STATISTICAL STUDIES ON PROTEIN POLYMORPHISM IN NATURAL POPULATIONS I. DISTRIBUTION OF SINGLE LOCUS HETEROZYGOSITY[1]

PAUL A. FUERST, RANAJIT CHAKRABORTY AND MASATOSHI NEI

Center for Demographic and Population Genetics, University of Texas at Houston, Texas 77030

Manuscript received October 15, 1976
Revised copy received January 1, 1977

ABSTRACT

Surveying the literature, the frequency distribution of single-locus heterozygosity among protein loci was examined in 95 vertebrate and 34 invertebrate species with the aim of testing the validity of the mutation-drift hypothesis. This distribution did not differ significantly from that expected under the mutation-drift hypothesis for any of the species examined when tested by the Kolmogorov-Smirnov goodness-of-fit statistic. The agreement between the observed interlocus variance of heterozygosity and its theoretical expectation was also satisfactory. There was an indication that variation in the mutation rate among loci inflates the interlocus variance of heterozygosity. The variance of heterozygosity for a homologous locus among different species was also studied. This variance generally agreed with the theoretical value very well, though in some groups of Drosophila species there was a significant discrepancy. The observed relationship between average heterozygosity and the proportion of polymorphic loci was in good agreement with the theoretical relationship. It was concluded that, with respect to the pattern of distribution of heterozygosity, the majority of data on protein polymorphisms are consistent with the mutation-drift hypothesis. After examining alternative possible explanations involving selection, it was concluded that the present data cannot be explained adequately without considering a large effect of random genetic drift, whether there is selection or not.

IT is now well recognized that most natural populations contain a large amount of genetic variation at the protein level (see LEWONTIN 1974; NEI 1975; and SELANDER 1976 for reviews). KIMURA (1968) and KIMURA and OHTA (1971) (see also ROBERTSON 1967 and CROW 1968) proposed that the majority of this variation is selectively neutral or nearly neutral. The simplicity of the assumptions of this hypothesis, often called the neutral mutation hypothesis or the mutation-drift hypothesis, results in a theory which is very powerful in predicting the evolutionary changes of populations. The hypothesis is concerned with the behavior of a "majority" of the genes that are incorporated into the population during evolution and does not deny the existence of deleterious genes or of a small proportion of advantageous or overdominant genes. In fact, KIMURA and OHTA (1973) maintain the view that the majority of fresh mutations are dele-

[1] This work was supported by grants from the Public Health Service and the National Science Foundation.

terious but, because of their deleterious effects, they are quickly eliminated from the population and contribute little to the genetic variability of a population. Because of this nature of the hypothesis, it is obvious that proper tests of its validity must be statistical. The demonstration of deterministic selection at a few loci cannot constitute proof against the hypothesis. (See NEI (1975) for the detailed properties of the mutation-drift hypothesis. We note that in many writings this hypothesis has been misinterpreted.)

Statistical testing of the mutation-drift hypothesis was initiated by KIMURA and OHTA (1971), who studied the relationship between average heterozygosity and the proportion of polymorphic loci. Subsequently, the relationship between the gene frequency and heterozygosity (YAMAZAKI and MARUYAMA 1972; CROW 1972; LATTER 1975), the ratio of the actual to the effective numbers of alleles per locus (JOHNSON 1972; KIRBY and HALLIDAY 1973; YAMAZAKI and MARUYAMA 1973), the relationship between the mean and variance of heterozygosity (NEI 1975), and the allele frequency distributions within populations (OHTA 1975) have been studied to see whether the data agree with predictions from the mutation-drift hypothesis. Many of these studies have been based on a relatively small amount of data derived from a few species, or when data from many species were used, little attention has been given to the effect of heterogeneous data. Since the number of publications of gene frequency data has increased tremendously in the past few years, it is now possible to conduct a more detailed statistical analysis.

One might criticize this statistical approach on the grounds that agreement between data and theory is not itself proof of the mutation-drift hypothesis, since the same data might also be explained by some combination of various selective genes. This criticism is valid, though certain types of selective hypotheses can easily be ruled out. However, the testing of many different predictions of the mutation-drift hypothesis will increase the probability of rejecting the hypothesis, if it is truly incorrect. Of course, even with a detailed series of tests there is a certain probability that the mutation-drift hypothesis will not be rejected. Obviously, this approach is the same as the classical statistical testing of a null hypothesis. As long as the null hypothesis of neutral mutations is not rejected, we can use it as a provisional theory. Clearly, a small degree of deviation of the data from the predictions of the theory of purely neutral mutations is not damaging to the hypothesis, since it is concerned with the majority of genes, as mentioned earlier. This indicates that the tests of the mutation-drift hypothesis by using the theory of purely neutral mutations are somewhat severe.

With this philosophy in mind, we have conducted an extensive series of statistical analyses of gene frequency data to test the null hypothesis of neutral mutations. The results obtained will be reported in this series of papers. In this first paper we shall examine several intrapopulational properties of heterozygosity, *i.e.*, the relationship between the mean and variance of heterozygosity, the distribution of heterozygosity, and the relationship between average heterozygosity and proportion of polymorphic loci. A preliminary result of this study

was reported previously (NEI, FUERST and CHAKRABORTY 1976b). Here the results of a more refined and extensive data analysis will be presented. We shall also present some theoretical background required for the data analysis. We note that heterozygosity is the most appropriate measure of genetic variability of a population (NEI 1975) and this quantity is affected only slightly by the existence of deleterious alleles, which are irrelevant to evolution. Therefore, it is important to know the properties of this quantity.

MATERIALS AND METHODS

The data used in this study were collected from the literature using two criteria. First, we have limited ourselves to populations which have been surveyed for at least twenty electrophoretically detectable loci. Second, we have considered only populations in which at least thirty genomes were examined at each locus. (When the variance of heterozygosity was computed from the variation among different species for a given locus, these criteria were not used, as will be mentioned later.) These criteria are based on NEI and ROYCHOUDHURY's (1974a) study of the sampling variance of heterozygosity. They showed that it is important to study many loci in estimating the mean and variance of heterozygosity, whereas the number of individuals is relatively unimportant. Our previous analysis (NEI, FUERST and CHAKRABORTY 1976b) used slightly less stringent criteria (15 loci). We believe that the minimum number of 20 loci will provide more accurate results. Since many experimental reports have appeared since our previous study, the number of species which satisfy the revised criteria remains large. We have used data from 129 species and 6 differentiated subspecies in animals. (They are listed in Tables 2 and 3.)

Throughout this paper, the heterozygosity at a locus has been computed by the formula $h = 1 - \Sigma x_i^2$, where x_i is the frequency of the ith allele at the locus. Namely, in this paper the word *heterozygosity* is used in the sense of *gene diversity* as defined by NEI (1975, p. 129). The average heterozygosity of a population was estimated by the average value ($\hat{H}$) of h over all loci examined. The variance of heterozygosity was computed from the variation of heterozygosity among loci in a population as well as from the variation among species for homologous loci. The estimate [$\hat{V}(h)$] of this variance was obtained by subtracting the intralocus sampling variance from the total interlocus variance of heterozygosity according to the procedure by NEI and ROYCHOUDHURY (1974a). When a species was represented by more than one population, average values of the statistics were used. In such cases, however, only populations satisfying our minimum criteria were included.

DATA ANALYSIS

Mean and variance of heterozygosity within species

Theoretical background: Using the infinite allele model of neutral mutation, KIMURA and CROW (1964) showed that the expected heterozygosity in an equilibrium population is given by

$$H = M/(1 + M), \tag{1}$$

where $M = 4N_e v$, in which N_e is the effective population size, and v is the mutation rate per locus per generation. WATTERSON (1974), STEWART (1976), and LI and NEI (1975) have shown that the variance of single locus heterozygosity is given by

$$V(h) = \frac{2M}{(M+1)^2(M+2)(M+3)} . \tag{2}$$

Formulae (1) and (2) enable us to test the mutation-drift hypothesis without knowing the values of N_e and v separately. Namely, if we estimate M from the average heterozygosity over many loci, we can compute the theoretical variance by using (2). This theoretical variance may be compared with the observed variance. The relationship between (1) and (2) is given by the solid lines in Figures 1 and 2.

One might question the applicability of the above formulae, since all data used in this study were obtained by electrophoresis. OHTA and KIMURA's (1973) stepwise mutation model is presumably more appropriate for such electrophoretic data. With this model, the expected heterozygosity is given by $H = 1 - 1/\sqrt{1 + 8N_e v}$. MORAN (1975) recently obtained a formula for the variance of heterozygosity for this model. It is given in an integral form but can be evaluated by numerical integration. The relationship between H and $V(h)$ for the stepwise mutation model is also given in Figures 1 and 2 (broken lines). In practice, the change in electrophoretic mobility may not be strictly stepwise (JOHNSON 1974). In this case the relationship between H and $V(h)$ is expected to lie between the two models.

Another factor that would affect the relationship between H and $V(h)$ is variation in the mutation rate among loci, since $V(h)$ is estimated from the

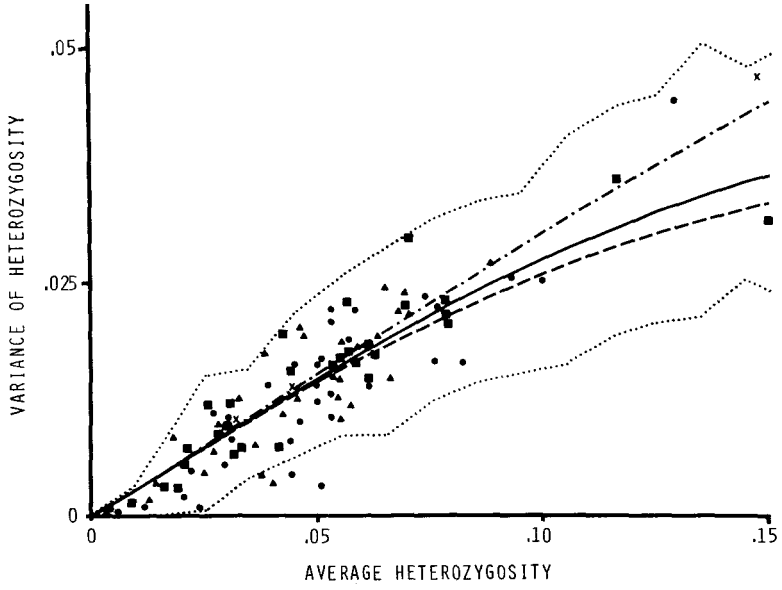

FIGURE 1.—Relationships between average heterozygosities ($\hat{H}$) and the interlocus variances of heterozygosity [$\hat{V}(h)$] for vertebrate species. Ninety-five species and one differentiated subspecies were used. ——— : theoretical relationship for the infinite allele model. — — : theoretical relationship for the stepwise mutation model. —•— : theoretical relationship for the infinite allele model with varying mutation rate (coefficient of variation of mutation rate = 1.0). ····· : 95% significance intervals of the variance obtained by the method described in the text. ■ mammals; ● reptiles; ▲ fishes; × amphibians.

interlocus variance of heterozygosity in a population. Studying the rate of amino acid substitution in 19 proteins during evolution and the molecular weights of 119 proteins in mammals, NEI, CHAKRABORTY and FUERST (1976a) suggested that the mutation rate per locus is distributed roughly as a gamma distribution with the coefficient of variation of about 1. If we accept this suggestion, the relationship between H and $V(h)$ can be computed by using their formulae for the infinite allele model with varying mutation rate. The results obtained are presented for comparison in Figures 1 and 2 (chain-block lines). Clearly, the variance of heterozygosity for a given value of H is larger in this model than in the infinite allele model with constant mutation rate. We have not computed the relationship between H and $V(h)$ for the stepwise mutation model with varying mutation rate, but it is expected to lie between the broken and chain-block lines in Figures 1 and 2. In practice, we do not know the actual magnitude of variation in mutation rate, but would expect that if the mutation-drift hypoth-

