Genet. Res., Camb. (1977), **29**, *pp.* 223–230
With 1 *text-figure*
Printed in Great Britain

The correlation of rare alleles with heterozygosity: determination of the correlation for the neutral models*

BY WALTER F. EANES† AND RICHARD K. KOEHN
Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794, U.S.A.

(*Received* 10 *December* 1976)

SUMMARY

We have shown in an earlier paper that there is routinely a large correlation between the heterozygosity of common ($P > 0{\cdot}05$) alleles and the number of rare ($P \leqslant 0{\cdot}05$) alleles at allozyme loci in *Drosophila*. We postulated that this correlation might be due to a high rate of intragenic recombination. While these correlations are large enough to be significantly different from zero, their relation to the mean correlations expected under the neutrality models is unknown. This paper reports the findings of a computer analysis determining the correlation for neutral allele pools as specified by the infinite-allele and charge-state models.

In the analysis, mean correlations for a range of $N_e\mu$ values and sample sizes of 100 and 1000 genes varied from a high of 0·284 to a low of −0·780. For the particular values of $N_e\mu$ relating to the heterozygosity of *Drosophila* allozymes in natural populations, tests of the empirical correlations to the means expected under the neutral models are made. Most empirical correlations are significantly different from the mean correlations under the neutrality models.

1. INTRODUCTION

We have reported the persistence of a large correlation between the number of 'rare' alleles and the heterozygosity of so-called 'common' alleles at allozyme loci in natural populations of *Drosophila*. Since there is no *a priori* reason to expect such a large correlation due to sampling from a multinomial distribution, we postulated that the rare variants were being regularly generated by higher than expected rates of mutation or intragenic recombination (Koehn & Eanes, 1976) involving the parental (common) alleles. However, the magnitude of expected correlations due to sampling from neutral allele pools were unknown. We report here the expected distribution of correlations in samples from neutral allele pools described by the infinite-allele (Kimura & Crow, 1964) or charge-state models of

* This is contribution number 176 from the Program in Ecology and Evolution at the State University of New York at Stony Brook.

† Present address: Institute of Ecology and Genetics, University of Aarhus, DK-8000, Aarhus C, Denmark. After 1 September 1977, Museum of Comparative Zoology, Harvard University, Cambridge, Mass. 02138.

the neutral theory (Ohta & Kimura, 1973). The distributions were simulated with a spectrum of values of effective population sizes, mutation rates, and sample size for both models.

2. BACKGROUND

In samples from natural populations of *Drosophila* there is a consistent, large correlation between the heterozygosity of alleles having observed frequencies greater than 0·05, and the total number of alleles having frequencies *persistently* less than or equal to 0·05; here termed 'common' and 'rare' respectively. The distinction between commonness and rareness is arbitrary, and for practical reasons has been set at $P = 0{\cdot}05$ here. The data reported earlier (Koehn & Eanes, 1976) came from a variety of published and unpublished sources, and were generally available only in the form of allele frequencies. Only data sets averaging 200 or more genes sampled per locus and having at least 10 loci per set were considered in our analyses. Since the number of rare alleles detected at any locus is heavily sample size dependent, all correlations were computed by weighting loci or localities by specific sample sizes.

For each data set the expected heterozygosity of common alleles and respective numbers of rare alleles for each locus were computed. For example, by the definition above, a locus having five alleles with relative frequencies 0·813, 0·116, 0·060, 0·006 and 0·005 has two rare alleles and an expected heterozygosity of the common alleles computed as $1-(0{\cdot}813)^2+(0{\cdot}116)^2+(0{\cdot}060)^2=0{\cdot}322$. The common allele heterozygosity was arc–sine transformed to normalize the variable (Sokal & Rohlf, 1969). For loci with only two alleles, where one allele was less than 0·05 in frequency, the heterozygosity was still computed, but the number of rare alleles was listed as zero. For loci with multiple alleles, but only one allele having a frequency greater than 0·05, the heterozygosity was computed between the most common allele and second most common allele. Thus, a locus was considered monomorphic only if one electrophoretic allele was observed. This treatment had very little effect on the correlation and excluded the contradiction of a locus having zero heterozygosity when alternate alleles are observed.

A weighted Pearson product-moment correlation coefficient was calculated over loci for each data set between the expected heterozygosity of the common alleles and the observed number of rare alleles at each locus. The weight applied in each case was the number of genes sampled per locus.

The above correlation was determined for 28 data sets representing 16 *Drosophila* species. Statistically significant correlations were observed in 20 of the 28 sets and all data sets with 17 or more loci sampled exhibited significant correlations (Table 1).

