Genet. Res., Camb. (1981), **38**, *pp.* 137–156
With 2 text-figures
Printed in Great Britain

The distribution of heterozygosity in temperate and tropical species of *Drosophila*

By B. D. H. LATTER

Faculty of Agriculture, University of Sydney, 2006, Australia

(*Received* 13 *May* 1980 *and in revised form* 3 *March* 1981)

SUMMARY

Electrophoretic surveys for nine species of *Drosophila* have been summarized in terms of the relative contribution to heterozygosity of each of ten gene frequency classes, the mean frequency of heterozygotes within subpopulations, and the degree of genetic divergence between subpopulations. It has been shown that the neutral model proposed by Kimura, and modified by Ohta to include the accumulation of slightly disadvantageous mutations, is capable of explaining all features of the data. The consistent difference between group I and group II enzymes can be explained by a difference in the average intensity of selection against mutational variants in the two groups. A highly significant difference between the temperate and tropical species in the distribution of heterozygosity appears to be due to the smaller effective breeding population sizes in the case of the temperate species.

1. INTRODUCTION

An extensive array of gene frequency data for electrophoretic variants in natural populations of *Drosophila* is now available for comparison with alternative genetic models. In this paper, gene frequency distributions for nine species are summarized in terms of the relative contribution to heterozygosity of each of ten gene frequency classes, and the observed distributions compared initially with those predicted by models involving selective substitution of alleles, and by models postulating strictly neutral mutational variation. Neither class of model fits the data adequately, but a computer model involving mutants of very slight selective disadvantage is shown to provide a satisfactory approximation.

Broadly speaking there are three types of model which have been proposed to account for the observed genetic variation in natural populations of outbreeding species. The *balance hypothesis* suggests that current gene frequencies are primarily the result of natural selection, polymorphic loci displaying equilibrium allele frequencies maintained by the selective superiority of heterozygotes, by frequency-dependent selection, or by some other form of balancing selection. Genetic variation is therefore supposed to be adaptive, and deleterious mutants are found only at extremely low frequencies. Amino acid substitutions in a protein in the

course of evolution can be explained by changes in environmental or competitive conditions, leading to selection in favour of a particular homozygote at a previously polymorphic locus.

The *selective substitution hypothesis* maintains that all amino acid changes in a protein over evolutionary time are due to the selective replacement of one allele by another. Current polymorphisms are therefore considered as transient states in this process, with one homozygote showing a slight selective advantage over all other genotypes at the locus concerned. Disadvantageous mutants on the other hand would be expected to be found only at extremely low frequencies.

The *neutral hypothesis* proposes that a large proportion of the amino acid replacements in a protein in the course of evolution are in fact non-adaptive, due to the random fixation of alleles which are not subject to selective forces. Current polymorphisms are interpreted as transient states in this process of non-selective allelic substitution. Disadvantageous mutants are maintained at extremely low frequencies by selection.

It is not possible to test the applicability of the balance hypothesis by an examination of observed gene frequency distributions alone. The model can always be reconciled with the data by an appropriate choice of the mode of balancing selection and of the selection coefficients supposed to be maintaining the equilibrium, which will therefore usually be different for each locus examined. The hypothesis is however open to experimental test by perturbation of gene frequencies from their presumptive equilibrium values.

The aim of this paper is therefore to determine how adequately the simpler models involving selective substitution and neutral variation can account for the gene frequency distributions observed in species of *Drosophila*, and to indicate the simplest explanation for the discrepancies detected. The data used in this study are the same as those involved in an earlier analysis (Latter, 1976), but the distributions of heterozygosity are here presented in terms of ten gene frequency classes instead of five, and the analyses take account of the existence of geographic differentiation in the species surveyed. In addition, a separate analysis is presented for temperate and tropical species of *Drosophila*.

2. MODELS AND THEORY

Two basic models of allelic variation have been used. The infinite allele model assumes that mutation to novel and individually distinguishable alleles occurs with frequency μ per generation (Kimura & Crow, 1964). The charge class model, on the other hand, proposes that mutation occurs with frequency μ to novel alleles, of which a proportion α is not distinguishable electrophoretically from the parent allele, a proportion β gives rise to a polypeptide differing by one unit of electric charge from the parental form, and a proportion γ differs by two units of charge. Positive and negative changes in charge are assumed to be equally probable, and charge differences alone are supposed to determine relative electrophoretic mobility (Ohta & Kimura, 1973). It has been argued by Johnson (1974) that the infinite

allele and charge class models represent the two extremes of allelic non-identification when electrophoretic techniques are used.

For each of these two models of detectable genetic variation, the relative contribution of each gene frequency class to heterozygosity has been determined theoretically, or by computer simulation, either for a single panmictic diploid population of constant effective breeding size N, or for an island model involving k subpopulations of effective size N each exchanging a total of Nm individuals equally with the remaining $k-1$ subpopulations every generation (Latter, 1973). More elaborate models of geographic population structure have not been examined in this study, in view of the minor degree of population subdivision observed in the *Drosophila* species surveyed, and the slight effects on the distribution of heterozygosity found with the island model by comparison with a single panmictic population.

Only three types of natural selection are discussed in this paper, the models involving either neutral allelic variation, selectively advantageous alleles, or slightly deleterious mutations, together with combinations of two or all three of these categories. Neutral models assume that all genotypes at a locus are of equal viability and reproductive ability, or more precisely that selective differences are negligible by comparison with the reciprocal of the effective size of the breeding population. The basic theory for neutral models in the case of a single panmictic population is now well known. The theoretical equilibrium gene frequency distribution for the infinite alleles model in a finite population was derived by Wright (1966), and that for the charge class model with single step mutations by Kimura & Ohta (1975). Extension of these models to the case of a subdivided population for comparison with *Drosophila* electrophoretic surveys has been accomplished in this study by computer simulation.

The contribution to heterozygosity of an advantageous mutant between its occurrence and final loss or fixation in a population has been discussed by Maruyama (1972) and Yamazaki & Maruyama (1972). If the gene has a selective advantage s in heterozygotes and $2s$ in homozygotes, the frequency of heterozygotes expected in the population at a mean gene frequency g is the same for all values of q except those in the vicinity of unity, provided s is small and kNs is of the order of 20 or greater. This property is independent of the geographical structure of the population if the pattern of mating is locally random, and migration does not change the mean gene frequency of the whole operation.

