Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 2419–2423, April 1990
Evolution

A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism

(population genetics/allelic genealogy/coalescence/balancing selection/major histocompatibility complex)

NAOYUKI TAKAHATA

National Institute of Genetics, Mishima 411, Japan

Communicated by Motoo Kimura, December 27, 1989 (received for review November 7, 1989)

ABSTRACT **Different alleles undergoing strong symmetric balancing selection show a simple genealogical structure (allelic genealogy), similar to the gene genealogy described by the coalescence process for a sample of neutral genes randomly drawn from a panmictic population at equilibrium. The only difference between the two genealogies lies in the different time scales. An approximate scaling factor for allelic genealogy relative to that of neutral gene genealogy is $\{\sqrt{S}/(2M)\}\cdot[\ln\{S/(16\pi M^2)\}]^{-3/2}$, where $M = Nu$ and $S = 2Ns$ (N, effective population size; u, mutation rate to selected alleles per locus per generation; s, selection coefficient). The larger the value of $\sqrt{S}/M$ (≥ 100), the larger the scaling factor. These findings, supported by simulation results, allow one to apply the theoretical results of the coalescence process directly to the allelic genealogy. Combined with the trans-species evolution of the major histocompatibility complex polymorphism for which balancing selection is believed to be responsible, allelic genealogy predicts that the number of breeding individuals in the human population could not be as small as 50–100 at any time of its evolutionary history. The analysis appears to contradict the founder principle as being important in recent mammalian evolution.**

Gene genealogy describes the ancestral relationships of genes at a locus, a simple consequence of the random loss or multiplication of genes in the reproduction process occurring in a finite population. When a gene is multiplied more than once and the products are transmitted to later generations, such a multiplication appears in a diagram of gene genealogy as a *divergence* when looked at forward in time and as a *coalescence* when looked at backward in time. The diagram represents identity by descent and the allelic states of the genes are irrelevant. Under neutrality (1), the theory of gene genealogy is called coalescent (2) and provides an efficient mathematical tool for studying various population genetics problems (see, e.g., ref. 3). On the other hand, the ancestral relationships among different allelic lines are called allelic genealogy (4). Obviously, an allelic divergence occurs only when a gene in an older line mutates to form a new one, and this process may or may not accompany a gene divergence. Of primary interest is how to describe which line was derived from which other line, particularly when selection is involved. Allelic genealogy thus focuses on the ancestral history in a sample that contains different lines; any relationship among genes that belong to the same line is irrelevant. It differs from the genealogy described by the lines-of-descent process (5) in which coalescences of genes occurring within each line and the age of each line are the main interest, but the allelic genealogy are not taken into account.

Although allelic genealogy was developed to examine quantitatively the extraordinary polymorphism of the major histocompatibility complex (MHC) molecules (6–10), the study was largely based on computer simulation (4). In this paper, I show that there exists a simple mathematical structure of allelic genealogy under strong symmetric balancing selection and use this theory to discuss the evolutionary implication of trans-species mode of MHC polymorphisms (7).

By balancing selection (one of the most efficient mechanisms to maintain polymorphism), I mean a collection of different selection schemes, all of which lead to the mean gene-frequency change given below. It has been shown that for this equation of the gene-frequency change, there can exist two fundamentally different selection models (4). In fact for any allele-frequency equation there are many alternative fitness models (11). This is unfortunate for the study of population genetics because simply observing a gene-frequency change cannot identify the underlying mechanism. However, this is not the concern here (see ref. 4).

Under balancing selection, alleles (allelic lines) can persist for a much longer time than neutral alleles, even in a relatively small population (12, 13). Those selected alleles may therefore differ from each other by more than one nucleotide change. When new alleles are produced each with an initial frequency of $1/(2N)$, where N is the effective population size, they differ from their parental ones by single changes. If new alleles happen to become common without further changes and their parental alleles still survive to that time, some pairs among common alleles can mutate to each other by single changes. However, such a rate would be of the order of the per-nucleotide mutation frequency, much smaller than the per-locus frequency (u) at which new alleles are produced. More importantly, new alleles thus produced will eventually replace old ones and play a significant role in the long-term evolution of molecules that experience even strong balancing selection and weak genetic drift. Thus the infinite-allele model of Kimura and Crow (14) (see also ref. 15), which ignores mutual changes among common alleles but incorporates the constant production of new alleles, seems most appropriate for describing allelic genealogy.

