Proc. Natl. Acad. Sci. USA
Vol. 76, No. 12, pp. 6500–6504, December 1979
Genetics

Human heterozygosity: A new estimate

(fibroblast polypeptides/two-dimensional gel electrophoresis/double-label autoradiography)

EDWIN H. MCCONKEY, BEVERLEY J. TAYLOR*, AND DUC PHAN†

Department of Molecular Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309

Communicated by Frank H. Ruddle, September 4, 1979

ABSTRACT **Several hundred polypeptides from four human diploid fibroblast cell lines were compared by high-resolution two-dimensional polyacrylamide gel electrophoresis and double-label autoradiography under conditions where allelic products that differ by a single charged amino acid would be distinguished. The average heterozygosity represented by this set of gene products appears to be less than 1% for changes involving charged amino acids.**

The nature and extent of genetic diversity among humans is a matter of great theoretical and practical significance. Our current views on this subject depend heavily on the summaries by Harris and colleagues (1, 2). Of the 87 enzymes that had been surveyed in Europeans up to 1977, 24 were polymorphic; the average heterozygosity for all 87 loci was 6.3%. Because only electrophoretically separable enzyme variants were detected in these surveys and because more than two-thirds of possible single-base changes in a gene should not change the net charge of the protein, it can be argued that total average heterozygosity must be at least 20%. Similar values of average heterozygosity have been calculated for other mammals (3), considerably higher estimates have been made for invertebrates (3), and recent studies on *Drosophila* indicate that heterozygosity at some loci may be rampant (e.g., ref. 4).

The development of high-resolution two-dimensional gel electrophoresis by O'Farrell (5) made it possible to separate as many as 1000 polypeptides on a single slab gel, and there is abundant evidence that polypeptides differing by a single charged amino acid can be clearly resolved (6–8). These facts suggested to one of us several years ago that an attempt to measure heterozygosity by comparison of O'Farrell gel patterns of polypeptides from cells of different individuals would be fruitful. Side-by-side comparison of complex patterns, however, is fatiguing and inaccurate. The problem of finding a few spots unique to one two-dimensional gel pattern among hundreds of spots common to both patterns in any pairwise comparison has been solved by the development of double-label autoradiography (9). This method permits the simultaneous electrophoresis of ^{3}H-labeled and ^{14}C-labeled protein mixtures on the same gel, followed by a two-step autoradiographic procedure that identifies the polypeptides that are unique to the ^{3}H-labeled preparation. We have used this technique to compare approximately 400 fibroblast polypeptides from several individuals. We find that the heterozygosity of the genes represented by this set of polypeptides is significantly less than one would have expected from the enzyme surveys (1, 2).

MATERIALS AND METHODS

Cell cultures were obtained from the Institute for Medical Research (Camden, NJ) (1MR90 and GM738) or the American Type Culture Collection (Rockville, MD) (MRC5 and WI38).

All cells were grown in F12 medium with 10% calf serum. Labeling was carried out for 12–16 hr in leucine-free F12 medium supplemented with [^{3}H]leucine (40 μCi/ml, 5 Ci/mmol) or with [^{14}C]leucine (2.5 μCi/ml, 300 Ci/mol) (1 Ci = 3.7×10^{10} becquerels). Cells were harvested by scraping with a rubber policeman, rinsed with 0.14 M NaCl/0.01 M sodium phosphate, pH 7.0, and pelleted in a conical tube which was then carefully wiped dry. Fresh or frozen cell pellets were lysed in freshly made 10 M urea, with vigorous mixing on a Vortex. The protein concentration of a 5- to 10-μl sample of homogenate was measured with the Bio-Rad Reagent (bulletin 1069), and 1% sodium dodecyl sulfate/10 M urea was added so that the final mass ratio of detergent to protein was 1.5:1. If necessary, 10 M urea was added so that the final protein concentration did not exceed 1 mg/ml. The homogenate was then made 0.1 M in acetic acid (final pH 4.0–4.2) and 5 mM in $ZnSO_4$; S1 nuclease (Sigma or P-L Biochemicals) was added at 2 units of enzyme per μg of homogenate protein, and the mixture was incubated at 45°C for 5 min, then chilled. An equal volume of 1 M acetic acid/1% 2-mercaptoethanol was added, followed by 10 homogenate volumes of cold acetone, and the proteins were allowed to precipitate at −20°C overnight. The precipitate was collected by centrifugation at 10,000 × *g* for 15 min.

