# POPULATION BOTTLENECKS AND NONEQUILIBRIUM MODELS IN POPULATION GENETICS. I. ALLELE NUMBERS WHEN POPULATIONS EVOLVE FROM ZERO VARIABILITY

TAKEO MARUYAMA* AND PAUL A. FUERST†

*National Institute of Genetics, Mishima 411, Japan, and †Department of Genetics, The Ohio State University, Columbus, Ohio 43210*

Manuscript received October 11, 1983
Revised copy accepted July 3, 1984

ABSTRACT

A simple numerical method was developed for the mean number and average age of alleles in a population that was initiated with no genetic variation following a sudden population expansion. The methods are used to examine the question of whether allele numbers are elevated compared with values seen in equilibrium populations having equivalent gene diversity. Excess allele numbers in expanding populations were found to be the rule. This was true whether the population began with zero variation or with low levels of variation in either of two initial distributions (initially an equilibrium allele frequency distribution or initially with loci occurring in only two classes of variation). Although the increase of alleles may persist for only a short time, when compared with the time which is required for approach to final equilibrium, the increase may be long when measured in absolute generation numbers. The pattern of increase in very rare alleles (those present only once in a sample) and the persistence of the original allele were also investigated.

THE effects of population size fluctuations upon the genetic variability maintained in natural populations has been of interest to geneticists for many years (WRIGHT 1938). Especially important is the situation in which a population goes through a severe, but temporary, population restriction, termed a population "bottleneck." The theoretical analysis of bottlenecks is relevant not simply because of the light it can shed upon nonequilibrium situations in population genetics but because the theory can have a very concrete practical application in the study of speciation (CARSON 1968), for the conservation of genetic resources in rare or endangered species and in genetic resource conservation of cultivated animals and plants.

In this series of papers we will study various nonequilibrium population genetic models that have a bearing upon the bottleneck problem. NEI, MARUYAMA and CHAKRABORTY (1975) pioneered the study of bottleneck effects, using a model in which a severe size reduction occurred suddenly in a population that was in equilibrium between mutation and genetic drift. Large amounts of the genetic variability in the population was lost following the bottleneck, and NEI, MARU-

YAMA and CHAKRABORTY (1975) were able to derive explicit formulas for the amount of gene diversity (genic heterozygosity) maintained in the population as a function of several population parameters, including bottleneck size and population growth rate.

Unlike gene diversity, some other statistics of variation, including those directly related to the number of alleles in a population, are not easily calculated in the extreme nonequilibrium situation. NEI and LI (1976) have obtained the distribution of the gene frequency spectrum following a sudden change in the population size. Their methods illustrate one difficulty that faces those interested in the analysis of the nonequilibrium situation. Although natural populations fluctuate in size, they do not usually undergo severe reductions, followed immediately by restoration to the original population size. Rather, there is a gradual increase in population size. Such changes are usually modeled by the approach of NEI, MARUYAMA and CHAKRABORTY (1975), but their model does not easily provide information on allele numbers; other methods may need to be employed. Among these is the use of alternative computer simulation models, such as those based on Ito's stochastic integrals (MARUYAMA 1980, 1981, 1982, 1983; MARUYAMA and NEI 1981; NEI, MARUYAMA and WU 1983). This will provide us with representative results but not exact solutions to the problems of interest. Such simulation methods will be examined in detail in subsequent papers in this series.

Alternatively, we can study nonequilibrium problems by using methods that will provide unique solutions but that do not approximate the natural situation as closely as the computer stochastic integral methods permit. By employing contrasting models, we may be able to make statements concerning possible outcomes in more realistic situations. Two processes define the bottleneck situation: population reduction at the bottleneck (which causes the loss of genetic variability) and population expansion following the bottleneck (which ultimately restores the variation through mutation). The nonequilibrium analysis of change of variability can similarly break the problem into (1) the decrease of population size from a previously large equilibrium population to a small steady population size and (2) the instantaneous increase of the population from a small size to a large steady state population. We can also extend this approach by considering (3) the combination of the two processes into a cycle of population size changes and, finally (4) the relaxation of the assumption of instantaneous change. In this paper, and the following series, we will consider these problems and offer some solutions to them.

NEI and LI (1976) studied the allele frequency distribution in this situation and obtained formulas for the gene frequency spectrum as an eigenfunction expansion of the time-dependent solution, involving hypergeometric functions. GRIFFITHS (1979a,b, 1980) gave a complete mathematical analysis on the distribution of the allelic frequencies in a sample taken from a transient population. GRIFFITHS (1979b, 1980) gave formulas for the mean number of different alleles expected to appear in a finite sample of genes taken from a population evolving from an arbitrary initial condition, including totally homoallelic cases. Griffith's formulas are expressed in series expansions by orthogonal polynomials on the Dirichlet distribution. Although the analyses are mathematically exact and beau-

tiful, actual calculation requires a computer and is fairly complicated. WATTERSON (1983) obtained a slightly simpler formula for the frequency spectrum of genes in a transient population. Computation using Watterson's formula is relatively easy, but it requires handling terms of large factorials if the sample size is large.

In this paper, we will present a simple numerical method for calculation of the mean number and age of alleles in a sample taken from an evolving population. The method enables us to compute the number of different alleles each with a given frequency. The initial condition does not need to be entirely homoallelic, although most of the cases examined in the paper assume zero variability initially. These methods are based upon use of difference equations, which approximate a Kolmogorov backward equation representing a classical diffusion process.

Since gene diversity and allele number are two important measures and have become available through electrophoretic studies in natural populations, we believe that the methods of this simple analysis will prove useful. When a natural population is surveyed for genetic variation, its history is usually unknown. Consequently the question of whether or not it could be in equilibrium between mutation and genetic drift is unanswerable. Because of this, we will also examine the relationship between the number of alleles in nonequilibrium situations and the observed population heterozygosity, which is measurable in any natural population. This comparison will allow us to determine the relative effects caused by bottlenecks that could be observed in actual surveys.

