Genet. Res., Camb. (1979), **34**, *pp.* 253–267
With 3 text-figures
Printed in Great Britain

Some sampling properties of selectively neutral alleles

Effects of variability of mutation rates

BY RANAJIT CHAKRABORTY AND PAUL A. FUERST

Center for Demographic and Population Genetics,
University of Texas Health Science Center, Houston, Texas

(*Received* 6 *July* 1979)

SUMMARY

Some sampling properties related with the mean and variance of the number of alleles and single locus heterozygosity are derived to study the effect of variations in mutation rate of selectively neutral alleles. The correlation between single locus heterozygosity and the number of alleles is also derived. Monte Carlo simulation is conducted to examine the effect of stepwise mutations. The relevance of these results in estimating the population parameter, $4N_e v$, is discussed in connexion with neutralist–selectionist controversy over the maintenance of genetic variability in natural populations.

1. INTRODUCTION

In the classical infinite allele model of population genetics (Wright, 1949; Kimura & Crow, 1964) an implicit assumption is made that the mutation rates of all loci are equal. Clearly this assumption is made largely for mathematical convenience. Recent studies suggest that an enormous amount of variability in per locus mutation rates exists for the spectrum of proteins and enzymes usually studied by electrophoresis (Nei, Chakraborty & Fuerst, 1976; Nei, Fuerst & Chakraborty, 1978; Chakraborty, Fuerst & Nei, 1978; Koehn & Eanes, 1978). We were thus led to propose a modification of the classical infinite allele model in which variability of mutation rates at different loci is incorporated (Nei *et al.* 1976). By considering the variability of subunit molecular weight of mammalian enzymes (Darnell & Klotz, 1975) and the variation of amino acid substitution rates per protein for over forty proteins (Dayhoff, 1976; Wilson, Carlson & White, 1977) it was suggested that the variability of mutation rates in natural populations can be represented as a first approximation by a gamma distribution. Under this assumption, the distribution of allele frequencies for a collection of loci in a random mating population of effective size N is given by

$$\Phi(x) = \frac{\bar{M}x^{-1}(1-x)^{-1}}{[1-\bar{M}\alpha\ln(1-x)^{\alpha+1}]}, \tag{1}$$

where $\bar{M} = 4N\bar{v}$ represents the product of four times the effective population size and average mutation rate over all loci, and $\alpha = \bar{v}^2/V_v$, V_v being the variance of mutation rate over loci (Nei *et al.* 1976).

0016-6723/79/2828-7710 $01.00

In the present paper we shall present further results of the varying mutation model. In particular we shall be concerned with the effect of a varying mutation rate upon the number of alleles at a locus, and the relationship between number of alleles and heterozygosity. Consideration of the variance of the number of alleles at different loci is important to any interpretation of the reported positive association between number of alleles and molecular weight at loci in natural populations (Koehn & Eanes, 1977, 1978). The relationship between heterozygosity and allele number will be considered as it relates to strategies of estimating the parameter $4N_e v$ from electrophoretic data. We shall also consider the problem of detecting differences in average mutation rates between monomorphic and polymorphic loci within a population.

2. MEAN AND VARIANCE OF OBSERVED NUMBER OF ALLELES

For a constant mutation rate Ewens (1972) worked out the expectation and variance of k, the observed number of alleles in a simple of n genes under the infinite allele model which are given by

$$E_M(k) = \sum_{i=0}^{n-1} \frac{M}{M+i} \tag{2}$$

and

$$V_M(k) = E_M(k) - \sum_{i=0}^{n-1} \frac{M^2}{(M+i)^2}, \tag{3}$$

respectively (Ewens, 1972).

When the mutation rate varies according to a gamma distribution with parameters $\alpha = \bar{v}^2/V_v$ and $\beta = \alpha/\bar{M}$, the expectation of the observed number of alleles in a sample of n genes is given by

$$\begin{aligned} E(k) &= E_f[E_M(k)] \\ &= 1 + \sum_{j=1}^{n-1} H(j), \end{aligned}$$

where $E_f(\cdot)$ is the expectation operator over the distribution of mutation rate, and

$$H(j) = \frac{(j\beta)^\alpha}{\Gamma(\alpha)} \int_0^1 \frac{y^\alpha}{(1-y)^{\alpha+1}} e^{-j\beta y/(1-y)} dy$$

(Nei *et al.* 1976). The variance of k in such a case is given by

$$\begin{aligned} V(k) &= E_f[V_M(k)] + V_f[E_M(k)] \\ &= \{2 - E(k)\}\{E(k) - 1\} \\ &\quad + 2 \sum_{\substack{i,j=1 \\ (j>i)}}^{n-1} \frac{(i\beta)^\alpha}{\Gamma(\alpha)} \int_0^1 \frac{i}{j + (i-j)y} \left(\frac{y}{1-y}\right)^{\alpha+1} e^{-i\beta y/(1-y)} dy. \end{aligned} \tag{4}$$

It may be noted that both $E(k)$ and $V(k)$ depend upon the average value of M (averaged over all loci), $\bar{M}$ as well as the number of individuals sampled. For the constant mutation model ($\alpha = \infty$) $E(k)$ as well as $V(k)$ can be computed directly using equations (3) and (4) whereas the same quantities for the varying model have

to be obtained by numerical integration. In Fig. 1 we present the relationship between these two quantities [$V(k)$ versus $E(k)$] for both constant and varying mutation model for two different sample sizes ($n = 100$ and 500). Note that the relationship between $E(k)$ and $V(k)$ is less sensitive to variations of sample size (number of genes sampled, n) when mutation rates are constant ($\alpha = \infty$) than when they vary especially when the expectation of the observed number of alleles is particularly large. For example, when $E(k)$ is approximately 5, by increasing the sample size from 100 to 1000 (not shown in Fig. 1), $V(k)$ is increased by approximately 8 % when $\alpha = \infty$, 10 % when $\alpha = 2$, and 15 % if $\alpha = 1$.