Samples were dissolved in O'Farrell's buffer (5) and centrifuged for 10 min at 10,000 × *g* at 25°C just before electrofocusing gels were loaded. Two-dimensional electrophoresis was done according to O'Farrell (5), except that the second-dimension gels were only 0.4 mm thick (1/64 inch). Slab gels were impregnated with scintillator (10), dried, and subjected to two-step autoradiography (9).

Two-dimensional gel patterns are presented with the acidic proteins on the left and the basic proteins on the right. Although this is the reverse of the original orientation (5), it allows the pH scale to go from low numbers on the left to high numbers on the right and thus follows the traditional Cartesian coordinate system (11). This orientation also corresponds to the conventional orientation of patterns from two-dimensional gel systems for basic proteins (e.g., refs. 12 and 13), where the most basic proteins have moved farthest to the right. We have also adopted one of the arbitrary pI standards suggested by Anderson and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

* Present address: Division of Biology, California Institute of Technology, Pasadena, CA 91125.
† Present address: Stanford University School of Medicine, M105, Stanford, CA 94305.

Hickman (11) in order to avoid the ambiguity inherent in the estimation of pH in a solution containing 9.5 M urea. Creatine phosphokinase (Worthington) was boiled in 10 M urea for 1, 2, 3, 4, or 6 min to create populations of molecules with various degrees of carbamoylation. Portions of each mixture were pooled to provide approximately equal amounts of each charge isomer from 1 to about 25 carbamoylated lysines; 10–12 μg of the creatine phosphokinase mixture can be included with any sample during electrophoresis, and the position of polypeptides from the sample can be referred to in the creatine phosphokinase charge train. Molecular weights are based upon the following standards: β-galactosidase, 130,000; phosphorylase *a*, 94,000; bovine serum albumin, 68,000; actin, 42,000; and carbonic anhydrase, 29,000 (14). The molecular weight scale in the figures is most accurate for the central portion of each pattern; various degrees of distortion occur at the edges.

RESULTS

Three human diploid lung fibroblast lines (IMR90, MRC5, and WI38) and HeLa cells (heteroploid, derived from cervical carcinoma) were labeled with [^{3}H]leucine or [^{14}C]leucine, all possible pairwise combinations were made, the mixtures were subjected to two-dimensional electrophoresis (5), and the components that were unique to the ^{3}H-labeled preparation in each mixture were identified by double-label autoradiography (9). The human skin fibroblast line, GM738, was similarly compared with the lung fibroblast line, MRC5. Double-label autoradiography is a two-step autoradiographic procedure: the first film yields a fluorogram, which records polypeptides labeled with either isotope; the second film records only the ^{14}C-labeled polypeptide (9). Accordingly, the pattern on the fluorogram will include some spots that are missing from the ^{14}C autoradiogram if there are any polypeptides unique to the cells that were labeled with ^{3}H. The power of the technique arises from the fact that two patterns are present on the same gel. Thus, all the spots common to both patterns serve as a vast array of reference points by which the presence of a few spots unique to the ^{3}H-labeled pattern can be detected with great confidence.

We included HeLa cells in our survey as a type of control; HeLa cells originated from an adenocarcinoma of the cervix (15). Moreover, their heteroploid nature, their growth characteristics, and the fact that the donor was an American Black, should all increase the probability that the O'Farrell gel patterns of their polypeptides would differ from the polypeptide patterns of diploid fibroblasts whose donors were Caucasian. This expectation was fulfilled; the double-label autoradiographic comparisons of HeLa polypeptides with diploid fibroblast polypeptides all showed that 2–3% of the polypeptides differed qualitatively (Table 1). Fig. 1*A* is an example of the comparisons of HeLa cells and fibroblasts; several major spots unique to the HeLa cell preparation are obvious, and many quantitative differences can be seen (black spots with white halos).