Analysis of a Diffusion Model

The model of symmetric balancing selection considered here assumes that the marginal fitness of allele A_k with frequency x_k is $1 - sx_k$. For this marginal fitness, selection can be either overdominant (heterotic) or frequency-dependent (4, 11). In the former case, s is the selective disadvantage of homozygotes relative to heterozygotes. The latter case occurs in the simplest frequency-dependent selection model which assumes that the relative fitness of A_k (multiplicative for diploid organisms) decreases by sx_k when its frequency is x_k. It is a model of the so-called minority advantage; the rarer, the more fit. In either case, the mean fitness of the population is given by $1 - sF = \Sigma_k(1 - sx_k)x_k$, where $F = \Sigma_k x_k^2$. In a

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: MHC, major histocompatibility complex.

diffusion approximation of the change in gene frequency x of a particular allele under the infinite-allele model and the above balancing selection, the mean change is given by $a(x) = -2Mx - Sx(x - F)$, where $M = Nu$, $S = 2Ns$, and S is to be read as $S/(1 - sF)$ when s is not small, while the variance is given by $b(x) = x(1 - x)$. Here, time is measured in units of $2N$ generations. In what follows, we further approximate the variance by $b(x) = x$ (14). As noted, this approximation is quite accurate for large S and facilitates the mathematical treatment considerably.

The above diffusion model has received much attention (12–14) and I first review some available results pertinent to the present purpose. Denote by $\Phi(x)dx$ the expected number of alleles whose equilibrium frequency is in the range x to $x + dx$. It is given by (14)

$$\Phi(x) = 4M\,e^{-Sx(x-2m)}x^{-1}, \tag{1}$$

where $m = F - 2M/S$ and F is assumed to be a constant. The function **1** has a local minimum at $x = \delta = 1/(2Sm)$ and a local maximum at $x = m - \delta$. For $S >> M$, $\Phi(\delta) \approx 4M\,e/\delta$ ($e \approx 2.718\ldots$) and $\Phi(m - \delta) \approx 4M \exp(Sm^2)/m$.

The number of rare alleles in the frequency range $1/(2N)$ to δ ($<<1$) becomes

$$n_r = \int_{1/(2N)}^{\delta} \Phi(x)dx \approx 4M(e - 1), \tag{2}$$

and the number of common alleles in the frequency range δ to 1 becomes

$$n_c = \int_{\delta}^{1} \Phi(x)dx \approx 4M \exp(Sm^2)\sqrt{\frac{\pi}{Sm^2}}\,. \tag{3}$$

On the other hand, F is computed by integrating $F = \int_{1/(2N)}^{1}\Phi(x)x^2dx$ or similar equations (14, 16). But since the integral is approximately $4Mm \exp(Sm^2)\sqrt{\pi/S}$ and $m \approx F$ for $S >> M$, we have

$$4M \exp(Sm^2)\sqrt{\frac{\pi}{S}} \approx 1. \tag{4}$$

Eq. **4** immediately yields the closed formula of F to be (14)

$$m \approx F \approx \sqrt{\frac{1}{2S}\ln\left\{\frac{S}{16\pi M^2}\right\}}. \tag{5}$$

Since from $n_c \approx 1/F$, the common alleles have an equal average frequency given by Eq. **5**. The actual number of alleles n_a becomes $n_r + n_c \approx 4M(e - 1) + 1/m$. In the above and subsequent calculations, the Laplace method for large S is extensively used.

Of interest are the expected times until a particular common allele, starting at moderate frequency x, becomes lost [$t(x)$] and a particular rare allele, starting at small frequency x, becomes common [$t^*(x)$]. There is only one exit boundary at $x = 0$ due to random genetic drift and mutation. The diffusion theory (17–19) allows us to write $t(x)$ as

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

where $\Psi(x) = \exp\{S(x - m)^2\}$. It is not easy to determine the above integrals exactly, but for large S we can derive a reasonably accurate result, which is given by

$$t(m) \approx \exp(Sm^2)\sqrt{\frac{2\pi}{S^3m^4}} = \frac{\sqrt{2}}{4MSm^2}, \qquad (x = m). \tag{6}$$

