VARIANCE COMPONENT ANALYSIS OF ALLOZYME FREQUENCY DATA FROM EASTERN POPULATIONS OF *DROSOPHILA MELANOGASTER*

DIANA B. SMITH, CHARLES H. LANGLEY AND FRANK M. JOHNSON

National Institutes of Health, National Institute of Environmental Health Sciences, Research Trinagle Park, North Carolina

Manuscript received May 13, 1977
Revised copy received September 6, 1977

ABSTRACT

Gene frequency variation at eight polymorphic allozyme loci in *Drosophila melanogaster* populations in North Carolina and the east coast of the United States were analyzed utilizing the variance component estimation procedures suggested by COCKERHAM (1969, 1973). These variance components were used to estimate correlations of genes within small geographic regions. The average (over loci) correlation between genes in the same individual within subpopulations was estimated to be 0.033. That between genes in the same subpopulation in different individuals was estimated to be very small, although significantly different from zero. The macrogeographic variation measured by the correlation of genes sampled from the same local region was large for some loci and smaller for others. This variation was also analyzed by correlation with latitude and longitude. Several previously recognized clines were identified as were several new clines.—These results were interpreted as indicating either some degree of nonrandom mating and local breeding unit isolation or a low frequency of null alleles. The geographic and temporal variation has no simple interpretation.

THE nature of population structure in Drosophila and especially in *D. melanogaster* bears indirectly on many facets of population genetics. We report here a detailed analysis of allozyme frequencies at eight loci (*Est-C*, *Est-6*, *Adh*, *αGpdh*, *Acph*, *Odh*, *Mdh*, *Pgm*) in over one hundred population samples. The structure that we have detected is presented in terms of correlations between genes, based on estimated variance components.

Consider a hierarchy of subpopulations: subpopulations grouped into regions and regions forming the total. As suggested by COCKERHAM (1969, 1973), let one allele take the value 1 and its alternative allele take the value 0. Then there are gene frequencies in the individual (0, 0.5, or 1), in the subpopulations, in the regions, and in the total. There are also variances in these frequencies. COCKERHAM (1969, 1973) showed that the variation in gene frequency could be divided into components. σ_w^2 is the component attributable to variation in genes within individuals. σ_{b1}^2 is the component attributable to variation among individuals within subpopulations. σ_{b2}^2 is that component due to gene variation among sub-

populations and σ^2_{b3} due to variation among regions. Straightforward statistical analysis gives adequate estimates of these components and procedures for testing their significance (COCKERHAM 1973). Intraclass correlations associated with these components were shown to have correspondence to the earlier measures of relatedness of genes and individuals (COCKERHAM 1969, 1973). It is these correlations that we report.

The correlation we report is f_1, the correlation between genes within individuals within subpopulations; F, the correlation between genes within subpopulations; f_2, the correlation of genes in different individuals within subpopulations within regions; and Θ_2, the correlation of genes between subpopulations in the same region. These correlations reflect the relatedness of genes in different levels of the hierarchy.

METHODS

The *D. melanogaster* were collected in the years 1970 through 1973. Collection sites were over much of the eastern part of the United States. As Figure 1 shows, most of the sampling was in North Carolina. The collection procedure is described in JOHNSON and BURROWS (1976). In most instances banana traps were sampled for several consecutive days, but never for more than a week. Since the sampling was not done with a hierarchical type of study in mind, the distribution of samples in time and space is not as uniform as one would like. The samples fall into natural geographic clusters, as shown in Figure 1. Each sample has been assigned the location of the nearest city or town to two decimal places of degree longitude and latitude. These assignments are referred to as "regions" throughout this text. "Subpopulations" are individual collections within "regions." Thus, subpopulations from the same region may have been collected hundreds of meters apart in the same week or kilometers apart in different years. Although more subpopulations were actually collected, we have utilized in our analysis only those subpopulations with fifty or more individuals scored for a particular locus. Thus, the number of subpopulations for the eight loci are different. A preliminary test of homogeneity of subpopulations from the same region and time showed that only one set of subpopulations could be pooled (Edenton, N.C., 1972). These were pooled; all other subpopulations constitute single collections. A listing of the data is available upon request.

Our analysis deals with the most frequent allele at each locus. All the loci have one predominant allele. Analysis of multiple alleles would complicate the procedure and contribute relatively little information because of small frequencies and sample sizes.

Besides the natural population collections, we have also analyzed data from a population cage and collections made indoors. The cage population was founded in 1972 with over 250 recently collected isofemale lines from Florida. It was maintained with discrete generations on the usual North Carolina State University media at 25°. The indoor collections came from sweet potato warehouses, pickle and vinegar factories, and a produce market in North Carolina.

The allozymic scoring was done utilizing starch gel electrophoresis and various histochemical stains. These techniques can be accessed through JOHNSON and SCHAFFER (1973).

COCKERHAM (1973) suggested four statistical methods for the analysis of gene frequencies in a hierarchically structured population. These are average symmetrical products, average symmetrical squared differences, the analysis of variance (ANOVA), and a corresponding chi-square analysis. The method suggested as best, when the variances between samples and between higher levels of the heirarchy are relatively small, was the weighted analysis of variance. Our estimates are based on such a procedure.

Following COCKERHAM's general model, variance components of gene frequencies were estimated:

σ^2_{b3} = the variance component attributable to differences in regions.

