ± 31.5 (67.9)
0.4	50	2×10^{-3}	9.08 ± 1.44 (9.59)	0.859 ± 0.018 (0.872)	20.3 ± 11.8 (19.5)	11.7 ± 6.1 (11.2)
	100	1×10^{-3}	11.67 ± 1.65 (12.25)	0.890 ± 0.012 (0.899)	21.8 ± 11.3 (22.1)	12.3 ± 5.6 (12.3)
	200	5×10^{-4}	15.23 ± 1.89 (15.71)	0.916 ± 0.008 (0.921)	27.4 ± 14.1 (25.8)	15.0 ± 6.8 (14.0)
4.0	50	2×10^{-2}	19.49 ± 2.76 (15.42)	0.913 ± 0.014 (0.928)	7.34 ± 4.10 (22.6)	3.90 ± 1.92 (12.15)
	100	1×10^{-2}	25.39 ± 3.19 (21.66)	0.931 ± 0.009 (0.938)	8.07 ± 4.07 (14.9)	4.32 ± 2.00 (7.94)
	200	5×10^{-3}	31.83 ± 3.84 (28.84)	0.945 ± 0.007 (0.949)	9.45 ± 4.69 (12.6)	4.89 ± 2.18 (6.64)

Mean values are given ± SD across 1,000 replicates. Analytical expectations for n_a, H, T_c, T_d, computed according to Equations 4, 5, 10 and 9, respectively.

[a] Times are given in units of N generations.

50, 100 and 200 with various mutation rates are given in Table 2. Values observed for the mean actual number of alleles in the population (n_a) and the mean overall heterozygosity (H) are increasing with N and u. The observed values are close to the expectations computed according to YOKOYAMA and NEI (1989), using the value for F obtained by solving equation (5). Our results are in agreement with simulation results in previous studies (EWENS and EWENS 1966; KIMURA 1966; CLARK 1993) and they confirm the validity of the diffusion approximations for this model. The mean coalescence time of all alleles (T_c) and the average pairwise divergence times between alleles (T_d), both expressed in units of N generations, are increasing with N but decreasing with u. The observed values are remarkably close to the expected values obtained using Equations 10 and 9, respectively, except for the case with the highest mutation rate ($4Nu$ = 4), for which the validity of the theory is no longer assured because $\sqrt{2Na/Nu} < 25$.

CLARK (1993, Table II) reported numerical values of T_c and T_d averaged over 100 replicates for the case of $N = 100$. These were interpreted as supporting the hypothesis that gametophytic self-incompatibility is similar to overdominance with homozygote lethality. Although the order of magnitude of the results are similar to ours, some discrepancies appear. For example, his values of T_d for u ranging from 10^{-2} to 10^{-4} are all higher than ours, but the reverse is true for $u = 10^{-5}$. This is of particular interest because his simulation results for $u = 10^{-5}$ are close to the expectation under overdominant selection with $s = 1.0$ whereas our observed value is closer to the expected value according to (8). We think that the difference between CLARK's simulation results and those presented here can be accounted for by a bias in CLARK's simulations, and thus that the fit to the overdominance model is purely coincidental. The bias arises for the following reason. In our simulations the average time to coalescence of all genes in a population with $N = 100$ and $u = 10^{-5}$ was approximately 56,000 generations, and the determination of T_c and T_d was done on average after 170,000 generations. If the simulation is stopped too soon after the first coalescence event then T_c and T_d will be underestimated because the genealogy is shorter than a random genealogy (data not shown, see also TAJIMA 1990). All simulations in CLARK (1993) were stopped after 40000 generations so an underestimation of the coalescence times for $u = 10^{-5}$ seems likely. For the other values of u studied by CLARK (1993), there is little bias because the fixation times are shorter, and hence his results are consistent with our analytical expectations.

Figure 1 illustrates the differences in T_c between a neutral locus and an S locus, as obtained by simulations for a population with $N = 100$ and mutation rates varying between 10^{-2} and 10^{-5}. For high values of $4Nu$ the time scale of the allelic genealogies at the S locus is of the same order of magnitude as for a neutral locus. However, as $4Nu$ decreases, the coalescence times in the allelic genealogy increase dramatically for the S locus while they decrease for the neutral locus. TAKAHATA (1991) derives formula for the coalescent properties of allelic genealogies at a neutral locus showing that T_c is proportional to Nu. For a gametophytic self-incompatibility locus, however, successive mutation events within an allelic lineage are necessary for the spread of the lineage because the frequency of individual alleles is constrained by a frequency-dependent mechanism. As a result, the time scale of the allelic genealogies will increase with decreasing values of the mutation rate.