Models involving selective substitution can most easily be examined in computer simulation studies by assuming that natural selection favours an optimal level of enzyme activity, the optimum changing directionally at a constant rate with time due to systematic changes in the environment. A convenient representative model is one in which fitness declines as the square of the deviation of enzyme activity from optimal, the spectrum of mutant effects being normally distributed with unit variance about the mean activity of the parental allele, and allelic effects on enzyme activity being additive (Latter, 1972, 1976). The relative importance of advantageous and disadvantageous mutants will then be determined primarily by

the rate of change in the position of the optimum, Δ_{opt}, on the scale of genotypic values, and the intensity of selection for the optimum, s. The fitness of a genotype with activity differing by d units from the optimum in any given generation is assumed to be $1-sd^2$.

The simplest model of slightly disadvantageous mutations assumes that each mutant has a selective disadvantage s relative to the parent allele from which it was derived, with all heterozygotes being exactly intermediate in fitness between the corresponding two homozygotes. If the mutational sequence is $A_1 \rightarrow A_2 \rightarrow A_3$, the relative fitness values of the heterozygotes are then $A_1A_2:A_1A_3:A_2A_3 = 1-s:1-2s:1-3s$, and those of the homozygotes are $A_1A_1:A_2A_2:A_3A_3 = 1:1-2s:1-4s$. The theory of Li (1978, 1979) deals with a related model involving a finite array of neutral and deleterious alleles in a panmictic population, with equal mutation rates from each allelic state to every other possible allele. Provided the deleterious alleles all remain rare in the population, Li's model is very similar to that outlined above. However, solutions have not yet been provided for the charge class model, nor for populations involving geographic differentiation. The required distributions of heterozygosity have therefore been approximated by computer simulation.

In the case of charge class models used in this study, selective effects are assumed to be associated with the individual genotype and not the charge class itself, charge differences *per se* having no effect on reproductive fitness. Electrophoretic classes will nevertheless usually differ in mean selective value, due to differences in the mutational history of the alleles in each class. King & Ohta (1975) have derived deterministic solutions for the expected distribution of electromorph frequencies at equilibrium in an infinite population, as a function of the ratio s/μ.

3. DATA AND ANALYSES

(i) *Data*

Data from the following nine species have been analysed: *D. bifasciata* (Saura, 1974), *D. equinoxialis* (Ayala *et al.* 1972*a*, 1974), *D. paulistorum* (Richmond, 1972), *D. pavani* (Kojima *et al.* 1972), *D. pseudoobscura* (Prakash, Lewontin & Hubby, 1969). *D. robusta* (Prakash, 1973), *D. subobscura* (Saura *et al.* 1973), *D. tropicalis* (Ayala *et al.* 1974) and *D. willistoni* (Ayala *et al.* 1971, 1972*b*, 1974). This survey of published electromorph frequencies has been restricted to those studies in which three or more separate localities have been sampled, and to those alleles reaching a frequency of 0·01 in at least one locality. The data for *D. pseudoobscura* from Bogota were not included in the analysis with those from North American localities. The survey has been restricted to enzyme polymorphisms, a total of 25 enzymes being involved in the 12 reported studies.

Gillespie & Kojima (1968) have pointed out that the enzymes assayed in electrophoretic studies can be divided into two groups differing in mean heterozygosity in *Drosophila* populations. Group I enzymes are those which take part in glycolysis, the citric acid cycle, or the hexose monophosphate shunt, or whose substrates are in one of these pathways (Kojima, Gillespie & Tobari, 1970): group

II enzymes are all those not included in the former category. In view of the consistently different levels of heterozygosity shown by group I and group II enzymes in all species involved in this study except *D. pavani*, these two sets of loci have been analysed separately in addition to the combined analysis. The enzymes are grouped as follows:

Group I. Aldolase, fumarase, glucose-6-phosphate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, α-glycerophosphate dehydrogenase, hexokinase, isocitrate dehydrogenase, malate dehydrogenase, malic enzyme, phosphoglucoisomerase, phosphoglucomutase and triosephosphate isomerase.

Group II. Acid phosphatase, adenylate kinase, alcohol dehydrogenase, aldehyde oxidase, alkaline phosphatase, amylase, esterase, glutamate oxaloacetate aminotransferase, leucine aminopeptidase, octanol dehydrogenase, peptidase, tetrazolium oxidase and xanthine dehydrogenase.

The data have also been subdivided into two groups on the basis of species range, the tropical species *D. equinoxialis*, *D. paulistorum*. *D. tropicalis* and *D. willistoni* having a geographic range lying almost completely between the Tropics of Cancer and Capricorn, and the temperate species *D. bifasciata*, *D. pavani*, *D. pseudoobscura*, *D. robusta* and *D. subobscura* being found almost entirely outside this range.

(ii) *The distribution of heterozygosity*

The following calculations have been performed separately for group I, group II and all enzymes combined, for each of the 12 surveys. Let q_{ij} denote the observed frequency of the ith allele at a given locus in the jth locality, and n_j the sample size in that locality. The mean frequency of the allele in the whole population, $\bar{q}_i$, has been calculated as

$$\bar{q}_i = \frac{1}{M}\sum_j n_j q_{ij}, \tag{1}$$

where $M = \sum_j n_j$. The contribution of the ith allele to heterozygosity, h_i, is given by

$$h_i = \frac{1}{M}\sum_j n_j q_{ij}(1-q_{ij}), \tag{2}$$

assuming random mating within each locality. The proportion of heterozygotes expected at this locus is then

$$H_w = \sum_i h_i = 1 - \frac{1}{M}\left(\sum_i \sum_j n_j q_{ij}^2\right). \tag{3}$$

The mean level of heterozygosity for the species is

$$\bar{H}_w = \frac{1}{l}\sum H_w, \tag{4}$$

where summation is over all l loci surveyed, including those found to be monomorphic.