In the above, Eq. **4** is used. On the other hand, $t^*(x)$ for $x < \delta$ may be computed as the expected time until a rare allele first hits an interior point, say δ, before extinction. Although a rare allele tends to stay near the boundary $x = 0$ for a short time, it quickly increases the frequency up to $x = m - \delta$ once it becomes close to $x = \delta$. I conjecture that the only fate of a rare allele that increases its frequency up to δ is to become a common allele. Thus, to compute $t^*(x)$, we use the formula (18, 19)

$$t^*(x) = 2\left[\frac{u_0(x)}{u_1(x)}\int_0^x \frac{u_1(y)dy}{b(y)\Psi(y)}\int_0^y \Psi(z)dz + \int_x^\delta \frac{u_1(y)dy}{b(y)\Psi(y)}\int_y^\delta \Psi(z)dz\right],$$

where $u_1(x)$ is the fixation probability at $x = \delta$ given by $\int_0^x\Psi(y)dy/\int_0^\delta\Psi(y)dy$ and $u_0(x) = 1 - u_1(x)$. The approximate expression of $t^*(x)$ for $x << \delta$ may be found as

$$t^*(x) \approx \frac{e - 1 + e^{-1}}{Sm}, \tag{7}$$

which is independent of x.

There are n_c common and n_r rare alleles in the equilibrium population so that to obtain the expected times until one of the common alleles becomes extinct and one of the rare alleles becomes common, we divide $t(m)$ by n_c and $t^*(x)$ by n_r. We then have

$$\frac{t(m)}{n_c} \approx \frac{\sqrt{2}}{4MSm} \text{ and } \frac{t^*(x)}{n_r} \approx \frac{1.21}{4MSm} \tag{8}$$

from Eqs. **2**, **3**, **6**, and **7**. Eqs. **8** confirm the conjecture about the fate of a rare allele that hits $x = \delta$. They imply that there is a precise balance between two events, one in which a common allele goes to extinction and the other in which a rare allele replaces that common allele, and that the average turnover rate (r) of allelic lines in the population is the reciprocal of either of these values. Since our approximation for $t(m)$ is more accurate than that for $t^*(x)$, we define r as

$$r = 2\sqrt{2}MSm \approx 2M\sqrt{S\ln\left\{\frac{S}{16\pi M^2}\right\}}. \tag{9}$$

If we consider the mean rate of accumulation of mutations in the entire population α, we need to further divide r by the number of common alleles. This is because during a turnover of allelic lines, only a single line can increase the number of mutations by one while the others cannot. It is convenient to measure α in units of $1/u$ generations, in which case $\alpha = 1$ is expected under neutrality (1). Thus we obtain a simple result for α;

$$\alpha \approx 2\sqrt{2}MSm/(2Nun_c) = \frac{\sqrt{2}}{2}\ln\left\{\frac{S}{16\pi M^2}\right\}. \tag{10}$$

Allelic Genealogy

If a population evolves according to the above symmetric model of strong selection, relatively weak mutation, and drift, it it possible to construct an allelic genealogy. Suppose first that there are n common allelic lines and that we randomly sample i such lines from the current equilibrium population. The preceding analysis suggests that the time T at which the most recent turnover of allelic lines occurred is exponentially distributed with rate r in Eq. **9** (20). Assume therefore that this happened T ago (in units of $2N$ generations) and that a new line replaced one of the common lines. This new line, denoted by L, is likely to be a descendant of one of n common lines having existed around T ago. Designate the parental line by P. If P has been lost by the time of sampling, there is no possibility that any pair of common lines in a sample diverged T ago. This probability is $1/n$ since any common line is equally likely to become lost. If on the other hand P has survived, there is such a possibility that P and L diverged T ago. It is necessary, however, that both P and L are chosen in a sample, otherwise the divergence cannot be traced in the sampled allelic lines. The probability that the sample does not contain both P and L is $1 - {}_iC_2/{}_nC_2$ where C is the binomial coefficient. Hence we obtain the probability that two lines in the sample did not diverge T ago as

$$d_i = \frac{1}{n} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{{}_iC_2}{{}_nC_2}\right) = 1 - \frac{i(i-1)}{n^2}, \qquad [11]$$

while the divergence occurs with probability $1 - d_i$.

