Editorial

A REVISED ESTIMATE OF THE AMOUNT OF GENETIC VARIATION IN HUMAN PROTEINS: IMPLICATIONS FOR THE DISTRIBUTION OF DNA POLYMORPHISMS

The advent of inexpensive and convenient techniques for one-dimensional electrophoresis (1-DE) some 30 years ago quickly led to the demonstration of far more concealed genetic variability than had previously been assumed. Thus, Harris in 1970 [1], summarizing electrophoretic studies, many personal, of some 20 different, arbitrarily chosen human enzymes, the products of some 27 loci, found that polymorphisms for electrophoretic variants had been encountered at six of these loci in Caucasoids and that the average Caucasoid was heterozygous at 5.4% of the loci, an estimate he later increased in the light of additional studies to 6.3% (the Index of Heterozygosity). This perceived heterozygosity was very unevenly distributed across loci, the products of some loci being essentially monomorphic, others being represented by as many as 16 different electromorphs (e.g., PGM_1), with the individual electromorph frequencies ranging from restriction to a single family to frequencies in the neighborhood of .4. By the application of further biochemical techniques, such as sequential electrophoresis, isoelectric focusing, or thermostability studies, many of the common electromorphs could be subtyped, an exercise that created an even higher Index of Heterozygosity.

The electrophoretic approach, as Harris and others pointed out, would not be expected, in principle, to detect amino acid substitutions that do not alter molecular charge or, with any precision, variants characterized by decreased or absent enzyme activity. Since it can readily be calculated that for each nucleotide mutation resulting in a charge change in a polypeptide there should be at least two more resulting in "silent" amino acid substitutions ([2, 3], inter alia), it could be presumed that there was substantially more genetic variation than electrophoresis was detecting—even though the amount revealed by electrophoresis was already straining aspects of current genetic theory regarding the maintenance of variation in populations.

A number of relatively recent technical advances now render feasible population surveys of some of the previously studied proteins for the frequency of variants other than electromorphs, and the advent of two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) [4–6], coupled with autoradiography or sensitive, silver-based protein stains [7–9], permits the extension of the search for variation to a variety of other polypeptides. In addition, the widespread application of the "restriction-site" endonucleases to the study of variation at the DNA level is rapidly providing a set of data on which the results of the studies on polypeptides bear importantly.

This editorial will: (1) consider the impact of a number of recent studies for our view of the extent of the genetic variation exhibited by human proteins, (2) develop a revised estimate of the Index of Heterozygosity for loci encoding for

these proteins, (3) consider how this estimate can be reconciled with the emerging data on DNA polymorphisms, and (4) suggest some basic issues created by all these new data.

The Frequency of "Enzyme-Deficiency" Variants in Human Populations

There are numerous examples of "enzyme deficiency" variants in human populations, some in polymorphic frequencies. The compendium edited by Stanbury et al. [10] is a storehouse of information arising from clinical studies. The findings of such studies, however, do not provide data generally applicable to populations. Such data can result only from a proper survey for heterozygotes for such variants, such surveys avoiding bias by the selection of enzymes for study only because (1) simple and accurate colorimetric or similar methods to determine enzyme levels are available, and (2) the coefficient of variation ([SD/mean] × 100) is of the order of 11%. This latter constraint creates a high probability of differentiating between persons with normal enzyme levels and those with 50% of normal, that is, heterozygotes for an allele whose product (if any) has little or no activity. The surveys that have been undertaken have been greatly facilitated by development of the centrifugal fast analyzer [11–13].

A survey of nine electrophoretically normal enzymes for activity values approximately 3 SD below normal in a sample of newborn infants and adults in Ann Arbor, Michigan, yielded 20 variants in 7,951 determinations, an average frequency of 2.5/1,000 for the enzymes surveyed [14]. (The enzymes surveyed were adenylate kinase [AK, E.C.2.7.4.3], aspartate amino transferase [GOT_1, E.C.2.6.1.1.], glucosephosphate isomerase [GPI, E.C.5.3.1.9], lactate dehydrogenase [LDH, E.C.1.1.1.27], malate dehydrogenase [MDH, E.C.2.7.2.3], phosphoglycerate kinase [PGK, E.C.2.7.2.3], pyruvate kinase [PK, E.C.2.7.1.40], triosephosphate isomerase [TPI, E.C.5.3.1.1], and glucose-6-phosphate dehydrogenase [G6PD, E.C.1.1.1.49].) The ethnic composition of the sample was approximately 90% Caucasoid and 10% American black, with a sprinkling of Orientals. All the deficiencies were shown to be inherited by the presence of the deficiency in a parent; the essentially bimodal nature of enzyme activity levels in the population strongly suggests simple Mendelian inheritance. Twelve of the variants involved TPI; seven of these were in the 146 blacks in the sample, in which ethnic group this variant is, thus, by definition a polymorphism [15]. (The possibility, of course, exists that several different variants contribute to this polymorphism.) A similar survey of Japanese, which included two additional enzymes (glyceraldehyde phosphate dehydrogenase [GAPD, E.C.1.2.1.12] and hexokinase [HK, E.C.2.7.1.1]), yielded among 26,634 determinations 59 enzyme deficiency variants confirmed by repeat blood samples and an additional 11 presumed variants in which a second sample could not be obtained for verification. For the 52 instances in which family studies were possible, again, in all instances, one parent exhibited a similar (qualitative) deficiency. The unweighted average frequency of variants was 2.4/1,000 determinations [16]. A third such study has been conducted on Amerindian samples (11 enzymes); among 6,741 determinations, there were 11 instances of approximately half-normal values, all inherited, a frequency of 1.5/1,000 determinations [17]. The average frequency in these three studies is 2.1/1,000 determinations.

