CHAPTER 3 / GENIC VARIATION IN NATURAL POPULATIONS

A METHODOLOGICAL PROGRAM

By characterizing the struggle to measure variation as a struggle between two opposing views, the classical and balance theories of population genetics, I have in part obscured the underlying problem of making a functional whole out of theory and observation. Suppose that the experiments of Wallace and of Mukai had proved beyond any reasonable doubt that new mutations were, on the average, heterotic, or that we found the evidence from the great variety of selection experiments absolutely compelling. Then the balance hypothesis would be admitted and we would know that there was a tremendous amount of genetic heterozygosity at most loci. Yet even then we would be no closer to an empirically and dynamically sufficient genetic description of populations.

Knowing that populations in general must be highly heterozygous will not tell us how fast or how far natural selection can go in changing a particular character, nor will it allow us to make any inferences about the past history of populations and geographical races, or about the genetic processes in speciation. The answers to these latter questions, which comprise the task that evolutionary geneticists have set themselves, require the estimation of the quantities that appear in population genetic theory; they require the characterization of *gene frequencies*. The whole of population genetic theory remains an abstract exercise unless the frequencies of alternative

alleles at various loci can be determined in different populations and at different times in the history of a given population. We cannot escape from the fact that the theory of evolutionary genetics is a theory of the historical changes in the frequencies of genotypes. The observations of evolutionary genetics must, then, also be observations on the frequencies of genotypes and their variations in time and space.

I have already discussed at length (p. 21) the deep structural difficulties that face us in any attempt to estimate allelic frequencies for evolutionarily significant variation. For phenotypes of evolutionary interest, like size, shape, metabolic rates, and probabilities of survival and reproduction, the average effects of gene substitutions are small compared with the variation from environmental fluctuation. The counting of genotypes in a population, however, requires that the differences in phenotype produced by allelic substitutions be large enough to allow an unambiguous classification of individuals into genetic classes. Any technique that is to enumerate genotypes in populations must satisfy the following program:

1. Phenotypic differences caused by the substitution of one allele for another at a single locus must be detectable as an unambiguous difference between *individuals*.

2. Allelic substitutions at one locus must be distinguishable in their effects from allelic substitution at other loci.

3. All, or a very large fraction of, allelic substitutions at a locus must be detectable and distinguishable from each other, irrespective of the intensity or range of their physiological effects.

4. The loci that are amenable to attack must be a random sample of genes with respect to physiological effects and with respect to the amount of genetic variation that exists at the locus.

Requirements 1 and 2 really amount to requiring that there be a one-to-one correspondence between phenotype and genotype so that ordinary Mendelian genetic analysis can be carried out on the phenotypes and so that allelic frequencies can be estimated in populations by counting individuals. It is important that gene substitutions be detectable even in heterozygous condition, that there be imperfect dominance. Although Mendelian analysis obviously can be carried out on alleles showing complete dominance (after all, Mendel did it), dominance makes population sampling much more difficult and introduces a bias. If, for example, a dominant allele were in

frequency 0.8 in a population and ten different recessive alleles were in frequency 0.02 each, the dominant phenotype would make up 96 percent of the population and no one of the recessive homozygotes or combinations would be as frequent as 0.1 percent. In any reasonable sample the variation at the locus would be greatly underestimated, and in a sample of 100 individuals, for example, there would be a 30 percent chance that all would have the dominant phenotype.

Requirement 3 is our demand that the detectable variation be the "stuff of evolution," the genetic basis of the subtle changes in development and physiology that make up the bulk of evolutionary change. This demand is, at first sight, in contradiction to requirements 1 and 2 and lies at the root of our methodological problem.

Requirement 4 arises because we can never hope to characterize the tens or hundreds of thousands of genes in the genome of a higher organism, so we must be able to sample a relatively small number of loci as being representative of the kind and amount of genetic variation in the genome as a whole. At first sight, requirement 4 is paradoxical, for it demands that we sample loci irrespective of their variation in a population. An unbiased estimate of the amount of heterozygosity in a population must include a proportionate sample of those loci that are invariant. But the science of genetics is built upon *differences*. If all organisms were identical and showed no heritable variation, there would be no science of genetics; indeed, there would be no problem of genetics. While requirements 1 and 2 are in conflict with requirement 3, requirement 4 is in contradiction to Mendelian genetics as a whole. It is little wonder that the struggle to measure variation has been such an unhappy one.

The sorts of variation reviewed in chapter 2 fail in one way of another to satisfy the demands of a correct method. Visible mutations and lethals are in reasonable accord with requirements with 1 and 2 although the problems of dominance and the necessity for allelism tests of lethals, and even of visibles, make gene counting tedious. However, lethals and visibles are neither a random sample of allelic substitution nor a random sample of loci since they are of such drastic effect. The same objection applies to the classic visible polymorphisms such as banding in snails (Lamotte, 1951), pattern polymorphism in Lepidoptera (Ford, 1953) and ladybirds (Timofeeff-Ressovsky, 1940), or to strongly selected biochemical poly-

morphisms such as sickle-cell anemia or thalassemia in man (Allison, 1955). The issue of whether it is possible and reasonable to extrapolate from visibles and lethals to substitutions of small effect lies at the basis of the disagreement between the classical and balance schools. On the other hand, the distribution of chromosomal viabilities and fertilities certainly includes a broad range of both drastic and subtle effects, but the method of chromosomal replication makes it impossible to detect single allelic substitution at single loci (requirement 1) and was specially designed to cope with the fact that the genotype of single individuals could not be characterized (requirement 2).

A hint of the kind of observations that would satisfy our methodological program is given by human blood group polymorphisms. The technique for the detection of these polymorphisms, reaction against specific antisera, resolves the methodological contradiction between the demand for Mendelizing distinguishable loci and the demand for a random sample of allelic effects. The classification of individuals according to their blood group genotypes is unambiguous, except for certain problems of dominance, and there is no confusion between loci. The physiological and developmental functions of human blood groups are unknown, and the method for detecting them does not depend on their physiological effect but only on the molecular configuration of glycolipids bound to the red-cell membranes. These differences are presumably the direct result of small differences in the glycosyl transfer enzymes coded by the blood group genes, so that primary gene effects are being detected, irrespective of the subsequent biological effects of the blood group substance. Although the immune reaction undoubtedly detects even small changes in molecular configuration, we are not certain that it detects all of them; at any particular stage of the development of technique, some genetic variations go unnoticed. Thus, the difference between the A_1 and A_2 subgroups was not worked out until 30 years after the discovery of the ABO blood groups.

Unfortunately, the antigen-antibody method of detecting blood groups guarantees that criterion 4 of the methodological program will not be fulfilled, because the detection of a locus for a blood group substance depends absolutely on the existence of a difference. A total of 33 genes have been detected that code for human blood groups, because at least one variant individual has been discovered

for each gene. But how many genes are there for which no variants have been discovered? Since the detection of a blood group depends upon the existence of antigenic variation within it, we cannot know. Once again, the methodological program is too demanding.

THE METHOD

The solution to our dilemma lies in the development of molecular genetics. It appears certain that the sequence of nucleotides that make up a structural gene is translated with a high degree of accuracy into a sequence of amino acids making up a polypeptide chain.* Therefore, putting aside redundant nucleotide substitutions, any change in the base sequence of a DNA molecule will be reflected in a substitution, deletion, or addition of an amino acid in the polypeptide chain coded by the gene in which the alteration has occurred. Since the polypeptide chains from a single gene or occasionally from two genes make up a species of enzyme or structural protein, any change in the amino acid sequence of an enzyme or structural protein can be directly ascribed to an allelic substitution at the locus coding for the polypeptide in which the change has occurred. Moreover, if an individual is a heterozygote, both forms of the enzyme or protein will appear since both forms of the gene will be transcribed and translated into protein. The exceptions to this rule of no dominance are the chain-terminating mutations or other DNA changes that interrupt or suppress transcription or translation (which represent only 3 percent of all single-step base substitutions) and the rare severe deletions.

The amino acid sequence of proteins is a phenotype that satisfies all the requirements of the methodological program. A single allelic substitution is detectable unambiguously since it results in a discrete change in phenotype—the substitution, deletion, or addition of an amino acid. Except for redundant code changes, which are irrelevant for our problem, every substitution is detectably different, and the gene effects of different loci cannot be confused with each other

*But not necessarily with *perfect* accuracy. Thus, suppressor genes are known to introduce a high noise level into the translation of all other genes, by coding for unusual transfer RNAs.

since they code different proteins. The physiological or morphogenetic effect of allelic substitution, on the other hand, may be arbitrarily small and environment may interact in an arbitrary way with the gene products to determine the total phenotype of the organism, yet this interaction will be irrelevant to the detection of the genetic difference. The conflict between the discrete phenotypic effects demanded by Mendelism and the subtle phenotypic differences relevant to evolution is resolved by looking directly at the gene products and not at their physiological and morphogenetic effects.

Finally, the apparent paradox of trying to detect invariant genes is also resolved. Most enzymes and proteins are the products of single genes, although some, like hemoglobin and lactate dehydrogenase in vertebrates, may be composed of polypeptide chains from two different genes, joined by covalent or weaker bonds. If a particular protein should be invariant in its structure in a population, this can be equated to an invariant gene as a first approximation, although it might, in fact, indicate two invariant genes. Molecular genetics, by equating one gene to one polypeptide and thus, in most cases, to one protein, makes it possible to detect genes without variation. Genetics, which began by considering the hereditary transmission of differences between organisms, has reached the stage where it no longer depends for its inferences upon those differences. The conditions for progress in genetics in the past have been eliminated by the very progress that those conditions made possible.

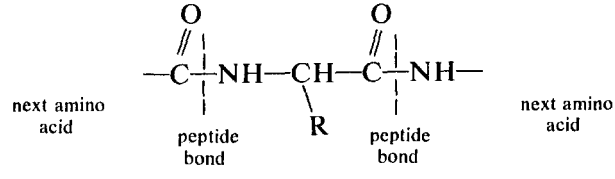
How can we turn this theoretical knowledge about the amino acid sequence of proteins into a practical program for measuring variation? At the present time it is not feasible to use the primary amino acid sequence of proteins directly as a phenotype. To determine the phenotype of an individual with respect to a single locus would require a high stage of purification of the protein in question, followed by a sequence determination, or at least a peptide "fingerprint" study. Current technology is simply not far enough advanced to allow such a procedure to be carried out on hundreds of individuals for scores of proteins, with a reasonable expenditure of labor. We are forced to consider some characterization of proteins that is sensitive to single amino acid substitutions but allows reasonably rapid examination of large numbers of individuals and many proteins.

One possibility is immunological techniques, by which many impure samples can be tested against a standard antiserum induced

against a highly purified and homogeneous protein. There are two difficulties with this procedure. First, although it might be possible to detect differences from the standard, it is much more difficult to characterize differences among variants, and impossible to tell heterozygotes from homozygotes. Second, antigen-antibody reactions against single amino acid substitutions differ quantitatively rather than qualitatively, so the clear-cut and simple amino acid sequence difference is converted into a continuously varying character. For example, Salthe (1969) attempted to use micro-complement fixation to characterize genetic variation for lactate dehydrogenase in frogs, and although he certainly demonstrated heterogeneity within and between populations, the resolution of the technique was inadequate for classifying genotypes.

The alternative to this biological detection scheme is to use the physicochemical properties of the proteins, and it is to this alternative that geneticists have turned.

When condensed into a polypeptide chain, amino acids are in the general form



where R is specific to the amino acid. The 20 common amino acids in proteins fall into three groups with respect to this R side chain. Sixteen of them, listed at the left in table 18, have non-ionizable (nonpolar) R chains. They are electrostatically neutral. Two,

TABLE 18
Classification of amino acids by side groups, with standard abbreviations

Neutral	*Neutral*	*Basic (positive)*
Alanine (Ala)	Methionine (Met)	Arginine (Arg)
Asparagine (Asn)	Phenylalanine (Phe)	Lysine (Lys)
Cysteine (Cys)	Proline (Pro)	
Glutamine (Gln)	Serine (Ser)	
Glycine (Gly)	Threonine (Thr)	*Acidic (Negative)*
Histidine (His)	Tryptophan (Try)	Glutamic acid (Glu)
Isoleucine (Ile)	Tyrosine (Tyr)	Aspartic acid (Asp)
Leucine (Leu)	Valine (Val)	

arginine and lysine, have an ammonia group that is in a dynamic equilibrium between a neutral and a positively charged form, depending on the concentration of hydrogen ions in its immediate milieu

$$H^+ + -C-NH_2 \rightleftharpoons -C-NH_3^+$$

The third group, consisting of aspartic and glutamic acids, has a carboxylic acid R group and so is in a dynamic equilibrium between neutral and negatively charged forms

$$H^+ + -C-\overset{\overset{O}{\parallel}}{C}-O^- \rightleftharpoons -C-\overset{\overset{O}{\parallel}}{C}-OH$$

A polypeptide made up of a mixture of the three types of amino acid will have a net negative or positive charge, depending on the balance of charges and the folding of the molecule. As the pH is lowered (hydrogen ion concentration is increased), more and more of the NH_2 groups will become positively charged NH_3^+ ions, while the acidic COO^- ions will be saturated and become neutral. The result is that the polypeptide as a whole will take on a positive charge. The reverse will happen as the pH is raised (decreasing the concentration of hydrogen ions). The point at which the negative and positive charges just balance out to give a neutral polypeptide is the iso-electric point, which for most proteins in animals is slightly alkaline, around pH 8.