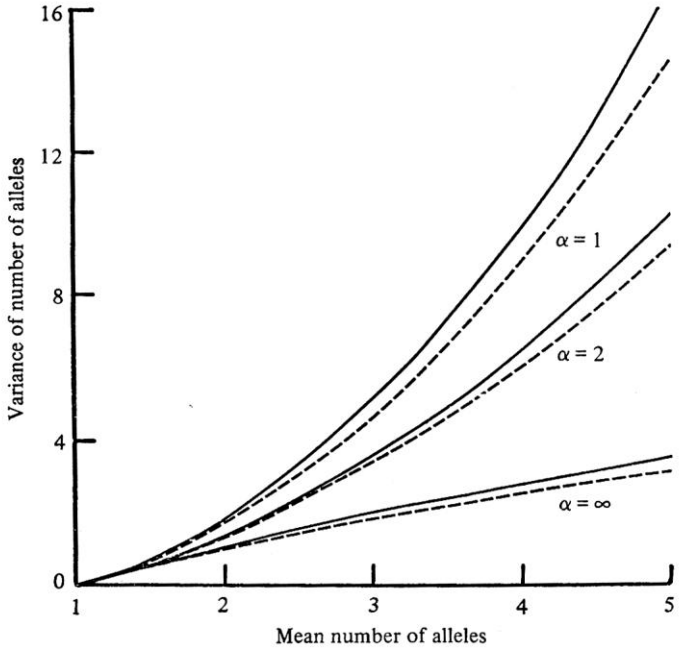

Fig. 1. Relationships between mean number of alleles and interlocus variance of number of alleles under the infinite allele model with varying mutation rate in an equilibrium population. α is the inverse of squared coefficient of variation of mutation rates over loci and n is the number of genes sampled per locus. The solid lines are for $n = 500$ and the dashed lines for $n = 100$.

Fig. 1 may be used to examine the relationship between mutation rate at a locus and the number of observed alleles in different organisms. Koehn & Eanes (1977) and Eanes & Koehn (1978) recently examined such a relationship indirectly by relating the observed number of alleles at a locus with subunit molecular weight of the enzyme. To interpret their results, we assume that the per codon mutation rate is constant for all loci. It then follows that the mutation rate at a locus will be directly proportional to the size of the cistron, and therefore subunit molecular weight would bear a direct relationship with the mutation rate at a locus. Of course if neutral mutation rate is not perfectly correlated with molecular size our calculated relationships will be altered to some degree. We will consider this point in greater

detail in the sequel. From study of equations (2)–(4) we can determine how much of the variation of the observed number of alleles can be explained by the relationship of mutation rate, v, with k, the observed number of alleles.

The amount of variability in the observed number of alleles that can be accounted for by a relationship with differences in mutation rate can be expressed as

$$V(k) - E_f[V_M(k)].$$

Thus, the proportion of variance of k which is attributable to variation in mutation rate is given by

$$R^2(k) = \frac{V_f[E_M(k)]}{V(k)}. \tag{5}$$

Clearly, $R^2(k)$ is also a function of $\bar{M}$ and sample size. In practice, however, $\bar{M}$ is not an observable quantity. One quantity which is observable is average heterozygosity. In Fig. 2 we present the relationship of $R^2(k)$ with the average heterozygosity ($\bar{H}$) estimated from values of $\bar{M}$ for $\alpha = 1$ and 2 (solid and dashed lines). For comparison, Fig. 2 also shows (dotted lines) the equivalent relationship

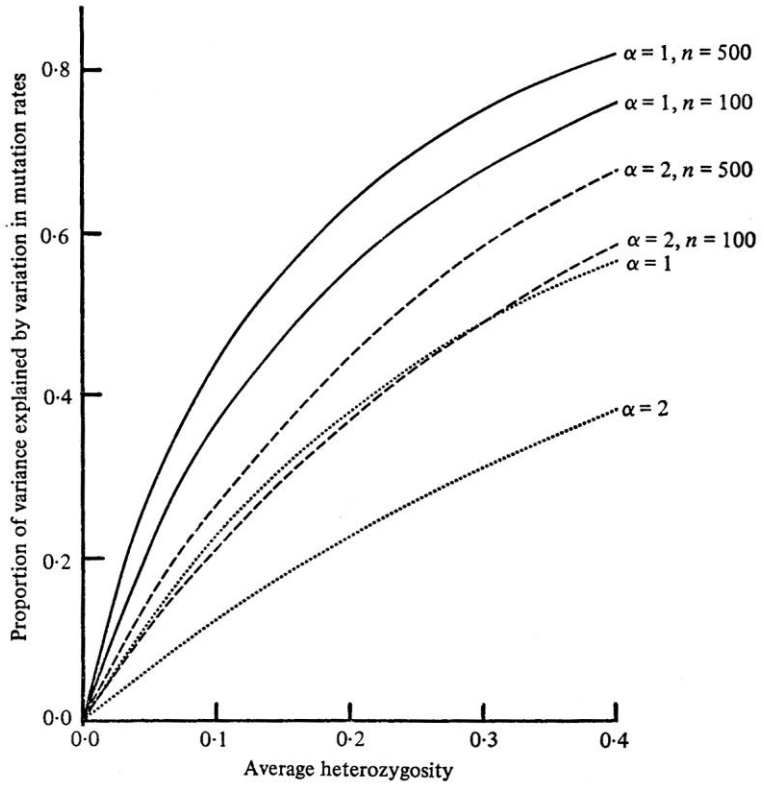

Fig. 2. Squared correlation (proportion of variance explained) between mutation rate and single-locus heterozygosity and between mutation rate and number of alleles observed at locus (in surveys of n genes/locus) as a function of the average heterozygosity in the population under the infinited allele model with varying mutation rate.