Since it was possible that the large observed correlations could be at least partly due to a parallel increase in both the number of rare and common alleles, the correlations were computed after excluding all loci with greater than two common alleles. This approach had no effect on the relative levels of the correlation coefficients.

If intragenic crossing over were a significant force in determining the numbers of rare alleles at electrophoretic loci, several confounding factors capable of obscuring the correlation reported here might be expected. For example, heterogeneity among loci in the number and position of variable sites, overall cistron size, extracistronic interference of recombination and finally selection, could each conceivably disrupt any predicted correlation, even if intracistronic recombination were a common event. However, it might be assumed that these factors are

Table 1. *Rare alleles and heterozygosity in* Drosophila

(Only those data sets possessing at least 17 loci are shown.)

Species	Ref.*	No. of loci	$\bar{N}$/locus	r	P	r^2
D. equinoxialis (Ven.)	1	30	343·2	0·864	< 0·001	0·747
D. willistoni (Ven.)	1	29	494·9	0·762	< 0·001	0·581
D. tropicalis (Ven.)	1	28	145·4	0·535	< 0·005	0·286
D. willistoni (C.)	2	25	2034·0	0·908	< 0·001	0·825
D. robusta	3	22	394·3	0·885	< 0·001	0·783
D. willistoni (Trin.)	2	21	332·8	0·974	< 0·001	0·949
D. tropicalis (S.D.)	4	21	124·5	0·773	< 0·001	0·598
D. nebulosa (Sant.)	4	20	119·0	0·816	< 0·001	0·666
D. tropicalis (Sant.)	4	20	105·6	0·645	< 0·005	0·416
D. bifasciata	5	19	368·7	0·860	< 0·001	0·740
D. nebulosa (S.D.)	4	19	74·4	0·456	< 0·05	0·208
D. tropicalis (May.)	4	19	72·1	0·621	< 0·01	0·386
D. tropicalis (Barr.)	4	17	51·3	0·771	< 0·001	0·594

* (1) Ayala *et al.* 1974; (2) Ayala *et al.* 1972; (3) Prakash, 1973; (4) Ayala, unpubl.; (5) Saura, 1974.

Table 2. *Correlation between heterozygosity of common alleles and the observed number of rare alleles at seven loci irrespective of species*

Locus	No. of loci	r	r^2
Acph	25	0·948*	0·899
Lap	21	0·981*	0·962
Adh	19	0·926*	0·857
Pgm	19	0·696*	0·484
Xdh	17	0·840*	0·706
αGpdh	16	0·980*	0·960
Idh	13	6·770**	0·593

* $P \leqslant 0{\cdot}001$. ** $P < 0{\cdot}005$.

relatively uniform across species lines for homologous cistrons. In fact, there is good evidence that the among-locus variance in allelic diversity for allozymes in *Drosophila* is at present best explained as a simple function of the molecular structure of those enzymes (Koehn & Eanes, 1977; Ward, 1977). Thus, allelic diversity *per se* may have a larger cistron-specific component than species-specific component. In fact, correlations between common allele heterozygosity and

numbers of rare alleles computed for homologous enzyme-loci across species lines are very large (Table 2).

Two widely accepted neutral models (Kimura & Crow, 1964; Ohta & Kimura, 1973) both predict specific numbers and frequencies of electrophoretic alleles from the product of the effective population size and the per cistron, per generation mutation rate. Ewens (1972) described the sampling distribution of neutral alleles under the assumptions of the older Kimura–Crow, or infinite-allele model, in his development of a test of empirical frequency distributions against distributions expected theoretically under that neutral model. More recently, the distribution of neutral allele frequencies under the ladder-rung or stepwise charge-state model has been described by Kimura & Ohta (1975). We may use both of these theoretical allele distributions to generate neutral allele pools from which we may sample and compute the correlation we have just described.