It is a simple matter to derive a formula for the event that a pair of lines in a sample of size $i(\leq n)$ diverged exactly K allelic turnovers ago. The value of K is a random variable and follows a geometric distribution; i.e., for $K = k(1, 2, 3, \ldots)$,

$$g_k = (1 - d_i)d_i^{k-1} = \frac{i(i-1)}{n^2}\left\{1 - \frac{i(i-1)}{n^2}\right\}^{k-1}. \qquad [12]$$

If an allelic divergence occurred K allelic turnovers ago, the number of distinct lines in that sample reduces to $i - 1$. In those $i - 1$ lines, Eq. **12** with i replaced $i - 1$ remains true. Obviously this process can be repeated until we find a single line for the first time. In constructing the coalescence process, Eq. **12** with $4N$ instead of n^2 is used: For large N, it reduces to the exponential density (2, 3). Our situation is different since n^2 is not necessarily large, but the conclusion is the same, as seen below.

For a given K, we define a random variable $S_K = \Sigma_{j=1}^{K} T_j$ (divergence time of alleles) in which values of T_j are mutually independent and follow a common exponential density with rate r in Eq. **9**. Thus the conditional probability density of S_K is given by a gamma function; that is,

$$\text{Prob}\{S_K = t|K = k\} = \frac{r(rt)^{k-1}}{(k-1)!}e^{-rt}. \qquad [13]$$

Taking this expectation with respect to g_k, we find the unconditional probability density of S_K to be an exponential function:

$$f_i(t) = \sum_{k=1}^{\infty} \text{Prob}\{S_K = t|K = k\}g_k = \alpha_i\exp\{-\alpha_i t\}, \qquad [14a]$$

$$\alpha_i = r(1 - d_i) = 2\sqrt{2}MSm\,\frac{i(i-1)}{n^2} \text{ for } 2 \leq i \leq n. \qquad [14b]$$

Eq. **14a** is also valid for further allelic divergences in the past so that the process of allelic genealogy in a sample of size i is completely determined by $f_j(t)(2 \leq j \leq i)$.

When we compare Eqs. **14** with the corresponding formulas for the coalescence process of randomly sampled neutral genes, we arrive at the main result of this paper. In the coalescence process, the time between successive coalescences is also exponentially distributed with rate $i(i - 1)/2$ in units of $2N$ generations. Hence we can interchange the factors, $2\sqrt{2}MSm/n^2$ and $1/2$, changing the time scales appropriately, and assert that the allelic genealogy and the coalescence process are identical in mathematical structure. Stated another way, all theoretical results obtained thus far for the coalescence process of neutral genes can be used after rescaling the time unit. The topological structure of the allelic genealogy is also the same as that of the coalescence process. This simplicity is a result of simplicity in the assumptions of the model and has been noted in ref. 21. Now that n in Eq. **14** can be regarded as $n_c \approx 1/m$, the rescaling factor from the coalescence process to the allelic genealogy becomes

$$f_s \approx \frac{\sqrt{2}}{8MSm^3} = \frac{\sqrt{S}}{2M}\left[\ln\left\{\frac{S}{16\pi M^2}\right\}\right]^{-3/2}, \qquad [15]$$

which is invariant if $\sqrt{S}/M$ is kept constant.

The probability density of the coalescence time at which i neutral genes are descended from j ancestral genes (22) and the probability of the number of distinct genes at a given past time (2, 23) are well known. Tavaré (3) provides a lucid review of this area. In units of $2N$ generations, the mean and variance of the coalescence time are especially simple and are given by $2(1 - 1/i)$ and $\Sigma_{j=1}^{i-1}[2/\{j(j + 1)\}]^2$, respectively. Furthermore, the mean number of distinct genes having existed $2Nt$ generations ago is given by (3)

$$\sum_{j=1}^{i} \frac{(2i-1)i_{[j]}}{i_{(j)}}\, e^{-j(j-1)t/2}, \qquad [16]$$

where $i_{(j)} = i(i + 1) \ldots (i + j - 1)$ and $i_{[j]} = i(i - 1) \ldots (i - j + 1)$.