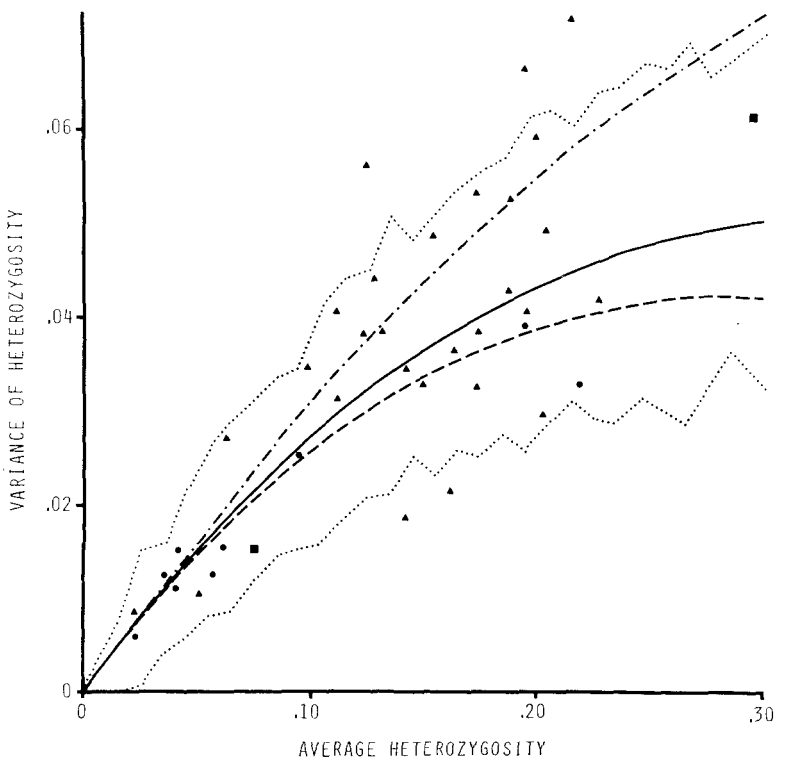

FIGURE 2.—Relationships between average heterozygosities ($\hat{H}$) and the interlocus variances of heterozygosity [$\hat{V}(h)$] for invertebrate species. Thirty-four species and five differentiated subspecies were used. ——— : theoretical relationship for the infinite allele model. — — : theoretical relationship for the stepwise mutation model. — • — : theoretical relationship for the infinite allele model with varying mutation rate (coefficient of variation of mutation rate = 1.0). ····· : 95% significance intervals of the variance obtained by the method described in the text. ▲ Drosophila; ■ non-Drosophila insects; ● non-insect invertebrates.

esis is correct, the observed variance will be scattered around these three curves. As long as the average heterozygosity in a population remains low, as is usually the case with most outbreeding species, the differences among the three theoretical curves are relatively small.

It should be mentioned that the relationships between H and $V(h)$ for all of the three models mentioned above depend on the assumption that the population is in equilibrium with respect to the effects of mutation and random genetic drift. In practice, the population size of a species would vary considerably in the evolutionary process, and thus the values of H and $V(h)$ would not stay constant. However, LI and NEI's (1975) study indicates that with a change of population size, H and $V(h)$ generally change in such a way that the ratio of $V(h)$ to H immediately after the change of population size is not altered drastically compared with that of an equilibrium population having the same value of H. This suggests that our formulae are roughly applicable even in nonequilibrium populations.

As mentioned earlier, we have used only data with a minimum number of 20 loci. However, even with this number, estimates of the variance of heterozygosity are expected to have a large standard error. In order to test the significance of departures of the estimates of variance $[\hat{V}(h)]$ from the expected values, we determined approximate significance limits of $\hat{V}(h)$ for the infinite allele model with constant mutation rate by means of computer simulation. F. M. STEWART (APPENDIX) has developed a method to obtain a random set of allele frequencies for a locus with the infinite allele model. If we use this method, the heterozygosity for a random locus can easily be determined with the aid of a computer. "Populations" consisting of samples of $2n = 100$ genes at twenty loci were generated for each of seven values of *expected* heterozygosity (0.01, 0.05, 0.1, 0.15, 0.2, 0.25, and 0.3) by using this method. The observed mean and variance of heterozygosity were calculated for each of these sample populations. Populations were then classified according to their *observed* mean heterozygosity, and the distribution of $\hat{V}(h)$ was constructed for each class of average heterozygosity spanning an interval of 0.01 from 0.00 to 0.3. From this distribution, 95 and 99% significance limits—the limits outside which $\hat{V}(h)$ is significantly different from the theoretical value—were empirically determined. Seven-hundred fifty replicate populations were generated for each of the seven expected heterozygosities used. The 95% empirical significance limits thus obtained are given in Figures 1 and 2. It is clear that the significance interval for $\hat{V}(h)$ increases roughly linearly as $\hat{V}(h)$ increases. Note that our significance limits are approximate since they depend on the expected heterozygosities used. However, slight changes of the expected heterozygosities in the region of $0.01 \sim 0.3$ do not appear to alter the significance limits drastically, as long as they are widely distributed. It has been noted that our empirical significance limits for $\hat{H} = 0.03 \sim 0.3$ are actually very close to those obtained under the assumption of the normal distribution of $\hat{V}(h)$ for a given value of $\hat{H}$. Evidently, the central limit theorem in statistics applies to this case.

We have not studied the significance limits for the stepwise mutation model.

However, since the theoretical relationship between H and $V(h)$ for this model is similar to that for the infinite allele model, we believe that the magnitudes of significance limits are more or less the same for the two models, as long as there is no variation in mutation rate. When mutation rate varies among loci with the coefficient of variation of one, the variance of heterozygosity is larger than that for the case of constant mutation rate. In the varying mutation model the significance interval for $\hat{V}(h)$ is also expected to be larger than that for the latter case and probably roughly proportional to the value of $\hat{V}(h)$. We have not attempted to determine the significance interval for the varying mutation model, since it requires excessive computer time and our test is approximate anyway.

Results: Data from vertebrates and invertebrates were analyzed separately, since they often differ in average heterozygosity (SELANDER and KAUFMAN 1973; POWELL 1975). Our vertebrate data came from 28 mammalian species and one subspecies (mean of average heterozygosities ($\hat{H}$) = 0.052; range 0.009 ~ 0.152), 33 reptile species (0.048; 0.004 ~ 0.129), 31 fish species (0.044; 0.0005 ~ 0.089) and 3 amphibians (0.074; 0.032 ~ 0.147). (There were five vertebrate and three invertebrate species in which no polymorphism was observed, but they were not included in this study.) The observed relationships between the mean and variance of heterozygosity for vertebrates are shown in Figure 1, together with the theoretical curves for the three models discussed earlier. (For those species in which the distribution of heterozygosity was tested, the values of $\hat{H}$ and $\hat{V}(h)$ are presented in Tables 2 and 3.) The agreement between the theoretical and observed variances is excellent for most of the species presented in Figure 1. In no mammalian species was $\hat{V}(h)$ significantly different from the expected value. Two fish (*Trematomus bernacchii* and *Gibbonsia metzi*; SOMERO and SOULÉ 1974) and two reptiles (*Anolis marmoratus* and *A. sabanus*; GORMAN and KIM 1976) had variances which were significantly smaller than the expected variances under the infinite allele model with constant mutation rate. The variances in *A. marmoratus* and *T. bernacchii* were lower than the 99% significance limits. Despite these few deviant observations, the great majority of data from vertebrates are in an excellent agreement with the predictions of the mutation-drift hypothesis. Such a general agreement is especially meaningful in view of the diverse types of organisms used.

ANTHONY BROWN has asked whether or not the significance interval of $\hat{V}(h)$ for a given value of $\hat{H}$ is small compared with the maximum possible range of $\hat{V}(h)$. Evidently, the smallest possible value of $\hat{V}(h)$ is 0 for any $\hat{H}$, while the maximum possible value for a given value of $\hat{H}$ is given by $\hat{H}(1 - \hat{H})$. The latter value is 0.048 for $\hat{H} = 0.05$, 0.09 for $\hat{H} = 0.1$, and 0.16 for $\hat{H} = 0.2$. Therefore, the maximum possible ranges of $\hat{V}(h)$ are far greater than the significance intervals given in Figures 1 and 2.

For the invertebrates, 34 species and 5 differentiated subspecies were used. Of these, 23 species and the 5 subspecies belonged to the genus Drosophila (overall average of $\hat{H} = 0.150$; range 0.023 ~ 0.216). The remaining invertebrates came from a heterogeneous taxonomic group of eleven species with average heterozygosities ranging from 0.022 ~ 0.294. Figure 2 shows the relationship

between the mean and variance of heterozygosity for all invertebrate species. It is clear that, as with vertebrates, the data generally agree with the expectations from the mutation-drift hypothesis. In five species the observed variances deviate significantly from the expected values of the infinite allele model with constant mutation rate. Two species of Hawaiian Drosophila (*D. adiostola* and *D. nigra*; AYALA 1975) have variances below the 99% significance limits, while two other Hawaiian species (*D. mimica* and *D. engyochracea*; STEINER unpublished) and a Western Pacific Drosophila species (*D. malerkotliana pallens*; YANG, WHEELER and BOCK 1972) deviate in the positive direction. None of the variances of non-Drosophila species showed a significant deviation.

NEI, FUERST and CHAKRABORTY (1976b) noted that as the average heterozygosity increases the variance of heterozygosity tends to become larger than the theoretical expectations of either the infinite allele model or the stepwise mutation model when no interlocus variation in mutation rate is assumed. They took this as evidence for the existence of interlocus variation of mutation rate. In fact, their observed variances agreed with the theoretical expectations obtained under the assumption that mutation rate varies among loci according to the gamma distribution with the coefficient of variation of one. Figures 1 and 2 show that the same tendency also exists in the data analyzed here. In vertebrates this tendency is less clear than in invertebrates, since the majority of average heterozygosity estimates lie in the range where the expected differences among the models are very small. If mutation rate in fact varies with the coefficient of variation of one, the significance limits of $\hat{V}(h)$ should be considered around the chain-block line. In invertebrates this would probably make the deviations of the three points above the upper dotted line statistically insignificant, while two more points would become significantly smaller than 95% lower limits. Note that the significance interval of $\hat{V}(h)$ for a given value of $\hat{H}$ is expected to be larger in this case than in the case of constant mutation rate.

NEI, FUERST and CHAKRABORTY (1976b) also noticed that the variance of heterozygosity tends to be smaller than the expected when the average heterozygosity is close to 0. The same tendency is observed in Figures 1 and 2. It is clear from our empirical significance limits that this is caused by sampling error. In fact, the averages of $\hat{V}(h)$ in our computer simulation for $\hat{H} = 0 \sim 0.01$, $0.01 \sim 0.02$, $0.02 \sim 0.03$, and $0.03 \sim 0.04$ were 0.0004, 0.0034, 0.0077, and 0.0101 compared with the theoretical expectations of 0.0017, 0.0049 0.0079, and 0.0109, respectively. For $\hat{H} > 0.04$, however, the sample mean of $\hat{V}(h)$ agreed well with the theoretical value. Sample means of $\hat{V}(h)$ smaller than the theoretical value are explained by the fact that the distribution of heterozygosity is extremely skewed when H is small (Table 1). If a relatively small number of loci (20 in our simulation) are sampled from this distribution, highly heterozygous loci are often unrepresented in the sample. Therefore, the sample variance is expected to be smaller than the theoretical value.