between $\bar{H}$, the average heterozygosity, and the proportion of explained variance of heterozygosity ($R^2(h)$), which can be accounted for by its relationship with mutation rate. $R^2(h)$ is given by

$$R^2(h) = \frac{V_f[E_M(h)]}{V(h)}, \tag{6}$$

where

$$V(h) = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^1 \frac{2-y^2}{(2-y)(3-2y)} \frac{y^\alpha}{(1-y)^{\alpha+1}} e^{-\beta y/(1-y)} dy - H^2(1)$$

(Nei *et al.* 1976), and

$$V_f[E_M(h)] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{M^{\alpha-1}}{(1+M)^2} e^{-\beta M} dM - (1-\bar{H})^2. \tag{7}$$

The expression in (7) can be shown to be mathematically equivalent to the covariance of heterozygosities between two populations at steady state (Chakraborty *et al.* 1978).

As seen in Fig. 2, the squared correlation between v and k can be 1·5–2 times as large as the corresponding squared correlation between v and h. For an organism with average heterozygosity of 8 %, the expected correlation between v and k can be 0·62 in a sample of 500 genes, whereas the expected correlation between v and h in such a population would be only 0·44 if the mutation rate has a coefficient of variation equal to one. It seems therefore reasonable that in man, for example, Koehn & Eanes (1978) found a significant relationship between the subunit molecular weight and the number of alleles whereas Harris, Hopkinson & Edwards (1977) found the relationship between subunit molecular weight and heterozygosity to be statistically insignificant. Of course, there are several other factors which may decrease the observed correlations between the number of alleles or heterozygosity and the subunit molecular weight, as emphasized in Nei *et al.* (1978). We have reported elsewhere a study of the relationship between molecular weight, number of alleles and heterozygosity (Chakraborty, Fuerst & Nei, 1980). This included an examination of data taken from the literature on 122 species of vertebrates and Drosophila. There was a significant tendency for the correlation between molecular weight and heterozygosity to be weaker than the correlation between molecular weight and number of alleles, as predicted by the results presented here. It was found that 78 of the 122 species showed this relationship, despite the large sampling error expected for both measures. The reader is directed to our other paper for details of the analysis.

One might, however, argue that in the above derivations we have considered only the infinite allele model, and that electrophoretic date may not strictly follow these expectations. Chakraborty *et al.* (1978) and Nei *et al.* (1978) studied analytically the equivalent expressions for (6) and (7) under the stepwise mutation model. These studies showed that the proportion of variance of heterozygosity explained by the variation of mutation rate under such a model is slightly smaller than the corresponding expectations under the infinite allele model. However, the analytical treatment of variance of the observed number of alleles in a sample under the stepwise mutation model does not seem to be so easy since the sampling theory of the stepwise mutation model is not yet available.

To obtain the expected correlation between the mutation rate and the observed number of alleles under the stepwise mutation model, we conducted a Monte Carlo simulation following the procedure described in Chakraborty (1977). Table 1 presents the expected proportion of variances of heterozygosity and of the observed number of alleles which are explained by their relationship with the mutation rate under both models. Two thousand replicates were examined for each of 5 levels of

Table 1. *Proportion of variances of heterozygosity and observed number of alleles explained by variation of mutation rate under the two models of neutral mutations*

Average heterozygosity	Infinite allele model				Stepwise mutation model			
	$\bar{k}$	$R^2(h)$	$R^2(k)$	$r(h, k)$	$\bar{k}$	$R^2(h)$	$R^2(k)$	$r(h, k)$
0·011	1·071	0·032	0·065	0·696	1·071	0·038	0·084	0·707
0·044	1·295	0·115	0·224	0·715	1·246	0·102	0·175	0·714
0·105	1·772	0·238	0·424	0·740	1·577	0·223	0·332	0·746
0·215	2·877	0·396	0·633	0·765	2·201	0·327	0·449	0·778
0·452	7·083	0·610	0·839	0·769	3·754	0·309	0·408	0·790

average heterozygosity with $\alpha = 1$. The results under the infinite allele model (with $\alpha = 1$) are obtained from the analytical formulae given above. It is clear that for the same average heterozygosity a smaller proportion of variance of any of these two statistics (h or k) is explained by variations in mutation rate under the stepwise mutation model. The two models are, however, very similar in their expectations for small average heterozygosity values, a pattern seen for several other parameters as well (e.g. see Ohta & Kimura, 1975; Chakraborty, 1977).

3. CORRELATION BETWEEN HETEROZYGOSITY AND OBSERVED NUMBER OF ALLELES IN A SAMPLE

The analysis just described indicates that the relationship between the mutation rate and heterozygosity is at least qualitatively similar to that between the mutation rate and the observed number of alleles. The principal difference that emerges is in the quantitative magnitude of the effect of varying mutation; for the number of alleles a greater percentage of its variation is explained by its relationship with the mutation rate. This qualitative similarity is expected if the number of alleles at a locus and the heterozygosity are correlated. Intuitively this may be obvious, but no formal theories have so far been advanced. In this section, we obtain such correlations for several neutral models.