The total contribution to heterozygosity of alleles at particular mean gene frequencies can also be evaluated to give a distribution of heterozygosity as a function of gene frequency. The individual values of h_i which are summed in equations (3) and (4) can be grouped into classes on the basis of the corresponding values of $\bar{q}_i$, and class means $\bar{H}_r(\bar{q})$ obtained by summing the values of h_i within

Table 1. *Gene frequency classes and the corresponding weighting factors used in deriving the distribution of heterozygosity*

	Class (r)									
	1	2	3	4	5	6	7	8	9	10
Gene frequency interval	0·001–0·050	0·051–0·100	0·101–0·200	0·201–0·300	0·301–0·500	0·501–0·700	0·701–0·800	0·801–0·900	0·901–0·950	0·951–0·999
Width of interval	0·05	0·05	0·10	0·10	0·20	0·20	0·10	0·10	0·05	0·05
Weighting factor W_r	20	20	10	10	5	5	10	10	20	20

https://doi.org/10.1017/S0016672300020498 Published online by Cambridge University Press

classes and dividing by l, the number of loci surveyed. In calculating the class means in this study, each value of h_i has been given weight $(a-1)/a$ where a is the number of alleles segregating at the locus concerned, since allele frequencies at each locus are linearly dependent.

Table 1 sets out the class intervals used in this analysis, the width of each interval, and the weighting factor W_r, needed to adjust for the unequal class intervals used in the grouping of mean gene frequencies. The relative contribution to heterozygosity, C_r, of each gene frequency class is then given by

$$C_r = \frac{H_r(\bar{q})}{\Sigma H_r(\bar{q})} . W_r, \tag{5}$$

where the summation in the denominator is over all gene frequency classes.

The values of C_r from different studies have been combined by taking a weighted mean for each class, with weights proportional to the number of independent alleles, i.e. the total number of alleles minus the number of loci. In the case of the superspecies *D. paulistorum*, the values of C_r were calculated separately for each semi-species and combined by the same weighting procedure. The values of $\bar{H}_w$ from different studies have similarly been combined by taking a weighted mean, with weights proportional to the number of loci (both monomorphic and polymorphic) upon which the estimate is based.

The standard error of a weighted mean

$$\bar{x} = \left(\sum_{i=1}^{n} w_i x_i\right) \Big/ \sum_{j=1}^{n} w_i$$

has been calculated as

$$\text{S.E.}(\bar{x}) = (s^2/\Sigma w_i)^{\frac{1}{2}}, \tag{6}$$

where

$$s^2 = \frac{1}{n-1}[\Sigma w_i x_i^2 - \bar{x}^2 \Sigma w_i]. \tag{7}$$

This formula is strictly appropriate only when the values of x_i involved have the same expectation.

(iii) *Population differentiation*

The measure of genetic differentiation adoped in this study is the parameter ϕ^* defined by Latter (1973), which relates the level of heterozygosity within subpopulations, $\bar{H}_w$, to that which would be observed in hybrids between subpopulations, $\bar{H}_b$. The possible range of values of ϕ^* extends from zero when all subpopulations are identical, to unity when all subpopulations are homozygous at all loci but not genetically identical.

If q_{ij} denotes the frequency of the ith allele in the jth locality at a particular locus as before, the expected frequency of heterozygotes in crosses between subpopulations has been calculated as

$$H_b = 1 - \frac{1}{L} \sum_{i} \sum_{j \neq j'} n_j n_{j'} q_{ij} q_{ij'} \tag{8}$$

corresponding to equation 3, where $L = \sum_{j \neq j'} n_j n_{j'}$. The mean over all loci is then

$$\bar{H}_b = \frac{1}{l} \Sigma H_b \tag{9}$$

and ϕ^*, the measure of population differentiation, is

$$\phi^* = 1 - \bar{H}_w / \bar{H}_b. \tag{10}$$

For ease of calculation, the following formulae are to be preferred in evaluating H_b. The value of L is given by

$$L = M^2 - \sum_j n_j^2$$

and H_b by

$$H_b = 1 - \frac{1}{L} \sum_i \left(M^2 \bar{q}_i^2 - \sum_j n_j^2 q_{ij}^2 \right). \tag{11}$$

Estimates of ϕ^* from different species or semispecies have been combined by the use of a weighted mean, with weights proportional to the number of loci contributing to each estimate.

4. RESULTS

(i) *Heterozygosity and population differentiation*

A summary of mean levels of heterozygosity for group I, group II and all enzymes combined is presented in Table 2, together with the measure of population differentiation, ϕ^*. Also given in the second column are species abbreviations which will be used in subsequent tables. The levels of heterozygosity for group I enzymes are appreciably lower than those for group II enzymes in all species except *D. pavani*, the means over all species being 0·102 and 0·226 respectively.

The measure of genetic divergence between localities, ϕ^*, could not be evaluated from the incomplete data for *D. willistoni* reported by Ayala *et al.* (1971). The estimates for other species range from 0·023 to 0·133 with an overall mean of $0{\cdot}075 \pm 0{\cdot}010$, with no relationship discernible between $\bar{H}_w$ and ϕ^*. The estimates of ϕ^* for individual species based on group I and group II enzymes are in each case very similar, the weighted mean over species from group I enzymes being $0{\cdot}061 \pm 0{\cdot}013$, and that for group II enzymes being $0{\cdot}079 \pm 0{\cdot}012$, the difference being non-significant.

The five temperate species *D. bifasciata*, *D. pavani*, *D. pseudoobscura*, *D. robusta* and *D. subobscura* can be seen from Table 3 to differ only slightly in mean heterozygosity from the remaining tropical species, though the agreement must be considered somewhat fortuitous in view of the wide range of heterozygosity levels in the temperate group. It will be shown in the next section that the two groups of species nevertheless differ appreciably in the distribution of heterozygosity as a function of mean gene frequency. Table 3 also shows the extent of population differentiation throughout the species range to be very similar for temperate and tropical species, corresponding to a value of Nm for the island model of the order of 2–3 depending on the number of subpopulations involved (Latter, 1973).