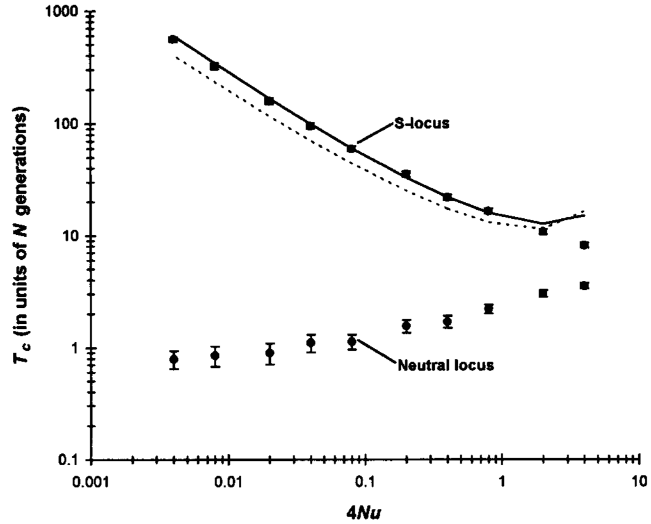

FIGURE 1.—Comparison of the coalescence times of all alleles in the population (T_c) between a neutral locus and a gametophytic self-incompatibility locus, as computed by simulation for a population with $N = 100$ and mutation rates between 10^{-2} and 10^{-5}. The analytical expectations for T_c, given by equation (10) are shown with f_s computed according to (8), solid curve, and to (12), dashed curve. Times are given in units of N generations. Error bars indicate 99.9% confidence intervals over 1,000 replicates.

Figure 1 also shows the analytical predictions for T_c computed according to Equation 10, using f_s calculated from Equation 8 (solid curve) and f_s calculated for the overdominant model with lethal homozygotes (after Equation 12; dashed curve). Once again, with an exception for high mutation rates where this approach is not expected to apply, the numerical results indicated with error bars in Figure 1 strongly support the use of Equation 8.

Further support for our theory is given by analyzing the genealogical structure of samples of S alleles from our simulations. We computed the time intervals between successive coalescent events in random samples of alleles. For a random sample of neutral genes, the expectation for each time interval is $4N/i(i-1)$ with i being the number of branches in the genealogy during that interval (HUDSON 1983; TAJIMA 1983). According to our results this formula can be applied to an S allele genealogy after multiplication by the rescaling factor f_s, which gives

$$T(i) = 4Nf_s/i(i-1) \qquad (13)$$

with $T(i)$ being the time interval during which there are i distinct lineages in the allelic genealogy. Figure 2 shows the numerical results for these time intervals, averaged over 1000 replicates, for random samples of ten common alleles drawn from simulation runs with $N = 500$ and $u = 0.0005$ ($n_a = 26.8$; $H = 0.951$; $n_c = 20.2$; $f_s = 5.08$). We used a larger population size than before so that sampling theory will apply. The solid line shows the analytical expectations using f_s as obtained through (8) and the broken line represents the expectations under the overdominant model with lethal homozygotes (using Equation 12). The simulation results shown with error bars in Figure 2 support our derivation of f_s for gametophytic self-incompatibility. That there is close agreement with the expectations supports the conclusion of TAKAHATA (1990) that the topology of an allelic genealogy under balancing selection is similar to that of a neutral gene genealogy but with a different time scale.

