An Estimate of Average Heterozygosity in Man

R. C. LEWONTIN

Department of Zoology,
University of Chicago,
Chicago, Illinois 60637.

A FUNDAMENTAL datum in the investigation of the population genetics of any species is the degree of genetic heterozygosity within a population. This heterozygosity can be characterized by the proportion of loci in the genome that are polymorphic, that is, for which alternate alleles are segregating in high frequency, and more precisely by the proportion of loci at which a randomly chosen individual is heterozygous.

Using electrophoretically separable protein variants, Lewontin and Hubby (1966) were able to show that in populations of *Drosophila pseudoobscura* about 30% of all loci have significant segregation of alternate alleles and that about 12% of loci are heterozygous in a typical individual. Hubby and Lewontin (1966) point out that such an estimate can be made only if genetic variation is detectable at single loci and if loci for study are picked at random with respect to how variable they may be. Thus, at first sight, it would seem that an estimate of heterozygosity in man would require testing of electrophoretic differences in proteins detected by their enzymatic activities, just as was the case for Drosophila. Indeed Harris (1966) has used exactly this method in a preliminary survey of ten enzymes of human blood.

The best-known group of human genes would seem to fail Hubby and Lewontin's second criterion. The antigens of human red cells are discovered only when a difference between two individuals exists. Thus, there would be no way of detecting nonpolymorphic loci, and any estimate of average heterozygosity based on such loci would be hopelessly biased. There is one way around this bias, however. Presumably, every locus specifying an antigen is mutable. Therefore, if enough individual bloods could be examined, even a rare mutant could be picked up and, using this mutant as a propositus for a family study, the genetics of the locus in question could be established. In the terms used by workers in this field, "private" and "public" blood group factors would allow us to detect nonpolymorphic loci. I am greatly indebted to Sheldon Reed for suggesting to me that enough human bloods have now been examined to make this approach interesting.

The simple fact that private and public factors are discoverable when very large numbers of bloods are examined is not in itself sufficient to overcome the bias, since polymorphic genes would be the first and most easily discovered. Rare variants will be seen only as the number of bloods examined becomes larger and larger, so that at any particular time the sample of loci is biased toward polymorphic loci; but this bias will grow smaller as the number of bloods examined grows larger. Eventually, when all antigen-

Received April 3, 1967.

TABLE 1. HUMAN BLOOD GROUPS WITH THEIR DATES OF DISCOVERY, GENE FREQUENCY OF THE MOST COMMON ALLELE, HETEROZYGOSITY PER LOCUS, AND CUMULATIVE PROPORTION OF LOCI POLYMORPHIC IN THE ENGLISH POPULATION

Data from Race and Sanger (1962).

	Blood group	Year	Frequency of most common allele	Heterozygosity at locus	Cumulative heterozygosity	Proportion polymorphic
1	ABO	1900	.437	.512	.512	1.00
2	MNS	1927	.389	.700	.606	1.00
3	P	1927	.540	.497	.569	1.00
4	Se	1930	.523	.499	.552	1.00
5	Rh	1940	.407	.662	.574	1.00
6	Lu	1945	.961	.075	.491	1.00
7	K	1946	.936	.122	.438	1.00
8	Le	1946	.815	.301	.421	1.00
9	Levay	1946	~1.00	~0	.374	.889
10	Jobbins	1947	~1.00	~0	.337	.800
11	Fy	1950	.549	.520	.353	.818
12	Jk	1951	.514	.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	.999	.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	.999	.002	.209	.476
22	Di	1955	~1.00	~0	.199	.454
23	Yt	1956	.995	.010	.191	.434
24	Js	1958	~1.00	~0	.183	.417
25	Sw	1959	.999	.002	.176	.400
26	Ge	1960	~1.00	~0	.169	.384
27	Good	1960	~1.00	~0	.163	.370
28	Au	1961	.576	.489	.175	.393
29	Lan	1961	~1.00	~0	.168	.380
30	Bi	1961	~1.00	~0	.163	.366
31	Xg	1962	.644	.458	.173	.387
32	Sm	1962	~1.00	~0	.167	.375
33	Tr	1962	~1.00	~0	.162	.364

specifying loci are known, the bias would disappear. This line of argument suggests that if the average heterozygosity estimated in man on the basis of blood groups is plotted against the year in which the estimate was made, the curve should start high and come down toward some asymptote. This paper shows the results of such a graphing.