3. THE COMPUTER ANALYSIS

The scheme used to construct the neutral allele pools was suggested by Nei, Maruyama & Chakraborty (1975). In that procedure the interval (0, 1) is first subdivided into n equal subintervals; where n equals the number of genes sampled per locus. The ordinate, $f(x_i)$, for each allele frequency class, x_i, is computed by substituting the mean value for each subinterval into the equilibrium density function $\Phi(x)$. For the infinite-allele model,

$$\Phi(x) = 4N_e\mu x^{-1}(1-x)^{4Ne\mu-1}, \tag{1}$$

where N_e is the effective population size and μ is the per cistron, per generation mutation rate (Ewens, 1972). In the case of the charge-state model,

$$\Phi(x) = \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha)\,\Gamma(\beta+1)}(1-x)^{\alpha-1}x^{\beta-1}, \tag{2}$$

where $\alpha = 4N_e\mu$, $\beta = 2N_e\mu b$ and

$$b = \frac{1+4N_e\mu-\sqrt{(1+8N_e\mu)}}{2N_e\mu[\sqrt{(1+8N_e\mu)}-1]} \quad \text{(Kimura \& Ohta, 1975).}$$

For the discrete density functions, cumulative frequency distributions are computed as

$$F(x_i) = \sum_{j=1}^{i} f(x_j). \tag{3}$$

To determine the number of copies of a given allelic type, say A_k, random numbers are sequentially drawn uniformly between 0 and 1 and the smallest value of $F(x_i)$ equal to or exceeding the kth random number is the number of copies of the A_k allele in the total sample of n. Succeeding alleles are extracted until the total number of genes sampled equals or exceeds n. When the total number exceeds n, as it usually does, the number of genes of the last allelic type sampled is reduced by the difference with n. This process is repeated for each locus. Sampling in this fashion, the number of alleles and their respective

frequencies should approximate neutral allele pools as specified by (1) or (2), depending on the model of interest, and assuming this method of sampling has introduced no bias. In this fashion, for a given value of $N_e\mu$, 20 loci were sampled. After discarding all fixed loci, the correlation described above was computed. A mean correlation was determined from 100 replications. Sample sizes of both 100 or 1000 genes were considered.

Levels of allozyme heterozygosity in natural populations of *Drosophila* are predicted by certain product values of $N_e\mu$, depending on the model followed. For the infinite-allele model this value corresponds approximately to $N_e\mu = 0{\cdot}056$ and for the charge-state model a slightly larger value of $N_e\mu = 0{\cdot}062$. For both of these values 500 replications with two sample sizes of 100 and 1000 genes per locus were run to construct a distribution of correlations from which to assess the probability of our empirical correlations having come from the specific neutral allele pools as defined by either model.

4. RESULTS

Table 3 lists the mean correlation ($\bar{r}$) for seven values of $N_e\mu$ with sample sizes of 100 and 1000 genes per locus. Average correlations are contrasted for both the infinite-allele and charge-state models. $\bar{r}$ is slightly higher with samples of 1000 genes. In both models the values of $\bar{r}$ are positive for small magnitudes of $N_e\mu$ and diminish as $N_e\mu$ increases. The largest average positive correlation observed $\bar{r} = 0{\cdot}284$ was for the infinite-allele model, corresponding to $N_e\mu = 0{\cdot}01$.

Table 3. *Mean correlations $\bar{r}$ described in text, for selected values of $N_e\mu$, and two sample sizes, for both the infinite-allele (A) and charge-state (B) neutral models*

	$n = 100$		$n = 1000$	
$N_e\mu$	$\bar{r}$ (A)	$\bar{r}$ (B)	$\bar{r}$ (A)	$\bar{r}$ (B)
0·010	0·191	0·185	0·284	0·229
0·056*	0·163	—	0·276	—
0·062*	—	0·165	—	0·159
0·100	0·171	0·153	0·244	0·135
0·200	0·044	0·016	0·131	0·139
0·500	−0·402	−0·135	−0·285	−0·078
1·000	−0·752	−0·322	−0·780	−0·276

* $N_e\mu$ values commensurate with magnitudes of allozyme heterozygosity levels in *Drosophila* under the two models discussed in the text.

Having described the general response of the correlation to a range of $N_e\mu$ values with two different sample sizes for each model, we have then determined whether our observed correlations from natural populations are expected from sampling the neutral allele pools defined by the models (Fig. 1).

Inspection of the distribution of empirical correlation coefficients demonstrates that most fall well outside the two expected neutral distributions. We conclude that the large empirical correlations we have observed in natural populations of

Drosophila between the heterozygosity of 'common' type alleles and the numbers of rare alleles are not predicted by either of the two most widely investigated neutral models.