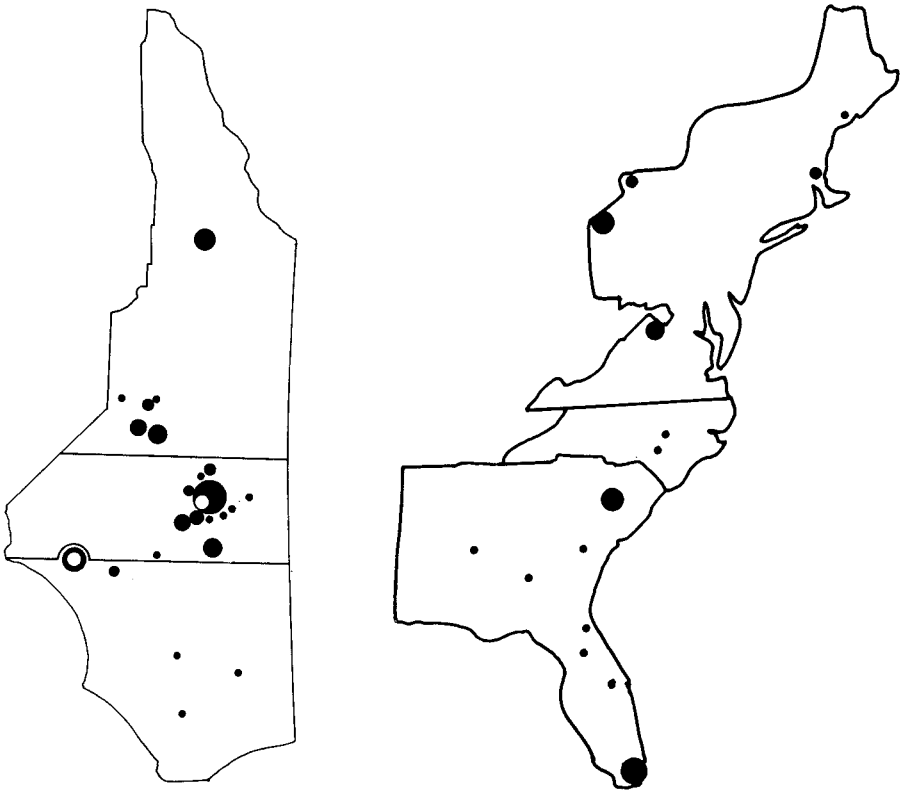

FIGURE 1.—Area of circles represents relative number of samples from that locality with the smallest circle in each map representing one sample. Circles within circles indicate two collections from slightly different localities. Figure 1a (left) is North Carolina divided into three sections—east, central, and west as used in the analysis of variance. Figure 1b (right) represents the east coast and includes two samples from North Carolina.

σ^2_{b2} = the variance component attributable to differences in subpopulations in regions.

σ^2_{b1} = the variance component attributable to differences in individuals in subpopulations.

σ^2_w = the variance component attributable to differences in alleles in individuals.

$\sigma^2 = \sigma^2_w + \sigma^2_{b1} + \sigma^2_{b2} + \sigma^2_{b3} = p(1-p)$.

Using our estimated variance components calculated from the expected mean squares, we then compute:

$f_1 = \sigma^2_{b1} / (\sigma^2_{b1} + \sigma^2_w)$; the correlation between genes within individuals

within subpopulations which is analogous to WRIGHT's F_{IS} (WRIGHT 1951). This intraclass correlation, unlike the others we estimate, may legitimately take on negative values.

$F = (\sigma^2_{b1} + \sigma^2_{b2} + \sigma^2_{b3})/\sigma^2$; the intraclass correlation of genes within subpopulations, analogous to WRIGHT's F_{IT} (WRIGHT 1951).

$\Theta_1 = (\sigma^2_{b2} + \sigma^2_{b3})/\sigma^2$; the intraclass correlation of genes between individuals in the same subpopulation.

$\Theta_2 = \sigma^2_{b3}/\sigma^2$; the intraclass correlation of genes between subpopulations in the same region.

$f_2 = \sigma^2_{b2}/(\sigma^2 - \sigma^2_{b3})$; the intraclass correlation of genes in different individuals within subpopulations within regions, that is:

$$f_2 = (\Theta_1 - \Theta_2)/(1 - \Theta_2)$$

In some parts of our analysis we have considered a less structured situation. This case will be referred to as the "reduced model". This analysis assumes a single region, *i.e.*, only variance components σ^2_w, σ^2_{b1}, and σ^2_{b2} are estimated. In this reduced model, Θ_2 is assumed to be zero.

Besides the ANOVA, we also compute chi-square values as suggested by COCKERHAM (1973). These give us a test of the significance of our estimates obtained from the ANOVA. The chi-square values are as follows:

Test of $f_1=0$. We compute $\chi^2_{f_1}$ with 1 degree of freedom.

Test of all $f_{ij}=0$, where f_{ij} is the single subpopulation parameter for the correlation of genes within individuals within subpopulations (see COCKERHAM 1973). We compute χ^2_{total} with M degrees of freedom.

Test of all $f_{ij} = f_1$, that is, that the variance of f_1 is not significant. We compute $\chi^2_{f_1} = \chi^2_{\text{total}} - \chi^2_{f_1}$ with $M-1$ degrees of freedom.

Test of $f_2=0$ for the general model or $\Theta_1=0$ for our reduced model. We compute χ^2_1 or $\chi^2_{f_2}$ with $M-1$ degrees of freedom.

Test of $\Theta_2=0$ for the general model is $\chi^2_{\Theta_2}$ with $I-1$ degree of freedom.

Besides the ANOVA and chi-square statistics, we also computed linear and rank correlations between gene frequencies and scalers (temporal and spatial). These procedures can be found in many textbooks on statistical methods.