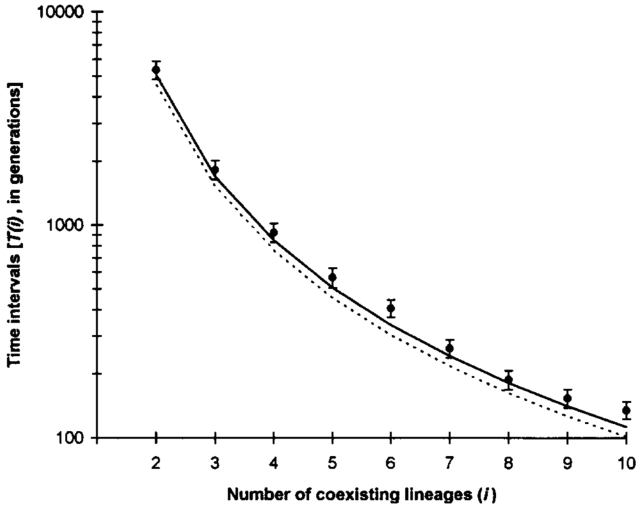

FIGURE 2.—Time intervals, in units of generations, between successive coalescent events in a random sample of ten alleles at a gametophytic self-incompatibility locus for a population with $N = 500$ and $u = 0.0005$. Comparison of simulation data with analytical predictions given by equation (13) with f_s computed according to (8), solid curve, and (12), dashed curve. Error bars indicate 99.9% confidence intervals over 1,000 replicates.

Genealogies of gene copies sampled within the same allelic line: According to the model of gametophytic self-incompatibility used here, different gene copies of the same functional allele have identical reproductive success. Thus each allelic line can be envisioned as a population of neutral gene copies. We would then expect that a gene genealogy based on a sample of gene copies from the same allelic line resemble a neutral gene genealogy and this could provide a test of the model if appropriate sequence data become available. For a neutral locus the properties of gene genealogies under nested subsampling have been studied by HUDSON and KAPLAN (1986). For the overdominant model TAKAHATA (1990) gives the following treatment of this problem in the case of strong selection, large population size and low mutation rate. We have n_c common allelic lines in the population with on average $2N_c = 2N/n_c$ gene copies belonging to each allelic line. TAKAHATA (1990) has called the quantity $2N_c$ the "effective gene number" because it is the average number of copies of each allelic line. The crude expectations for the coalescence time of all gene copies, T_{cw}, and for the average pairwise divergence time between gene copies within individual lines, T_{dw} are then, respectively, $4N_c(1 - 1/2N_c)$ and $2N_c$ ac-

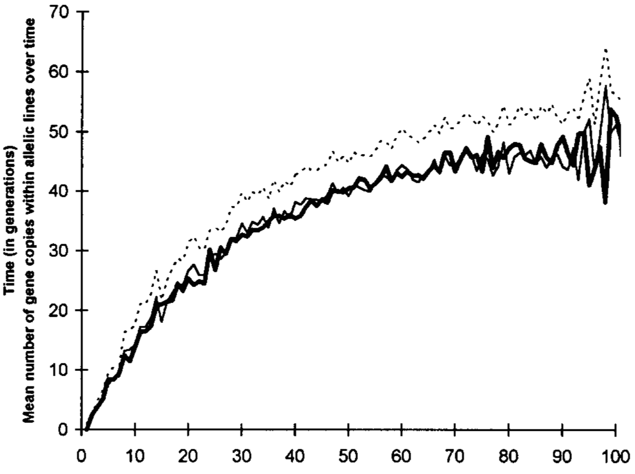

FIGURE 3.—Variation in the time scale of genealogies of different gene copies of the same allelic line as a function of the allelic frequency at the time of sampling. All allelic lines produced during 1,000 runs with $N = 500$ and $u = 0.0005$ were categorized according to their frequency at sampling time and the average of the following statistics were computed within each category: the average pairwise divergence time between gene copies sampled within the same allelic line, T_{dw} (bold solid curve), in units of generations; the harmonic (light solid curve) and arithmetic (dashed curve) means of the number of gene copies of each allelic line over time.