If an allelic change at a locus results in the replacement of an amino acid in one group in table 18 with an amino acid from another, the iso-electric point of the protein will be altered, as will the net charge on the protein at any given pH. For example, a single-step change in the DNA codon AAC to the codon AAA results in the substitution of the positively charged lysine for the neutral asparagine. An even more drastic single-step change is from AAG to GAG, with resulting substitution of a negatively charged glutamic acid for the positively charged lysine. Such changes in net charge can be used to separate proteins and thus to identify the products of allelic forms of the same gene. The technique by which this separation is achieved is *gel electrophoresis.*

Figure 7 is a diagram of a typical gel electrophoresis apparatus during the course of an experiment. It consists essentially of a slab

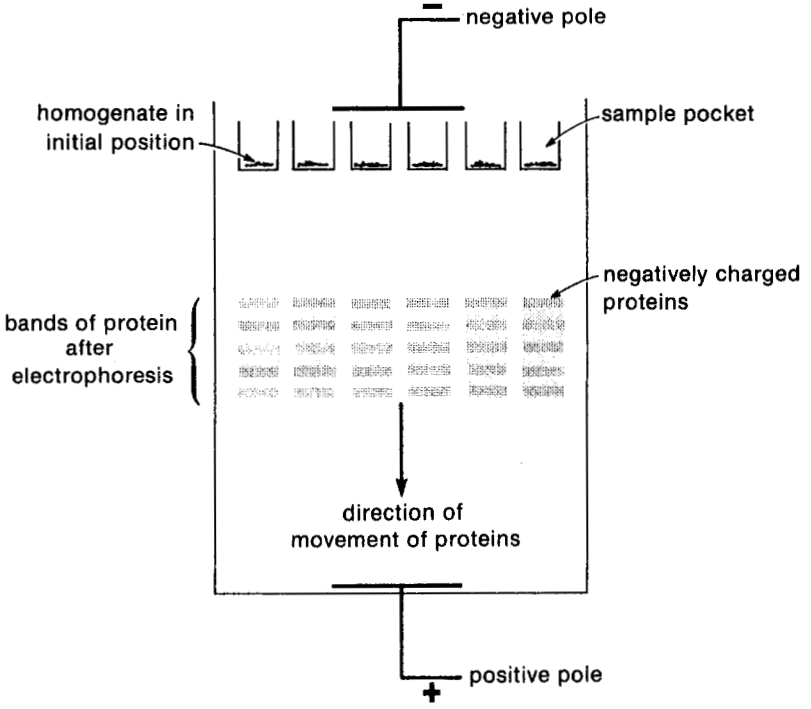


FIGURE 7

Diagram of a vertical slab gel-electrophoresis aparatus.

of some jellylike material (starch, agar, or a synthetic polymer) whose two ends are in contact with the opposite poles of an electric potential. Material for electrophoresis is introduced into wells at one end of the gel and any charged molecules will move down along the gel under the force of the electric field applied. The gel is enclosed in a cooling jacket to prevent its overheating and the consequent denaturation of the proteins moving through it. In practice, a bit of tissue, or, in the case of Drosophila, a whole individual, is ground up to break down the cells, the solid material is centrifuged or otherwise removed, and the liquid phase containing a mixture of the soluble proteins of the organism is placed in one of the slots in the gel. The pH of the gel and the grinding solution is adjusted to be on the alkaline side of the iso-electric point of the proteins being studied, so that the proteins will be negatively charged and migrate toward the positive pole of the apparatus. The speed of migration of any particular protein will depend upon its molecular size and the

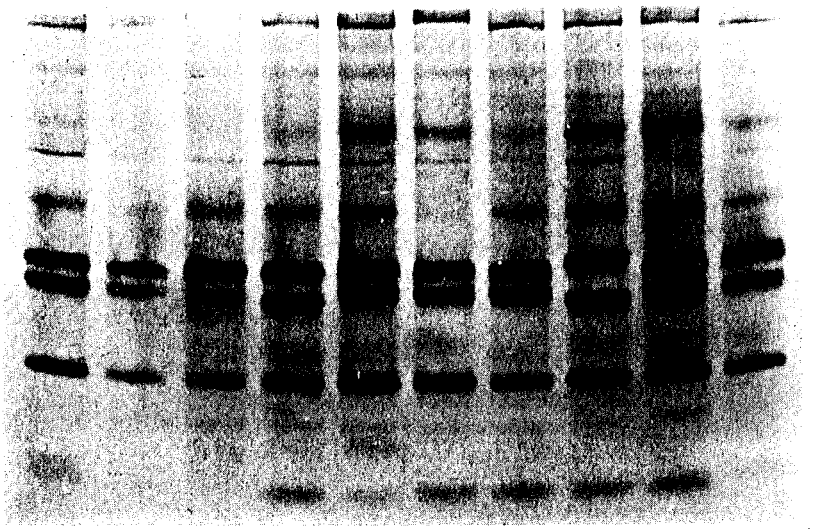


FIGURE 8

Larval hemolymph proteins of *D. pseudoobscura* after electrophoresis, stained with coomassie blue. Each vertical column contains the proteins of a single larva, each band being a different protein.

net charge it carries. After a high voltage has been applied to the gel for one to two hours, the various proteins that migrated at different rates will be concentrated at different points along the gel, as indicated in the diagram, although of course they are invisible. The problem then is to visualize them.

Proteins in high concentration can be seen simply by staining the gel with a dye that is a general protein stain. Figure 8 shows such a gel made from *Drosophila pseudoobscura* larvae. There are a large number of bands, and the high repeatability of the method is shown by the identity of the pattern of bands from larva to larva.

Enzymes, however, generally exist in such low concentrations that there is insufficient material in a single fly or tissue sample to be visualized with a general protein stain. Each enzyme can be found on the gel by immersing the slab in a bath containing a substrate for that enzyme and a dye that will be bound or colorized at the place where the substrate is split. For example, the oxidized form of a dye may be colorless, but may become colored if it is reduced by electrons, which are passed to the dye when a dehydrogenase enzyme

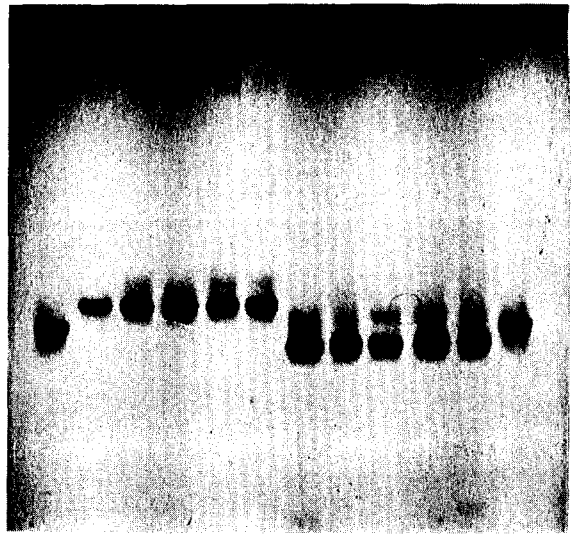


FIGURE 9

Adult esterases from *D. pseudoobscura*. *Sample 1*, standard strain; *samples 2–6*, five individuals from a second strain; *samples 7–11*, five individuals from a third strain; *sample 12*, standard strain. From Hubby and Lewontin (1966). Reprinted with permission of *Genetics*.

and its cofactor split. Although many reactions are more complicated and require various intermediates, the principle is the same for all. The result is a colored band in the gel at the place where the particular enzyme is concentrated. Several different enzymes can be visualized on the same gel, provided the conditions for the reactions are not mutually exclusive or inhibitory. Figure 9 depicts a gel of *D. pseudoobscura* adults stained for an esterase enzyme. Each sample is an individual fly. Samples 1 and 12 are a standard strain, while samples 2 to 6 and 7 to 11 are from two different homozygous strains. Again we see the high repeatability within strains and a clear-cut difference between strains, which turned out, on later analysis, to be the result of allelic differences at a single locus. Thus the electrophoretic mobility of an enzymatic protein is a repeatable phenotype that gives discrete, unambiguous differences between genotypes.

The method is capable of distinguishing different homozygotes not only from each other but also from heterozygotes. Figure 10 is the result of comparing individuals of a strain segregating at the

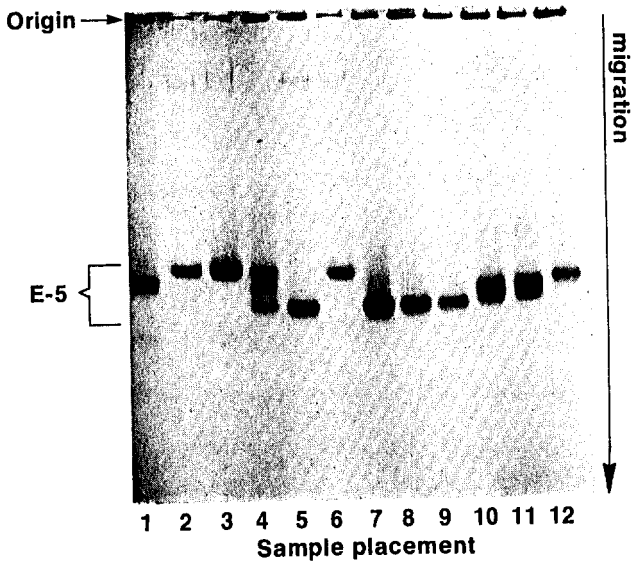


FIGURE 10

Adult esterases from *D. pseudoobscura*. Sample 1, standard strain *est-5*$^{1.00}$/*est-5*$^{1.00}$; samples 2 and 3, *est-5*$^{0.95}$/*est-5*$^{0.95}$; sample 4, *est-5*$^{0.95}$/*est-5*$^{1.12}$; sample 5, *est-5*$^{1.12}$/*est-5*$^{1.12}$; sample 6, *est-5*$^{0.95}$/*est-5*$^{0.95}$. From Hubby and Lewontin (1966). Reprinted with permission of *Genetics*.

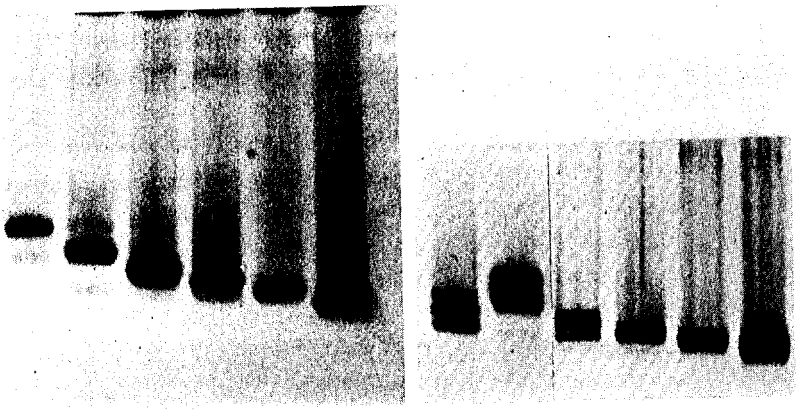


FIGURE 11

Allozyme phenotypes. Left, homozygotes for six different alleles at the *esterase-5* locus in *Drosophila pseudoobscura*; right, several different heterozygotes between alleles. From Hubby and Lewontin (1966).

esterase-5 locus in *D. pseudoobscura*. Sample 1 is again a homozygote standard strain, and the allele it carries has been designated *est-5*$^{1.00}$, the superscript standing for the relative electrophoretic mobility of the protein specified by that allele. Samples 2, 3, and 6 have a more slowly moving protein, while sample 5 moves faster than the standard. Sample 4 shows three bands, two identical with the fast- and slow-moving forms and one intermediate between them. Further genetic analysis of individuals from the strain from which this last sample comes shows that it is segregating for two alleles, *est-5*$^{0.95}$ and *est-5*$^{1.12}$, and that sample 4 is the heterozygote, *est-5*$^{0.95}$/*est-5*$^{1.12}$. The presence of the intermediate band in the hybrid indicates that esterase-5 is a dimeric enzyme, so that in heterozygotes three sorts of dimers are produced: homo dimer *0.95—0.95*, with the same mobility as the dimer made by the *est-5*$^{0.95}$/*est-5*$^{0.95}$ homozygote; homo dimer *1.12–1.12*, with the same mobility as the dimer made by the *est-5*$^{1.12}$/*est-5*$^{1.12}$ homozygote; and a hetero dimer *0.95–1.12*, with a mobility halfway between. Even if the enzyme were monomeric so that "hybrid" molecules were not produced, a heterozygote would still produce two different forms of the enzyme, each corresponding to one of the homozygotes. There is no dominance at the level of production of the polypeptides, and each allele in the heterozygote functions to make its product. The different enzyme forms produced by different alleles at the same locus have been called *allozymes* by Prakash, Lewontin, and Hubby (1969) to distinguish them from the more general phenomenon of *isozymes* that are the different molecular forms of an enzyme arising from any cause (Markert and Møller, 1959). Figure 11a shows the allozymes of homozygotes from six different alleles recovered from a population of *D. pseudoobscura*, and figure 11b gives some of the heterozygous combinations between them. When the mobility of two allozymes is very close, the three bands in the heterozygote cannot be resolved but appear as a single thick band. Nevertheless, heterozygotes are distinguishable from homozygotes.

THE FIRST APPLICATIONS OF THE METHOD

The fourth point in our methodological program requires that we score loci at random with respect to the amount of variation present in populations. All that is required in order to apply gel elec-

trophoresis of proteins and enzymes to the problem of the amount of heterozygosity in natural populations is a large repertoire of dye-coupled enzyme-substrate reactions that have been accumulated for reasons other than the study of polymorphisms. By about 1964, an appropriately large repertoire existed, the methods having been developed to study the ontogeny of enzymes, their tissue localization, their sensitivity to inhibitors, their molecular weight, their purification, but not, in general, their genetic variation. Yet, in the course of accumulating these methods, a large amount of genetic variation turned up incidentally. A survey of the literature by Shaw (1965) recovered 16 different enzymes in some 20 species of organisms, including flagellates, insects, amphibia, birds, and mammals, that had revealed some variation in electrophoretic mobility. Of 58 cases, 35 were reported as "polymorphic," but the meaning of this term was vague in most cases since genetic studies had not been done, or population samples had not been taken. Nevertheless, the stage had clearly been reached at which an appropriate tool, gel electrophoresis, existed to fulfill the exacting methodological program required for an answer to the central problem of evolutionary genetics.

Two groups of workers simultaneously, and completely independently, set about estimating the heterozygosity and allelic distributions in natural populations by carrying out large-scale surveys of electrophoretic variation in randomly chosen enzymes and proteins. One group, in London, studied ten randomly chosen human enzymes in the English population (Harris, 1966). They found three to be polymorphic. Red-cell acid phosphatase had three alleles with frequencies 0.36, 0.60, and 0.04; phosphoglucomutase had two alleles with frequencies 0.74 and 0.26; and adenylate kinase (not yet completely analyzed in 1966) had two alleles at frequencies 0.05 and 0.95. The proportion of individuals heterozygous for these three loci was then 0.509, 0.385, and 0.095, respectively; averaging these with the seven monomorphic loci gives an estimate of heterozygosity in the English population of 0.099.

The second group, in Chicago, made a more extensive study in *Drosophila pseudoobscura*, including eight enzymes and ten larval hemolymph proteins (Hubby and Lewontin, 1966; Lewontin and Hubby, 1966). Their initial survey was not made on natural populations directly, but on 43 strains derived from single fertilized females

taken from 5 natural populations. Of these strains, 33 had been in the laboratory for approximately five years (75 generations) and the remaining 10 were the F_2 and F_3 generations from newly trapped wild females.

Table 19 shows the results of this study in detail. The most striking feature that is immediately apparent is the great genetic variation present in these strains. Of the 18 different gene loci represented, 7 have electrophoretic variants and, of these, 5 loci have three or more alleles represented. This variation is all the more striking because of the small number of strains examined and the fact that, with the exception of *alkaline phosphatase-4*, all the variable loci have more than one variant strain, from more than one population. Clearly, the variation is widespread in natural populations. An interesting feature of the variation is the persistence of some segregation in strains that have been maintained in the laboratory for 75 generations with repeated episodes of breeding from very small numbers of parents. Among the variable loci, the sample from Strawberry Canyon, California, fresh from the wild, is segregating in 43 percent of the strains although among the strains with long laboratory histories only 9 percent are still segregating. Yet this persistent segregation is not at random over loci, most of it being at the *pt-7* and *pt-8* loci.

An extremely important aspect of the observations in table 19 is that, on genetic analysis, the variation between strains turned out in every case to be genetic and unifactorial. That is, *for every variable protein and enzyme, the variation was the result of the segregation of alleles at single loci. This fact is the cornerstone of the method.* It allows us to calculate allelic frequencies at each locus and therefore heterozygosities in the population, but most important, it allows us to equate each of the invariant proteins, as a close approximation, to an invariant locus. Unless we can make such an equation, the entire procedure is useless for estimating the proportion of loci that is polymorphic and the amount of heterozygosity per locus that exists in a population, since the number of monomorphic loci would be unknown. Genetic analysis is therefore essential, especially in view of the fact that minor variations in experimental conditions can in some cases cause enough variation in electrophoretic mobility to obscure or mimic genetic variation (G.B. Johnson, 1971).