I have shown that those formulas can be applied directly to the allelic genealogy, but the time scale is now in units of $2Nf_s$ generations. For example, we have the mean allelic divergence time between two lines that are randomly selected ($i = 2$) as $E\{T_p\} \approx 2Nf_s$ (generations) and that of the most distantly related lines ($i = n$) as $E\{T_c\} \approx 4Nf_s(1 - 1/n)$ (generations). The mean divergence time $E\{T_d\}$ averaged over all pairwise comparisons of common lines equals $E\{T_p\}$ (22). $E\{T_p\} = E\{T_d\}$ and $E\{T_c\}$ with $n = n_c = 1/F$ suggest a way to examine our main result. That is that the ratio $R_H = E\{T_c\}/E\{T_d\}$ simply becomes $2H$ ($H = 1 - F$), irrespective of the rescaling factor f_s. The average number of nucleotide changes that can accumulate in an allelic line during $E\{T_x\}$ (the subscript x stands for either c or d) may be computed by $rE\{T_x\}/n_c$. Denoting it by $E\{D_x\}$, we get $E\{D_d\} \approx n_c$ and $E\{D_c\} \approx 2Hn_c$.

Discussion and Conclusion

We first examine the accuracy of the theory. Takahata and Nei (4) have conducted a time-consuming simulation to observe F, n_a, $E\{T_c\}$, $E\{T_d\}$, and α. Although the value of s (=0.5) taken in their simulation might be too large and that of N (=200) too small, it is interesting to compare the two results. Table 1 shows that the theory is generally in fairly good agreement with simulation results, even though the value of Ns (=100) used is not enormous. This might be unexpected because we have used crude approximations in various places in manipulating the diffusion equation. There are some discrepancies, however, about allelic divergence times in particular. Part of the reason for this is that the selection coefficient $s = 0.5$ in the simulation ($S = 200$) is not small enough that sF in Eq. **1** can be neglected. If S is converted to $S/(1 - sF)$, the value of S in the theory may actually be 220, in which case and for $M = 0.001$, $E\{T_c\}/N \approx 402$ and $E\{T_d\}/N \approx 248$ (Table 1). Another cause comes from Eq. **5**, which tends to overestimate F to some extent (4). If the true value of F is 0.188 for $S = 200$ and $M = 0.001$, f_s in Eq. **15** would be 12% larger than that expected from Eq. **5**. Then $E\{T_c\}/N$ and $E\{T_d\}/N$ become 450 and 277, respectively, which are closer to the simulation values. Using the variance $b(x) = x$ instead of $x(1 - x)$ is also an approximation. However, these causes should not affect the R_H value and indeed the expected values agree well with those obtained by the simulation even when f_s is not accurately estimated. The

Table 1. Comparison between simulation results and theoretical values for various values of M

M	F	n_a	$E\{T_c\}$	$E\{T_d\}$	α	R_H
			M = 0.001			
Simulation	0.188	5.9	468	311	10.2	1.50
Theory	0.195	5.1	384	238	10.7	1.61
			M = 0.01			
Simulation	0.158	7.2	80.8	51.6	7.7	1.57
Theory	0.163	6.2	68.7	40.9	7.5	1.67
			M = 0.1			
Simulation	0.123	10.9	19.3	10.8	4.3	1.83
Theory	0.122	9.4	17.1	9.8	3.0	1.74

In Takahata and Nei's simulation (4), N = 200 and s = 0.5 were used and the number of replicates for each set of parameter values was 20. The values of $E\{T_c\}$ and $E\{T_d\}$ are measured in units of N generations. Results for M = 1.0 are not shown, because, in this case, there are always several rare alleles and the theory does not provide accurate predictions, particularly for the divergence time of alleles.

same conclusion was drawn about $E\{D_d\}$ and $E\{D_c\}$. As shown in Fig. 1, the expected values of D_d and D_c are 5.9 and 9.7, and those observed in this particular replicate were 7.4 and 9.8, respectively. Considering this, I conclude that the present theory can satisfactorily describe most aspects of allelic genealogy when $\sqrt{S/M} \geq 100$ (data not shown) and that the genealogical structure is exactly the same as that of the coalescence process. The only important difference between the two theories is their different time scales; allelic genealogy is magnified by a factor f_s relative to neutral gene genealogy.