RESULTS

Analysis of variance

Our first analysis is of the more than 100 samples from North Carolina. We structured the analysis in various ways so that possible relationships or inconsistencies might become more apparent.

In Table 1 is an analysis of the North Carolina data with our highest level in the ANOVA being regions (longitude-latitude). Note that m_1 in this table is the number of regions, N_t is the total number of flies, $\overline{N}$ is the average number per subpopulation (sample), and s^2_N is the variance of the number of individuals per subpopulation. At the right margin of this and following ANOVA tables are the

TABLE 1

*North Carolina collections: analysis of gene frequencies**

		αGdph	*Mdh*	*Adh*	*Est-6*	*Pgm*	*Est-C*	*Odh*	*Acph*		
Correlation	p	0.8307	0.9650	0.6974	0.6144	0.8533	0.9117	0.9201	0.9513		
between genes	m_1	23	24	24	24	20	24	24	24		
	N_t	12778	13371	13580	13169	8186	11450	12675	13102		
	$\bar{N}$	125.3	127.3	128.1	126.6	115.3	118.0	124.3	124.8		
	$s^2{}_N$	8119.3	9516.8	8728.2	9630.1	2453.5	8873.2	8025.1	8363.8		
Within regions	σ_2	−0.0008	−0.0008[b]	0.0014[c]	0.0005[c]	−0.0006	0.0032[c]	0.0005[c]	−0.0000[c]	$\sigma_2{}^c = 0.0004$	se = 0.0005
Within subpopu-	m	102	105	106	104	71	97	102	105		
lations in regions	f_2	0.0024[c]	0.0058[c]	0.0015[b]	0.0048[c]	0.0024[b]	0.0053[c]	0.0025[c]	0.0071[c]	$\bar{f}_2{}^c = 0.0040$	se = 0.0007
Within individuals	f_1	0.0167	0.0207[a]	0.0157	0.0453[c3]	−0.0135	0.0299[c3]	0.0360[c3]	0.1157[c3]	$\bar{f}_1{}^{c3} = 0.0333$	se = 0.0133
F		0.0183	0.0257	0.0185	0.0506	−0.0117	0.0382	0.0389	0.1219	$\bar{F} = 0.0376$	se = 0.0138
Raleigh	m	28	28	29	28	12	26	26	29		
Within individuals	f_1	0.0153	0.0209	0.0495[b]	0.0429[a]	0.0213	0.0642[b]	0.0472[a]	0.1346[c]	$\bar{f}_1{}^c = 0.0495$	se = 0.0136
Within subpopulations	Θ_1	0.0083[c]	0.0147[c]	0.0031[a]	0.0141[c]	0.0014	0.0118[c]	0.0036[c]	0.0133[c]	$\bar{\Theta}_1{}^c = 0.0088$	se = 0.0014
Not Raleigh	m	74	77	77	76	59	71	76	76		
Within individuals	f_1	0.0171	0.0207[a]	0.0057	0.0463[c]	−0.0194	0.0207[2]	0.0325[c2]	0.1108[c3]	$\bar{f}_1{}^{c3} = 0.0293$	se = 0.0135
Within subpopulations	Θ_1	−0.0002	0.0015[b]	0.0026[c]	0.0028[c]	0.0019[b]	0.0072[c]	0.0024[c]	0.0054[c]	$\bar{\Theta}_1{}^c = 0.0030$	se = 0.0008

* See text for explanation of symbols.
Mean significance; a, $P \leq 0.05$; b, $P \leq 0.01$; c, $P \leq 0.001$.
Variance significance; 1, $P \leq 0.05$; 2, $P \leq 0.01$; 3, $P \leq 0.001$.

means and variances across loci of f_1, f_2, Θ_2 and F. (The statistical significance indicated is determined for the summed chi-squares (across loci) rather than the mean itself.) We can see in Table 1 that six out of the the eight loci have significant Θ_2 values. The mean Θ_2 over loci is 0.0004. The summed $\chi^2_{\Theta_2}$ is highly significant. f_2 is the intraclass correlation of genes in different individuals within subpopulations within regions and is significant for all loci. Note that $\bar{f}_2$ is ten times larger than $\bar{\Theta}_2$. The correlation of gene frequencies within subpopulations is much larger than that between subpopulations within regions. In absolute value our variance among subpopulations is small but definitely larger than that expected from simple binomial sampling.

From Table 1, we see the f_1 values are mostly positive (seven out of eight) and five of these are significant. The average f_1 is significantly different from zero ($\bar{f}_1$=0.033). The same five loci have significant variances of f_1 as shown by their $\chi^2_{f_1}/df$. (This variance of f_1 suggests that it would be difficult to get a good estimate of f_1 with a small number of samples.) Note the high f_1 values (see below). These results clearly indicate an overall discrepancy from the expectation for the Hardy-Weinberg model in favor of too many homozygotes.

The bottom of Table 1 breaks down the data into Raleigh samples *versus* the rest of the state. This is a logical subdivision as about 25% of the samples in North Carolina are from Raleigh. This also allows us to study in more detail the changes that take place in one locality. Our Θ_1 is the correlation of genes in different individuals within subpopulations (reduced model). The Θ_1 values for within Raleigh are as statistically significant as those not in Raleigh, but do appear to be more than twice as large. There seems to be a general consistency between "Raleigh" and "not Raleigh" for the f_1 values, with one exception. The f_1 estimate for *Adh* is 0.0495 in "Raleigh" *versus* 0.0057 in "not Raleigh". Note the consistency in the *Acph* f_1 for "Raleigh" *versus* "not Raleigh" samples (0.1346 *versus* 0.1108).