It should be mentioned that the proteins examined for genetic polymorphism varied considerably from survey to survey. At least 58 different proteins were examined in one or more of the studies included here. Certain proteins such as

esterases, malate dehydrogenase, phosphoglucomutase, and isocitrate dehydrogenase were examined in over 80% of the surveys. Other proteins such as aldehyde oxidase and hydroxybutyrate dehydrogenase were studied only in invertebrates, while hemoglobin, transferrin, albumin, haptoglobin, and others were studied exclusively in vertebrates. These differences in the proteins studied may have introduced some heterogeneity into the observed relationship between the mean and variance of heterozygosity among different organisms. For instance, it was found that the observed variances in the *Drosophila willistoni* group studied by AYALA and his associates tend to be smaller than the expected value, while the data reported by PRAKASH and others for several North American Drosophila species tend to show a variance larger than the expected. Possibly contributing to this difference is the fact that the studies by AYALA's group include about 16% esterase loci and no nonenzymatic proteins, while PRAKASH's investigations include only 6% esterases and 38% nonenzymatic proteins. Nevertheless, the overall agreement between data and theory is very good, so that the effect of differences in proteins used seems to be generally small.

Distribution of single-locus heterozygosity within species

Theoretical distribution: In an equilibrium population, heterozygosity is expected to vary widely among loci owing to random genetic drift, even if the mutation rate is the same for all loci. It is therefore of interest to determine whether the actual distribution agrees with the theoretical distribution of the mutation-drift hypothesis. Theoretically, such a test is more powerful than a test of the variance alone, if a sufficient number of loci are available. Some earlier studies (*e.g.*, NEI and ROYCHOUDHURY 1974b; NEI, FUERST and CHAKRABORTY 1976b) have indicated that the agreement between the theoretical and observed distributions is qualitatively satisfactory in a number of species. In this paper we intend to test the agreement quantitatively.

Since no analytical formula for the theoretical distribution is available, we obtained it by using STEWART's method (APPENDIX). For each of the same seven H values used for deriving the confidence limits of $\hat{V}(h)$, heterozygosities for 50,000 random loci were determined with sample size of 100 genes. Using these sample heterozygosities the theoretical distributions were obtained empirically. They are given in Table 1. For very low average heterozygosities, the frequency distribution is strongly skewed towards zero. Thus, when $H = 0.01$, over 96% of the loci show heterozygosities less than 0.05. The frequency of loci in this class declines as the average heterozygosity increases but still accounts for 20% of the total frequency when H reaches 0.3, the upper bound of H so far observed in natural populations of outbreeding species. For all values of H between $0.01 \sim 0.3$ a second smaller peak exists in the heterozygosity class $0.45 \sim 0.50$. Existence of this peak has been predicted by STEWART (1976) in his study of the theoretical distribution of heterozygosity for the three allele model. Our results show an additional intermediate peak that occurs in the heterozygosity class $0.25 \sim 0.30$. This peak is very small but consistent for all values of average heterozygosity in the range of $H = 0.01 \sim 0.25$. The large number of samples used in our computation makes us confident of the reality of this peak, although it is not dis-

cernible when the average heterozygosity is 0.3. Our finding is in agreement with STEWART's speculation concerning the existence of peaks in the probability density function of heterozygosity.

We have not examined the distribution of single locus heterozygosities with the stepwise mutation model. Some studies on this distribution have been made by CHAKRABORTY (1977) by means of Monte Carlo simulation. His results indicate that the distribution of heterozygosity is similar to that of the infinite allele model when H is equal to or less than 0.3. LI (1976) has shown that when there are nonstepwise changes in electrophoretic mobility, population parameters such as average heterozygosity rapidly approach the values for the infinite allele model as the proportion of nonstepwise changes increases. In the light of the available data we believe that the distributions for the infinite allele model presented here can be used for testing the mutation-drift hypothesis. The effect of interlocus variation in mutation rate has not been explored theoretically, since the data currently available can be explained by the mutation-drift hypothesis without considering the variation in mutation rate, as will be seen below.

Results: The observed frequency distribution of single locus heterozygosity was constructed for all species. When there were several populations surveyed in the same species, each population was treated separately and an average distribution for the species was obtained. To test the goodness-of-fit of the observed distribution to the theoretical, we employed the Kolmogorov-Smirnov test. As mentioned earlier, we generated the theoretical distributions of heterozygosity for

TABLE 1

Theoretical distributions of single locus heterozygosity (h) *for various values of expected heterozygosity* [$H = M/(1 + M)$]

h	H	0.01	0.05	0.10	0.20	0.30
$0 \leq 0.05$		.9621	.8284	.6667	.4063	.2121
$0.05 < h \leq 0.10$		.0081	.0362	.0635	.0866	.0840
$0.10 < h \leq 0.15$		.0047	.0202	.0364	.0548	.0580
$0.15 < h \leq 0.20$		.0040	.0148	.0276	.0502	.0514
$0.20 < h \leq 0.25$		.0029	.0115	.0241	.0397	.0490
$0.25 < h \leq 0.30$		.0034	.0135	.0255	.0433	.0511
$0.30 < h \leq 0.35$		.0020	.0122	.0227	.0404	.0514
$0.35 < h \leq 0.40$		.0039	.0122	.0250	.0447	.0548
$0.40 < h \leq 0.45$		.0029	.0164	.0297	.0489	.0625
$0.45 < h \leq 0.50$		.0053	.0272	.0494	.0786	.0926
$0.50 < h \leq 0.55$		.0004	.0034	.0137	.0415	.0761
$0.55 < h \leq 0.60$		.0001	.0019	.0073	.0258	.0550
$0.60 < h \leq 0.65$		——	.0014	.0054	.0215	.0476
$0.65 < h \leq 0.70$		——	.0006	.0024	.0131	.0342
$0.70 < h \leq 0.75$		——	.0001	.0006	.0039	.0155
$0.75 < h \leq 0.80$		——	——	.0001	.0007	.0041
$0.80 < h \leq 0.85$		——	——	——	——	.0004

These distributions were obtained by evaluating the heterozygosities for 50,000 random loci for each value of H using Stewart's method (Appendix). See the text for details.

seven values of H. The goodness-of-fit of the theory to data was tested only for species whose average heterozygosity was in the range of ± 0.01 about one of these seven values of H. There were 68 such species or subspecies.

The results of the Kolmogorov-Smirnov test of the difference between the theoretical and observed distributions are given in Tables 2 and 3. The number

TABLE 2

Means ($\hat{H}$) *and variances* [$\hat{V}(h)$] *of heterozygosity, and tests of the agreement between the theoretical and observed distributions of heterozygosity in terms of the Kolmogorov-Smirnov statistic* (D)*—vertebrate species*

Species	Populations	Loci	$\hat{H}$	$\hat{V}(h)$	D	Sources
MAMMALS						
Macaca fuscata	15	21	.016	.004	.045	Nozawa *et al.* (1975)
M. cyclopis	1	29	.041	.007	.061	Nozawa *et al.* (unpublished)
Peromyscus polionotus	10	32	.054	.016	.019	Selander *et al.* (1971)
P. floridanus	3	38	.055	.017	.023	Smith *et al.* (1973)
P. boylii	5	21	.021	.006	.048	Avise *et al.* (1974a)
P. eremicus (Eastern race)	7	24	.058	.017	.061	Avise *et al.* (1974b)
P. eremicus (Western race)	4	24	.009	.001	.014	Avise *et al.* (1974b)
Myotis velifer	2	25	.152	.032	.051	Straney *et al.* (1976b)
Rattus rattus	4	37	.044	.016	.050	Patton *et al.* (1975)
Thomomys talpoides	10	31	.056	.017	.026	Nevo *et al.* (1974)
REPTILES						
Anolis luciae	1	22	.093	.026	.097	Yang *et al.* (1974)
A. griseus	1	22	.020	.002	.107	Yang *et al.* (1974)
A. blanquillanus	1	22	.053	.022	.056	Yang *et al.* (1974)
A. roquet	1	22	.058	.022	.101	Yang *et al.* (1974)
A. trinitatis	1	22	.061	.014	.192	Yang *et al.* (1974)
A. bimaculatus	1	22	.053	.021	.047	Gorman and Kim (1976)
A. leachi	1	22	.044	.008	.063	Gorman and Kim (1976)
A. oculatus	1	22	.050	.012	.035	Gorman and Kim (1976)
A. lividus	1	22	.053	.013	.056	Gorman and Kim (1976)
A. marmoratus	1	22	.051	.003	.192	Gorman and Kim (1976)
A. sabanus	1	22	.044	.005	.147	Gorman and Kim (1976)
A. gingivinus	1	22	.100	.026	.030	Gorman and Kim (1976)
A. schwartzi	1	22	.053	.011	.049	Gorman and Kim (1976)
A. wattsi	1	22	.046	.010	.049	Gorman and Kim (1976)
A. carolinensis	4	23	.051	.017	.031	Webster *et al.* (1972)
A. distichus	1	23	.050	.017	.037	Webster *et al.* (1972)
A. sagrei	2	23	.012	.001	.071	Webster *et al.* (1972)
Cnemidophorus tigris	2	21	.050	.014	.051	Parker and Selander (1976)
C. septemvittatus	2	21	.057	.018	.044	Parker and Selander (1976)
C. sexlineatus	2	21	.045	.016	.021	Parker and Selander (1976)
Bipes canaliculatus	1	22	.004	.0002	.030	Kim *et al.* (1976)
AMPHIBIANS						
Bufo viridis	10	26	.147	.048	.125	Nevo *et al.* (1975)
Plethodon cinereus	14	24	.044	.014	.032	Highton and Webster (1976)

TABLE 2—Continued

Species	Populations	Loci	$\hat{H}$	$\hat{V}(h)$	D	Sources
FISH						
Menidia menidia	5	24	.053	.015	.010	JOHNSON (1975)
M. peninsulae	5	24	.055	.019	.029	JOHNSON (1975)
M. beryllina	8	24	.045	.013	.024	JOHNSON (1975)
Hesperoleucus symmetricus	1	24	.054	.015	.066	AVISE and AYALA (1976)
Mylopharodon conocephalus	1	24	.002	.0001	.030	AVISE and AYALA (1976)
Ptychocheilus grandis	1	24	.012	.013	.045	AVISE and AYALA (1976)
Orthodon microlepidotus	1	24	.015	.004	.027	AVISE and AYALA (1976)
Gila bicolor	1	24	.058	.018	.073	AVISE and AYALA (1976)
Trematomus bernacchii	1	26	.040	.004	.098	SOMERO and SOULÉ (1974)
Gillichthys mirabilis	1	29	.054	.013	.049	SOMERO and SOULÉ (1974)
Leuresthes tenuis	1	33	.042	.011	.033	SOMERO and SOULÉ (1974)
Abudefduf troschelii	1	20	.057	.012	.179	SOMERO and SOULÉ (1974)
Bathygobius ramosus	1	26	.055	.010	.061	GORMAN *et al.* (1976)
Sebastes caurinus	1	25	.018	.008	.031	JOHNSON *et al* (1973)
Lepidopsetta bilineata	1	23	.046	.020	.048	JOHNSON and UTTER (1976)
Platichthys stellatus	1	21	.047	.020	.044	JOHNSON and UTTER (1976)
Cymatogaster aggregata	1	23	.0005	.0001	.038	JOHNSON and UTTER (1976)

