THE EFFECT OF INTRAGENIC RECOMBINATION ON THE NUMBER OF ALLELES IN A FINITE POPULATION*

CURTIS STROBECK AND KENNETH MORGAN

Department of Genetics, University of Alberta, Edmonton, Alberta, Canada T6G 2E9

Manuscript received August 24, 1977
Revised copy received November 28, 1977

ABSTRACT

A two-site infinite allele model is constructed to study the effect of intragenic recombination on the number of neutral alleles and the distribution of their frequencies in a finite population. The results of theory and Monte Carlo simulation of the two-site model demonstrate that intragenic recombination significantly increases the mean and variance of the number of alleles when the rates of mutation and recombination are as large as the reciprocal of the population size. Data from natural populations indicate that this may be a significant process in generating variation and determining its distribution.

IT has been known for some time that intragenic recombination might be as powerful an agent for generating new variation as is mutation (WATT 1972). In addition, KOEHN and EANES (1976) conjectured that such a process would maintain a number of rare alleles in excess of that predicted by neutral allele theory alone. The belief in what is called the classical view of evolution (LEWONTIN 1974) probably accounts for a lack of interest among population geneticists in intragenic recombination; for if the population is homozygous at most loci, then intragenic recombination has no effect. Although it was known that a large amount of genetic variation might not be detected by electrophoretic surveys of natural populations (LEWONTIN and HUBBY 1966), it is just now becoming clear how much variation is concealed (COYNE 1976; SINGH, LEWONTIN and FELTON 1976).

If there exists a large amount of variation, intragenic recombination would be important in population genetics only if the rate is large enough to be an effective means of generating additional variation. Therefore, it is important to know what rate of intragenic recombination is required to generate an increased amount of variation in a finite population. Furthermore, even if intragenic recombination generates an increased amount of variation, it must be shown that the effect of intragenic recombination is qualitatively different from that of an increased mutation rate. Therefore, it is important to know what the distribution of allele frequencies is in a population in which intragenic recombination is an important source of variation.

Intragenic recombination can occur both by crossing over and by a non-

* This work was supported by National Research Council of Canada, Grant No. A0502.

reciprocal process, gene conversion. If gene conversion is not symmetric, then it acts almost like meiotic drive (GUTZ and LESLIE 1976). If gene conversion is symmetric, it can be thought of as a reduced rate of reciprocal recombination from a population point of view. Since the interest here is to determine the rate of intragenic recombination necessary to increase the number of alleles and affect the distribution of alleles, it is assumed throughout this paper that gene conversion is symmetric.

A model of intragenic recombination with a large number of sites at which mutation can occur is probably impossible to construct. However, many of the above question can be answered by using a two-site model. If intragenic recombination is symmetric, then such a model is equivalent to a two-locus model with each locus considered as a mutable site and each gametic type as a different allele.

THEORY

The two-site model used is equivalent to the infinite allele (KIMURA and CROW 1964) two-locus model of random union of gametes (KARLIN and MCGREGOR 1968). It assumes a finite population of $2N$ gametes. In order to generate a gamete in the next generation, two gametes are selected at random with replacement from the existing $2N$ gametes. For any two arbitrary gametes denoted by a_1b_1 and a_2b_2, one of the four meiotic products a_1b_1, a_1b_2, a_2b_1, and a_2b_2 is selected with probability $\frac{1}{2}(1-r)$, $\frac{1}{2}r$, $\frac{1}{2}r$, and $\frac{1}{2}(1-r)$, respectively, where r is the recombination value. This process is then repeated until $2N$ new gametes have been generated.

Five inbreeding coefficients or descent measures are needed to describe the behavior of the system from one generation to the next. Three of the inbreeding coefficients involve two gametes chosen at random without replacement from the $2N$ gametes; one coefficient with three gametes; and one coefficient with four gametes. If an arbitrary gamete is denoted by a_ib_i, then the five inbreeding coefficients are:

For two gametes a_ib_i $(i=1,2)$:

$\Phi_A = P(a_1 \equiv a_2)$ = Probability a_1 is identical by descent to a_2

$\Phi_B = P(b_1 \equiv b_2)$ = Probability b_1 is identical by descent to b_2

$\Phi_{AB} = P(a_1 \equiv a_2,\ b_1 \equiv b_2)$ = Probability a_1 is identical by descent to a_2 and b_1 is identical by descent to b_2

For three gametes a_ib_i $(i=1,2,3)$:

$\Gamma_{AB} = P(a_1 \equiv a_2,\ b_1 \equiv b_3)$ = Probability a_1 is identical by descent to a_2 and b_1 is identical by descent to b_3

For four gametes a_ib_i $(i=1,2,3,4)$:

$\Delta_{AB} = P(a_1 \equiv a_3,\ b_2 \equiv b_4)$ = Probability a_1 is identical by descent to a_3 and b_2 is identical by descent to b_4

These are the same five variables that have been used for the two-locus model with random mating of zygotes (WEIR and COCKERHAM 1974; SERANT 1974).