The results of the comparisons of fibroblasts with other fibroblasts were strikingly different. In one series of experiments, we chose to compare cells that were as similar as possible because our early expectation was that we would find a large number of differences due to heterozygosity (1, 2). We chose the diploid fibroblast lines IMR90, MRC5, and WI38, all of which came from lung tissue of Caucasian fetuses. It is unlikely that the donors of these cell lines were related. The cultures were established in different laboratories at different times and, in addition, MRC5 cells are male whereas IMR90 and WI38 cells are female. All cells were labeled at approximately the same "age" (30–35 passages in culture) and approximately the same stage of growth (about one cell doubling prior to confluency). The results were surprising; fewer than 0.5% of 400 spots on the O'Farrell gel patterns differed, regardless of the pair of cells being compared or the labeling protocol (Table 1).

Table 1. Comparison of two-dimensional gel patterns of polypeptides from various human cell lines

Cells compared	Total spots	Spots containing ^{3}H only
^{3}H-HeLa/^{14}C-IMR90	500	11
^{3}H-IMR90/^{14}C-HeLa	530	12
^{3}H-HeLa/^{14}C-MRC5	490	11
^{3}H-MRC5/^{14}C-HeLa	600	14
^{3}H-HeLa/^{14}C-WI38	530	13
^{3}H-WI38/^{14}C-HeLa	570	14
^{3}H-IMR90/^{14}C-MRC5	390	1
^{3}H-MRC5/^{14}C-IMR90	380	0
^{3}H-IMR90/^{14}C-WI38	400	2
^{3}H-WI38/^{14}C-IMR90	410	0
^{3}H-MRC5/^{14}C-WI38	420	2
^{3}H-WI38/^{14}C-MRC5	390	1
^{3}H-GM738/^{14}C-MRC5	580	5
^{3}H-MRC5/^{14}C-GM738	570	2

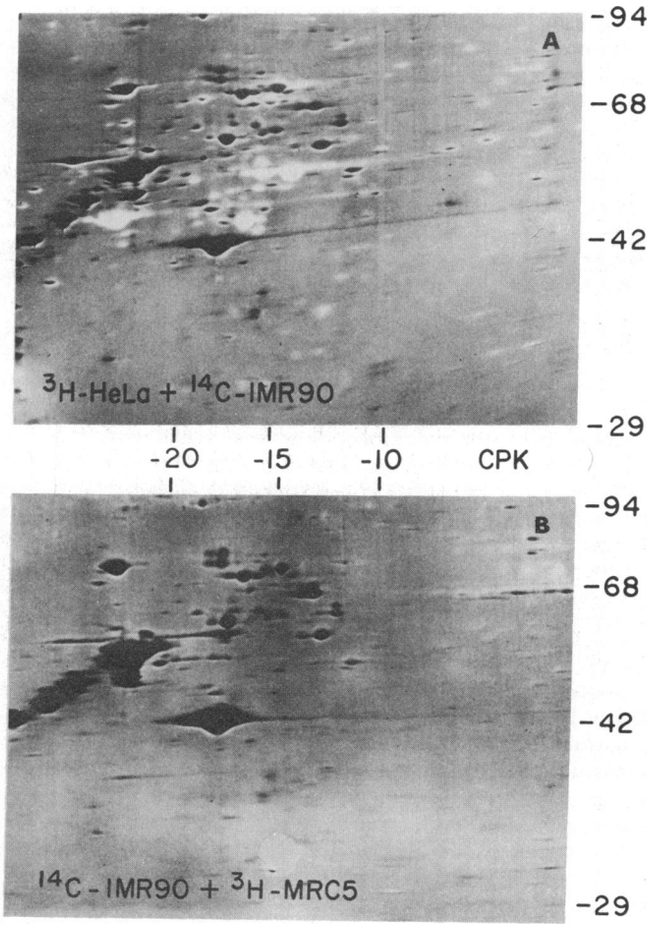

FIG. 1. Comparison of polypeptides from cells of human origin by double-label autoradiography (9). Extracts of total cellular protein were prepared, then fractionated by two-dimensional polyacrylamide gel electrophoresis (5). The horizontal scale is an internal isoelectric point reference series (see *Materials and Methods* and ref. 11); the vertical scale is in units of molecular weight $\times 10^{-3}$. CPK, creatine phosphokinase. (*A*) ^{3}H-Hela and ^{14}C-IMR90; (*B*) ^{14}C-IMR90 and ^{3}H-MRC5. Both panels show overlays, consisting of a ^{14}C autoradiogram on top of a negative of the corresponding ^{3}H and ^{14}C fluorogram.

Fig. 1*B* shows an example of the data; in this case, there were no differences in the 380 spots that were scored.