Table 2. *Mean levels of heterozygosity* ($\bar{H}_w$) *and population differentiation* (ϕ^*)

Species	Abbr.	No. loci	Heterozygosity ($\bar{H}_w$) Group I	Group II	Total	ϕ^*
D. bifasciata	bif.	21	0·123	0·325	0·238	0·103
D. equinoxialis	equi.	23	0·133	0·270	0·210	0·063
D. equinoxialis	equi.	30	0·135	0·197	0·170	0·094
D. paulistorum	paul.	13	0·008	0·263	0·204	0·060
D. pavani	pav.	16	0·171	0·170	0·171	0·133
D. pseudoobscura	pseu.	13	0·037	0·182	0·160	0·023
D. robusta	rob.	28	0·038	0·165	0·110	0·071
D. subobscura	sub.	18	0·122	0·280	0·210	0·064
D. tropicalis	trop.	30	0·083	0·191	0·141	0·122
D. willistoni	will.	24	0·073	0·191	0·142	—
D. willistoni	will.	25	0·124	0·268	0·199	0·047
D. willistoni	will.	31	0·099	0·243	0·179	0·033
Means		22·7	0·102 ±0·012	0·226 ±0·015	0·174 ±0·011	0·075 ±0·010

Table 3. *Mean population parameters for temperate and tropical species of* Drosophila

Species group	Heterozygosity ($\bar{H}_w$) Group I	Group II	Total	ϕ^*
Temperate species	0·100±0·025	0·222±0·033	0·174±0·024	0·080±0·017
Tropical species	0·104±0·012	0·228±0·014	0·175±0·010	0·072±0·014

(ii) *Distribution of Heterozygosity*

The values of C_r given by equation (5) for each species are presented in Table 4, together with the number of independent alleles contributing to the estimated distribution of heterozygosity. To provide a single parameter upon which comparisons between species may be based, the proportion of heterozygosity due to alleles with central mean frequencies in the range 0·2–0·8, i.e.

$$P_c = \sum_{r=4}^{7} C_r / W_r \tag{12}$$

is also given for each survey in the final column of the table. Where more than one survey has been made of the same species, as in the case of *D. equinoxialis* and *D. willistoni*, the independent values of P_c are in good agreement, but the differences between species are quite considerable, ranging from 0·31 for *D. willistoni* to 0·70 for *D. bifasciata*.

The distributions of heterozygosity summarized in Table 5 show that the temperate species have a higher mean value of P_c than the tropical species for both group I and group II enzymes, the means over all enzymes of 0·597±0·041 and 0·361±0·024 being statistically significant at the 0·001 level of probability. The

Table 4. *Distribution of heterozygosity over ten gene frequency classes (C_r), and proportion of heterozygosity due to alleles with mean frequencies in the range 0·2–0·8 (P_c)*

Species	No. indep. alleles	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	P_c
bif.	54	1·60	1·78	0·87	1·26	1·43	1·08	0·74	0·00	0·56	0·30	0·702
equi.	58	2·28	1·74	1·50	0·28	0·94	0·52	0·63	1·38	1·70	0·88	0·382
equi.	107	3·98	2·14	0·67	0·55	1·00	0·64	0·28	1·06	0·82	1·38	0·410
paul.	40	2·60	1·18	1·54	0·37	0·89	1·04	0·87	0·38	0·52	1·70	0·509
pav.	12	0·64	0·56	2·01	1·31	0·76	1·58	0·00	1·24	0·00	0·34	0·599
pseu.	30	3·80	3·22	0·00	2·44	0·53	0·55	0·77	0·00	0·82	1·42	0·537
rob.	37	3·18	1·08	1·26	0·54	0·68	0·69	1·45	0·61	1·48	1·06	0·472
sub.	54	2·50	2·66	0·00	1·62	0·87	1·36	0·00	0·55	0·74	0·80	0·609
trop.	78	3·34	2·42	1·53	0·39	0·62	0·56	0·81	0·90	1·14	1·14	0·356
will.	31	2·78	2·04	1·59	0·00	0·94	0·64	0·00	1·23	2·98	0·22	0·317
will.	67	3·94	2·36	1·09	1·12	0·00	0·58	0·71	1·29	1·80	1·16	0·299
will.	134	3·40	3·20	1·39	0·62	0·54	0·72	0·00	0·83	1·64	1·06	0·313

Table 5. *Average distributions of heterozygosity for temperate and tropical species based on group I enzymes, group II enzymes and all enzymes combined*

Enzymes	Species	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	P_c
Group I	*Temperate*	3·35	1·48	0·68	1·50	0·34	1·16	—	1·12	1·20	1·38	0·449
		±0·88	±0·94	±0·34	±0·36	±0·23	±0·35		±0·52	±0·38	±0·84	±0·135
	Tropical	4·95	2·63	0·69	0·16	0·23	0·24	0·55	1·46	2·89	1·90	0·166
		±0·51	±0·61	±0·34	±0·13	±0·14	±0·14	±0·22	±0·28	±0·43	±0·65	±0·069
	All species	4·54	2·34	0·69	0·50	0·26	0·47	0·41	1·38	2·46	1·77	0·238
		±0·47	±0·51	±0·29	±0·22	±0·11	±0·18	±0·16	±0·24	±0·37	±0·49	±0·071
Group II	*Temperate*	2·05	2·32	0·60	1·37	1·21	0·98	0·92	—	0·61	0·49	0·666
		±0·51	±0·43	±0·30	±0·35	±0·23	±0·25	±0·35		±0·17	±0·15	±0·036
	Tropical	2·54	2·23	1·55	0·74	0·90	0·87	0·34	0·76	0·65	0·73	0·462
		±0·32	±0·19	±0·19	±0·21	±0·14	±0·11	±0·17	±0·21	±0·24	±0·15	±0·023
	All species	2·40	2·25	1·29	0·91	0·98	0·90	0·50	0·55	0·64	0·67	0·517
		±0·26	±0·18	±0·20	±0·19	±0·12	±0·11	±0·17	±0·17	±0·16	±0·11	±0·033
All enzymes	*Temperate*	2·47	2·05	0·63	1·41	0·93	1·03	0·62	0·36	0·80	0·78	0·597
		±0·45	±0·41	±0·32	±0·29	±0·17	±0·17	±0·26	±0·18	±0·19	±0·20	±0·041
	Tropical	3·36	2·36	1·26	0·54	0·67	0·66	0·41	1·00	1·42	1·13	0·361
		±0·23	±0·24	±0·14	±0·11	±0·13	±0·05	±0·14	±0·11	±0·24	±0·13	±0·024
	All species	3·12	2·28	1·09	0·77	0·74	0·76	0·47	0·83	1·25	1·04	0·424
		±0·24	±0·20	±0·16	±0·16	±0·10	±0·08	±0·12	±0·12	±0·18	±0·11	±0·037

Table 6. *Theoretical distributions of heterozygosity for the neutral model in a single panmictic population at equilibrium*