From the data of table 19, allelic frequencies in each population

TABLE 19

Number of strains from each population of *Drosophilla pseudoobscura* that are either homozygous or segregating for various alleles at different loci

Locus	*Allele*	*Strawberry Canyon*[a]	*Wild Rose*[a]	*Cimarron*[b]	*Mather*[a]	*Flagstaff*[c]	*Bogotá*[d]
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*	—

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*[f] *alleles*	*3*[g] *alleles*	*2 alleles*	*2*[h] *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*
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*

Note: From Lewontin and Hubby (1966).

[a]California. [b]Colorado. [c]Arizona. [d]Colombia. [e]Both loci segregating in the same strain. [f]Three strains segregating. [g]One strain segregating. [h]Two strains segregating.

can be estimated by assuming that the strains were roughly representative of the populations from which they were sampled. The frequencies are very rough estimates indeed, since so few strains were taken from each locality and since, except for Strawberry Canyon, there was a long history of laboratory culture. These frequency estimates, which are not worthy of serious consideration in themselves, can, however, in turn be used to estimate the proportion of the genome that is heterozygous in each population, a figure that is a great deal more reliable than the individual allele frequencies.

Table 20 gives both the proportion of loci polymorphic and the average heterozygosity per individual. A locus was counted as polymorphic only if a variant allele appeared in more than one strain out of all 43 examined. The variation is impressive. A third of all loci are polymorphic, and the average individual is a heterozygote at one out of eight of his loci, estimates that are remarkably similar to Harris's preliminary results for man. Even at those loci for which an individual is homozygous he is likely to differ from another randomly chosen individual in the population, since about one-third of all loci are segregating.

The estimates of variation become more impressive, indeed staggering, when one realizes that a majority of amino acid substitutions do *not* involve charge changes. From table 18 it is easy to show that 136 out of the possible 380 amino acid substitutions, or 36

TABLE 20

Proportion of loci, out of 18, that are polymorphic and proportion of genome that is estimated to be heterozygous in an average individual, for 5 populations of *D. pseudoobscura*

Population	*Proportion of loci polymorphic*	*Proportion of genome heterozygous per individual*
Strawberry Canyon, Calif.	.33	.148
Wild Rose, Calif.	.28	.106
Cimarron, Colo.	.28	.099
Mather, Calif.	.33	.143
Flagstaff, Ariz.	.28	.081
Average	.30	.115

Note: From Lewontin and Hubby (1966).

percent, involve a charge change. If we consider only single-base substitutions in DNA, then of the 399 nonredundant single-step missense mutations possible, 128, or 32 percent, involve a change in charge. These values take no account, however, of the inequalities in amino acid content of proteins. Although the empirical formulas differ considerably from protein to protein, a correction for amino acid content gives values of 26–28 percent for the proportion of substitutions that would be charge changes. But these calculations are of uncertain worth because in actual practice many neutral amino acid substitutions may make sufficient conformational or resonance changes to make the substitution detectable electrophoretically. For example, all of the substitutions in cytochrome *c* over the whole range of organisms from yeast to man, involving multiple substitutions at 70 percent of all amino acid sites, have not caused any change in the arithmetical sum of charges, all cytochromes having compensatory substitutions. Yet, with sufficient diligence, it is possible to distinguish these "neutral" substitutions electrophoretically. Given that Hubby and Lewontin did not undertake extensive investigations of various buffer systems and other technical alterations that might have further resolved apparent identities, we must assume that they detected only about a third of the total sequence variation actually present. On this basis, the average heterozygosity per locus for natural populations of *D. pseudoobscura* is about 35 percent, and essentially every gene is polymorphic. At least on the face of it, the classical hypothesis of population structure is firmly and directly refuted, and the balance theory is revealed as correct.

Even if we take Lewontin and Hubby's and Harris's values directly, without trying to correct for the undetected variation, natural populations of flies and men are immensely variable genetically, members of the same population and even the same family differing from each other at thousands of loci. The first direct measurements of genetic variation in natural populations have proved, in Dobzhansky's words, that the norm "is not a transcendental constant, standing above or beyond the multiform reality" (1955). As we shall see later, they may have proved too much.

SURVEYS OF VARIATION

The first estimates of heterozygosity in natural populations were followed by much more extensive and accurate surveys for both

man and *D. pseudoobscura*, as well as by a large number of studies in other organisms. There is considerable variation in the kind of information that can be obtained from the different surveys, depending on their experimental methods and sampling techniques. One of the earliest, a preliminary report on what was eventually an extensive study of *D. ananassae*, showed 9 polymorphic loci out of 20, representing six different enzyme functions (Johnson, et al., 1966), but no gene frequencies were given.

A study of ten egg white proteins in domestic chicken flocks and breeds (Baker, 1968) showed that four loci were polymorphic in at least one breed and that 24 out of 37 breeds were segregating for at least one locus of the four. Because only a few, often only two or three, birds could be tested from most breeds, no quantitative estimate of heterozygosity can be made from the bulk of the sample. A few flocks were more extensively sampled, however, and the results are given in table 21. While allele frequencies may occasionally differ widely between two flocks of the same breed or between two breeds that are closely related, the average heterozygosity per individual is a characteristic of breeds rather than individual flocks. It is not at all clear what is meant by a "popula-

TABLE 21

Frequency of one of the two alleles segregating at each of the four polymorphic egg-white protein loci among ten loci tested in several flocks of domestic chickens

Flock		*n*	*Ovalbumin*^{S}	G_1^A	G_2^D	*Tf*^{S}	*Heterozygosity per individual*
Australorp	A	19	1.00	0.63	1.00	1.00	0.047
	B	18	1.00	0.17	0.94	1.00	0.040
Lt. Sussex	A	51	1.00	0.49	0.61	0.71	0.139
	B[a]	24	1.00	0.52	0.79	0.98	0.087
Maran[b]	A[a]	24	0.81	0.83	0.65	1.00	0.105
	B	30	0.77	0.87	0.42	0.95	0.116
N. Holland Blue[a]		25	0.86	0.24	0.62	1.00	0.108
Barnevelder[a]		22	1.00	0.86	0.52	1.00	0.074
Welsummer[a]		15	1.00	0.63	0.93	1.00	0.060
Average							0.086

Note: From Baker (1968).

[a]Flocks kept by the same breeder.

[b]Bracketed breeds were derived from a common ancestral type in recent times.

tion" of domesticated chickens, and so it is difficult to say what universe is being sampled.

One difficulty with the interpretation of surveys is illustrated by the very large studies of Ruddle et al. (1969) and Roderick et al. (1972) on inbred and feral mouse populations. These workers studied 16 enzyme loci in 35 inbred strains collected from a variety of laboratories, and 17 loci in 6 collections of mice from the wild. Among the inbred lines, not a single locus was constant over all strains, 4 loci showed a single variant strain, and the remaining 12 loci varied among many strains. But the loci studied were chosen because they had already been shown in published reports to have some genetic variation. Thus it is impossible to assess the relevance of this variation for our problem. In the samples from natural populations, 5 out of 17 loci were variable, and the average heterozygosity per population varied from 7.6 to 14.7 percent. But again the loci were previously known to have some genetic variation in laboratory strains.

No survey of protein variation is completely free of difficulties and biases, but reasonably reliable estimates of genetic variation in natural populations require:

1. Adequate sample sizes, large enough to provide about 50 wild genomes for each locus.

2. Genomes tested either directly from natural populations or only a few generations removed from the original samples.

3. A *large* sample of loci, more important even than a large sample of individuals, because of the very considerable differences in heterozygosities at different loci. For example, Harris's original estimate of 0.099 for heterozygosity in man, from the first ten loci looked at, turned out to be 50 percent too large (Harris and Hopkinson 1972).

4. A *diverse* sample of loci, not weighted heavily by one or two enzymatic functions. It may be very misleading, for example, if in a sample of 14 loci, 6 code for enzymes that are nonspecific esterases, attacking the same range of nonbiological substrates (Ayala et al., 1970) or if 10 out of 18 genes code for high concentration globular proteins of unknown function (Lewontin and Hubby, 1966). Of course, it may not matter, but with our present rudimentary understanding of the distribution of variation over different kinds of loci, a survey that is spread over the widest possible range of enzymatic functions is preferable.

5. An *unbiased* sample of loci, not chosen consciously or unconsciously because of known variability. In the very first surveys this was not a problem, but as time goes on and the literature comes to be filled with reports on studies of enzyme polymorphism, it becomes progressively more difficult to choose a sample of loci in a new organism or population that is *not* tainted by the workers' knowledge about what has gone before. Geneticists like variation and find genetic uniformity rather dull. The excitement of seeing a new genetic segregation in a new organism is real and seductive. It is surely no accident that, in virtually every organism examined in the last six years, even when only a few enzymes are chosen the nonspecific esterases turn up and are nearly always polymorphic. Often the bias is unconscious and only appears as trends in data. If, for example, the estimated heterozygosity goes down as more and more loci are looked at in an organism, or if many populations have been examined (but only some loci in each) and a negative correlation between the heterozygosity and the number of loci tested is found, then there is strong reason to suspect an unconscious bias toward variable loci. A. D. Hershey is reported to have described heaven as "finding an experiment that works and doing it over and over and over." Population geneticists too have found paradise.

I have summarized in table 22 the data known to me from those allozyme surveys that seem to approach most closely the requirements of adequacy and unbiasedness. None is perfect, and others, not included, are nearly as good. I have chosen arbitrarily to include only surveys of at least 18 loci. The table shows the proportion of loci that are polymorphic, defined as the proportion of loci in which the most common allele does not exceed a frequency of 0.99, and the estimated heterozygosity per locus per individual in the population. Since the proportion of "polymorphic" loci is somewhat arbitrary and has a high variance (because a relatively small number of loci are tested), the heterozygosity per individual is the more informative figure and no survey has been included from which this value cannot be estimated.

The original results of Lewontin and Hubby and of Harris are completely confirmed. In a dozen species shown in table 22, between 20 and 86 percent of loci are segregating within a population, and the heterozygosity per individual falls in a relatively narrow range over all species, between 5.6 and 18.4 percent. The median

TABLE 22
Surveys of genic heterozygosity in a number of organisms

Species	*Number of populations*	*Number of loci*	*Proportion of loci polymorphic per population*	*Hetero-zygosity per locus*	*Standard error of hetero-zygosity*	*Reference*
Homo sapiens	1	71	.28	.067	.018	Harris and Hopkinson (1972)
Mus musculus musculus	4	41	.29	.091	.023	Selander, Hunt, and Yang (1969)
M. m. brevirostris	1	40	.30	.110		Selander and Yang(1969)
M. m. domesticus	2	41	.20	.056	.022	Selander, Hunt and Yang (1969)
Peromyscus polionotus	7 (regions)	32	.23	.057	.014	Selander et al. (1971)
Drosophila pseudoobscura	10	24	.43	.128	.041	Prakash, Lewontin, and Hubby (1969); Prakash, Lewontin, and Crumpacker (1973)
D. persimilis	1	24	.25	.106	.040	Prakash (1969)
D. obscura	3 (regions)	30	.53	.108	.030	Lakovaara and Saura (1971a)
D. subobscura	6	31	.47	.076	.024	Lakovarra and Saura (1971b)
D. willistoni	2–21	28	.86	.184	.032	Ayala et al. (1972)
	10	20	.81	.175	.039	Ayala, Powell, and Dobzhansky (1971)
D. melanogaster	1	19	.42	.119	.037	Kojima, Gillespie, and Tobari (1970)
D. simulans	1	18	.61	.160	.052	Kojima, Gillespie, and Tobari (1970)
Limulus polyphemus	4	25	.25	.061	.024	Selander et al.(1970)

proportion of polymorphic loci is 30 percent and the median heterozygosity per individual is 10.6 percent, with man somewhat on the low side and *Drosophila pseudoobscura* somewhat on the high side of the median variation. Thus we will not be far off if we characterize sexually reproducing species of animals as being polymorphic for a third of their genes, and individuals within the species as being heterozygous for about 10 percent of their loci. Again we must bear in mind that, for the genes sampled, these are *minimum* estimates since they are based on only those gene substitutions that are detectable electrophoretically.

The standard errors of average heterozygosity given in table 22 are meant to emphasize the importance of examining a large number of loci. These standard errors are calculated from the variance among loci, after the heterozygosities are averaged over populations within a species, and they are between one-fifth and two-fifths as large as the mean heterozygosities. Such large standard errors arise because different loci differ markedly in their genetic variation, with the majority being monomorphic, and the polymorphic minority being very heterozygous.

In table 23 the average heterozygosities for different loci in man and *Drosophila pseudoobscura* have been listed in ascending order. Although the range of heterozygotes for *Drosophila* is greater, there is a strong suggestion of bimodality among the 20 polymorphic genes in man. Because of such heterogeneity, it is impossible to interpret differences in heterozygosity between species until large numbers of loci have been tested. For example, two earlier studies of *D. melanogaster* and *D. simulans*, one based on 10 loci (O'Brien and MacIntyre, 1969) and one on 6 loci (Berger, 1970), gave heterozygosities for *D. melanogaster* of 0.23 in both cases, but only 0.07 and 0.00 for *D. simulans*. Ingenious explanations can be made for the apparent lack of heterozygosity in *D. simulans*, based upon the differences in population structure and economy of the two species. There is even a temptation to speculate about genetic heterogeneity as an adaptive strategy that is optimal for *D. melanogaster*, whereas homozygosity is a better strategy for *D. simulans*. But all such speculations are idle in view of the larger study of Kojima, Gillespie, and Tobari (1970), cited in table 22, which completely reverses the apparent facts, showing *D. simulans* to be more heterozygous than *D. melanogaster*. Nor is the difference in het-

TABLE 23
Average heterozygosity in populations of *Drosophila pseudoobscura* and man, for different loci

D. pseudoobscura		*Man*	
Locus	*Heterozygosity*	*Locus*	*Heterozygosity*
12 monomorphic loci	.000	51 monomorphic loci	.00
		Peptidase C	.02
Acetaldehyde oxidase	.012	Peptidase D	.02
		Glutamate-oxaloacetate transaminase	.03
Protein-7	.063	Leucocyte hexokinase	.05
Protein-13	.070	6-Phosphogluconate dehydrogenase	.05
Malic dehydrogenase	.102	Alcohol dehydrogenase-2	.07
Octanol dehydrogenase	.109	Adenylate kinase	.09
		Pancreatic amylase	.09
Leucine aminopeptidase	.155	Adenosine deaminase	.11
Protein-10	.229	Galactose-1-phosphate uridyl transferase	.11
Protein-12	.234	Acetyl cholinesterase	.23
α-Amylase-1	.353	Mitochondrial malic enzyme	.30
Xanthine dehydrogenase	.492	Phosphoglucomutase-1	.36
		Peptidase A	.37
Protein-8	.513	Phosphoglucomutase-3	.38
Esterase-5	.741	Pepsinogen	.47
		Alcohol dehydrogenase-3	.48
		Glutamate-pyruvate transaminase	.50
		RBC acid phosphatase	.52
		Placental alkaline phosphatase	.53

Note: Data from Prakash, Lewontin, and Crumpacker, (1972) and Harris and Hopkinson (1972).

erozygosity between the two species significant (0.160 as against 0.119), being less than one standard error.

Surveys of polymorphism have not yet reached the stage at which comparisons of total heterozygosity between species can be meaningful; if such comparisons are to be made for comparative evolutionary studies, something on the order of 100 randomly chosen loci will be required. Only in *Homo sapiens* has anything like a sufficient number of genes been studied, and unfortunately only Europeans have been extensively surveyed.