In all of these analyses the subpopulations within a region were often sampled in different months in several different years. This raises the question as to what effect there is of time. Table 2 shows the average (over loci) of Θ_2 (years) and f_2 (subpopulations in years) for western, central and eastern North Carolina and for the entire state. In all cases the Θ_2 values are significant, *i.e.*, there is significant variation in gene frequencies from year to year within the three divisions of the state. Also, Table 2 gives the gene frequencies for years and their estimated standard errors from the analysis of variance. Negative values for standard errors reflect their smallness and the fact that these are estimates based on expected mean squares. In general, there are significant differences between years in these three divisions of the state. An examination of yearly gene frequencies in Table 2 indicates that some of these changes may be directional. Later we will show by correlation analysis that this is true.

In order to see how much of the variance between subpopulations within regions is between subpopulations at any one time (collections from the same longitude, latitude, and time), a further analysis was made (Table 3). The

TABLE 2

Gene frequency correlations within years (Θ_2) and subpopuolatins in years (f_2) in eastern, central and western North Carolina

	Years		Subpopulations					
	Θ_2	se_{Θ_2}	f_2	se_{f_2}	χ^2/df for Θ_2		χ^2/df for f_2	
N.C. east	0.0005	0.0023	0.0025	0.0074	2.25	$P < 0.05$	1.674	$P < 0.0005$
N.C. central	0.0011	0.0013	0.0042	0.0013	6.14	$P < 0.0005$	2.099	$P < 0.0005$
N.C. west	0.0017	0.0023	0.0019	0.0011	5.00	$P < 0.0005$	1.444	$P < 0.0005$
N.C. total	0.0012	0.0013	0.0036	0.0099	10.52	$P < 0.0005$	1.931	$P < 0.0005$

Analysis of gene frequencies over years

Years	*αGdph*		*Mdh*		*Adh*		*Est-6*		*Pgm*		*Est-C*		*Odh*		*Acph*	
	p	*se*	*p*	*se*	*p*	*se*	*p*	*se*	*p*	*se*	*p*	*se*	*p*	*se*	*p*	*se*
1970	0.8219	0.0051	0.9606	0.0034	0.6770	0.0038	0.6246	0.0094	——	——	0.9370	0.0030	0.9218	0.0031	0.9445	0.0048
1971	0.8206	+	0.9632	0.0049	0.7119	0.0124	0.6063	0.0092	0.8364	+	0.9171	0.0053	0.9217	+	0.9250	+
1972	0.8359	0.0001	0.9678	0.0010	0.7072	0.0027	0.6122	0.0029	0.8562	0.0023	0.8928	0.0023	0.9170	0.0023	0.9540	0.0018
1973	0.8404	+	0.9677	0.0012	0.7083	+	0.5957	0.0069	0.8478	0.0032	0.9001	0.0137	0.9252	+	0.9612	+

se_{Θ_2} = Sample variance over loci of Θ_2.

se = ANOVA estimate of standard error.

\+ = Negative estimate.

TABLE 3

Comparisons of correlations within concurrent subpopulations with that within subpopulations in regions

	αGdph	*Mdh*	*Adh*	*Est-6*	*Pgm*	*Est-C*	*Odh*	*Acph*
Concurrent subpopulations	0.0026	0.0043	0.0018	0.0027	0.0033	0.0059	0.0026	0.0084
Subpopulations in regions	0.0024	0.0058	0.0015	0.0048	0.0024	0.0053	0.0025	0.0071

intraclass correlation for concurrent subpopulations is compared with that for all subpopulations in region. We see that the average over loci for concurrent collections is about the same as that for all subpopulations in a region. This suggests that most of the variance among subpopulations within regions is between concurrent collections. The samples are clearly differentiated within regions at any one time although the magnitude of this difference is small.

We would like to know if the values of the estimates we obtained in North Carolina are typical of *D. melanogaster* along the East Coast. Table 4 is an analysis of collections from Florida and up the coast to Maine, including two of the North Carolina samples. The reduced model was applied. Seven out of eight Θ_1 values are significant. Of the four significant f_1 values, three of these (*Est-6*, *Est-C*, and *Acph*) were also significant in North Carolina. If we look at the estimates of f_1 for those states north of North Carolina and those south of North Carolina (bottom of Table 4), we find that the southern subpopulations appear to have larger f_1 values. There may be a difference between the northern and southern states for *Adh* ($f_1 = 0.0336$ *versus* $f_1 = 0.0487$). Note that for f_1 for *Acph* is significant in the southern states ($f_1 = 0.1073$) but not significant in the north ($f_1 = 0.0034$).

We have two different collections left for study. These are indoor collections and the cage samples. JOHNSON and BURROWS (1976) analyzed *Adh* gene frequencies for samples from warehouses, factories, and a Farmer's Market, all in North Carolina. We applied the analysis of variance for the reduced model to these collections (Table 5) for all eight loci, and found only one f_1 significant (0.1044 for *Acph*). Perhaps this is because of the smaller number of subpopulations. All the θ_1 values are significant, which is not surprising for samples from such diverse sources. However, the overall similarity in frequencies and the f_1 value for *Acph* suggest little differentiation of these populations from the surrounding natural populations.

We would like to compare cages with natural populations. JOHNSON had set up a cage from Miami collections (December 1972) and Table 6 shows the analysis for five samples starting in May 1973 and continuing through January 1974. From these we find little evidence of inbreeding, but a great deal of evidence that gene frequencies are changing ($\theta_1 = 0.0380$; $\bar{f} = 0.0106$).

