THEORETICAL POPULATION BIOLOGY 23, 183–201 (1983)

Properties of a Neutral Allele Model with Intragenic Recombination

RICHARD R. HUDSON

Department of Genetics, University of California, Davis, California 95616

Received July 12, 1982

An infinite-site neutral allele model with crossing-over possible at any of an infinite number of sites is studied. A formula for the variance of the number of segregating sites in a sample of gametes is obtained. An approximate expression for the expected homozygosity is also derived. Simulation results are presented to indicate the accuracy of the approximations. The results concerning the number of segregating sites and the expected homozygosity indicate that a two-locus model and the infinite-site model behave similarly for $4Nu \leqslant 2$ and $r \leqslant 5u$, where N is the population size, u is the neutral mutation rate, and r is the recombination rate. Simulations of a two-locus model and a four-locus model were also carried out to determine the effect of intragenic recombination on the homozygosity test of Watterson (*Genetics* **85**, 789–814; **88**, 405–417) and on the number of unique alleles in a sample. The results indicate that for $4Nu \leqslant 2$ and $r \leqslant 10u$, the effect of recombination is quite small.

1. INTRODUCTION

A variety of neutral models have been extensively studied (for a review see Ewens, 1979.) Many properties of these models, however, are known only under the generally unrealistic assumption that no intragenic recombination occurs. It is known that intragenic recombination can have a significant effect on the sampling properties of a neutral allele model (Strobeck and Morgan, 1978; Morgan and Strobeck, 1979). Clearly, for estimation and hypothesis testing more realistic models need to be considered. Some insight into the effects of intragenic recombination can be gained from the analysis of two-locus models. In this case, each locus of the two-locus model is considered to be a sublocus; the two subloci together constitute a global locus at which intragenic recombination can occur. Strobeck and Morgan (1978) obtained an expression for the expected homozygosity under a two-locus model. Griffiths (1981) found (among other things) the variance of the number of segregating sites in a sample of two genes for a two-locus model. It is unclear, however, how well these two-locus models reflect the behavior of loci that consist of hundreds of nucleotides. Actual genetic loci are

0040-5809/83 $3.00

hundreds of nucleotides long and it may be that neutral mutations can occur at many of the nucleotide sites and that crossing-over can occur between all of them. It is therefore appropriate to consider a model with many sites between which crossing-over can occur.

I report here properties of an infinite-site neutral model with crossing-over possible at any of an infinite number of sites distributed along the locus. An analytic expression for the variance of the number of segregating sites in a sample of two genes is derived. An approximate expression for larger samples is also presented. An approximate formula for the mean homozygosity is derived, which is accurate for large recombination rates and small mutation rates. Monte Carlo simulation results are presented to indicate the adequacy of the approximations.

I also report here the results of a simulation study of a four-locus model and a two-locus model. The study was carried out to examine the effects of recombination on the homozygosity test of Watterson (1977, 1978b) and on the number of unique alleles in a sample. The statistic used in the homozygosity test is the sample homozygosity $\hat{F}$ conditioned on k, the number of alleles in the sample. Under the no-recombination neutral model the distribution of $\hat{F} \mid k$ is independent of unknown parameters (Ewens, 1972). Strobeck and Morgan (1978) showed, with simulations of a two-locus model, that the sampling theory of neutral alleles with recombination is not the same as the sampling theory that assumes no recombination. Their results give no direct indication of what the effects of recombination are on the distribution of $\hat{F} \mid k$. Morgan and Strobeck (1979) used Monte Carlo simulations of the two-locus model to assess the effects of recombination on $\hat{F} \mid k$. Based on one particular set of parameter values, they concluded that intragenic recombination increases the conditional population homozygosity and increases the number of unique alleles conditional on k. The validity of their generalization is open to question. I have used simulations of a four-locus model to estimate the mean, variance, and critical values of $\hat{F} \mid k$ for several combinations of recombination rates and mutation rates. It is found that intragenic recombination sometimes increases and sometimes decreases the mean conditional homozygosity depending on the particular parameter values used. Similarly, simulations of a two-locus model indicate that recombination sometimes increases the mean number of unique alleles and sometimes decreases the number of unique alleles depending on the parameter values and k. The results presented in Sections 2 and 3 suggest that the four-locus model behaves essentially as the infinite-site model over an important range of recombination and mutation rates which will be specified later.

The properties of the infinite-site model are obtained by first considering an m-locus model analogous to the two-locus model studied by Strobeck and Morgan (1978) and Griffiths (1981). The m linearly arranged loci are

considered to be "subloci" that together constitute the global locus whose properties we wish to study. Each sublocus follows the infinite site model of Kimura (1969). Recombination does not occur within subloci, but occurs between adjacent subloci with rate $r/(m-1)$ per generation per gamete. Random union of gametes (Karlin and MacGregor, 1968) is assumed. The population is assumed to remain at a constant size of $2N$ gametes. r is assumed small so that the occurrence of more than one cross-over per gamete per replication can be ignored. With this assumption the recombination rate between the most distant subloci is r. The number of mutations per replication per sublocus is assumed Poisson distributed with mean u/m. Mutations are assumed to occur at sites not currently segregating in the population, and so produce unique alleles at the sublocus and global locus levels. Gametic types are equivalent to alleles at the global locus. Although this m-locus model is considered primarily as a means to obtain results concerning an infinite-site model, it may be useful by itself. Since the coding sequences of many gene products of eukaryotes are interrupted by introns (Gilbert, 1978), the m-locus model may be useful when considering variation in products coded by m exons and $m-1$ introns.

The infinite-site model is obtained by letting m tend to infinity with u and r held constant. Although in this model each of the infinite number of subloci is assumed to be an infinite-site locus itself, the model will clearly provide a good approximation to a neutral locus made up of a large number of nucleotides provided the the product of the population size and the mutation rate per nucleotide site per replication is small. For some theoretical results concerning this point see Golding and Strobeck (1982).

