Frequency and distribution of rare electrophoretic mobility variants in a population of human newborns in Ann Arbor, Michigan

H. W. MOHRENWEISER*, K. H. WURZINGER AND J. V. NEEL

Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109

SUMMARY

We have summarized the frequency and distribution of the rare variants encountered during the screening of 258815 allele products, the products of 51 different loci, in 3242 predominantly Caucasian (88%) newborns. Seventy-nine different rare variants, representing 187 occurrences, were identified. Almost 60% (46 of 79) of the rare variants occurred as singletons while another 20% were seen in two unrelated individuals. No rare variants were detected at 18 loci while no variants, either rare or polymorphic, were detected at 14 loci. More rare variants were identified at loci that were classified as polymorphic and also at loci where the gene products exist as a monomer. A positive relationship was observed between variant frequency, either classes or copies, and subunit molecular mass.

INTRODUCTION

Each new study of the level of genetic variation among individuals/populations highlights the diversity in the human gene pool. The variation observed is of two general classes. The first is variation which is polymorphic, that is, a few variants at a locus are widely distributed among many individuals in a population. In contrast, rare variants are, by definition, detected among a relatively few individuals in the population, the usual definition being that a phenotype with a frequency of less than 2% is considered a rare variant. The latter are more likely to represent the spectrum of relatively recent mutational events than are the polymorphisms, which for a variety of reasons represent the few survivors from among many mutational events occurring in the more remote past.

The number of studies that have surveyed, by electrophoretic techniques, a wide range of loci, including loci expected to be monomorphic, has been quite limited (Harris *et al.* 1974; Neel *et al.* 1978; Neel *et al.* 1980). Within the context of a pilot study concerned with the feasibility of using electrophoretic techniques to detect germinal mutations in a human population (Neel *et al.* 1987) we have accumulated, during the last 10 years, an extensive data base useful for an analysis of rare variant frequency at 51 loci. The loci studied were selected because of the availability of electrophoretic techniques permitting high resolution of their gene products rather than on the bases of anticipated variation, biological function or other restrictive criteria. This data base is especially suited for addressing questions regarding rare variant frequencies in a human population and the extent of variation in rare variant frequency among loci of the human genome.

* Current address: Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

DATA BASE

Cord blood samples were collected from 3456 newborns and their parents at Women's Hospital at the University of Michigan during the period of this study. To the extent possible, samples were collected from all newborns where informed consent and a blood sample from both parents could be obtained. In our previous publication on this series (Neel *et al.* 1987), one of each pair of identical twins was excluded but otherwise all siblings were allowed to remain in the series as each is an independent test of mutation. This resulted in 277747 locus tests. Seven percent of the samples were from siblings, including non-identical twins, as determined by examination of mothers' hospital identification numbers. For the present series, to eliminate non-binomial variation in statistical tests, only the first child in a family has been included in this data base, which reduces the number of newborns in the series to 3242 and the number of locus tests to 258815. The population is 88% Caucasian and 7% Black while the remaining 5% of the population includes individuals of mixed or unknown racial origin and a very limited number of individuals with other racial backgrounds. The number of unrelated newborns studied ranges from 2827 for PGM1 to 632 for BG7S in the Caucasian sample and from 41 for BG7S to 242 for several loci in the Black sample. Differences in the number of determinations (individuals studied) reflect the sequential addition of new loci as electrophoretic techniques were developed.

A question arises as to the definition of a 'rare variant' when one is screening representatives of different ethnic groups. This is especially a consideration when, as in this series, some 20–30% of the gene pool of the American Black group is known to be derived from the Caucasoid group, with a more limited gene flow in the other direction (summary in Reed, 1969). In the present paper, we have elected not to count as a rare variant in either group, any allele with a frequency greater than 0·01 in either of the two groups. This convention, however, does influence the direct comparison of our average variant frequencies with those of other investigators, who also face the problem of ethnic admixture, albeit usually in a less pronounced form. Accordingly, the impact of this convention on the estimated variant frequency will also be discussed.