(i) *Infinite allele model with constant mutation rate*

Let h and k denote the observed heterozygosity and the number of alleles in a sample of n genes at a locus chosen from a population. Let $f = 1 - h$ denote the sample homozygosity. According to Ewens (1972), the expectation of k, $E(k)$, is given by equation (2) whereas the expectation of f, $E(f)$, is

$$E(f) = E_k[E(f|k)], \tag{8}$$

where $E_k(\cdot)$ is the expectation over the distribution of k in the sample, and $E(f|k)$

is the expectation of the distribution of f given k in the sample. The value of $E(f|k)$, as obtained by Watterson (1977), is given by

$$E(f|k) = \frac{1}{n} + \left(1 - \frac{1}{n}\right) G(k),$$

where

$$G(k) = \sum_{l=1}^{k} S_n^{(l)} / S_n^{(k)},$$

$S_n^{(l)}$ being Stirling numbers of the first kind, and Stewart (1977) gives in explicit terms the distribution of k in the sample from Ewens' (1972) sampling theory. Thus,

$$\begin{aligned} E(f) &= \sum_{k=0}^{n} \left\{\frac{1}{n} + \left(1 - \frac{1}{n}\right) G(k)\right\} \frac{\Gamma(M)\, M^k (-1)^{n-k}}{\Gamma(n+M)} S_n^{(k)} \\ &= \frac{1}{n} + \left(1 - \frac{1}{n}\right) / (1+M). \end{aligned} \tag{9}$$

By similar computations, the expectation of the product of sample homozygosity and the number of alleles, $E(fk)$, is obtained as

$$\begin{aligned} E(fk) &= E_k[E(fk|k)] \\ &= \sum_{k=0}^{n} k \left[\frac{1}{n} + \left(1 - \frac{1}{n}\right) G(k)\right] (-1)^{n-k} \frac{\Gamma(M)\, M^k S_n^{(k)}}{\Gamma(n+m)} \\ &= \frac{E(k)}{n} + \left(1 - \frac{1}{n}\right) \left[\frac{E(k)}{1+M} + \frac{M}{(1+M)^2}\right], \end{aligned} \tag{10}$$

where $E(k)$ is as given by equation (2). Using (9) and (10) we thus obtain the covariance between h and k as

$$\mathrm{Cov}\,(h, k) = \left(1 - \frac{1}{n}\right) M/(1+M)^2. \tag{11}$$

To obtain the correlation between these two quantities we then need to compute $V(h)$, which in turn is

$$\begin{aligned} V(h) = V(f) &= E_k[V(f|k)] + V_k[E(f|k)] \\ &= \frac{n-1}{2n^3} \left[\frac{5n(n-1)+2}{M+1} - \frac{8(n-1)(n-2)}{M+2} + \frac{3(n-2)(n-3)}{M+3}\right] \\ &\quad - \frac{n-1}{n^2} \left(2 + \frac{n-1}{M+1}\right) / (M+1) \end{aligned} \tag{12}$$

using formula (4.3.9) of Watterson (1977) and some algebraic simplifications.

Some numerical values of the correlation between heterozygosity and number of alleles for various values of M and sample size n together with the sample average heterozygosity and the expected number of alleles are given in Table 2. It is clear that in a small sample the correlation between k and h is quite strong and that as sample size increases the correlation coefficient decreases slightly. Furthermore, Table 2 also indicates that in populations with larger average heterozygosity the correlation between h and k is weaker as compared to the same correlation in popula-

tions with lower genetic variability. The actual decrease in $r(h, k)$ as the average heterozygosity increases is very small, however, even smaller than the decrease resulting from increased sample size.

Table 2. *Correlation (r) between heterozygosity and observed number of alleles in a sample of n genes and the expectations of these two quantities for various M values and sample sizes (n) under the infinite allele model with constant mutation rates of neutral mutations*

(The average heterozygosity in the population is denoted by H.)

		$n = 25$			$n = 100$			$n = 1000$		
M	H	$E(h)$	$E(k)$	$r(h, k)$	$E(h)$	$E(k)$	$r(h, k)$	$E(h)$	$E(k)$	$r(h, k)$
0·01	0·010	0·010	1·038	0·855	0·010	1·052	0·752	0·010	1·075	0·630
0·05	0·048	0·046	1·185	0·848	0·047	1·255	0·744	0·048	1·370	0·621
0·10	0·091	0·087	1·363	0·841	0·090	1·502	0·734	0·091	1·733	0·611
0·15	0·130	0·125	1·534	0·833	0·129	1·743	0·725	0·130	2·089	0·602
0·20	0·167	0·160	1·699	0·827	0·165	1·978	0·717	0·167	2·439	0·593
0·30	0·231	0·222	2·014	0·815	0·228	2·432	0·702	0·231	3·123	0·577
0·40	0·286	0·274	2·311	0·804	0·283	2·866	0·689	0·286	3·788	0·563
0·50	0·333	0·320	2·591	0·795	0·330	3·284	0·678	0·333	4·436	0·551
1·00	0·500	0·480	3·816	0·762	0·495	5·187	0·637	0·500	7·485	0·506