It should be noted that the random union of gametes model considered here is not equivalent (except for $r=0$) to the perhaps more biologically realistic random union of zygotes model (Kimura, 1963). The two models behave similarly in many respects. For a review see Ewens (1979, Section 3.8). The results presented in this paper are expected to apply approximately to the random union of zygotes model, but no evidence to support this expectation is presented in this paper.

The simulation algorithm used in this study is of special interest, as it can be used for studying sampling properties of the neutral model other than those properties reported here. For example, properties concerning the number of segregating sites and the genealogical relationships of sampled genes could be studied. The algorithm does not require computer representations of entire populations or the time-consuming process of sampling to produce successive generations of gamete populations. Section 5 contains a description of the algorithm.

2. Segregating Sites

Consider the m-locus model described in the Introduction. Let s_i be the number of segregating sites at the ith sublocus in a random sample of two genes. From Watterson (1975), $\mathrm{Var}(s_i) = \theta/m + \theta^2/m^2$, where $\theta = 4Nu$. Griffiths (1981) obtained the following formula for the covariance of the number of segregating sites at two linked loci:

$$\mathrm{Cov}(s_i, s_j) = (\theta/m)^2 f(C), \tag{1}$$

where $f(C) = (C + 18)/(C^2 + 13C + 18)$ and C equals $4N$ times the recombination rate between the two loci. Thus the variance of $S = \sum_{i=1}^{m} s_i$ is

$$\begin{aligned}\mathrm{Var}(S) &= \sum_{i=1}^{m} \mathrm{Var}(s_i) + \sum_{i \neq j} \mathrm{Cov}(s_i, s_j) \\ &= \theta + \theta^2/m + \frac{2\theta^2}{m} \sum_{j=1}^{m-1} (m-j) f\left(\frac{jR}{m-1}\right),\end{aligned} \tag{2}$$

where $R = 4Nr$. The infinite-site result, obtained by letting m tend to infinity in (2), is

$$\mathrm{Var}(S) = \theta + \theta^2 V(R), \tag{3}$$

where

$$V(R) = \frac{2}{R^2} \int_0^R f(z)(R - z)\, dz. \tag{4}$$

The integral can be evaluated explicitly to obtain a rather bulky expression. $V(R)$ is a monotonically decreasing function with $\lim_{R \to 0} V(R) = 1$, $V(5.89) \approx 0.5$, and $V(\infty) = 0$. Under the infinite site model with $\theta = 1$, the variance of S is just $1 + V(R)$, which is plotted in Fig. 1. Also plotted is the variance of S for the two-locus model from Griffiths (1981). For $R < 5$, the variance of S is approximately the same for the two-locus model and the infinite-site model.

It is of interest to know the variance of the number of segregating sites for larger sample sizes. Consider first the two-locus model. For the case $R = 0$, Watterson (1975) gives the variance of S as $\theta \sum_{j=1}^{n-1} 1/j + \theta^2 \sum_{j=1}^{n-1} 1/j^2$, where n is the sample size, and similarly the variance of s_i is $(\theta/2) \sum_{j=1}^{n-1} 1/j + (\theta/2)^2 \sum_{j=1}^{n-1} 1/j^2$. Clearly, since the variance of S equals twice the variance of s_1 plus twice the covariance of s_1 and s_2, the covariance of s_1 and s_2 when $R = 0$ is $(\theta/2)^2 \sum_{j=1}^{n-1} 1/j^2$. If the covariance decreases with increasing R in approximately the same way for larger

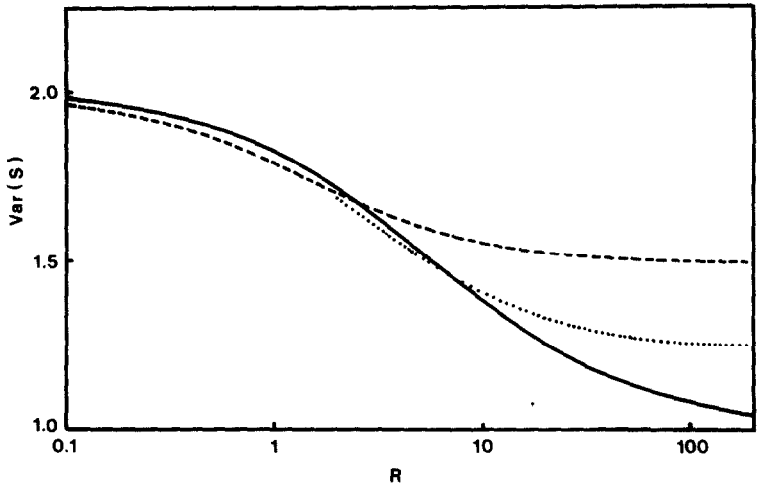

FIG. 1. The variance of the number of segregating sites in a sample of two gametes as a function of $R = 4Nr$ under the two-locus model (– – –), under the four-locus model (...), and under the infinite-site model (—) ($\theta = 1$).

samples as it does for sample size two (as in (1)), then the covariance would be given approximately as

$$\mathrm{Cov}(s_1, s_2) \simeq (\theta/2)^2 \left(\sum_{j=1}^{n-1} 1/j^2 \right) f(R). \tag{5}$$

It will be shown in Section 5.1 that $\mathrm{Cov}(s_1, s_2)$ is indeed proportional to θ^2. An interpretation of the constant of proportionality and a means of estimating it with Monte Carlo simulations are also described there. Estimates of the constant of proportionality obtained from simulations are shown in Fig. 2 for $n = 100$ and several values of R. Also plotted as a function of R is the putative constant of proportionality in (5): $(\sum_{j=1}^{n-1} 1/j^2)$ $f(R)/4$. Simulations with smaller values of n were also carried out with similar results. For $n \leqslant 100$ at least, it appears that the approximation (5) is quite good.