TABLE 4

*East Coast collections: analysis of gene frequencies**

	αGdph	*Mdh*	*Adh*	*Est-6*	*Pgm*	*Est-C*	*Odh*	*Acph*	
East Coast									
p	0.8479	0.9731	0.7038	0.5895	0.8708	0.9089	0.9028	0.9433	
m	30	29	33	29	24	27	28	33	
N	4532	4687	5015	4699	2946	4177	4328	4871	
$\bar{N}$	151.1	161.1	152.0	162.0	122.8	154.7	154.6	147.6	
$s^2{}_N$	9346	11412	11089	11524	3264	10472	9999	10836	
f_1	0.0314[a]	0.0120[3]	0.0046	0.0329[a]	0.0147	0.0597[c]	0.0228	0.0564[b3]	$\bar{f}_1{}^{c3} = 0.0293$ se $= 0.0071$
Θ_1	0.0069[c]	0.0012	0.0556[c]	0.0059[c]	0.0193[c]	0.0182[c]	0.0210[c]	0.0239[c]	$\bar{\Theta}_1{}^{c} = 0.0190$ se $= 0.0060$
F	0.0381	0.0133	0.0600	0.0386	0.0337	0.0769	0.0433	0.0789	$\bar{F} = 0.0478$ se $= 0.0074$
North of NC									
m	11	11	11	11	9	11	10	11	
f_1	0.0106	−0.0034	−0.0336	0.0363	0.0110	0.0347	0.0469	0.0034	$\bar{f}_1 = 0.0132$ se $= 0.0092$
Θ_1	0.0069[c]	0.0007	0.0267[c]	0.0028[a]	0.0142[c]	0.0172[c]	0.0187[c]	0.00599[c]	$\bar{\Theta}_1{}^{c} = 0.0184$ se $= 0.0067$
South of NC									
m	17	16	20	16	13	14	16	20	
f_1	0.0380	0.0212[c]	0.0487[a]	0.0366	0.0233	0.0701[b]	0.0191	0.1073[c3]	$\bar{f}_1{}^{c3} = 0.0455$ se $= 0.0106$
Θ_1	0.0073[c]	0.0025[a]	0.0257[c]	0.0101[c]	0.0086[b]	0.0140[c]	0.0128[c]	−0.0000	$\bar{\Theta}_1{}^{c} = 0.0101$ se $= 0.0028$

* See text for explanation of symbols.
Mean significance; a, $P \leq 0.05$; b, $P \leq 0.01$; c, $P \leq 0.001$.
Variance significance; 1, $P \leq 0.05$; 2, $P \leq 0.01$; 3, $P \leq 0.001$.

TABLE 5

*North Carolina indoor collections: analysis of gene frequencies**

	αGdph	*Mdh*	*Adh*	*Est-6*	*Pgm*	*Est-C*	*Odh*	*Acph*			
p	0.8368	0.9722	0.6874	0.5906	0.8714	0.9211	0.9025	0.9564			
m	15	14	15	15	12	12	15	15			
N	2184	2159	2252	2806	1511	1662	2226	2190			
$\bar{N}$	145.6	154.2	150.1	139.1	125.9	138.5	148.4	146.0			
$s^2{}_N$	10739	13315	12779	11809	1786	16438	12903	13078			
f_1	−0.0016	0.0164	0.0046	0.0053	0.0150	0.0390[1]	0.0155[1]	0.1044[c1]	$\bar{f}_1$	$= 0.0248$	se $= 0.0122$
Θ_1	0.0047[b]	0.0071[c]	0.0283[c]	0.0144[c]	0.0031	0.0092[c]	0.0157[c]	0.0040[b]	$\bar{\Theta}_1{}^c$	$= 0.0108$	se $= 0.0030$
F	0.0031	0.0235	0.0328	0.0196	0.0108	0.0478	0.0310	0.1080	$\bar{F}$	$= 0.0355$	se $= 0.0113$

* See text for explanation of symbols.
Mean significance; a, $P \leq 0.05$; b, $P \leq 0.01$; c, $P \leq 0.001$.
Variance significance; 1, $P \leq 0.05$; 2, $P \leq 0.01$; 3, $P \leq 0.001$.

TABLE 6

*Miami cage samples (5/73 to 1/74): analysis of gene frequencies**

	αGdph	*Mdh*	*Adh*	*Est-6*	*Pgm*	*Est-C*	*Odh*	*Acph*			
p	0.9218	0.9721	0.7658	0.5363	0.7617	0.8461	0.8687	0.9565			
m	5	5	5	5	5	5	5	5			
N	1209	1220	1247	1158	1196	1079	1123	1137			
$\bar{N}$	241.8	244.0	249.4	231.6	239.2	215.8	224.6	227.4			
$s^2{}_N$	5260	3906	4852	3760	5008	2768	3939	2658			
f_1	0.0382	0.0221	0.0179	0.0284	0.0100	−0.0041	−0.0206	−0.0097	$\bar{f}_1$	= 0.0106	se = 0.0072
Θ_1	0.0200[c]	0.0130[c]	0.0441[c]	0.1059[c]	0.0814[c]	0.0276[c]	0.0032[c]	0.0086[c]	$\bar{\Theta}_1{}^c$	= 0.0380	se = 0.0131
F	0.0574	0.0349	0.0612	0.1313	0.0906	0.0237	−0.0174	−0.0100	$\bar{F}$	= 0.0476	se = 0.0171

* See text for explanation of symbols.
Mean significance; a, $P \leq 0.05$; b, $P \leq 0.01$; c, $P \leq 0.001$.
Variance significance; 1, $P \leq 0.05$; 2, $P \leq 0.01$; 3, $P \leq 0.001$.

