A MOLECULAR APPROACH TO THE STUDY OF GENIC HETEROZYGOSITY IN NATURAL POPULATIONS. II. AMOUNT OF VARIATION AND DEGREE OF HETEROZYGOSITY IN NATURAL POPULATIONS OF *DROSOPHILA PSEUDOOBSCURA*[1]

R. C. LEWONTIN AND J. L. HUBBY

Department of Zoology, University of Chicago, Chicago, Illinois

Received March 30, 1966

AS pointed out in the first paper of this series (HUBBY and LEWONTIN 1966), no one knows at the present time the kinds and frequencies of variant alleles present in natural populations of any organism, with the exception of certain special classes of genes. For human populations we know a good deal about certain polymorphisms for blood cell antigens, serum proteins, and metabolic disorders of various kinds but we can hardly regard these, *a priori*, as typical of the genome as a whole. Clearly we need a method that will randomly sample the genome and detect a major proportion of the individual allelic substitutions that are segregating in a population. In our previous paper, we discussed a method for accomplishing this end by means of a study of electrophoretic variants at a large number of loci and we showed that the variation picked up by this method behaves in a simple Mendelian fashion so that phenotypes can be equated to homozygous and heterozygous genotypes at single loci.

It is the purpose of this second paper to show the results of an application of the method to a series of samples chosen from natural populations of *Drosophila pseudoobscura.* In particular, we will show that there is a considerable amount of genic variation segregating in all of the populations studied and that the real variation in these populations must be greater than we are able to demonstrate. This study does not make clear what balance of forces is responsible for the genetic variation observed, but it does make clear the kind and amount of variation at the genic level that we need to explain.

An exactly similar method has recently been applied by HARRIS (1966) for the enzymes of human blood. In a preliminary report on ten randomly chosen enzymes, HARRIS describes two as definitely polymorphic genetically and a third as phenotypically polymorphic but with insufficient genetic data so far. Clearly these methods are applicable to any organism of macroscopic dimensions.

The Populations Studied

We have chosen populations of *D. pseudoobscura* for a number of reasons. This species is not commensal with man, as is *D. melanogaster*, and so can be

[1] The work reported here was supported in part by grants from the National Science Foundation (GB 3112 and GB 3213) and the Public Health Service (GM 11216).

said to be truly "wild." It has a wide distribution in Western North and Central America from British Columbia to Guatemala with a recently discovered outlier as far south as Bogotá, Colombia. *D. pseudoobscura* is genetically well known, at least to the extent of having marker genes and inversions on all of its four major chromosomes, and there exists a vast literature on the population genetics of the inversion systems on chromosome 3 of this species by DOBZHANSKY and his school. No species of Drosophila is really well understood in its ecological aspects, but for *D. pseudoobscura* 30 years of study of natural populations has led to a fair knowledge of population size fluctuation, kind of vegetation with which the species is associated, diurnal activity and temperature tolerance. Numerous samples from wild populations exist in the laboratory, and new samples are constantly becoming available. All of these reasons suggested to us that *D. pseudoobscura* would be a good species for our first survey of natural genic variation. It seemed to us that the variation found within and between populations of this species ought to be typical of a common, relatively widespread, sexually reproducing organism.

The populations in this study are represented by a number of separate lines each stemming from a single fertilized female caught in nature. For example, nine separate single-female lines maintained separately in the laboratory since 1957 represent the population from Flagstaff, Arizona. Because we were unable to get fresh samples (except for one case) we preferred these separate lines to any mixed population. Such separate lines may each suffer homozygosis because of inbreeding, but the differences *between* lines will preserve some portion of the original population variance. If the lines had been pooled and kept since 1957 as a mixture, more of the variability originally introduced would have been lost. As our results will show, most, but not all, lines are in fact homozygous but differences between lines have been preserved. Nevertheless, the loss of variation because of inbreeding needs to be kept in mind when we analyze the results.

The population samples in the study were as follows: (1) Flagstaff, Arizona. Nine lines collected in a ponderosa pine forest above 5,000 feet elevation in 1957. The natural population is virtually pure for the Arrowhead gene arrangement on the third chromosome and all lines are Arrowhead homozygotes (see DOBZHANSKY and EPLING 1944). (2) Mather, California. Seven lines collected between 1957 and 1960 in a Transition Zone forest at 4,600 feet elevation. This population is highly polymorphic for third chromosome inversions in nature. All strains used were homozygous Arrowhead (see DOBZHANSKY, 1948). (3) Wildrose, California. Ten strains collected in 1957 in the Panamint Range at 8,000 feet elevation in a piñon Juniper forest. The population is highly polymorphic for inversions, but the strains tested were all homozygous Arrowhead (see DOBZHANSKY and EPLING 1944). (4) Cimarron, Colorado. Six lines collected in a *Quercus gambelii* grove at about 7,000 feet elevation in 1960. All lines are homozygous Arrowhead. (5) Strawberry Canyon (Berkeley), California. Ten strains from a much larger collection made in 1965 at an elevation of 800 feet. This population is highly polymorphic for third chromosome inversions, and the strains used were also polymorphic, being the F_2 and F_3 from the wild females. (6) A single strain from Bogotá, Colombia. A much larger sample is planned for this extreme outlier of the species range, but the single strain collected in 1960 was included since it was available. The population occurs between 8,000 and 10,000 feet elevation and has two inversions, Santa Cruz and Treeline in proportions 65:35 (see DOBZHANSKY *et al.* 1963).