## THE MODEL AND ANALYSIS BASED ON DIFFERENCE EQUATION

Consider a locus at which infinitely many alleles are possible, with every mutant being new to the population. Further assume that every gene, irrespective of its allelic state, mutates to a new allele at a rate of $v$, where $v$ is a constant. The population size is assumed to be finite, with all alleles being selectively neutral and differentiable from one another (*i.e.*, the infinite allele model of WRIGHT 1948 and KIMURA and CROW 1964). The population size, $N$, is constant, so that for a diploid organism the total number of genes in the population is $2N$. The population is assumed to be homoallelic at each locus under consideration at time $t = 0$. Time can be measured either in absolute number of generations or in terms of $2N$ generations. Since every mutant is new, every existing allele will have a unique entry time into the population; at a given observation point, each allele has a well-defined time of persistence measured beginning at its time of entry. We call the persistence time the "age" of the allele.

Assuming that the population size is sufficiently large, we can approximate the sampling process by diffusion models. This allows us to use the various mathematical tools available for the analysis of diffusion processes, notably the Kolmogorov backward equation. Results based on the diffusion approximation can be compared to those obtained by Watterson's formula for the discrete Markov chain which describes the original dynamics of the evolving population.

Consider a particular allele, $A$, and let $\phi(t, x, y)$ be the probability density that the frequency of $A$ is $y$ at time $t$, given that it is $x$ at time $t = 0$. Then $\phi(t, x, y)$ satisfies

$$\frac{\partial \phi}{\partial t} = \frac{x(1-x)}{2}\frac{\partial^2 \phi}{\partial x^2} - 2Nvx\frac{\partial \phi}{\partial x} \tag{1}$$

where $\phi = \phi(t, x, y)$. Time in (1) is measured in units of $2N$ generations. Following EWENS (1972), let

$$f(y) = 1 - (1-y)^M$$

where $y$ is the variable defined above and $M$ is the sample size. We can define

$$u(t, x) = \int_0^1 \phi(t, x, y) f(y) dy. \tag{2}$$

Function $u(t, x)$ is the probability that allele $A$ will appear at least once in the sample taken at time $t$.

For the starting frequency of $A$, $x$ is 1 if $A$ is the allele in the original homoallelic population, and $x$ is zero if $A$ is the allele entering the population at time $t > 0$. More generally, if allele $A$ has an intermediate frequency at the beginning of the process, $x$ is equal to the value of the frequency of $A$ at time $t = 0$. This can be the case following a bottleneck when a population loses most of its genetic variability but retains alleles at intermediate frequencies at a few loci.

Applying the integration defined by (2) to $\phi(t, x, y)$ in (1), we get the following Kolmogorov backward equation

$$\frac{\partial u(t, x)}{\partial t} = \frac{x(1-x)}{2}\frac{\partial^2 u(t, x)}{\partial x^2} - 2Nvx\frac{\partial u(t, x)}{\partial x} \tag{3}$$

with the initial and boundary conditions

$$u(t, 0) = \frac{du(t, x)}{dx}\bigg|_{x=1} = 0 \tag{4}$$

and

$$u(0, x) = 1 - (1-x)^M. \tag{5}$$

We can express the solution of (3) satisfying (4) and (5) in a series expansion by eigenfunctions of the operator on the right side of (3), since it admits Jacobi orthogonal polynomials as its eigenfunctions. This approach has been taken by KIMURA (1955), CROW and KIMURA (1956) and NEI and LI (1976) for similar problems. These authors expressed solutions in a series of hypergeometric functions. Such analytic solutions expressed as series of transcendental functions have many nice features; particularly useful is the known asymptotic behavior of the process. There are, however, problems such as those being considered here that require numerical solutions obtained using a computer. Although all information concerning the fate of genes in the population can theoretically be obtained from the fundamental solutions of (1) expressed in series of eigenfunctions, it is often more straightforward to integrate (3) numerically.

Our approach to the problem of allele number and age is, therefore, numerical integration of (3) with the boundary conditions given in (4) and (5). The integration takes place in the two-variable space of $t$ and $x$, where $x$ ranges

between 0 and 1 and $t$ varies from 0 to some finite value. We can denote this space by $[(0, \infty) \times (0, 1)]$. Regarding this space as a two-dimensional lattice in which the width between two adjacent points along the time axis is $\Delta t$ and that between two adjacent points along the frequency axis is $\Delta x$, we can consider $u(t, x)$ defined on these lattice points and denote its value as $u_{ij}$. If we consider (3) as a difference equation rather than a differential equation, we have the following approximations

$$\frac{\partial u(t, x)}{\partial t} \simeq \frac{u_{i+1,j} - u_{ij}}{\Delta t}$$

$$\frac{\partial u(t, x)}{\partial x} \simeq \frac{u_{i,j+1} - u_{i,j-1}}{2\Delta x}$$

and

$$\frac{\partial^2 u(t, x)}{\partial x^2} \simeq \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{(\Delta x)^2}.$$

Substituting these approximations into (3) and rearranging, we have

$$u_{i+1,j} = u_{ij} + \frac{\Delta t}{2}\left[x_j(1 - x_j)\frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{(\Delta x)^2}\right] - \Delta t\left[2Nvx_j\frac{u_{i,j+1} - u_{i,j-1}}{2\Delta x}\right] \quad (6)$$

where $x_j$ is the value of $x$ at the corresponding points on the lattice (MARUYAMA 1977). For the initial condition we set

$$u_{oj} = 1 - \left(1 - \frac{j}{l}\right)^M$$

and for the boundary condition

$$\begin{aligned} u_{io} &= 0 \\ u_{i,l} &= u_{i,l-1}, \end{aligned} \quad (7)$$

where $l$ corresponds to the largest value of $j$ in the lattice or $l\Delta x = 1$. Using the difference equation (6), $u_{ij}$ can be readily computed for $i = 1, 2, 3 \ldots$. Then, it is known from the theory of differential equations that, as $\Delta t$ and $\Delta x$ go to zero, $u_{ij}$ converges to the particular solution of (3). We need to keep the $x_j(1 - x_j)\Delta t/(2(\Delta x)^2)$ less than $\frac{1}{2}$ to have convergence of $u_{ij}$ to the correct solution.

Among the $u_{ij}$'s computed, we are interested in $u_{il}$ and $u_{i1}$ representing, respectively, the probability that an allele existing in the original population has not been lost and the probability that an allele entering the population has not been lost at time $i(t = i\Delta t)$.

Examination of (3) demonstrates that, in the diffusion approximation, the population size, $N$, does not exist as a separate parameter, since $N$ enters the equation as a component of $Nv$. However if we partition the space variable $x$ into $l(j = 0, 1, 2 \ldots, l)$ in the difference equation (6), $u_{i1}$ represents information on the fate of a mutant present only once at the time it entered a population of $l$ genes. Similarly, $u_{il}$ gives information for the allele originally occupying the entire population.