Let us apply the theory to some of the unusual features of MHC polymorphisms (6–10). As pointed out (8–10), such features are always related to the antigen recognition site in a MHC molecule that is composed of 57 amino acid residues and to which processed foreign peptides can bind. Most extraordinary is the long persistence time of polymorphic alleles. One instance clearly shows that a pair of polymorphic alleles predated the divergence between mice and rats, *ca.* 13 million years ago (10). This characteristic specific to MHC molecules is designated as trans-species evolution of polymorphism (7, 24). It is enormously long compared with $4N$ generations for the mean coalescence time of neutral genes (3). In fact, if there are two or more generations per year in these rodents and if $N = 10^5$ (25), such alleles must have diverged at least $260N$ generations ago. To examine whether this figure is compatible with the present theory, we need to estimate the mutation rate and selection coefficient. It was shown that the synonymous substitution rate (2.9×10^{-9} per site per year) in the region of MHC is somewhat lower than that in other regions (26). Assume, under neutrality, that this synonymous substitution rate is an estimate of the mutation rate per site. Since about 80% nucleotide changes are nonsynonymous, assume also that the relevant mutation rate (u) of amino acid replacements per antigen recognition site can be estimated as $2.9 \times 10^{-9} \times 57 \times 3 \times 0.8 \approx 3 \times 10^{-7}$ per year and therefore about 10^{-7} per generation. If all amino acid replacements at the antigen recognition site are not selected for (6, 9), the estimate of u here and in the discussion below should be decreased appropriately. On the other hand, there is little information about the value of s. Suppose, however, that $Ns \geq 10$ is satisfied for selection to be effective. Table 2 shows then that s = 0.001 is too small, suggesting that s = 0.01 or more. If s = 0.01 in rodents, then $F \approx 0.057$ ($n_c \approx 17.6$) and $T_c/N \approx 182 \pm 98$, which are compatible with the observations. But we must assume a lower mutation rate, stronger selection, larger population size, or all three if the persistence time of polymorphic alleles is even longer.

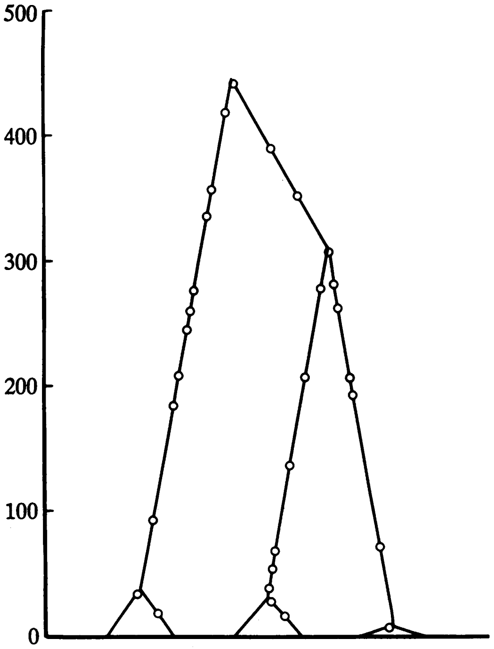

FIG. 1. Computer-generated allelic genealogy for S = 266 and M = 0.001. There are six allelic lines at the time of sampling. Each allelic turnover, a new line starting with frequency $1/(2N)$, is indicated by a small circle. The time scale of the ordinate is in units of $2N$ generations. Note that the coalescence time between genes belonging to the same allelic line cannot exceed the most recent allelic divergence time but coalescences of genes between different lines occur prior to the divergence of these lines.

During such a long persistence time, allelic lines must have accumulated not only selected but also neutral mutations and must have had more descendant allelic lines. It may therefore be interesting to infer their relative contributions. Using Eqs. **3** and **5** for $N = 10^5$, $s = 0.01$, and $u = 10^{-7}$, we have $E\{D_c\} \approx 2Hn_c = 33$ for selected mutations while the expected number of neutral mutations $E\{T_c\}v$ is $193Nv$ (v, neutral mutation rate per locus per generation). Even if v is as large as 10^{-6} so that $Nv = 0.1$, selected mutations would have contributed more to differentiate MHC molecules in terms of nucleotide differences among alleles. Such a rapid accumulation of selected mutations is reflected in Eq. **10**; i.e., $\alpha \approx 9.1$ for this case.

We may also ask how many different neutral alleles there are in each selected line on the average. The average age of each selected line (from a tip to the nearest circle along a line in Fig. 1) is given by a reciprocal of α in Eq. **10**. It becomes $11N$ generations in our example. This might be sufficiently large to apply sampling theory (27). The bottleneck phase during which a rare allele becomes common is much shorter than the persistence time of a common allele, the proportion being $t^*(x)/(n_r t(m)) \approx t(m)/(n_c t(m)) \approx m$. Furthermore, the

Table 2. Expected values of the rescaling factor f_s, the persistence time of polymorphic alleles $E\{T_c\}/N$, and the number of common alleles n_c.