cording to coalescence theory (KINGMAN 1982). However TAKAHATA (1990) did not verify these predictions using simulation data. Because of the genealogical information provided by our simulations we were able to address this problem numerically. We tested the observed time scales of the genealogies of gene copies within allelic lines against the neutral expectation, although we didn't simulate neutral mutations arising within allelic lines. Within each common allelic class, we computed the coalescence time of all gene copies, T_{cw}, the average pairwise divergence time between gene copies, T_{dw}, and noted the number of gene copies of the allele at the time of subsampling. We report the results for $N = 500$ and $u = 0.0005$ ($n_c = 20.2$ and $2N_c = 2N/n_c = 49.5$). The results, averaged over all 30,174 allelic lines produced during 1,000 runs, are as follows: $T_{cw} = 66.5 \pm 0.2$ (SE) generations; $T_{dw} = 36.1 \pm 0.1$ generations. Using the expectations given above these values for T_{cw} and T_{dw} give estimates of the average number of gene copies within allelic lines of 34.3 ± 0.1 (SE) and 36.1 ± 0.1, respectively. Both of these values are lower than 49.5, the expected value of $2N_c$. Because pooling allelic lines with different frequencies can be misleading we computed the same statistics but averaged only over allelic lines having the same number of gene copies at the time of sampling (Figure 3). The results show that the time scales of the gene genealogies are increasing with the number of gene copies. For the 557 allelic lines with exactly $2N_c$ gene copies, the results are as follows: $T_{cw} = 73.2 \pm 1.7$ (SE) generations; $T_{dw} = 39.9 \pm 0.8$ generations. The computed estimates of the average number of gene copies based on T_{cw} and T_{dw} are 37.6 ± 0.9 (SE) and 39.9 ± 0.8, respectively. Again both of these values are lower than the expected value of 49.5.

We suggest that two factors could account for these discrepancies. First some allelic lines of a given frequency would have recently evolved. Because the initial spread of a new allele is very fast (10 generations on average for becoming a common allele for $N = 500$ and $u = 0.0005$) the genealogy of such lines should be similar to genealogies under rapid population growth (SLATKIN and HUDSON 1991) and thus would lead to shorter coalescence times. However the average age of a common allele is high (488 ± 3 (SE) generations) and only a small fraction of the alleles ($\approx F = 4.4\%$) are expected to have evolved recently, so that this could only partially explain the smaller coalescence times observed. The second factor is that the number of gene copies within an allelic line is not constant but is fluctuating under the effects of selection and genetic drift. When population sizes are varying through time, the effective population size is the harmonic mean of population sizes over time (WRIGHT 1938). We computed the harmonic and arithmetic means over generations of the number of gene copies of each allelic line until the time of common ancestry of the lineage. The results are as follows: the arithmetic mean is 41.8 ± 0.1 (SE) gene copies when averaged over all 30,174 common allelic lines and is 46.0 ± 0.5 gene copies when averaged only over the 557 allelic lines with $2N_c$ copies at sampling time; the harmonic mean values are respectively 36.5 ± 0.1 (SE) and 40.6 ± 0.5 gene copies in the first and second cases. These values for the harmonic mean of the number of gene copies are close to the computed estimates of the average number of gene copies based on T_{cw} and T_{dw} given above, while the arithmetic means are higher. Moreover, harmonic means seem to follow reasonably well the pattern of variation in T_{dw} with allelic frequency as shown in Figure 3.

These results support the suggestion that the gene genealogies within allelic lines have a time scale on the same order of magnitude as the harmonic mean of the number of gene copies within an allelic line. Moreover, they show that these time scales will depend on the frequency of the allelic line at the time of subsampling. This observation had already been reported by HUDSON and KAPLAN (1986) for nested subsampling at a neutral locus and points to the importance of having a knowledge of the allelic frequencies when testing sequence data from subsamples of gene copies within allelic lines.

Now that we know the effective size of populations of gene copies within allelic lines we can test the structure of their genealogies against the neutral expectation, although we cannot completely take into account the fluctuations in population size that we have found are important. We applied TAJIMA's (1989) test of the neutral mutation hypothesis on nucleotide sequences generated on the basis of the gene genealogies produced by our simulations. Using the genealogies of samples of ten gene copies within allelic lines, as described above, we generated sequences starting from the most recent common ancestor by applying the JUKES-CANTOR's model of evolution of nucleotide sequences (JUKES and CANTOR 1969). The JUKES-CANTOR's model assumes that substitutions occur randomly among the four types of nucleotides and that the rate of substitution in each of the three possible directions of change is α. The generated sequences were 1,000 nucleotides long and the value of α taken as 0.0333. A high substitution rate was necessary in order to generate enough segregating sites in these genealogies within such short time scales ($T_{cw} \approx 65$ generations). Then we computed the number of segregating sites, S, the average number of pairwise nucleotide differences between sequences, k, and TAJIMA's statistic D. Our results, averaged over 1091 allelic lines, for $N = 500$ and $u = 0.0005$ are as follows: $S = 21.04 \pm 0.35$ (SE); $k = 6.65 \pm 0.13$; $D = -0.65 \pm 0.02$; with 1025 out of 1091 (94.0%) tests not significantly different than the null hypothesis of neutrality ($P < 0.05$). This high proportion of non-significant tests was not found to be sensitive to changes in the value of α and the sequences were long enough so that there was a very low probability that the same site would mutate twice.