The variety of organisms in table 22 (man, mouse, Drosophila, and horseshoe crab) gives us some confidence in the generality of the result, and a number of less extensive surveys widen the taxonomic range even more, without contradicting the general picture. *Drosophila athabasca* and *D. affinis* (Kojima, Gillespie, and Tobari, 1970), *D. ananassae* and *D. nasuta* (Johnson et al., 1966; Gillespie and Kojima, 1968; Stone et al., 1968), and *D. mimica* (Rockwood, 1969), the harvester ant, *Pogonomyrmex barbatus* (Johnson et al., 1969), the snails, *Cepaea nemoralis* and *C. hortensis* (Manwell and Baker, 1968), the salamander, *Ambystoma maculatum* (T. Uzzell, unpublished data), the chicken (Baker, 1968), the pheasant, *Phasianus calchicus* (Baker et al., 1966), and the quail, *Coturnix coturnix* (Baker and Manwell, 1967) have all been shown to be polymorphic for a number of protein-specifying genes in natural populations, although the data do not allow accurate quantitative values to be calculated (comparable to those in table 22), usually because too few loci have been studied.

Thus far there has been no extensive survey of allozyme variation in plants, although the few studies that have been made leave no doubt that plants, too, are highly polymorphic. Marshall and Allard (1970) and Clegg and Allard (1972) have found clear polymorphism for several genes in two species of wild oats, *Avena barbata* and *A. fatua*, despite the fact that these are strongly inbreeding species. A survey of a population of *Oenothera biennis* revealed 5 out of 19 enzymes to be polymorphic, but every individual in the population was a heterozygote for these five genes, presumably because *O. biennis* is a permanent, multiple-translocation heterozygote and the polymorphic genes are markers of the translocated segments (Levin, Howland, and Steiner, 1972). *Taraxacum officinale*, a completely apomictic dandelion, also shows a good deal of allozyme variation between clones growing in different habitats (Solbrig, 1971), but no genetic analysis is possible.

With their immense variety of breeding systems, plants will be extremely important for comparative studies and for sorting out the forces influencing allozyme variation. They are also far more easily manipulated for population experiments, both in nature and in artificial cultivation, than are elusive and willful animals. In proportion to their potential, plants have been greatly neglected as material for studies of genetic variation.

The very consistency of the data in table 22 and the support they

receive from the other, less extensive, studies is a little worrisome. If the polymorphism so widely observed is truly related to the evolutionary processes that have molded and will mold the history of various species, there ought to be some variation among species in the degree of their genetic variation. In particular, there ought to be some species which, for reasons of their unique population structure or recent evolutionary history, are rather less variable then most. There must be species that have recently expanded from a very small genetic base, or have undergone severe selection for certain specialized genotypes, or whose breeding structure constantly gives rise to high local inbreeding, so that a substantial fraction of the genome should be monomorphic.

There are several reports of electrophoretic surveys in which abnormally low heterozygosity has been observed. Most, unfortunately, are not extensive or intensive enough for one to be certain that the difference is not the result of a small sample of genes or a difficulty in technique. The most extreme case is the report by Serov (1972) that 138 individuals of *Rattus rattus* from a wild population in Novosibirsk were all completely homogeneous for 21 loci. The sample of both loci and individuals seems large enough to make this observation reliable, but the lack of technical details, and the startling difference from the house mouse, despite the great similarity in natural economy between the two rodents, make this report tantalizing but uncertain.

Of 7 loci studied in three Southeast Asian populations of a macaque, *Macacus fascicularis* (Weiss and Goodman, 1972), 5 showed detectable genetic variation in one or more populations, but in most cases one allele was in overwhelming preponderance. In two localities, Thailand and the Philippines, only 2 out of 7 loci were polymorphic and the heterozygosity per individual was 0.019 and 0.034, respectively. In Malaysia, however, 3 loci were polymorphic and heterozygosity was 0.077, values that are well within the usual range.

Finally, a species of Pacific salmon, *Oncorhyncus keta*, is reported to be polymorphic for 11 to 14 percent of its loci, with heterozygosity lying between 0.0195 and 0.032 (Altukhov et al., 1972). Even these limits are uncertain, however, because a large number of monomorphic hemoglobin, lens protein, and muscle protein bands are included.

Three reports of low heterozygosity are convincing. A wide geo-

graphical sample of the cricket frog, *Acris crepitans*, in the United States showed this species to be polytypic but not very heterozygous (Dessauer and Nevo, 1969). Thus 12 out of 20 enzymes and proteins were variable in the species as a whole, but most of the 27 populations were fixed or nearly fixed for one allele or another. The average proportion of polymorphic loci per population was only 12 percent and 4 loci were completely homozygous within populations yet differed between populations. Three distinct geographical races could be discerned on the basis of the preponderant electrophoretic alleles, an Appalachian race, a Gulf race, and a Midwestern race; considering the large number of populations sampled and the reasonable number of loci, these divisions are probably real. *Acris crepitans* is probably a truly polytypic species with low heterozygosity within populations.

A second case concerns species of the kangaroo rat, *Dipodomys*. In a survey of 17 loci in several forms of this desert-dwelling rodent, Johnson and Selander (1971) studied large samples of two species, *D. merriami* and *D. ordii*, in a number of populations. For *D. merriami*, the proportion of loci polymorphic varied between populations from 12 to 29 percent and the heterozygosity was between 0.020 and 0.071, with a mean of 0.051. *Dipodomys ordii* was even less genetically varying, with polymorphism between 6 and 24 percent of loci and heterozygosity between 0.004 and 0.017 per individual, the mean being 0.0081.

A fossorial rodent, the mole rat, *Spalax ehrenbergi*, was surveyed for genetic variation at 17 loci by Nevo and Shaw (1972), with similar results. Polymorphism varied from 6 to 29 percent, with an average of 19 percent, and heterozygosity was from 0.018 to 0.056 in different geographic races, with a mean of 0.037.

With the possible exception of *Acris crepitans*, the data are at present insufficient to be sure that any of these species is significantly more homozygous than those in table 22, and we must bear in mind that, based on the earlier, less extensive studies, *Drosophila simulans* would have been regarded as a monomorphic species. I must emphasize again that we cannot yet compare the levels of heterozygosity in different species and be all all confident that the observed differences are real. Species as diverse physiologically, evolutionarily, and ecologically as man, mouse, and the phylogenetic "relic," *Limulus*, the horseshoe crab, are virtually identical in

both the proportion of loci that are polymorphic and the average heterozygosity per individual.

HOW REPRESENTATIVE ARE THE GENES?

The methodological program set out at the beginning of this chapter demanded that a satisfactory method for measuring genic variation be capable of detecting most or all of the allelic substitution at a locus, and that loci be chosen "at random" with respect to physiological function and degree of variation. Clearly, the method of electrophoretic separation of proteins does not detect all, or perhaps even most, of the allelic substitution at a locus. Perhaps it detects as little as one-quarter of all substitutions. Thus our estimates of heterozygosity are downwardly biased, perhaps by as much as four times, and this makes even more startling our estimate that a third of all loci are polymorphic and 10 percent of all genes are heterozygous in each individual. Are there any possible biases in the other direction, biases that may inflate our estimate of the total variation in the genome as a whole? We must consider the requirement that genes be sampled "at random" from the genome.

The constitution of the genome, in the sense of the functional array of gene products, is still a mystery in higher organisms. It is certain that many genes code for enzymes, and one by-product of the investigation of natural populations of Drosophila has been the localization of many new genes, coding specific enzymes, on the genetic map of different species. But the sperm of *D. melanogaster* contains enough DNA for 10^8 base pairs. With three base pairs per codon and an average of 150 amino acids per polypeptide, there is enough DNA to code about 2×10^5 polypeptides. For man, with 16 times as much DNA, there is enough to code 3×10^6 polypeptides. We can hardly believe that higher organisms are capable of manufacturing between a quarter million and 3 million different enzymes! Even to suppose that 10 percent of *Drosophila*'s DNA codes for enzymes (20,000 enzymes) stretches our credulity in the light of present knowledge of biochemistry. Of the remaining 90 percent of the DNA, some may not be informational at all, and we need not concern ourselves about this non-genic fraction. Yet a further fraction, perhaps as much as 50 percent, making up the "medium" and "fast" reannealing DNA, consists of a relatively small number of

different genes that are reduplicated scores, hundreds, or thousands of times. The function of this DNA is unknown. It might be assumed to code for proteins that are required in very large amounts, like hemoglobin or other respiratory pigments. Yet the hemoglobin genes are certainly not present in multiple copies in the basic haploid genome of vertebrates, since hemoglobin variants behave as simple all-or-none characters in inheritance. Some unknown fraction of the genome codes for so-called structural protein, the protein of membranes, cell organelles, muscle, and lens. There is no compelling reason to suppose that the majority of genes might not code for such molecules. Finally, there is no information on the genes in higher organisms that may be important in control of protein synthesis, in a manner analogous to the operon of bacteria. Some large fraction of the DNA of higher organisms may be of this type.

Thus far the "random sample" of genes tested for electrophoretic variation has been restricted to the genes coding enzyme proteins and a few nonenzymatic molecules such as egg white, serum albumins, and larval hemolymph proteins of Drosophila. We know nothing of structural protein or of controlling genes. Moreover, most but not all of the enzymes tested have been soluble enzymes easily extractable from the liquid phase of cells. A few particle-bound enzymes have been studied, and they do not appear to differ in degree of polymorphism from soluble enzymes. Thus, in man, two out of three mitochondrial enzymes are polymorphic (average heterozygosity 0.11), in *Peromyscus polionotus* none of three mitochondrial enzymes is polymorphic, while in *Limulus* two out of two are (average heterozygosity 0.14). It is likely, then, that soluble enzymes are representative of enzymes in general

Even among enzymes there may be differences between groups of functionally related molecules in their degree of heterozygosity. Gillespie and Kojima (1968) postulated that enzymes directly involved in energy metabolism would be less variable because their functions were more "essential" and needed to be more coordinated than would other, peripheral enzymes. An extensive examination of this thesis was made by Kojima, Gillespie, and Tobari (1970) for several species of *Drosophila* and for Selander and Yang's data on the house mouse. I have made the same comparison for man, using Harris and Hopkinson's data, and the results are shown in table 24. Group I enzymes include all those that catalyze steps in glycolysis

TABLE 24
Polymorphism and heterozygosity for enzymes of energy metabolism (group I) and other enzymes (group II) in *Drosophila*, mouse, and man

Species	Number of loci	Proportion polymorphic	Heterozygosity
Drosophila melanogaster[a]			
group I	11	.36	.094
group II	8	.50	.156
D. similans[a]			
group I	11	.36	.030
group II	7	1.00	.364
D. willistoni[b]			
group I	10	—	.112
group II	18	—	.223
Mus musculus[c]			
group I	17	.24	.089
group II	11	.45	.106
Homo sapiens[d]			
group I	24	.21	.048
group II	47	.32	.077

[a]Data are from Kojima, Gillespie, and Tobari (1970).
[b]Data are from Ayala et al. (1972).
[c]Data are from Selander and Yang (1969).
[d]Data are from Harris and Hopkinson (1972).

up to the beginning of the citric acid cycle, including the pentose "shunt." All other enzymes are in Group II. It does indeed appear that, in varying degrees, the enzymes of glycolysis are less variable than the others, but it should also be noted that glycolytic enzymes are overrepresented in the samples relative to the proportion of all enzymes known in metabolism. On this basis, the average genetic heterogeneity over the whole set of enzyme-specifying genes has been underestimated. So neither the distinction between particle-bound and soluble enzymes nor the distinction between functional groups of enzymes leads us to revise our estimates of heterozygosity downward. If anything, they are too conservative.

What of nonenzymatic proteins? There exists a large body of data on genetic polymorphism for human nonenzymatic proteins which at first sight is of no use to us. These are the data on blood groups. It is a cornerstone of our method that we choose our sample of loci without reference to how variable they are. Unfortunately, blood groups are detected from antibody-antigen reactions, so a blood

group locus remains undetected unless there is at least one variant in the population. Moreover, since blood groups are detected in routine cross-matching, the greater the variation, the more likely a locus is to be detected. This bias would disappear, however, if enough bloods could be tested, because presumably every blood group locus is mutable and therefore would eventually be detected as a rare variant. At first, when only a small number of bloods have been cross-matched, only the most polymorphic loci will appear, but as time goes on and a progressively larger number of tests have been run, the cumulative estimate of genetic variation should fall as more and more relatively invariant loci show up.

This historical method is applied to the English population in table 25 and figure 12 (Lewontin, 1967a). As expected, the first blood groups discovered were highly polymorphic, and the estimates of proportion of loci polymorphic and heterozygosity per individual were very high. As the number of individuals tested increased, and as sophisticated methods of analysis were developed, more and more nearly monomorphic loci were found, so that the estimates of polymorphism and heterozygosity reached an asymptote while the number of loci discerned rose steeply. The asymptotic values are approximately 36 percent for the proportion of loci polymorphic and 16 percent for the proportion of an individual genome that is heterozygous. To compare this heterozygosity with that estimated from electrophoretic variation, we must take into account that the antigens are not themselves the primary products of gene action but reflect changes in glycosyl transfer enzymes, and are not restricted to charge changes. On the assumption that electrophoretic changes constitute only about one-third of all amino acid substitution, we get from Harris and Hopkinson an estimate of total heterozygosity for enzymes of $3 \times 0.067 = 0.201$, which is in reasonable agreement with the value of 0.162 for the blood groups. Of course, the factor of 3 is somewhat arbitrary and cannot be taken too seriously. The point is that heterozygosity estimated from electrophoretic variation of enzymes is once more seen to be conservative.

While all the available evidence reinforces the conclusion that a very large fraction of genes in most sexually reproducing organisms is polymorphic, and that each individual in a population is heterozygous at a large fraction of his loci, there will always remain a

TABLE 25
Information on human blood groups

	Blood group	Year discovered	Frequency of most common allele	Hetero-zygosity at locus	Cumulative hetero-zygosity	Proportion poly-morphic
1	ABO	1900	0.437	0.512	.512	1.00
2	MNS	1927	0.389	0.700	.606	1.00
3	P	1927	0.540	0.497	.569	1.00
4	Se	1930	0.523	0.499	.552	1.00
5	Rh	1940	0.407	0.662	.574	1.00
6	Lu	1945	0.961	0.075	.491	1.00
7	K	1946	0.936	0.122	.438	1.00
8	Le	1946	0.815	0.301	.421	1.00
9	Levay	1946	~1.00	~0	.374	.889
10	Jobbins	1947	~1.00	~0	.337	.800
11	Fy	1950	0.549	0.520	.353	.818
12	Jk	1951	0.514	0.500	.366	.833
13	Becker	1951	~1.00	~0	.337	.769
14	Ven	1952	~1.00	~0	.313	.714
15	Vel	1952	~1.00	~0	.292	.667
16	H	1952	~1.00	~0	.274	.625
17	Wr	1953	0.999	0.002	.258	.588
18	Be	1953	~1.00	~0	.244	.556
19	Rm	1954	~1.00	~0	.231	.526
20	By	1955	~1.00	~0	.219	.500
21	Chr	1955	0.999	0.002	.209	.476
22	Di	1955	~1.00	~0	.199	.454
23	Yt	1956	0.995	0.010	.191	.434
24	Js	1958	~1.00	~0	.183	.417
25	Sw	1959	0.999	0.002	.176	.400
26	Ge	1960	~1.00	~0	.169	.384
27	Good	1960	~1.00	~0	.163	.370
28	Au	1961	0.576	0.489	.175	.393
29	Lan	1961	~1.00	~0	.168	.380
30	Bi	1961	~1.00	~0	.163	.366
31	Xg	1962	0.644	0.458	.173	.387
32	Sm	1962	~1.00	~0	.167	.375
33	Tr	1962	~1.00	~0	.162	.364

Note: From Lewontin (1967a). Data refer to the English population.

nagging doubt when we contemplate the few scores of genes examined in man in contrast with the 3 million genes that might be coded by the 3×10^9 nucleotides in the DNA of each of his sperm.