Fig. 2 compares the O'Farrell gel patterns of total cellular polypeptides from a skin fibroblast line (GM738) and a lung fibroblast line (MRC5) by double-label autoradiography. Skin and lung fibroblasts may not be metabolically identical (16), and their O'Farrell gel patterns reveal many more quantitative differences than were seen when any two lung fibroblast lines were compared (Fig. 1*B*). A few spots where the ^{14}C is much less than the ^{3}H are marked with white-edged arrows in Fig. 2; Fig. 2*A Inset* reproduces part of the ^{14}C autoradiogram (i.e., the upper film in the overlay shown in the main figure), so that some of these faint ^{14}C spots can be seen more clearly. The arrowheads in Fig. 2 identify polypeptides that appear to be exclusively labeled with ^{3}H; some of these may be examples of extreme quantitative differences. Spots that are not marked by arrowheads definitely contain ^{14}C, which is unambiguously present on the original autoradiograms but not necessarily visible in the figures. The striking fact is that there are so few qualitative differences in the polypeptides synthesized by these two cell lines (Table 1).

The most interesting spots in Fig. 2 are in the upper right corner of each pattern. In Fig. 2*A*, the arrowhead marks a small white spot just below a large, fuzzy spot that is primarily, but not exclusively, ^{3}H labeled. In Fig. 2*B*, the arrowhead marks a different faint white spot that is also below, but a few millimeters to the left of the large fuzzy spot. The spot that is white in Fig. 2*A* is black in Fig. 2*B*; the spot that is white in Fig. 2*B* is black in Fig. 2*A*. This is the behavior expected of two allelic gene products, each present in homozygous state in one of the two cell types being compared. Moreover, the two polypeptides represented by these spots have the same molecular weight and are separated in the isofocusing dimension by a distance that is consistent with a single charge difference. This is the only candidate case of a polymorphism involving two homozygotes that we have seen in the patterns in which IMR90, MRC5, WI38, and GM738 cell proteins were compared.

DISCUSSION

Double-label autoradiography allows easy visual detection of polypeptides that are uniquely labeled with ^{3}H in mixtures of ^{3}H-labeled and ^{14}C-labeled (or ^{35}S-labeled) proteins. In all of the experiments reported here, the input ratio of ^{3}H to ^{14}C was adjusted so that the sensitivities for fluorographic detection of both isotopes were equal (9). In any pairwise comparison, polypeptides that are unique to the ^{3}H-labeled cells can arise in several ways, including (*i*) the expression of alleles present in cell line A but absent in cell line B, (*ii*) differences in gene expression due to type and age of cell line, stage of growth or cell cycle, etc., and (*iii*) selective losses during preparative stages prior to mixing the two labeled preparations. In the comparison of IMR90, MRC5, and WI38 cells, we controlled for the second category of differences as much as possible by using three diploid fibroblast lines that were all derived from lung tissue of Caucasian fetuses, by labeling cells of similar ages (30–35 passages in culture), and by use of subconfluent cultures that appeared to be at virtually identical densities. The skin fibroblast cells (GM738) were at passage 10 when they were labeled; the MRC5 cells with which they were compared were at passage 22. This could have contributed to the differences observed. Selective losses during sample preparation could also lead to spurious differences on the autoradiograms. We think this does not occur commonly; we have never found any qualitative differences when we compare ^{3}H-labeled and ^{14}C-labeled proteins from the same cell line (not shown).

There are also factors that can lead to an underestimate of the differences between two populations of proteins compared by our method. Evidently, we cannot detect differences that do not change the isoelectric point or size of a protein. However, there is strong evidence that all charge changes *will* be detected, at least for polypeptides with isoelectric points in the range studied here (roughly pH 5–7). Carbamoylation of lysine residues produces a series of well-resolved spots from any polypeptide, each of which must differ from its neighbors by a unit charge (6, 11). Translational errors, which presumably substituted glutamine for histidine and lysine for asparagine, also led to well-resolved charge isomers of actin and other proteins (8). There is no reason to doubt that changes to or from arginine, aspartic acid, or glutamic acid would be equally well resolved, at least for proteins with pIs in the midrange. Cryptic charge changes should not occur; all polypeptides are extensively denatured in the 9.5 M urea used for electrofocusing in the O'Farrell system. Finally, human transferrin variants that appear to be due to simple mendelian mutations have been resolved on O'Farrell gels (7), although proof that these variants are due to single amino acid substitutions is lacking.