In analyzing the number of alleles observed in a sample of size $M$, we consider two sources of information. One is the probability, $u_{il}$, that the original allele persists in the population. The second is the probability of persistence of those alleles entering the population between time zero and time $t$. Since we assume that every mutant has a constant rate of mutation, $v$, there will be on the average $2Nv$ new mutants entering each generation, and $(2Nv \times 2N)$ new mutants in a unit time.

A mutant entering the population at time $\xi$ will have the probability $u\left(t - \xi, x = \frac{1}{2N}\right)$ of persisting at time $t$. Therefore, the total number of alleles that enter the population between time $\xi$ and $\xi + d\xi$ and remain until time $t$ is equal to

$$4N^2vu\left(t - \xi, \frac{1}{2N}\right) d\xi.$$

Approximating this in terms of $u_{ij}$, we have

$$4N^2vu_{\gamma-\xi,1}\Delta t$$

where $t = \gamma\Delta t$, with $\gamma = 0, 1, 2, \ldots$ and $\xi = 0, 1, 2, \ldots, \gamma$. Therefore, the total number of alleles, $n(t, M)$, which will be found in a sample of $M$ genes taken randomly from a large population, will be equal to

$$n(t, M) = u_{\gamma l} + 4N^2v\Delta t \sum_{\xi=1}^{\gamma} u_{\gamma-\xi,1}. \tag{8}$$

Despite the simplicity of the computational procedure, formula (8) provides a good approximation to the discrete Markov chain model of the Fisher-Wright type. A few examples are given in Table 1 where the average numbers of alleles expected to be found in a sample taken from an evolving population are compared for the two models.

The average age of an allele can be calculated from $u_{ij}$ or $u(t, x)$. Note that $u\left(t - \xi, \frac{1}{2N}\right)$ is the probability that a mutant arising at time $t - \xi$ is retained in the population at time $t$, and that its age at time $t$ is $t - \xi$. Therefore, the average age of all extant alleles in the population at time $t$ is equal to

$$\left[4N^2v\int_0^t \xi u\left(t - \xi, \frac{1}{2N}\right) d\xi + tu(t, 1)\right] \Big/ \left[4N^2v\int_0^t u\left(t - \xi, \frac{1}{2N}\right) d\xi + u(t, 1)\right]$$

In terms of the solution of the difference equations, the average age, $A_g(t)$, at time $t$ is given by

$$A_g(t) = \frac{u_{\gamma,l}t + 4N^2v\Delta t \sum_{\xi=1}^{l} \xi u_{l-\xi,1}}{u_{\gamma,l} + 4N^2v\Delta t \sum_{\xi=1}^{l} u_{\gamma-\xi,1}}. \tag{9}$$

TABLE 1

*Comparison of the average number of alleles calculated by formula (8) and by WATTERSON'S (1983) formula.*

| | $4Nv$ (sample size 10) | | | $4Nv$ (sample size 20) | | |
|---|---|---|---|---|---|---|
| Time | 1 | 5 | 20 | 0.5 | 2 | 10 |
| 0.025 | | | | | | |
| (8) | 1.12 | 1.56 | 3.05 | 1.11 | 1.43 | 3.04 |
| (W) | 1.12 | 1.58 | 3.11 | 1.11 | 1.44 | 3.12 |
| 0.05 | | | | | | |
| (8) | 1.22 | 2.04 | 4.53 | 1.20 | 1.78 | 4.57 |
| (W) | 1.22 | 2.07 | 4.62 | 1.20 | 1.80 | 4.69 |
| 0.10 | | | | | | |
| (8) | 1.40 | 2.83 | 6.40 | 1.34 | 2.32 | 6.70 |
| (W) | 1.41 | 2.86 | 6.54 | 1.35 | 2.36 | 6.88 |
| 0.20 | | | | | | |
| (8) | 1.67 | 3.92 | 7.79 | 1.54 | 3.06 | 9.08 |
| (W) | 1.67 | 3.98 | 7.98 | 1.55 | 3.12 | 9.32 |
| 0.5 | | | | | | |
| (8) | 2.19 | 5.34 | 8.05 | 1.88 | 4.22 | 10.90 |
| (W) | 2.23 | 5.45 | 8.28 | 1.91 | 4.31 | 11.22 |
| 1.0 | | | | | | |
| (8) | 2.60 | 5.66 | 8.05 | 2.15 | 4.92 | 10.99 |
| (W) | 2.66 | 5.82 | 8.28 | 2.20 | 5.05 | 11.33 |

Time, In units of $2N$ generations; (8), formula (8); (W), WATTERSON'S (1983) formula.

Formula (9) yields a good approximation for the average age of alleles simulated by using the Markov chain with $M$ genes. It should be remembered that the average age in our analysis is defined with respect to the beginning of the mutational process. The original allele in the population will have been carried over from some preexisting population, but its age is still calculated only from time $t = 0$. Because of this definition of age, the age of alleles during the early phase of the process must be strongly related to the time since the start of the population.

## TRANSIENT EXCESS OF ALLELES

One of the major purposes of the present study was to observe the time-dependent dynamics of the number of alleles found in a sample of genes taken from a population. In nature, we cannot determine the equilibrium state of a population. In practice, we assume that the population being studied is at equilibrium and that we can, therefore, use equilibrium expectations to test various statistics generated by genetic models (see CHAKRABORTY, FUERST and NEI 1978). Studies have shown that the asymptotic rate at which the frequency spectrum approaches its steady state distribution will be the same for all regions

of the spectrum. This rate of approach is equal to $(2v + 1/2N)$ per generation, as shown by EWENS and KIRBY (1975), KARLIN and AVNI (1975) and NEI and LI (1976). The studies of NEI and LI (1976), and those presented here, deal with the transient behavior of the process far from an asymptotic value. Based on the pattern of graphs of the transient allele frequency spectrum, NEI and LI argue that since the spectrum has a sharp peak at 0 at the beginning of the process rare alleles in a transient state will greatly exceed the numbers expected if the population was at a steady state condition with an equivalent average heterozygosity.