(ii) *Infinite allele model with varying mutation rate*

Using the same sampling theory the variances and covariances of h and k can also be derived when the mutation rate as described in the previous section varies according to a given probability distribution. As was done previously, we continue to use a gamma distribution to represent the variation of mutation rates. The expectation of the sample homozygosity is then given by

$$E(f) = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \left[\frac{1}{n} + \left(1-\frac{1}{n}\right)/(1+M)\right] e^{-\beta M} M^{\alpha-1} dM$$

$$= \frac{1}{n} + \left(1-\frac{1}{n}\right)\beta \int_0^\infty e^{-y} \frac{y^{\alpha-1}}{\beta+y} dy/\Gamma(\alpha), \tag{13}$$

whereas the variance of sample homozygosity, $V(f)$, is derived using (9) and (12) as

$$V(f) = V_f\left[\frac{1}{n} + \left(1-\frac{1}{n}\right)/(1+M)\right] + E_f\left[\frac{n-1}{2n^3}\left\{\frac{5n(n-1)+2}{M+1} - \frac{8(n-1)(n-2)}{M+2}\right.\right.$$

$$\left.\left. + \frac{3(n-2)(n-3)}{M+3}\right\} - \frac{n-1}{n^2}\left(2+\frac{n+1}{M+1}\right)/(M+1)\right]$$

$$= \frac{n-1}{2n^3}\left[(5n^2-9n+2)\epsilon_1 - 8(n-1)(n-2)\epsilon_2 + 3(n-2)(n-3)\epsilon_3\right]$$

$$- \left(1-\frac{1}{n}\right)^2 \epsilon_1^2, \tag{14}$$

where

$$\epsilon_i = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^1 \left[e^{-\beta i y/(1-y)} (iy)^{\alpha-1}/(1-y)^\alpha\right] dy.$$

The covariance of h and k in this case is given by

$$\text{Cov}(h, k) = \frac{n-1}{n}\frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \frac{M^\alpha e^{-\beta M}}{(M+1)^2}dM + \frac{n-1}{n}\sum_{i=0}^{n-1}\frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \frac{M^{\alpha+1}e^{-\beta M}}{(M+1)(M+i)}dM$$

$$-\frac{n-1}{n}\left\{\frac{\beta^\alpha}{\Gamma(\alpha)}\right\}^2 \int_0^\infty \frac{M^\alpha e^{-\beta M}}{(1+M)}\sum_{i=0}^{n-1}\int_0^\infty \frac{M^\alpha e^{-\beta M}}{M+i}dM$$

$$= \frac{n-1}{n}\left[\frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \frac{M^\alpha e^{-\beta M}}{(1+M)^2}dM + \sum_{i=0}^{n-1} I_{1i} - I_{01}\sum_{i=0}^{n-1} I_{0i}\right],$$

where

$$I_{ij} = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \frac{M^{\alpha+1}e^{-\beta M}}{(M+i)(M+j)}dM. \tag{15}$$

Using (4), (14) and (15) the correlation between h and k with varying mutation was obtained. Some numerical computations are shown in Table 1 and Fig. 3. In contrast to Table 2, Fig. 3 indicates that as the average heterozygosity increases, the correlation coefficient between the number of alleles and the heterozygosity at a locus increases slightly when the mutation rate varies according to a gamma distribution. Note that this relationship is the reverse when the mutation rate remains the same at all loci. To obtain the results in Table 1 the mutation rate is assumed to be distributed as a gamma variate with coefficient of variation of unity.

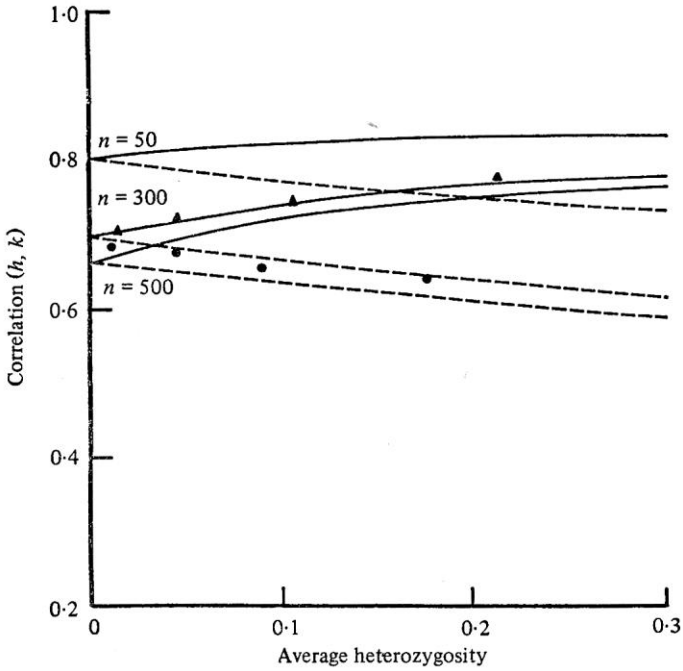

Fig. 3. Correlation between single locus heterozygosity and the number of alleles at a locus as a function of the average heterozygosity in the population. n = number of genes sampled per locus, smooth lines are for the infinite allele model with $\alpha = 1$ and dotted lines are for the infinite allele model with $\alpha = \infty$; ▲ and ● are from Monte Carlo simulation of the stepwise mutation model with $n = 300$, $\alpha = 1$ and $\alpha = \infty$, respectively.

As mentioned in the previous section, a sampling theory is not yet available for a step mutation model. To obtain the correlation between h and k we therefore conducted a Monte Carlo simulation of the charge-change model following Chakraborty (1977). In each generation $n = 300$ genes were sampled. Four cases were investigated each with 4000 replications with M (or average M) values 0·01, 0·05, 0·10 and 0·20. In Fig. 3 and Table 1 we indicate the relationship between $E(h)$ and Corr (h, k) under the step mutation model for constant (solid circles) as well as varying mutation rates (solid triangles). From these limited results it appears that Corr (h, k) under the step mutation model is essentially the same as that seen for the infinite

Table 3. *Relationship between heterozygosity and number of alleles in natural populations*

(The data is taken from that reported by Chakraborty, Fuerst & Nei (1980). n = number of genes sampled.)