Additional species shown in Figure 5: *Peromyscus californicus, P. sejugis* (AVISE *et al.* 1974a), *P. pectoralis* (AVISE *et al.* 1974b), *Mus musculus* (SELANDER *et al.* 1969), *Thomomys bottae, T. umbrinus* (PATTON *et al.* 1972), *Geomys bursarius, G. personatus* (KIM 1972), *Geomys arenarius* (SELANDER *et al.* 1974), *Eutamias panamintinus* (KAUFMAN *et al.* 1973), *Dipodomys heermanni* (PATTON *et al.* 1976), *Halichoerus grypus, Phoca vitulina* (MCDERMID and BONNER 1975), *Macaca mulatta, M. irus* (NOZAWA, unpublished), *Macrotus waterhousii, M. californicus* (GREENBAUM and BAKER 1976), *Myotis californicus, Pipistrellus hesperus* (STRANEY *et al.* 1976a), *Anolis richardi, A. bonairensis, A. extremus, A. aeneus* (YANG *et al.* 1974), *Anolis ferreus, A. nubilus, A. pogus* (GORMAN and KIM 1976), *Anolis grahami* (TAYLOR and GORMAN 1975), *Anolis cristatellus* (GORMAN, unpublished), *Bipes biporus* (KIM *et al.* 1976), *Sceloporus grammicus* (two chromosomal species) (HALL and SELANDER 1973), *Plethodon serratus* (HIGHTON and WEBSTER 1976), *Menidia audens, M. extensa* (JOHNSON 1975), *Cichlasoma cyanoguttatum* (SAGE and SELANDER 1975), *Zoarces viviparus* (FRYDENBERG and SIMONSEN 1973), *Lavinia exilicauda, Pogonichthys macrolepidotus, Notemigonus crysoleucus* (AVISE and AYALA 1976), *Trematomus hansoni, Gibbonsia metzi, Mugil cephalus* (SOMERO and SOULÉ 1974), *Sebastes alutus, S. elongatus* (JOHNSON *et al.* 1973), *Bathygobius andrei, B. soporator* (GORMAN *et al.* 1976).

Species with no electrophoretic variability: *Dipodomys californicus* (PATTON *et al.* 1976), *Geomys tropicalis* (SELANDER *et al.* 1974), *Mirounga angustirostris* (BONNELL and SELANDER 1974), *Peromyscus dickeyi* (AVISE *et al.* 1974a), *Anolis angusticeps* (WEBSTER *et al.* 1972).

of loci used for a species ranges from 20 to 42. The value of the Kolmogorov-Smirnov D statistic at the 5% significance level is 0.29 for 20 loci and 0.21 for 42. All the D values in Tables 2 and 3 are smaller than these values. Therefore, in none of the 68 species examined is the observed distribution significantly different from the expected. For some species the observed distribution was computed by pooling data from a number of populations, so that the number of observations (heterozygosities) was larger than the number of loci examined. For example, in *Peromyscus polionotus* the total number of observations (heterozygosities) was 320. All of these observations are not independent of each other, because of a high correlation in the heterozygosity value among related populations. However, a liberal test can be done by assuming independence of the observations. Even such tests indicated that the difference between the theo-

TABLE 3

Means ($\hat{H}$) *and variances* [$\hat{V}(h)$] *of heterozygosity, and tests of the agreement between the theoretical and observed distributions of heterozygosity in terms of the Kolmogorov-Smirnov statistic* (D)*—invertebrate species*

Species	Populations	Loci	$\hat{H}$	$\hat{V}(h)$	D	Sources
Drosophila melanogaster	1	23	.147	.033	.090	BAND (1975)
D. tropicalis	4	29	.142	.035	.052	AYALA *et al.* (1974b)
D. paulistorum (Amazonian)	2	31	.195	.041	.035	AYALA *et al.* (1974b)
D. paulistorum (Orinocan)	2	31	.201	.050	.054	AYALA *et al.* (1974b)
D. adiostola	1	31	.142	.019	.186	AYALA (1975)
D. nigra	1	31	.160	.022	.170	AYALA (1975)
D. crassifemur	1	31	.204	.029	.124	AYALA (1975)
D. malerkotliana pallens	1	23	.194	.067	.115	YANG *et al.* (1972)
D. malerkotliana malerkotliana	1	23	.154	.049	.127	YANG *et al.* (1972)
D. bipectinata	1	23	.199	.060	.151	YANG *et al.* (1972)
D. pseudoobscura bogotana	1	24	.051	.011	.078	PRAKASH *et al.* (1969)
D. persimilis	1	24	.092	.036	.125	PRAKASH (1969)
Otiorrhynchus scaber	1	23	.294	.062	.047	SUOMALAINEN and SAURA (1973)
Homarus americanus	4	42	.040	.011	.044	TRACEY *et al.* (1975)
Phoronopsis viridis	3	38	.094	.026	.037	AYALA *et al.* (1974c)
Nearchaster aciculosus	1	24	.195	.039	.118	AYALA *et al.* (1975b)
Asterias forbesi	1	27	.041	.016	.061	SCHOPF and MURPHY (1973)
Euphausia superba	1	36	.057	.013	.078	AYALA *et al.* (1975c)

Additional species shown in Figure 3 are: *Drosophila willistoni, D. equinoxialis, D. nebulosa* (AYALA *et al.* 1974b), *D. planitibia* (AYALA 1975), *Drosophila equinoxialis caribbensis* (AYALA *et al.* 1974a), *Drosophila willistoni quechua* (AYALA and TRACEY 1973), *D. parabipectinata, D. pseudoananassae* (YANG *et al.* 1972), *D. obscura* (LAKOVAARA and SAURA 1971), *D. subobscura* (ZOUROS *et al.* 1974), *D. pseudoobscura* (PRAKASH *et al.* 1969), *D. robusta* (PRAKASH 1973a), *D. busckii* (PRAKASH 1973b), *D. buzzatii* (BARKER and MULLEY 1976), *D. mimica, D. engyochracea* (STEINER unpublished), *Philaenus spumarius* (SAURA *et al.* 1973), *Tridacna maxima* (AYALA *et al.* 1973), *Liothyrella notorcadensis* (AYALA *et al.* 1975a), *Asterias vulgaris* (SCHOPF and MURPHY 1973), *Limulus polyphemus* (SELANDER *et al.* 1970).

Species with no electrophoretic variability: *Lasioglosum zephyrum, Augochlora pura,* and *Bombus americanorum* (SNYDER 1974).

retical and observed distributions of heterozygosity is not statistically significant in any case.

These results are somewhat inconsistent with the conclusion obtained from the study of variance of heterozygosity, since in the latter study the variance of heterozygosity was significantly different from the expected in 4 out of the 96 vertebrate species or subspecies and 5 out of the 39 invertebrate species or subspecies examined. This discrepancy is most likely due to the fact that the Kolmogorov-Smirnov test is not generally very powerful, although it is known to be more powerful than alternative parametric tests when the number of observations is small. It is noted, however, that the D statistic for the deviant species tends to be larger than for the other species listed in Tables 2 and 3. (The Kolmogorov-Smirnov test was not conducted for *G. metzi, D. mimica,* and *D. engyochracea.*)

Let us now examine the distributions of heterozygosity for the deviant species,

excluding *G. metzi*, *D. mimica*, and *D. engyochracea* for which the theoretical distributions are not available. The observed distributions are given in Figures 3 to 5 together with the theoretical distributions. The three vertebrate species (*A. marmoratus*, *A. sabanus*, and *T. bernacchii*), with average heterozygosities between 0.04 and 0.05, show virtually the same pattern of distribution (Figure 3). Compared with the theoretical distribution, the observed distributions have a lower frequency for the heterozygosity class 0, and loci showing a heterozygosity near 0.5 are virtually absent. Compensating for these deficiencies are increased frequencies of mildly heterozygous classes of loci. A similar pattern is observed with the two species of Drosophila (*D. adiostola* and *D. nigra*, average heterozygosity between 0.14 and 0.16), in which the observed variance was smaller than the expected (Figure 4). In this case, however, the deficiency of the class of heterozygosity near 0.5 is not so great. In contrast, in *Drosophila malerkotliana pallens*, where the observed variance was significantly larger than the expected, the frequencies of heterozygosity classes 0 and above 0.5 are slightly higher than expected (Figure 5). In this species, however, gene frequencies were computed by pooling four island populations in Southeast Asia (Luzon, Mindanao, Palawan, and Borneo; YANG, WHEELER and BOCK 1972), so that the large variance for this species may be due to genetic heterogeneity of the populations studied.

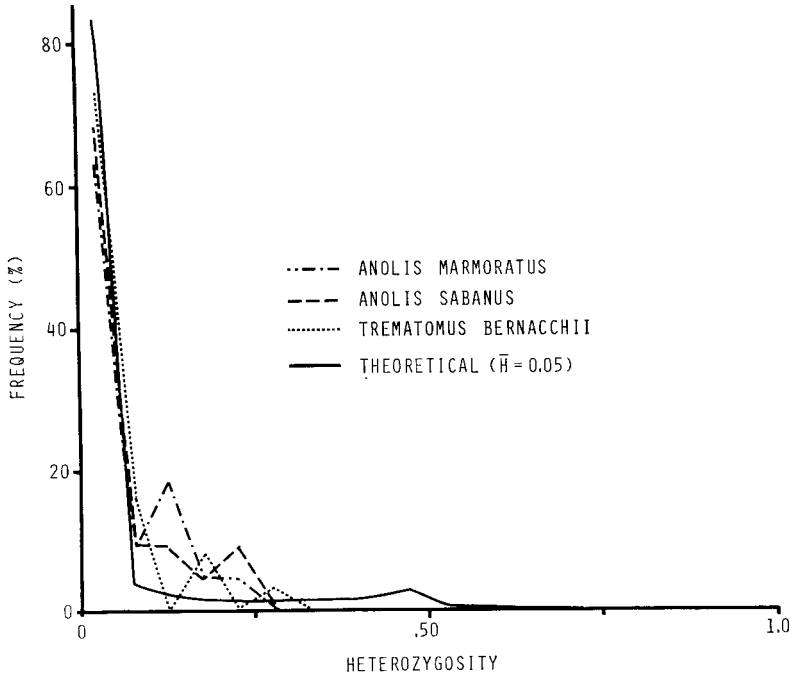

FIGURE 3.—Frequency distributions of single locus heterozygosity for three vertebrate species in which the observed interlocus variance of heterozygosity was significantly lower than the theoretical value. The theoretical distribution corresponds to that given in Table 1 for $\dot{H} = 0.05$.

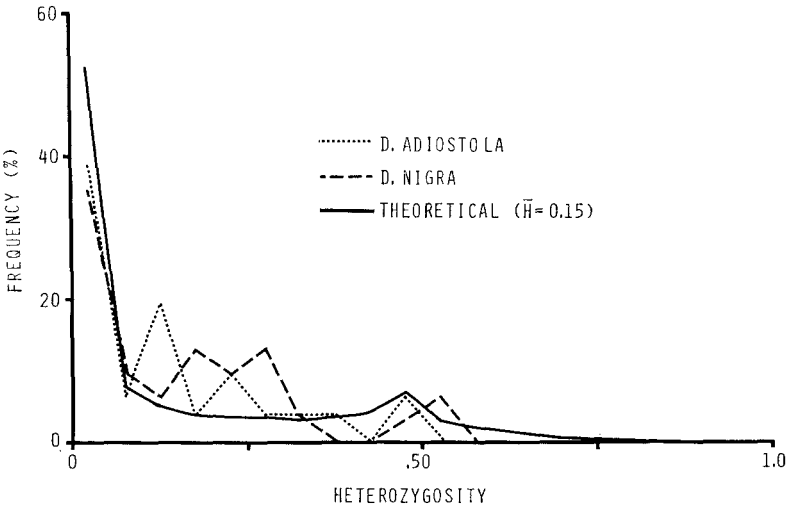

FIGURE 4.—Frequency distributions of single locus heterozygosity for two Hawaiian Drosophila species in which the observed interlocus variance of heterozygosity was significantly lower than the theoretical value. The theoretical distribution was obtained as described in the text for $\hat{H} = 0.15$.