Using the same approach as was used to obtain (2) but for sample sizes greater than 2, and using the approximation (5) for the covariances, it follows for the m-locus model that

$$\mathrm{Var}(S) \simeq \theta \left(\sum_{j=1}^{n-1} 1/j \right) + \frac{\theta^2}{m} \sum_{j=1}^{n-1} 1/j^2 + \frac{2\theta^2}{m^2} \left(\sum_{j=1}^{n-1} 1/j^2 \right) \sum_{j=1}^{m-1} f\left(\frac{jR}{m-1} \right) (m-j), \tag{6}$$

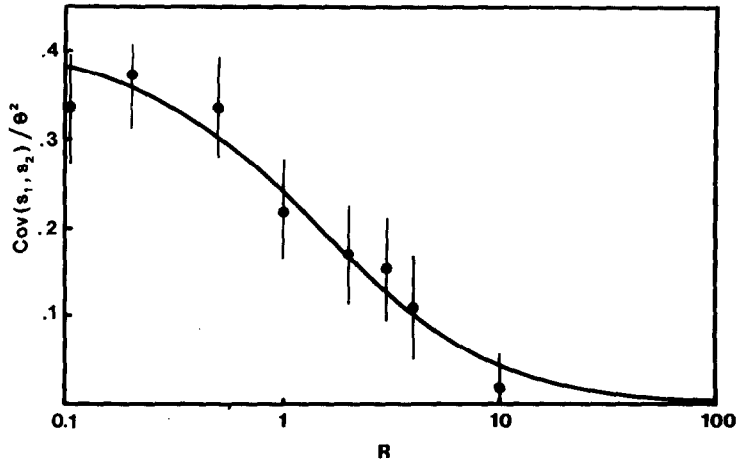

FIG. 2. The covariance of the number of segregating sites at two loci in a sample of 100 gametes as a function of $R = 4Nr$. The dots are estimates of the covariance obtained by simulation as described in Section 5.1. Ninety-five percent confidence intervals are indicated. The curve is the putative approximate relationship given by (5).

and for the infinite-site model

$$\mathrm{Var}(S) \simeq \theta \left(\sum_{j=1}^{n-1} 1/j \right) + \theta^2 \left(\sum_{j=1}^{n-1} 1/j^2 \right) V(R). \tag{7}$$

3. HOMOZYGOSITY

An approximate expression for the homozygosity under the infinite-site model is obtained in this section. To illustrate the approach, I start by considering a single locus (no-recombination) model, with neutral mutation rate u. Two genes sampled from the population have a most recent common ancestor (MRCA) at some time, t generations back. Under the infinite alleles model (and other Wright–Fisher-type neutral models), t is geometrically distributed with mean $2N$. Consequently $t/2N$ is approximately exponentially distributed with mean 1. Conditional on t, the number of mutations that occurred on the line of descent from the MRCA to the sampled genes is Poisson distributed with mean $2ut$. Therefore, conditional on t, the probability of identity of the two genes is $\exp[-\theta(t/2N)]$. So the unconditional probability of identity of the two genes is the expectation of $\exp[-\theta(t/2N)]$, which is just the moment generating function of the exponential random variable $t/2N$. The moment generating function of an exponential random variable with mean one is just $1/(1+\theta)$. This is the

familiar result of Kimura and Crow (1964) concerning the expected homozygosity.

Now consider the m-locus model and a sample of two gametes from the population. Denote by t_i the time of the MRCA of the sampled genes at the ith sublocus. Because of linkage, the t_i's are not independent. If at each sublocus the mutation rate is u/m, then conditional on all the t_i, $i = 1,..., m$, the probability of identity of the two gametes at all subloci is $\exp(-\theta T)$, where $T = \sum_i (t_i/2Nm)$. The expected homozygosity (global locus) is just the unconditional expectation of $\exp(-\theta T)$. The expection of $\exp(-\theta T)$ can be expanded about $T = 1$ as

$$E[\exp(-\theta T)] = e^{-\theta} - \theta e^{-\theta} E(T-1) + (\theta^2/2)\, e^{-\theta} E(T-1)^2 + H, \quad (8)$$

where H is the sum of the higher-order terms. The expectation of T is 1. The variance of T depends on the recombination rate between the subloci. Note that if the subloci are unlinked so that the t_i's are independent, the variance of T is $1/m$, which tends to zero as m tends to infinity. Since T is identically 1 for an infinite number of unlinked subloci with total mutation rate u, the expected (global-locus) homozygosity is $\exp(-\theta)$. This is not a new result (e.g., see Ewens, 1979, p. 239). For linked loci, the variance of T can be simply obtained as follows. Since S, the number of segregating sites in the sample of two gametes, when conditioned on T is Poisson distributed, there is a simple relationship between the variance of S and the variance of T:

$$\begin{aligned} \mathrm{Var}(S) &= E[\mathrm{Var}(S \mid T)] + \mathrm{Var}[E(S \mid T)] \\ &= E(\theta T) + \mathrm{Var}(\theta T) \\ &= \theta + \theta^2 \mathrm{Var}(T). \end{aligned} \quad (9)$$

From (2) and (9) we have

$$\mathrm{Var}(T) = 1/m + \frac{2}{m^2} \sum_{j=1}^{m-1} f\left(\frac{jR}{m-1}\right)(m-j) \quad (10)$$

for the m-locus model, and

$$\mathrm{Var}(T) = V(R) \quad (11)$$

for the infinite-site model ($V(R)$ is defined by (4)).