TRANS-SPECIFIC POLYMORPHISMS IN THE SOLANACEAE

Nucleotide sequences of S alleles from several species of the Solanaceae suggest that some of the S locus polymorphism existed prior to their divergence and has been maintained to the present (IOERGER *et al.* 1990). The time of divergence between these species has been estimated at 30 to 40 million years ago, using sequence data from ribulose bisphosphate carboxylase and an estimate of the synonymous substitution rate of 6.6×10^{-9} substitutions per site per year (references in IOERGER *et al.* 1990). Figure 4A shows values of T_c, the coalescence time of all alleles, computed according to Equation 10 for a range of population sizes and mutation rates. We have no information on population sizes but a very crude estimate of u, the mutation rate to new functional alleles per locus per generation, can be calculated as follows (TAKAHATA 1990).

Assume that we know the number of amino acids involved in the recognition process, that any change in these amino acids produces a new functional allele, and that the substitution rate per site is similar for synonymous and non-synonymous sites and is equal to 6.6×10^{-9} substitutions per site per year. Then u = no. of amino acids $\times 3 \times 0.8 \times 6.6\ 10^{-9}$, because on average 80% of nucleotide substitutions are non-synonymous. By studying the variation in substitution rates within the S locus for several species of Solanaceae, CLARK and KAO (1991) showed that the amino acids between position 40 and 80 had a higher non-synonymous substitution rate than other regions of the locus. This region is thought to be a good candidate for a recognition site determining S allele specificity (SIMS 1993) because it includes two hydrophilic hypervariable domains and it is flanked by two highly conserved ribonuclease active site regions as well as by two cysteine residues which form a disulfide bridge. Taking, as a crude approximation, the number of selected amino acids as 40, and assuming a generation time of one year, we have $u = 6.4\ 10^{-7}$ per locus per generation.

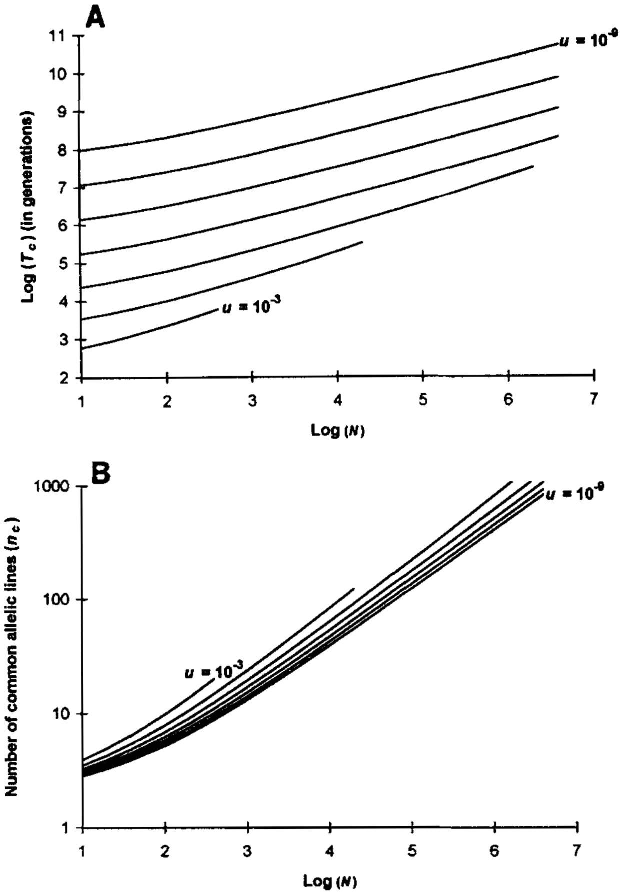

FIGURE 4.—Values of (A) the coalescence time of all alleles at a gametophytic self-incompatibility locus, T_c, and (B) of the number of common allelic lines, n_c, plotted against the population size, N, for a range of mutation rates per locus per generation, u. T_c values are given in units of generations and are computed after equation (10); n_c is computed as $1/F$ with F obtained through Equation 5.

