A STUDY OF LINKAGE DISEQUILIBRIUM IN POPULATIONS OF *DROSOPHILA MELANOGASTER*

BRIAN CHARLESWORTH AND DEBORAH CHARLESWORTH

Department of Genetics, University of Liverpool, Liverpool L69 3BX, England

Manuscript received July 18, 1972

ABSTRACT

This paper presents the results of a study of linkage disequilibrium between five polymorphic enzyme genes located on chromosome 3 of *D. melanogaster*. Three sets of chromosomes were examined: two represented samples from successive years of the same natural population, and one came from a large laboratory population. Out of the thirty possible tests for linkage disequilibrium between pairs of loci, two were significant at the 5% level and two at the 1% level. This result cannot reasonably be ascribed to chance alone. The pairs of loci that had a significant correlation in one sample had higher than average correlations in the other samples (though not necessarily in the same direction); this effect was highly significant statistically. There was no tendency for the high correlations to be associated with tightness of linkage between the loci concerned. All five loci were involved in at least one significant effect. It was concluded that these results are difficult to explain on the neutral allele theory of protein polymorphism, but are consistent with the concept of selective control of allele frequencies.

UNTIL recently, most studies of protein polymorphisms in natural populations have been concerned with determining gene frequencies for one or more loci, and have not usually considered the problem of whether or not alleles at different loci are correlated in frequency (linkage disequilibrium). Some investigations of associations between loci in chromosomes from natural population have, however, been reported. KOJIMA, GILLESPIE and TOBARI (1970) studied two second chromosome and three third chromosome enzymes of *Drosophila melanogaster*. They found highly significant disequilibrium between one locus on each chromosome and inversions on the same chromosome. MUKAI, METTLER and CHIGUSA (1970; 1971) studied three loci and two inversions on the second chromosome of *D. melanogaster*. Significant associations were found between the inversions and two of the loci, but there was no definite evidence for associations between the loci themselves. PRAKASH and LEWONTIN (1968; 1971) found very strong associations between third-chromosome inversions and third-chromosome genes in *D. pseudoobscura* and *D. persimilis*. These associations were, in most cases, consistent from population to population and even across the species barrier. NAIR and BRNCIC (1971) have reported an association between genes and inversions in *D. pavani*. It does not appear, however, that cases of linkage disequilibrium between enzyme loci within a particular gene arrangement have been reported up to now, except for the case of transient linkage disequilibrium in *D. melanogaster* reported by O'BRIEN and MACINTYRE (1971).

There are two reasons why studies of associations between genes in natural populations are of interest. The first is that, on theoretical grounds, FRANKLIN and LEWONTIN (1970) have proposed that non-random associations between selectively-maintained linked genes may be the rule rather than the exception. The second reason is that it is increasingly apparent that data on the geographical distribution of allele frequencies cannot alone provide an answer to the question of whether protein polymorphisms are maintained by selection or are neutral, in the absence of more detailed information about gene-flow and population structure than is likely to be obtainable by present-day methods. Findings of consistent linkage disequilibrium between enzyme loci or enzyme and inversions provide strong evidence that protein polymorphisms are selectively maintained (PRAKASH and LEWONTIN 1968; 1971).

In this paper, we present the results of an experiment to detect possible linkage disequilibrium between five polymorphic enzyme loci, located around the middle of chromosome 3 of *D. melanogaster*.

MATERIALS AND METHODS

Single male flies were used to start lines. Single chromosomes were extracted from these males, using Ubx^{130} (LINDSLEY and GRELL 1967) as a balancer. At the end of the extraction procedure, the lines were maintained as balanced stocks, or homozygous for the extracted chromosomes.

Three population samples were used. *M69* and *M70* were supplied to us by DR. P. T. IVES, and were trapped from his Amherst population in 1969 and 1970. The *M69* sample gave 98 lines and the *M70* samples 102 lines. The *S* sample was derived from a large artificial population descended from several hundred flies collected at Rochester, New York, and maintained for several years by DR. S. SAUL. It consisted of 265 lines.

Every line was scored for its genotype with respect to five enzyme loci, using starch and acrylamide gel electrophoresis. Lethal-bearing chromosomes were scored as well as chromosomes viable when homozygous. The enzymes studied and the methods used were as follows.