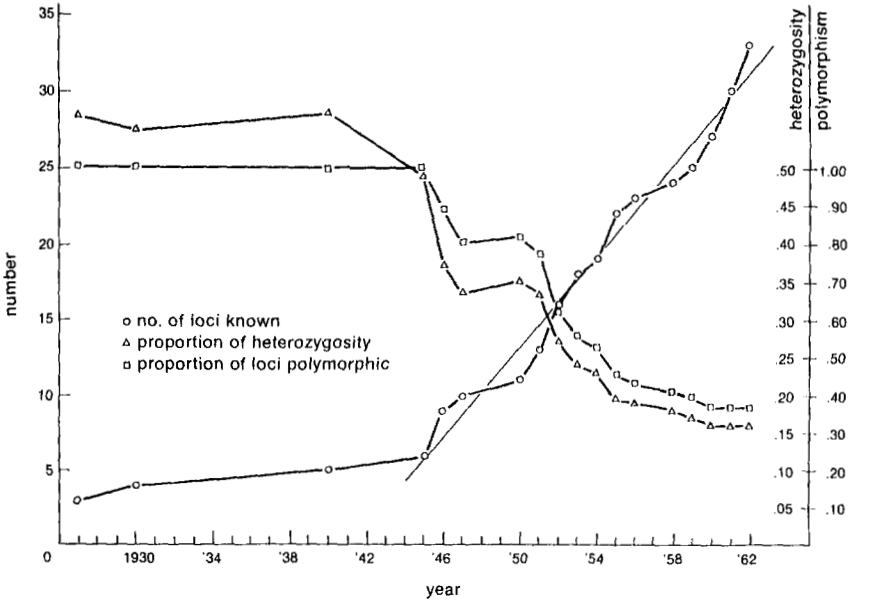


FIGURE 12

Relations between year of discovery and blood group polymorphism. From Lewontin (1967a).

A CLOSE LOOK AT *DROSOPHILA PSEUDOOBSCURA*

After the initial and somewhat unsatisfactory survey of lines of *D. pseudoobscura*, the Chicago group began a systematic survey of a large number of populations in order to see in detail the pattern of allelic variation over the range of the species. *Drosophila pseudoobscura* is distributed over western North America and Central America from southern British Columbia to Guatemala and from the Pacific coast to the Rocky Mountains and central plateaus of Texas and Mexico (figure 13). In addition, there is a large, disjunct population in the highlands of Colombia around Bogotá. The species is typically found in cool forests, being restricted in higher elevations in the more southerly or arid parts of the range. It is found essentially at sea level in northern California, but not below 5500 feet in Guatemala or below 7000 feet in Colombia. The species occurs in the semi-

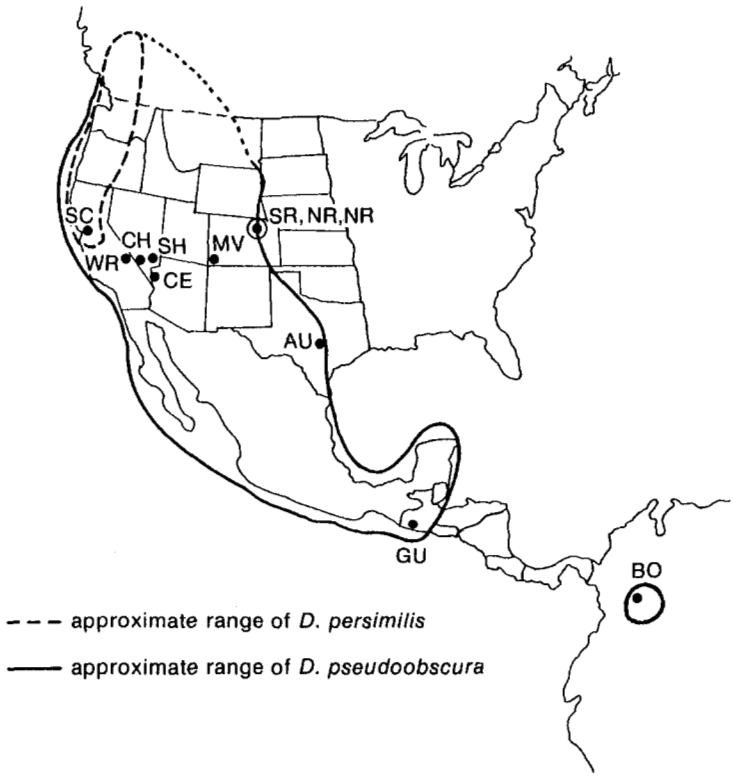


FIGURE 13

Distribution range of *Drosophila pseudoobscura* and *D. persimilis*, and location of populations sampled. Abbreviations are those used in the text.

isolated mountain ranges of the Basin and Ranges region of the western United States, but in the spring it can be found around small oases even in Death Valley and the Mojave Desert. Whether such populations persist, even as aestivating forms, during the summer heat is doubtful. The general impression given is of a widespread species that is almost continuous in its distribution during the favorable spring and fall seasons but that contracts markedly to isolated, cool, moist foci in the hot and dry seasons, at least in the temperate part of the distribution. Since the flies are easy to entice into baited traps when they are active, but extremely elusive in their free state, nothing

is known about their actual distribution in cold winter months, nor have their food sources or breeding sites been found, although they certainly feed on yeast and bacteria (Dobzhansky and Epling, 1944).

Collections designed to span the biogeographic and ecological range of the species have been made from a variety of populations. These are shown on the map in figure 13. One population, SC, is at the center of abundance of the species and has a rich inversion polymorphism on the third chromosome. Five populations, WR, CH, SH, CE, and MV, are from isolated mountain ranges and plateaus in the more arid part of the distribution, forming an east-west transect through the distribution. Two population groups, the cluster SR, HR and NR, and the population AU, are from the extreme eastern ecological boundary of the species range, and the GU sample is from the extreme southern boundary of the main distribution, in the highlands of Guatemala. The last population, BO, is isolated from the rest of the species range by 1500 miles.

Of the 24 loci examined (Prakash, Lewontin, and Hubby, 1969), 11 are monomorphic and homogeneous over the entire species range. The allelic frequencies of the 13 polymorphic loci are given in tables 26A–J and 27A–E. The 10 loci represented in table 26 show three general features. The first, and most obvious, is that heterozygosity at these loci is spread throughout the entire species range. Except for the 2 local polymorphisms, *acid phosphatase-4* and *aldehyde oxidase*, polymorphic loci are generally polymorphic. Second, the allelic frequencies at these loci show a remarkable uniformity over the main contiguous distribution of the species. There is no evidence of polytypy or of geographical races. When one allele is strongly predominant in frequency, this predominance is universal. Moreover, when there are two frequent alleles, as in *xanthine dehydrogenase* (table 26H) with roughly a 3:1 ratio of the two major alleles, or *pt-8* (table 26I) with a 1:1 polymorphism of the two major alleles, these ratios are characteristic of all the population. Even the *esterase-5* locus (table 26J), for which as many as 10 different alleles may be segregating, has a rough similarity, with allele *1.00* being the most frequent in all populations, allele *1.07* being the second most frequent in almost all cases and allele *0.95* being in third or fourth place in all cases but one.

The uniformity of order and of relative frequency of alleles from populations as distant as California, Colorado, and Texas, including

TABLE 26

Frequency of alternative alleles of ten polymorphic loci in various populations of *Drosophila pseudoobscura*

	Populations											
Allele	*SC*	*WR*	*CH*	*SH*	*CE*	*MV*	*SR*	*HR*	*NR*	*AU*	*GU*	*BO*
A. Larval acid phosphatase-4, X chromosome												
0.93	—	—	—	—	—	—	—	—	—	.028	—	—
1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.860	1.000	1.000
1.05	—	—	—	—	—	—	—	—	—	.112	—	—
B. Acetaldehyde oxidase-2, chromosome II												
0.90	.010	—	—	—	—	—	—	—	—	—	—	—
0.93	.030	—	—	—	—	—	—	—	—	—	—	.050
1.00	.940	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.830
1.02	.020	—	—	—	—	—	—	—	—	—	—	.120
C. Malic dehydrogenase, chromosome IV												
0.80	—	—	—	—	.011	—	.019	—	—	—	—	—
1.00	.969	1.000	.936	.929	.954	.948	.962	.882	.904	.957	.727	1.00
1.20	.031	—	.064	.071	.034	.052	.019	.118	.096	.043	.273	—
D. Octanol dehydrogenase-1, chromosome II												
null	—	—	.014	—	—	—	—	—	—	—	—	—
0.75	—	.026	—	—	—	—	—	—	—	—	—	—
0.86	—	—	—	.056	.020	.013	—	—	—	—	—	—
1.00	.977	.871	.951	.902	.939	.961	.965	.972	.885	1.000	1.000	1.000
1.05	—	.077	.014	.028	.010	—	.035	—	.115	—	—	—
1.22	.023	.026	.021	.014	.031	.026	—	.028	—	—	—	—
E. Leucine aminopeptidase, autosome												
0.83	—	.012	—	.012	.009	—	—	—	—	—	—	—
0.90	.008	—	.016	—	—	.025	—	—	.019	.043	.036	—
0.95	.050	.024	.039	.024	.018	.008	—	—	—	.022	—	—
1.00	.892	.916	.897	.893	.954	.940	.875	1.000	.923	.870	.964	.947
1.10	.050	.048	.048	.071	.018	.025	.125	—	.058	.054	—	.054

Allele	Populations											
	SC	WR	CH	SH	CE	MV	SR	HR	NR	AU	GU	BO
F. Pt-7, chromosome II												
0.68	.005	—	—	—	—	—	—	—	—	—	—	—
0.73	.005	—	.021	.014		.009	.040	.025	—	.012	—	.050
0.75	.954	.950	.979	.971	.987	.955	.960	.950	1.000	.966	1.000	.925
0.77	.036	.050	—	.014	.013	.036	—	.025	—	.023	—	.025
G. Pt-13, autosomal												
1.23	.057	—	.082	—	.045	.025	.058	—	.022	.022	—	—
1.30	.943	1.000	.918	1.000	.940	.975	.942	1.000	.935	.978	1.000	.725
1.37	—	—	—	—	.015	—	—	—	.043	—	—	.275
H. Pt-8, chromosome II												
0.80	.014	.025	.008	—	—	.009	.019	.027	—	.011	—	.870
0.81	.472	.450	.600	.514	.473	.410	.539	.595	.480	.441	.625	.093
0.83	.514	.525	.392	.472	.527	.576	.442	.378	.480	.512	.375	.037
0.85	—	—	—	.014	—	—	—	—	.040	.035	—	—
I. Esterase 5, X chromosome												
null	—	—	—	—	.016	—	—	—	—	—	—	—
0.85	—	.013	.007	.047	—	.035	—	—	—	—	—	—
0.90	—	.027	.030	—	—	—	—	—	—	.015	—	—
0.95	.123	.149	.096	.140	.114	.113	.237	.216	.114	.031	.158	.026
0.97	—	.027	.015	—	—	—	—	—	—	.031	—	—
1.00	.424	.460	.356	.419	.317	.365	.474	.486	.341	.292	.579	.974
1.02	.014	.013	.022	—	—	.048	—	.108	.182	.108	.053	—
1.03	.080	—	—	—	—	.039	—	—	—	—	—	—
1.04	.004	.041	.193	.198	.211	.104	.017	.135	.045	.154	—	—
1.07	.193	.243	.200	.174	.260	.196	.271	.054	.273	.262	.210	—
1.09	.009	—	.007	—	—	—	—	—	—	—	—	—
1.12	.132	.027	—	.023	.081	.100	—	—	.045	.046	—	—
1.16	.019	—	.073	—	—	—	—	—	—	.062	—	—

J. Xanthine dehydrogenase, chromosome II												
0.90	.053	—	.007	—	—	.016	.035	—	—	.018	—	—
0.92	.074	—	.030	.040	—	.073	.089	.026	.020	.036	—	—
0.97	—	.133	.098	.077	.012	—	—	—	—	—	—	—
0.99	.263	.200	.188	.173	.131	.300	.286	.289	.220	.232	.278	—
1.00	.600	.667	.647	.710	.857	.581	.555	.633	.720	.661	.722	1.000
1.02	.010	—	.030	—	—	.030	.035	.052	.040	.053	—	—

Note: Data are from Prakash, Lewontin, and Hubby (1969) and Prakash, Lewontin, and Crumpacker (1973).
SC = Strawberry Canyon, Calif.; WR = Wild Rose, Calif.; CH = Charleston, Nev.; SH = Sheep Range, Nev.; CE = Cerbat, Ariz.; MV = Mesa Verde, Colo.; SR = State Recreation, Colo.; HR = Hardin Ranch, Colo.; NR = Nelson Ranch, Colo.; AU = Austin, Tex.; GU = Guatemala; BO = Bogotá, Colombia.

TABLE 27

Allelic frequencies of three polymorphic loci on chromosome III of *Drosophila pseudoobscura* from the same populations as in table 26

	Populations											
Allele	*SC*	*WR*	*CH*	*SH*	*CE*	*MV*	*SR*	*HR*	*NR*	*AU*	*GU*	*BO*
A. Pt-10												
1.02	.005		.007	.014	.015	.022	—	—	—	.010	—	—
1.04	.615	.898	.945	.943	.985	.970	.770	.694	.308	.935	—	—
1.06	.380	.102	.048	.043	—	.008	.230	.306	.692	.054	1.000	1.000
B. α-Amylase												
.74	.030	—	—									
.84	.290	.206	.090	.172	.194	.211	.380	.391	.548	.125	1.00	1.00
1.00	.680	.794	.910	.828	.806	.789	.620	.609	.452	.875	—	—
C. Pt-12												
1.18	.550	.736	.750	.792	.733	.940	1.000	1.000	.972	.900	1.000	1.000
1.20	.450	.264	.250	.208	.267	.060	—	—	.028	.100	0	0

central and peripheral populations, prosperous and ecologically marginal ones, is a fact that must figure critically in any explanation of the vast heterozygosity in the species Different explanatory hypotheses about the origin and fate of genetic variation make different predictions about the similarity between ecologically and geographically diverse populations, as we shall see in chapter 5.

The third feature of the observation is that Bogotá (BO), the isolated Andean population, is an exception to the widespread polymorphism and uniformity of gene frequencies seen in the other populations. For the more polymorphic loci (table 26D–J) the BO population is markedly less heterozygous than the average population from the main part of the distribution, strikingly so for the species' three most polymorphic loci (tables 26H–J). For the *pt-8* locus an allele that is rather rare elsewhere is the preponderant one in Bogotá, but for the other loci it is the most common allele in the species that predominates in Bogotá as well. The most striking case is *esterase-5*, which is immensely polymorphic everywhere except in Bogotá, where allele *1.00* is virtually monomorphic. The depauperate genetic variation in the large, prosperous, but isolated Andean population must also figure strongly in any theory of variation. That it is indeed the isolation of Bogotá, rather than its tropical environment, that is responsible for its unusual characteristics is shown by the data from Guatemala (GU). An extremely small sample was captured there, so the lack of observed polymorphism among the less variable loci is to be expected. For the more variable loci, however, GU is quite as heterozygous as the rest of the main distribution (table 26E–I), in contrast to the homogeneity at these loci in Bogotá.