This is probably an upper bound for u because it has been suggested that more than one mutation would be required to change allelic identity (CLARK 1993). From Figure 4A, one can see that for $u = 10^{-6}$ and 10^{-7}, population sizes of the order of 4×10^5 and 2×10^4, respectively, are compatible with divergence times around 40 million years. Also it appears that the time scales of the allelic genealogies are more sensitive to changes in u than to changes in N. Figure 4B shows the number of common allelic lines maintained at equilibrium within populations, n_c, for the same range of parameter values. Figure 4B shows that the number of alleles, unlike the coalescence times, are more sensitive to changes in N than to changes in u. For $u = 10^{-6}$ and $N = 4 \times 10^5$, the expected value for n_c is 375; and it is 66 for $u = 10^{-7}$ and $N = 2 \times 10^{4.}$ We do not yet have any estimate of the number of different alleles at the species level so that we cannot comment on these values. Nevertheless, there are a few estimates within populations which showed 30–45 alleles in populations of actual size of the order of thousands of individuals (Atwood 1944; Emerson 1939; Lawrence and O'Donnell 1981). It is important to note that these numbers represent the actual numbers of alleles in the samples. Using the original data from the three studies cited above, we computed the effective number of alleles [as $1/F$, using the correction of Nei (1978) for small sample sizes] and found values between 13 and 20 [$\bar{x} = 16.8 \pm 2.4$ (SD) for six samples]. By looking at Figure 4b one can see that these values, if used as estimations of n_c, are compatible with a mutation rate of the order of 10^{-6}–10^{-7} per locus per generation and $N = 1000$.

Moreover we note that for $N = 1000$ and u varying between 10^{-6} and 10^{-7}, we expect the coalescence time of all alleles to be of the order of 10^6 to 10^7 generations (Figure 4A) so that a substantial divergence among alleles can occur even within relatively small isolated populations. This result is noteworthy because a recent investigation of the S allele sequence diversity in a wild tomato relative (*Lycopersicon peruvianum* Mill.), using Southern-hybridization methods, detected high levels of sequence divergence both within and among natural populations (Rivers *et al.* 1993).

Referring to the genealogies of gene copies sampled within the same allelic line, and taking $N = 4 \times 10^5$ and $u = 10^{-6}$ (and thus $n_c = 375$), we can compute an estimate of their time scales as the effective gene number $(2N_c) = 2 \times 4 \times 10^5/375 = 2133$. This number represents an estimate of the average pairwise divergence time in generations between gene copies of an allelic line. It illustrates the important difference in time scale of the two genealogical processes occurring at the S locus. On the one hand, distinct allelic lines are diverging on a time scale of 10^7 to 10^8 generations, leading to *trans*-specific polymorphism, whereas, on the other hand, all gene copies within the same allelic line are coalescing within 4,000 generations. In comparison, genes at a neutral locus would coalesce within $4N = 1.6 \times 10^6$ generations and thus are not expected to show any excess of shared polymorphism among the species investigated.

CONCLUSIONS

To conclude, our results show that for gametophytic self-incompatibility, genealogies of random samples of S alleles, as well as of random samples of gene copies within a particular allelic line, have a genealogical structure similar to genealogies of neutral genes but with different time scales. Although the model seems to account for both the transspecific polymorphism of S alleles within the Solanaceae, and the occurrence of high levels of sequence divergence at the population level in wild tomato, there are still insufficient data to provide a thorough test. One way to test the validity of the model would be to randomly sample alleles within populations and study their genealogical structure using molecular sequences and phylogenetic inference. However, as pointed out by Clark and Kao (1991) the great level of divergence among sequences both for synonymous and non-synonymous substitutions would make such analyses difficult. It must also be noted that the effect of two additional complications, *i.e.*, population subdivision and intragenic recombination, have not yet been explicitly incorporated into the model. As theory refines, methods of analysis improve and more data are gathered, we think that the study of genealogies of S alleles will give valuable insights into the population genetics and paleo-genetics of Angiosperm species with gametophytic self-incompatibility.