		s			
N		0.0001	0.001	0.01	0.1
10^4	f_s	—	4.1	7.5	16.0
	$E\{T_c\}/N$	—	10.4	26.0	60.8
	n_c	—	2.8	7.3	20.1
10^5	f_s	9.4	20.5	48.3	119
	$E\{T_c\}/N$	20.4	68.7	182	468
	n_c	2.2	6.2	17.6	51.3

The mutation rate u per locus per generation is assumed to be 4.5×10^{-6} for $N = 10^4$ and 10^{-7} for $N = 10^5$ to mimic the situation of humans and of rodents, respectively. —, Ns is too small for the theory to be valid.

effective gene number in each selected line is $2N_c \approx 2N/n_c$, in terms of which $11N$ is about $186N_c$ generations. Therefore the mean number of neutral alleles in each line is likely to be at its equilibrium value. In the population of a selected line, the corresponding value of M ($=N_cv$) is approximately 0.005 so the mean number of neutral alleles is only about 1.15 (18, 27). However, the number in the whole population may be $1.15 \times n_c \approx 20.0$, which is much larger than the value of 3.79 expected from a panmictic population with $Nv = 0.1$. Clearly, balancing selection plays the same role in the evolution of neutral genes as does subdivision in a population. This tremendous increase of genetic variability due to neutral alleles is, however, consistent with their smaller contribution to the divergence of alleles in terms of nucleotide changes; neutral alleles within a selected line are closely related to each other and neutral mutations between different selected lines accumulate more slowly than selected mutations.

As another example of our main concern, consider the human population with long-term generation time and population size being 15 years and 10^4, respectively (25). If humans diverged from chimpanzees or gorillas *ca.* 5 million years ago, this amounts to $33N$ generations. Assume that $s \geq 0.01$ and $u = 4.5 \times 10^{-6}$ per antigen recognition site ($M = 0.045$) because of the long generation time. If s is not as large as 0.1 as might be argued from the monomorphism at MHC loci in small isolated populations (8, 9, 28), N must be larger than 10^4 for the theory to account for HLA polymorphisms. If $N = 10^4$, the divergence time of the human lineage corresponds to 2.1 for $s = 0.01$ and 1.0 for $s = 0.1$ (Table 2). A recent attempt for estimating the number of individuals at speciation (29) makes use of a large number and long persistence times of polymorphic alleles at MHC loci. My approach to this problem is as follows.

First, to estimate the average number (k_f) of distinct lines in a sample of size i that existed 5 million years ago, we use Eq. **16** or the distribution itself for the number of distinct lines (3, 22) if the maximum likelihood approach is preferred. For $i = 47$ as in HLA-B (29), $k_f = 1.4$ for $s = 0.01$ and 2.3 for $s = 0.1$. It is thus likely that there were at least two distinct allelic lines of the sample of $i = 47$ when the human lineage originated. Now, suppose that there was a severe bottleneck in the founding population of human lineage as described in *Genesis*. The strength of a bottleneck can be determined by not only the reduced population size (N_b) but also the duration time (t_b). The finding of shared polymorphisms between humans and chimpanzees imposes a strong requirement on the value of N_b. A necessary condition for shared polymorphisms is that plural alleles were passed on from the common ancestral species. For two neutral genes (assuming that $N_bs < 1$), this probability is given by $\exp\{-t_b/(2N_b)\}$ (3, 30). For it to be as high as 0.99, $N_b = 50t_b$ is necessary. Hence, even for a one-generation bottleneck ($t_b = 1$), the founding population must have consisted of at least 50 breeding individuals. This conclusion is reinforced if the mutation rate u in the human lineage is smaller than that in the rodent (31). If $u = 10^{-6}$, the human divergence time becomes 0.77 ($k_f = 2.8$) for $s = 0.01$ and 0.33 ($k_f = 5.7$) for $s = 0.1$. The larger the number of shared alleles and the longer the bottleneck phase, the more founding individuals required.

Of course, there is no good reason to believe that there was only one such bottleneck event along the human lineage. If N were reduced such that $Ns < 1$ during these 5 million years, population genetics theory (17–19) predicts that effects of selection must have ceased and that polymorphism would have been lost. Hence, the passing on of MHC polymorphisms through such events, if ever, requires that, at any moment in the course of human evolution, the number of individuals is as large as $Ns = 10$ or more. This implies that N is at least of the order of $10/s$ and that it may well be larger than 100. In other words, though rather redundant, the evolution of *Homo sapiens sapiens* or speciation in general could take place without any genetic revolution mediated directly through extreme founder or bottleneck effects (32).