The Frequency of Thermostability Variants in Human Populations: Implications for the Frequency of "Silent" Amino Acid Substitutions

Numerous studies of the thermostability of the electrophoretic variants discovered in various *Drosophila* species have resulted in a subdivision of these electromorphs into distinct types on the basis of their response to heat. An analysis of a variety of reports [18–23] suggests that at least two-thirds, and possibly three-quarters, of the variants recognized by electrophoresis exhibit easily detected thermolability, a relatively common electromorph typically being subdivisible into three to five variants. Studies of *Peromyscus* electrophoretic variants yielded comparable results [24]. These same papers demonstrate the ability to detect thermolability variants whose electrophoretic behavior is normal, but the design of the studies renders a generalization as to the total heterozygosity revealed by these techniques difficult.

Comparable studies on human material are much less extensive and, for the most part, of necessity, based on heterozygotes. Satoh and Mohrenweiser [25] found that among 20 individuals heterozygous for an electrophoretic variant of phosphoglucose isomerase, termed GPI $4_{HIR\,1}$, three subtypes could be distinguished on the basis of thermostability studies: one stable, one labile, and one very labile. Wurzinger and Mohrenweiser [26] demonstrated that the majority of the electrophoretic variants of glutamic oxaloacetic transaminase exhibited altered thermostability. Thus, the human data also suggest that a high proportion of variants presumably due to amino acid substitutions that alter molecular charge also alter thermostability.

These various findings suggested that a survey of an unselected series of electrophoretically normal enzymes for thermostability should reveal some considerable fraction of the "silent" amino acid substitutions, that is, the substitutions not detectable by electrophoresis. We have now carried out two such surveys. For these screening purposes, enzyme activity has been studied at three temperatures and a thermostability variant has been defined as a variant characterized by a percent remaining activity after incubation that was more than 2.5 SD below the mean at the intermediate incubation temperature and also at one of the other two temperatures. Empirical justification for this definition is provided by the fact that of the previously referenced human electromorphic heterozygotes analyzed for thermostability all showed departures from normal of this or greater magnitude. It is, incidentally, recognized that in the surveys of human populations genetic proof that a thermolabile variant involves the structural gene in question (rather than an associated locus) is usually impossible.

In the first survey, placental blood samples from 100 newborn infants from the Maternity Service of the University of Michigan Hospitals were examined for thermolabile variants of eight erythrocyte enzymes, these a subset of the nine examined for activity variants (see above) [27]. Three such variants were identified, none of which was detectable by standard electrophoretic techniques, with a frequency of 3.8/1,000. The genetic nature of all variants was confirmed by family studies; monogenic inheritance is assumed from the clear-cut departure from normal in the affected parent.

The second survey involves a series of Japanese, examined with respect to seven of the eight enzymes studied in Ann Arbor, Michigan [28]. The number

of determinations per enzyme varied from 399 to 1,049, for a total of 5,930 determinations. Electrophoretic variants were considered separately. A preliminary analysis of the findings reveals an unweighted variant frequency of 2.4/1,000 determinations; when family studies were possible, all variants were also present in the father or mother.

The Lower Estimates of Heterozygosity from Two-Dimensional Electrophoresis of the Abundant Soluble Proteins

The advent of two-dimensional electrophoresis (2-DE) some 8 years ago provided a new approach to the question of the heterozygosity index. In this procedure, the proteins of a cell type, tissue, or body fluid are solubilized and then separated in two dimensions, the first on the basis of charge by isoelectric focusing and the second on the basis of molecular weight by electrophoresis in the presence of dodecyl sulfate. Since the solubilization mixture contains urea in a 4–9 M concentration, any multimeric proteins are dissociated into their individual subunits. The polypeptides are visualized either through autoradiography or by the application of Coomassie Blue or sensitive, silver-based stains. Not all the proteins of the cell are visualized: some do not enter the gel, some migrate through and off the gel, and some are present in such small amounts that (unless the gel is grossly overloaded with respect to other proteins) they cannot be visualized with clarity. Thus, even this technology fails to demonstrate some cellular proteins.

Thus far, seven studies of human material have been reported by other investigators. We present them in chronological order. McConkey et al. [29] examined the heterozygosity revealed in four human diploid fibroblast lines by double-label autoradiography. Since the number of moieties scored differed from line to line, exact numerators and denominators were not computed, but the authors conclude that "among several hundred polypeptides . . . the average heterozygosity represented by this set of gene products appears to be less than 1% for changes involving charged amino acids." Walton et al. [30], studying five lines of normal human fibroblasts in which the proteins were, as above, scored from double-labeled autoradiographs, and again employing an approach in which it is difficult to specify numerators and denominators, conclude that "only about 1.2% of the proteins of different cell lines were found to differ in their electrophoretic mobility. This corresponds to an average heterozygosity of approximately 0.6%." Smith et al. [31] found "no genic variation" in a survey of 25 human kidneys with respect to 83 proteins, the positions of the proteins being identified by Coomassie Blue R-250 staining. Hamaguchi et al. [32, 33], contrasting approximately 250 polypeptides in phytohemagglutinin-stimulated peripheral blood lymphocyte preparations from three unrelated healthy males and using double-label autoradiography, observed only three variable polypeptides; their findings suggest an Index of Heterozygosity of $0.5 \pm 0.3\%$. More recently [34], studying the "100 or so most intensely Coomassie Blue-stained" polypeptides from PHA-stimulated peripheral blood lymphocytes, they identified four polymorphic polypeptides, but given the ambiguity with which the exact number of proteins being scored is presented, the data do not permit a precise calculation of the Index of Heterozygosity. Comings [35], employing Coomassie Blue R-250 staining of 2-D PAGE preparations of 145 brains obtained from patients who had died of a variety of