METHODS

Electrophoretic methods for most of the proteins studied have been previously described (Neel *et al.* 1980; Barrantes *et al.* 1982; Naidu *et al.* 1985; Long *et al.* 1986). The electrophoretic system for inosine triphosphatase (ITPA), enolase (ENO1), diaphorase (DIA) and acetaldehyde dehydrogenase (ALDH) was the FUM/GPT buffer system described by Long *et al.* (1986). ITPA, DIA and ENO1 were stained as described by Harris & Hopkinson (1976) while ALDH was stained as described by Pearse (1972). Prealbumin (PALB) was typed on a 7% acrylamide gel with a gel buffer of 1·25 M-Tris-HCl, pH 9·1, and a tank buffer of 0·04 M-Tris glycine, pH 8·5. PALB was localized by staining with Coomassie Blue. 7-S beta globulin (BG7S) was typed as described by Haupt & Böhm (1977). The method for apolipoprotein A (APOA1) was modified from Haupt & Böhm (1977). The first step was electrophoresis of plasma, following treatment with Triton X-100 and cetyltrimethylammonium bromide, on agarose gels with a Tris glycine buffer. The second step involved isoelectric focusing (pH 4·0–6·5 gradient) of the proteins of the agarose gel section containing the APOA1 on a polyacrylamide gel containing 8 M urea. The

APOA1 bands were visualized by staining with Coomassie Blue. Sample storage and preparation of haemolysates were as described by Neel *et al.* (1980).

RESULTS

The data on number of determinations, classes of rare variants, total number of copies of rare variants and number of copies of each class of rare variants for Caucasians and Blacks are in Table 1*a* and 1*b*, respectively. Tables 1*a*, *b* also present the allele counts (multiply determinations by two except for sex linked loci) for the polymorphic alleles in the major ethnic groups. The genetic nature of all rare and low frequency polymorphic variants was determined by establishing the occurrence of the variant in one of the child's parents. Although the electrophoretic mobilities of the variants have been rigorously compared, using standard techniques, to establish the minimum number of electrophoretic classes, no systematic attempt has been made to make direct comparison of these variants with those reported by others. Therefore, the variants have not been assigned names unless they have been the subject of other publications concerned with their biochemical characteristics. These would include variants at the following loci: *CRPL*, Mohrenweiser & Decker, 1982; *GOTS*, Wurzinger & Mohrenweiser, 1982; *TPI*, Asakawa & Mohrenweiser, 1982; *TF*, Fujita *et al.* 1985; *GPI*, Mohrenweiser *et al.* 1987; *G6PD*, Mohrenweiser & Fielek, in preparation. The allele frequencies at the polymorphic loci are very similar to the previously published frequencies for these two ethnic groups in the United States (Mourant *et al.* 1976; Silver, 1982; Tills *et al.* 1983).

In the Caucasian sample, 64 different products of rare alleles, distributed among 154 individuals, were identified from a total of 114560 determinations (1·3 variants/1000 determinations). Seventeen different products of rare alleles, occurring in 27 copies (individuals), were detected in 9634 determinations among the Black sample (2·8 variants/1000 determinations). The frequency of different rare variants is higher in the Black sample (1·8 variant classes/1000 determinations) than in the Caucasian sample (0·6 variant classes/1000 determinations) ($\chi^2 = 12{\cdot}9$, D.F. = 1, $P < 0{\cdot}01$). In the total (pooled) sample, 79 different rare variants occurred in a total of 187 copies in the screening of 258815 alleles (1·4 variants/1000 determinations). If one does not exclude those variants which were found to be polymorphic in the other ethnic group, then the variant frequencies are as follows: in the Caucasian sample, there were 72 different rare allele products distributed among 255 individuals, resulting in a frequency of 0·6 variant classes/1000 determinations; in the Black population, there were 18 different rare allele products distributed among 28 individuals with a resulting frequency of 1·8 variant classes/1000 determinations. Thus, the higher number of different rare alleles in Blacks is independent of the method of defining rare variants. By contrast, the number of variants encountered becomes 2·2/1000 determinations for Caucasians and 2·9/1000 determinations for Blacks. Eighty percent of the increase in number of variants in the Caucasian group is contributed by the CRPL*MI and PEPD*2 alleles, which along with the TF*DCHI variant, account for 45% of all of the rare variants in the Caucasian group.