TABLE 7a

North Carolina collections: linear correlations

	αGdph	*Mdh*	*Adh*	*Est-6*	*Pgm*	*Est-C*	*Odh*	*Acph*
Latitude	—0.26*	+0.11	—0.25*	+0.14	+0.01	+0.10	—0.04	+0.02
Longitude	—0.09	+0.13	+0.03	+0.32*	—0.12	+0.05	+0.03	—0.22*
Month	+0.01	+0.01	+0.21*	—0.11	—0.04	—0.07	+0.13	—0.04
— Raleigh	—0.02	—0.08	+0.17*	+0.06	—0.25	+0.13	—0.04	+0.01
— not Raleigh	—0.06	—0.02	+0.20	—0.23*	+0.03	—0.07	+0.13	—0.07
Time	+0.29*	+0.20*	+0.26*	—0.14	+0.05	—0.53*	+0.02	+0.25*
— Raleigh	+0.44*	+0.45*	+0.31	+0.23	+0.22	—0.39*	+0.19	+0.46*
— not Raleigh	+0.08	—0.07	+0.21	—0.44*	+0.02	—0.56*	—0.13	+0.14

Time = month + 12× (year—1970).
* $P \leq 0.05$.

Linear and rank correlation analysis

Correlations between gene frequencies and latitude for the East Coast collections were reported earlier by JOHNSON and SCHAFFER (1973). For the North Carolina samples we applied both rank and linear correlation analysis to investigate the relationship between gene frequencies and scalars (latitude, longitude, month in year, and months over years). Table 7a and 7b give the linear and rank correlations respectively. Note that the two methods give overall similar results, even though one assumes normality and the other method (Kendall's rank correlation analysis) does not assume normality. We see that *Adh* has a definite cline with latitude, and *Est-C* and *αGpdh* also have probable clines with latitude. This is in agreement with the East Coast clines reported by JOHNSON and SCHAFFER (1973). *Est-6* has a strong longitudinal cline and to a lesser extent a latitudinal cline. *Acph* exhibits an apparent longitudinal cline, as does *αGpdh.*

Temporal correlation suggests *Adh* has a seasonal (monthly) change not generally shown by any of the other loci. This was suggested by JOHNSON and BURROWS (1976) utilizing part of these data. The *Est-6* monthly change may be due to a seasonal effect not seen in Raleigh, or it is due to some inadvertent bias

TABLE 7b

North Carolina collections: rank correlations

	αGdph	*Mdh*	*Adh*	*Est-6*	*Pgm*	*Est-C*	*Odh*	*Acph*
Latitude	—0.13	+0.05	—0.20*	+0.14*	+0.02	+0.16*	—0.06	+0.03
Longitude	—0.14*	+0.03	+0.01	+0.15*	—0.08	+0.01	—0.01	—0.11
Month	—0.03	—0.01	+0.15*	—0.08	—0.04	—0.05	+0.09	—0.05
— Raleigh	—0.13	—0.20	+0.10	—0.01	—0.17	—0.14	+0.03	—0.18
— not Raleigh	—0.03	+0.00	+0.16*	—0.14	+0.02	—0.06	+0.08	—0.02
Time	+0.15*	+0.10	+0.23*	—0.17*	+0.02	—0.33*	+0.02	+0.13
— Raleigh	+0.32*	+0.34*	+0.32*	+0.13	+0.07	—0.26	+0.15	+0.39*
— not Raleigh	+0.10	—0.01	+0.18*	—0.30*	—0.03	—0.36*	—0.06	+0.11

* $P \leq 0.05$.

in sampling since *Est-6* does show a geographic cline. Gene frequencies change definitely with months over years (time) as shown by the Raleigh data and also by the analysis of variance (Table 3). Some of these seasonal and yearly changes may be locality-specific.

It is especially interesting to note that *αGpdh*, *Mdh*, and *Adh* show simultaneous increase in frequency with time. These three loci are all linked in the left arm of chromosome two. The overall results in this table suggest significant associations between gene frequencies and scalar quantities. The geographic clines are slight but significant and several are well established from previous studies.

Population sizes vary substantially with season and latitude, *i.e.*, average temperature or some related climatic variables. If the observed deficiency of heterozygotes were due solely to the finite sizes of random breeding populations, the f_1 values would be expected to correlate negatively with population sizes. Contrary to this, we observed higher f_1 values in the southern subpopulations from along the East Coast. We also examined the North Carolina samples for correlation of f_1 with season since local population sizes appear much smaller in the spring and seem to peak in the fall. No significant correlations or patterns were observed in the analysis (not shown).

DISCUSSION

The results of this study are fundamentally descriptive. Much of what is understood about the genetic variation of natural populations is based on observations from Drosophila. It is from this view that we examined the genetic structure of the *D. melanogaster* populations through allozyme frequencies. Beyond the primary task of description, we also attempted to address specific questions when the opportunity presented itself.

The data in this study were not collected with the study of hierarchial population structure in mind. Sampling was neither stratified nor random with respect to geography and time. Since these collections fall into natural sets, such as regions or years, the hierarchial analysis suggested by COCKERHAM (1969, 1973) is appropriate. Interpretation of the various correlations is always ambiguous unless complete historical and demographic information is available. But these correlations do reflect the relatedness of genes. The actual causal forces that shaped the pedigrees reflected in the correlations can not be discerned. All that can be stated are the estimates of correlations and their interpretation as crude measures of relatedness.

The collections used in this study were flies collected in several traps (meters apart) during a period of several days. The individuals must represent a group that had the potential to interbreed although the presence of the trap may have introduced artificial congregation. We can think of the collection as a sample from a subpopulation that has some particular history of mating. This mating structure may or may not have been random. The correlation of allelic genes within individuals is a measure of the inbreeding that took place in the previous generations. We are, of course, neglecting the role of selection for the time being.