We can directly compare the number of alleles appearing in a sample of $M$ genes taken from an evolving population with the expected number based on steady state assumptions. If the population starts from zero variability at time $t = 0$, the level of heterozygosity, $H_t$, is given by

$$H_t = \frac{4Nv}{1 + 4Nv}(1 - e^{-(1+4Nv)t}) \tag{10}$$

where $t$ measures time in units of $2N$ generations (NEI, MARUYAMA and CHAKRABORTY 1975). It is easy to show that $H_\infty = 4Nv/(1 + 4Nv)$, which is the equilibrium level of heterozygosity obtained by KIMURA and CROW (1964), and that $H_0 = 0$.

To compare the number of alleles in an evolving population with the number expected in a steady state population, we used (10) to obtain the heterozygosity at time $t$ in the evolving population. We can then equate the value of $H_t$ to $\hat{\theta}/(1 - \hat{\theta})$ where $\hat{\theta}$ denotes the $4Nv$ value of the assumed steady state population with equivalent gene diversity. Then the number of alleles in a sample of $M$ genes taken from such a population, denoted by $n_*(t, M)$, is given by

$$n_*(t, M) = 1 + \frac{\hat{\theta}}{\hat{\theta} + 1} + \frac{\hat{\theta}}{\hat{\theta} + 2} + \ldots\ldots + \frac{\hat{\theta}}{\hat{\theta} + M - 1} \tag{11}$$

where $\hat{\theta} = H_t/(1 - H_t)$ (EWENS 1972).

We have compared the actual number of alleles, $n(t, M)$, calculated from (8), to $n_*(t, M)$ of formula (11) for various values of $t$ and $4Nv$. Figure 1 shows the number of alleles that will be found in a sample of 200 genes at various times in an evolving population. The dashed curves in Figure 1 represent the expected number of alleles, based on an assumption that the population was in mutation-drift equilibrium. Since each dashed curve is below the corresponding solid curve, it is clear that the number of alleles grows faster than the heterozygosity. As a consequence there will be an apparent excess of alleles observed in the population if it were assumed that the population under study is actually in mutation-drift equilibrium. A population starting from zero variability must accumulate new mutants, and each new mutant must begin at low frequency. Such mutants can appear in a sample, especially if the sample is large enough, but they contribute little to the average heterozygosity. This results in a comparison in which the steady state expectation predicts many fewer rare alleles than are observed. The difference between the two expectations can be large, partic-

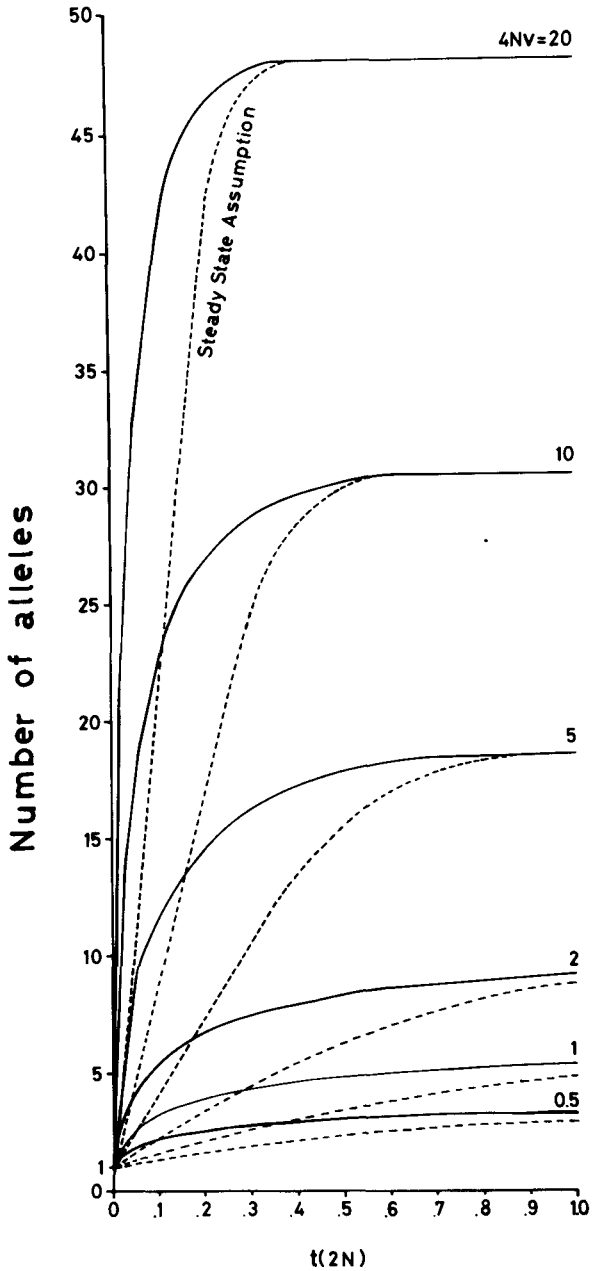


FIGURE 1.—Average number of alleles in a sample of 200 genes taken from an evolving population (solid curves). Dashed curves, The number of alleles based on the assumption that the population is in equilibrium at the observed level of heterozygosity. Time, In units of $2N$ generations.

ularly when $4Nv$ is large. However, the difference persists for a relatively short period of time. For instance, if $4Nv = 20$, which is a rather large value, the maximum difference between the two expected numbers of alleles can be as large as 20 alleles or more, but a difference of this magnitude lasts less than $0.2N$