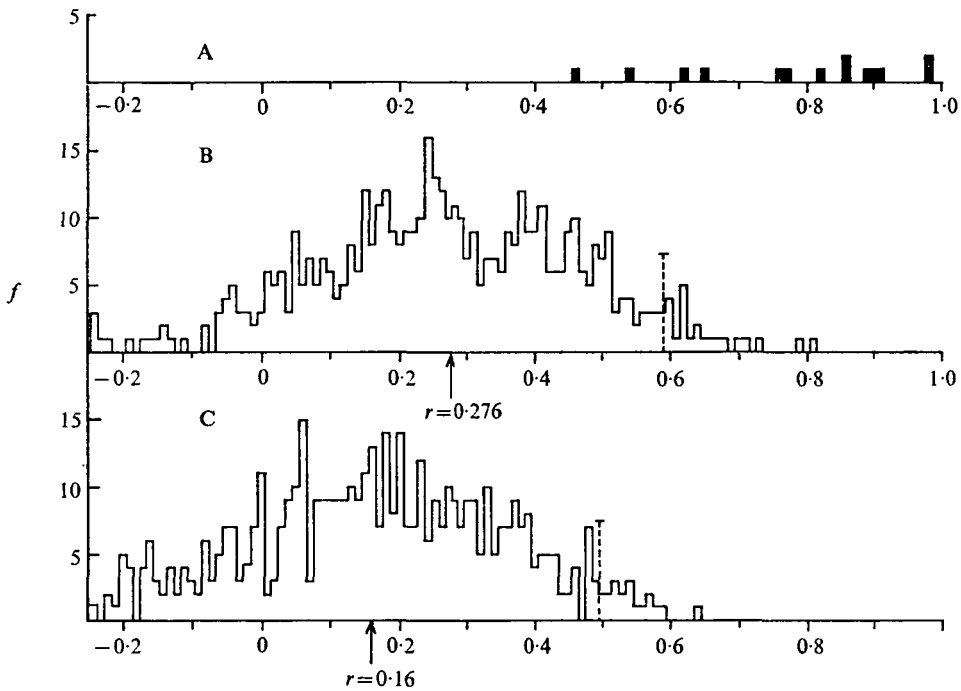

Fig. 1. The distribution of r in samples from natural populations of *Drosophila* (A) and from neutral allele pools produced from the infinite-allele (B) and charge-state (C) models. Distributions B and C were computed with 500 replications each. Correlation coefficients were computed for samples of twenty loci and 1000 genes per locus. Correlations in A are for *Drosophila* data with at least 20 loci, but various sample sizes. Vertical dashed lines in B and C represent the approximate 5% rejection region ($P < 0{\cdot}05$) for a test of the null hypothesis of an observed r having been sampled from the neutral allele pools specified in the text.

5. IMPLICATIONS

The infinite allele model of Kimura and Crow has been shown to be an inadequate description of the patterns of allelic diversity in electrophoretic data (e.g. Johnson, 1972; Johnson & Feldman, 1973). Recently this model has been replaced by the charge-state model, which is presumably more realistic. The latter model assumes that alleles will exist as a limited number of charge classes, rather than as a nearly infinite set of potential electrophoretic alleles. It also predicts fewer numbers of alleles for given magnitudes of $N_e\mu$ and larger ratios of effective numbers of alleles ($n_e = 1/\Sigma x_i^2$; where x_i equals the frequency of the ith allele) to actual observed numbers (n_e/n_a), thereby correcting a deficiency of the earlier model (Ohta & Kimura, 1974).

While the new stepwise charge-state model is in better accordance with certain

qualitative and quantitative observations in empirical data, neither model predicts the large correlations we have determined to exist between the heterozygosity of 'common' alleles and number of 'rare' type alleles in *Drosophila*. In comparing these observations to the expectations of models, we have assumed that these real loci are at mutation-drift equilibria and we do not know if this large correlation could be generated by a series of loci lying at various distances from a theoretical mutation-drift equilibrium. Furthermore, a condition where μ is equal in all loci no doubt is unrealistic (Koehn & Eanes, 1977), but a mixed model has not been investigated here. However, the larger average correlations computed for homologous enzyme-loci (Table 2) weakens this explanation.

The common alleles reported here have in numerous instances been cited as possessing characteristics compatible with a selective maintenance (e.g. clines, differences in kinetic properties, etc.). However, there are at present no formal selection models to account for the correlation described here. The selective status of rare variants in *Drosophila* is unknown and would be logistically difficult to estimate in the laboratory.

Whatever the underlying mechanisms, our results suggest that rare electrophoretic variants are being produced at a rate dependent on the relative frequencies of potential parental alleles. It seems apparent that both intragenic recombination and mutation rate need to be examined more extensively.