The natural and laboratory history of these various strains is thus rather different. Two, Cimarron and Flagstaff, are from the eastern part of the species distribution where chromosomal (inversion) variability is low. All but Strawberry Canyon have been in the laboratory for 5 to 8 years' as separate strains, while Strawberry Canyon is a fresh sample from nature, and is polymorphic for inversions. One strain, Bogotá, represents a geographically remote population that surely represents the extreme southern part of the species distribution. All in all, the sample was chosen to give a diversity of histories so that the results could be given some generality.

The laboratory maintenance of all strains was the same. They were kept at 18°C in half-pint culture bottles with an average of about 50 parents each generation, but with considerable variation in size. At times in their culture, most, if not all, suffered one or more extreme breeding size bottlenecks. Thus, there has been inbreeding to an unknown extent. At the culture temperature of 18°C, there is little or no difference in selective values among third chromosome inversion types, although nothing can be said in this respect about other segregating gene systems.

RESULTS

The methods of electrophoretic separation and detection of enzyme systems are fully explained by HUBBY and LEWONTIN (1966) and we will take it as demonstrated in that paper that the phenotypes we see are reflective of simple allelic substitutions at single genetic loci. Therefore, in what follows in this paper, we will refer to "alleles" and "loci" without again referring to the phenotypic appearance of the electrophoretic gels.

In every case, five or more individuals were tested from each strain. A strain is classified as homozygous for an allele if all individuals tested were homozygous, while the strain is classified as segregating for two alleles if any of the individuals was heterozygous or if homozygotes of two different kinds were found. The notation *.95/1.07*, for example, means that the allele *.95* and the allele *1.07* for a gene were found segregating among the tested individuals of the strain. Throughout we use the relative electrophoretic mobilities as names of alternate alleles (see HUBBY and LEWONTIN 1966).

The observations are summarized in Table 1. The body of the table shows the number of strains (not individuals) either homozygous or segregating for various alleles at various loci. Of the ten enzyme systems discussed in HUBBY and LEWONTIN (1966), two (*ap-1* and *ap-2*) are not included here because they appeared on the gels infrequently and are not sufficiently reliable to be used in a population study. For the same reason, only ten of the 13 larval proteins are included in the present study. The decision whether to include a band in the study was made solely on the basis of reliability, and independently of whether it showed electrophoretic variants.

The entry in Table 1 for Leucine aminopeptidase (*lap*) is different in meaning from the others. The relative mobilities of the variant forms are so close for this locus that it is not possible to make the proper cross assignments between populations. There are at least four alleles at the locus, but we do not at present know unambiguously which are present in which populations. Therefore, in Table 1 we have simply indicated how many alleles are present among the strains of that population.

Table 1 shows some remarkable results. First, of the 18 loci represented, there is some genetic variation in some population for nine of them. Second, genetic variation is found in more than one population for seven of the loci: malic dehydrogenase (*mdh*), esterase-5 (*e-5*), leucine aminopeptidase (*lap*), alkaline phosphatase-7 (*ap-7*), *pt-7*, *pt-8* and *pt-10*. This variation in more than one popu-

TABLE 1

Number of strains from each population either homozygous or segregating for various alleles at different loci

Locus	Allele	Strawberry Canyon	Wildrose	Cimarron	Mather	Flagstaff	Bogotá
esterase-5	*.85*	0	0	0	1	0	0
	.95	0	1	0	1	1	0
	1.00	0	3	3	0	4	1
	1.03	0	1	0	2	0	0
	1.07	0	0	2	1	4	0
	1.12	0	1	0	2	0	0
	.95/1.00	1	0	0	0	0	0
	.95/1.07	1	0	0	0	0	0
	.95/1.12	0	0	1	0	0	0
	1.00/1.07	4	1	0	0	0	0
	1.00/1.12	3	1	0	0	0	0
	1.03/1.07	1	1	0	0	0	0
	1.03/1.12	0	1	0	0	0	0
	1.07/1.12	1	0	0	0	0	0
malic dehydrogenase	*.90*	0	0	0	1	0	0
	1.00	6	10	6	4	8	1
	1.11	2	0	0	0	0	0
	1.22	0	0	0	0	1	0
	.90/1.00	0	0	0	2	0	0
	1.00/1.11	2	0	0	0	0	0
glucose-6-phospate dehydrogenase	*1.00*	9	10	4	6	9	1
alkaline phosphatase-4	*.93*	0	0	0	0	1	.
	1.00	9	11	6	7	8	.
alkaline phosphatase-6	+	9	10	5	7	9	.
	—/+	0	0	1*	0	0	.
alkaline phosphatase-7	+	9	9	5	7	9	.
	—/+	0	1	1*	0	0	.
α-glycerophosphate dehydrogenase	*1.00*	10	10	6	6	8	1
leucine aminopepidase	*.95* *.97* *1.00* *1.02*	2† alleles	3‡ alleles	2 alleles	2§ alleles	3 alleles	1 allele
pt-4	*.45*	10	10	6	6	8	1
pt-5	*.55*	1	4	4	6	2	1
pt-6	*.62*	10	10	6	6	8	1
pt-7	*.73*	0	0	0	0	1	0
	.75	9	10	5	5	6	1
	.77	0	0	0	0	0	0
	.73/.75	0	0	0	0	1	0
	.75/.77	1	0	1	1	0	0
pt-8	*.80*	0	0	0	0	0	1
	.81	2	2	3	2	1	0
	.83	1	4	1	1	5	0
	.81/83	7	4	2	3	2	0
pt-9	*.90*	3	8	4	1	0	0