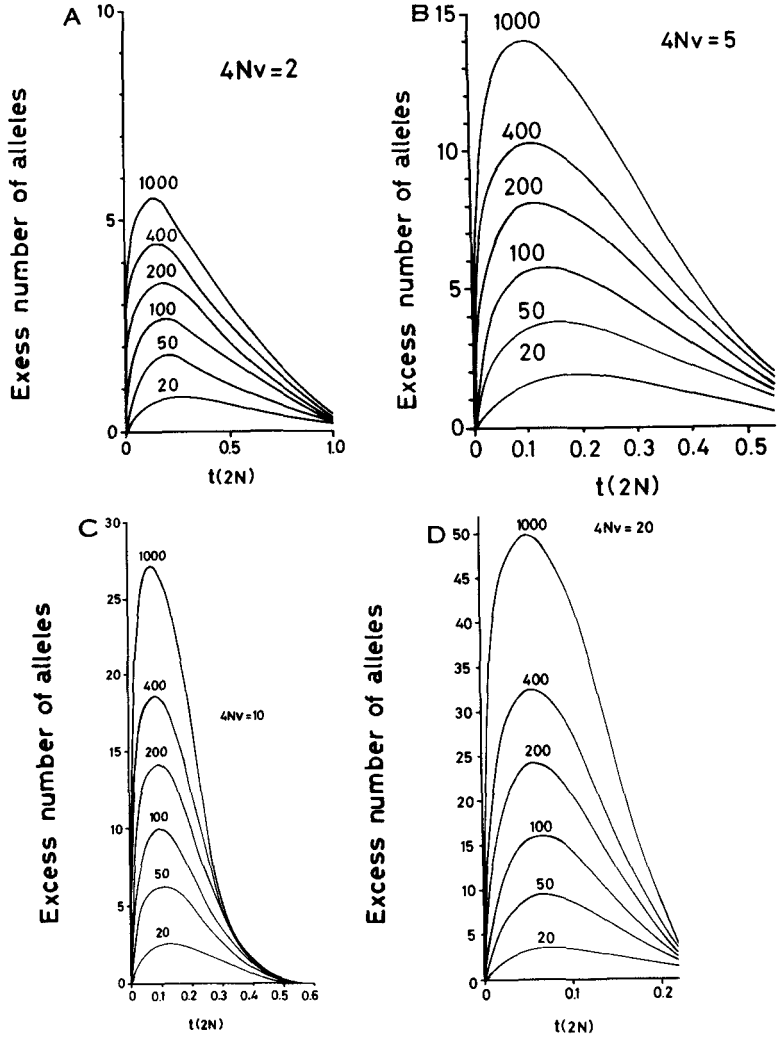


FIGURE 2.—Excess number of alleles to be seen between the actual number of alleles and that based on the equilibrium assumption. Numbers over the curves represent the sample size. a, $4Nv = 2$; b, $4Nv = 5$; c, $4Nv = 10$; d, $4Nv = 20$.

generations. If $4Nv = 5$, a significant difference persists for about $0.5N$ generations.

The difference in the number of alleles expected for the two assumptions (evolving populations *vs.* steady state) is shown in Figure 2, for various combinations of $4Nv$ and sample size. The graphs reveal several features of biological importance. The difference between the two expectations becomes larger as the sample size increases. For instance, when $4Nv = 10$, the maximum difference changes from approximately 2.5 alleles, when $M = 20$ genes, to about 27 alleles when $M = 1000$. This has an important implication for survey design. When an excess of the number of alleles is observed, the amount of the excess increases as the sample size increases, if the population being sampled is actually not in

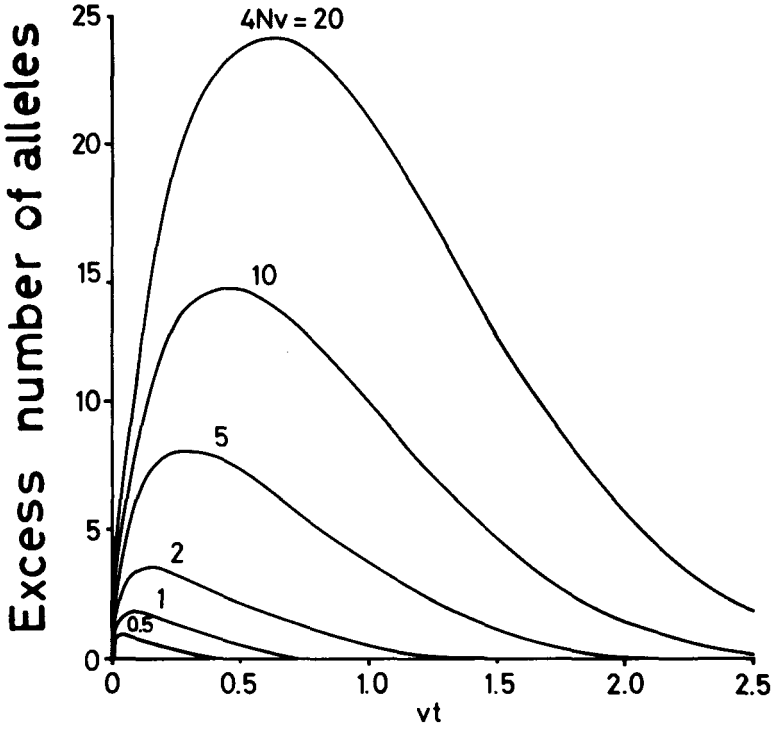


FIGURE 3.—The excess number of alleles and the $4Nv$ value. vt, The mutation rate ($v$) times the number of generation ($t$).

steady state equilibrium. Studies such as those of OHTA (1976) and CHAKRABORTY, FUERST and NEI (1978) did not analyze the effect of increasing sample size on the excess of alleles observed. Future studies should take this into account.

Figures 1 and 2 reveal that the difference from steady state expectations increases most rapidly soon after the population begins to evolve. This tendency is most pronounced when the sample size and the value of $4Nv$ are both large. The peak occurs slightly later (in terms of $2N$ generations) as $4Nv$ decreases. The excess lasts longer for a large value of $M$.

Although a study of Figure 2 suggests that the excess of total alleles will persist only for a short time (in terms of $2N$ generations), this period (in terms of absolute generations) may be substantial. In Figure 3 we present an example that compares the excesses observed in populations with different values of $N$ when we fix the value of the mutation rate. As can be seen in Figure 3, the actual duration of an excess will strongly depend upon the size of the ultimate equilibrium population. With a large population this excess may be sustained over a long evolutionary time.

## THE NUMBER OF ALLELES EACH PRESENT SINGLY

Among the statistics that we can test, some of the most interesting deal with alleles that appear only rarely, or even only once, in the sample. Each generation

$2Nv$ new mutant alleles are introduced into the population. Of course, most of these new mutants become extinct within a few generations, before reaching a frequency high enough to be included in a small sample. When $2Nv$ is large, the population accumulates mutant genes rather rapidly, and some will appear in survey samples. In a previous paper (MARUYAMA and FUERST 1983), we have shown that, if $4Nv$ is of the order of 0.5 to 5, the first arrival time for a new mutant allele to reach an intermediate frequency is not large. This first arrival time is approximately equal to a small (0.1–0.3) fraction of $N$ generations. From this, we expect that the frequency classes close to zero might attain an equilibrium number of alleles more rapidly than regions farther from zero. More alleles may be found at low frequencies in an evolving population than in a population that is in equilibrium between loss of alleles due to drift and input of alleles due to mutation. Consequently, there will be more alleles in samples of genes taken from each population. NEI (1979) investigated this problem for a steady state population. Here, we will present a study of this problem for an evolving population.