Mutation rates have been routinely determined for homozygous or haploid cistrons. Rates of intragenic recombination (especially gene conversion) have been estimated for non-allozyme loci as high as 10^{-2} to 10^{-5} in *Neurospora* and yeast (Stadler & Kariya, 1969; Fogel, Hurst & Mortimer, 1971, and others). For allozyme loci rates of intragenic recombination of the order of 10^{-2} to 10^{-5} have been either demonstrated (Chovnick *et al.* 1970) or postulated from the apparent high mutation rates observed in heterozygous parents (Ohno *et al.* 1969; Wright & Atherton, 1968; Tsuno, 1975). Whatever the rate necessary to produce such a correlation the dynamics are complex (Watt, 1972).

The inability of these neutral models to predict the observed high correlation does not necessarily deny the selective equivalency of these alleles, only the inability of the current models to explain observable patterns of allelic diversity. New neutral models should attempt to cope with this correlation. Perhaps the consideration of intragenic recombination or heterozygote-dependent mutation rates as additional forces would permit this adjustment.

We would like to thank Freddy B. Christiansen for commenting on the original manuscript. During the tenure of this study R. K. Koehn was supported by USPHS grant GM-28963, and W. F. Eanes by both USPHS Grant GM-21133 to R. K. Koehn and NSF Grant B393000 to R. G. Turner. Computer time was supplied by the Department of Ecology and Evolution, State University of New York at Stony Brook.

REFERENCES

Ayala, F. J., Powell, J. R., Tracey, M. L., Mourao, C. A. & Perez-Salas, S. (1972). Enzyme variability in the *Drosophila willistoni* group. IV. Genic variation in natural populations of *D. willistoni*. *Genetics* **70**, 113–139.

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

Chovnick, A., Ballantyne, G. H., Baillie, D. L. & Holm, D. G. (1970). Gene conversion in higher organisms: half-tetrad analysis of recombination within the rosy cistron of *Drosophila melanogaster*. *Genetics* **66**, 315–329.

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

Fogel, S., Hurst, D. D. & Mortimer, R. K. (1971). Gene conversion in unselected tetrads from multi-point crosses. In *Stadler Symposium*, vol. 1, pp. 89–110. Missouri Agricultural Experimental Station, Columbia.

Johnson, G. B. (1972). Enzyme polymorphisms. Evidence that they are not selectively neutral. *Nature, New Biology* **237**, 170–171.

Johnson, G. B. & Feldman, M. W. (1973). On the hypothesis that polymorphic enzyme alleles are selectively neutral. I. The evenness of the allele frequency distribution. *Theoretical Population Biology* **4**, 209–221.

Kimura, M. & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–728.

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

Koehn, R. K. & Eanes, W. F. (1976). An analysis of allelic diversity in natural populations of *Drosophila*: The correlation of rare alleles with heterozygosity. In *Population Genetics and Ecology* (ed. S. Karlin and E. Nevo), pp. 377–390. New York: Academic Press.

Koehn, R. K. & Eanes, W. F. (1977). Subunit size and genetic variation in natural populations of *Drosophila*. *Theoretical Population Biology* (in the Press).

Nei, M., Maruyama, T. & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10.

Ohno, S., Stenius, C., Christian, L. & Schipmann, G. (1969). *De novo* mutation-like events observed at the 6PGD locus of the Japanese quail, and the principle of polymorphism breeding more polymorphism. *Biochemical Genetics* **3**, 417–428.

Ohta, T. & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research* **22**, 201–204.

Ohta, T. & Kimura, M. (1974). Simulation studies on electrophoretically detectable genetic variability in a finite population. *Genetics* **76**, 615–624.

Prakash, S. (1973). Patterns of gene variation in central and marginal populations of *Drosophila robusta*. *Genetics* **75**, 347–369.

Saura, A. (1974). Genic variation in Scandinavian populations of *Drosophila bifasciata*. *Hereditas* **76**, 161–172.

Sokal, R. R. & Rohlf, F. J. (1969). *Biometry*. San Francisco: W. H. Freeman.

Stadler, D. R. & Kariya, B. (1969). Intragenic recombination at the *mtr* locus of *Neurospora* with segregation at an unselected site. *Genetics* **63**, 291–316.

Tsuno, K. (1975). Esterase gene frequency differences and linkage equilibrium in *Drosophila virilis* from different ecological habitats. *Genetics* **80**, 585–594.

Ward, R. D. (1977). Relationship between enzyme heterozygosity and quaternary structure. *Biochemical Genetics* **15**, 123–135.

Watt, W. B. (1972). Intragenic recombination as a source of population genetic variability. *American Naturalist* **106**, 737–753.

Wright, J. E. & Atherton, L. (1968). Genetic control of interallelic recombination at the LDH B locus in brook trout. *Genetics* **60**, 240.