neurological diseases, observed polymorphisms in only two of the 176 polypeptides that he scored. Eleven presumed heterozygotes were encountered among the 24,600 polypeptides examined, resulting in an Index of Heterozygosity of 0.04 ± 0.01%. Klose et al. [36] likewise detected less than 1% heterozygosity in the proteins of freshly cultured human fibroblasts and hair-follicle cells. Finally, Goldman and Merril [37], employing ^{14}C-labeling of phytohemagglutinin-stimulated human lymphocytes, observed that of 186 proteins scored for variation in 28 persons 19 exhibited variation, for an average heterozygosity of 2.4 ± 0.2%.

Our own experience with the frequency of detection of heterozygosity in 2-D PAGE preparations, employing the silver-staining technique, has been somewhat different. A preliminary study suggested that 2-D PAGE resolved roughly 80% of the variants visualized by a variety of 1-DE approaches specially tailored to the specific protein under study [38]. Until now, we have examined three types of preparations for genetic variation: plasma, erythrocyte lysate (membrane removed), and platelets. Among a total of 20 plasma polypeptides chosen for scoring because of a relatively isolated position on the gel and a rather high staining density, the Index of Heterogeneity in 63 individuals was 6.2 ± 0.7% [39]. Among 46 polypeptides of red cell lysates, selected for scoring on the same basis, the Index of Heterozygosity in 27 individuals was 3.1 ± 0.5% [40]. Finally, among 52 polypeptides of platelets, similarly selected, the Index was 2.0 ± 0.4% [41]. All variants were confirmed as heritable by family studies. Average heterozygosity for the three types of preparations was 3.8 ± 0.3%. The possible reasons for the obvious discrepancy among these findings and those of other investigators, as well as the lack of agreement with the studies employing 1-DE, include differences in the cell types and gene products under examination, different techniques, greater care on our part to select polypeptides that can be scored with accuracy, and the routine examination of material from nuclear families (child, father, mother).

That the explanation may involve several of the above reasons is suggested by the findings of Singh and Coulthart [42], who examined the abundant soluble proteins of *Drosophila melanogaster* and *Drosophila pseudoobscura* (i.e., the types of proteins visualized in 2-D PAGE) by 1-DE, using 5% polyacrylamide gel slabs. They found the average heterozygosity to be approximately intermediate between the results of 1-DE on enzyme proteins and 2-D PAGE on the more highly represented soluble proteins, confirming our experience. Further studies are obviously needed to determine the amount of heterozygosity revealed by this relatively new technique. While for a specific protein visualized on a gel 2-D PAGE may not be as efficient in detecting electrophoretic variants as 1-DE specifically tailored to that protein, the technique, on the other hand, is detecting a class of variants in the molecular-weight axis that may not be detected by 1-DE [39]. For present purposes, we suggest that a preliminary estimate of the Index for human proteins revealed by this methodology should be 3%–4%.

The Total Index of Heterozygosity for Protein Variants

The foregoing data permit a more precise evaluation than previously of the average heterozygosity characterizing a series of human genetic loci encoding

for proteins, unselected with reference to their variability. Not unexpectedly, several issues immediately present themselves.

First, how representative are the loci whose products have been examined? This question is especially relevant with reference to the enzymes examined for the presence of thermostability and deficiency variants. Table 1 presents the findings in this laboratory with respect to the frequency of electromorphic variants of these same enzymes; the samples studied for enzyme deficiency and thermolability variants are a subset of a larger sample studied for electrophoretic variants. Note that the Index of Heterozygosity for electromorphs of this series of enzymes is quite low (1.1%). This was not a deliberate choice: the enzymes biochemically suitable for studies of deficiency and thermolability variants were less variable electrophoretically than the average. On the other hand, it seems reasonable to assume that the ratio of detected heterozygotes for electromorphic variants to detected heterozygotes for thermolability variants not characterized by abnormal electrophoretic mobility would be similar across loci. For these data, that ratio can be calculated to be approximately 3:1.

Second, what proportion of the total variation is detected by electrophoresis? Johnson [43, 44] was apparently the first to present data indicating that conformational as well as charge changes in protein molecules could alter electrophoretic mobility. Several investigators have now extended that observation to a study of the electrophoretic behavior of variant hemoglobin molecules in which different amino acid substitutions resulting in the same charge change have occurred [45, 46]. Differing mobilities between identically charged hemoglobin molecules can often be demonstrated by careful measurement. The majority of these are so small, however, that they would probably go undetected in routine laboratory screening. Nevertheless, Bonhomme and Selander [47] suggested, on the basis of such demonstrations, that standard electrophoresis detects on average ap-

TABLE 1

THE FREQUENCY OF POLYMORPHISMS OBSERVED IN OUR LABORATORY AMONG THE EIGHT OF THE NINE ENZYMES SELECTED FOR THE DETECTION OF ENZYME DEFICIENCY AND/OR THERMOLABILE VARIANTS FOR WHICH 1DE ELECTROPHORETIC DATA ARE ALSO AVAILABLE