As noted above, under the infinite-site model, the variance of T (and higher moments) tend to zero as R tends to infinity. Consequently, for R large we can expect H to be small, especially for smaller values of θ. So for

large R and small θ the following approximation is expected to be accurate for the infinite-site model:

$$E(F) = E[\exp(-\theta T)] \simeq e^{-\theta}[1 + (\theta^2/2)\, V(R)]. \tag{12}$$

In Fig. 3, I have plotted (12) for θ equal to 0.5, 0.75, 1.0, and 2.0. Also

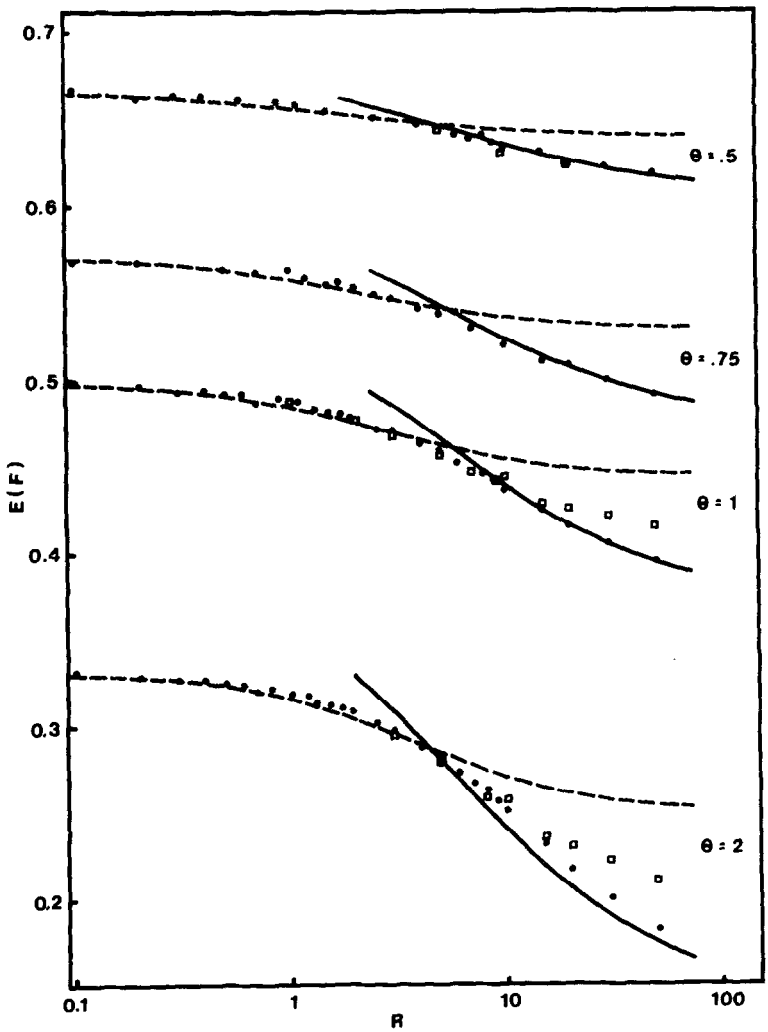

FIG. 3. Expected homozygosity as a function of $R = 4Nr$. – – –, Two-locus theory (Strobeck and Morgan, 1978); —, plots of (12), an approximate relationship for the infinite-site model; estimates from simulations: ●, infinite-site model; □, four-locus model.

plotted is $E(F)$ for the two-locus model (Eq. (5) of Strobeck and Morgan, 1978). Also shown are estimates of $E(F)$ for the infinite-site model and four-locus model obtained from simulations. The simulation algorithm is described in Section 5. It can be seen that for $R \lesssim 5$, the two-locus model provides a good approximation to the infinite-site model. For $R \geqslant 5$ and $\theta \leqslant 2$, approximation (12) is quite accurate. Also note that for $R \leqslant 30$ and $\theta \leqslant 2$, the four-locus model has approximately the same expected homozygosity as the infinite-site model.

4. The Homozygosity Test and the Number of Unique Alleles

The results of Strobeck and Morgan (1978) and Morgan and Strobeck (1979) raise doubts about the use of the homozygosity test, especially for highly variable loci. Morgan and Strobeck (1979) showed, under the two-locus model with $2N = 200$, $\theta = 4$, and $R = 40$, that the mean population homozygosity and the number of unique alleles (conditional on 25 alleles being present in the population) is higher than under the no-recombination model. To investigate the generality of this effect and its significance for applications of the homozygosity test, simulations of a two-locus and a four-locus model were carried out.

The effect of recombination on the expected number of unique alleles was studied with the two-locus model. The results for $\theta = 4$, $R = 40$, and sample size of 200 are shown in Fig. 4. For $k < 25$, the expected number of unique alleles conditional on k is increased due to recombination. For $k > 25$, the expected number of unique alleles is decreased due to recombination. Morgan and Strobeck's result ($k = 25$) is also shown on the figure. They found significantly more unique alleles than I did. Perhaps this discrepancy reflects the difference between population statistics considered by Morgan and Strobeck and the sample statistics reported here.

Additional simulations showed that higher rates of recombination result in an increased range of k values for which the mean number of unique alleles was increased; with lower rates of recombination there is an increased range of k values for which the mean number of unique alleles is decreased. It is difficult to generalize from the small number of simulations that have been carried out. It is, however, clear that the effects of recombination are complex and that the mean number of unique alleles conditional on k can be increased or decreased by recombination depending on the values of θ, R, and k.