TABLE 1—Continued

Number of strains from each population either homozygous or segregating for various alleles at different loci

Locus	Allele	Strawberry Canyon	Wildrose	Cimarron	Mather	Flagstaff	Bogotá
pt-10	*1.02*	0	0	0	0	0	0
	1.04	4	9	6	4	8	0
	1.06	0	0	0	0	0	1
	1.02/1.04	0	1	0	0	0	0
	1.04/1.06	6	0	0	2	0	0
pt-11	*1.12*	4	10	6	6	8	.
pt-12	*1.18*	5	10	6	6	8	1
pt-13	*1.30*	7	10	6	6	8	1

* Both loci segregating in the same strain.
† *Three strains segregating.*
‡ One strain segregating.
§ Two strains segregating.

lation must be characterized as polymorphism in the usual sense because variant alleles occur with some appreciable frequency in more than an isolated case.

Third, and most remarkable of all, is the widespread occurrence of segregation in strains that have been in the laboratory for as many as seven years. As might be expected, the Strawberry Canyon strains are segregating at those loci that are polymorphic. In fact, not a single strain of Strawberry Canyon is homozygous for an allele of *e-5*. But four strains of Wildrose are also segregating for alleles at this locus, as is one strain of Cimarron. Most striking of all is the case of the *.81/.83* polymorphism at the *pt-8* locus where there are segregating strains in every population (not including the single strain from Bogotá). Despite the segregation at many of these loci, Table 1 definitely gives the impression of an effect of inbreeding over the many generations during which the strains have been maintained in the laboratory. The Strawberry Canyon strains segregate far more frequently than any of the others, and, in general, more of the genetic variation in the other populations is between homozygous strains.

Fourth, the genotype of the single strain from Bogotá is sometimes unusual. In most cases, the Bogotá strain is homozygous for the allele most commonly found in other localities. This is not the case for *pt-8*, however, where Bogotá is homozygous for an allele not found elsewhere, and *pt-10* where Bogotá is homozygous for one of the less common alleles.

In order to make the pattern of genic variation simpler to perceive, Table 2 has been constructed from the data in Table 1. In Table 2 *very approximate* gene frequencies are calculated for the alleles shown in Table 1 by using the following convention. Each of the original strains carried four independent doses of each gene when it was brought into culture. A large proportion of the strains still have more than one of these original doses since so many strains are still polymorphic and therefore carry at least two of the original four alleles. How many of the original alleles are still represented in any strain can only be guessed at, however. We make an arbitrary convention that each line shall be counted

TABLE 2

Approximate gene frequencies calculated from the data of Table 1

Locus	Allele	Strawberry Canyon	Wildrose	Cimarron	Mather	Flagstaff	Bogotá
esterase-5	*.85*	0	0	0	.14	0	0
	.95	.09	.10	.08	.14	.11	0
	1.00	.36	.40	.50	0	.44	x
	1.03	.05	.20	0	.29	0	0
	1.07	.32	.10	.33	.14	.44	0
	1.12	.18	.20	.08	.29	0	0
malic dehydrogenase	*.90*	0	0	0	.29	0	0
	1.00	.70	1.00	1.00	.71	.89	x
	1.11	.30	0	0	0	0	0
	1.22	0	0	0	0	.11	0
glucose-6-phosphate dehydrogenase	*1.00*	1.00	1.00	1.00	1.00	1.00	x
alkaline phosphatase-4	*.93*	0	0	0	0	.11	.
	1.00	1.00	1.00	1.00	1.00	.88	.
alkaline phosphatase-6	+	1.00	1.00	.92	1.00	1.00	.
	—	0	0	.08	0	0	.
alkaline phosphatase-7	+	1.00	.95	.92	1.00	1.00	.
	—	0	.05	.08	0	0	.
α-glycerophosphate dehydrogenase	*1.00*	1.00	1.00	1.00	1.00	1.00	x
leucine aminopeptidase	*.95* *.97* *1.00* *1.02*	2 alleles	3 alleles	2 alleles	2 alleles	3 alleles	1 allele
pt-4	*.45*	1.00	1.00	1.00	1.00	1.00	x
pt-5	*.55*	1.00	1.00	1.00	1.00	1.00	x
pt-6	*.62*	1.00	1.00	1.00	1.00	1.00	x
pt-7	*.73*	0	0	0	0	.19	0
	.75	.95	1.00	.92	.92	.81	x
	.77	.05	0	.08	.08	0	0
pt-8	*.80*	0	0	0	0	0	x
	.81	.55	.40	.67	.58	.25	0
	.83	.45	.60	.33	.42	.75	0
pt-9	*.90*	1.00	1.00	1.00	1.00	1.00	x
	1.02	0	.05	0	0	0	0
pt-10	*1.04*	.70	.95	1.00	.83	1.00	0
	1.06	.30	0	0	.17	0	x
pt-11	*1.12*	1.00	1.00	1.00	1.00	1.00	x
pt-12	*1.18*	1.00	1.00	1.00	1.00	1.00	x
pt-13	*1.30*	1.00	1.00	1.00	1.00	1.00	x