	PHENOTYPES					
ENZYME	Normal	Heterozygous for slow V	Heterozygous for fast V	Homozygous for slow V	Homozygous for fast V	Σ
AK	2,272	190	2	2	0	2,466
GOT	2,477	2	7	0	0	2,486
GPI	2,355	4	9	0	0	2,368
LDH B	2,367	1	...	...	...	2,368
MDH	2,368	0	0	0	0	2,368
PGK(M)	833	0	0	0	0	833
(F)	756	0	0	0	0	756
TPI	2,365	0	2	0	0	2,367
G6PD(M)	1,015	0	0	3	2*	1,020
(F)	932	0	3	0	1	936

NOTE: Only Caucasoids are included. V is the abbreviation for variant.
* Hemizygosity.

proximately 50% of the alleles at structural gene loci in the house mouse. Ayala [48] goes further, suggesting that "standard electrophoresis may detect most of the protein variation present in natural populations." From these various statements, it would seem that standard electrophoresis should detect some 25%–50% of uncharged amino acid substitutions.

Third, what proportion of the silent amino acid substitutions present in electrophoretically normal enzymes are detected by surveys for thermolability? This question can only be treated accurately when the thermostability of a series of non-charge-change amino acid substitutions ascertained through protein sequencing or some other effective approach has been determined. As noted above, the mutational process should give rise to at least two uncharged amino acid substitutions for each charged substitution. A priori, one would not expect the selective forces acting on these two types of substitutions to be greatly different, but if differences did exist, one would expect uncharged substitutions to be less disruptive to molecular functioning than charged ones. Thus, polymorphisms for substitutions not altering molecular charge should accumulate to at least the same degree as for charged ones. That thermostability variants can occur in human populations in polymorphic proportions is demonstrated by numerous incidental findings. For instance, Scozzari et al. [49] identified a segregant of the complex PGM_1 locus [50] that confers heat lability; 11% of the alleles at this locus were characterized by this property.

Our own surveys have failed to reveal any polymorphism for thermolability variants among electrophoretically normal samples. Most biochemists with whom I have discussed the point suggest on general principles that amino acid substitutions that alter molecular charge would be more apt to alter thermostability than uncharged substitutions. On the other hand, given the high proportion of charge-change substitutions that do alter thermostability, one is inclined to speculate that a considerable proportion of non-charge-change substitutions not detected by electrophoresis should influence thermostability to an extent detectable by the techniques employed. Let us make the arbitrary assumption that thermostability studies should detect about one-third of the non-charge-change amino acid substitutions not detected by electrophoresis. This implies that the combination of electrophoresis and thermostability studies detects no less than one-half of the non-charged amino acid substitutions, and possibly as much as two-thirds.

Fourth, what proportion of variants characterized by absence of enzyme activity can be detected in the heterozygous condition? While there is room for considerable equivocation concerning the proportion of silent amino acid substitutions being detected by current techniques, there is much less uncertainty regarding the enzyme deficiency variants in this series. We assume detection to be quite efficient for carefully selected enzymes. Only one polymorphism in electrophoretically normal enzymes was detected in our work, but other examples of such polymorphisms detected in the course of surveys (often involving very specific subpopulations) are provided by red cell peptidase A [51], adenosine deaminase [52], and lactate dehydrogenase [53, 54]. Clinically oriented studies, of course, have defined a number of such polymorphisms, the most notable being those involving hexosaminidase A and the thalassemias, both α- and β-type. It would be premature to

attempt to assign an average locus figure for deficiency variants, but in the context of developing a heterozygosity index based on proteins, to be contrasted with one based on nucleotide substitutions, I suggest the appropriate index for this type of variant can scarcely on the average exceed 0.5%.

Fifth, how much does the subdivision of electromorphs into thermostability classes alter the Index of Heterozygosity? Because some individuals apparently homozygous for a variant are actually heterozygous, subdivision increases the Index. The magnitude of the increase depends on the number of alleles revealed by the subdivision and the frequency of the variant alleles. At the average levels of heterozygosity for electromorphs in human populations, apparent homozygotes for variants are so infrequent that the increase in heterozygosity from this cause will be small.

As noted above, on the basis of electrophoretic studies, the Index of Heterozygosity in humans for a series of 50 loci encoding for enzymes is currently set at 6.3%. The data we have reviewed may be conservatively interpreted as indicating that the number of non-charge-change amino acid substitutions *not* detected by electrophoresis and thermostability studies is likely to be < the number detected. If we apply the 3:1 ratio derived earlier, then for a series of typical enzymes, the frequency of detected electromorphs and/or thermostability variants becomes 8.4%; this should be one-half to two-thirds of the variation due to amino acid substitutions. Heterozygosity for enzyme deficiency variants will probably not, on average, exceed 0.5%. Thus, it seems unlikely the per-locus Index of Heterozygosity in humans for enzyme loci (even after correction for subtypes of electromorphs) is greater than about 17%–18%, with a more probable value about 15%. If we assume that 2-D PAGE is detecting *about* the same proportion of the genetic variation in the abundant soluble cellular proteins as 1-DE, then, the comparable figure that issues from the 2-D PAGE studies is about 9.0%. The average of these two values, which we will take as the current best estimate of the Human Index of Heterozygosity for loci encoding for proteins, is 12%–13%.