If there is no mutation, the recurrence equations for the five inbreeding coefficients are:

$$\Phi_A' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\Phi_A$$

$$\Phi_B' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\Phi_B$$

$$\Phi_{AB}' = (1-r)^2\left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\Phi_{AB}\right] + 2r(1-r)\Gamma + r^2\Delta \qquad (1)$$

$$\Gamma_{AB}' = (1-r)\Gamma + r\Delta$$

$$\Delta_{AB}' = \Delta$$

where

$$\Gamma = \frac{1}{(2N)^2}\{1 + (2N-1)(\Phi_A+\Phi_B+\Phi_{AB}) + (2N-1)(2N-2)\Gamma_{AB}\}$$

$$\Delta = \frac{1}{(2N)^3}\{2N + 2(2N-1)(\Phi_A+\Phi_B+\Phi_{AB}) + (2N-1)(2N-2)(\Phi_A+\Phi_B+4\Gamma_{AB}) + (2N-1)(2N-2)(2N-3)\Delta_{AB}\}$$

which can be obtained from Table 1 and Table 2.

The equilibrium values of the five inbreeding coefficients are

$$\hat{\Phi}_A = \hat{\Phi}_B = \hat{\Phi}_{AB} = \hat{\Gamma}_{AB} = \hat{\Delta}_{AB} = 1$$

if there is no mutation; that is, the population is homozygous at both sites. The

TABLE 1

The probability of occurrence and the value of the inbreeding coefficient for each arrangement of three gametes randomly sampled with replacement from a population of 2N gametes (α, β, and γ are different gametes)

Arrangement 1	2	3	Probability	$P(a_1 \equiv a_2, b_1 \equiv b_3)$
α	α	α	$\frac{1}{4N^2}$	1
α	β	α	$\frac{2N-1}{4N^2}$	Φ_A
α	α	β	$\frac{2N-1}{4N^2}$	Φ_B
α	β	β	$\frac{2N-1}{4N^2}$	Φ_{AB}
α	β	γ	$\frac{(2N-1)(2N-2)}{4N^2}$	Γ_{AB}

TABLE 2

The probability of occurrence and the value of the inbreeding coefficient for each arrangement of four gametes randomly sampled with replacement from a population of 2N gametes (α, β, γ, and δ are different gametes)

Arrangement(s) 1 3 / 2 4	Probability	$P(a_1 \equiv a_3, b_2 \equiv b_4)$
$\alpha\ \alpha$ / $\alpha\ \alpha$	$\frac{1}{8N^3}$	1
$\alpha\ \alpha$ / $\beta\ \beta$	$\frac{2N-1}{8N^3}$	1
$\alpha\ \beta\quad \beta\ \alpha$ / $\alpha\ \alpha\quad \alpha\ \alpha$	$\frac{2N-1}{4N^3}$	Φ_A
$\alpha\ \alpha\quad \alpha\ \alpha$ / $\alpha\ \beta\quad \beta\ \alpha$	$\frac{2N-1}{4N^3}$	Φ_B
$\alpha\ \beta\quad \alpha\ \beta$ / $\alpha\ \beta\quad \beta\ \alpha$	$\frac{2N-1}{4N^3}$	Φ_{AB}
$\beta\ \gamma$ / $\alpha\ \alpha$	$\frac{(2N-1)(2N-2)}{8N^3}$	Φ_A
$\alpha\ \alpha$ / $\beta\ \gamma$	$\frac{(2N-1)(2N-2)}{8N^3}$	Φ_B
$\alpha\ \beta\quad \alpha\ \beta$ / $\alpha\ \gamma\quad \gamma\ \alpha$ / $\beta\ \alpha\quad \beta\ \alpha$ / $\alpha\ \gamma\quad \gamma\ \alpha$	$\frac{(2N-1)(2N-2)}{2N^3}$	Γ_{AB}
$\alpha\ \beta$ / $\gamma\ \delta$	$\frac{(2N-1)(2N-2)(2N-3)}{8N^3}$	Δ_{AB}

rate to homozygosity is $1-\lambda_0$, where λ_0 is the largest eigenvalue of the Jacobian evaluated at this equilibrium. The eigenvalues are:

$$\lambda_1 = \lambda_2 = 1 - \frac{1}{2N}$$

and λ_3, λ_4, and λ_5 are $\left(1 - \frac{1}{2N}\right)$ times the roots of the cubic equation

$$\begin{aligned} &2N^3x^3 - \\ &N(2N^2r^2-6N^2r+6N^2-2Nr^2+8Nr-7N+r^2-4r+3)x^2 + \\ &(1-r)(N-1)(2N^2r^2-6N^2r+6N^2-4Nr^2+9Nr-8N+r^2-3r+3)x - \\ &(1-r)^3(N-1)^2(2N-3) = 0 \end{aligned} \tag{2}$$

which is the same polynomial as equations (16) and (17) in KARLIN and

McGregor (1968). They have shown that the largest root of equation (2) decreases from $\left(1-\frac{1}{2N}\right)$ to $\left(1-\frac{1}{2N}\right)^2$ as r varies from 0 to 1.