The distribution of $\hat{F} \mid k$ was studied with the four-locus model. The four-locus model is expected to reflect more closely the behavior of a realistic many-sites model. Estimates of the mean, variance, and critical values of $\hat{F} \mid k$ obtained from simulations using the four-locus model are shown in

TABLE I

Estimates of the Mean, Variance, and Critical Values of $\hat{F} \mid k$ for Samples of 100 under the Four-Locus Model with Various Values of θ and R^a

k	θ	R	$E(\hat{F} \mid k)$	Var	$\hat{C}_{2.5}$	(CI)	$\hat{C}_{5.}$	(CI)	$\hat{C}_{97.5}$	(CI)	N
3	—	0:	0.671	0.033	0.369[b]	—	0.398[b]	—	NS	—	—
	0.5	5:	0.683	0.031	0.386	(0.371–0.398)	0.430	(0.414–0.435)	NS	—	1897
	1.0	10:	0.696	0.031	0.381	(0.366–0.399)	0.415	(0.398–0.434)	NS	—	877
	2.0	20:	0.695	0.027	—	—	—	—	—	—	49
5	—	0:	0.490	0.025	0.263[b]	—	0.283[b]	—	0.849[b]	—	—
	0.5	5:	0.488	0.023	0.264	(0.253–0.275)	0.287	(0.275–0.295)	0.831	(0.813–0.848)	997
	1.0	10:	0.518	0.027	0.279	(0.271–0.286)	0.295	(0.290–0.303)	0.866	(0.848–0.885)	1452
	2.0	20:	0.551	0.028	0.300	(0.261–0.314)	0.316	(0.300–0.344)	0.866	(0.848–0.866)	272
7	—	0:	0.376	0.017	0.20	—	0.21	—	0.71[b]	—	1000
	0.5	5:	0.359	0.016	0.203	(0.176–0.212)	0.214	(0.201–0.218)	0.679	(0.625–0.830)	267
	1.0	10:	0.382	0.017	0.211	(0.203–0.216)	0.223	(0.217–0.227)	0.698	(0.679–0.727)	1225
	2.0	20:	0.420	0.018	0.225	(0.209–0.232)	0.246	(0.229–0.252)	0.728	(0.712–0.777)	707

10	—	0:	0.271	0.009	0.15	—	0.16	—	0.48	—	1000
	1.0	10:	0.259	0.0074	0.153	(0.143–0.156)	0.161	(0.154–0.167)	0.504	(0.460–0.523)	528
	2.0	20:	0.294	0.011	0.159	(0.154–0.162)	0.171	(0.163–0.176)	0.573	(0.535–0.604)	1170
	4.0	40:	0.328	0.0098	—	—	—	—	—	—	42
15	—	0:	0.176	0.0033	0.11	—	0.11	—	0.33	—	1000
	2.0	20:	0.174	0.0029	0.107	(0.105–0.109)	0.113	(0.110–0.115)	0.307	(0.288–0.333)	979
	4.0	40:	0.206	0.0051	0.120	(0.111–0.124)	0.124	(0.117–0.136)	0.419	(0.373–0.460)	182
20	—	0:	0.125	0.0013	0.08	—	0.08	—	0.22	—	1000
	2.0	20:	0.120	0.0010	0.080	(0.076–0.082)	0.082	(0.081–0.084)	0.200	(0.185–0.225)	467
	4.0	40:	0.137	0.0017	0.079	(0.077–0.083)	0.087	(0.079–0.092)	0.240	(0.222–0.269)	397
25	—	0:	0.094	0.0006	0.06	—	0.07	—	0.15	—	1000
	2.0	20:	0.081	0.0003	—	—	0.062	(0.060–0.065)	—	—	131
	4.0	40:	0.097	0.0007	0.064	(0.061–0.066)	0.067	(0.064–0.068)	0.164	(0.155–0.183)	629

[a] The expectation and variance of $\hat{F} \mid k$ for $R = 0$ are from Ewens (1979, Appendix D). The critical values for $R = 0$, except for those marked *b*, are from Ewens (1979, Appendix C), and were based on 1000 independent samples drawn from the theoretical distribution for the no-recombination model. $\hat{C}_x$ is an estimate of C_x, the critical value of the distribution of $\hat{F} \mid k$ such that the probability that $\hat{F} \mid k$ is less than C_x is $x/100$. CI is the 95% confidence interval of the preceding estimate of C_x. N is the number of samples which were used to obtain the estimates.

[b] Exact value obtained from the theoretical distribution of Ewens (1972).

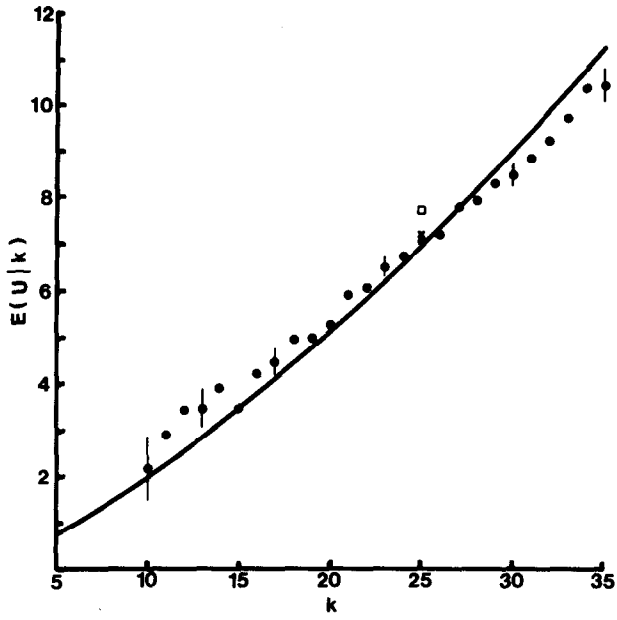

FIG. 4. Expected number of unique alleles (U) given k, the number of alleles in the sample, for the two-locus model and $\theta = 4$. —, With no recombination ($R = 0$) exact results are available; ●, estimates from simulations with $R = 40$ and sample size $n = 200$; vertical bar, ± 2 standard errors; results of Morgan and Strobeck's simulation for $R = 40$ (□) and $R = 0$ (×), considering an entire population ($2N = 200$).