The Early Results from Restriction-Site Endonucleases

With the advent of the widespread use of restriction-site endonucleases and the recognition of restriction fragment length polymorphisms (RFLPs), a new level of study of genetic polymorphisms was established. The first attempt to calculate average heterozygosity, by Jeffreys [55], based on the human β-globin gene complex, yielded an estimate of an average of one variable site per every 100 nucleotides. This estimate was amended to 1 in 200 by Ewens et al., leading to the suggestion of one heterozygous nucleotide pair per every 1,000 nucleotide pairs [56]. Subsequently, the estimate of heterozygosity for this complex has been further amended, to 1 in 500 nucleotide pairs [57]. A recent study of the human albumin gene yielded a similar figure for the frequency of RFLPs, namely, 1 in 95 nucleotide sites [58].

In addition to RFLPs presumably due to single nucleotide substitutions, polymorphisms due to DNA insertions, deletions, and rearrangements are also being observed. One of the first such studies, involving a probe of unknown function some 16,000 nucleotides (16 kb) in length characterized by DNA rearrangements,

yielded at least eight haplotypes, and more than 75% of 43 individuals examined were heterozygotes [59]. Shortly thereafter, a fragment length polymorphism some 700 basepairs (bp) upstream from the single human insulin gene was identified, the polymorphism apparently due to insertions of DNA of two general-size classes: below 600 and 1,600–2,000 bp [60, 61]. The exact number of alleles is not clear, but at least 63% of the 52 samples exhibited heterozygosity. How variations of this type will ultimately be shown to relate to a protein product is unclear, but the example of the thalassemias reveals the importance of events outside exons to gene expression.

Although these unusual situations are instructive, what is important in the present context is the expectation of heterozygosity for RFLPs in exons on the assumption that these RFLPs are randomly distributed in the genome. Let us assume that the average protein scored in the past for genetic variation has a mol. wt. of 45,000 and is composed of about 400 amino acids, coded by 1,200 nucleotides. The data just reviewed on the β-globin gene complex and the albumin gene indicate that if the RFLPs were randomly distributed throughout the exons, there should be six polymorphic sites in the DNA representing these exons. In the absence of "linkage disequilibrium," there should be 2^6 haplotypes. In fact, however, there will be linkage disequilibrium, often marked [cf. especially 62], and any calculation of the expected heterozygosity at a locus depends not only on the degree of this disequilibrium, but also, of course, on the frequencies of the various haplotypes. The intuitive expectation is for a very high level of heterozygosity at such a locus, certainly the equivalent of the findings of Wyman and White [59] quoted above. We are thus confronted with the paradox that as estimates of the Index of Heterozygosity for proteins fall, estimates of DNA heterozygosity rise: there seems to be a clear discrepancy between the data on DNA and the results of the protein studies, a discrepancy that is enhanced when it is recalled that the enzyme-deficiency variants, like the thalassemias [57], may result from changes in both flanking and intron regions.

The most obvious solution of this paradox leads to a prediction: that these RFLPs will not be found to be distributed randomly along the DNA, but will tend to cluster either in noncoding regions (i.e., outside exons) or to involve substitutions in "uninformative" coding positions. One is probably justified in an even stronger statement: either the biochemical techniques have been *very* inefficient in the detection of protein variants or the polymorphic sites of DNA are *very* nonrandomly distributed. Direct evidence on this point is now becoming available. In his pioneering study, Jeffreys [55] noted that all three of the "restriction enzyme cleavage site variants detected probably lie within intervening sequences" (p. 6) and raised the question of a nonrandom distribution. Subsequently, Kazazian et al. [57] pointed out that of 12 RFLPs identified in the β-globin complex, seven were in flanking DNA; three, in intervening sequences; one, in a pseudogene sequence; and only one in the coding sequence of the β-gene. Since the β-globin complex involves 50 kilobases (kb) of DNA, of which the expressed gene sequences occupy 15%, these findings are in the direction of nonrandomness. However, the first really convincing evidence for nonrandomness at this locus has now been supplied by Poncz et al. [63], on the basis of establishing the continuous DNA

sequence of a 16.5-kb-pair region within the β-globin complex, encompassing the linked δβ-globin gene cluster of humans. From a comparison with the published data of others, involving 7,101 nucleotides, two polymorphic sites were detected in the 885 nucleotides comprising the six exons of the δ and β genes (0.22%), whereas there were 36 such sites in the remaining 6,216 nucleotides (0.58%). The frequency for the entire region (38/7,101; 0.54%) is in good agreement with the earlier estimates. While these data constitute a small and perhaps atypical sample from which to operate, it is clear that the degree of nonrandomness suggested by this observation should reduce substantially the disproportion between the heterozygosity observed at the DNA level and that observed at the protein level. Perhaps the most striking evidence to date for the nonrandomness to which the protein data point emerges from recent studies of the alcohol dehydrogenase locus of *Drosophila melanogaster* [64]: among 11 cloned "genes" at this locus, 2,721 bp in length, of which 765 bp are in exons, 43 "polymorphisms" (i.e., nucleotide differences) were detected, only one of which was associated with an amino acid substitution.

There are at least two indirect approaches to the question of the nonrandom distribution of RFLPs. The first emerges from the evidence for a relatively higher rate of evolutionary change in introns than in exons (reviewed in [65]). These substitutions should have been preceded by (transient) polymorphisms, in consequence of which trend polymorphisms might even now be expected to be more common in introns. The second approach emerges from a consideration of the types of nucleotide substitutions in the codons of exons. If during evolution nucleotide substitutions occurred at random in exons, the ratio of substitutions that result in an altered amino acid to those which do not should be about 3:1. A review of the substitutions that had occurred during the evolution of the avian and mammalian α and β hemoglobin genes by Czelusniak et al. [66] revealed ratios that averaged below 1. This figure is an approximation, depending on a reconstruction of events that cannot be documented, but the principle, of an excess of substitutions not reflected in a change in polypeptide composition, seems established. As in the case of substitutions in introns, these changes should have been preceded by nucleic acid polymorphisms that would not be reflected in protein variation.