Phosphoglucomutase was done on starch gels by the method of HOPKINSON and HARRIS (1969). Larval alkaline phosphatase was run on starch gels as described by O'BRIEN and MACINTYRE (1968), and stained as described by BECKMAN and JOHNSON (1964). Esterase-6 was run on acrylamide gel in Tris/HCl buffer (46 gms Tris/1; adjusted to pH 8.0). It was found necessary to use a split gel technique. The gel for about 2 cm from the pockets contained 8% acrylamide; the remainder was 6%; this greatly improved resolution. The gels were stained according to the method of HUBBY and LEWONTIN (1966). Xanthine dehydrogenase was done according to the method of PRAKASH, LEWONTIN and HUBBY (1969), except that KCN was omitted from the staining mixture, and it was found helpful to cool the staining mixture for half an hour before incubation with the gel. Aldehyde oxidase was run on the same Tris/HCl buffer as for esterase-6. It was stained according to the recipe of PRAKASH, LEWONTIN and HUBBY (1969), except that benzaldehyde was used as a substrate in place of acetaldehyde (we are grateful to DR. J. H. GILLESPIE for suggesting this change to us). Homogenates of flies were prepared as described by HUBBY and LEWONTIN (1966).

Alleles at a locus were designated by their mobilities relative to an arbitrarily chosen common allele. References for the map positions of each locus may be found in O'BRIEN and MACINTYRE (1971).

Sixty-six of the lines from the *M69* sample were screened for inversions of the third chromosome by crossing with a wild-type stock containing the standard sequence and examining squashes of the larval salivary glands. Two of these lines contained the inversion *In (3R) C*. This inversion covers the region 92D1-E1 to 100 F2-3 on the salivary map (LINDSLEY and GRELL 1967). Crossing

over in *3R* is almost completely suppressed by this inversion. Sixty-nine lines from the *M70* sample were examined; two carried *In (3R) C*, and four *In (3R) P*. *In (3R) P* covers the region 89C2-3 to 36A18-19, and again almost completely suppresses crossing over in *3R*. No inversions were found in the 61 lines from the *S* sample which we examined.

RESULTS

The loci which we have studied are given in Table 1, which shows their map

TABLE 1

Loci examined and their allele frequencies

Esterase-6

(Locus *1*. Map position 36.8)

Population	Alleles 1.00	1.10	Total number of lines
M69	.558	.442	95
M70	.495	.505	101
S	.712	.288	264

Phosphoglucomutase

(Locus *2*. Map position 43.4)

Population	Alleles 0.70	1.00	1.20	1.30	Total number of lines
M69	.093	.794	.113	—	97
M70	.059	.705	.236	—	102
S	.015	.625	.106	.254	264

Larval alkaline phosphatase

(Locus *3*. Map position 46.3)

Population	Alleles 1.00	1.10	Total number of lines
M69	.465	.535	86
M70	.370	.630	100
S	.321	.679	261

Xanthine dehydrogenase

(Locus *4*. Map position 52.0)

Population	Alleles 1.00	1.04	1.06	Total number of lines
M69	.755	.204	.041	98
M70	.647	.265	.088	102
S	.762	.234	.004	265

Aldehyde oxidase

(Locus *5*. Map position 56.6)

Population	Alleles 1.00	1.04	1.10	Total number of lines
M69	.904	.064	.032	93
M70	.782	.168	.050	101
S	.929	.068	.003	265

TABLE 2

Correlation coefficients between pairs of loci (the sign of the correlation coefficients is such that an association between the left-hand allele of the first locus and the right-hand allele of the second gives a positive r)