	n = 100–300		n = 300–600		n = 600–2000	
Average heterozygosity	No. of species	Average $r(h, k)$	No. of species	Average $r(h, k)$	No. of species	Average $r(h, k)$
0·025–0·075	46	0·753	10	0·687	5	0·615
0·075–0·125	31	0·775	8	0·727	3	0·578
0·125–0·175	6	0·750	3	0·747	4	0·687
0·175–0·225	7	0·798	2	0·768	3	0·765

allele case. We have also examined the observed relationship between number of alleles and heterozygosity in a large body of data on electrophoretic surveys collected from the literature. (See Chakraborty *et al.* 1980, for a full description of data.) Species with 100 or more gene products determined at 20 or more loci were included. The results of this study for 128 populations with heterozygosity values above 0·025 are shown in Table 3. The data represent several different average heterozygosity and average sample size ranges. If we compare these results with Fig. 3 we see that the agreement between our theoretical prediction and the average correlations for the various species groups is good. Clearly evident in the data is the tendency for Corr (h, k) to decrease with increasing sample size. Less clear is the question of whether a tendency exists for Corr (h, k) to rise with increasing average heterozygosity, although even here the data do show some trend in this direction.

4. AVERAGE MUTATION RATES IN POLYMORPHIC VERSUS MONOMORPHIC LOCI

We have thus far considered two measures of intrapopulation genic variability, namely, the observed number of alleles and the average heterozygosity. Harris, Hopkinson & Edwards (1977) used yet a third approach when they considered the difference in average molecular weights between polymorphic and monomorphic loci in their study of electrophoretic variation in humans. It should be obvious that the expected difference in molecular weight can only be predicted following a consideration of the effect of varying mutation rates between loci. In discussing the

efficiency of this test let us first compare the average mutation rates for the polymorphic with those for the monomorphic loci, and obtain the minimum sample size needed to establish a significant difference between them.

We define a locus as polymorphic when the frequency of the most common allele is less than $1-q$, where q is a small quantity. The expected proportion of polymorphic loci is then given by

$$1-\int_{1-q}^{1}\Phi(x)\,dx,$$

where $\Phi(x)$ is as given in (1). In a previous publication (Nei *et al.* 1976) we obtained a closed expression for this proportion, which may as well be obtained by taking weighted average of $1-q^M$ over the variation of M since for a locus with a particular mutation rate v, the probability of polymorphism is given by $1-q^M$ (Kimura & Ohta, 1971). Thus, the expected proportion of polymorphic loci is

$$\begin{aligned}&=\int_0^\infty (1-q^M)\frac{\beta^\alpha}{\Gamma(\alpha)}e^{-\beta M}M^{\alpha-1}\,dM\\&=1-\left(\frac{a}{\alpha-\bar{M}\log q}\right)^\alpha\end{aligned}\tag{16}$$

It may now be shown that the expected mutation rate for the class of polymorphic loci will be

$$\begin{aligned}\bar{v}_p&=\frac{\beta^\alpha}{4N\Gamma(\alpha)}\int_0^\infty Me^{-\beta M}M^{\alpha-1}(1-q^M)\,dM\Big/\left[1-\left(\frac{\alpha}{\alpha-\bar{M}\log q}\right)^\alpha\right]\\&=\bar{v}[1-P^{\alpha+1}]/[1-P^\alpha],\end{aligned}\tag{17}$$

when $P=\alpha/(\alpha-\bar{M}\log q)$, and $\bar{v}$ is the average mutation rate for all loci. The variance of the mutation rate for the class of polymorphic loci is then given by

$$\begin{aligned}\sigma_p^2&=\frac{\beta^\alpha}{(4N)^2\Gamma(\alpha)}\int_0^\infty M^2e^{-\beta M}M^{\alpha-1}(1-q^M)\,dM\Big/\left[1-\left(\frac{\alpha}{\alpha-\bar{M}\log q}\right)^\alpha\right]-\bar{v}_p^2\\&=\bar{v}^2(\alpha+1)(1-P^{\alpha+2})/[\alpha(1-P^\alpha)]-\bar{v}^2[1-P^{\alpha+1}]^2/[1-P^\alpha]^2.\end{aligned}\tag{18}$$

Similarly, for the monomorphic loci, the mean and variance of the mutation rate are given by

$$\begin{aligned}\bar{v}_m&=\frac{\beta^\alpha}{4N\Gamma(\alpha)}\int_0^\infty Me^{-\beta M}M^{\alpha-1}q^M\,dM/P\\&=P\bar{v}\end{aligned}\tag{19}$$

and

$$\begin{aligned}\sigma_m^2&=\frac{\beta^\alpha}{(4N)^2\Gamma(\alpha)}\int_0^\infty M^2e^{-\beta M}M^{\alpha-1}q^M\,dM/P^\alpha-\bar{v}_m^2\\&=\bar{v}^2P^2/\alpha,\end{aligned}\tag{20}$$

when P is as defined in (17).

From expressions (16)–(20) we can state that if a random sample of n loci are examined, a statistically significant (at, say, 5% level) difference between the

mutation rates of the polymorphic and monomorphic classes of loci would be observed if

$$n > [(Z_{0\cdot05} - Z_\gamma)\left(\frac{\sigma_p^2}{1-P} + \frac{\sigma_m^2}{P}\right)/(\bar{v}_p - \bar{v}_m)]^2, \tag{21}$$

where Z_γ is defined by Prob $[Z > Z_\gamma \,|\, Z \sim N(0,1)] = \gamma$.

In (21) we appealed to the central limit theorem since the sample distribution for the mutation rates is not Gaussian. The same inequality for the total sample size

Table 4. *The minimum number of loci to be examined to observe significant differences (at the 1 % and 5 % levels) in the average mutation rates among the monomorphic and polymorphic classes of loci for different average heterozygosities*

(The coefficient of variation of the distribution of overall mutation rate is taken as unity.)