It will be several years before a sufficient number of genes have been sequenced for a critical test with human material of the agreement between the revised estimate of expressed heterozygosity developed in this paper and the distribution (and individual frequencies) of the RFLPs. Even without this critical test, however, the emerging data are beginning to draw attention to another set of issues. An unevenly distributed pattern of RFLPs throughout the DNA reflects both the site of the mutational process and the action of selection. If mutation is more or less randomly spaced along the DNA, then an uneven distribution of RFLPs reflects the action of selection. The example of the thalassemias quoted above demonstrates that this selection is by no means restricted to the effects of mutations in exons. Thus, the level of selection one must postulate to achieve this nonrandomness raises even further questions than those already before us concerning the manner in which an animal with the low reproductive potential of *Homo sapiens* but

mutation rates approaching 1×10^{-5}/locus per generation accommodates such selection (cf. [67]). If mutation is not randomly distributed—and some such nonrandomness cannot be excluded by the present data—then, how nonrandom?

The difficulty in acquiring the data necessary for a decision is illustrated by a recent paper of Kazazian et al. [68], who have drawn attention to the unusual number of nucleotide alterations leading to the thalassemia trait (as well as hemoglobin S and C) located in the nine nucleotides comprising codons 6-8 of the β-globin gene and who have suggested that this region appears to be "particularly susceptible to mutations affecting nucleotide number" (p. 1031). While the suggestion may be correct, it must be borne in mind that the traits resulting from these substitutions have been the object of strong positive selection; the argument assumes that a similar mutation at any other location in the β-globin gene would be similarly selected for. An unbiased judgment requires detection of the mutational event immediately following its occurrence.

Assuming that pronounced nonrandomness will indeed characterize the distribution of RFLPs and nucleotide polymorphisms, we are, thus, once again, confronted, this time at the ultimate genetic level, by the need for any true understanding of the basis of our genetic architecture to unravel the complex interaction between patterns of mutation and patterns of selection. Given the well-known difficulties in studying selective differentials in modern human populations (which differentials might not correspond to those of the past), the study of the nature of mutation at the DNA level appears the more immediately promising approach. Such developments would dovetail nicely with current efforts to develop more efficient techniques for the study of mutation resulting in protein abnormalities [69]; the results of the latter technology will serve, of course, as a partial check on the results of the former.

ACKNOWLEDGMENTS

The financial support of the Department of Energy and The National Cancer Institute in the work reported from this laboratory is gratefully acknowledged.

JAMES V. NEEL

University of Michigan Medical School
Ann Arbor

REFERENCES

1. HARRIS H: *The Principles of Human Biochemical Genetics.* Amsterdam, North-Holland, 1970
2. MACCLUER J, quoted in SHAW CR: Electrophoretic variation in enzymes. *Science* 149:936–943, 1965
3. MARSHALL DR, BROWN ADH: The charge-state model of protein polymorphism in natural populations. *J Mol Evol* 6:149–163, 1975
4. O'FARRELL PH: High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 250:4007–4021, 1975
5. KLOSE J: Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissue. A novel approach to testing for induced point mutations in mammals. *Humangenetik* 26:231–243, 1975
6. SCHEEL G: Two-dimensional gel analysis of soluble proteins. Characterization of guinea pig exocrine pancreatic proteins. *J Mol Chem* 250:5375–5385, 1975