The method of double-label autoradiography will also lead to a small underestimate of heterozygosity. This is diagrammed in Table 2, which shows two extreme situations. When the maximum possible heterozygosity for a two-allele system (p = 0.5, q = 0.5) occurs at a given locus, double-label autoradiography will lead to a calculated heterozygosity that is only 75% of the true value; when the heterozygosity at a given locus is the minimum that can classify a locus as "polymorphic" (p = 0.99, q = 0.01), double-label autoradiography estimates the heterozygosity accurately to the third decimal place. For intermediate levels of heterozygosity, intermediate underestimates will occur. The underestimates arise because some pairwise comparisons will involve two heterozygotes at a given locus instead of two different homozygotes or one homozygote and one heterozygote; the former, being identical, will produce no unique ^{3}H-labeled spots on the autoradiograms.

It seems, therefore, that technical factors cannot account for our finding that less than 1% of the gene loci represented by the

Table 2. Detection of polymorphisms by double-label autoradiography

		^{3}H				^{3}H		
		AA	*Aa*	*aa*		*AA*	*Aa*	*aa*
	AA	0.0625	<u>0.1250</u>	<u>0.0625</u>	*AA*	0.9606	<u>0.0194</u>	<u>0.0001</u>
^{14}C	*Aa*	0.1250	0.2500	0.1250	*Aa*	0.0194	0.0004	<0.0001
	aa	<u>0.0625</u>	<u>0.1250</u>	0.0625	*aa*	<u>0.0001</u>	<u><0.0001</u>	<0.0001
		p = 0.5, q = 0.5				p = 0.99, q = 0.01		
		True heterozygosity = $2pq$ = 0.5				True heterozygosity = $2pq$ = 0.0198		
		Apparent heterozygosity = 0.375				Apparent heterozygosity = 0.0196		

The numbers are the expected frequencies for all possible pairs of genotypes at a single locus with two alleles. Underlined numbers designate the genotype pairs for which double-label autoradiography would detect a gene product unique to the ^{3}H-labeled protein preparation. Allele frequencies, p and q, apply to alleles *A* and *a*, respectively.