Table 1 is a list of all the human blood group antigens known to be genetic, ordered by year of their discovery. The frequency of the most common allele at each locus is given, as currently estimated in the white population of Great Britain (chiefly England). For each locus the average heterozygosity is calculated from the gene frequencies, again in English populations, under the assumption of random mating, and the next to last column of the table

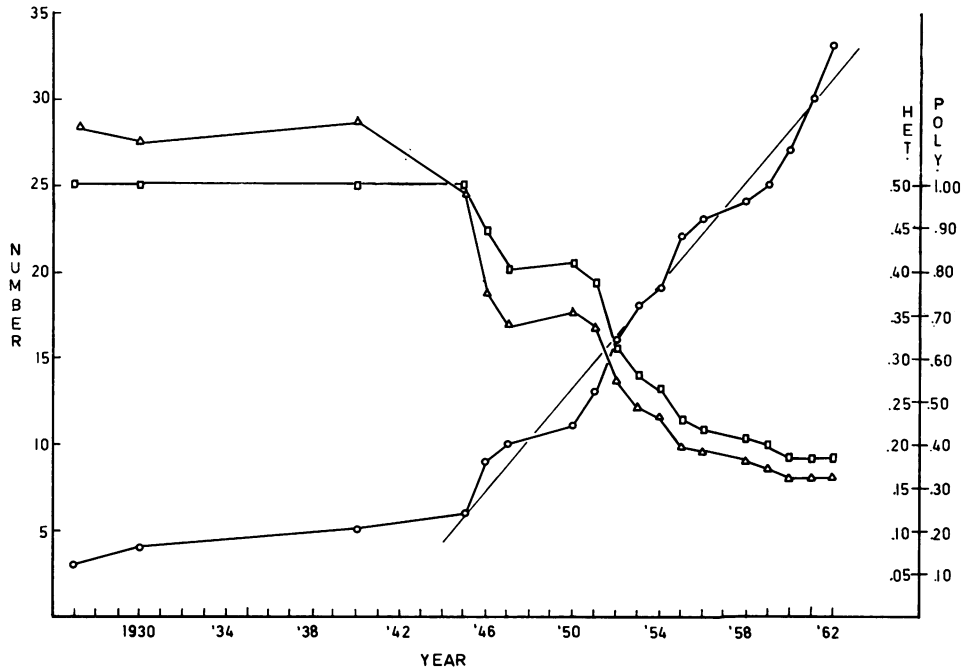

FIG. 1. Relations between year (abscissa) and three variables from Table 1. Circles show number of loci known; triangles show proportion of heterozygosity; squares show proportion of loci polymorphic. Left ordinate: number of loci. Right ordinate: proportion of polymorphic and proportion of heterozygosity.

gives the *cumulative average heterozygosity* estimated over all loci discovered up until that time. The last column gives the proportion of all loci known up to that year that are polymorphic, using as a criterion of polymorphism that a variant allele is in frequency greater than 0.01. The table was constructed from the data given in the last edition of Race and Sanger (1962).

One difficulty in interpreting the data in Table 1 arises because not all possible tests for allelism have been made for private and public antigens. The amount of cross-testing that has been done is variable from factor to factor, and the possibility remains that some private or public factors will be found to be mutants at the same locus as others. Such discoveries of allelism will, of course, *increase* the estimate of polymorphism and heterozygosity since it will decrease the number of monomorphic loci.

As expected, Table 1 shows a strong bias towards polymorphic loci in the early years followed by a long series of private and public factors. It is important to note, however, that polymorphic loci are still being discovered, as evidenced by the discovery of the Au system in 1961 and the Xg system in 1962.