7. MERRIL CR, SWITZER RC, VAN KEUREN ML: Trace polypeptides in cellular extracts and human body fluids detected by two-dimensional electrophoresis and a highly sensitive silver stain. *Proc Natl Acad Sci USA* 76:4335–4339, 1979
8. SAMMONS DW, ADAMS LD, NISHIZAWA EE: A silver-based color development system for staining of polypeptides in polyacrylamide gels. *Electrophoresis* 2:135–141, 1981
9. WRAY W, BOULIKAS T, WRAY VP, ET AL.: Silver staining of proteins in polyacrylamide gels. *Anal Biochem* 118:197–203, 1981
10. STANBURY JB, WYNGAARDEN JB, FREDRICKSON DS, ET AL.: *The Metabolic Basis of Inherited Disease*, 5th ed. New York, McGraw-Hill, 1983
11. ANDERSON NG: Computer interfaced fast analyzer. *Science* 166:317–324, 1969
12. BURTIS CA, MAILEN JC, JOHNSON WF, ET AL.: Development of a miniature fast analyzer. *Clin Chem* 18:753–761, 1972
13. FIELEK S, MOHRENWEISER HW: Erythrocyte enzyme deficiencies assessed with a miniature centrifugal analyzer. *Clin Chem* 25:384–388, 1979
14. MOHRENWEISER HW: Frequency of enzyme deficiency variants in erythrocytes of newborn infants. *Proc Natl Acad Sci USA* 78:5046–5050, 1981
15. MOHRENWEISER HW, FIELEK S: Elevated frequency of carriers for triosephosphate isomerase deficiency in newborn infants. *Pediatr Res* 16:960–963, 1982
16. SATOH C, NEEL JV, YAMASHITA A, GORIKI K, FUJITA M, HAMILTON HB: The frequency among Japanese of heterozygotes for deficiency variants of 11 enzymes. *Am J Hum Genet* 35:656–674, 1983
17. MOHRENWEISER HW, NEEL JV: A "disproportion" between the frequency of rare electromorphs and enzyme deficiency variants in Amerindians. *Am J Hum Genet* 36:655–662, 1984
18. BERNSTEIN SC, THROCKMORTON LH, HUBBY JL: Still more genetic variability in natural populations. *Proc Natl Acad Sci USA* 70:3928–3931, 1973
19. SINGH RS, HUBBY JL, THROCKMORTON LH: The study of genetic variation by electrophoretic and heat denaturation techniques at the octanol dehydrogenase locus in members of the *Drosophila virilis* group. *Genetics* 80:637–650, 1975
20. SINGH RS, LEWONTIN RC, FELTON AA: Genetic heterogeneity within electrophoretic 'alleles' of xanthin dehydrogenase in *Drosophila pseudoobscura*. *Genetics* 84:609–629, 1976
21. TRIPPA G, LOVERRE A, CATAMO A: Thermostability studies for investigating non-electrophoretic polymorphic alleles in *Drosophila melanogaster*. *Nature* 260:42–44, 1976
22. SAMPSELL B: Isolation and genetic characterization of alcohol dehydrogenase thermostability variants occurring in natural populations of *Drosophila melanogaster*. *Biochem Genet* 15:971–988, 1977
23. COCHRANE BJ, RICHMOND RC: Studies of esterase-6 in *Drosophila melanogaster*. II. The genetics and frequency distributions of naturally occurring variants studied by electrophoretic and heat stability criteria. *Genetics* 93:461–478, 1979
24. AQUADRO CF, AVISE JC: An assessment of "hidden" heterogeneity within electromorphs at three enzyme loci in deer mice. *Genetics* 102:269–284, 1982
25. SATOH C, MOHRENWEISER HW: Genetic heterogeneity within an electrophoretic phenotype of phosphoglucose isomerase in a Japanese population. *Ann Hum Genet* 42:283–292, 1979
26. WURZINGER KH, MOHRENWEISER HW: Studies on the genetic and non-genetic (physiological) variation of human erythrocyte glutamic oxaloacetic transaminase. *Ann Hum Genet* 46:191–201, 1982
27. MOHRENWEISER HW, NEEL JV: Frequency of thermostability variants: estimation of total "rare" variant frequency in human populations. *Proc Natl Acad Sci USA* 78:5729–5733, 1981
28. SATOH C, NEEL JV, GORIKI K, ET AL.: Frequency of thermolability variants in Japanese. Submitted for publication

29. McConkey EH, Taylor BJ, Phan D: Human heterozygosity: a new estimate. *Proc Natl Acad Sci USA*:6500–6504, 1979
30. Walton KE, Styer D, Guenstein EI: Genetic polymorphism in normal human fibroblasts as analyzed by two-dimensional polyacrylamide gel electrophoresis. *J Mol Biol* 254:7951–7960, 1979
31. Smith CR, Racine RR, Langley CH: Lack of genic variation in the abundant proteins of human kidney. *Genetics* 96:967–974, 1981
32. Hamaguchi H, Ohta A, Mukai R, et al.: Genetic analysis of human lymphocyte proteins by two-dimensional gel electrophoresis: I. Detection of genetic variant polypeptides in PHA-stimulated peripheral blood lymphocytes. *Hum Genet* 59:215–220, 1981
33. Hamaguchi H, Yamada M, Noguchi A, et al.: Genetic analysis of human lymphocyte proteins by two-dimensional gel electrophoresis: II. Genetic polymorphism of lymphocyte cytosol 64K polypeptide. *Hum Genet* 60:176–180, 1982
34. Hamaguchi H, Yamada M, Shibasaki M, et al.: Genetic analysis of human lymphocyte proteins by two-dimensional gel electrophoresis. *Hum Genet* 62:142–147, 1982
35. Comings DE: Two-dimensional gel electrophoresis of human brain proteins. III. Genetic and non-genetic variations in 145 brains. *Clin Chem* 28:798–804, 1982
36. Klose J, Willers I, Singh S, et al.: Two-dimensional electrophoresis of soluble and structure-bound proteins from cultured human fibroblasts and hair root cells: qualitative and quantitative variation. *Hum Genet* 63:262–267, 1983
37. Goldman D, Merril CR: Human lymphocyte polymorphisms detected by quantitative two-dimensional electrophoresis. *Am J Hum Genet* 35:827–837, 1983
38. Wanner LA, Neel JV, Meisler MH: Separation of allelic variants by two-dimensional electrophoresis. *Am J Hum Genet* 34:209–215, 1982
39. Rosenblum BB, Neel JV, Hanash SM: Two-dimensional electrophoresis of plasma polypeptides reveals "high" heterozygosity indices. *Proc Natl Acad Sci USA* 80:5002–5006, 1983
40. Rosenblum BB, Neel JV, Hanash SM, Joseph JL, Yew N: Identification of genetic variants in erythrocyte lysate by two-dimensional gel electrophoresis. *Am J Hum Genet* 36:601–612, 1984
41. Hanash SM, Rosenblum BB, Neel JV, et al.: Genetic analysis of fifty-two platelet polypeptides detected in two-dimensional polyacrylamide gels. Manuscript in preparation
42. Singh RS, Coulthart MB: Genic variation in abundant soluble proteins of *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genetics* 102:437–453, 1982
43. Johnson GB: On the estimation of effective number of alleles from electrophoretic data. *Genetics* 78:771–776, 1974
44. Johnson GB: Evaluation of the stepwise mutation model of electrophoretic mobility: comparison of the gel sieving behavior of alleles at the esterase-5 locus of *Drosophila pseudoobscura*. *Genetics* 87:139–157, 1977
45. Ramshaw JAM, Coyne JA, Lewontin RC: The sensitivity of gel electrophoresis as a detector of genetic variation. *Genetics* 93:1019–1037, 1979
46. Fuerst PA, Ferrell RE: The stepwise mutation model: an experimental evaluation using hemoglobin variants. *Genetics* 94:185–201, 1980
47. Bonhomme F, Selander RR: The extent of allelic diversity underlying electrophoretic protein variation in the house mouse, in *Origins of Inbred Mice*, edited by Morse HC III, New York, Academic Press, 1978, pp 569–589
48. Ayala FJ: Genetic variation in natural populations: problem of electrophoretically cryptic alleles. *Proc Natl Acad Sci USA* 79:550–554, 1982
49. Scozzari R, Trippa G, Santachiara-Benerecetti AS, et al.: Further genetic heterogeneity of human red cell phosphoglucomutase-1: a non-electrophoretic polymorphism. *Ann Hum Genet* 45:313–322, 1981