With the approach based on numerical integration of (3), we set the initial condition

$$u_{oj} = M\left(\frac{j}{l}\right)\left(1 - \frac{j}{l}\right)^{M-1}. \tag{12}$$

Using the difference equation (6), the initial condition (12) and the boundary condition (7), we can calculate $u_{ij}$ which is the probability that the particular allele will be present only once in the sample, given that the frequency of that allele is $j\Delta x$ at time zero. The remainder of the calculations needed for this problem are the same as those used for the total number of alleles. Namely,

$$u_{\gamma l} + 4N^2v\Delta t \sum_{i=1}^{l} u_{\gamma-i,1} \tag{13}$$

gives the number of alleles each of which will be present singly. The average number of singly present alleles in a sample of $M$ genes taken from an equilibrium population is given by

$$\frac{4NvM}{M + 4Nv - 1} \tag{14}$$

while, similarly, the average number of alleles each present $k$ times is

$$\frac{4NvM(M-1)\ldots(M-k+1)}{K(M+4Nv-1)(M+4Nv-2)\ldots(M+4Nv-k)}.$$

We can examine how alleles singly present are found in the population at various times after the start of the mutational process. Examples are given in Figure 4. The population rapidly reaches a plateau with respect to the number of such alleles, and this number changes little after the plateau is attained. Since formula (14) gives easily the expected number of singly present alleles we calculated the excess of singly present alleles assuming the population is at a steady state with the measured amount of variability.

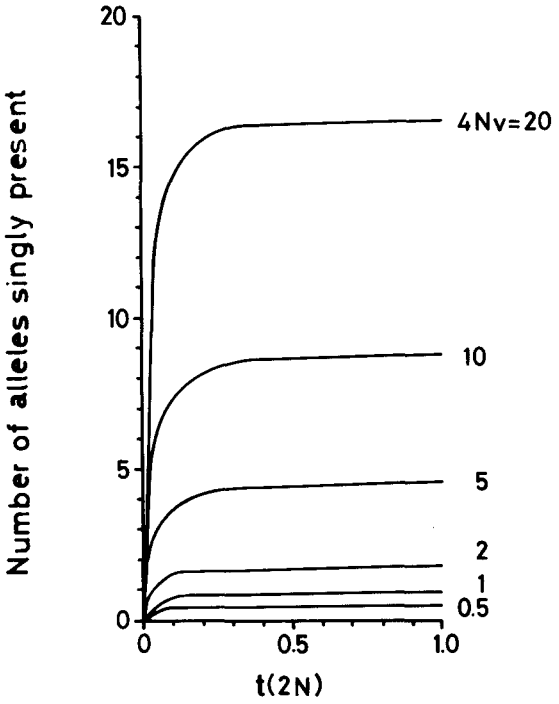


FIGURE 4.—The number of alleles each present singly in sample. $M = 100$.

Assuming that the population is in a steady state at the observed level of genetic variability, we calculated from formulas (13) and (14) the excess of singly present alleles. Figure 5 presents some examples of the mean excess for sample size 100. It was found that, as $4Nv$ increases, the excess of singly present alleles makes up a larger fraction of the total excess. When $4Nv$ is larger than 5, and when the sample size is reasonably large, more than a half of the total excess is due to the excess of singly present alleles. This confirms the emphasis of NEI and LI (1976) that in an expanding population the number of rare alleles should be in great excess.

## EXPANSION FROM A NONHOMOALLELIC STATE

In the preceding sections, we assumed that the population under consideration expanded from a completely homoallelic population. This describes one extreme case of a postbottleneck population. The model can be used to place bounds on the more realistic case in which genetic variability is not lost completely. It is, however, not difficult to extend the analysis to situations in which the starting population has nonzero variability. Such a model was analyzed by NEI and LI (1976) using an eigenfunction series expansion. Here, we will present an analysis based on the difference equation method.

Suppose that $\phi(x)$ is the frequency spectrum in a starting population. Then, the number of alleles in an evolving population can be obtained using formula (8), where $u_{\gamma,l}$ is replaced by

$$\sum_{j=1}^{M} u_{\gamma,j}\phi(x_j) \tag{15}$$

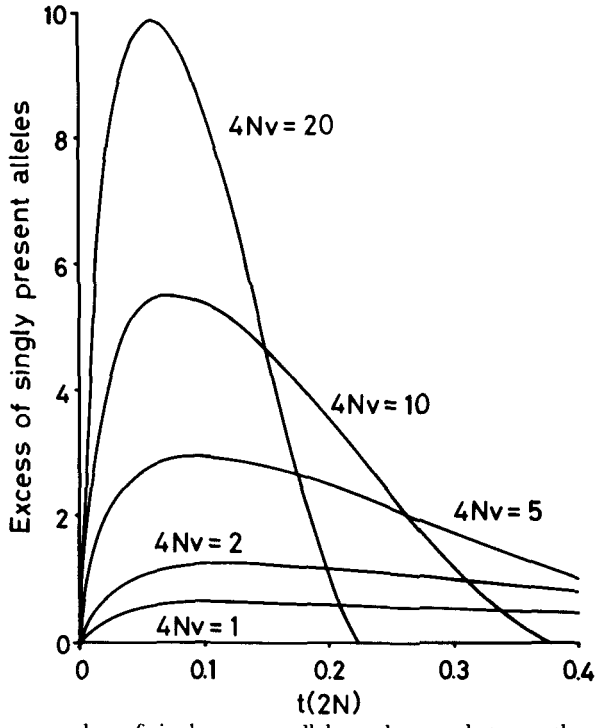


FIGURE 5.—Excess number of singly present alleles to be seen between the actual number and that based on the equilibrium assumption. $M = 100$.

and $u_{\gamma j}$ is the solution of difference equation (3) satisfying the boundary conditions (4) and (5). The quantity given by (15) is simply the total number of alleles carried over from the founder population which appear in a sample of $M$ genes sampled at time $t$.

If the founder population is in equilibrium between mutation and drift, and with population size equal to $N_o$, the frequency spectrum at time ($t = 0$) is given by

$$\phi(x) = 4N_o v x^{-1}(1 - x)^{4N_o v - 1}$$

where $\phi(x)dx$ is the number of alleles whose frequency at $t = 0$ is in the range ($x$, $x + dx$).