The correlations within individuals (f_1) we observed indicates some amount of apparent inbreeding: mean $f_1 = 0.033$. If this observed f_1 were due wholly to inbreeding, a question arises with respect to the population genetics of recessive lethals. In the theoretical formulations for the expected chromosomal frequency of recessive lethals, Q, the heterozygous effect on fitness, h and the amount of inbreeding f add together: $Q = U/(h+f)$ where U is the chromosomal recessive lethal mutation rate (CROW and KIMURA 1970). This value is insensitive to population size when $(h+f) \geq\geq U$ as in this case (CROW and KIMURA 1970). In their review of the literature on recessive lethal frequencies and mutation rates, CROW and TEMIN (1964) estimated h. They assumed that f was zero and that the decrement in frequency of recessive lethals from the expected square root of the mutation rate was wholly due to the heterozygous deleterious effects. They cited the study of HIRAIZUMI and CROW (1960) in which h was directly estaimated to be 0.01. All these earlier observations complement each other to form a plausible picture where the theoretical predictions and experimental observations concur. Our observation calls this concordance into question. CROW and TEMIN (1964), and indeed many others, have assumed no inbreeding in natural populations of *D. melanogaster*. If $f_1 = 0.033$ is substituted into the formula then h is close to zero, given the observed Q and U per chromosome. Thus there is some inconsistency generated by our results; h and f can not both be equal to 0.03 within the framework of the theory and the observed mutation frequencies and rates.

This apparent inconsistency has several potential resolutions. The first may be that the simplistic model of independent of loci in mutation and fitness effects is incorrect. A second potential resolution may lie in the selective effects of allozymes or closely linked loci. CHRISTIANSEN, FRYDENBERG and SIMONSEN (1973) have observed an excess of homozygotes at an allozyme locus in natural populations of the eel pout (a fish). Their analysis ruled out nonrandom mating or other population structure effects as the source. They concluded that underdominant natural selection was operating. The result presented here *may* be interpreted as evidence of selection favoring homozygotes.

A third resolution might be the existence of low frequency null alleles at the allozyme loci we examined. Heterozygotes for such null alleles would be scored as homozygotes, thus elevating the frequency of homozygotes above expectation. Consider a three allele case:

allozyme 1 has frequency $= p$
allozyme 2 has frequency $= q$
and the null allele has frequency $= r$.

With $f = 0$ in the population we have a frequency of electrophoretic heterozygotes of $2pq/(1-r^2)$ and an expected frequency based on the sample allele frequencies [$p = p/(1-r)$ and $q = q/(1-r)$] of $2pq/(1-r)^2$. Assuming that individuals homozygous for the null are ignored, we can calculate the relationship between the null allele frequency and the estimate of the within individual correlation, $f = 2r/(1+r)$. For our mean value of $\bar{f}_1 = 0.033$ the required fre-

quency of a null allele would be 0.017. This does not seem to us to be an unreasonably large value. Even the high *Acph* mean f_1 value can be explained by a null frequency of 0.061. This would predict a null homozygote frequency of 0.004. This frequency of individuals that gave no enzyme activity would go undetected in samples of the size in this study.

SINGH (1976) has suggested that the frequency of nulls might indeed be lower for allozymes with single substrates, Group I, as compared to those with multiple substrates, Group II (see GILLESPIE and KOJIMA 1968 and GILLESPIE and LANGLEY 1974). Table 8 shows the mean (over loci and samples) of f_1 for Group I (*αGpdh*, *Mdh*, and *Pgm*) and Group II. Group II shows definitely higher f_1 values than Group I, except for the cages. Even if we were to remove *Acph*, the means would remain in the same relative direction. Using the three largest independent collections (Raleigh, not Raleigh, and East Coast) we tested the significance of this trend using the Mann-Whitney U test. The pooled z scores from this test on the f_1 values is significant ($P \leq 0.02$).

If the f_1 values are due to population structure, either in sampling or mating, we have two available interpretations. The first would be some form of positive assortative mating or inbreeding such that individuals of similar genotypes simply mate more frequently than expected. The second is that the samples represent subpopulations with internal subdivisions, "isolates" (COCKERHAM 1973). In this case the f_1 values are the sums of the correlation of alleles within individuals and the correlation of alleles within isolates. It is difficult to rule out one or the other of these phenomena as the cause of the positive f_1 values. The lack of any equivalent level of correlation at the higher levels of the hierarchical structure suggests that the f_1 values may represent a local mating structure or very transient isolate formation.

Much of the theory of population subdivision assumes constant migration patterns. It is quite likely that natural populations of Drosophila have seasonal variation in average migration as food availability varies and also substantial variance in migration distance per individual per generation. Since no appropriate model is available, not to mention appropriate observations of migration,

TABLE 8

Comparison of Group I with Group II enzymes

f statistics for	Group I Mean	Group I *se*	Group II Mean	Group II *se*	Mann-Whitney U test
North Carolina	0.0080	0.0108	0.0486	0.0174	
— Raleigh	0.0192	0.0019	0.0677	0.0171	$P=0.018$ ⎫
— not Raleigh	0.0061	0.0128	0.0432	0.0182	$P=0.098$ ⎬ $P \leq 0.02$
East Coast	0.0194	0.0061	0.0353	0.0104	$P=0.196$ ⎭
— North	0.0061	0.0047	0.0200	0.0156	
— South	0.0275	0.0053	0.0564	0.0152	
Indoors	0.0099	0.0058	0.0338	0.0187	
Miami cage	0.0234	0.0082	0.0024	0.0090	

we must conclude that the correlations within individuals and within samples are not inconsistent with random drift and migration of neutral alleles.