Model	$\bar{H}_w$	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	P_c
Infinite	0·102	1·09	1·11	1·09	1·08	1·05	1·00	0·95	0·90	0·83	0·70	0·614
allele	0·174	1·18	1·19	1·17	1·14	1·09	1·00	0·90	0·81	0·70	0·53	0·622
	0·226	1·26	1·27	1·23	1·19	1·11	0·99	0·86	0·74	0·60	0·42	0·625
Charge	0·102	0·93	1·02	1·05	1·06	1·06	1·03	0·99	0·94	0·87	0·73	0·624
class*	0·174	0·90	1·03	1·09	1·12	1·11	1·06	0·97	0·87	0·75	0·55	0·643
	0·226	0·87	1·05	1·13	1·16	1·16	1·07	0·95	0·81	0·65	0·43	0·657

* Single step charge class model of Kimura & Ohta (1975).

unweighted means are very similar in magnitude, being $0{\cdot}584 \pm 0{\cdot}038$ and $0{\cdot}369 \pm 0{\cdot}028$ respectively, which also differ significantly at the 0·001 level. Many of the individual estimates of C_r in Table 5 have large standard errors, particularly those for group I enzymes in temperate species, due to the small number of group I enzymes studied in three of the five species. However, the distributions of heterozygosity for temperate and tropical species based on all enzymes are distinctively different, with far higher frequencies of alleles in the first three and last three classes in the tropical group. This is in spite of the fact that the two groups of species show very similar mean levels of heterozygosity (Table 3).

The contrast between the distributions of heterozygosity for group I and group II enzymes in Table 5 is equally striking, the mean values of P_c over all species being $0{\cdot}238 \pm 0{\cdot}071$ for group I enzymes, and $0{\cdot}517 \pm 0{\cdot}033$ for group II enzymes. In this instance the difference in the distribution of heterozygosity clearly parallels the difference in mean heterozygosity shown in Table 3.

(iii) *Single population models*

In spite of the evidence of heterogeneity afforded by Table 5, it is convenient to consider first the extent of agreement between simple theoretical models and distribution of heterozygosity observed for *all enzymes* and *all species* (Table 5, last row). Comparison with the predictions of Wright (1966) for the neutral infinite allele model, and Kimura & Ohta (1975) for the neutral charge class model, at equilibrium in a single panmictic population (Table 6, $\bar{H}_w = 0{\cdot}174$) shows highly significant discrepancies in each case. The observed mean value of $P_c = 0{\cdot}424 \pm 0{\cdot}037$ is significantly different at the 0·001 level from those predicted by the two neutral models (0·622 and 0·643 respectively).

Detailed comparisons of the distributions of Table 5 with the corresponding distributions of Table 6 ($\bar{H}_w = 0{\cdot}102$ for group I, 0·226 for group II) show an excess of heterozygosity contributed by alleles in the first two classes ($\bar{q} \leqslant 0{\cdot}100$) in all classifications of the *Drosophila* data. The departures are statistically highly significant for both groups of enzymes in the tropical species, but are not significant for either group in the case of the temperate species. The distribution for all enzymes combined in the temperate species shows an excess of heterozygosity in class 1 which is on the borderline of significance when compared with the charge class model, but the observed value of P_c is not significantly different from that predicted by either model.

Table 7 summarizes the distributions of heterozygosity observed in computer populations for the model of selective allele substitution in response to a changing optimum. Comparisons with Table 6 reveal only minor differences between the neutral and selective substitution models, the latter showing an increase in the frequency of rare alleles at the expense of those at intermediate frequencies due to selection for the current optimum. The distributions of heterozygosity for the tropical *Drosophila* species show highly significant departures from the corresponding distributions of Table 7. The temperate species, on the other hand, differ significantly only in respect of the absence of observed gene frequencies in the

Table 7. *Distributions of heterozygosity with natural selection for a directionally changing optimum in a single panmictic population**

Model	$N\mu$	Δ_{opt}	$\bar{H}_w$	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	P_c
Charge class	0·3	1×10^{-3}	0·196 ±0·008	1·20 ±0·06	1·22 ±0·05	1·14 ±0·04	1·08 ±0·02	1·00 ±0·06	0·96 ±0·03	0·99 ±0·05	0·96 ±0·05	0·85 ±0·06	0·56 ±0·04	0·599 ±0·017
Charge class	0·3	2×10^{-4}	0·190 ±0·011	1·21 ±0·07	1·24 ±0·08	1·18 ±0·07	1·10 ±0·03	1·00 ±0·08	0·96 ±0·06	0·92 ±0·04	0·93 ±0·08	0·82 ±0·07	0·61 ±0·05	0·594 ±0·026
Infinite allele	0·1	1×10^{-3}	0·236 ±0·011	1·56 ±0·07	1·38 ±0·04	1·29 ±0·03	1·17 ±0·04	1·08 ±0·05	0·90 ±0·03	0·84 ±0·03	0·77 ±0·04	0·61 ±0·04	0·41 ±0·04	0·596 ±0·012

* *Each distribution based on* 8 *replicates of duration* 50000 *generations, with* $N = 500$, $Ns = 2$. Charge class models based on $\alpha = 0{\cdot}66$, $\beta = 0{\cdot}32$, $\gamma = 0{\cdot}02$ (Latter, 1975).

Table 8. *Distributions of heterozygosity with slightly disadvantageous mutations, in a single panmictic population assuming the infinite allele model**

Ns	Reps	$\bar{H}_w$	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	P_c
1	10	0·168 ±0·009	2·17 ±0·11	1·82 ±0·06	1·35 ±0·04	0·95 ±0·03	0·68 ±0·06	0·70 ±0·03	0·91 ±0·04	1·10 ±0·04	1·12 ±0·07	0·76 ±0·08	0·462 ±0·015
2	10	0·180 ±0·005	3·61 ±0·06	2·45 ±0·06	1·34 ±0·03	0·57 ±0·03	0·25 ±0·01	0·48 ±0·03	1·00 ±0·04	1·42 ±0·04	1·50 ±0·07	0·84 ±0·04	0·304 ±0·008
5	3	0·182 ±0·005	6·64 ±0·16	2·50 ±0·04	0·63 ±0·06	0·08 ±0·02	0·01 ±0·01	0·12 ±0·03	0·76 ±0·06	2·09 ±0·06	2·42 ±0·09	0·83 ±0·05	0·108 ±0·009
10	3	0·181 ±0·005	9·11 ±0·13	1·21 ±0·10	0·10 ±0·03	—	—	0·02 ±0·01	0·33 ±0·08	2·41 ±0·03	3·32 ±0·21	0·60 ±0·04	0·037 ±0·008

* *Based on replicates each of duration* 50000 *generations, with* $N = 500$, $s/\mu = 10$.

interval 0·8–0·9 for group II enzymes. This is without doubt a sampling accident due to the small number of species so far studied.