In addition to private and public factors, there are also genes that are essentially homozygous in the English population but are segregating in Africans and their descendents (the Js system) or in Mongoloid Asians and

their descendents (the Di system). These "racial" genes contribute to the number of loci that are polymorphic in the species as a whole but monomorphic within a population.

In Fig. 1, information from Table 1 has been plotted against year. Circles show the increase in the number of genes known with time, the triangles show the change in estimated heterozygosity, and the squares show the proportion of loci polymorphic up to that time. To save space, the curves begin in 1927, the ABO system being included at that point.

The results are striking. Since 1945, the number of loci known has risen linearly with a slope of about 1.6 loci per year, shown by the light line drawn through the points. There is a suggestion of two cusps in the curve, as if the rate of discovery had been exponential until 1955, after which it began a new exponential increase. At any rate, it is clear that the rate of discovery of blood groups did not fall off during that period. The estimated heterozygosity and proportion of loci polymorphic remained high until 1945, when the rapid increase in discovery of genes began. Since that time, their estimates have fallen but with a constantly decreasing slope so that they are approaching or have reached an asymptote. For heterozygosity, the asymptotic value is in the neighborhood of 0.16, while the proportion of polymorphic loci has leveled off at approximately 0.36.

One must exercise some care in judging these asymptotes. If all polymorphic loci had been discovered by say 1951, and since then only monomorphic loci were found, the estimate of heterozygosity and of proportion of loci polymorphic would fall off with ever-decreasing slope and approach zero asymptotically. That is, the fact that the two curves are concave upward rather than linear is a simple arithmetic property of the preponderance of monomorphic loci discovered since 1951. However, the occasional discovery of a new polymorphic locus, such as occurred in 1961 and 1962, prevents the curves from falling asymptotically to zero. If we can assume that the rate of discovery of monomorphic and polymorphic loci is now at equilibrium, as Fig. 1 suggests, then the current apparent asymptotes are the correct ones. Moreover, the number of cistrons controlling antigenic sites on blood cells can hardly be infinite. If we suppose that another 33 are still to be discovered and if, as is unlikely, all 33 should turn out to be monomorphic, the final estimates of polymorphism and heterozygosity would be half what they are now.

The apparent equilibrium in the discovery of polymorphic and monomorphic genes raises an interesting problem. Why have not *all* the polymorphic loci long since been discovered? The answer is that special techniques were required to reveal segregation of these loci because the antigens produced do not stimulate antibody formation under normal conditions. This same difficulty, however, must also apply to monomorphic genes as well. That is, some proportion of private antigens go undetected because they are poor inducers of antibody. Thus the nondetection bias arising from low antibody stimulation is likely to apply to both polymorphic and monomorphic loci, while the bias of sampling that favors polymorphic loci is decreasing with time.

What is remarkable about these results is the similarity to the values found by Lewontin and Hubby (1966) for natural populations of *Drosophila pseudoobscura*. They showed that about 0.30 of all loci are polymorphic in a given population and the average heterozygosity per individual was about 0.12.

Whether red blood cell antigen loci in man can be regarded as a random sample of the same kind as enzyme-controlling genes is a matter for speculation. It may indeed be that red blood cell antigens are a special class of substances that, because of natural selection, are maintained at a higher (or lower) polymorphism than other genes. However, Harris (1966), on a preliminary basis, found three of ten enzymes of human blood to be polymorphic.

SUMMARY

By examining the time sequence of the discovery of red blood cell antigens in man, it has been shown that the proportion of loci for which the English population is polymorphic is about one-third and that the probable heterozygosity per individual is about 0.16. These estimates make use of the asymptotic loss of bias toward polymorphism as the number of loci discovered increases. There is a remarkable similarity of these values to the estimates of polymorphism and heterozygosity in Drosophila and those made by Harris (1966) for enzymes of human blood.

REFERENCES

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

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

LEWONTIN, R. C., AND HUBBY, J. L. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54: 595–609.

RACE, R. R., AND SANGER, R. 1962. *Blood Groups in Man*, 4th Ed. Oxford: Blackwell.