Therefore the largest eigenvalue is $\lambda_0 = \left(1-\frac{1}{2N}\right)$.

If the mutation rate at each of the two sites is ν and each mutation is unique (the infinite allele model of Kimura and Crow, 1964), then the recurrence equations are:

$$\begin{aligned}
\Phi_A' &= (1-\nu)^2\left[\frac{1}{2N} + \left(1-\frac{1}{2N}\right)\Phi_A\right] \\
\Phi_B' &= (1-\nu)^2\left[\frac{1}{2N} + \left(1-\frac{1}{2N}\right)\Phi_B\right] \\
\Phi_{AB}' &= (1-\nu)^4\left\{(1-r)^2\left[\frac{1}{2N} + \left(1-\frac{1}{2N}\right)\Phi_{AB}\right] + 2r(1-r)\Gamma + r^2\Delta\right\} \qquad (3) \\
\Gamma_{AB}' &= (1-\nu)^4[(1-r)\Gamma + r\Delta] \\
\Delta_{AB}' &= (1-\nu)^4\Delta
\end{aligned}$$

where Γ and Δ are as defined above. The equilibrium values can be obtained using Cramer's rule. Three different cases are considered:

1. $N >> 1$, $\nu \simeq O\left(\frac{1}{N}\right)$, and $r << \nu$

$$\begin{aligned}
\hat{\Phi}_A &\simeq \frac{1}{1+4N\nu} \\
\hat{\Phi}_B &\simeq \frac{1}{1+4N\nu} \qquad (4) \\
\hat{\Phi}_{AB} &\simeq \frac{1}{1+8N\nu} \\
\hat{\Gamma}_{AB} &\simeq \frac{3+20N\nu}{(1+4N\nu)(1+8N\nu)(3+8N\nu)} \\
\hat{\Delta}_{AB} &\simeq \frac{9+72N\nu+64N^2\nu^2}{(1+4N\nu)(1+8N\nu)(3+4N\nu)(3+8N\nu)}
\end{aligned}$$

2. $N >> 1$, $\nu \simeq O\left(\frac{1}{N}\right)$, $r \simeq \nu$

$$\begin{aligned}
\hat{\Phi}_A &\simeq \frac{1}{1+4N\nu} \\
\hat{\Phi}_B &\simeq \frac{1}{1+4N\nu} \qquad (5)
\end{aligned}$$

$$\hat{\Phi}_{AB} \simeq \frac{128N^3\nu^3+32N^3\nu^2r+176N^2\nu^2+48N^2\nu r+8N^2r^2+72N\nu+26Nr+9}{(1+4N\nu)(256N^3\nu^3+192N^3\nu^2r+32N^3\nu r^2+320N^2\nu^2+152N^2\nu r+8N^2r^2+108N\nu+26Nr+9)}$$

$$\hat{\Gamma}_{AB} \simeq \frac{80N^2\nu^2+48N^2\nu r+8N^2r^2+72N\nu+26Nr+9}{(1+4N\nu)(256N^3\nu^3+192N^3\nu^2r+32N^3\nu r^2+320N^2\nu^2+152N^2\nu r+8N^2r^2+108N\nu+26Nr+9)}$$

$$\hat{\Delta}_{AB} \simeq \frac{64N^2\nu^2+48N^2\nu r+8N^2r^2+72N\nu+26Nr+9}{(1+4N\nu)(256N^3\nu^3+192N^3\nu^2r+32N^3\nu r^2+320N^2\nu^2+152N^2\nu r+8N^2r^2+108N\nu+26Nr+9)}$$

3. $N >> 1$, $\nu \simeq O\left(\frac{1}{N}\right)$, $r >> \nu$

$$\hat{\Phi}_A \simeq \frac{1}{1+4N\nu}$$

$$\hat{\Phi}_B \simeq \frac{1}{1+4N\nu}$$

$$\hat{\Phi}_{AB} \simeq \frac{1}{(1+4N\nu)^2} \qquad (6)$$

$$\hat{\Gamma}_{AB} \simeq \frac{1}{(1+4N\nu)^2}$$

$$\hat{\Delta}_{AB} \simeq \frac{1}{(1+4N\nu)^2}$$

Since in all three cases it is assumed that $N >> 1$, these results are the same as those obtained by SERANT (1974) for the two-locus model with random mating of zygotes.