We turn finally to models involving slightly disadvantageous mutations, in an attempt to account for the excess contribution of rare alleles to the observed distribution of heterozygosity in *Drosophila* populations. The distributions presented in Table 8 for a single panmictic population in equilibrium are based on the infinite allele model with the selective disadvantage of individual mutations (s) equal to 10 times the mutation rate (μ). This ratio is expected to give a mean frequency of heterozygotes close to that observed for all enzymes combined in both temperate and tropical species of *Drosophila* (Table 3). None of the distributions in Table 8 fits the *Drosophila* data completely satisfactorily, but $Ns = 1$ and $Ns = 2$ give distributions which are clearly very similar to those observed for the temperate and tropical species respectively (Table 5). It may be noted that the chance fixation of deleterious mutants was detected only in the computer populations with $Ns = 1$, a total of 13 mutants being fixed in 500000 generations, a rate of $0{\cdot}013 \pm 0{\cdot}003$ mutants per N generations.

(iv) *Models with population differentiation*

It has previously been shown that the distributions of heterozygosity for group I and group II enzymes in temperate species do not depart significantly from those predicted for neutral models in a single panmictic population. The same is true of populations composed of subpopulations of equal size, exchanging migrants at a constant rate corresponding to $Nm = 2$. The first two populations listed in Table 9, for example, are virtually identical to the charge class models of Table 6 with the same mean levels of heterozygosity, and the same is true of neutral infinite allele models.

Table 10 gives parameter combinations and observed summary statistics for theoretical models involving neutral or disadvantageous mutations which approximate the *Drosophila* data satisfactorily. The best fit to the data for group I enzymes in the temperate species is given by the infinite allele model with $kNs = 1{\cdot}5$ (Fig. 1*a*), but there are marked irregularities in the *Drosophila* data due to the small number of group I enzymes included in the surveys of three of the species. The infinite allele model with $kNs = 3$ has been found to give the most satisfactory fit to the data for group I enzymes in the tropical species (Fig. 1*b*), and the charge class model with $kNs = 4$ does not differ significantly in any respect from the data.

Fig. 2 shows the distributions of heterozygosity for group II enzymes and the closest fit to the data for the temperate and tropical species. In the case of the temperate species the neutral infinite allele model provides the best fit (Table 10). A charge class model in which 55 % of loci are subject to selection with $kNs = 4$, and 45 % of loci are neutral (Table 10) comes close to fitting the data, but there is a deficiency of heterozygotes due to alleles at intermediate frequencies in the model ($P_c = 0{\cdot}562$) by comparison with the data ($P_c = 0{\cdot}666$) which is statistically significant at the 0·05 level.

The only model which has been found to fit the group II enzyme data in the

Table 9. *Neutral charge class models with migration between subpopulations of equal size*
($N = 20$, $k = 5$, $Nm = 2$, $\alpha = 0{\cdot}66$, $\beta = 0{\cdot}32$, $\gamma = 0{\cdot}02$)

$kN\mu$	$\bar{H}_w$	ϕ^*	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	P_c
0·08	0·098	0·097	0·89	1·05	1·02	1·04	1·09	1·06	0·96	0·90	0·87	0·76	0·630
	±0·005	±0·001	±0·04	±0·04	±0·04	±0·03	±0·03	±0·04	±0·03	±0·03	±0·05	±0·05	±0·010
0·24	0·218	0·097	0·85	1·10	1·17	1·15	1·16	1·08	0·93	0·78	0·63	0·42	0·654
	±0·005	±0·001	±0·02	±0·02	±0·03	±0·02	±0·03	±0·03	±0·01	±0·03	±0·03	±0·02	±0·008
0·55	0·356	0·096	0·83	1·16	1·30	1·36	1·34	1·07	0·77	0·52	0·32	0·14	0·695
	±0·005	±0·001	±0·02	±0·04	±0·02	±0·02	±0·03	±0·03	±0·02	±0·04	±0·02	±0·01	±0·008

https://doi.org/10.1017/S0016672300020498 Published online by Cambridge University Press

Table 10. *Models involving disadvantageous or neutral mutations with migration between subpopulations of equal size*

($N = 20$, $k = 5$, $Nm = 2$.)

Model	kNs	$kN\mu$	$\bar{H}_w$	ϕ^*	P_c	Drosophila data
Infinite allele	0	0·08	0·239 ±0·007	0·097 ±0·001	0·634 ±0·007	Temperate group II
	1·5	0·08	0·098 ±0·002	0·093 ±0·001	0·338 ±0·014	Temperate group I
	3	0·14	0·084 ±0·002	0·086 ±0·001	0·127 ±0·007	Tropical group I
	3, 0†	0·14	0·198 ±0·006	0·096 ±0·001	0·503 ±0·015	Tropical group II
Charge class	4	0·55	0·086 ±0·002	0·083 ±0·001	0·107 ±0·007	Tropical group I
	4, 0†	0·55	0·208 ±0·002	0·089 ±0·001	0·562 ±0·006	Temperate group II

† Combined distributions for 55% of loci subject to selection and 45% neutral loci.