Loci	Correlation between alleles		Correlations		
	First locus	Second locus	*M69* population	*M70* population	*S* population
1 and 2	1.00, 1.10	1.00, 1.20	+ 0.0985	+ 0.0678	— 0.0379
		1.00, 1.30	——	——	— 0.0301
		1.20, 1.30	——	——	+ 0.0193
1 and 3	1.00, 1.10	1.00, 1.10	— 0.0487	— 0.0054	— 0.0536
1 and 4	1.00, 1.10	1.00, 1.04	+ 0.0598	+ 0.2148*	+ 0.0920
1 and 5	1.00, 1.10	1.00, (1.04 and 1.10)	— 0.0709	+ 0.0428	— 0.0799
2 and 3	1.00, 1.20	1.00, 1.10	— 0.1781	— 0.0998	+ 0.0865
	1.00, 1.30				+ 0.2400†
	1.20, 1.30				+ 0.1531
2 and 4	1.00, 1.20	1.00, 1.04	+ 0.2254*	— 0.1224	— 0.0619
	1.00, 1.30				+ 0.1210
	1.20, 1.30				+ 0.1914
2 and 5	1.00, 1.20	1.00, (1.04 and 1.10)	— 0.0919	+ 0.0022	— 0.0712
	1.00, 1.30				— 0.1070
	1.20, 1.30				— 0.0153
3 and 4	1.00, 1.10	1.00, 1.04	— 0.1656	— 0.0011	— 0.0006
3 and 5	1.00, 1.10	1.00, (1.04 and 1.10)	— 0.1101	+ 0.1099	— 0.1832†
4 and 5	1.00, 1.04	1.00, (1.04 and 1.10)	+ 0.0919	+ 0.1119	— 0.0086

* $p < 0.05$
† $p < 0.01$

positions, the alleles which we detected, and their frequencies. The *M69* and *M70* data show differences in allele frequencies for loci *2* and 5 which are significant at the 5% level, even though *M69* and *M70* are samples from successive years of the same population.

To test for linkage disequilibrium between pairs of loci, tables of the numbers of flies with the appropriate genotypes were drawn up, and tested for non-random associations of alleles by the contingency χ^2 method. Some of the low-frequency alleles (*0.70* at locus *2*, and *1.06* at locus *4*) were omitted from this analysis, in order to avoid very low expected numbers in the contingency tables. Alleles *1.04* and *1.10* at locus 5 were pooled, since both were rare. All the loci reduced to two-allele systems, except for locus *2* in the *S* population, which had three alleles.

The strength of the associations between loci was measured (for pairs with two alleles at each locus) by the product moment correlation in gene frequency between the loci, *r*. If the loci have alleles *A*, *a* and *B*, *b* respectively, with frequencies p_A, q_A and p_B, q_B, and if the frequencies of the chromosome types *A B*, *A b*, *a B* and *a b* are g_1, g_2, g_3, g_4, then we have

$$r = \frac{g_1 g_4 - g_2 g_3}{\sqrt{p_A q_A p_B q_B}}$$

r has the advantage over D (the coefficient of linkage disequilibrium, equal to $g_1g_4 - g_2g_3$), as a measure of the strength of the association between loci, in that it is very much less sensitive to the effect of gene frequency.

For combinations involving locus *2* of the *S* population, the correlation coefficients were calculated for each of the three possible allele combinations. These three correlations are obviously not independent of one another.

The correlation coefficients for each of the ten pairwise combinations of loci in each of the three samples are shown in Table 2. Of the thirty corresponding x^2 tests, four results significant at the 5% level or above were obtained. In the *M69* sample, loci *2* and *4* gave a value of r of 0.225. This corresponds to a $x^2_{(1)}$ of 4.22, significant at the 5% level. In the *M70* sample, loci *1* and *4* gave a value of r of 0.215 ($x^2_{(1)} = 4.25$, $0.02 < p < 0.05$). In the *S* population, loci *2* and *3* gave a $x^2_{(2)}$ of 13.54, $p < 0.01$, and loci *3* and *5* gave an r of -0.183, $x^2_{(1)} = 8.73$, $p < 0.01$.

With this number of tests, there is clearly a good chance of picking up some nominally significant results even if there is no real linkage disequilibrium. It is unlikely that the results described above are due to this. If we confine ourselves to the two results where $p < 0.01$, the smaller of the two, x^2, had a probability level of about 0.006, by interpolation. Using the binominal distribution, the probability of getting two such values out of thirty is about 0.014. At least some of the linkage disequilibria detected are therefore real effects. Further evidence discussed below suggests that probably all are real. If this is so, then every one of the five loci we have studied is involved in at least one significant disequilibrium. There is no sign that disequilibrium is confined to the closest pairs of loci. Loci *2* and *3* are 2.9 map units apart (this is the closest pair in the set which we studied); loci *2* and *4* are 8.6 units apart, which is the 5th closest (out of 10); loci *3* and 5, at 10.3 units apart, are the 7th closest; loci *1* and *4*, at 15.2 units apart, are the 9th closest. Moreover, there is no significant decrease in the values of the correlation coefficients for pairs of loci as map separation increases. Figure 1 shows the means of the squared correlation coefficients for the 10 loci pairs, plotted against map separation. (For the combinations involving locus *2* in the *S* population, which has 3 alleles, we used the mean of the squares of the 3 correlation coefficients obtained by breaking down the 2×3 tables, as the contribution to the overall mean square.)