A very different picture emerges when we look at the three loci of table 27. In place of the uniformity of polymorphism and gene frequency over the main continental distribution of the species, with BO being homozygous for the most common species allele, there is considerable variation in allele frequency from population to population. For two of the three loci, BO is homozygous for an allele that is in a minority elsewhere, and for all three loci GU is identical with BO. There are apparently five geographic races: California, Death Valley, eastern Colorado, Texas, and tropical. The three loci of table 27 are on the third chromosome of *D. pseudoobscura*. This chromosome element is highly polymorphic for a series of overlap-

ping inversions that, when heterozygous, reduce recombination on the third chromosome to less than 1 percent of its normal value (Dobzhansky and Epling, 1948). The frequencies of the inversions show strong geographical differentiation (Dobzhansky and Epling, 1944), so that if the different alleles at the *pt-10*, *pt-12*, and *α-amylase* loci were closely associated with different inversions, the geographical variation in allelic frequencies might be explained.

Prakash and Lewontin (1968, 1971) determined the allelic composition of a variety of gene arrangements sampled from several populations, and their results are given in table 28. There is a very strong association at the *pt-10* locus between the allele *1.04* and the inversion types Standard, Arrowhead, and Pikes Peak, whereas allele *1.06* characterizes the other arrangements. At the *α-amylase* locus one of the alleles, *1.00*, is associated with the three arrangements, Standard, Arrowhead, and Pikes Peak, but the other arrangements are predominantly *0.84*, although the association is not as strong as at the *pt-10* locus. For *pt-12* the variation is much simpler. The Standard gene arrangement is 80 percent $pt\text{-}12^{1.20}$, while that allele is practically absent from all other inversions. These patterns of association are even more striking when the evolutionary relations between the chromosome types are considered. Because the arrangements differ from each other by single overlapping inversions, it is possible to deduce their phyletic relationships. Figure 14 shows that Standard, Arrowhead, and Pikes Peak belong to a cluster of related inversions, the "Standard phylad," connected by single inversion steps but separated from the "Santa Cruz" phylad of inversions by a hypothetical arrangement that has never been found in nature. Moreover, the Standard arrangement is shared between *D. pseudoobscura* and its sibling species, *D. persimilis*, which has its own cluster of inversions, also part of the Standard phylad. Thus it appears that the alleles $pt\text{-}10^{1.04}$ and $\alpha\text{-}amylase^{1.00}$ characterize the Standard phylad of gene arrangements while their alternate alleles are associated with the Santa Cruz phylad. These connections are further borne out by the fact that *D. persimilis* is homozygous for $pt\text{-}10^{1.04}$ and is segregating for $\alpha\text{-}amylase^{1.00}$ and for a number of alleles that are unique to *D. persimilis*, but is virtually without $\alpha\text{-}amylase^{0.84}$, as we expect for a species whose inversions belong entirely to the Standard phylad (Prakash, 1969). The association between specific inversions must then date back to the original split

TABLE 28
Allelic frequencies in different gene arrangements of chromosome III of *Drosophila pseudoobscura*

Gene arrangement	Allele	Populations					
		MA	*SC*	*MV*	*AU*	*CE*	*BO*
A. Pt-10							
Standard	*1.04*	1.00	1.00	×	1.00	1.00	—
	1.06	—	—	—	—	—	—
Arrowhead	*0.94*	0.10	—	0.02	—	—	—
	1.02	—	—	0.97	—	0.02	—
	1.04	0.90	1.00	0.01	1.00	0.98	—
	1.06	—	—	—	—	—	—
Pikes Peak	*1.02*	—	—	—	0.01	—	—
	1.04	1.00	—	1.00	0.99	1.00	—
Santa Cruz	*1.06*	—	1.00	—	—	—	1.00
Chiricahua	*1.04*	0.50	—	—	—	—	—
	1.06	0.50	1.00	—	×	×	—
Treeline	*1.04*	—	—	—	0.33	—	—
	1.06	1.00	1.00	—	0.66	—	1.00
Olympic	*1.06*	—	—	—	×	—	—
Estes Park	*1.06*	—	—	—	—	1.00	—
B. α-Amylase							
Standard	*0.84*	0.05	0.12	—	0.80	—	—
	0.92	0.10	—	—	—	—	—
	1.00	0.85	0.88	—	0.20	×	—
Arrowhead	*0.84*	0.05	0.30	0.21	0.29	0.22	—
	1.00	0.95	0.70	0.79	0.71	0.78	—
Pikes Peak	*0.84*	—	—	—	—	0.04	—
	1.00	1.00	—	—	1.00	0.96	—
Santa Cruz	*0.84*	—	1.00	—	—	—	1.00

between the two phylads and must predate the divergence of *D. persimilis* from *D. pseudoobscura* and the current geographic distribution of these species, which at the most conservative estimate is Arcto-Tertiary, about one million years ago.

The associations of the alleles of *pt-10*, *α-amylase*, and *pt-12* do indeed explain the frequency differences between the populations in table 27. For example, both Guatemala and Bogotá are segregating

Gene arrangement	Allele	Populations					
		MA	SC	MV	AU	CE	BO
α-Amylase (cont).							
Chiricahua	*0.84*	0.36	0.36	—	×	—	—
	1.00	0.64	0.64	—	—	—	—
Treeline	*0.74*	—	0.14	—	—	—	—
	0.84	0.90	0.79	–	1.00	1.00	1.00
	1.00	0.10	0.07	—	—	—	—
Olympic	*0.84*	—	—	—	—	—	×
Estes Park	*0.84*	—	—	—	—	1.00	—
C. *Pt-12*							
Standard	*1.18*	0.20					
	1.20	0.80					
Arrowhead	*1.18*	0.95					
	1.20	0.05					
Pikes Peak	*1.18*	1.0					
	1.20	—					
Chiricahua	*1.18*	1.0					
	1.20	—					
Treeline	*1.18*	0.95					
	1.20	0.05					

Note: Data are from Prakash and Lewontin (1968, 1971) and Prakash, Lewontin, and Crumpacker (1973). When only one chromosome was examined, an X indicates presence of an allele.

MA = Mather, Calif.; SC = Strawberry Canyon, Calif.; MV = Mesa Verde, Colo.; AU = Austin, Tex.; CE = Cerbat, Ariz.; BO = Bogotá, Colombia.

only for inversions of the Santa Cruz phylad and so are expected to be virtually homozygous for $pt\text{-}10^{1.06}$, $\alpha\text{-}amylase^{0.84}$, and $pt\text{-}12^{1.18}$, in contrast to the low frequency of $pt\text{-}10^{1.06}$ and $\alpha\text{-}amylase^{0.84}$ in the rest of the species, in which the Standard phylad is much more common. In general there is an excellent and detailed match between the observed frequencies of alleles for these three loci in different populations and the frequencies predicted from the relations in table

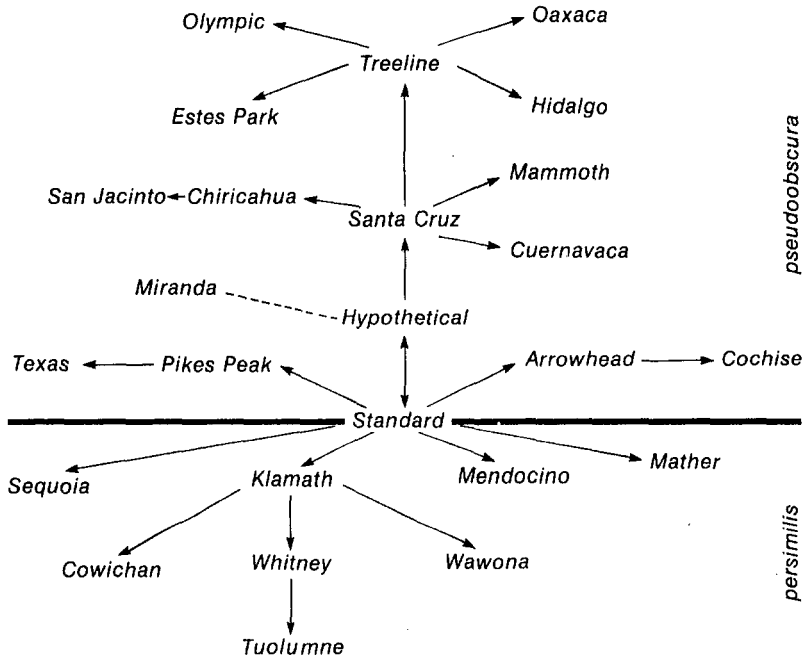


FIGURE 14

Phylogenetic relationships among the arrangements on the third chromosome of *Drosophila pseudoobscura* and *D. persimilis*. Adapted from Dobzhansky and Epling (1944).

28 together with the known frequencies of inversions given in table 29 (Prakash and Lewontin, 1968, 1971; Prakash, Lewontin, and Crumpacker, 1973). Note especially the striking similarity between the total frequency of the Standard phylad in table 29 with the frequency of $pt\text{-}10^{1.04}$ in table 27A.

There is some suggestion, although it has not been critically tested, that variation among populations in the frequency of alleles at the sex-linked *esterase-5* locus (see table 26J) may be the result of similar association between alleles and the "sex-ratio" inversions of the X chromosome (Prakash, Lewontin, and Crumpacker, 1973).

The association of specific alleles with inversions also has the consequence that the alleles of two different loci are not at random with respect to each other. For example, the frequency of gametes of the genotype $\underline{pt\text{-}10^{1.06},\ \alpha\text{-}amylase^{1.00}}$ in the Strawberry Canyon population should be $(0.320)\ (0.680) = 0.258$, according to the

TABLE 29

Frequencies of different gene arrangements of chromosome III in populations of *D. pseudoobscura*

	Populations											
Gene arrangement	*SC*	*WR*	*CH*	*SH*	*CE*	*MV*	*SR*	*HR*	*NR*	*AU*	*GU*	*BO*
Standard	.47	.34	.28	.35	.38	.01	.03	—	.09	.06	—	—
Arrowhead	.09	.44	.67	.60	.52	.98	.42	.40	.28	.08	—	—
Pikes Peak	.03	.06	.02	—	.05	.01	.27	.21	.18	.78	—	—
Total Standard phylad	.59	.84	.97	.95	.95	1.00	.75	.61	.55	.92	—	—
Chiricahua	.19	.11	.03	.05	.04	—	—	.02	—	.01	—	—
Treeline	.14	.04	—	—	.01	—	—	.04	.17	.06	.65	.37
Estes Park	.06	.01	—	—	—	—	.25	.33	.24	—	—	—
Santa Cruz	.01	—	—	—	—	—	—	—	—	—	.30	.63
Olympic	.01	—	—	—	—	—	—	—	.04	.01	—	—
Oaxaca	—	—	—	—	—	—	—	—	—	—	.05	—
Total Santa Cruz phylad	.41	.16	.03	.05	.05	—	.25	.39	.45	.08	1.00	1.00
Proportion of heterokaryotypes	.71	.67	.47	.52	.58	0.04	.69	.69	.79	.38	.49	.47

Note: Data are from Mayhew et al. (1966), Strickberger and Wills (1966), and Prakash, Lewontin, and Crumpacker (1973). The abbreviations are the same as in table 26.

allelic frequencies in table 27. However, because of the complete association of *pt-10*$^{1.06}$ with the Santa Cruz phylad of inversions and the strong association of *α-amylase*$^{1.00}$ with the Standard phylad, the frequency of the gamete is only 0.131. If we consider the triple combination *pt-10*$^{1.06}$, *α-amylase*$^{1.00}$, *pt-12*$^{1.20}$, the frequency would be (0.380) (0.680) (0.450) = 0.116 if the loci were assorted at random, but the real frequency is only 0.005; that is, practically absent.

In chapter 6 I shall discuss the importance of these associations between alleles at different loci as indications of the forces controlling the genetic variation, and as elements in the dynamic theory of genetic change. Clearly, if such associations are frequent, then a sufficient set of dimensions for a description of evolutionary processes must be related to gametic types rather than to allelic frequencies at individual loci.

All the data on *D. pseudoobscura* are summarized in table 30. Once again the lack of any difference among the populations of the main continental distribution is clear. What differences do exist in average heterozygosity per locus disappear completely when inversion heterozygosity is discounted. The two extremes, Strawberry Canyon with 16 percent heterozygosity and Guatemala with 8 percent, turn out to be essentially the same if third-chromosome genes are removed from the calculation. Only the isolated population of Bogotá is clearly differentiated from the rest, departing more than ten standard deviations from the average heterozygosity of the main continental distribution. Yet these continental populations include the complete biogeographical range of the species, from the year-round abundant population of Strawberry Canyon through the populations living at the eastern ecological margin of the species, detectable only in favorable seasons, to the very sparse population of the highlands of Guatemala, where a week of collecting produced only eight individuals of *D. pseudoobscura* among numerous Drosophila, including large numbers of its close relative, *D. azteca*. Moreover, the similarity among the continental populations lies not only at the gross level of average heterozygosity per locus, but at the fine level of close similarity of allelic frequencies.

This uniformity of genetic composition contrasts sharply with the considerable variation in frequencies of third chromosome arrangements among populations (table 29). Since the extensive work of Dobzhansky and his colleagues has shown that these gene arrange-

TABLE 30

Proportion of loci polymorphic out of 24 examined, average number of alleles per locus with frequency greater than 1 percent, and proportion of genome estimated as heterozygous in 12 populations of *Drosophila pseudoobscura*

Population	*Number of polymorphic Loci*	*Proportion of loci polymorphic*	*Average number of alleles per locus*	*Proportion of genome heterozygous*	
				including chromosome III	*excluding chromosome III*
Strawberry Canyon, Calif.	12	.50	2.29	.161	.116
Wild Rose, Calif.	9	.38	1.92	.129	.105
Charleston, Nevada	11	.46	2.21	.126	.113
Sheep Range, Nevada	10	.42	1.96	.125	.108
Cerbat, Arizona	11	.46	1.92	.112	.093
Mesa Verde, Colorado	11	.46	2.12	.117	.109
State Recreation, Colo.	10	.42	1.71	.131	.110
Hardin Ranch, Colo.	8	.33	1.62	.123	.098
Nelson Ranch, Colo.	10	.42	1.79	.140	.113
Austin, Texas	11	.46	2.21	.126	.119
Guatemala	(5)[a]	(.21)[a]	(1.29)[a]	.081	.092
Bogotá, Colombia	6	.25	1.37	.051	.058
Mean excluding Bogotá and Guatemala	10.3	.43	1.97	.129 ± .0058	.108 ± .0034
Grand Mean	9.5	.40	1.86	.119 ± .0081	.103 ± .0048

Note: Data are from Prakash, Lewontin, and Hubby (1969) and Prakash, Lewontin, and Crumpacker (1973).
[a]Sample size was so small that it produced spuriously low values.

ments respond to selection (see Dobzhansky, 1970, chapter 5, for a review), it must be that the populations we have sampled are living under physical and biotic conditions that differ from population to population in a way that is significant to the physiology of the species. Our interpretation of the genic variation we have seen must account for this divergence in the apparent action of natural selection between chromosomal arrangements and allelic substitutions at individual loci not associated with inversions.