Average heterozygosity	$q = 0\cdot01$			$q = 0\cdot05$		
	Power = 0·5	0·75	0·9	0·5	0·75	0·9
			1 % level test			
0·02	123	205	296	186	310	448
0·05	44	74	107	67	110	158
0·10	25	42	61	36	60	86
0·15	18	29	42	23	39	56
0·20	14	23	32	18	30	42
0·25	11	19	27	14	23	33
			5 % level test			
0·02	61	123	195	93	185	295
0·05	23	45	72	35	68	105
0·10	12	25	40	18	36	57
0·15	9	18	28	12	24	38
0·20	7	14	21	9	18	28
0·25	6	11	17	7	15	23

(number of loci) would hold for testing the difference between the average molecular weights for these two classes of loci if the molecular weights are strictly proportional to the mutation rates. We must note that our study of the relationship between amino acid substitution rates and molecular weight indicates that a correlation of no greater than 0·5 exists between mutation rate and molecular size (Nei *et al.* 1978). In Table 4 we present some values for these critical sample sizes. Note that in (21) the critical sample size is a function of the probability of monomorphism which is in turn functionally related to the average heterozygosity, $\bar{H}$. In Table 4 therefore we compute the critical sample size (right side of inequality 21) for several average heterozygosities for $q = 0\cdot01$ and 0·05 and two levels of significance (1 % and 5 %). In all these computations the coefficient of variation of the overall distribution of mutation rate is taken as unity. It can be seen that in populations with larger average heterozygosity a smaller number of loci need to be studied to observe significant differences in the average mutation rates between the two classes of loci than is required for populations with smaller average heterozygosity. If the variance of the distribution of mutation rate is smaller than we have assumed ($\alpha > 1$ in equations 16–20), the number of loci needed to detect significant differences

would be larger than the ones presented in Table 4. Of more importance, however, is the fact that the correlation between the mutation rate and the size of a molecule is not perfect (Nei *et al.* 1976, 1978). In this case the actual efficiency of tests based on differences in molecular weight between classes of alleles would be even weaker than indicated by our results. Given these considerations, the negative findings of Harris *et al.* (1977) are not greatly surprising. Nevertheless, we have noted elsewhere that the electrophoretic data from humans does exhibit a general tendency to be inconsistent with several of the predictions of the mutation-drift hypothesis (Nei *et al.* 1978; Chakraborty *et al.* 1980).

5. DISCUSSION

We have demonstrated in this paper that under selective neutrality a high correlation exists between heterozygosity and the number of alleles at a locus. These two statistics provide alternative estimates of the parameter $4N_e v$ for testing the mutation-drift hypothesis, a situation which in recent years has generated controversy regarding the efficiency of estimating $4N_e v$. From statistical considerations Ewens (1972) and Watterson (1978) argue that number of alleles is the most efficient statistic for estimating the above parameter. This is true as long as we attempt to obtain a locus-specific estimate. On the contrary, for testing the neutral mutation hypothesis (which, we contend, is 'majority-rule' – postulating that the majority of genic variability of a population is mainly due to neutral or nearly neutral mutations) heterozygosity per locus may be a more appropriate measure (Fuerst, Chakraborty & Nei, 1977; Li, 1979), since it directly gives an average estimate of $4N_e v$. This averaging has an added advantage in the sense that heterozygosity as an estimator of $4N_e v$ has robustness in the presence of rare deleterious alleles whereas this is not true of an estimator using the number of alleles (Li, 1979). The question of whether a locus-specific test of neutrality or a test of the simultaneous behaviour of many loci (sometimes called 'bulk test') should be preferred is closely associated with this controversy and is discussed in detail elsewhere (for example, see Fuerst *et al.* 1977; Chakraborty *et al.* 1978; Li, 1979). To cite a few examples to show how different the alternate estimates can be, we use the genic variability observed in some human studies. Neel *et al.* (1978) recently estimated the locus-specific $4N_e v$ values from their survey of genetic variants of 22 proteins in Japanese populations from the observed number of alleles at each locus. From the 17 variable loci (in each of which the observed number of alleles is more than one) the $4N_e v$ estimate as obtained from number of alleles was $0{\cdot}350 \pm 0{\cdot}089$. If we compute the heterozygosities at these loci and use $\hat{h}/(1-\hat{h})$ as an estimator of $4N_e v$, where $\hat{h}$ is the average heterozygosity (0·103, in this case), the $4N_e v$ estimate turns out to be $0{\cdot}116 \pm 0{\cdot}044$. In Harris, Hopkinson & Robson's (1974) survey of 43 enzyme loci from Caucasians of the British Islands the corresponding estimates of $4N_e v$ are $0{\cdot}312 \pm 0{\cdot}078$ on the basis of number of alleles and $0{\cdot}133 \pm 0{\cdot}040$ based on average heterozygosity, where both of these are based on 28 variable loci. Thus, in both cases the average $4N_e v$ estimates are higher when number of alleles are used to estimate this quantity. In fact, Ewens (1972) proposed this estimator only to provide, as

we said before, a locus-specific $4N_e v$ value. The heterozygosity measure, on the other hand, has never been advocated as a locus-specific estimator since this quantity when computed for a single locus has a large stochastic error (Nei & Roychoudhury, 1974; Li & Nei, 1975). Since point estimation of $4N_e v$ is possible only for variable loci using number of alleles, it is obvious that it provides a conditional estimate on the supposition that the number of alleles observed in the sample is greater than one. It may, therefore, be worthwhile to evaluate the expectation of this conditional estimator as compared to the true value of $M = 4N_e v$. Let $\hat{M}_k$ denote the estimator of M given that in a sample of n genes from this locus more than one allele is observed. Then we have

$$M = Q(1).0 + (1 - Q(1)).E(\hat{M}_k),$$

where $Q(1)$ is the probability of finding one allele at this locus, given by

$$Q(1) = \frac{\Gamma(M).M.(n-1)!}{\Gamma(n+M)} = \prod_{i=1}^{n-1}\left(\frac{i}{M+i}\right).$$