As mentioned earlier, our analyses involved comparisons of data with theoretical expectations for three equilibrium models. Nonequilibrium situations may lead to some changes in the relationship between H and $V(h)$, although we expect that the changes are relatively minor. It is interesting that all of the seven terrestrial species in which the observed relationship deviated significantly

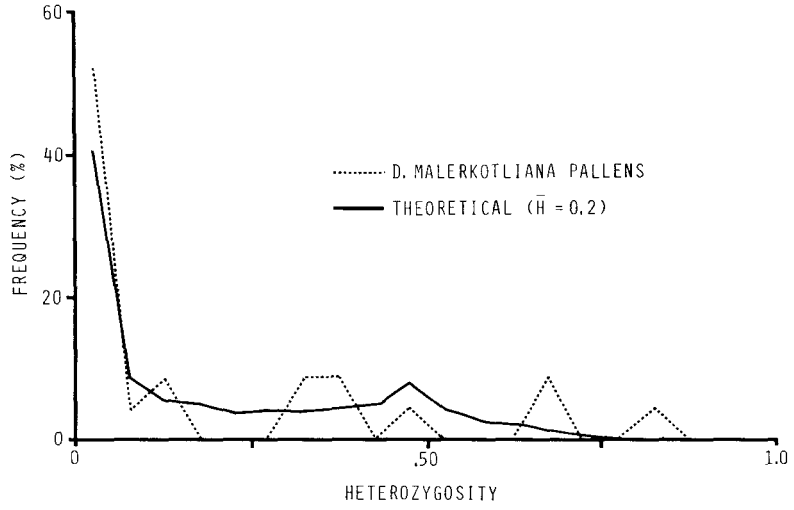

FIGURE 5.—Frequency distribution of single locus heterozygosity for *Drosophila malerkotliana pallens* in which the observed interlocus variance of heterozygosity was significantly greater than the theoretical value. The theoretical distribution corresponds to that given in Table 1 for $\hat{H} = 0.20$.

from the theoretical values exist on islands. It is possible that the observed deviations are in part due to nonequilibrium conditions.

At any rate, there are altogether 9 cases in which the observed variance of heterozygosity was significantly different from the expected value under the infinite allele model with constant mutation rate at the 5% level. Since the total number of species or subspecies examined is 135, the proportion of deviant cases is only slightly higher than that expected by chance alone. Therefore, we may conclude that the distribution of heterozygosity generally agrees with the theoretical distribution expected under the mutation-drift hypothesis.

Variance of heterozygosity among species

If the variation of single locus heterozygosity is caused mainly by random genetic drift, we would expect that the heterozygosity at a given locus will vary from species to species if there is no historical correlation in heterozygosity among species. At a locus the mutation rate should be more or less the same for related species, but the effective population size may vary with species. Therefore, if the mutation-drift hypothesis is correct, we would expect that the variance of heterozygosity among species is equal to or larger than the theoretical value given by formula (2).

It should be noted, however, that if the time after divergence of species is short, the heterozygosities in related species will be correlated, and consequently the variance among species will be reduced. If the effective population size (N_e) is the same for two species, the rate of decay of the correlation of heterozygosity per generation is $2v + 1/(2N_e)$, unless the initial heterozygosity is equal to the equilibrium value (Li and Nei 1975). In the latter case it is $4v + 1/(N_e)$. In Drosophila species with the average heterozygosity of about 0.16, the mutation-drift hypothesis predicts that $v = 10^{-8}$ and $N_e = 4 \times 10^6$ approximately for electrophoretic alleles (Nei and Li 1975). Therefore, it might take about $4N_e = 1.6 \times 10^7$ generations or 1.6×10^6 years for the correlation to disappear in Drosophila. In practice, we do not know divergence time for most of the species in which the heterozygosities for homologous loci were studied. It is likely, however, that the correlation has not disappeared completely at least for a sizable number of species used in our study.

Keeping in mind this difficulty, we have computed the interspecific variance of heterozygosity for homologous loci in order to see whether the data are consistent with the mutation-drift hypothesis. We have gathered those data for which homologous genetic loci were studied for 7 or more species.

For the vertebrates, data were available for a total of 125 loci from various groups of mammals, fish, or reptiles. In mammals we were able to study 13 loci in 8 species of *Peromyscus* (Avise *et al.* 1974b), 14 loci in 10 species of *Dipodomys* (Johnson and Selander 1971) and 17 loci in 7 species of *Macaca* (K. Nozawa, unpublished). In fish 17 loci were studied in 9 species in the cyprinid subfamily *Leuciscinae* (Avise and Ayala 1976), 10 loci in 10 species of *Lepomis* (Avise and Smith 1974) and 5 loci in 8 species of salmon and trout (Utter, Allendorf and Hodgins 1973). Reptiles are represented solely by Anolis liz-

ards. This data includes 16 loci in 9 species of the *roquet* species group (YANG, SOULÉ and GORMAN 1974), 18 loci in 13 species of the *bimaculatus* species group (GORMAN and KIM 1976) and 15 loci in 15 species of the *cristatellus* group (G. C. GORMAN, unpublished). In the invertebrates information on 47 loci, all from Drosophila, was available for study. These included 26 loci for 12 species of the *Drosophila willistoni* group and the Hawaiian Drosophila (AYALA *et al.* 1974a; AYALA 1975), 13 loci for 14 species of the *Drosophila repleta* group (R. H. RICHARDSON, unpublished), 3 loci for 12 species of Drosophilids (SABATH 1975), and 5 loci for 28 Hawaiian Drosophila (ROCKWOOD *et al.* 1971). In all of these studies the homology of the loci were inferred by the authors. Some species are included despite small sample size (16 genes) at a locus in order to increase the number of species examined for a locus.

In vertebrates the observed variance agrees with the expected value in a majority of the loci examined (Figure 6). Thirty loci with average heterozygosity below 0.02 are not shown in Figure 6. All of these loci had variances below the theoretical curve for the infinite allele model. Note that in this analysis the effect of sampling error is larger than that in Figures 1 and 2, since the variance was computed from a smaller number of observations (species). When the average heterozygosity is small, the sampling error occurs more often in the downward direction than in the upward direction, as noted earlier. For the Anolis

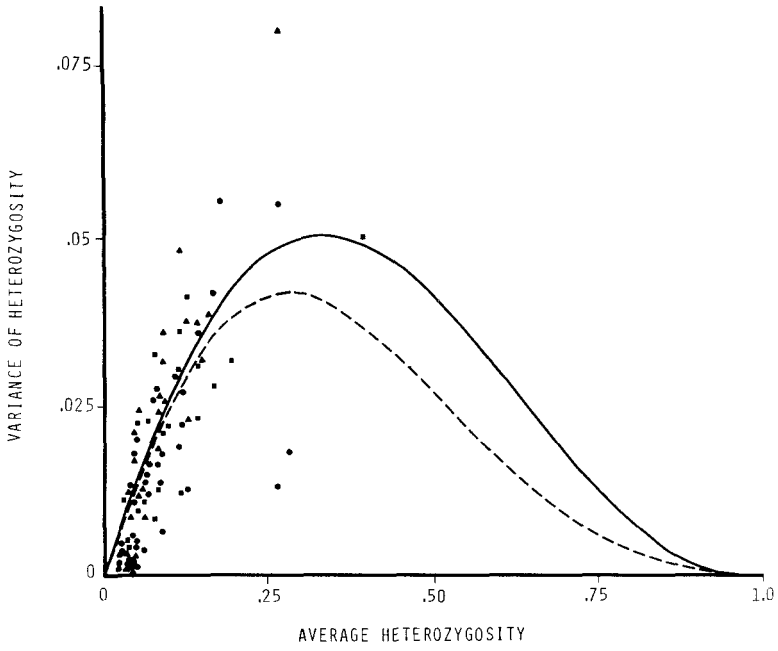

FIGURE 6.—Relationships between the means and variances of heterozygosity for homologous loci among different species of vertebrates.

——— : theoretical relationship for the infinite allele model.

— — : theoretical relationship for the stepwise mutation model.

■ mammals; ● lizards; ▲ fishes.

lizards there is a general tendency for the variance to be lower than the expected. This lowering of observed variance could be due to some sort of selective force, but it can also be explained by the historical correlation of heterozygosity, since the genetic distances among the species used are not always large (GORMAN and KIM 1976).

The Drosophila data used for this test can be divided into three groups, *i.e.*, AYALA's data, RICHARDSON's data, and the other miscellaneous data. They are represented by different symbols in Figure 7. It is clear that RICHARDSON's data and the miscellaneous data agree with the predictions from the mutation-drift hypothesis. There is some tendency for the variance to be higher than the expected. However, if the effective population size varies from species to species, this is exactly what is expected. AYALA's data deviate considerably from the other data. In 25 out of the 26 loci examined the variance is smaller than that expected under the stepwise mutation model. It is not likely that the main cause for this low variance is the historical correlation of heterozygosity, since the genetic distance among the species used is not small (AYALA *et al.* 1974b; AYALA 1975). We are thus led to conclude that AYALA's data are not consistent with the mutation-drift hypothesis. The possible factors responsible for this low variance will be discussed later.

Average heterozygosity and proportion of polymorphic loci

We shall finally examine the relationship between average heterozygosity and the proportion of polymorphic loci. A locus is called polymorphic if the fre-

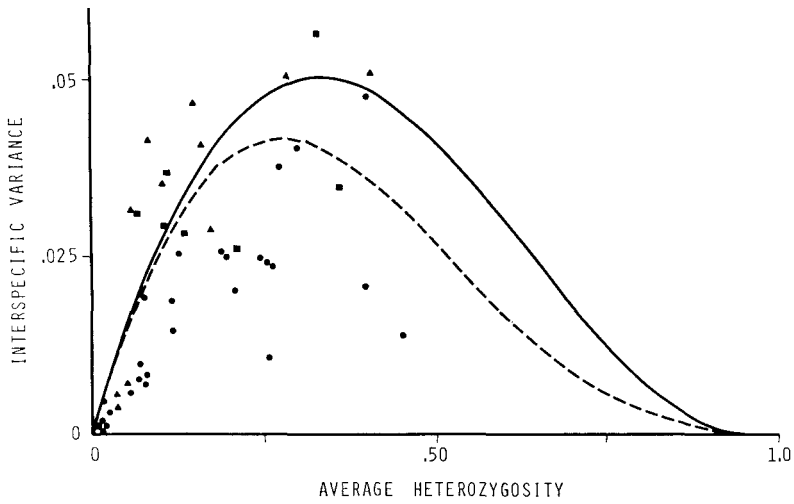

FIGURE 7.—Relationships between the means and variances of heterozygosity for homologous loci among different species of Drosophila.

——: theoretical relationship for the infinite allele model.

— —: theoretical relationship for the stepwise mutation model. ● data from the *Drosophila willistoni* group and Hawaiian Drosophila compiled by F. J. AYALA; ▲ data from the *Drosophila repleta* group compiled by R. H. RICHARDSON; ■ data from miscellaneous Drosophilid groups as described in the text.

quency of the most common allele is equal to or less than $1 - q$, where q is a small quantity. The most commonly used value of q is 0.01. For the infinite allele model, KIMURA (1971) showed that the proportion of polymorphic loci (P) is related to the average heterozygosity (H) by

$$P = 1 - q^{H/(1-H)} . \quad (3)$$

Analogous formulae have been obtained for the stepwise mutation model (KIMURA and OHTA 1975) and the varying mutation model (NEI, CHAKRABORTY and FUERST 1976a). The theoretical relationships for the three models for $q = 0.05$ and 0.01 are presented in Figures 8 and 9, respectively.