GEOGRAPHICAL PATTERNS IN OTHER SPECIES

The pattern of geographical variation in *Drosophila pseudoobscura* has four features: (1) monomorphic loci are identically monomorphic over all populations; (2) most polymorphic loci show no pronounced geographical differentiation; (3) loci associated with inversions show patterns of geographic differentiation characteristic of the inversions; and (4) a completely isolated population is more homozygous than populations in the contiguous range of the species. Several other species show patterns that are consonant with the situation in *D. pseudoobscura*, but there are some interesting differences as well.

Drosophila willistoni is widely distributed in the islands of the Caribbean, in tropical Central and South America as far south as northern Argentina, and as far north as southern Florida. A comparison of four continental populations from Colombia with six Caribbean island populations for 20 loci shows the same kind of overall similarity as in *D. pseudoobscura* (Ayala, Powell, and Dobzhansky, 1971). Of 16 polymorphic loci, 14 have closely similar allele frequencies in all populations, but for 2 there is a marked and consistent difference in frequencies between the islands and the mainland. The average heterozygosity on the islands is 0.162 as compared to 0.184 on the mainland but, because of the small number of localities tested, this difference, while suggestive, is not significant.

One of the most interesting features of the genetics of *D. willistoni* is the rich inversion polymorphism on all of its chromosome arms. Unlike the large overlapping inversions that form the third chromosome polymorphism in *D. pseudoobscura*, the inversions of *D. willistoni* are small, numerous, and often nonoverlapping (da Cunha, Burla, and Dobzhansky, 1950). Unfortunately this has made it im-

possible to estimate the frequencies of individual inversions in various populations, homozygotes being virtually indistinguishable from each other in polytene chromosomes, but heterozygotes can be counted. Average inversion heterozygosity varies widely over the species, with central populations heterozygous for many inversions (9.36 per female in Mojolinho, Goiás, Brazil) but marginal and island populations nearly homozygous (0.27 heterozygous inversions per female on the island of St. Vincent). Although the continental populations in the study of Ayala, Powell, and Dobzhansky (1971) were not examined for inversions, we can judge from their position in the distribution that they are heterozygous for about 7 inversions per female, in contrast to the islands, which vary from 0.23 inversions per female on tiny St. Kitts to 3.18 per female on the very large and proximate island of Trinidad (Dobzhansky, 1957).

The lack of a major difference in genic heterozygosity between the mainland and the islands seems at first in contradiction to the pattern of inversion variation, especially since the inversions are uniformly spread over all the chromosomes and so, presumably, include all or most of the gene loci. However, the connection between inversion heterozygosity and genic heterozygosity only becomes powerful when there is an extreme association between genes and inversions. For example, suppose alleles A and a were in frequencies 0.8 and 0.2, respectively, in one chromosomal type, whereas they were in reverse frequencies, 0.2 and 0.8, in an inverted chromosome. Then populations completely homozygous for either chromosomal arrangement would have a genic heterozygosity of $2(0.8)(0.2) = 0.32$, while a population with equal frequencies of the two chromosomal arrangements would have a genic heterozygosity of 0.50. Moreover, the effect of chromosomal heterozygosity is not cumulative. That is, if there were 18 inversions covering 18 different parts of the genome, each with a gene having the same 0.8:0.2 association, then a population in which all 18 were segregating would still only have an average genic heterozygosity of $(18)(0.50)/18 = 0.50$, although the average number of heterozygous inversions per individual would be 9. The completely structurally monomorphic population, with zero inversions heterozygous per individual, would, nevertheless, be genically heterozygous for 0.32 of its genome. In the absence of detailed matching of gene alleles with inversions in *D. willistoni*, it is therefore impossible to say whether associations like those in *D.*

pseudoobscura exist. The fact that two loci show marked and consistent frequency differences between the islands and the continent suggests that these two may be strongly correlated with particular inversions, some of which become unusually frequent on the islands. For example, inversion IIL-*a* is absent on the continent but 40 percent of flies on Trinidad are heterozygous for it.

The extensive survey of *Drosophila obscura* from Finland replicates the pattern of *D. pseudoobscura*. Populations in southern and central Finland and marginal populations in Lapland were studied for 30 loci (Lakovaara and Saura, 1971a), 16 of which were polymorphic. The 14 monomorphic loci were identical in all populations, and there was little variation in allele frequency among the polymorphic loci; for one esterase locus the most common allele in central Finland was absent in Lapland, replaced in northern Lapland by an allele unique to the north, and for a second esterase locus the commonest allele in central Finland was rare in the south. There was no local differentiation in average heterozygosity, and the marginal populations of Lapland were just as heterozygous as other populations.

Lakovaara and Saura's study of *Drosophila subobscura* gave similar results (1971b). Of 15 polymorphic loci, 9 showed no differences among populations, 2 showed strong differentiation from population to population, and 3 were homozygous in an island population. In these latter three cases an allele that is otherwise rare was fixed on the islands, a curious circumstance and quite different from the Bogotá population of *D. pseudoobscura*, in which the common allele in the rest of the species usually was common or fixed in the isolated population. The three fixed loci in *D. subobscura* do not show any geographical variation outside the islands and are among the most heterozygous loci in the study.

Drosphila subobscura is distributed from Persia to Finland. The Lakovaara and Saura study included only the northern margin and thus more extensive differentiation might exist in the species range as a whole. Extreme southern and northern populations are monomorphic for different inversions, whereas Central European populations are highly polymorphic, so there is a good chance that the species as a whole may be more differentiated in allelic composition (Sperlich and Feuerbach, 1969).

A study of five loci in *Mus musculus*, the house mouse, from 16

North American regions, including an island population from Jamaica, again showed the island population to be virtually monomorphic as compared with the mainland ones (Selander, Yang, and Hunt, 1969). There is fairly general uniformity of pattern over the continental populations, but some large regional similarities can also be detected. Thus, although the major allele at the *est-2* locus is in about the same frequency everywhere, the minor component consists of a single allele in the Northeast but three low-frequency alleles in the Southwest. As another example, there is a 1:1 polymorphism for the hemoglobin β–chain locus in California and Arizona, while for the rest of the continent the polymorphism is about 4:1. Thus there is some correlation between neighboring regions and some differentiation between distant places, but no major variation and no obvious pattern.

Two cases are known in which much more extensive and regular differentiation of gene frequencies occurs. The first is in *Peromyscus polionotus*, the beach mouse, from the Gulf Coast of the United States (Selander et al., 1971). The results of the survey of 32 loci are shown in table 31 and figure 15. Each circle, representing a population, is filled in in proportion to the heterozygosity per individual, a full circle denoting a heterozygosity of 0.10. There is no apparent reason why the population of peninsular Florida should be 75 percent more heterozygous than the main continental part of the distribution. The western beach populations, however, are geographically isolated from the rest of the species and are living in an atypical habitat, grassy beach dunes, as opposed to

TABLE 31
Average heterozygosity per individual in mainland and insular populations of *Peromyscus polionotus*

Locality	*Heterozygosity*
Western beach	
Santa Rosa Island (4, 5)	.0195
Peninsulas (1–3, 6)	.0323
Florida panhandle (7–12)	.0523
South Carolina and Georgia (13–21)	.0506
Peninsular Florida (23–28)	.0810
Eastern beach islands (29, 30)	.0825

Note: Data are from Selander et al. (1971). Population numbers in parentheses refer to figure 15.

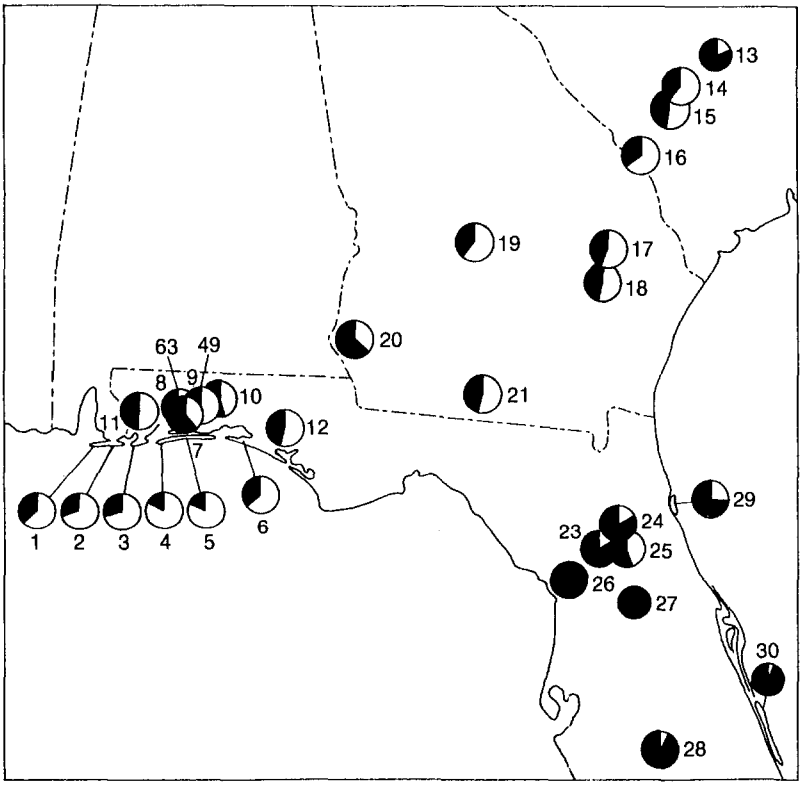


FIGURE 15

The average heterozygosity per population in *Peromyscus polionotus*. The dark area of each circle is proportional to 10× the heterozygosity per individual, a completely darkened circle representing a heterozygosity of 0.10. Adapted from Selander et al. (1971).

the usual abandoned fields that are characteristic of *P. polionotus*. Yet both of these features are characteristic also of the two Florida beach island populations, although isolation from the mainland may be less and population sizes are probably larger. A second characteristic of the western beach island and peninsular populations is a considerably greater heterogeneity of allelic frequencies between populations than for the mainland. Both this heterogeneity and the lower heterozygosity point to isolation and random fluctuation in gene frequencies as important agents in determining the genetic variation of the beach mice.

The second case shows even greater variation among populations than does *Peromyscus*. Stone and colleagues (1968) and F. M. Johnson (1971) studied samples of *Drosophila ananassae* and *D. nasuta* from island groups that stretched 4500 miles across the western Pacific, from Samoa to the Philippines. The island groups are separated by 200 to 2000 miles of open water, whereas islands within a group are 50 to 60 miles apart at most or, in the case of the two Philippine samples, separated by 800 miles of almost continuous land surface.

As usual, there were no loci that were fixed for different alleles in different island groups, but there were many polymorphic loci at which large and consistent difference in allele frequencies occurred between island groups. The largest differences observed were for *acid phosphatase* in *D. ananassae*, in which one of the five alleles had a frequency of 0.81–0.93 in Samoa but only 0.05–0.08 in Palau and 0.025 in Ponape. Somewhat less extreme was *esterase-F* in *D. nasuta*. Here an allele with a frequency of 0.48–0.55 in Samoa had a frequency of 0.07–0.16 in Fiji, only 450 miles away. Not all polymorphic loci were variable, however. The major allele of *leucine aminopeptidase* varied only from 0.79 to 0.96 over the entire set of islands, with most of the variation being contributed by two island groups. The two Philippine localities were remarkably similar in each case despite the 800-mile separation.

Once again we see that isolation seems to be a necessary although not sufficient condition for marked differentiation in allele frequencies. On the other hand, there were no appreciable differences in heterozygosity per individual when averaged over loci, nor were the differences that did occur related to island size or isolation. The highest heterozygosities were found on the vast Philippine archipelago, only 400 miles from the Asian mainland, and on the miniscule island of Yap in mid-Pacific.

The most extreme case of geographical divergence is the polytypic species *Acris crepitans* (Dessauer and Nevo, 1969) that I discussed in the preceding section. An average of only 12 percent of its loci were polymorphic per population, but 60 percent of the loci showed significant genetic variation over the whole species. Four out of 20 loci were completely homozygous within populations yet differed between populations. Three geographical races could be clearly demarcated—an Appalachian, a Gulf, and a Midwestern

race—and the question is open whether these geographical races are indeed exchanging genes in nature, or whether they are in some degree sexually isolated from each other.

An analogous situation exists for the two subspecies *Mus musculus musculus* and *M. musculus domesticus* in northern and southern Denmark, studied by Selander, Hunt, and Yang (1969). Both subspecies are heterozygous (see table 22) but there are profound differences in gene frequencies between them. Twenty loci were identically monomorphic in the two subspecies, but every variable locus was clearly differentiated between the groups. It seems almost certain that there is strong, perhaps complete, reproductive isolation between the two neighboring subspecies so that, as in *Acris*, we are dealing with separately evolving entities rather than a single species.

MARGINAL POPULATIONS

It is often observed that populations at the margins of a species' range are less phenotypically variable than those near the center. Mayr (1963) notes that a study of phenotypically polymorphic species "reveals almost invariably that the degree of polymorphism decreases toward the border of the species, and that the peripheral populations are not infrequently monomorphic."

The accuracy of the generalization is improved if a distinction is made between geographically peripheral and ecologically marginal populations. The area inhabited by a species is a spatio-temporal mosaic of favorable and unfavorable habitats. As the frequency in space and time of the favorable "patches" grows less and less along some geographical gradient, there finally comes a point at which the species is unable to maintain a population. This is a species border, and it will be ill-defined and fluctuate as a result of fluctuations in the environment and in the numbers of immigrants that reach it from the center of the species distribution. As well as being geographically peripheral, populations at this border are also ecologically marginal. Yet the boundaries of the species need not all be of the ecologically marginal sort. For a terrestrial species the seashore is always a periphery, but the populations living there are not necessarily living on an ecological margin in our sense. Thus the Pacific Ocean is the western boundary of the distribution of *Drosophila pseudoobscura*,

but the population living in the San Francisco Bay area is one of the most prosperous and polymorphic in the entire species range.

Alternatively, a species may have an internal "margin," a region of low frequency of favorable habitats surrounded by very favorable regions. An isolated lowland surrounded by continuously increasing elevations is an example. For *Drosophila pseudoobscura* the Isthmus of Tehuantepec in southern Mexico is such an internal margin, and the species is rare or absent there although present in reasonable numbers on either side. Of course there is a danger of circularity in this reasoning, since the ecological requirements of most species are not well known, and a border may be judged an ecological margin if the species is monomorphic there but merely an abrupt border if the species populations are locally polymorphic. The best escape from this circularity is to make the judgment from average population size and its temporal variation, compared with the center of the species.

The greater monomorphism of marginal populations extends to chromosomal as well as morphological variation. I have already pointed out that inversion heterozygosity in *Drosophila willistoni* is very high in central Brazil, with more than 9 inversions heterozygous per female, whereas the species is virtually monomorphic in the islands of the Lesser Antilles. There are few inversions at the northernmost end of the distribution in Florida (2.06 per female) and at the southern end in extreme southern Brazil and northern Argentina, as well as along the southeastern Atlantic coast of Brazil and the western Pacific coast of Ecuador. The degree of chromosomal polymorphism correlates well with an index of environmental diversity based on climate, vegetation, and closely related competitors (da Cunha and Dobzhansky, 1954; da Cunha et al., 1959). We have seen, however, that genic polymorphism in *D. willistoni* is only slightly less on the Caribbean islands than on the mainland of South America so that there seems to be no effect, or at most a slight effect, of marginality on variation at the genic level.