The proportional bias of $\hat{M}_k$, therefore, will be given by $Q(1)/(1-Q(1))$ which would vary from locus to locus if sample size (n) varies over loci. For example, if $n = 8000$, the proportional bias can be of the order of 63 % of the estimate of M when $4N_e v$ is about 0·10 for the locus. This, of course, does not account for the difference between the two alternative estimates (which is two and one-half to three fold in human surveys). In both surveys a large number of rare alleles (alleles with a frequency less than 1 %, for example) are observed in the sample, some of which may be slightly deleterious. These rare alleles, as we contend, affect estimates of M when the number of alleles is used but do not affect estimates using heterozygosity since the average heterozygosity is only marginally affected by the presence of rare alleles.

Finally, we wish to comment on our use of the gamma distribution to represent the variability in mutation rate among loci. The gamma distribution is notable for its extreme robustness. If our principal purpose were to provide a statistical fit to the true distribution of mutation rate this might be a distinct disadvantage. It must be stressed that this has not been our purpose. Current biological information on locus-specific mutation rates is extremely limited. Given this paucity of data, the primary purpose of our studies on varying mutation has been to provide insights into the potential effects of differences in mutation rates. From this point of view the robustness of the gamma distribution is of distinct advantage. The fact that useful differences between the constant and varying mutation model have been identified justifies this approach (Fuerst *et al.* 1977; Chakraborty *et al.* 1978). We anticipate that future advances in molecular biology and mutation rate monitoring may provide a more specific formulation than provided by our earlier studies, but we feel that our general conclusions will remain unaltered.

Research supported by U.S. Public Health Service grants GM 20293, GM 19513 and CA 19311 from the U.S. National Institutes of Health.

REFERENCES

CHAKRABORTY, R. (1977). Simulation results with stepwise mutation model and their interpretations. *Journal of Molecular Evolution* **9**, 313–322.

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

CHAKRABORTY, R., FUERST, P. A. & NEI, M. (1980). Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies within populations. *Genetics*. (in the Press).

DARNALL, D. W. & KLOTZ, I. M. (1975). Subunit constitution of proteins: A table. *Archives of Biochemistry and Biophysics* **166**, 651–682.

DAYHOFF, M. O. (1976). *Atlas of Protein Sequences and Structure*, vol. 5, Suppl. 2. National Biomedical Research Foundation, Washington.

EANES, W. F. & KOEHN, R. K. (1978). The relationship between subunit size and the number of rare electrophoretic alleles in human enzymes. *Biochemical Genetics* **16**, 971–985.

EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

FUERST, P. A., CHAKRABORTY, R. & NEI, M. (1977). Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics* **86**, 455–483.

HARRIS, H., HOPKINSON, D. A. & EDWARDS, Y. H. (1977). Polymorphism and the subunit structure of enzymes: A contribution to the neutralist-selectionist controversy. *Proceedings of the National Academy of Sciences, U.S.A.* **74**, 698–701.

HARRIS, H., HOPKINSON, D. A. & ROBSON, E. B. (1974). The incidence of rare alleles determining electrophoretic variants: Data on 43 enzyme loci in man. *Annals of Human Genetics, London* **37**, 237–253.

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

KIMURA, M. & OHTA, T. (1971). *Theoretical Aspects of Population Genetics*. Princeton University Press.

KOEHN, R. K. & EANES, W. F. (1977). Subunit size and genetic variation of enzymes in natural populations of Drosophila. *Theoretical Population Biology* **11**, 330–341.

KOEHN, R. K. & EANES, W. F. (1978). Molecular structure and protein variation within and among populations. *Evolutionary Biology* **11**, 39–100.

LI, W.-H. (1979). Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. *Genetics* (in the Press).

LI, W.-H. & NEI, M. (1975). Drift variances of heterozygosity and genetic distance in transient states. *Genetical Research* **25**, 229–248.

NEEL, J. V., UEDA, N., SATOH, C., FERRELL, R. E., TANIS, R. J. & HAMILTON, H. B. (1978). The frequency in Japanese of genetic variants of 22 proteins. V. Summary and comparison with data on Caucasians from the British Isles. *Annals of Human Genetics, London*.

NEI, M., CHAKRABORTY, R. & FUERST, P. A. (1976). Infinite allele model with varying mutation rate. *Proceedings of the National Academy of Sciences, U.S.A.* **73**, 4164–4168.

NEI, M., FUERST, P. A. & CHAKRABORTY, R. (1978). Subunit molecular weight and genetic variability of proteins in natural populations. *Proceedings of the National Academy of Sciences, U.S.A.* **75**, 3359–3362.

NEI, M. & ROYCHOUDHURY, A. K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics* **76**, 379–390.

OHTA, T. & KIMURA, M. (1975). Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proceedings of the National Academy of Sciences, U.S.A.* **72**, 2761–2764.

STEWART, F. M. (1977). Computer algorithm for obtaining a random set of allele frequencies for a locus in an equilibrium population. *Genetics* **86**, 482–483.

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

WILSON, A. C., CARLSON, S. S. & WHITE, T. J. (1977). Biochemical evolution. *Annual Review of Biochemistry* **46**, 573–639.

WRIGHT, S. (1949). Genetics of populations. *Encyclopedia Britannica*, 14th ed. **10**, 111–112.