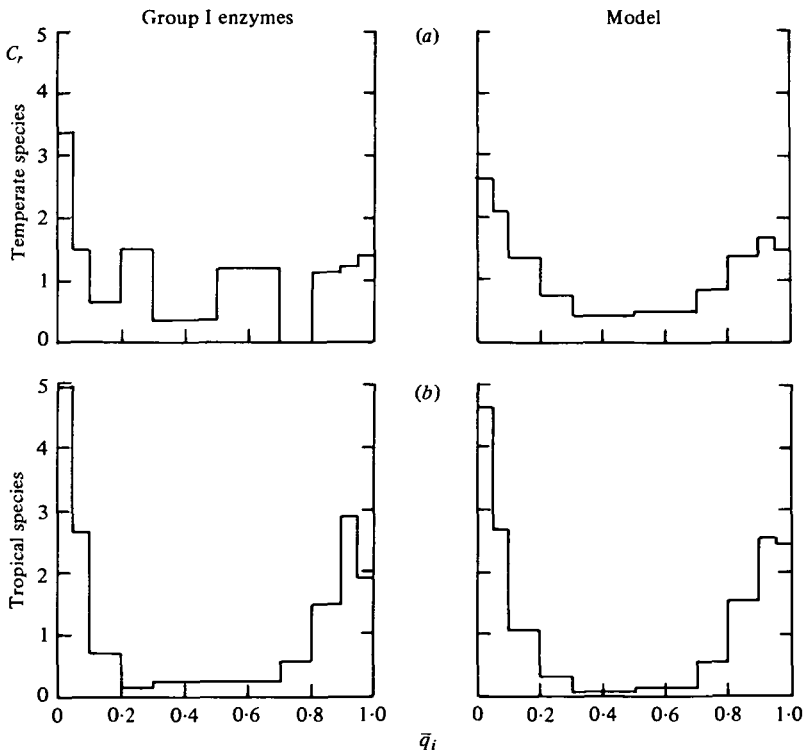

Fig. 1. Distributions of heterozygosity (C_r) as a function of mean gene frequency ($\bar{q}_i$) for group I enzymes in *Drosophila* species, and the corresponding models of best fit. Both theoretical distributions are given by the infinite allele model, with slightly disadvantageous mutations and migration between subpopulations of equal size. (*a*) Temperate species: $kNs = 1{\cdot}5$, $kN\mu = 0{\cdot}08$ (Table 10). (*b*) Tropical species: $kNs = 3$, $kN\mu = 0{\cdot}14$ (Table 10).

tropical species is an infinite allele model in which 55 % of loci are subject to selection with $kNs = 3$, and 45 % of loci are neutral (Table 10). The neutral loci in the model correspond to such enzymes as xanthine dehydrogenase, aldehyde oxidase, adenylate kinase, some of the esterases, and other highly variable loci. The following statistics summarize the properties of neutral infinite allele loci with $kN\mu = 0{\cdot}14$, $Nm = 2$: $\bar{H}_w = 0{\cdot}338 \pm 0{\cdot}013$, $\phi^* = 0{\cdot}097 \pm 0{\cdot}001$, $P_c = 0{\cdot}617 \pm 0{\cdot}015$.

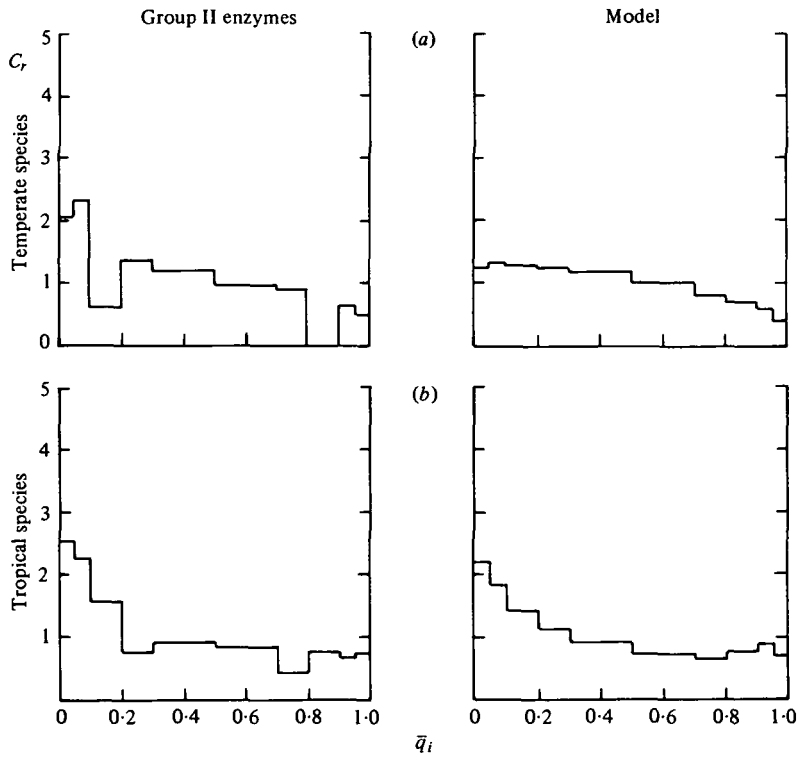

Fig. 2. Distributions of heterozygosity for group II enzymes and the models of best fit to the data. Both theoretical distributions are given by the infinite allele model, with migration between subpopulations of equal size. (*a*) Temperate species: $kNs = 0$, $kN\mu = 0{\cdot}08$ (Table 10). (*b*) Tropical species: $kNs = 3$ (55 % loci) and $kNs = 0$ (45 % loci), $kN\mu = 0{\cdot}14$ (Table 10).

5. DISCUSSION

The objective of this analysis has been the identification of the simplest models capable of explaining the observed distribution of heterozygosity in a number of species of *Drosophila*. It has been shown that the neutral model proposed by Kimura (1968, 1969), and modified by Ohta (1974, 1976) to include the accumulation of slightly disadvantageous mutations, is capable of explaining all features of the data. The models depicted in Figs. 1 and 2 involve only neutral loci and those subject to selection in favour of the parental allele at the expense of newly arisen mutations.

The highest intensity of selection involved in the theoretical distributions of best fit to the data corresponds to $kNs = 3$ for group I enzymes in the tropical *Drosophila* species (Fig. 1*b*), where k denotes the number of panmictic subpopulations of breeding size N, and s is the selective disadvantage of new mutants as heterozygotes. This intensity of selection is sufficient to ensure that only a minute fraction of the observed variation will ever be involved in the process of gene substitution, unless drastic bottlenecks in species population size occur.

A particularly striking feature of the *Drosophila* data is the contrast between the distributions of heterozygosity displayed by group I and group II enzymes, associated with a marked difference in the mean frequency of heterozygotes (Tables 3, 5). The distributions plotted in Figs. 1 and 2 clearly indicate that this difference can be satisfactorily explained by a higher mean intensity of selection against mutational variants in the case of group I enzymes. The difference between the two groups of enzymes cannot be explained on the basis of inherent differences in mutation rate alone, in view of the pronounced departure of the distributions for tropical group I and group II enzymes from expectations based on the neutral models (Tables 5, 6).