HILL (1975) used the two-locus infinite allele model with random union of gametes to study linkage disequilibrium in a finite population. The results in equation (10) of HILL and the results in (5) of this paper can be obtained from each other by a linear transformation. This is true since, if N is large, then:

$$\Phi_A \simeq 1 - H_A = 1 - \sum_{i \neq j} p_i p_j$$

$$\Phi_B \simeq 1 - H_B = 1 - \sum_{i \neq j} q_i q_j$$

$$\Phi_{AB} \simeq \sum_{i,j} f^2{}_{ij} = \sum_{i,j} (p_i q_j + D_{ij})^2$$

$$= 1 - H_A - H_B + H_A H_B + 2\sum_{i,j} p_i q_j D_{ij} + \sum_{i,j} D^2{}_{ij}$$

$$\Gamma_{AB} \simeq \sum_{i,j} p_i q_j f_{ij} = \sum_{i,j} p_i q_j (p_i q_j + D_{ij})$$

$$= 1 - H_A - H_B + H_A H_B + \sum_{i,j} p_i q_j D_{ij}$$

$$\Delta_{AB} \simeq \sum_{i,j} p_i q_j p_i q_j = 1 - H_A - H_B + H_A H_B$$

where p_i is the frequency of the i^{th} allele at the A locus, q_j is the frequency of the j^{th} allele at the B locus and $f_{ij} = p_i q_j + D_{ij}$ is the frequency of the gamete containing the i^{th} allele at the A locus and the j^{th} allele at the B locus.

The total mutation rate of the gene for the model of intragenic recombination is $\mu = 2\nu$ and, therefore, the effective number of alleles is $n_e = \frac{1}{\Phi_{AB}}$. The effective number of alleles for $N >> 1$ and $\mu \simeq O\left(\frac{1}{N}\right)$ is:

$1+4N\mu$ if $r << \mu$

$$\frac{(1+2N\mu)(32N^3\mu^3+48N^3\mu^2r+16N^3\mu r^2+80N^2\mu^2+76N^2\mu r+8N^2r^2+54N\mu+26Nr+9)}{16N^3\mu^3+8N^3\mu^2r+44N^2\mu^2+24N^2\mu r+8N^2r^2+36N\mu+26Nr+9} \quad \text{if } r \simeq \mu \qquad (7)$$

$$(1+2N\mu)^2 \quad \text{if } r >> \mu.$$

The effective number of alleles for various values of $N\mu$ and for several values of the ratio of μ to r is plotted in Figure 1. It is seen that:

(1) The effect of intragenic recombination increases as $N\mu$ increases;

(2) for small values of $N\mu$, there is essentially no effect; and

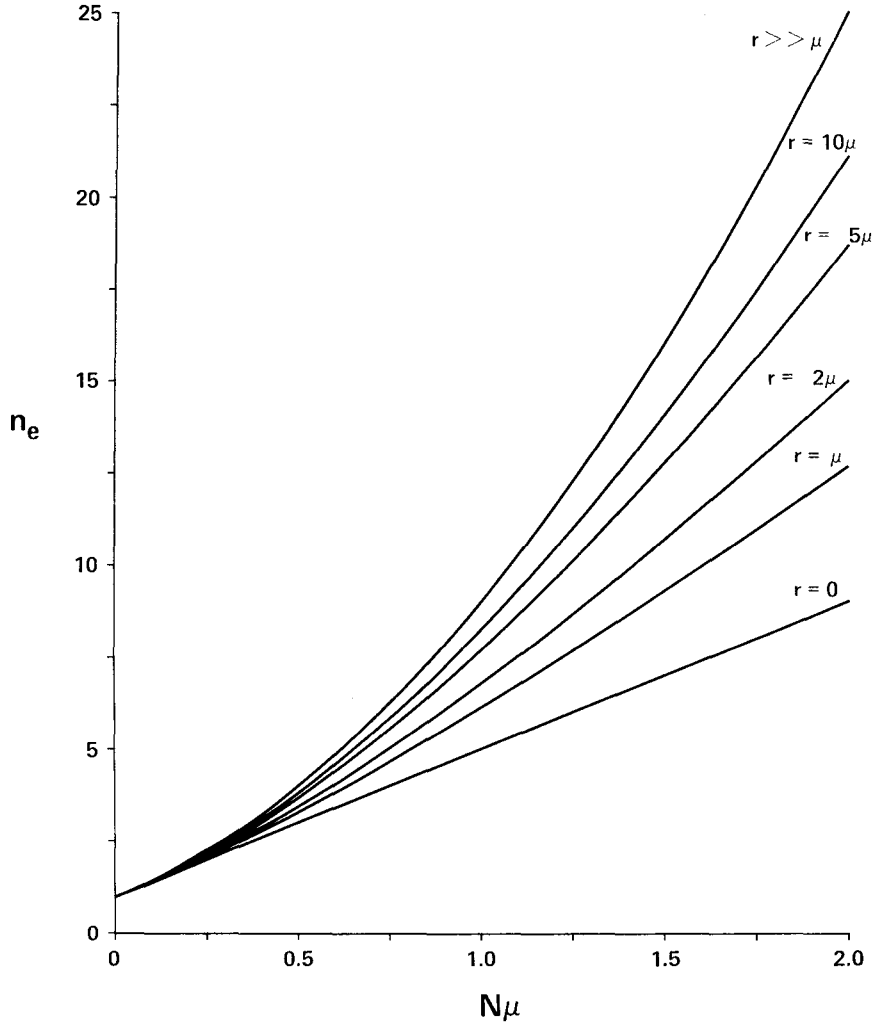

FIGURE 1.—The effective number of alleles for the two-site infinite allele model with intragenic recombination.