If the history of organisms is inscribed in the chromosomes, we can hope to decipher it, if only partially. Population dynamics could be read from polymorphic loci but not from monomorphic loci. In case of neutral polymorphic loci, population history and dynamics may be traced back to about $4N$ generations ago. This period might be too short for some organisms with short generation times and small population sizes as compared to their life times in the evolutionary scene. In contrast, polymorphic loci that have been maintained by balancing selection are worth scrutiny in the present context and, among such candidates (25, 33), MHC loci will provide a unique opportunity for studying *palaeopopulation biology*.

I thank J. F. Crow, J. Gillespie, N. Kaplan, M. Kimura, J. Klein, M. Nei, M. Slatkin, and J. Tomizawa for their comments and stimulating discussions on genealogy. Special thanks are due to J. Klein and M. Nei who I believe raised a right question about the evolution of MHC molecules.

1. Kimura, M. (1968) *Nature (London)* **217**, 624–626.
2. Kingman, J. F. C. (1982) *Stoch. Processes Appl.* **13**, 235–248.
3. Tavaré, S. (1984) *Theor. Popul. Biol.* **26**, 119–164.
4. Takahata, N. & Nei, M. (1990) *Genetics*, in press.
5. Griffiths, R. C. (1980) *Theor. Popul. Biol.* **17**, 37–50.
6. Klein, J. (1986) *Natural History of the Major Histocompatibility Complex* (Wiley, New York).
7. Klein, J. (1980) in *Immunology 80: Progress in Immunology IV*, eds. Fougereau, K. & Dausset, J. (Academic, London), pp. 239–253.
8. Figueroa, F. & Klein, J. (1988) in *H-2 Antigens: Genes, Molecules, Function*, ed. David, C. S. (Plenum, New York), pp. 61–76.
9. Hughes, A. & Nei, M. (1988) *Nature (London)* **335**, 167–170.
10. Figueroa, F., Günther, E. & Klein, J. (1988) *Nature (London)* **335**, 265–267.
11. Denniston, C. & Crow, J. F. (1990) *Genetics*, in press.
12. Wright, S. (1960) *Biometrics* **16**, 61–85.
13. Maruyama, T. & Nei, M. (1981) *Genetics* **98**, 441–459.
14. Kimura, M. & Crow, J. F. (1964) *Genetics* **49**, 725–738.
15. Crow, J. F. (1989) *Genetics* **121**, 631–634.
16. Yokoyama, S. & Nei, M. (1979) *Genetics* **91**, 609–626.
17. Crow, J. F. & Kimura, M. (1970) *An Introduction to Population Genetics Theory* (Harper & Row, New York).
18. Ewens, W. J. (1980) *Mathematical Population Genetics* (Springer, Berlin).
19. Karlin, S. & Taylor, H. M. (1981) *A Second Course in Stochastic Processes* (Academic, New York).
20. Gillespie, J. H. (1983) *Am. Nat.* **121**, 691–708.
21. Gillespie, J. H. (1989) *Am. Nat.* **134**, 638–658.
22. Takahata, N. & Nei, M. (1985) *Genetics* **110**, 325–344.
23. Watterson, G. A. (1984) *Theor. Popul. Biol.* **26**, 77–92.
24. Klein, J. & Takahata, N. (1990) *Immunol. Rev.*, in press.
25. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
26. Hayashida, H. & Miyata, T. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2671–2675.
27. Ewens, W. J. (1972) *Theor. Popul. Biol.* **3**, 87–122.
28. O'Brien, S. J., Roelke, M. E., Marker, L., Neuman, A., Winkler, C. A., Meltzer, D., Colly, L., Greumann, J. F., Bush, M. & Wildt, D. E. (1985) *Science* **227**, 1428–1434.
29. Klein, J., Kasahara, M., Gutknecht, J. & Figueroa, F. (1989) *Chem. Immunol.* **49**, 35–50.
30. Takahata, N. (1989) *Genetics* **122**, 957–966.
31. Wu, C.-I. & Li, W.-H. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1741–1745.
32. Mayr, E. (1977) *Populations, Species, and Evolution* (Harvard Univ. Press, Cambridge, MA).
33. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).