Several of the alleles designated in Tables 1*a*, *b* have been subdivided by utilizing additional electrophoretic or thermostability techniques (Wurzinger & Mohrenweiser, 1982; Fujita *et al.* 1985). Therefore, the estimate of 79 different rare alleles is an absolute minimum estimate as this estimate is based upon variant classes identified by the original electrophoretic procedures

Table 1*a*. *Frequency of rare electrophoretic mobility variants in a sample of unrelated Caucasian newborns*

				Polymorphic alleles			Rare alleles[b]		
Protein	EC number	Locus	N^a	Allele/number of copies			Classes	Copies	Copies/class
Acetaldehyde dehydrogenase	1.2.1.3	ALDH	1122	ALDH*1/2240		—	0	—	—
Acid phosphatase 1	3.1.3.2	ACCP 1	2825	ACP*B/3429	ACP*A/1953	ACP*C/266	2	2	1, 1
Adenosine deaminase	3.5.4.4	ADA	2807	ADA*1/5352	ADA*2/262		0	—	—
Adenylate kinase	2.7.4.3	AK1	2802	AK1*1/5374	AK1*1/228		1	2	2
Albumin		ALB	2542	ALB*1/5081		—	2	2	1, 1
Apolipoprotein A1		APOA1	965	APOA1*1/1929		—	1	1	1
Carbonic anhydrase-1	4.2.1.1	CA1	671	CA1*1/1342		—	0	—	—
Carbonic anhydrase-2	4.2.1.1	CA2	883	CA2*1/1765	CA2*2/1		0	—	—
Ceruloplasmin		CRPL	2737	CRPL*B/5419	CRPL*A/0 CRPL*CB/7	CRPL*MI/46	2	2	1, 1
Diaphorase	1.6.*.*	DIA	634	DIA*1/1264		—	2	4	2, 2
Enolase 1	4.2.1.11	ENO1	1123	ENO1*1/2246		—	0	—	—
Esterase A-1	3.1.1.1	ESA1	2819	ESA1*1/5635		—	1	3	3
Esterase A-2	3.1.1.1	ESA2	2819	ESA2*1/5637		—	1	1	1
Esterase A-3	3.1.1.1	ESA3	2819	ESA3*1/5638		—	0	—	—
Esterase A-C	3.1.1.1	ESAC	2819	ESAC*1/5637		—	1	1	1
Esterase B	3.1.1.1	ESB	1663	ESB*1/3326		—	0	—	—
Esterase D	3.1.1.1	ESD	2827	ESD*1/4698	ESD*2/668		0	—	—
Fumarate hydratase	4.2.1.2	FUM	959	FUM*1/1918		—	0	—	—
Gal-1-phosphate uridyl transferase	2.7.7.12	GALT	2768	GALT*1/5036	GALT*D/501		1	1	1
Glucose-6-phosphate dehydrogenase	1.1.1.49	G6PD	2243	G6PD*B/3304	G6PD*A/6		3	5	3, 1, 1
Glucosephosphate isomerase	5.3.1.9	GPI	2807	GPI*1/5598		—	5	16	11, 2, 1, 1, 1
Glutamate-oxaloacetic transaminase(S)	2.6.1.1	GOTS	2823	GOTS*1/5636		—	5	10	4, 2, 2[c], 1, 1
Glutamate-pyruvate transaminase(S)	2.6.1.2	GPTS	1162	GPTS*1/1246	GPTS*2/1072		2	6	4, 2