(3) a recombination value of the same order of magnitude as the mutation rate can have a large effect even for moderate values of $N\mu$.

MONTE CARLO SIMULATION

Since the effective number of alleles is a function of both the number of alleles and their frequencies, the increase in the effective number of alleles shown in Figure 1 might conceal a much greater change in the number of alleles. In order to derive the formulas for the mean and variance of the number of alleles and the variance of the homozygosity, the distribution of allelic frequencies is required. Because we are unable to obtain a formula for the allelic distribution, a Monte Carlo simulation of the model for intragenic recombination was undertaken.

In the Monte Carlo simulation, five pseudo-random numbers are used to generate each new gamete from the existing gametes. Two are used to select two parental gametes from the $2N$ existing gametes with replacement, one is used to determine which one of the four meiotic products is selected, and two are used to determine if a mutation occurs at each of the two mutable sites. This sequence is then repeated until $2N$ new gametes are selected, which then become the $2N$ gametes in the next generation.

For each run, the initial population was assumed to consist entirely of one allele. In order to insure that we were sampling from a stationary process, the first sample recorded was the $12N^{\text{th}}$ generation. Also to insure that the covariance between consecutive samples was minimal, the samples were taken $2N$ generations apart. For each sample the number of alleles and their frequencies were recorded. A total of 95 samples were analyzed for each run.

Two population sizes were used: $2N = 100$ and $2N = 200$. For each population size runs were made with $4N\mu = 1$ and $4N\mu = 4$ ($\mu = 2\nu = 0.005$ and $\mu = 2\nu = 0.02$ for $2N = 100$; $\mu = 2\nu = 0.0025$ and $\mu = 2\nu = 0.01$ for $2N = 200$). For each combination of population size and mutation rate, five different recombination values were used: $r = 0$, $r = \mu$, $r = 2\mu$, $r = 5\mu$, and $r = 10\mu$. Small values of the population size were chosen in the interest of economy for computer simulation, although they imply unrealistically high rates of mutation and intragenic recombination. Two replicate runs were made of each of the twenty combinations of population size, mutation rate, and recombination value.

The number of alleles and their frequencies at each of the two sites were also recorded for each of the 95 samples in every run. These data can be used as an internal check to see if the simulations are giving results that approximate those expected from the sampling theory of selectively neutral alleles (EWENS 1972). Although the simulations provide estimates of the population parameters, we choose to compare them to the sample statistics, with $n = N$, for the mean and variance of the number k of alleles present in any generation, as was done by EWENS and GILLESPIE (1974) in their neutral allele simulations. In Figure 2, the histograms for the observed and expected distributions of the number of alleles at one of the sites are shown for the four cases: (1) $2N = 100$ and