One strain = 2 alleles. No gene frequency estimate can be made for Bogotá, so the allele present is marked with an x.

equally and, since many of the strains are segregating, each allele in such lines is given a weight of one half. So, for example, in Strawberry Canyon, for locus *pt-8*, there are two strains homozygous for allele *.81*, seven strains segregating *.81*/*.83*, and one strain homozygous *.83*. Then the gene frequency of allele *.81* is $q_{.81} = (2 + 7/2)/(2 + 7 + 1) = .55$. Such a method can give only a very crude estimate of the frequency of alleles in the original sample brought into the laboratory, except for Strawberry Canyon where the sample was examined in the F_2 and F_3 generations from the wild. Since these original samples were themselves small, we cannot take our gene frequency estimation in Table 2 too seriously. They are meant only to give a qualitative picture of the variation, yet they show certain patterns and on the basis of these crude estimates we can characterize the variation at each locus as falling into certain broad categories.

1. *Monomorphism.* In a sufficiently large population, no locus can be completely without variant alleles. However, we class as monomorphic those loci that are without variation in our sample and those with only a single variant allele in a single strain. It might be argued that the presence of even a single variant allele in such a small sample as ours is evidence that in the population this variant is at reasonably high frequency. Nevertheless, we prefer to err on the side of conservatism and class such isolated variants as newly arisen mutations that have not yet been eliminated from the population by natural selection or genetic drift. Using the criterion that a variant must be present in more than one strain in more than one population in order for a population to be considered polymorphic, we find 11 out of 18 loci monomorphic. Of these, nine are completely without variation in our sample: glucose-6-phosphate dehydrogenase, α-glycerol phosphate dehydrogenase, *pt-4*, *pt-5*, *pt-6*, *pt-9*, *pt-11*, *pt-12*, and *pt-13*. The other two, alkaline phosphatase-4 and alkaline phosphatase-6 each have a single variant allele in a single strain. In the case of alkaline phosphatase-4, the strain is homozygous for the variant allele so it is likely that it has been in the strain for some time, probably from the original sample from the wild. Nevertheless, we do not count this locus as polymorphic.

2. *Widespread polymorphism with one allele in high frequency.* In this class there are three loci in our sample: *ap-7* which has the same variant allele in two different geographical regions but in low frequency, *pt-7* which is similar, but has the polymorphism more widespread and which also has a second variant allele restricted to one population, and *pt-10* which is like *pt-7* except that the rarer allele is found fixed in the Bogotá strain. These three loci are clearly polymorphic, but one allele in each case is found in high frequency in every population and so may be considered the "typical" allele. For *pt-10* the "type" concept is shaky since in Strawberry Canyon the atypical allele is in a frequency of 30% and the allele is fixed in the single Bogotá strain.

3. *Ubiquitous polymorphism with no wild type.* This class includes three loci. The most extreme case is the esterase-5 gene which has six alleles so far recovered. Populations are segregating for between three and five of these and no one allele is most common. Allele *1.00* comes close to being most common, but it is completely lacking in the Mather sample. Only one allele, *.85*, is restricted to a single population, all others being found in a minimum of three populations. *pt-8* has about a 50:50 polymorphism of alleles *.81* and *.83* in all the populations and this is related to the fact that all populations had some strains still segregating for these two alleles. In addition, *pt-8* has a unique allele in Bogotá. *Leucine amino-*

peptidase appears to fall in this group, although there is some suspicion, not yet confirmed, that allele *1.00* is most common in all populations.

4. *Local indigenous polymorphism.* Only one locus is completely of this sort, malic dehydrogenase. Three of the five populations have a local variant in high frequency, but it is a different variant in each case. Allele *1.00* would appear to be a "type" allele or at least a most common form. In addition to *mdh*, we have already noted an occasional local variant, such as the allele *.80* of *pt-8* in the Bogotá strain, the allele *.73* of *pt-7* found only in Flakstaff, and the allele *.85* of esterase-5 known only from Mather. In these last two cases, it is impossible to distinguish them from the single homozygous variant of alkaline phosphatase-4 which we have classed as nonpolymorphic.