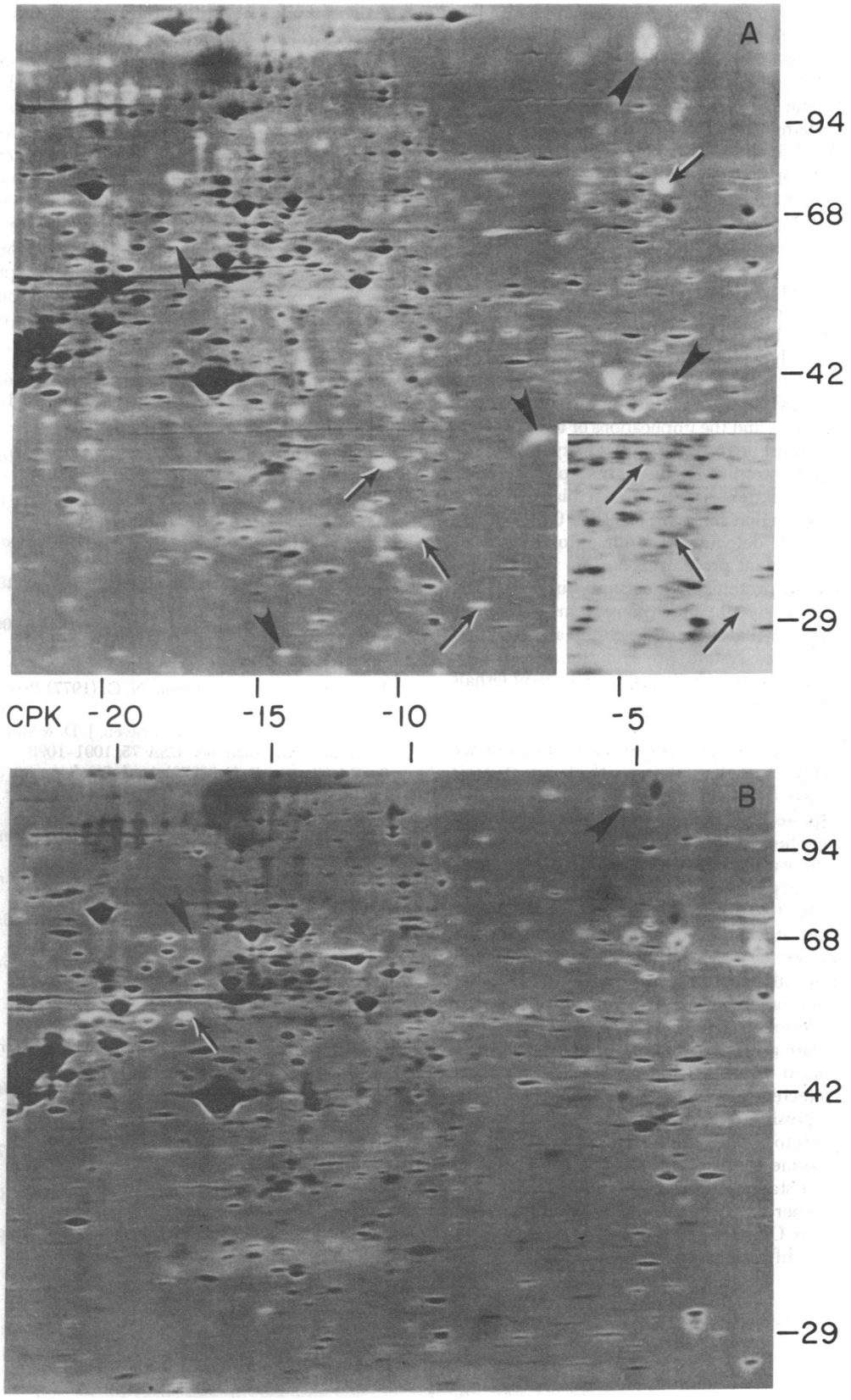

FIG. 2. Comparison of polypeptides from lung fibroblasts (MRC5) and skin fibroblasts (GM738) from different individuals by double-label autoradiography (9). Extracts of total cellular protein were prepared, then fractionated by two-dimensional polyacrylamide gel electrophoresis (5). The horizontal scale is an internal isoelectric point reference series (see *Materials and Methods* and ref. 11); the vertical scale is in units of molecular weight $\times 10^{-3}$. CPK, creatine phosphokinase. (*A*) The main section is an overlay, consisting of the ^{14}C-MRC5 autoradiogram on top of a negative of the ^{3}H-GM738 and ^{14}C-MRC5 fluorogram; (*Inset*) part of the ^{14}C autoradiogram. (*B*) As in *A*, except that the ^{3}H is from MRC5 cells and the ^{14}C is from GM738 cells.

400–500 polypeptides in our survey appear to be polymorphic, whereas extrapolation of enzyme survey data (1, 2) led us to expect approximately 6% differences. To be more precise, we would expect every 106 spots on the O'Farrell gels to represent the products of 100 gene loci, if the average heterozygosity per locus were 6%, for changes in the DNA that change the net charge on the protein. It is also true that one gene product may lead to more than one spot on a two-dimensional gel due to posttranslational modification. However, this would not affect the expected frequency of differences unless the products of polymorphic genes were much less likely or much more likely to be modified than the products of monomorphic loci. We note that a similar comparison of human fibroblast polypeptides by two-dimensional gel electrophoresis, with similar results, has been briefly reported (17), and a recent study of a smaller sample of *Drosophila* polypeptides (18) also reveals significantly less polymorphism than enzyme surveys.

What is the biological significance of the discrepancy between the enzyme surveys and the implications of the O'Farrell gel patterns? First, it must be recognized that enzyme surveys and two-dimensional electrophoresis of total cell proteins have different biases. In the latter case, the major bias is toward abundance; i.e., proteins that make up less than 0.01% of the total were probably not detectable on our autoradiograms. Many of the most abundant proteins in our gel patterns are undoubtedly structural proteins, like actin, tubulin, etc., although histones and nearly all ribosomal proteins are not included. We think it likely that structural proteins and enzymes that are part of multimeric systems are evolutionarily conservative. Many mutations at those loci must be dominant lethals because the defective proteins that they produce inactivate an entire structure (e.g., microtubules, fatty acid synthase, protein synthesis initiation complexes, etc.) regardless of the presence of the normal gene product. If proteins with stringent steric requirements are a significant portion of total cellular proteins, then overall average heterozygosity would be expected to be less than the average heterozygosity of soluble enzymes (19).