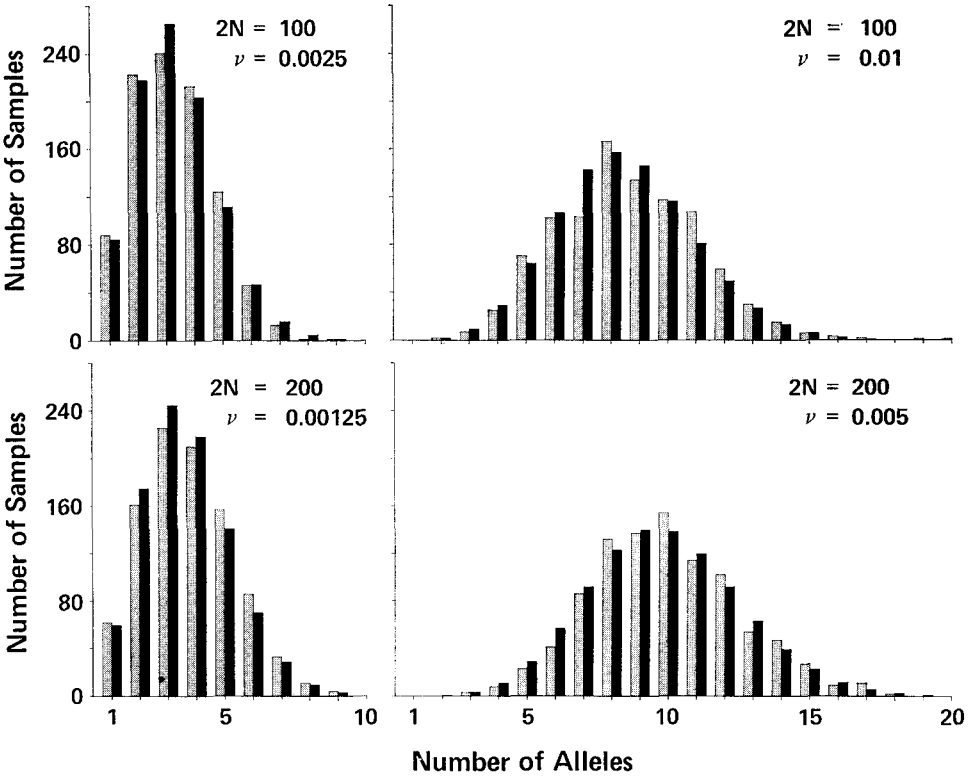

FIGURE 2.—Results of Monte Carlo simulation of the one-site infinite allele model. Frequency distribution of number of alleles: observed—shaded bars; expected—solid bars.

$\theta = 4Nv = 0.5$; (2) $2N = 100$ and $\theta = 4Nv = 2$; (3) $2N = 200$ and $\theta = 4Nv = 0.5$; and (4) $2N = 200$ and $\theta = 4Nv = 2$. For each of the four cases there are 950 samples (95 samples from 10 runs). The expected distribution was obtained using the program given in Appendix 3 of EWENS (1972). In each case it is seen that there is close agreement between the expected and observed distributions.

In Table 3 the mean and variance of the number of alleles for the gene (two-site) in each run are shown. For $r = 0$, the expected sample mean and variance of the number of alleles are (EWENS 1972):

Case		Mean	Variance
$2N$=100	θ=$4N\mu$=1	5.19	3.55
$2N$=100	θ=$4N\mu$=4	13.53	9.15
$2N$=200	θ=$4N\mu$=1	5.88	4.24
$2N$=200	θ=$4N\mu$=4	16.24	11.78

Except for the runs with $r = 0$, there is no theory to which these values can be compared. However, some important observations can be made from these data. It is seen that the mean number of alleles increases with increasing r. With

TABLE 3

Mean and variance of number of alleles from Monte Carlo simulation of the two-site infinite allele model

			$r=0$		$r=\mu$		$r=2\mu$		$r=5\mu$		$r=10\mu$	
$4N\mu$	2N	μ	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
1.0	100	0.005	5.29	5.21	5.59	3.14	6.16	4.43	6.08	5.35	6.45	6.63
			5.40	3.84	5.53	4.42	5.93	5.03	5.86	5.40	6.07	4.58
	200	0.0025	6.20	6.50	6.62	5.77	6.89	5.92	6.86	9.52	7.18	8.68
			6.21	3.66	6.36	4.51	6.48	6.17	6.40	6.03	7.89	9.46
4.0	100	0.02	13.93	9.30	16.94	13.12	18.45	19.25	20.31	33.83	22.54	35.49
			14.47	9.83	17.42	13.12	18.96	20.70	21.83	31.91	24.22	47.58
	200	0.01	17.07	10.75	20.42	15.14	22.06	24.32	25.74	50.11	28.23	67.41
			16.58	8.91	20.33	17.97	22.14	24.01	24.84	42.99	28.28	60.95

TABLE 4

Mean and variance of homozygosity from Monte Carlo simulation of the two-site infinite allele model

$4N\mu$	2N	μ	$r=0$ Mean	$r=0$ Variance	$r=\mu$ Mean	$r=\mu$ Variance	$r=2\mu$ Mean	$r=2\mu$ Variance	$r=5\mu$ Mean	$r=5\mu$ Variance	$r=10\mu$ Mean	$r=10\mu$ Variance
1.0	100	0.005	0.497	0.0498	0.486	0.0337	0.432	0.0442	0.440	0.0328	0.464	0.0470
			0.489	0.0473	0.508	0.0450	0.509	0.0437	0.480	0.0538	0.441	0.0298
	200	0.0025	0.482	0.0422	0.466	0.0443	0.447	0.0340	0.455	0.0362	0.455	0.0384
			0.487	0.0397	0.474	0.0375	0.487	0.0422	0.532	0.0511	0.409	0.0398
4.0	100	0.02	0.214	0.0069	0.159	0.0036	0.154	0.0068	0.148	0.0078	0.130	0.0041
			0.192	0.0056	0.170	0.0078	0.148	0.0029	0.132	0.0043	0.123	0.0048
	200	0.01	0.206	0.0093	0.172	0.0080	0.155	0.0058	0.132	0.0040	0.127	0.0047
			0.207	0.0083	0.169	0.0067	0.147	0.0041	0.138	0.0054	0.125	0.0054

$4N\mu = 1$, there is only a slight increase in the number of alleles with increasing r, while with $4N\mu = 4$ the increase is dramatic. Second, especially with $4N\mu = 4$, the variance increases much faster than the mean. In fact, the ratio of the variance to the mean becomes much greater than one. From the equation for the sample variance in EWENS (equation 24; 1972), it is easily seen that the ratio of the sample variance to the sample mean for neutral alleles is always less than one. This indicates that the sampling theory of selectively neutral alleles does not apply to the case of intragenic recombination.