Figure 1 shows that the four points that correspond to the four loci-pairs which had significant disequilibria in one population are markedly above the other points. This is not simply due to the inclusion of the significant (and therefore high-valued) correlation coefficients in these average values. For the pairs of loci that had a significant correlation in one population, the squared values of the correlation coefficients for the other two populations are also strikingly high, and differ significantly ($p < 0.01$) from the values for pairs of loci that had no significant correlations in any of the populations. The test of significance was carried out by the non-parametric MANN-WHITNEY U-test. This result is not due simply to a correlation between the *M69* and *M70* populations, since a significant

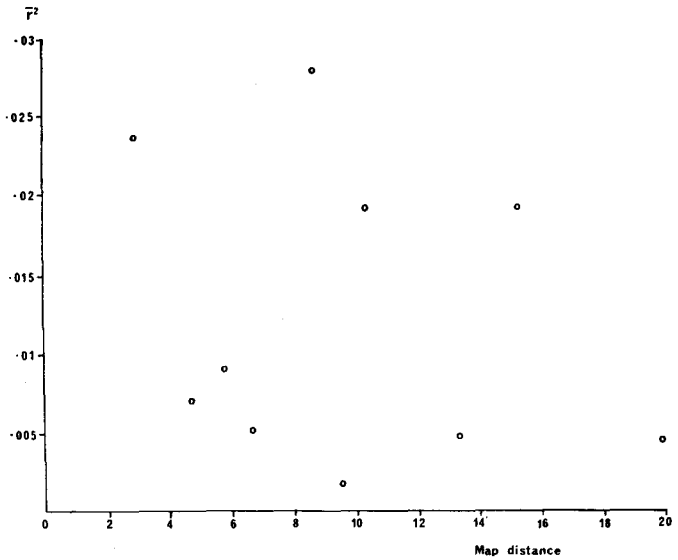

FIGURE 1.—This figure shows the mean value over the three populations of the squares of the correlation coefficients for pairs of loci, as a function of the map distance between them.

($p < 0.01$) result is still obtained when the *M69* and *M70* are pooled, and the test repeated. We also have evidence that *M69* and *M70* are to a large extent independent populations, since they differ significantly in gene frequency at loci *2* and *5*, and in the linkage disequilibrium between loci *2* and *4*, which is positive in the 1969 sample and negative in the 1970 one.

The general picture obtained from these pairwise tests is that high correlations between certain pairs of loci tend to occur in all three separate populations, although the direction of the associations may differ from population to population (see also Table 2). There is no tendency for high correlation coefficients to be restricted to the most closely-linked pairs of genes.

We also tested the data for the existence of correlations of higher order than 2×2. Using the gene frequency estimates, we calculated the expected numbers (on the hypothesis of no associations) of each genotype, and compared them with the observed numbers, by a 2xj χ^2 test. The χ^2 values for *M69* and *M70* were not significant, but the *S* data gave a value of 67.4 for 46 degrees of freedom, which is significant at the 5% level. This could, however, all be accounted for by the pairwise associations detected in this data. There is, therefore, no evidence for higher-order correlations in this data.

It remains to comment on the relations between the gene loci and the inversions detected in the *M69* and *M70* samples. Both of the *In(3R)C* chromosomes in *M69* had the same genotype (*1.10, 1.20, 1.00, 1.00, 1.00*), which has a probability of 0.001 of occurring on the hypothesis of random association between genes and the inversion. There were no such deviations from random expectation for the *In(3R)C* and *In(3R)P* chromosomes in the *M70* population. In neither population was there any evidence for correlation between individual loci and the

inversions, or that associations between genes and inversions contributed to the genic correlations which were detected. The simplest interpretation for the nature of the *In(3R)C* chromosomes in the *M69* population is that they represent the descendants of a single inversion chromosome introduced into that population.