This work was supported by a N.A.T.O. postdoctoral fellowship to X.V. and a grant from the National Institutes of Health to M.S. We thank A. G. Clark, M. Uyenoyama, S. Otto and J. Wakeley for helpful discussions on this topic. We also thank D. Charlesworth and two anonymous reviewers for many improvements in the manuscript.

LITERATURE CITED

Atwood, S. S., 1944 Oppositional alleles in natural populations of *Trifolium repens*. Genetics **29:** 428–435.

Charlesworth, D., 1985 Distribution of dioecy and self-incompatibility in angiosperms, pp. 237–268 in *Evolution, Essays in Honour of John Maynard Smith*, edited by P. J. Greenwood, P. H. Harvey and M. Slatkin. Cambridge University Press, Cambridge.

Clark, A. G., 1993 Evolutionary inferences from molecular characterization of self-incompatibility alleles, pp. 79–108 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, Mass.

Clark, A. G., and T.-H. Kao, 1991 Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc. Natl. Acad. Sci. USA **88:** 9823–9827.

De Nettancourt, D., 1977 *Incompatibility in Angiosperms*. Springer, New York.

Emerson, S., 1939 A preliminary survey of the *Oenothera organensis* population. Genetics **24:** 524–537.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Ewens, W. J., and P. M. Ewens, 1966 The maintenance of alleles by mutation-Monte Carlo results for normal and self-sterility populations. Heredity **21:** 371–378.

Fisher, R. A., 1958 *The Genetical Theory of Natural Selection*, second revised edition. Dover Publications, New York.

Hudson, R. R., 1983 Testing the constant-rate neutral model with protein sequence data. Evolution **37:** 203–217.

Hudson, R. R., and N. L. Kaplan, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–170.

Ioerger, T. R., A. G. Clark and T.-H. Kao, 1990 Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc. Natl. Acad. Sci. USA **87:** 9732–9735.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.

KAMBOJ, R. K., and J. F. JACKSON, 1986 Self-incompatibility alleles control a low molecular weight basic protein in pistils of *Petunia hybrida*. Theor. Appl. Genet. **71:** 815–818.

KIMURA, M., 1966 Simulation studies on the number of self-sterility alleles maintained in a small population. Annu. Rept. Natl. Inst. Genet. Jpn. **16:** 86–88.

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.

LAWRENCE, M. J., and S. O'DONNELL, 1981 The population genetics of the self-incompatibility polymorphism in Papaver rhoeas. III. The number and frequency of S-alleles in two further natural populations (R102 and R104). Heredity **47:** 53–61.

MARUYAMA, T., and M. NEI, 1981 Genetic variability maintained by mutation and overdominant selection in finite populations. Genetics **98:** 444–459.

MAYO, A., 1966 On the problem of self-incompatibility alleles. Biometrics **22:** 11–120.

NEI, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89:** 583–590.

RIVERS, B. A., R. BERNATZKY, S. J. ROBINSON and W. JAHNEN-DECHENT, 1993 Molecular diversity at the self-incompatibility locus is a salient feature in natural populations of wild tomato (*Lycopersicon peruvianum*). Mol. Gen. Genet. **238:** 419–427.

SIMS, T. L., 1993 Genetic regulation of self-incompatibility. Crit. Rev. Plant Sci. **12:** 129–167.

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., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

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

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., 1991 A trend in population genetics theory, pp. 27–47 in *New Aspects of The Genetics of Molecular Evolution*, edited by M. KIMURA and N. TAKAHATA. Springer-Verlag, New York.

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.

UYENOYAMA, M. K., 1988 On the evolution of genetic incompatibility systems: incompatibility as a mechanism for the regulation of outcrossing distance, pp. 212–232 in *The Evolution of Sex*, edited by R. E. MICHOD and B. R. LEVIN. Sinauer Associates, Sunderland, Mass.

WRIGHT, S., 1938 Size of a population and breeding structure in relation to evolution. Science **87:** 430–431.

WRIGHT, S., 1939 The distribution of self-sterility 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 given size: a reexamination. Biometrics **16:** 61–85.

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

YOKOYAMA, S., and M. NEI, 1979 Population dynamics of sex-determining alleles in honey bees and self-incompatibility alleles in plants. Genetics **91:** 609–626.

Communicating editor: A. G. CLARK