Of course, for the case of a true bottleneck, the population will not be in an equilibrium state because of the sampling process involved in going through the bottleneck. This does not mean that we cannot obtain the allele frequency spectrum for such cases, and in fact we will present results obtained using stochastic integrals in a future paper (P. A. FUERST and T. MARUYAMA, unpublished results). If the initial spectrum is known, even if this is not for an equilibrium population, it can be used in (14). For instance, if the initial population has two alleles with frequency $x_o$ and $(1 - x_o)$, formula (15) reduces to

$$u_{\gamma j_1} + u_{\gamma j_2}$$

where $j_1 \Delta x = x_o$ and $j_2 \Delta x = 1 - x_o$.

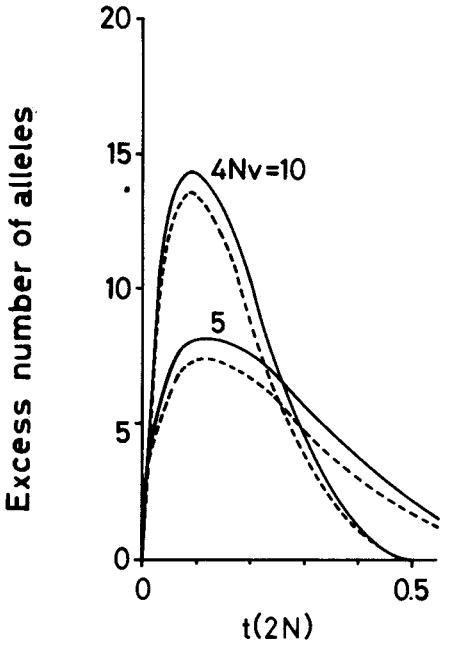


FIGURE 6.—The excess number of alleles calculated on the assumption that the starting population has 0.1 heterozygosity. Solid curves, Starting populations are in equilibrium; Dashed curves, starting populations consist of loci of zero variability and of those of 0.5 heterozygosity.

We have used this approach to study two extreme cases of a population expanding from nonzero variability. In the first case, we assume that the population is small but in equilibrium between mutation and drift.

We can, therefore, use the allele frequency spectrum of the equilibrium population as a starting point for the study. In the second case, we assume that the population is not in equilibrium. To do this, we use the most extreme case possible; one or more loci are started with two alleles, each having frequency 0.5, with all remaining loci started in a homoallelic state. This would be analogous to sampling a single founding individual to form a population.

In both situations which we studied, the populations rapidly accumulate new mutations. It should be kept in mind that, especially in the situation when the population is initiated in a nonequilibrium state, the average heterozygosity of the population will be considerably larger than it would be in the case of a recently homoallelic population. In addition, the number of rare alleles in the first few generations would be less than expected for a population with equivalent average heterozygosity. Nevertheless, the results are changed only negligibly from those obtained for homoallelic populations. An excess in the number of alleles is seen for all values of $4Nv$ studied, compared with the number of alleles expected in an equivalent equilibrium population. The slight deficiency of alleles initially seen for the population started at nonequilibrium is so transient, and of such low magnitude, that it has not been plotted in these figures. In Figure 6 the excess in total number of alleles for the initial populations in equilibrium with low variability is given by the solid curves, whereas the excess for the

populations in nonequilibrium state with the same heterozygosity level is given by the dashed curves.

It can be seen from these figures that the availability of preexisting genetic variability does little to dampen the increase in the number of alleles that are to be found in a rapidly expanded population. The increase in the number of alleles follows a similar time course for all comparable cases shown in Figures 3 and 4. The increases in the number of alleles are almost the same when there is a small amount of variation as when there is no variation in the population. The situations used to generate the model for Figure 6 with low variability, and that for nonequilibrium low variability, represent the extreme types of populations which are significantly different from those in Figure 2, in which the population begins with no variation. It appears, therefore, that Nei's speculation that there will be an increase in the number of alleles following a population bottleneck may be correct, at least when the population is very rapidly expanded following the restriction. In a subsequent study (T. MARUYAMA and P. A. FUERST, unpublished results), we will show that under some situations the action of restricting population size will result in a transient decrease in the expected number of alleles in a population. There are thus interacting forces that are determining allele numbers in a population. However, this may last only a very short time, because the loss of alleles due to population reduction will dissipate much faster than the time period during which the excess of alleles is observed due to population expansion. We are carrying out extensive computer simulations based on the stochastic integral method to investigate the spectral change in a bottleneck population. We also intend to apply the analysis developed here to reveal more detailed features of the transient behavior of the number of alleles in a postbottleneck population.

## DISCUSSION

A point of biological interest concerns the probability that the allele that existed in the original homoallelic population is found in a sample of genes taken from the evolving population. This allele is the only connection between the evolving population and the original population from which it was derived. This probability is equal to the first term, $u_{\gamma,l}$, on the right side of equation (8). This is directly related to the incidence of finding a common allele in two evolving populations derived from a common ancestral homoallelic population. Figure 7 shows the relationship between the probability of retaining the original allele and the value of $4Nv$, assuming a sample size, $M$, of 200 genes. Examining other sample sizes, we found that the probability of retaining the original allele is nearly independent of sample size ($M$), once $M$ has attained even a fairly small value such as 20 (ten individuals). Unless the sample is very small, the probability of persistence decays faster as $4Nv$ gets larger. It also shows that the time at which the persistence probability is equal to 0.5 is, approximately, inversely proportional to the value of $4Nv$. Note that the area under the curve in Figure 7 is equal to the average age of a mutant whose frequency is 1.0 when it is introduced. This is because the age is a time-reversible process, the time required

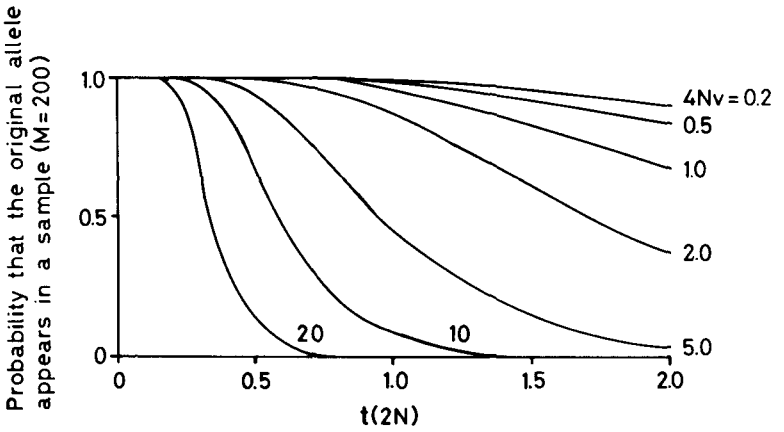


FIGURE 7.—Probability that an originally present allele appears in a sample taken from an evolving population.

for $x$ to change from a low value ($x = 0$) to 1 being the same as that for the reversed process (MARUYAMA and KIMURA 1974; WATTERSON 1977; NAGASAWA and MARUYAMA 1979).