5. *Local pure races.* A class of variation that is completely lacking in our sample of loci is the local pure race. In no case do we find some populations homozygous for one allele and other populations homozygous for a different one. We expect such a pattern if the alleles were functionally equivalent isoalleles not under any natural selection pressure. The failure to find such cases is important to our hypotheses about the forces responsible for the observed variation.

To sum up these classes, out of 18 loci included in the population study, seven are clearly polymorphic in more than one population and two are represented by rare local variants in a single population which, to be conservative, are not considered polymorphic. Thus, conservatively 39% of loci are polymorphic. This takes account of all populations and does not give an estimate of the polymorphism in any given population, which will be less. Table 3 is a summary of the information for each population separately. The populations are very similar to each other in their degree of polymorphism with an average of 30% of the loci varying in each. It is interesting that Strawberry Canyon, a fresh sample from the wild, is not different from the others. We can assume that most of the variation from nature has been preserved in the laboratory stocks but has been converted to variation between strains by the inbreeding attendant on laboratory culture. Another point of interest is that the great similarity in *proportion* of loci polymorphic in each population is not entirely a result of identity of poly-

TABLE 3

Proportion of loci, out of 18, polymorphic and proportion of the genome estimated to be heterozygous in an average individual for each population studied

Population	No. of loci polymorphic	Proportion of loci polymorphic	Proportion of genome heterozygous per individual	Maximum proportion of genome heterozygous
Strawberry Canyon	6	.33	.148	.173
Wildrose	5	.28	.106	.156
Cimarron	5	.28	.099	.153
Mather	6	.33	.143	.173
Flagstaff	5	.28	.081	.120
Average	.	.30	.115	.155

morphisms. Thus, although Wildrose and Flagstaff are both polymorphic at five out of 18 loci, only three of these are common to both populations. Flagstaff is polymorphic at two loci, *mdh* and *pt-7*, for which Wildrose is monomorphic, but Wildrose is polymorphic for *ap-7* and *pt-10*, while Cimarron is monomorphic at these loci.

Yet another question that can be asked from the data is, "At what proportion of his loci will the average individual in a population be heterozygous?" In fact, this can be described without exaggeration as the central problem of experimental population genetics at the present time. A complete discussion of the conflicting results on this question is not possible here, but the issue is very clearly drawn by WALLACE (1958). The results reported by WALLACE in that paper, in previous papers (WALLACE 1956) and in subsequent works by WALLACE (1963), WALLACE and DOBZHANSKY (1962), DOBZHANSKY, KRIMBAS and KRIMBAS (1960), and many others, all point, although indirectly, toward a high level of heterozygosity in natural populations. On the other hand, theoretical considerations by KIMURA and CROW (1964) and experiments of HIRAIZUMI and CROW (1960), GREENBERG and CROW (1960), MULLER and FALK (1961) and FALK (1961) among others, point in the opposite direction. These latter authors interpret their results as showing that the proportion of loci heterozygous in a typical individual from a population will be quite small and that polymorphic loci will represent a small minority of all genes.

Our data enable us to estimate the proportion of heterozygosity per individual directly. This is estimated in the next to the last column of Table 3 for each population separately. This estimate is made by taking the gene frequencies of all the alleles at a locus in a population, calculating the expected frequencies of heterozygotes from the Hardy-Weinberg proportions, and then averaging over all loci for each population separately. For example, at the *e-5* locus in Flagstaff there are three alleles at frequency .44, .44, and .11, respectively. The expected frequency of heterozygotes at this locus in Flagstaff is then given by:

$$\textit{Proportion heterozygotes} = 2(.11)(.44) + 2(.11)(.44) + 2(.44)(.44) = .581.$$

This value is then averaged with similarly derived values from each of the other loci for Flagstaff, including the monomorphic ones which contribute no heterozygosity. Obviously, for a given number of alleles the proportion of heterozygosity is maximized when all are in equal frequency. In such a case

$$\textit{maximum proportion heterozygosity} = (n-1)/n$$

where n is the number of alleles present. This value is given for comparison in the last column of Table 3.

As Table 3 shows, between 8% and 15% of the loci in an average individual from one of these populations will be in a heterozygous state and this is not very different from the maximum heterozygosity expected from the number of alleles actually segregating in the population. It is interesting that the two populations with the lowest amount of chromosomal polymorphism, Flagstaff and Cimarron (DOBZHANSKY and EPLING 1944) also have a slightly lower genic heterozygosity than the chromosomally highly polymorphic populations of Mather, Strawberry Canyon, and Wildrose. More extensive data on chromosomally polymorphic and monomorphic populations are being taken now.

DISCUSSION

Biases: Before we attempt to explain the amount of polymorphism shown in Table 3, we need to ask what the biases in our experiment are. There are four sources of bias in our estimates and they are all in the same direction.