Most of the variation among samples is seen between samples from the same locality at any one time. This variation is not large in magnitude. There is not an obvious experimental source of such variation. Perhaps it can be interpreted as reflecting the magnitude of local population variation in frequency due either to simple drift or transient selection at loci linked to allozyme loci.

The variation in gene frequencies over North Carolina is also small. Nevertheless clines can be detected. Some of these can be recognized as parts of well known latitudinal, continental clines (such as *Adh* and *Est-C*), while others show longitudinal clines (*Est-6* and *Acph*) that have not been previously reported. The role of polymorphic inversions in allozymic variation is not altogether clear. It may be that inversions cannot account for much of the allozyme frequency variation (R. A. VOELKER and C. C. COCKERHAM, personal communication).

We also found changes in gene frequency seasonally and with time (with successive months). Again the simplest interpretation may be seasonal variation in the frequency of *In2Lt* to which Adh^S is completely linked. The analysis in Table 2 indicated a clear year-to-year variation in gene frequencies at several loci and the correlation analyses in Tables 7a and 7b indicate some directional changes. Most interesting, perhaps, is the simultaneous increase of *αGdph*, *Mdh*, and *Adh* in the Raleigh area over the four year period. Since these three loci are frequently observed to be in linkage disequilibrium with *In2Lt* such a concerted change might be expected. The frequency of *In2Lt* does appear to have increased in Raleigh during the period of this study (R. A. VOELKER, personal communication).

Another type of cline and seasonal association we expected to detect was that with f_1. If the correlation of alleles within individuals were at all associated with population size, we expected f_1 to increase in the north and in the spring. This was not observed. Perhaps the positive f_1 are due solely to null alleles. Alternatively we could admit we know even less than we had hoped about the natural history and demography of *D. melanogaster*.

We included some data from indoor collections and a cage for comparison. The indoor collections are well within the range of frequencies observed in the surrounding natural populations. The f_1 values do not appear to be any larger than those in natural populations. These indoor populations may have been recently established by migrants from the natural populations.

The cage data are interesting in two ways. First the f_1 values are not larger, if as large, as those in the population from which it was established. This suggests that inbreeding is not more common in the cage. Also we would expect rare nulls to be present in the cage if they are causing the positive f_1 values in the natural populations. The cage does not look so different from the natural populations on the sample level, *i.e.*, the f_1 values are not significantly different. But the variation in gene frequency still occurring after a year in the cage suggests considerable selection. This cage was founded from over 250 isofemale lines and maintained in discrete generations without any obvious bottlenecks. The

most plausible explanation might be that there was considerable selection on the polymorphic inversions in the founding sample and the allozyme frequency changes reflect linkage disequilibria between allozymes and linked inversions. In any case it suggests that newly established cages may not come to equilibrium rapidly and that considerable natural selection occurs during this process.

The analysis of allozyme frequency variation in natural populations has not proven critical in resolving fundamental questions about the biological significance of this variation. The analysis reported here shows clearly that *D. melanogaster* is only slightly structured. Individuals *appear* to be mildly inbred. The source of the apparent inbreeding may be an error in classification due to null alleles. Geographic and temporal trends are apparent. The trends may be attributed to linkage disequilibria with polymorphic inversions or other closely linked and strongly selected polymorphic loci.

We wish to thank Dr. C. Clark Cockerham and Dr. Bruce S. Weir for their helpful discussion and criticism.

LITERATURE CITED

Christiansen, F. B., O. Frydenberg and V. Simonsen, 1973 Genetics of *Zoarces* populations. IV. Selection component analysis including mother-offspring combinations. Genetics **76:** 339–366.

Cockerham, C. Clark, 1969 Variance of gene frequencies. Evolution **23:** 72–84. ——, 1973 Analysis of gene frequencies. Genetics **74:** 679–700.

Crow, J. F. and M Kimura, 1970 *An Introduction to Population Genetics Theory.* Harper and Row, New York.

Crow, J. F. and R G. Temin, 1964 Evidence for the partial dominance of recessive lethal genes in natural populations of Drosophila. Am. Naturalist **98:** 21–33.

Gillespie, J. H. and K. Kojima, 1968 The degree of polymorphisms in enzymes involved in energy production compared to that in non-specific enzymes in two *Drosophila ananassae* populations. Proc. Nat. Acad. Sci. **61:** 582–585.

Gillespie, J. H. and C. H. Langley, 1974 A general model to account for enzyme variation in natural populations. Genetics **76:** 837–884.

Hiraizumi, Y. and J. F. Crow, 1960 Heterozygous effects on viability, fertility, rate of development, and longevity of Drosophila chromosomes that are lethal when homozygous. Genetics **45:** 1071–1083.

Johnson, F. M. and P. M. Burrows, 1976 Isozyme variability in species of the genus *Drosophila* VIII: The alcohol dehydrogenase polymorphism in North Carolina populations of *Drosophila melanogaster.* Biochem. Genet. **14:** 47–58.

Johnson, F. M. and H. E. Schaffer, 1973 Isozyme variability in species of the genus Drosophila. VII: Genotype-environment relationships in populations of *Drosophila melanogaster* from the Eastern United States. Biochem. Genet. **10:** 149–163.

Singh, R. S., 1976 Substrate-specific enzyme variation in natural populations of *Drosophila pseudoobscura.* Genetics **82:** 507–526.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Corresponding editor: J. F. Kidwell