A different bias may apply to the data on enzyme polymorphisms (1, 2). Most large-scale enzyme surveys are done on enzymes that are easy to assay, and easy assays often depend upon synthetic substrates. Johnson (20) has pointed out that enzymes that accept a wide variety of substrates are much more polymorphic than specific-substrate enzymes. The table in ref. 1 includes several esterases, peptidases, and phosphatases that belong to the variable-substrate class of enzymes. This may have led to an overestimate of average heterozygosity per locus. It must also be borne in mind that the molecular basis of the polymorphisms detected in enzyme surveys has rarely been determined. It is possible for changes that do not involve charged amino acids to alter the conformation of a protein so that the native enzyme is separable from the product of the standard allele on a starch gel or acrylamide gel of a given composition (21), whereas the denatured polypeptides would not be separable on O'Farrell gels. Finally, some types of posttranslational modifications may alter the conformation of a native protein (21) without affecting the net charge of the denatured polypeptide.

Other arguments could be advanced, but none would be conclusive. The technique that we have used for estimating genetic diversity in humans appears to be theoretically sound, but it needs to be applied much more extensively before any major generalizations are made. It will be particularly informative to identify the O'Farrell gel location of some of the enzymes that are known to be polymorphic (1, 2). However, it may ultimately be necessary to determine the molecular basis of many of these polymorphisms. Whatever the reason for the low frequency of differences among polypeptides from different fibroblasts examined by two-dimensional gel electrophoresis may be, it is clear that current views on average heterozygosity need to be critically re-examined.

Note Added in Proof. Walton *et al.* (22) have reported on a similar study of five human fibroblast lines. Their results and conclusions agree with ours.

1. Harris, H. & Hopkinson, D. A. (1972) *Ann. Human Genet.* **36,** 9–20.
2. Harris, H., Hopkinson, D. A. & Edwards, Y. H. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 698–701.
3. Selander, R. K. & Kaufman, D. W. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 1875–1877.
4. Singh, R. S., Lewontin, R. C. & Felton, A. A. (1976) *Genetics* **84,** 609–629.
5. O'Farrell, P. H. (1975) *J. Biol. Chem.* **250,** 4007–4021.
6. Steinberg, R. A., O'Farrell, P. H., Friedrich, U. & Coffino, P. (1977) *Cell* **10,** 381–391.
7. Anderson, L. & Anderson, N. G. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5421–5425.
8. Parker, J., Pollard, J. W., Friesen, J. D. & Stanners, C. P. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1091–1095.
9. McConkey, E. H. (1979) *Anal. Biochem.* **96,** 39–44.
10. Bonner, W. M. & Laskey, R. A. (1974) *Eur. J. Biochem.* **46,** 83–88.
11. Anderson, N. L. & Hickman, B. J. (1979) *Anal. Biochem.* **93,** 312–320.
12. Kaltschmidt, E. & Wittmann, H. G. (1970) *Anal. Biochem.* **36,** 401–412.
13. Mets, L. J. & Bogorad, L. L. (1974). *Anal. Biochem.* **57,** 200–210.
14. Weber, K. & Osborn, M. (1969) *J. Biol. Chem.* **244,** 4406–4412.
15. Jones, H. W., McKusick, V. A., Harper, P. S. & Wuu, K. D. (1971) *Obstet. Gynecol.* **38,** 945–949.
16. Schneider, E. L., Mitsui, Y., Au, K. S. & Shorr, S. S. (1977) *Exp. Cell Res.* **108,** 1–6.
17. Walton, K. & Gruenstein, E. (1977) *J. Cell Biol.* **75,** 388a.
18. Brown, A. J. L. & Langley, C. H. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 2381–2384.
19. Berger, E. M. & Weber, L. (1974) *Genetics* **78,** 1173–1183.
20. Johnson, G. B. (1974) *Science* **184,** 28–37.
21. Johnson, G. B. (1979) *Prog. Nucleic Acid Res. Mol. Biol.* **22,** 293–326.
22. Walton, K. E., Styer, D. & Gruenstein, E. I. (1979) *J. Biol. Chem.* **254,** 7951–7960.