In Table 4, the mean and variance of the homozygosity is shown for each run. Homozygosity is calculated as $\sum_i p_i^2$, where p_i is the frequency of the i^{th} allele. Expected values of the homozygosity are given by $\hat{\phi}_{AB}$ in equations (4), (5) and (6). The expected values for the two-site model for the 10 cases are given in Table 5. It is seen in Table 6 that the simulations are in good agreement with the expected values, although the observed values are consistently slightly larger than expected. In order to show that the distribution of alleles is not that predicted by the theory of neutral alleles, the observed variances of the homozygosity have been compared to those expected for selectively neutral alleles

$$\sigma^2 = \frac{2\theta}{(1+\theta)^2(2+\theta)(3+\theta)} \tag{8}$$

with $\theta = 4N\mu$ (WATTERSON 1974; STEWART 1976; KINGMAN 1977). Since $\theta = 4N\mu$ is only appropriate when $r = 0$, the formula above was used with $\theta = n_e - 1$, where the value of n_e is given by equations (7). Values of the variances calculated by equation (8) are given in Table 5. The ratio of the observed variance to the expected variance for each run is given in Table 6. Especially for $4N\mu = 4$, the ratio of the observed variance to that expected increases as r increases. KINGMAN (1977) has proved that whenever the distribution of alleles is that of EWENS' sampling theory, then the homozygosity must have the vari-

TABLE 5

Mean homozygosity calculated from the two-site model and variance of homozygosity calculated from the one-site model (see text)

$4N\mu$	Recombination rate	Mean (two-site)	Variance (one-site)
1.0	$r = 0$	0.5	0.0417
	$r = \mu$	0.4847	0.0401
	$r = 2\mu$	0.4758	0.0392
	$r = 5\mu$	0.4629	0.0378
	$r = 10\mu$	0.4552	0.037
4.0	$r = 0$	0.2	0.00762
	$r = \mu$	0.1633	0.00472
	$r = 2\mu$	0.1477	0.00369
	$r = 5\mu$	0.1301	0.00269
	$r = 10\mu$	0.1215	0.00226

TABLE 6

Ratios of observed (from Monte Carlo simulation) to expected means and variances of homozygosity; expected mean homozygosity calculated from the two-site model and expected variance calculated from the one-site model (see text)

			$r=0$		$r=\mu$		$r=2\mu$		$r=5\mu$		$r=10\mu$	
$4N\mu$	2N	μ	Ratio of means	Ratio of variances	Ratio of means	Ratio of variances	Ratio of means	Ratio of variances	Ratio of means	Ratio of variances	Ratio of means	Ratio of variances
1.0	100	0.005	0.99	1.19	1.00	0.84	0.91	1.13	0.95	0.87	1.02	1.27
			0.98	1.13	1.05	1.12	1.07	1.11	1.04	1.42	0.97	0.80
	200	0.0025	0.96	1.01	0.96	1.11	0.94	0.87	0.98	0.96	1.00	1.04
			0.97	0.95	0.98	0.94	1.02	1.08	1.15	1.35	0.90	1.08
4.0	100	0.02	1.07	0.90	0.97	0.76	1.04	1.83	1.14	2.91	1.07	1.81
			0.96	0.73	1.04	1.65	1.00	0.79	1.01	1.59	1.01	2.12
	200	0.01	1.03	1.22	1.05	1.69	1.05	1.56	1.02	1.48	1.05	2.08
			1.03	1.09	1.04	1.43	1.00	1.11	1.06	2.01	1.03	2.37

ance given by equation (8). These comparisons clearly demonstrate that the sampling theory of neutral alleles with intragenic recombination is not the same as the theory that assumes no intragenic recombination.

DISCUSSION

It has been demonstrated that intragenic recombination can significantly affect the number of neutral alleles and their distribution in a finite population. The requirements are that the rate of recombination be equal to or greater than the mutation rate, and the mutation rate be at least the same order of magnitude as the reciprocal of the population size. For example, when $4N\mu = 4$ and $r = 10\mu$, the effective number and the mean number of alleles are approximately twice as large as would be expected if there were no intragenic recombination. In addition, the variance of the number of alleles and the variance of the homozygosity are at least twice as large as would be expected.

By comparing data from natural populations to the results from the Monte Carlo simulation and the analytical theory developed in the preceding sections, it is possible to decide whether or not this process plays an important role in natural populations. Therefore, it is necessary to compare known intragenic recombination rates to mutation rates and also to show that $4N\mu$, for at least some loci, is large enough for intragenic recombination to have an effect.