DISCUSSION

Three types of explanation can be advanced for the occurrence of linkage disequilibrium between polymorphic genes: random genetic drift, gene flow between populations differing in gene frequencies at several loci, and selection.

1. *Random drift:* According to the theory of neutral alleles, reviewed by KIMURA and OHTA (1971), random genetic drift can induce linkage disequilibrium between pairs of linked loci segregating for selectively equivalent alleles in a finite population. According to KIMURA and OHTA, the mean value of r^2 over an array of isolated populations for two loci with recombination fraction c, in a population of effective size N_e, will be (at steady-state or at flux equilibrium) approximately

$$\bar{r}^2 = \frac{1}{4N_e c}.$$

This equation enables one to estimate N_e for pairs of loci for which $\bar{r}^2$ and c are known. Our data give us estimates of $\bar{r}^2$ for each pair of loci. c can also be determined for each pair, from the map positions of the loci. (In calculating c, the absence of crossing over in Drosophila males must be taken into account by halving the appropriate map measures; a correction for the presence of inversions should also be made, but this is trivial with the present data.) Our estimates of N_e obtained in this way vary between 170 and 8,200. Theoretically, all pairs of loci ought to give the same value of N_e.

It is difficult to tell whether this rather striking discrepancy means anything or not, since the distributions of r^2 and r on the hypothesis of neutrality are not known, and also the effects of immigration and rapidly changing population size (which must be major factors in a species such as *D. melanogaster*) are difficult to take into account. The general pattern of our results does seem, however, to be difficult to explain on the neutral allele theory, which predicts that all pairs of loci should behave in the same way with respect to the distribution of r, except for the effect of linkage. It is not obvious how, as in our data, four pairs of loci could have significantly higher values of r^2, over two or three populations, than the other pairs of loci; why the size of r^2 for these pairs of loci should be roughly consistent from population to population, and why there should be no relation between r^2 and map separation. Furthermore, the data in Table 1 shows that (with the exception of Locus *2*) the alleles present in the *M* and *S* populations are broadly similar in frequency. This is in accordance with many observations on Drosophila species (PRAKASH, LEWONTIN and HUBBY 1969; O'BRIEN and MACINTYRE 1968). This similarity implies, on the neutral allele theory, a sufficient amount of migration in relation to local population size to keep the species genetically homogeneous (KIMURA and OHTA 1971). This is difficult to reconcile with our finding of significant linkage disequilibrium.

2. *Gene flow:* Linkage disequilibrium can be generated by exchange of genes between populations with different allele frequencies at the loci concerned. This type of explanation also encounters difficulties in explaining the consistency in behaviour of r^2 from population to population, and the distinction between the four loci-pairs with high r^2 and the rest. Also, the *S* population was kept for over three years as a closed laboratory population before sampling. This implies that, if the linkage disequilibrium observed was due to the mixed nature of the *S* population, the initial correlations must have been much higher than the final ones.

This can be seen as follows. In an infinite population, with no selection, the following expression relates the value of r in generation n with its value in generation n-1:

$$r_n = (1\text{-}c)\ r_{n-1}$$

This follows from the corresponding standard formula for *D*, the coefficient of linkage disequilibrium.

Applying this expression to loci 3 and 5, which have a map separation of 10.3 and an observed r of 0.18, the initial value of r must have been 0.67, if it is assumed that 70 generations elapsed between the establishment of the closed population and the time of sampling. Such a high value is unlikely to be produced by migration between populations, particularly in view of the similarity in allele frequencies between populations of *D. melanogaster* (O'BRIEN and MACINTYRE 1968).