50. TAKAHASHI N, NEEL JV, SATOH C, ET AL.: A phylogeny for the principal alleles of the human phosphoglucomutase-1 locus. *Proc Natl Acad Sci USA* 79:6636–6640, 1982
51. LEWIS WHP: Common polymorphism of peptidase A. Electrophoretic variants associated with quantitative variation of red cell levels. *Ann Hum Genet* 36:267–271, 1973
52. JENKINS T, LANE AB, NURSE GT, ET AL.: Red cell adenosine deaminase (ADA) polymorphism in Southern Africa, with special reference to ADA deficiency among the Kung. *Ann Hum Genet* 42:425–433, 1979
53. TANIS RJ, NEEL JV, DEARAUZ RT: Two more "private" polymorphisms of Amerindian tribes: LDH_B, GUA 1 and ACP_1 $B_{GUA\,1}$ in the Guaymi of Panama. *Am J Hum Genet* 29:419–430, 1977
54. MOHRENWEISER HW, NOVOTNY JE: An enzymatically inactive variant of human lactate dehydrogenase—LDH_B GUA-1. *Biochim Biophys Acta* 702:90–98, 1982
55. JEFFREYS AJ: DNA sequence variants in the G_γ-, A_γ-, δ- and β-globin genes of man. *Cell* 18:1–10, 1979
56. EWENS WJ, SPIELMAN RS, HARRIS H: Estimation of genetic variation at the DNA level from restriction endonuclease data. *Proc Natl Acad Sci USA* 78:3748–3750, 1981
57. KAZAZIAN HH JR, CHAKRAVARTI A, ORKIN SH, ET AL.: DNA polymorphisms in the human β-globin gene cluster, in *Evolution of Genes and Proteins,* edited by NEI M, KOEHN, RK, Sunderland, Mass., Sinauer, 1983, pp 137–146
58. MURRAY JC, DEMOPULOS CM, LAWN RN, ET AL.: Molecular genetics of human serum albumin: restriction enzyme fragment length polymorphisms and analbuminemia. *Proc Natl Acad Sci USA* 80:5951–5955, 1983
59. WYMAN AR, WHITE R: A highly polymorphic locus in human DNA. *Proc Natl Acad Sci USA* 77:6754–6758, 1980
60. BELL GI, KAREM JH, RUTTER WJ: A polymorphic DNA region adjacent to the 5′ end of the human insulin gene. *Proc Natl Acad Sci USA* 78:5759–5763, 1981
61. BELL GI, SELBY MJ, RUTTER WJ: Sequence of a highly polymorphic DNA segment in the 5′ flanking region of the human insulin gene. *Nature* 295:31–35, 1982
62. BECH-HANSEN NT, LINSLEY PS, COX DW: Restriction fragment length polymorphisms associated with immunoglobulin C & g genes reveal linkage disequilibrium and genomic organization. *Proc Natl Acad Sci USA* 80:6952–6956, 1983
63. PONCZ M, SCHWARTZ E, BALLANTINE M, ET AL.: Nucleotide sequence analysis of the δβ-globin gene region in humans. *J Biol Chem* 258:11599–11609, 1983
64. KREITMAN, M: Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster. Nature* 304:412–417, 1983
65. PERLER F, ESTRATIADIS A, LOMEDICO P, ET AL.: The evolution of genes: the chicken preproinsulin gene. *Cell* 20:555–566, 1980
66. CZELUSNIAK J, GOODMAN M, HEWETT-EMMETT D, ET AL.: Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. *Nature* 298:297–300, 1982
67. NEEL JV: The wonder of our presence here: a commentary on the evolution and maintenance of human diversity. *Perspect Biol Med* 25:518–558, 1982
68. KAZAZIAN HH JR, ORKIN SH, BOEHM CD, SEXTON JP, ANTONARAKIS SE: β-Thalassemia due to a deletion of the nucleotide which is substituted in the β^S-globin gene. *Am J Hum Genet* 35:1028–1033, 1983
69. NEEL JV, ROSENBLUM BB, SING CF, ET AL.: Adapting two-dimensional gel electrophoresis to the study of human germline mutation rates, in *Methods and Applications of Two-Dimensional Gel Electrophoresis of Proteins,* edited by CELIS JE, BRAVO R, New York, Academic Press, 1984, pp 259–306