The rates of intragenic recombination have been estimated in several organisms. Data available for Drosophila should be representative. For the rudimentary locus, CARLSON (1971) has estimated that recombination varies between 7.6×10^{-4} to 5.2×10^{-6} depending on which alleles are being used. At the rosy locus, CHOVNICK, BALLANTYNE and HOLM (1971) found that the recombination rate for null alleles varied between 1.2×10^{-4} and 7.2×10^{-6}. (Since both experiments recovered only wild-type mutants, the recombination rates given here are twice those rates given in the two papers.) Estimated mutation rates vary between 10^{-4} and 10^{-10}; and for visible mutations at representative loci in Drosophila the rate varies between 10^{-5} and 10^{-6} (DOBZHANSKY 1970). Therefore, the rate of intragenic recombination is large enough to be a significant process in determining the number and distribution of alleles, as long as $4N\mu$ is greater than one.

Using either the number of alleles or the homozygosity to estimate the value of $4N\mu$, intragenic recombination should be an important factor in generating variation at loci such as esterase-5 and xanthine dehydrogenase (the rosy locus) in Drosophila, and the *HLA A* and *B* loci in man. In populations of *Drosophila pseudoobscura* (not including Guatemala and Bogotá), the homozygosity and number of alleles of esterase-5 vary from 0.20 to 0.32 and from 4 to 10, respectively (calculated from Table 26; LEWONTIN 1974); and for xanthine dehydrogenase from 0.16 to 0.61 and from 4 to 9, respectively (SINGH, LEWONTIN and FELTON 1976). Among the 146 genomes from all 12 populations, there were 37 alleles of xanthine dehydrogenase detected. In man, for *HLA A* and *B*, the number of alleles are at least 15 and 16, respectively, with corresponding homo-

zygosities of 0.15 and 0.11 for European caucasoids (BODMER, CANN and PIAZZA 1973).

The data on the rate of intragenic recombination and the value of $4N\mu$ imply that intragenic recombination plays a significant role in determining the distribution of neutral alleles in finite populations. In particular, the number and distribution of alleles at the xanthine dehydrogenase structural locus should be influenced by intragenic recombination and therefore any test of neutrality based on EWENS' sampling scheme would be inappropriate for this locus.

We would like to acknowledge an anonymous reviewer for helpful comments.

LITERATURE CITED

BODMER, W. F., H. CANN and A. PIAZZA, 1973 Differential genetic variability among polymorphisms as an indicator of natural selection. pp. 753–767. In: *Histocompatibility Testing 1972*. Edited by J. DAUSSET and J. COLOMBANI. Munksgaard, Copenhagen.

CARLSON, P. S., 1971 A genetic analysis of the rudimentary locus of *Drosophila melanogaster*. Genet. Res. **17**: 53–81.

CHOVNICK, A., G. H. BALLANTYNE and D. G. HOLM, 1971 Studies on gene conversion and its relationship to linked exchange in *Drosophila melanogaster*. Genetics **69**: 179–209.

COYNE, J. A., 1976 Lack of genic similarity between two sibling species of Drosophila as revealed by varied techniques. Genetics **84**: 593–607.

DOBZHANSKY, TH., 1970 *Genetics of the Evolutionary Process*. Columbia University Press, New York.

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

EWENS, W. J. and J. H. GILLESPIE, 1974 Some simulation results for the neutral allele model, with interpretations. Theor. Pop. Biol. **6**: 35–37.

GUTZ, H. and J. F. LESLIE, 1976 Gene conversion: A hitherto overlooked parameter in population genetics. Genetics **83**: 861–866.

HILL, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Pop. Biol. **8**: 117–126.

KARLIN, S. and J. MCGREGOR, 1968 Rates and probabilities of fixation for two locus random mating finite populations without selection. Genetics **58**: 141–159.

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., 1977 The population structure associated with the Ewens sampling formula. Theor. Pop. Biol. **11**: 274–283.

KOEHN, R. K. and W. F. EANES, 1976 An analysis of allelic diversity in natural populations of Drosophila: The correlation of rare alleles with heterozygosity. pp. 377–390. In: *Population Genetics and Ecology*. Edited by S. KARLIN and E. NEVO. Academic Press, New York.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

LEWONTIN, R. C. and J. L. HUBBY, 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics **54**: 595–609.

SERANT, D., 1974 Linkage and inbreeding coefficients in a finite random mating population. Theor. Pop. Biol. **6**: 251–263.

SINGH, R. S., R. C. LEWONTIN and A. A. FELTON, 1976 Genetic heterogeneity within electrophoretic "alleles" of xanthine dehydrogenase in *Drosophila pseudoobscura*. Genetics **84**: 609–629.

STEWART, F. M., 1976 Variability in the amount of heterozygosity maintained by neutral mutations. Theor. Pop. Biol. **9**: 188–201.

WATT, W. B., 1972 Intragenic recombination as a source of population genetic variability. Am. Naturalist **106**: 737–753.

WATTERSON, G. A., 1974 Models for the logarithmic species abundance distributions. Theor. Pop. Biol. **6**: 217–250.

WEIR, B. S. and C. C. COCKERHAM, 1974 Behavior of pairs of loci in finite monoecious populations. Theor. Pop. Biol. **6**: 323–354.

Corresponding editor: B. S. WEIR