Drosophila robusta is a species of the eastern half of the United States and southern Canada, following closely the distribution of the American elm, *Ulmus americanus*. A rich inversion polymorphism is spread throughout all the major chromosome arms. Populations at the center of the distribution, in Missouri and Tennessee, are heterozygous for 8 to 9 inversions per female, while the southernmost

extension of the range (in central Florida) has only about 1 inversion heterozygous per individual, and the westernmost extension in Nebraska is monomorphic (Carson, 1958). A survey by Prakash (1973) of 20 enzymes and proteins from Nebraska, Missouri, and Florida populations gave heterozygosities per individual of 0.13, 0.14, and 0.15, respectively. Thus, as in *D. willistoni*, there is no relation between genic polymorphism and marginality of the population, or between average genic heterozygosity and average inversion heterozygosity. This does not, however, rule out association between specific alleles and specific inversions except those associations that are nearly complete. There is, in fact, evidence of such associations in *D. robusta* (Prakash, 1973).

The relation between geography and inversion heterozygosity in *Drosophila pseudoobscura* is not simple and does not clearly follow any rule about marginal and central populations. The most chromosomally polymorphic populations are in California, as might be expected from the abundance of the species there, and at the eastern margin in Colorado. The only populations that approach monomorphism are those of Arizona, Utah, and western Colorado. All other populations, including the eastern margin in Texas, the central plateau of Mexico, and the southern margin in Guatemala, have an equal and intermediate degree of inversion heterozygosity. The bottom line in table 29 gives illustrative examples of these generalities. Strawberry Canyon, California, has 71 percent heterokaryotypes, Mesa Verde in Colorado only 4 percent, and Austin, Texas, and Guatemala have 38 and 49 percent, respectively. Nor can we invoke ecological marginality to explain these figures, since the population density of *D. pseudoobscura* is vastly greater in Mesa Verde than in Guatemala. These wide variations in the extent of heterokaryotypy in different populations make a sharp contrast with the relative uniformity of genic heterozygosity shown in table 28. Of course only one chromosome arm is involved in the chromosomal polymorphism of this species. Nevertheless, there is certainly no relationship between marginality or centrality of a population and its average *genic* heterozygosity in *D. pseudoobscura*, especially if the third chromosome is discounted.

There is some disagreement about the explanation for the greater inversion polymorphism seen in central populations of species like *D. willistoni* and *D. robusta*. Dobzhansky (1951) believes the inver-

sions are differentially adapted to various environmental conditions, so that central populations living in a more diverse environment will hold many inversions in stable equilibrium by natural selection, while in marginal environments one arrangement will be most fit because the range of environments is narrow. Carson (1959) believes, however, that special combinations of alleles must be selected in the stringent marginal environments, unlike those that are found in high frequency at the center. Therefore recombination is necessary at the margins in order to promote the formation of these new combinations, which are then driven to homozygosity by natural selection once they have arisen ("homoselection"). Both Dobzhansky and Carson agree that marginal environments demand more specialization than central ones. Their disagreement is only on whether the required specialized genotypes are already in existence, tied up in one or another of the inversions (Dobzhansky), or are novelties that must be produced by recombinational events (Carson).

My own view (Lewontin, 1957) has been quite different. First, in contrast to both Carson and Dobzhansky, I emphasize the *temporal instability* of marginal environments, so that the variation in selection at the margins is at least as great as it is at the center. Second, I have been skeptical (although on no evidence) whether *chromosomal* hetero- and homozygosity can be equated with *genic* hetero- or homozygosity. On the evidence that has now come to light from electrophoretic studies, it appears that such skepticism was justified. Marginal populations are as heterozygous as central ones, despite great differences in chromosomal polymorphism and despite associations of alleles with certain inversions. This does not mean that the chromosomal polymorphism is irrelevant. Quite the contrary. Dobzhansky seems to me correct in his view that in central populations with predictable, spatially diverse environments, a small number of distinct and diverse physiological and developmental modes will be selected. These modes are determined by coadapted genotypes tied up in inversions that prevent recombination. Carson seems to me right when he emphasizes the necessity of recombination in marginal populations in order to produce combinations of alleles that are not represented in the normal modes. But it is not some particular, specialized, homozygous genotype that is being selected in the marginal environment. In the highly unstable and

unpredictable environment of the margin, quite different genotypes are being selected at different times. It is not surprising, then, that genic heterozygosity is high, and remains high, because no particular genotype is favored for very long. The metaphor of the laboratory is sometimes used to describe marginal populations. They are thought of as performing genetic "experiments." If so, these are frustrating experiments, the very opposite of Hershey's Heaven.

RACIAL VARIATION IN MAN

It is a commonplace that the human species is divided into more or less easily distinguishable racial groups. After all, everyone can tell a black African from a Chinese from a European, while to an African, one lank-haired, gray-faced European must look pretty much like another. But we must be careful. The apparent homogeneity within races as compared to the "obvious" difference between them stems partly from the fact that our consciousness of racial differences is constantly being reinforced socially because racial distinctions serve economic and political ends, and partly because the very characters we use to distinguish races—skin color and texture, hair form, eye, nose, and lip shape—are those to which we are most keenly attuned for the purpose of distinguishing individuals. When

TABLE 32
Allelic frequencies at seven polymorphic loci in Europeans and black Africans

	Europeans			*Africans*		
Locus	*allele 1*	*allele 2*	*allele 3*	*allele 1*	*allele 2*	*allele 3*
Red cell acid phosphatase	.36	.60	.04	.17	.83	—
Phosphoglucomutase-1	.77	.23	—	.79	.21	—
Phosphoglucomutase-3	.74	.26	—	.37	.63	—
Adenylate kinase	.95	.05	—	1.00	—	—
Peptidase A	.76	—	.24	.90	.10	—
Peptidase D	.99	.01	—	.95	.03	.02
Adenosine deaminase	.94	.06	—	.97	.03	—
Average heterozygosity per individual		.068 ± .028			.052 ± .023	

Note: From Harris (1970), modified by later data.

TABLE 33

Examples of extreme differentiation and close similarity in blood group allele frequencies in three racial groups

Gene	*Alleles*	*Caucasoid*	*Negroid*	*Mongoloid*
Duffy	*Fy*	.0300	.9393	.0985
	Fy^a	.4208	.0607	.9015
	Fy^b	.5492	—	—
Rhesus	R_0	.0186	.7395	.0409
	R_1	.4036	.0256	.7591
	R_2	.1670	.0427	.1951
	r	.3820	.1184	.0049
	r'	.0049	.0707	0
	others	.0239	.0021	0
P	P_1	.5161	.8911	.1677
	P_2	.4839	.1089	.8323
Auberger	Au^a	.6213	.6419	
	Au	.3787	.3581	
Xg	Xg^a	.67	.55	.54
	Xg	.33	.45	.46
Secretor	*Se*	.5233	.5727	
	se	.4767	.4273	

Note: From the summary of Cavalli-Sforza and Bodmer (1971).

we examine allele frequencies at randomly chosen loci, we get a rather different picture.

Harris (1970) studied 27 enzyme loci in Europeans and black Africans. Of these, 20 loci were identically monomorphic in both groups. The 7 polymorphic loci are shown in table 32. There is no strong differentiation between the groups with exception of the *PGM-3* locus, at which opposite alleles are in the majority in whites and blacks.

Among the 34 known blood group genes, again no case is known in which fixation for different alleles has occurred in different races, but some of the polymorphic genes show a good deal of geographical variation. Summaries of the 15 polymorphic loci are given for "Caucasoids," "Negroids," and "Mongoloids" by Cavalli-Sforza and Bodmer (1971), of which 6 show marked racial differentiation. Table 33 gives the three most extreme examples of differentiation together with the three most homogeneous cases. The ABO system, which shows some large divergences, especially in groups like the

American Indians (some of which are nearly pure O), nevertheless betrays a strong clustering of frequencies (Brues, 1954). Figure 16 shows a representative sample of diverse human groups plotted on a trilineal diagram, showing the strong clustering of ABO allele frequencies around 20 percent I^A, 15 percent I^B and 65 percent i.

It is possible to partition the total genetic variation observed in man into components ascribable to different levels of population aggregation. We can calculate the heterozygosity within national groups from allele frequencies. Let us call the average heterozygosity within national groups H_0. Suppose we now consider a

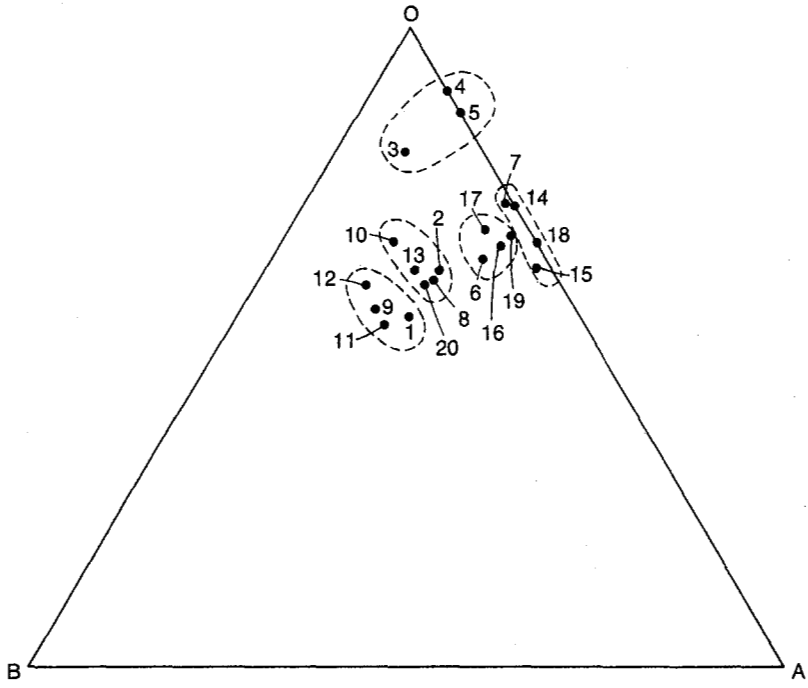


FIGURE 16

Trilineal diagram of the ABO blood group allele frequencies for human populations. Each axis represents one of the alleles (no distinction is made between A_1 and A_2) Each point is a human population. *1–3:* Africans; *4–7:* American Indians; *8–13:* Asians; *14, 15:* Australian Aborigines; *16–20:* Europeans. Dashed lines enclose arbitrary classes based on gene frequency, irrespective of "race." From Jacquard (1970).

hypothetical population made up by merging all the tribal or national populations within one race, for example, Ewe, Kikuyu, Pygmies, and others, making up one large African "race," and calculate the heterozygosity H_1 in this composite. If all the national groups were identical in allele frequencies, then the merger would have no effect and H_1 would be equal to H_0, but if there were some differentiation in allelic frequencies between nations or tribes, the merged group would be more heterozygous. In exactly the same way we could (on paper) merge all the world's races into a single large species group and calculate H_2, the total human species genetic diversity. The relative difference $(H_1 - H_0)/H_2$ is a measure of the proportion of the total genetic diversity that is due to differences between national or tribal groups, and $(H_2 - H_1)/H_2$ is the proportion of the total diversity that arises from racial differentiation.

At present 18 polymorphic genes have been well enough studied over a variety of human populations to make a reasonable calculation possible. Table 34 gives the values of H_0, H_1, and H_2 for these genes, together with the partition of the average heterozygosity. The actual values in the table are not heterozygosities but a nearly identical measure, the Shannon information, $H = -\Sigma\, p \ln p$ (Lewontin, 1972a). Although there is variation between loci in their relative contributions, the average values show that 85 percent of human genetic diversity is *within* national populations and only 7.5 percent between nations within races and 7.5 percent between major races. The relative division between the last two categories depends in part on how national groups are assigned to races and how many "races" are constructed. Are Hindi-speaking peoples of India Caucasoids or a separate group? Do they belong together with Vedic speakers? Where do the Turks or the Lapps belong? There are many possible ways of dividing groups, depending upon linguistic, morphological, or genetic criteria. But these decisions in no way affect the 85 percent variation within groups and only alter in a small way the relative sizes of the between-race and between-nation components.

In any case, the 85 percent within-population component is an underestimate, since all groups were equally weighted in the calculation. But small, isolated groups like the American Indians, Basques, Eskimos, and Australian Aborigines are usually the most deviant from the world average in allele frequencies. Then these very small

TABLE 34
Measure of human diversity within populations, H_0, within races, H_1, and for the whole species, H_2, for 17 loci

Locus	H_0	H_1	H_2	$\frac{H_0}{H_2}$	$\frac{H_1 - H_0}{H_2}$	$\frac{H_2 - H_1}{H_2}$
Haptoglobin	.888	.938	.994	.893	.050	.057
Lipoprotein Ag	.829	—	.994	.834	—	—
Lipoprotein Lp	.600	—	.639	.939	—	—
Xm	.866	—	.869	.997	—	—
Red cell acid phosphatase	.917	.977	.989	.927	.061	.012
6-Phosphogluconate dehydrogenase	.286	.305	.327	.875	.058	.067
Phosphoglucomutase	.714	.739	.758	.942	.033	.025
Adenylate kinase	.156	.160	.184	.848	.022	.130
Kidd	.724	.930	.977	.741	.211	.048
Duffy	.597	.695	.938	.636	.105	.259
Lewis	.960	.993	.994	.966	.033	.001
Kell	.170	.184	.189	.899	.074	.027
Lutheran	.106	.139	.153	.693	.215	.092
P	.949	.978	1.000	.949	.029	.022
MNS	1.591	1.663	1.746	.911	.041	.048
Rh	1.281	1.420	1.900	.674	.073	.253
ABO	1.126	1.204	1.241	.907	.063	.030
Average				.849 within populations	.075 between populations within races	.075 between races

Note: From Lewontin (1972a).

populations make a greatly excessive contribution to the between-group comparisons. Since most of the world's population is made up of Chinese, Indians, Europeans, and the recently hybridized populations of South America, who vary less from each other than do the small isolated groups, the correct proportion of human genetic variation that is within nations or tribes is closer to 95 than to 85 percent.

The taxonomic division of the human species into races places a completely disproportionate emphasis on a very small fraction of the total of human diversity. That scientists as well as nonscientists nevertheless continue to emphasize these genetically minor differences and find new "scientific" justifications for doing so is an indication of the power of socioeconomically based ideology over the supposed objectivity of knowledge.

Indeed the whole history of the problem of genetic variation is a vivid illustration of the role that deeply embedded ideological assumptions play in determining scientific "truth" and the direction of scientific inquiry. Those who, like Monod (1971), think that facts speak for themselves will suppose that the struggle between the classical and balanced schools is over, having been decisively concluded by the hard observations of the new molecular population genetics. But they will be wrong. The classical hypothesis has been developed in extended form, feeding upon, digesting, assimilating, and waxing fat on the very facts that were meant to give it fatal indigestion. It is not the facts but a world-view that is at issue, a divergence between those who, on the one hand, see the dynamical processes in populations as essentially conservative, purifying and protecting an adapted and rational *status quo* from the nonadaptive, corrupting, and irrational forces of random mutation, and those, on the other, for whom nature is process, and every existing order is unstable in the long run, who see as did Denis Diderot that "Tout change, tout passe, il n'y a que le tout qui reste."