The empirical relationship between H and P has been previously examined by KIMURA and OHTA (1971) and SELANDER (1976) for limited numbers of species. Their results show that the empirical relationship is not far from the theoretical one.

By using the criterion of $q = 0.05$, we examined the relationship between H and P for all the species and subspecies used in our earlier analyses. The results obtained are presented in Figure 8. The 95% significance limits given in this

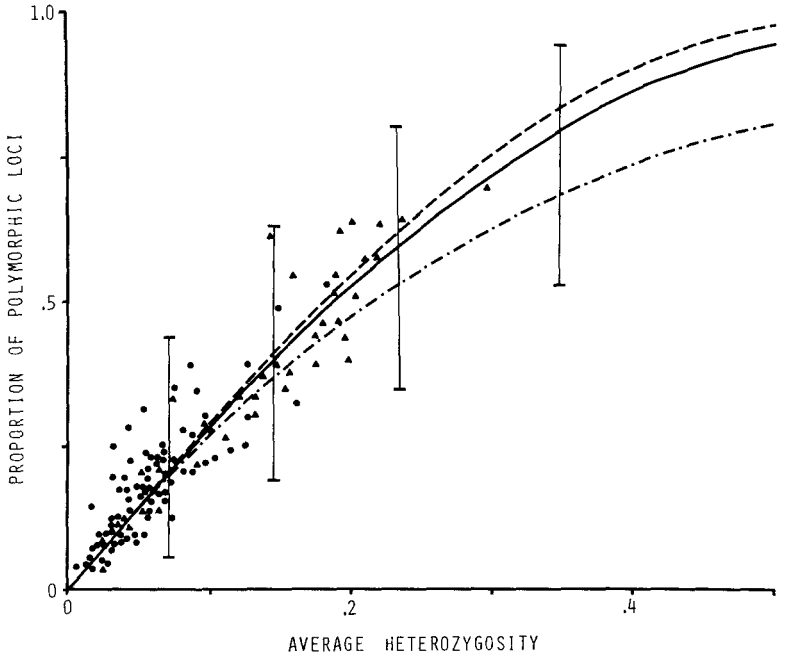

FIGURE 8.—Relationships between average heterozygosities and the proportions of polymorphic loci when the criterion of $q = 0.05$ was used.

——— : theoretical relationship for the infinite allele model.

— — : theoretical relationship for the infinite allele model with varying mutation rate (coefficient of variation of mutation rate = 1.0). The 95% significance intervals for the proportion of polymorphic loci (0.2, 0.4, 0.6, 0.8) for the infinite allele model were obtained by using the binomial sampling theory with sample size of 20 loci. ● vertebrate species; ▲ invertebrate species.

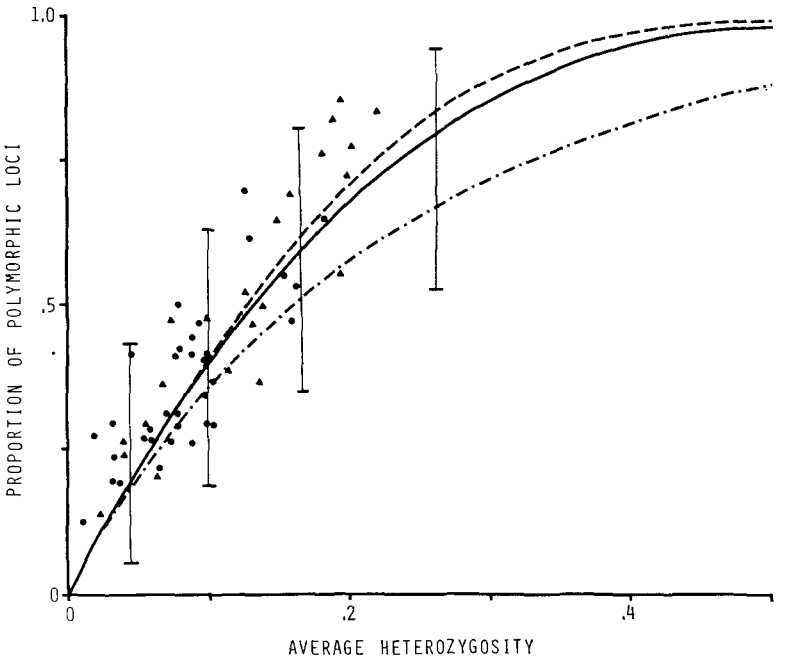

FIGURE 9.—Relationships between average heterozygosities and the proportions of polymorphic loci when the criterion of $q = 0.01$ was used.

——: theoretical relationship for the infinite allele model.

— —: theoretical relationship for the stepwise mutation model.

— • —: theoretical relationship for the infinite allele model with varying mutation rate (coefficient of variation of mutation rate = 1.0). The 95% significance intervals for the proportion of polymorphic loci (0.2, 0.4, 0.6, 0.8) for the infinite allele model were obtained by using the binomial sampling theory with sample size of 20 loci. ● vertebrate species; ▲ invertebrate species.

figure were obtained by the theory of binomial sampling of sample size 20 (20 loci). It is clear that the agreement between the theoretical and empirical relationships is excellent. It should be noted that the results for 33 vertebrate species with average heterozygosities between 0.03 and 0.08 are not shown in Figure 8 because of the high density of points surrounding the curves in this region. However, all of the species excluded had a $H - P$ relationship close to the theoretical one.

The relationship between H and P was also examined by using the criterion of $q = 0.01$. In this case only 34 vertebrate and 22 invertebrate species were used, since this criterion requires a sample size of at least 100 genomes. The results obtained are again consistent with the mutation-drift hypothesis, as will be seen from Figure 9.

DISCUSSION

We have seen that the interlocus variation of heterozygosity for protein loci agrees well with the value expected under the mutation-drift hypothesis. As men-

tioned earlier, however, agreement alone is not proof of this hypothesis. Let us now consider whether our observation can be explained by some other hypotheses or not. We shall discuss only those hypotheses which are directly related to the results obtained in the present paper.

The first hypothesis that may be invoked to explain our results is that each locus has an optimum level of heterozygosity determined by its metabolic function and the substrate availability for the enzyme encoded (KOJIMA, GILLESPIE and TOBARI 1970; GILLESPIE and LANGLEY 1974). In this hypothesis heterozygosity varies from locus to locus, but it is in principle independent of population size and mutation rate. In this hypothesis, the agreement of the observed variance with the expected variance of the mutation-drift hypothesis must be accidental. In practice, however, the agreement occurred in most of the 135 species and subspecies examined. Since the probability of the agreement occurring in so many species by chance alone is extremely small, this hypothesis is not adequate for explaining our observation. Furthermore, this hypothesis has another difficulty. Namely, according to this hypothesis, the heterozygosity for a given locus should not vary extensively among related species. In contrast, however, we have seen that the interspecific variance for a given average heterozygosity is almost as large as the interlocus variance, except in AYALA's Drosophila data.

The second possible explanation is OHTA's (1974) (see also LATTER 1975) hypothesis of slightly deleterious genes. In this hypothesis the level of heterozygosity for a locus is determined by the population size and the rate of mutations to allelic states at which gene function is only slightly impaired. In large populations such as in Drosophila, the average heterozygosity per locus is supposed to be determined mainly by the mutation-selection balance, and the interlocus variation of heterozygosity is a reflection of the differences in the number of subnormal allelic states among loci. In large populations, therefore, the variance of heterozygosity would not agree with that of the mutation-drift hypothesis except by chance. In practice, however, the agreement between the theoretical and observed variances is good in many "large population" species (with average heterozygosities of $0.1 \sim 0.3$), so that it is difficult to explain it by chance alone.

Of course, in a relatively small population with an $N_e s$ of about 1 or less, where s is the selection coefficient against the mutant genes, the relationship between the mean and variance of heterozygosity would not differ appreciably from that of the mutation-drift hypothesis. Using a model of centripetal selection, which is similar to OHTA's hypothesis of slightly deleterious mutations, LATTER (1972) studied the distribution of heterozygosity in finite populations by means of Monte Carlo simulation. His results indicate that slight selection against the mutant genes affects the distribution of heterozygosity for a given level of average heterozygosity only to a small extent in a relatively small population. It is important, however, to realize that the large variation of heterozygosity observed in this simulation has been generated mainly by random genetic drift. When $N_e s$ is relatively small, the population dynamics of slightly deleterious genes is essentially the same as that of neutral genes.

Nevertheless, our observation that the interspecific variance of heterozygosity for a given protein is often smaller than the value expected under the mutation-drift hypothesis may be an indication of the prevalence of slightly deleterious genes at least in some species. This is because in OHTA's hypothesis we would expect that as long as there is mutation-selection balance, the heterozygosity is more or less the same for all species. This explanation, however, is not flawless, since the alleles or electromorphs (KING and OHTA 1975) and their frequencies at a locus generally vary from species to species. Gene substitution seems to have occurred almost continuously in the evolutionary process. This problem will be discussed in detail in the following paper.

Recently, using a deterministic model of stepwise mutation with genic selection, MORAN (1976) has shown that, for a special set of mutation rates and selection coefficients, a locus under the mutation-selection balance can have the same heterozygosity value as the expected heterozygosity under the mutation-drift balance. In his model, however, the heterozygosity is uniquely determined and therefore has no variance. Furthermore, gene substitution is not expected to occur unless the pattern of selection changes. These theoretical predictions are quite contrary to our findings in this and following papers.

It is clear that the overdominance hypothesis has a difficulty similar to that of OHTA's hypothesis in explaining our data. (Of course, when $N(s_1 + s_2)$ is about 1 or less, in which s_1 and s_2 are the selection coefficients for homozygotes, it would again be difficult to distinguish between neutral and overdominant genes by our method; see WATTERSON 1977.) In addition to this, the average heterozygosity per locus seems to be too low in many species for the overdominance hypothesis to be of general importance (NEI 1975, p. 167). Similar comments apply to other types of balancing selection.

One might wonder if our results are explainable by a mixture of various types of deterministic selections. The answer is "probably yes," though it would require an elaborate combination of different types of genes. In the absence of the theoretical variance of heterozygosity for other types of genes, however, it is difficult to study this problem quantitatively. Nevertheless, it seems clear that whatever the selective mechanism is, the data cannot be explained without considering the effect of random genetic drift.

While the majority of the data used here were consistent with the predictions from the mutation-drift hypothesis, AYALA's data on the interspecific variance of heterozygosity for specific loci showed a significant deviation. This poses a problem for the mutation-drift hypothesis. The only way to accommodate this observation with the mutation-drift hypothesis seems to be to assume that the current identification of homologous loci and homologous electromorphs among different species are not perfect (BLAKE and OMOTO 1975) and that this imperfect identification introduces a distortion in the relationship between the mean and variance of heterozygosity. The degree of distortion may be protein-dependent, but we note that RICHARDSON's data, which are consistent with the mutation-drift hypothesis, were obtained from the same set of proteins studied by AYALA. At any rate, it is not easy to explain the deviation in this set of data at the present time. Appar-

ently, more data must be collected to study this problem before any definite conclusion is drawn.

As mentioned in the introductory part, a number of papers have been published about the statistical analyses of protein polymorphisms. In most of these studies data were taken from a small group of organisms. Our study attempts to use the largest set of data possible in order to find general statistical properties. In this paper we have not discussed all aspects of the data related to the mutation-drift hypothesis. However, concerning the results from statistical studies on heterozygosity within species, the conclusion is clear: The majority of protein polymorphisms can be explained by the mutation-drift hypothesis. We are aware that with respect to some other aspects there are a number of observations which are hard to explain by this hypothesis. A general discussion on the maintenance of protein polymorphism will be made in a later paper.