Table I. The results are for samples of 100 gametes and various combinations of θ and R. The simulations have been carried out for a very limited set of parameter values, but generally the behavior of $\hat{F} \mid k$ appears to parallel the behavior of the number of unique alleles conditional on k. It is clear from the table that $\hat{F} \mid k$ is dependent on θ and R, but for $\theta < 4$ and $R < 10\theta$ it appears that $\hat{F} \mid k$ is only weakly affected by recombination. The greatest effects seem to occur for values of k well above or well below the expected number of alleles for a given combination of θ and R. For k well below the expected number, the mean of $\hat{F} \mid k$ as well as the 2.5, 5, and 97.5% critical values are raised somewhat by higher recombination rates. When k is well above its expected value, the mean of $\hat{F} \mid k$ is reduced from its no-recombination value; the variance of $\hat{F} \mid k$ is also reduced. The net result seems to be that, for k well above its expected value, the 2.5 and 5% critical values are not detectably changed from the no-recombination values; the 97.5% critical value appears to be somewhat reduced.

It can be seen from the results of Sections 2 and 3 that the expected homozygosity and the variance of the number of segregating sites are nearly

the same under the four-locus model as under the infinite-site model, when $\theta \leqslant 2$ and $R \leqslant 20$. This suggests that the conclusions concerning the distribution of $\hat{F} \mid k$ for the four-locus may apply to the infinite-site model as well, when $\theta \leqslant 2$ and $R \leqslant 20$. If so, the following conclusions are justified when $\theta \leqslant 2$ and $R \leqslant 10\theta$. If one is testing the neutral model with the alternative hypothesis being a model which produces allelic frequencies more even than under the neutral model (perhaps a model with heterosis), the homozygosity test will be little affected by recombination; the small effect that is produced will make the test slightly conservative. When the alternative hypothesis is a model which produces less even allelic frequencies than the neutral model (e.g., a deleterious alleles model), one uses either the 95 or the 97.5% critical values, which may be higher or lower than the no-recombination values depending on k, θ, and R. However, in general, the effects of recombination are not large and will not explain the appearance of a great excess of rare alleles.

For larger values of θ and R it is difficult to extrapolate from the results presented so far, but individual samples can be evaluated as follows. Consider the sample of 21 chromosomes of *Drosophila persimilis* examined by Coyne (1976). In his sample he found 10 alleles at the *X dh* locus, nine of which were present only once in the sample. The no-recombination neutral model can be rejected with this sample using the homozygosity test (Watterson, 1978a). To determine if the sample homozygosity observed could be accounted for by a neutral model with recombination, a simulation was carried out with an eight-locus model with θ equal to 5 and R equal to 50. Ten thousand samples of size 21 were generated. The mean number of alleles per sample was 14.7. Only 6% of the samples contained 10 or fewer alleles. Recall that it is when the observed k is well below the expected k that recombination tends to produce an "excess" of rare alleles as seen in the sample of Coyne. None of the 289 samples which contained 10 alleles had the extreme F seen in the Coyne sample. For higher values of θ and R, one is very unlikely to see 10 or fewer alleles. It can be concluded that for $R < 10\theta$, and assuming stationarity, the neutral model even with recombination cannot account for the Coyne sample.

5. The Computer Algorithm

5.1. *Two-Locus Model*

The simulations of Strobeck and Morgan were carried out by representing entire populations in the computer and repeatedly producing one entire generation from the preceding one to get populations in statistical

equilibrium. Their method is very time-consuming and restricted them to consideration of relatively small population sizes. I describe here an efficient method of producing independent samples of genes or gametes from populations evolving according to the Wright–Fisher neutral model. The method does not require representations of entire populations or the repeated sampling of gametes to produce generation after generation of gamete pools. The core of the method is the generation of the history or pedigree of a sample of gametes. The history of the sampled gametes consists of a collection of "family trees," one tree for each locus. The tree for a locus specifies which sampled genes are most closely related and also when the most recent common ancestors (MRCAs) of the sampled genes occurred. The generation of the tree for a single locus, with no intragenic recombination, is simple and is described by Hudson (1983). For linked loci, the topologies and the lengths of the branches of the trees for the different loci are correlated. The generation of these correlated trees requires a more complex algorithm that is described in the following paragraphs. Once the trees are generated the number of mutations that occur on each branch of each of the trees is easily generated because the number of mutations is Poisson distributed when conditioned on the trees. After the mutations are generated, the number of segregating sites, number of alleles, and $\hat{F}$ can be tabulated.

Before specifying exactly how the trees are generated some definitions are necessary. I consider the same two-locus model studied by Strobeck and Morgan. "Generation t" signifies the population t generations before the present generation. We consider a sample of n gametes from the current population (generation 0). Those gametes of generation t that contain genetic material at either locus which is directly ancestral to genetic material of the sampled gametes will be referred to as ancestral gametes of generation t. Let $g(t)$ denote the number of ancestral gametes of generation t. $g(0)$ is n, the number of gametes in the sample. *If any two ancestral gametes of generation t have a common ancestor in generation $t+1$, we say that the event* CA *has occurred in generation $t+1$.* An ancestral gamete may contain directly ancestral genetic material at one or two loci. This directly ancestral genetic material at a locus will be referred to as an ancestral gene. If an ancestral gamete of generation t contains two ancestral genes, with probability r (the recombination rate), the gamete is the recombinant descendent of two ancestral gametes of generation $t+1$. *When any one of the ancestral gametes of generation t is the recombinant descendent of two ancestral gametes of generation $t+1$, we say that the event* RE *has occurred in generation $t+1$.* Denote by $d(t)$ the number of ancestral gametes of generation t that contain ancestor genes at both loci. In Fig. 5 an example history of a sample of two gametes is shown.