1. The method of electrophoretic separation detects only some of the differences between proteins. Many amino acid substitutions may occur in a protein without making a detectable difference in the net charge. We do not know what proportion of substitutions we are detecting but it is probably on the order of one half. Depending upon the protein, different results have been observed. For tryptophan synthetase about 7/9 of all mutations tested are electrophoretically detectable (HENNING and YANOFSKY 1963), but none of the forms of cytochrome-c are electrophoretically separable despite extensive amino-acid substitution over the plant and animal kingdoms (MARGOLIASH, personal communication). Presumably in the latter case, net charge is critical to proper function. At any rate, our estimate of the number of variant alleles is clearly on the low side.

2. Our lines have preserved only a portion of the variation originally present in them when they were taken from nature. Because of the inbreeding effect of maintaining small populations with occasional bottle necks in breeding size, some of the alleles originally present must have been lost. This causes our estimate of variation to be on the low side.

3. The original lines were only a small sample of the natural populations. We have tested very few lines, as few as six in the case of Cimarron, so that we are only sampling a portion of the natural variation. Alleles at frequencies of say 5% or 10% may easily be lacking in such samples. Again our experiment underestimates the variation within each population.

4. We have deliberately excluded as polymorphic two loci in which only a single variant allele was found. This coupled with the fact that only five individuals were surveyed in each strain will leave out of account real polymorphisms at low frequencies. Had we included the two rare variants in Table 3, both Cimarron and Flagstaff would have had 33% of loci polymorphic which would change the overall average to 32%. The proportion of loci heterozygous per individual in these populations would be increased from .09 and .081 to .107 and .092, respectively, bringing the average over all populations to 12%, a very small change.

All these sources of bias cause us to underestimate the proportion of loci polymorphic and the proportion of heterozygous loci per individual, but by how much we cannot say. At present we are studying a large sample of over 100 F_1 lines from females caught in Strawberry Canyon over the course of a year. This study will eliminate biases 2 and 3 above and give us an appropriate correction for our present estimates.

One other possible source of bias is in the choice of enzyme assays. If there were some subtle reason that the enzymes we have chosen to use tended to be more or less genetically variable than loci in general, our results would not be referable to the genome as a whole. Our chief protection against this sort of bias is in the use of the larval proteins in addition to the specific enzyme assays. Both of these classes of genes give about the same degree of polymorphism: three out of ten polymorphic loci for larval proteins and four out of eight for the enzymes. While it might be argued that the very existence of a published method for the detection of an enzyme on a gel is a bias in favor of variable enzymes, no such argument can be made for the larval proteins, all of which are developed on the same gel by a general protein stain. Moreover, two of the enzymes, malic dehydrogenase and α-glycerophosphate dehydrogenase, were developed in this laboratory simply because suitable coupling methods are known for dehydrogenases.

In order to avoid the bias that might arise from considering only a particular enzyme function, we have deliberately not assayed a large number of proteins associated with similar functions. For example, there are ten different sites of esterase activity, presumably representing ten different genes, but we have only assayed the one with the greatest activity. To load our sample with more esterases might introduce a bias if there were some reason why esterase loci were more or less polymorphic than other genes.

The source of the variation: It is not possible in this paper to examine in detail all of the alternative explanations possible for the large amount of genic variation we have observed in natural populations. Our observations do require explanation and we already have some evidence from the observations themselves.

Genetic variation is destroyed by two forces: genetic drift in populations under going periodic size reductions and selection against recessive or partly dominant deleterious genes. Genetic variation is increased or maintained by three factors: mutation, migration between populations with different gene frequencies, and balancing selection usually of the form of selection in favor of heterozygotes. On the basis of combination of these factors, we can distinguish three main possibilities to explain the variation we have seen.

(1) The alleles we have detected have no relevance to natural selection but are adaptively equivalent isoalleles. In such a case, genetic drift will drive populations to homozygosity, but will be resisted by recurrent mutation and migration. We have some idea of the effective breeding size, N, in populations of *D. pseudoobscura* from the experiments of DOBZHANSKY and WRIGHT (1941, 1943) and WRIGHT, DOBZHANSKY, and HOVANITZ (1942). Various estimates agree that "panmictic unit" has an effective size, N, of between 500 and 1,000 in the Mount San Jacinto populations where the species is most dense and successful. At Wildrose the population size is between one-fifth and one-tenth of that at Mount San Jacinto and, although there is no published evidence, the same is true at Cimarron where flies are rare even in summer. For the dense populations the conclusion of DOBZHANSKY and WRIGHT (1943) is that "the effective size of the panmictic unit in *D. pseudoobscura* turns out to be so large that but little permanent differentiation can be expected in a continuous population of this species owing to

genetic drift alone." For Cimarron and Wildrose, however, this is not true, yet we find these populations with the same average heterozygosity as other populations. The lack of any loci showing pure local races in nature is against the selective equivalence of isoalleles. It can be argued, however, that genetic drift in the marginal populations is producing local pure races but that migration from the other populations and mutation (of unknown magnitude for these alleles) is preventing differentiation. As a matter of fact, very little migration, of the order of one individual per generation, will effectively prevent homozygosis by drift. We must also take account of the observation that many lines in the laboratory are still segregating for several loci and that effective population size of these lines has been very small and migration (contamination) close to nil. The continued segregation of alleles in the laboratory might be caused by mutation rates much higher for isoalleles than for dysgenic alleles, and we are checking the mutation rate for a few alleles. All in all, however, complete selective neutrality is not a satisfactory explanation of all the observations.