We wish to extend our grateful appreciation to the many investigators who provided us with unpublished gene frequency data. Special thanks go to Drs. R. H. Richardson, G. C. Gorman and K. Nozawa for their unpublished data which enabled us to examine the interspecific variance of heterozygosity. Other investigators who generously provided gene frequency data supplementing previously published reports include Drs. J. C. Avise, M. C. Baker, A. G. Johnson, E. D. Parker, M. H. Smith, G. N. Somero, and W. W. M. Steiner. Without their kind cooperation the present work would not have been completed. We are also grateful for Dr. Wen-Hsiung Li for his comments on the first draft.

LITERATURE CITED

Avise, J. C. and F. J. Ayala, 1976 Genetic differentiation in speciose versus depauperate phylads: evidence from the California minnows. Evolution **30**: 46–58.

Avise, J. C. and M. H. Smith, 1974 Biochemical genetics of sunfish. II. Genic similarity between hybridizing species. Am. Naturalist **108**: 458–472.

Avise, J. C., M. H. Smith and R. K. Selander, 1974a Biochemical polymorphism and systematics in the genus *Peromyscus*. VI. The *boylii* species group. J. Mammalogy **55**: 751–763.

Avise, J. C., M. H. Smith, R. K. Selander, T. E. Lawlor and P. R. Ramsey, 1974b Biochemical polymorphism and systematics in the genus *Peromyscus*. V. Insular and mainland species of the subgenus *Haplomylomys*. Systemat. Zool. **23**: 226–238.

Ayala, F. J., 1975 Genetic differentiation during the speciation process. Evol. Biol. **8**: 1–78.

Ayala, F. J. and M. L. Tracey, 1973 Enzyme variability in the *Drosophila willistoni* group. VIII. Genetic differentiation and reproductive isolation between two subspecies. J. Heredity **64**: 120–124.

Ayala, F. J., D. Hedgecock, G. S. Zumwalt and J. W. Valentine, 1973 Genetic variation in *Tridacna maxima*, an ecological analog of some unsuccessful evolutionary lineages. Evolution **27**: 177–191.

Ayala, F. J., M. L. Tracey, L. G. Barr and J. G. Ehrenfeld, 1974a Genetic and reproductive differentiation of the subspecies, *Drosophila equinoxialis caribbensis*. Evolution **28**: 24–41.

Ayala, F. J., M. L. Tracey, L. G. Barr, J. F. McDonald and S. Pérez-Salas, 1974b Genetic variation in natural populations of five Drosophila species and the hypothesis of the selective neutrality of protein polymorphisms. Genetics **77**: 343–384.

Ayala, F. J., J. W. Valentine, L. G. Barr and G. S. Zumwalt, 1974c Genetic variability in a temperate intertidal phoronid, *Phoronopsis viridis*. Biochem. Genet. **11**: 413–427.

AYALA, F. J., J. W. VALENTINE, T. E. DELACA and G. S. ZUMWALT, 1975a Genetic variability of the Antarctic brachiopod *Liothyrella notorcadensis* and its bearing on mass extinction hypotheses. J. Paleon. **44:** 1–9.

AYALA, F. J., J. W. VALENTINE, D. HEDGECOCK and L. G. BARR, 1975b Deep-sea asteroids: high genetic variability in a stable environment. Evolution **29:** 203–212.

AYALA, F. J., J. W. VALENTINE and G. S. ZUMWALT, 1975c An electrophoretic study of the Antarctic zooplankter *Euphausia superba*. Limnol. and Ocean. **20:** 635–640.

BAND, H. T., 1975 A survey of isozyme polymorphism in a *Drosophila melanogaster* natural population. Genetics **80:** 761–771.

BARKER, J. S. F. and J. G. MULLEY, 1976 Isozyme variation in natural populations of *Drosophila buzzatii*. Evolution **30:** 213–233.

BLAKE, N. M. and K. OMOTO, 1975 Phosphoglucomutase types in the Asian-Pacific area: a critical review including new phenotypes. Ann. Hum. Genet. **38:** 251–273.

BONNELL, M. L. and R. K. SELANDER, 1974 Elephant seals: genetic variation and near extinction. Science **184:** 908–909.

CHAKRABORTY, R., 1977 Simulation results with stepwise mutation model and their interpretations. J. Molec. Evol. (In press).

CROW, J. F., 1968 The cost of evolution and genetic load. pp. 165–178. In: *Haldane and Modern Biology*. Edited by K. R. DRONAMRAJU. Johns Hopkins Press, Baltimore, Maryland. ——, 1972 Darwinian and non-Darwinian evolution. Proc. 6th Berkeley Symp. Math. Statist. and Probab. **V:** 1–22.

FRYDENBERG, O. and V. SIMONSEN, 1973 Genetics of *Zoarces* populations. V. Amount of protein polymorphism and degree of genic heterozygosity. Hereditas **75:** 221–232.

GILLESPIE, J. H. and C. H. LANGLEY, 1974 A general model to account for enzyme variation in natural populations. Genetics **76:** 837–848.

GORMAN, G. C. and Y. J. KIM, 1976 Anolis lizards of the eastern Caribbean: a case study in evolution. II. Genetic relationships and genetic variation of the *bimaculatus* group. Systemat. Zool. **25:** 62–77.

GORMAN, G. C., Y. J. KIM and R. RUBINOFF, 1976 Genetic relationships of three species of *Bathygobius* from the Atlantic and Pacific sides of Panama. Copeia, No. 2, 361–364.

GREENBAUM, I. F. and R. J. BAKER, 1976 Evolutionary relationships in *Macrotus* (Mammalia: Chiroptera): Biochemical variation and karyology. Systemat. Zool. **25:** 15–25.

HALL, W. P. and R. K. SELANDER, 1973 Hybridization of karyotypically differentiated populations in the *Sceloporus grammicus* complex (Iguanidae). Evolution **27:** 226–242.

HIGHTON, R. and T. P. WEBSTER, 1976 Geographic protein variation and divergence in populations of the salamander *Plethodon cinereus*. Evolution **30:** 33–45.

JOHNSON, A. G. and F. M. UTTER, 1976 Electrophoretic variation in intertidal and subtidal organisms in Puget Sound, Washington. Animal Blood Groups and Biochemical Genetics **7:** 3–14.

JOHNSON, A. G., F. M. UTTER and H. O. HODGINS, 1973 Estimate of genetic polymorphism and heterozygosity in three species of rockfish (Genus *Sebastes*). Comp. Biochem. Physiol. **44B:** 397–406.

JOHNSON, G. B., 1972 Evidence that enzyme polymorphisms are not selectively neutral. Nature New Biol. **237:** 170–171. ——, 1974 On the estimation of effective number of alleles from electrophoretic data. Genetics **78:** 771–776.

JOHNSON, M. S., 1975 Biochemical systematics of the atherinid genus *Menidia*. Copeia, No. 4, 662–691.

JOHNSON, W. E. and R. K. SELANDER, 1971 Protein variation and systematics in kangaroo rats (genus *Dipodomys*). Systemat. Zool. **20**: 377–405.

KAUFMAN, D. W., R. K. SELANDER and M. H. SMITH, 1973 Genic heterozygosity in a population of *Eutamias panamintinus*. J. Mammalogy **54**: 776–778.

KIM, Y. J., 1972 Studies of biochemical genetics and karyotypes in pocket gophers (family *Geomyidae*). Ph.D. Dissertation, Univ. of Texas, Austin, Texas.

KIM, Y. J., G. C. GORMAN, TH. PAPENFUSS and A. K. ROYCHOUDHURY, 1976 Genetic relationships and genetic variation in the Amphisbaenian genus *Bipes*. Copeia, No. 1, 120–124.

KIMURA, M., 1968 Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11**: 247–269. ——, 1971 Theoretical foundation of population genetics at the molecular level. Theoret. Pop. Biol. **2**: 174–208.

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

KIMURA, M. and T. OHTA, 1971 Protein polymorphism as a phase of molecular evolution. Nature **229**: 467–469, ——, 1973 Mutation and evolution at the molecular level. Genetics **73** (suppl.): 19–35. ——, 1975 Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. Proc. Natl. Acad. Sci. **72**: 2761–2764.

KING, J. L. and T. OHTA, 1975 Polyallelic mutational equilibria. Genetics **79**: 681–691.

KIRBY, G. C. and R. B. HALLIDAY, 1973 Another view of neutral alleles in natural populations. Nature **241**: 463–464.

KOJIMA, K., J. GILLESPIE and Y. N. TOBARI, 1970 A profile of Drosophila species' enzymes assayed by electrophoresis. I. Number of alleles, heterozygosities, and linkage disequilibrium in glucose-metabolizing systems and some other enzymes. Biochem. Genet. **4**: 627–637.

LAKOVAARA, S. and A. SAURA, 1971 Genetic variation in natural populations of *Drosophila obscura*. Genetics **69**: 377–384.

LATTER, B. D. H., 1972 Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. Genetics **70**: 475–490. ——, 1975 Influence of selection pressures on enzyme polymorphisms in Drosophila. Nature **257**: 590–592.

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

LI, W.-H., 1976 A mixed model of mutation for electrophoretic identity of proteins within and between populations. Genetics **83**: 423–432.

LI, W.-H. and M. NEI, 1975 Drift variances of heterozygosity and genetic distance in transient states. Genet. Res. **25**: 229–248.

MCDERMID, E. M. and W. N. BONNER, 1975 Red cell and serum protein systems of grey seals and habour seals. Comp. Biochem. Physiol. **50B**: 97–101.

MORAN, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theoret. Pop. Biol. **8**: 318–330. ——, 1976 A selective model for electrophoretic profiles in protein polymorphisms. Genet. Res. **28**: 47–53.

NEI, M., 1975 *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam and New York.

NEI, M. and W.-H. LI, 1975 Probability of identical monomorphism in related species. Genet. Res. **26**: 31–43.

NEI, M. and A. K. ROYCHOUDHURY, 1974a Sampling variances of heterozygosity and genetic distance. Genetics **76**: 379–390. ——, 1974b Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. Amer. J. Hum. Genet. **26**: 421–443.

NEI, M., R. CHAKRABORTY and P. A. FUERST, 1976a Infinite allele model with varying mutation rate. Proc Natl. Acad. Sci. **73**: 4164–4168.

NEI, M., P. A. FUERST and R. CHAKRABORTY, 1976b Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. Nature **262**: 491–493.

NEVO, E., Y. J. KIM, C. R. SHAW and C. S. THAELER, 1974 Genetic variation, selection and speciation in *Thomomys talpoides* pocket gophers. Evolution **28**: 1–23.

NEVO, E., H. C. DESSAUER and K.-C. CHUANG, 1975 Genetic variation as a test of natural selection. Proc. Natl. Acad. Sci. **72**: 2145–2149.

NOZAWA, K., T. SHOTAKE and Y. OKURA, 1975 Blood protein polymorphisms and population structure of the Japanese macaque, *Macaca fuscata fuscata*. pp. 225–241. In: *Isozymes, Vol. IV. Genetics and Evolution*. Edited by C. L. MARKERT. Academic Press, New York.

OHTA, T., 1974 Mutational pressure as the main cause of molecular evolution and polymorphism. Nature **252**: 351–354. ——, 1975 Statistical analyses of Drosophila and human protein polymorphisms. Proc. Natl. Acad. Sci. **72**: 3194–3196.

OHTA, T. and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22**: 201–204.