3. *Selection:* It is notorious that almost anything can be explained by natural selection, and the present results are no exception. Many years of theoretical investigations, reviewed by KOJIMA and LEWONTIN (1970), have demonstrated that deviations from additive interaction between loci in determining fitness can, with sufficiently strong selection and tight enough linkage, generate stable equilibria with linkage disequilibrium between the interacting loci. The precise pattern of associations among the loci will depend on the pattern of interactions. This type of explanation can clearly be applied to our results. The fact that the direction of the association between a pair of loci may differ from population to population (as with loci *2* and *4* for *M69* and *M70*) does not contradict a selectionist interpretation. It is a well-known property of 2-locus systems that there may often be two stable points with similar gene frequencies but with linkage disequilibria of opposite signs. A bottleneck in population size, such as *D. melanogaster* probably experiences in winter, might well be able to push a population from the neighbourhood of one stable point towards another.

It therefore appears that the interpretation which fits these data best is that the loci studied (which were each involved in one or other significant linkage disequilibrium effect) are under selective control, and that interactions occur between the loci in their effects on fitness in such a way that pairwise correlations are set up between some of them. There is no evidence for the extreme type of linkage disequilibrium effect proposed by FRANKLIN and LEWONTIN (1970), however.

This interpretation is consistent with the findings of PRAKASH and LEWONTIN (1968, 1971) on *D. pseudoobscura* and *D. persimilis*, for associations between

enzyme loci and inversions. It is obviously highly desirable for more data of this sort to be collected since the present results represent a very limited sample.

The electrophoretic part of this work was carried out while B.C. held a postdoctoral fellowship from the Ford Foundation grant for population biology to the Department of Biology, University of Chicago, and while D.C. was a research associate in the Department of Medicine, University of Chicago. We thank DRS. P. T. IVES and S. SAUL for providing us with flies, and R. C. LEWONTIN for his interest and encouragement.

LITERATURE CITED

BECKMAN, L. and F. M. JOHNSON, 1964 Variation in larval alkaline phosphatase controlled by *Aph* alleles in *Drosophila melanogaster*. Genetics **49**: 829–835.

FRANKLIN, I. and R. C. LEWONTIN, 1970 Is the gene the unit of selection? Genetics **65**: 707–734.

HOPKINSON, D. A. and H. HARRIS, 1969 Red cell acid phosphatase, phosphoglucomutase and adenylate kinase. In: *Biochemical Methods in Red Cell Genetics*. Edited by J. J. YUNIS. Academic Press, New York.

HUBBY, J. L. and R. C. LEWONTIN, 1966 A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. Genetics **54**: 577–594.

KIMURA, M. and T. OHTA, 1971 Theoretical aspects of population genetics. Princeton University Press, Princeton, N. J.

KOJIMA, K. and R. C. LEWONTIN, 1970 Evolutionary significance of linkage and epistasis. In: *Mathematical Topics in Population Genetics*, pp. 366–388, Edited by K. KOJIMA. Springer-Verlag, Berlin.

KOJIMA, K., J. H. 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.

LINDSLEY, D. L. and E. H. GRELL, 1967 Genetic Variations of *Drosophila melanogaster*. Carnegie Institution of Washington, Publ. No. 627, Washington, D.C.

MUKAI, T., L. E. METTLER and S. I. CHIGUSA, 1970 On the linkage equilibrium of isozymes in a Raleigh, N. C. population of *D. melanogaster*. Drosophila Information Service **45**: 77. ——, 1971 Linkage disequilibrium in a local population of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S. **68**: 1065–1069.

NAIR, P. S. and D. BRNCIC, 1971 Allelic variations within identical chromosome inversions. Amer. Nat. **105**: 291–294.

O'BRIEN, S. J. and R. J. MACINTYRE, 1968 An analysis of gene-enzyme variability in natural populations of *Drosophila melanogaster* and *D. simulans*. Amer. Nat. **103**: 97–112. ——, 1971 A biochemical genetic map of *D. melanogaster*. Drosophila Information Service **46**: 89–93. ——, 1971 Transient linkage disequilibrium in Drosophila. Nature **230**: 335–336.

PRAKASH, S. and R. C. LEWONTIN, 1968 A molecular approach to the study of genic heterozygosity in natural populations. III. Direct evidence of coadaptation in gene arrangements of Drosophila. Proc. Nat. Acad. Sci. U.S. **59**: 398–405. ——, 1971 A molecular approach to the study of genic heterozygosity in natural populations. V. Further direct evidence of coadaptation in inversions of Drosophila. Genetics **69**: 405–408.

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.