The process of model fitting for the tropical species has led us to propose two subgroups within the group II category, one composed of effectively neutral loci with mean heterozygosity of the order of $\overline{H}_w = 0{\cdot}34$, and the other made up of loci similar to those of group I (Table 10, Fig. 2*b*). All loci nevertheless are assumed to have the same intrinsic mutation rate, specified by $kN\mu = 0{\cdot}14$ for both group I and group II loci (Figs. 1*b*, 2*b*). The effectively neutral loci in group II include xanthine dehydrogenase and aldehyde oxidase, which are known to be highly variable in many *Drosophila* species. It may be concluded that the differences in the frequency of heterozygotes among groups of loci can be explained satisfactorily by differences in the intensity of selection against disadvantageous mutations. A similar conclusion has been reached by Koehn & Eanes (1977) from a study of variation in molecular subunit size and genetic variability in *Drosophila* populations, *viz.* that differences in *effective* mutation rates are due in part to the selective disadvantage of mutants which alter the quaternary structure of the functional enzyme molecule.

Another notable feature of the *Drosophila* data analyzed in this paper is the highly significant difference between the temperate and tropical species in the distribution of heterozygosity as a function of mean gene frequency, despite the similarity in the mean frequency of heterozygotes in the two groups of species (Tables 3, 5). The infinite allele theory of deleterious mutations in finite populations leads to the expectation that the ratio s/μ is of prime importance in determining the mean level of heterozygosity at equilibrium, and that kNs is the parameter which influences the shape of the gene frequency distribution (Table 8). The difference between the temperate and tropical species is therefore most readily understood as a consequence of the smaller effective breeding population sizes in the case of the temperate species, leaving the mean ratio s/μ unaltered, but reducing kNs and $kN\mu$ proportionally. The distributions for group I enzymes in

Fig. 1 suggest that effective population sizes for the tropical species are of the order of twice those of the temperate species, but the ratio could be considerably greater. Additional survey information is required before more precise estimates can be made.

REFERENCES

AYALA, F. J., POWELL, J. R. & DOBZHANSKY, TH. (1971). Polymorphisms in continental and island populations of *Drosophila willistoni*. *Proceedings of the National Academy of Sciences, U.S.A.* **68**, 2480–2483.

AYALA, F. J., POWELL, J. R. & TRACEY, M. L. (1972*a*). Enzyme variability in the *Drosophila willistoni* group. V. Genic variation in natural populations of *Drosophila equinoxialis*. *Genetical Research* **20**, 19–42.

AYALA, F. J., POWELL, J. R., TRACEY, M. L., MOURAO, C. A. & PEREZ-SALAS, S. (1972*b*). Enzyme variability in the *Drosophila willistoni* group. IV. Genic variation in natural populations of *Drosophila willistoni*. *Genetics* **70**, 113–139.

AYALA, F. J., TRACEY, M. L., BARR, L. G., MCDONALD, J. F. & PEREZ-SALAS, S. (1974). Genetic variation in natural populations of five *Drosophila* species and the hypothesis of the selective neutrality of protein polymorphisms. *Genetics* **77**, 343–384.

GILLESPIE, J. H. & KOJIMA, K. (1968). The degree of polymorphism in enzymes involved in energy production compared to that in nonspecific enzymes in two *Drosophila ananassae* populations. *Proceedings of the National Academy of Sciences, U.S.A.* **61**, 582–585.

JOHNSON, G. B. (1974). On the estimation of effective numbers of alleles from electrophoretic data. *Genetics* **78**, 771–776.

KIMURA, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

KIMURA, M. (1969). The rate of molecular evolution considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences, U.S.A.* **63**, 1181–1188.

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

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

KING, J. L. & OHTA, T. (1975). Polyallelic mutational equilibria. *Genetics* **79**, 681–691.

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

KOJIMA, K., GILLESPIE, J. H. & TOBARI, Y. N. (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. *Biochemical Genetics* **4**, 627–637.

KOJIMA, K., SMOUSE, P., YANG, S., NAIR, P. S. & BRNCIC, D. (1972). Isozyme frequency patterns in *Drosophila pavani* associated with geographical and seasonal variables. *Genetics* **72**, 721–731.

LATTER, B. D. H. (1972). Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. *Genetics* **70**, 475–490.

LATTER, B. D. H. (1973). The island model of population differentiation: a general solution. *Genetics* **73**, 147–157.

LATTER, B. D. H. (1975). Influence of selection pressures on enzyme polymorphisms in *Drosophila*. *Nature* **257**, 590–592.

LATTER, B. D. H. (1976). The intensity of selection for electrophoretic variants in natural population of *Drosophila*. In *Population Genetics and Ecology* (ed. S. Karlin and E. Nevo), pp. 391–410.

LI, W.-H. (1978). Maintenance of genetic variability under the joint effect of mutation, selection and random drift. *Genetics* **90**, 349–382.

LI, W.-H. (1979). Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. *Genetics* **92**, 647–667.

MARUYAMA, T. (1972). Some invariant properties of a geographically structured finite population: distribution of heterozygotes under irreversible mutation. *Genetical Research* **20**, 141–149.

OHTA, T. (1974). Mutational pressure as the main cause of molecular evolution and polymorphisms. *Nature* **252**, 351–354.

OHTA, T. (1976). Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theoretical Population Biology* **10**, 254–275.

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.

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

PRAKASH, S., LEWONTIN, R. C. & HUBBY, J. L. (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.

RICHMOND, R. C. (1972). Enzyme variability in the *Drosophila willistoni* group. III. Amounts of variability in the superspecies *D. paulistorum*. *Genetics* **70**, 87–112.

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

SAURA, A., LAKOVAARA, S., LOKKI, J. & LANKINEN, P. (1973). Genic variation in central and marginal populations of *Drosophila subobscura*. *Hereditas* **75**, 33–46.

WRIGHT, S. (1966). Polyallelic random drift in relation to evolution. *Proceedings of the National Academy of Sciences, U.S.A.* **55**, 1074–1081.

YAMAZAKI, T. & MARUYAMA, T. (1972). Evidence for the neutral hypothesis of protein polymorphism. *Science* **178**, 56–58.