(2) Selection tends to eliminate alternative alleles but mutation restores them. This hypothesis comes close to the neutral isoallele theory because our observed gene frequencies of alternate alleles would require that mutation rates and selection coefficients be of the same order of magnitude. That is, the equilibrium gene frequency for an allele selected against with intensity t in heterozygotes (we can ignore the rarer homozygotes) is approximately equal to u/t, where u is the mutation rate. Since our rarer alleles at each locus vary in frequency from 5% to 45%, u and t must be of about the same order of magnitude. This in turn suggests extraordinarily high mutation rates or very, very weak selection *on the average*. But an average selection coefficient of .001 implies that in some populations at some times the gene in question is selected for rather than against so that local pure race formation should be promoted. Again we must check to see that mutation rates are not higher than 10^{-3}.

(3) Selection is in favor of heterozygotes. This hypothesis satisfies all the objections to (1) and (2) above, since heterosis, if strong enough, can maintain genic variation in any size population, irrespective of mutation and migration. However, two different problems are raised by the assumption of nearly universal heterosis. First, unless we assume that the two homozygotes are very weakly selected against, in which case we are back effectively to alternatives (1) and (2), the total amount of differential selection in a population with many heterotic loci is tremendous. For example, suppose two alleles are maintained by selecting against both homozygotes to the extent of 10% each. Since half of all individuals are homozygotes at such a locus, there is a loss of 5% of the population's reproductive potential because of the locus alone. If our estimate is correct that one third of all loci are polymorphic, then something like 2,000 loci are being maintained polymorphic by heterosis. If the selection at each locus were reducing population fitness to 95% of maximum, the population's reproductive potential would be only $(.95)^{2000}$ of its maximum or about 10^{-46}. If each homozygote were 98% as fit as the heterozygote, the population's reproductive potential would be cut to 10^{-9}. In either case, the value is unbelievably low. While we cannot assign

an exact maximum reproductive value to the most fit multiple heterozygous genotype, it seems quite impossible that only one billionth of the reproductive capacity of a Drosophila population is being realized. No Drosophila female could conceivably lay two billion eggs in her lifetime.

There is a strong possibility that the intensity of heterosis decreases as the number of loci heterozygous increases (VANN 1966). This does not really solve the problem, however, since drift will fix loci until the heterosis per locus still segregating is high enough to resist random fixation.

We then have a dilemma. If we postulate weak selective forces, we cannot explain the observed variation in natural populations unless we invoke much larger mutation and migration rates than are now considered reasonable. If we postulate strong selection, we must assume an intolerable load of differential selection in the population.

Some most interesting numerical calculations have been made by KIMURA and CROW (1964) relating the mutation rate, population size, heterozygosity, and genetic load of isoallelic systems. Their conclusions on the theoretical implications of widespread heterosis are similar to ours. One possible resolution of this dilemma is to suppose that in any given environment, only a portion, say 10% or less, of the polymorphisms are actually under selection so that most polymorphisms are relics of previous selection. If this is coupled with a small amount of migration between populations sufficient to retard genetic drift between periods of selection, we might explain very large amounts of variation without intolerable genetic loads. Such a process needs to be explored theoretically, while tests for heterosis need to be made under controlled conditions in the laboratory for a variety of loci and environments. Such tests are now under way. One such test by MACINTYRE and WRIGHT (1966) on esterase alleles in *D. melanogaster* was ambiguous in its result, but pointed in the direction of selective neutrality for the alleles tested.

Second, if we are to postulate heterosis on such a wide scale, we must be able to explain the adaptive superiority of heterozygotes for so many different functions. Heterozygotes differ from homozygotes in an important respect: they have present in the same organism both forms of the protein, and, in some cases they also have a third form, the hybrid protein. Only some of our enzyme proteins and none of our larval proteins show hybrid enzyme formation, so that hybrid enzyme *per se* cannot lie at the basis of general heterosis. But variation in physicochemical characteristics of the same functional protein might very well enhance the flexibility of an organism living in a variable environment. One of the best evidences that such heteromorphy of protein structure is adaptive in evolution is the occurrence of polymeric proteins made up of very similar but not identical subunits. Obviously the genes responsible for the α and β subunits of hemoglobin or the subunits of lactic dehydrogenase tetramers must have arisen by a process of gene duplication since the polypeptides they produce are so similar in amino acid sequence. The advantage of duplicate genes with slight differentiation over a single gene with different alleles is that in the former case every individual in the population can have the advantage of polymorphism. Gene duplication provides the opportunity for fixed "heterozygosity" at the functional level while allelic variation always suffers from segregation of less fit homozygotes. Heterozygosis, then, is a suboptimal solution to the problem that duplicate genes solve optimally. An excellent presentation of this argument may be found in the last chapter of FINCHAM (1966).