The algorithm starts at the present and proceeds back in time generating

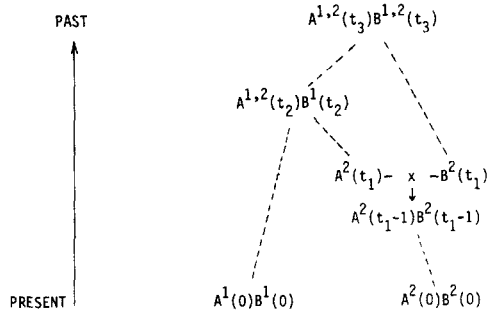

FIG. 5. A possible history of two gametes sampled from a population, illustrating RE and CA events as defined in the text. $A^i(t)$ denotes the direct ancestor in generation t of the A-locus gene of gamete i. In this history there is one RE event. It occurs at generation t_1 in which the gamete $A^2(t_1-1)B^2(t_1-1)$ is the recombinant descendent of the two gametes, $A^2(t_1)-$ and $-B^2(t_1)$. There are two CA events, one at generation t_2 and one at t_3. The function $g(t)$, defined in the text, takes on the value 2 for $0 < t < t_1$ and $t_2 < t < t_3$, and the value 3 for $t_1 < t < t_2$. The function $d(t)$, also defined in the text, is 2 for $0 < t < t_1$ and 1 for $t_1 < t$.

the times of successive RE and CA events. Given $g(t)$ and $d(t)$, the probability that neither CA nor RE occur in generation $t+1$ is

$$(1-r)^{d(t)} \prod_{i=1}^{g(t)-1} \left(1-\frac{i}{2N}\right) \simeq 1 - rd(t) - \frac{g(t)[g(t)-1]}{4N}, \qquad (13)$$

where the approximation requires that $g(t) \ll 2N$ and that r is small. Note that as long as no CA or RE occurs, g and d remain constant. This implies that the intervals between events of either type is geometrically distributed. If, for example, an event occurs at generation t, the time back until another event occurs is geometrically distributed with approximate mean of $4N/[4Nrd(t) + g(t)\{g(t)-1\}]$. If time is measured in units of generations divided by $4N$, the time interval is approximately exponentially distributed with mean $1/[4Nrd(t) + g(t)\{g(t)-1\}]$; these exponential random variables are easily generated on the computer.

Given that an event has occurred at a particular time, it is necessary to determine which type of event has occurred. Assuming that the probability of both events occurring in the same generation is negligible, the probability of RE given that CA or RE occurred is approximately $4Nrd(t)/[4Nrd(t) + g(t)\{g(t)-1\}]$; the probability of CA is approximately one minus this quantity.

As a result of each event, g and d change. Also, with each CA event it must be determined which, if any, of the sample genes have MRCAs as a result of the CA event. Recall that it is the occurrences of the MRCAs which are of primary interest. To determine who has an MRCA and how to change

$d(t)$, a list of ancestral gametes is maintained in the computer. Each element of the list is a representation of an ancestral gamete, indicating at which loci the gamete contains ancestor genes and also which of the sample gametes contain the genes descended from the ancestor genes. This list is updated as each event occurs to keep track of the composition of the ancestral gametes as each new time interval is considered.

When a CA event occurs, for example in generation t, two ancestral gametes are randomly chosen from the list and replaced by a single common ancestor gamete. Thus $g(t)$ equals $g(t-1)-1$. The value of $d(t)$ may or may not differ from $d(t-1)$ as a result of the CA event. This depends on the location of the ancestor genes on the ancestral gametes of generation $t-1$. As CA events are determined to have occurred, the occurrence of MRCAs of sampled genes are noted and the trees for each locus are thus constructed.

When an RE event occurs, one of the ancestral gametes with ancestor genes at both loci is randomly chosen. The chosen ancestral gamete is replaced by two gametes, one with one of the ancestor genes and the other with the other ancestor gene. Since the possibility of both an RE and CA event in the same generation is assumed small enough to be ignored, the portions of the parent gametes not contained in the recombinant descendent gamete are assumed not to be directly ancestral to sample genes. Thus $d(t)$ is one less that $d(t-1)$ as a result of the RE event. Also, $g(t)$ is $g(t-1)+1$ as a result of the RE event.

In summary, the trees are constructed by generating the times and order of occurrence of CA and RE events. The time intervals between events depend on the number of ancestral gametes and on how the ancestral genes are linked on the ancestral gametes. These quantities change through time as RE and CA events occur. When CA events occur, MRCAs of sample genes may occur; it is the time of occurrence of the MRCAs and the identity of the genes involved that constitute the information needed to construct the trees for each locus. As mentioned earlier, after the trees are generated, the mutations can easily be generated and F and k tabulated. The algorithm is easily extended to more loci.

To answer some questions, only the trees without the mutations need to be generated. For example, consider the problem of estimating the covariance of the number of segregating sites at two loci in a sample of n gametes. Let s_i, $i=1, 2$, denote the number of segregating sites at the ith locus. Consider the tree representing the history of the ith locus. Let t_i denote the total time (in units of $4N$) in this tree since the most ancient node, i.e., since the MRCA of the entire sample at the ith locus. Since the expectation of s_i, conditional on t_i, is $(\theta/2)\,t_i$, it is clear that

$$\operatorname{Cov}(s_1, s_2) = (\theta^2/4)\operatorname{Cov}(t_1, t_2). \tag{14}$$

To estimate the covariance of s_1 and s_2, one need only generate sample trees

and obtain an estimate of the covariance of t_1 and t_2; mutations need not be generated. It was this method that was used to obtain the estimates shown in Fig. 2.

A listing of a Fortran subroutine to generate random samples under this multilocus neutral model is available from the author.