We must be cautious in the interpretation of the results presented in this paper, particularly when extrapolating them to the bottleneck situation. NEI, MARUYAMA and CHAKRABORTY (1975) have shown that, even in a rather extreme case of bottleneck, a population does not lose all of its genetic variability but rather that a substantial fraction of the variability is retained in the population following the bottleneck. We will show in a subsequent paper of this series (P. A. FUERST and T. MARUYAMA, unpublished results) that the population easily loses its rare alleles but that many intermediate frequency (polymorphic) alleles are retained following the bottleneck. Therefore, as far as losses of gene diversity and alleles are concerned, the loss of alleles will be much more drastic. If all rare alleles, but many intermediate alleles, are lost from the population there may actually be a strong deficiency of alleles for a few generations following the bottleneck. Nevertheless, our studies on populations that begin at nonequilibrium, but with some variability, suggest that an excess of alleles will quickly appear.

The next paper in this series (T. MARUYAMA and P. A. FUERST, unpublished results) will deal with the model labeled (1) in the introduction, and show that a population size reduction does, in fact, yield a deficiency of alleles following the change in population size. Only a complete treatment of the problem, which we hope to provide with subsequent papers in this series, will allow the determination of the total change in allelic expectations.

We are grateful to M. NEI to whom we owe a number of substantial improvements in this paper. We would like also to thank two anonymous reviewers, T. GOJOBORI and G. WATTERSON, for their comments on an earlier version of this paper. This study was supported by research grant DEB-8110220 from the United States National Science Foundation and by grant 57120001 from the Japanese Ministry of Education, Science and Culture.

## LITERATURE CITED

CARSON, H. L., 1968 The population flush and its genetic consequences. pp. 123–137. In: *Population Biology and Evolution*, Edited by R. C. LEWONTIN. Syracuse University Press, New York.

CHAKRABORTY, R., P. A. FUERST and M. NEI, 1978 Statistical studies on protein polymorphism in natural populations. II. Gene differentiation between populations. Genetics **88:** 367–390.

CROW, J. F. and M. KIMURA, 1956 Some genetic problems in natural populations. pp. 1–22. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. University of California Press, Berkeley, California.

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

EWENS, W. J. and K. KIRBY, 1975 The eigenvalues of the neutral alleles processes. Theor. Pop. Biol. **7:** 212–220.

GRIFFITHS, R. C., 1979a A transition density expansion for a multi-allele diffusion model. Adv. Appl. Prob. **11:** 310–325.

GRIFFITHS, R. C., 1979b Exact sampling distributions from the infinite neutral alleles model. Adv. Appl. Prob. **11:** 326–354.

GRIFFITHS, R. C., 1980 Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theor. Pop. Biol. **17:** 37–50.

KARLIN, S. and H. AVNI, 1975 Derivation of the eigenvalues of the configuration process induced by a labeled direct product branching process. Theor. Pop. Biol. **7:** 221–228.

KIMURA, M., 1955 Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp. Quant. Biol. **20:** 33–53.

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

MARUYAMA, T., 1977 *Stochastic Problems in Population Genetics.* Springer-Verlag, New York.

MARUYAMA, T., 1980 On the overdominant model of population genetics. Adv. Appl. Prob. **12:** 274–275.

MARUYAMA, T., 1981 Stochastic problems in population genetics. pp. 154–161. In: *Stochastic Nonlinear Systems*, Edited by L. ARNOLD and R. LEFEVER. Springer-Verlag, New York.

MARUYAMA, T., 1982 Stochastic integrals and their application to population genetics. pp. 151–166. In: *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, Edited by M. KIMURA. Springer-Verlag, Berlin.

MARUYAMA, T., 1983 Stochastic theory of population genetics. Bull. Math. Biol. **45:** 521–554.

MARUYAMA, T. and P. A. FUERST, 1983 Analyses of the age of genes and the first arrival times in a finite population. Genetics **105:** 1041–1059.

MARUYAMA, T. and M. KIMURA, 1974 A note on the speed of gene frequency changes in reverse directions in a finite population. Evolution **28:** 151–163.

MARUYAMA, T. and M. NEI, 1981 Genetic variability maintained by mutation and overdominant selection in finite populations. Genetics **98:** 441–459.

NAGASAWA, M. and T. MARUYAMA, 1979 An application of time reversal of Markov processes to a problem of population genetics. Adv. Appl. Prob. **11:** 457–478.

NEI, M., 1979 Stochastic theory of population genetics and evolution. pp. 17–47. In: *Vito Volterra Symposium on Mathematical Models in Biology* (Lectures Notes in Biomathematics no. 39), Edited by C. BARIGOZZI. Springer-Verlag, New York.

NEI, M. and W. H. LI, 1976 The transient distribution of allele frequencies under mutation pressure. Genet. Res. **28:** 205–214.

NEI, M., T. MARUYAMA and R. CHAKRABORTY, 1975 The bottleneck effect and genetic variability in populations. Evolution **29:** 1–10.

NEI, M., MARUYAMA, T. and C. I. WU, 1983 Models of evolution of reproductive isolation. Genetics **103:** 557–579.

OHTA, T., 1976 The role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor. Pop. Biol. **10:** 254–275.

WATTERSON, G. A., 1977 Reversibility and the age of an allele. II. Two-allele models, with selection and mutation. Theor. Pop. Biol. **12:** 179–196.

WATTERSON, G. A., 1983 Allele frequencies after a bottleneck. Statistics Research Report no. 83, Monash University, Victoria.

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

WRIGHT, S., 1948 Genetics of populations. pp. 111, 111A–D, and 112. *Encyclopedia Britannica*, Ed. 14, Vol. 10.

Corresponding editor: M. NEI