We are greatly indebted to DR. SUMIKO NARISE and MR. ALAN NOVETSKY for their contribution to the survey of the strains. DR. CHRISTOPHER WILLS and MR. ALAN WICK have provided us most generously with flies from Strawberry Canyon. A number of illuminating comments and criticisms of the ideas were provided by BRUCE WALLACE, ROSS MACINTYRE, JAMES CROW, and HERMAN LEWIS, to all of whom we are most grateful.

SUMMARY

Using genetic differences in electrophoretic mobility, demonstrated by HUBBY and LEWONTIN (1966) to be single Mendelian alternatives, we have surveyed the allelic variation in samples from five natural populations of *D. pseudoobscura*. Out of 18 loci randomly chosen, seven are shown to be clearly polymorphic in more than one population and two loci were found to have a rare local variant segregating. Thus, 39% of loci in the genome are polymorphic over the whole species. The average population is polymorphic for 30% of all loci. The estimates of gene frequency at these loci enable us to estimate the proportion of all loci in an individual's genome that will be in heterozygous state. This value is between 8% and 15% for different populations, with an average of 12%. A suggestion of a relationship has been observed between the extent of this heterogeneity and the amount of inversion polymorphism in a population.—An examination of the various biases in the experiment shows that they all conspire to make our estimate of polymorphism and heterozygosity lower than the true value. There is no simple explanation for the maintenance of such large amounts of genic heterozygosity.

LITERATURE CITED

DOBZHANSKY, TH., 1948 Genetics of natural populations. XVI. Altitudinal and seasonal changes produced by natural selection in certain populations of *D. pseudoobscura* and *D. persimilis.* Genetics **33**: 158–176.

DOBZHANSKY, TH., and C. EPLING, 1944 Contributions to the genetics, taxonomy and ecology of *Drosophila pseudoobscura* and its relatives. Carnegie Inst. Wash. Publ. **554**:

DOBZHANSKY, TH., A. S. HUNTER, O. PAVLOSKY,, B. SPASSKY, and BRUCE WALLACE, 1963 Genetics of natural populations. XXXI. Genetics of an isolated marginal population of *Drosophila pseudoobscura.* Genetics **48**: 91–103.

DOBZHANSKY, TH., C. KRIMBAS, and M. G. KRIMBAS, 1960 Genetics of natural populations. XXX. Is the genetic load in *Drosophila pseudoobscura* a mutational or balanced load? Genetics **45**: 741–753.

DOBZHANSKY, TH., and S. WRIGHT, 1941 Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of *Drosophila pseudoobscura.* Genetics **26**: 23–51. —— 1943 Genetics of natural populations. X. Dispersion rates in *Drosophila pseudoobscura.* Genetics **28**: 304–340.

FALK, R., 1961 Are induced mutations in Drosophila overdominant? II. Experimental results. Genetics **46**: 737–757.

FINCHAM, J. R. S., 1966 *Genetic Complementation.* Benjamin, New York.

GREENBERG, R., and J. F. CROW, 1960 A comparison of the effect of lethal and detrimental chromosomes from Drosophila populations. Genetics **45**: 1153–1168.

HARRIS, H., 1966 Enzyme polymorphisms in man. Proc. Roy. Soc. Lond. B **164**: 298–310.

HENNING, V., and C. YANOFSKY, 1963 An electrophoretic study of mutationally altered A proteins of the tryptophan synthetase of *Escherichia coli.* J. Mol. Biol. **6**: 16–21.

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.

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

MACINTYRE, ROSS, and T. R. F. WRIGHT, 1966 Responses of esterase 6 alleles of *Drosophila melanogaster* and *D. simulans* to selection in experimental populations. Genetics **53**: 371–387.

MULLER, H. J., and R. FALK, 1961 Are induced mutations in Drosophila overdominant? I. Experimental design. Genetics **46**: 727–735.

VANN, E., 1966 The fate of X-ray induced chromosomal rearrangements introduced into laboratory populations of *D. melanogaster.* Am. Naturalist (in press)

WALLACE, B., 1956 Studies on irradiated populations of *D. melanogaster.* J. Genet. **54**: 280–293. —— 1958 The average effect of radiation induced mutations on viability in *D. melanogaster.* Evolution **12**: 532–552. —— 1963 Further data on the overdominance of induced mutations. Genetics **48**: 633–651.

WALLACE, B., and TH. DOBZHANSKY, 1962 Experimental proof of balanced genetic loads in Drosophila. Genetics **47**: 1027–1042.

WRIGHT, S., TH. DOBZHANSKY, and W. HOVANITZ, 1942 Genetics of natural populations. VII. The allelism of lethals in the third chromosome of *Drosophila pseudoobscura.* Genetics **27**: 363–394.