5.2. *Infinite-Site Model*

To obtain estimates of the expected homozygosity under the infinite-sites model, an algorithm similar to that described above can be used. Only samples of size two need be considered. Each of the sampled gametes is represented by the interval $[0, 1]$, which represents the length of chromosome that constitutes the locus being studied. As before, it is assumed that with probability r, a gamete of generation t is the recombinant descendent of two gametes of generation $t+1$. Given that a gamete is a recombinant descendent of two gametes of the previous generation, it is assumed that the cross-over which produced the recombinant is equally likely to have occurred at any point in the interval $[0, 1]$ that represents the locus. The mutation rate per replication of subintervals is assumed equal to u times the length of the subinterval. Just as in the two-locus algorithm, the algorithm generates a random sequence of CA and RE events which occur at time intervals which are exponentially distributed with mean that depends on the number of ancestral gametes and the recombination rate. Most recent events are considered first. As before, a list of ancestral gametes is maintained, but in this case, each ancestral gamete is represented by a collection of subintervals (subsets of $[0, 1]$) which are the pieces of the ancestral gamete which are directly ancestral to a corresponding piece of one or more sample gametes.

When an RE event occurs, for example in generation t, a randomly chosen ancestral gamete of generation $t-1$ is cut in two pieces at a randomly chosen cross-over point. Each of the two pieces is made part of a separate parental gamete of generation t. As before, the ends of the two parental gametes which do not appear in the descendent recombinant gamete are assumed not to be ancestral to sampled gametes.

When a CA event occurs, say in generation t, two ancestral gametes of generation $t-1$ are randomly picked and merged to form a common ancestral gamete of generation t. The representation of the merged gamete is actually just the union of the intervals representing the descendant gametes. The intersection of the intervals representing the two descendent gametes are subintervals of the sampled gametes which have an MRCA at generation t. Given that the subintervals, say $[a, b]$, of the two sampled gametes have an MRCA at generation t, the probability that the two sampled gametes are identical (i.e., no mutations have occurred in these subintervals in the descent from their MRCA) is just $\exp[-\theta(b-a)t/2N]$. The generation of CA and

RE events continues until MRCAs have been found for all the genetic material of the locus. Given the history of the sampled gametes, the probability that the two gametes are identical can be calculated as the product of exponentials, each of the form $\exp[-\theta l_i t_i/2N]$, where l_i is the length of the subintervals of the sampled gametes that have an MRCA at generation t_i. The estimates of the mean homozygosity shown in Fig. 3 were obtained by generating many such histories and calculating the mean probability that the two sampled gametes are identical.

6. Discussion and Conclusions

Actual genetic loci consist of many nucleotide sites. It may be that neutral mutations can occur at many of these sites, and that crossing-over can occur between many or all of them. Consequently, the most appropriate neutral model to consider may be a many-site model with recombination possible between all the sites. For this reason, the properties of the infinite-site model presented in the previous sections are of special interest. In general, many-site models with recombination are very difficult to analyze and few properties are known except in the extreme cases of free recombination or complete linkage between sites. The analyses and the simulations of this study suggest that a two-locus neutral model is a good approximation to an infinite-site neutral model for $\theta \leqslant 2$ and $4Nr \leqslant 5$. This means that the analyses of two-locus models, such as those of Strobeck and Morgan (1978) and Griffiths (1981) give us quantitative as well as qualitative information concerning the effects of intragenic recombination in many-sites models.

My simulation studies of the four-locus neutral model indicate that critical values of the statistic used in the homozygosity test are only weakly affected by recombination, at least for $\theta \leqslant 2$ and $R \leqslant 10\theta$. Depending on θ and k, the homozygosity conditional on k may be increased or decreased as a result of intragenic recombination. When testing the neutral hypothesis against the alternative hypothesis of heterosis, maintaining an even distribution of allele frequencies, the test becomes slightly conservative in the presence of recombination if one uses the no-recombination critical values. When the alternative hypothesis is a deleterious alleles model, for which many rare alleles are maintained, the situation is less clear cut. The appropriate critical values may be greater than or less than the no-recombination critical values depending on θ, R, and the number of alleles in the sample. In any event the effects of recombination are not great and will not explain the appearance of large numbers of rare alleles, at least when $\theta \leqslant 2$ and $R \leqslant 10\theta$.

ACKNOWLEDGMENTS

I thank John Gillespie and the Division of Environmental Studies Computational Facility for making the computer time available for this study. I am grateful to Michael Turelli for his helpful comments on an earlier version of the paper. Support for the author was provided by a Public Health Service training grant administered by the University of Pennsylvania.

REFERENCES

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

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

EWENS, W. J. 1979. "Mathematical Population Genetics," Springer-Verlag, Berlin/New York.

GILBERT, W. 1978. Why genes in pieces? *Nature (London)* **271**, 501.

GOLDING, G. B., AND STROBECK, C. 1982. The distribution of nucleotide site differences between two finite sequences, *Theor. Pop. Biol.* **22**, 96–107.

GRIFFITHS, R. C. 1981. Neutral two-locus multiple allele models with recombination, *Theor. Pop. Biol.* **19**, 169–186.

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

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

KIMURA, M. 1963. A probability method for treating inbreeding systems especially with linked genes, *Biometrics* **19**, 1–17.

KIMURA, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations, *Genetics* **61**, 893–903.

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

MORGAN, K., AND STROBECK, C. 1979. Is intragenic recombination a factor in the maintenance of genetic variation in natural populations? *Nature (London)* **277**, 383–384.

STROBECK, C., AND MORGAN, K. 1978. The effect of intragenic recombination on the number of alleles in a finite population, *Genetics* **88**, 829–844.

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

WATTERSON, G. A. 1977. Heterosis or neutrality? *Genetics* **85**, 789–814.

WATTERSON, G. A. 1978a. An analysis of multi-allelic data, *Genetics* **88**, 171–179.

WATTERSON, G. A. 1978b. The homozygosity test of neutrality, *Genetics* **88**, 405–417.