PARKER, E. D. and R. K. SELANDER, 1976 The organization of genetic diversity in the parthenogenetic lizard *Cnemidophorus tesselatus*. Genetics **84**: 791–805.

PATTON, J. L., R. K. SELANDER, and M. H. SMITH, 1972 Genic variation in hybridizing populations of gophers (genus *Thomomys*). Systemat. Zool. **21**: 263–270.

PATTON, J. L., S. Y. YANG and P. MYERS, 1975 Genetic and morphologic divergence among introduced rat populations (*Rattus rattus*) of the Galápagos Archipelago, Ecuador. Systemat. Zool. **24**: 296–310.

PATTON, J. L., H. MACARTHUR and S. Y. YANG, 1976 Systematic relationships of the four-toed populations of *Dipodomys heermanni*. J. Mammalogy **57**: 159–163.

POWELL, J. R., 1975 Protein variation in natural populations of animals. Evol. Biol. **8**: 79–119.

PRAKASH, S., 1969 Genic variation in a natural population of *Drosophila persimilis*. Proc. Natl. Acad. Sci. **62**: 778–784. ——, 1973a Patterns of gene variation in central and marginal populations of *Drosophila robusta*. Genetics **75**: 347–369. ——, 1973b Low gene variation in *Drosophila busckii*. Genetics **75**: 571–576.

PRAKASH, S., R. C. LEWONTIN and J. L. HUBBY, 1969 A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of *Drosophila pseudoobscura*. Genetics **61**: 841–858.

ROBERTSON, A. 1967 The nature of quantitative genetic variation. pp. 265–280. In: *Heritage from Mendel*. Edited by R. A. BRINK. Univ. of Wisconsin Press, Madison, Wisconsin.

ROCKWOOD, E. S., C. G. KANAPI, M. R. WHEELER and W. S. STONE, 1971 Allozyme changes during the evolution of Hawaiian Drosophila. Univ. Texas Publ. **7103**: 193–212.

SABATH, M. D., 1975 Enzyme variability in 12 sympatric drosophilid species (Genera: *Chymomyza, Leucophenga, Scaptomyza,* and *Drosophila*). Amer. Midland Naturalist **94**: 144–153.

SAGE, R. D. and R. K. SELANDER, 1975 Trophic radiation through polymorphism in cichlid fishes. Proc. Natl. Acad. Sci. U.S. **72**: 4669–4673.

SAURA, A., O. HALKKA and J. LOKKI, 1973 Enzyme gene heterozygosity in small island populations of *Philaenus spumarius* (L.) (Homoptera). Genetica **44**: 459–473.

SCHOPF, T. J. M. and L. S. MURPHY, 1973 Protein polymorphism of the hybridizing seastars *Asterias forbesi* and *Asterias vulgaris* and implications for their evolution. Biol. Bull. **145**: 589–597.

SELANDER, R. K., 1976 Genic variation in natural populations. pp. 21–45. In: *Molecular Evolution.* Edited by F. J. AYALA. Sinauer Assoc., Sunderland, Massachusetts.

SELANDER, R. K., W. G. HUNT and S. Y. YANG, 1969 Protein polymorphism and genic heterozygosity in two European subspecies of the house mouse. Evolution **23:** 379–390.

SELANDER, R. K. and D. W. KAUFMAN, 1973 Genic variability and strategies of adaptation in animals. Proc. Natl. Acad. Sci. U.S. **70:** 1875–1877.

SELANDER, R. K., D. W. KAUFMAN, R. J. BAKER and S. L. WILLIAMS, 1974 Genic and chromosomal differentiation in pocket gophers of the *Geomys bursarius* group. Evolution **28:** 557–564.

SELANDER, R. K., M. H. SMITH, S. Y. YANG, W. E. JOHNSON and J. B. GENTRY, 1971 Biochemical polymorphism and systematics in the genus *Peromyscus.* I. Variation in the old-field mouse (*Peromyscus polionotus*). Univ. Texas Publ. **7103:** 49–90.

SELANDER, R. K., S. Y. YANG, R. C. LEWONTIN and W. E. JOHNSON, 1970 Genetic variation in the horseshoe crab (*Limulus polyphemus*), a phylogenetic "relic." Evolution **24:** 402–414.

SMITH, M. H., R. K. SELANDER and W. E. JOHNSON, 1973 Biochemical polymorphism and systematics in the genus *Peromyscus.* III. Variation in the Florida deer mouse (*Peromyscus floridanus*), a Pleistocene relict. J. Mammalogy **54:** 1–13.

SNYDER, T. P., 1974 Lack of allozymic variability in three bee species. Evolution **28:** 687–689.

SOMERO, G. N. and M. SOULÉ, 1974 Genetic variation in marine fishes as a test of the niche-variation hypothesis. Nature **249:** 670–672.

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

STRANEY, D. O., M. J. O'FARRELL and M. H. SMITH, 1976a Biochemical genetics of *Myotis californicus* and *Pipistrellus hesperus* from southern Nevada. J. Mammalogy (In press).

STRANEY, D. O., M. H. SMITH, R. J. BAKER, and I. F. GREENBAUM, 1976b Biochemical variation and genic similarity of *Myotis velifer* and *Macrotus californicus.* Comp. Biochem. Physiol. **54B:** 243–248.

SUOMALAINEN, E. and A. SAURA, 1973 Genetic polymorphism and evolution in parthenogenetic animals. I. Polyploid *Curculionidae.* Genetics **74:** 489–508.

TAYLOR, C. E. and G. C. GORMAN, 1975 Population genetics of a "colonising" lizard: natural selection for allozyme morphs in *Anolis grahami.* Heredity **35:** 241–247.

TRACEY, M. L., K. NELSON, D. HEDGECOCK, R. A. SHLESER and M. L. PRESSICK, 1975 Biochemical genetics of lobsters: Genetic variation and the structure of American lobster (*Homarus americanus*) populations. J. Fish. Res. Brd. Canada **32:** 2091–2101.

UTTER, F. M., F. W. ALLENDORF and H. O. HODGINS, 1973 Genetic variability and relationships in Pacific salmon and related trout based on protein variations. Systemat. Zool. **22:** 257–270.

WATTERSON, G. A., 1974 Models for the logarithmic species abundance distributions. Theoret. Pop. Biol. **6:** 217–250. ——, 1977 Heterosis or neutrality? Genetics **85:** 789–814.

WEBSTER, T. P., R. K. SELANDER and S. Y. YANG, 1972 Genetic variability and similarity in the *Anolis* lizards of Bimini. Evolution **26:** 523–535.

YAMAZAKI, T. and T. MARUYAMA, 1972 Evidence for the neutral hypothesis of protein polymorphism. Science **178:** 56–58. ——, 1973 Evidence that enzyme polymorphisms are selectively neutral. Nature New Biol. **245:** 140–141.

YANG, S. Y., L. L. WHEELER and I. R. BOCK, 1972 Isozyme variations and phylogenetic relationships in the *Drosophila bipectinata* species complex. Univ. Texas Publ. **7213:** 213–227.

YANG, S. Y., M. SOULÉ and G. C. GORMAN, 1974 *Anolis* lizards of the eastern Caribbean: a case study in evolution. I. Genetic relationships, phylogeny, and colonization sequence of the *roquet* group. Systemat. Zool. **23**: 387–399.

ZOUROS, E., C. B. KRIMBAS, S. TSAKAS and M. LOUKAS, 1974 Genic *versus* chromosomal variation in natural populations of *Drosophila subobscura*. Genetics **78**: 1223–1244.

Corresponding editor: J. F. KIDWELL

APPENDIX

COMPUTER ALGORITHM FOR OBTAINING A RANDOM SET OF ALLELE FREQUENCIES FOR A LOCUS IN AN EQUILIBRIUM POPULATION

FRANK M. STEWART

Department of Mathematics, Brown University, Providence, Rhode Island

In a finite population the number of alleles and their frequencies may vary from locus to locus owing to random genetic drift. Using the infinite allele model, EWENS (Theoret. Pop. Biol. **3**: 87, 1972), has shown that the probability of finding k alleles, $A_1, A_2, \ldots, A_k$, in a sample of n genes at a locus is

$$Q(k) = \Gamma(M)M^k n!B(k,n)/\{\Gamma(n+M)k!\}\ , \tag{A1}$$

where $M = 4N_e v$, in which N_e and v are the effective population size and mutation rate, respectively, and

$$B(k,n) = \sum_{n_1,\ldots,n_k} (n_1 n_2 \ldots n_k)^{-1}$$

where the summation is taken over all possible permutations of $n_1, n_2, \ldots, n_k$ with the restriction of $n_i \geq 1$ and $\Sigma_i n_i = n$. EWENS' original formula is expressed in terms of the Stirling number of the first kind ($S_n^{(k)}$) rather than $B(k,n)$. These two quantities are related by

$$B(k,n) = (-1)^{n-k}\frac{k!}{n!}S_n^{(k)}\ .$$

He has also shown that the conditional probability, given n and k, that there are n_1 genes of A_1, n_2 genes of A_2, etc. is given by

$$P(n_1,\ldots,n_k) = [B(k,n)n_1 n_2 \ldots n_k]^{-1}\ . \tag{A2}$$

Our strategy of obtaining a random set of allele frequencies (numbers) for a locus is first to determine the number of alleles (k) in a sample at random and then choose a random set of $n_1, n_2, \ldots, n_k$ for a given value of k. The first process is accomplished by generating a random number of uniform distribution on $[0,1]$ and comparing it with the cumulative probabilities of (A1). The same principle is used in determining n_1, n_2, etc., but the actual computation is greatly simplified if we note the following properties of (A2). Namely, for a given value of k,

$$\begin{aligned} P(n_1) &= \sum_{n_2,\ldots,n_k} [B(k,n)n_1 n_2 \ldots n_k]^{-1} \\ &= [B(k,n)n_1]^{-1} \sum_{n_2,\ldots,n_k} [n_2 n_3 \ldots n_k]^{-1} \\ &= \frac{B(k-1,n-n_1)}{B(k,n)n_1}\ . \end{aligned} \tag{A3}$$

Similarly,

$$P(n_1,n_2) = \frac{B(k-2,n-n_1-n_2)}{B(k,n)n_1 n_2}\ .$$

In general, for $l \leq k$,

$$P(n_1, \ldots, n_l) = \frac{B(k-l, n-n_1 \ldots - n_l)}{B(k,n) n_1 \ldots n_l} . \tag{A4}$$

Furthermore,

$$P(n_l | n_1, \ldots, n_{l-1}) = \frac{B(k-l, n-n_1 \ldots - n_l)}{B(k-l+1, n-n_1 \ldots - n_{l-1}) n_l} . \tag{A5}$$

Therefore, the value of n_1 for a given sample can be determined by generating a random number on $[0,1]$ and comparing it with the cumulative probabilities of (A3). Namely, if A is the random number generated, successive sums, $P(1)$, $P(1) + P(2)$, . . . are computed until $P(1) + P(2) + \ldots + P(n_1)$ becomes equal to or exceeds A. The last value of n_1 is the value to be chosen. Once n_1 is chosen, n_2, n_3, etc. are determined by the same procedure but now using (A5).

For computing probabilities (A3) and (A5), we have to know the value of $B(i,j)$ for $i = 1$ to k and $j = 1$ to n. To compute $B(i,j)$, we use the recursion formula

$$B(i,j+1) = [iB(i-1,j) + jB(i,j)]/(j+1),$$

which can be obtained from the corresponding formula for the Stirling number of the first kind:

$$S^{(i)}_{j+1} = S^{(i-1)}_{j} - jS^{(i)}_{j} .$$

A computer program for the above computation has been written by F. M. STEWART and Y. TATENO. It is available by writing to M. NEI.