Glyoxalase-1	4.4.1.5	GLO1	1346	GLO1*1/1497	GLO1*2/1195		0	—	—
Haemoglobin A-1		HGBA1	2807	HGBA1*/5614		—	0	—	—
Haemoglobin A-2		HGBA2	2807	HGBA2*/5614		—	0	—	—
Haemoglobin B		HGBB	2785	HGBA2*/5569		—	1	1	1[c]
Haemoglobin GA		HGBGA	2807	HGBGA*/5614		—	0	—	—
Haemoglobin GG		HGBGG	2807	HGBGG*5614		—	0	—	—
Hexokinase-1	2.7.1.1	HK1	2737	HK1*/5472		—	2	2	1, 1
Hexokinase-2	2.7.1.1	HK2	2806	HK2*/5612		—	0	—	—
Isocitrate dehydrogenase(S)	1.1.1.42	ICDS	2719	ICDS*1/5436		—	2	2	1, 1
Inosine triphosphatase	3.6.1.19	ITPA	1122	ITPA*1/2244		—	0	—	—
Lactate dehydrogenase A	1.1.1.27	LDHA	2807	LDHA*1/5612		—	2	2	1, 1
Lactate dehydrogenase B	1.1.1.27	LDHB	2807	LDHB*1/5613		—	1	1	1
Malate dehydrogenase	1.1.1.27	MDH	2807	MDH*1/5614		—	0	—	—
Nucleoside phosphorylase	2.4.2.1	NP	2818	NP*1/5633		—	3	3	1, 1, 1
Peptidase A	3.4.11.*	PEPA	2819	PEPA*1/5629	PEPA*2/4		3	5	2, 2, 1
Peptidase B	3.4.11.*	PEPB	2824	PEPB*1/5641		—	3	7	5, 1, 1
Peptidase C	3.4.11.*	PEPC	2631	PEPC*1/5254		—	2	8	7, 1
Peptidase D	3.4.13.9	PEPD	2633	PEPD*1/5231	PEPD*2/31	PEPD*3/1	2	3	2, 1
Phosphogluconate dehydrogenase	1.1.1.44	PGD	2866	6PGD*A/5471	6PGD*C/139		2	2	1, 1
Phosphoglucomutase-1	2.7.5.1	PGM1	2827	PGM1*1/4352	PGM1*2/1299		3	3	1, 1, 1
Phosphoglucomutase-2	2.7.5.1	PGM2	2826	PGM2*1/5649		—	2	3	2[c], 1[c]
Phosphoglycerate kinase	2.7.2.3	PGK	1866	PGK*1/2752		—	0	—	—
Inorganic pyrophosphatase	3.6.1.1	PP	678	PP*1/1354		—	0	—	—
Prealbumin		PALB	795	PRLB*1/1599		—	0	—	—
7-S-Beta globulin		BG7S	632	BG7S*1/1264		—	0	—	—
Transferrin		TF	2737	TF*C/5420	TF*D1/5		5	49	36, 8, 2, 2, 1
Triosephosphate isomerase	5.3.1.1	TPI	2805	TPI*1/5609		—	1	1	1
Uroporphyrinogen synthetase	4.3.1.8	URO	1810	URO*1/3614		—	1	6	6

[a] N is number of determinations.

[b] Classes is number of different rare variants and copies is the total number of rare variants, while copies/class is the number of copies of each rare allele.

[c] Rare alleles that were identified in both ethnic groups.

Table 1*b*. *Frequency of rare electrophoretic mobility variants in a sample of unrelated Black newborns*

		Polymorphic alleles	Rare alleles		
Locus	N[a]	Allele/number of copies	Classes	Copies	Copies/class
ALDH	86	ALDH*1/172	0	—	—
ACP	241	ACP1*B/371 ACP1*A/103 ACP1*C/5	2	3	2, 1
ADA	238	ADA*1/471 ADA*2/5	0	—	—
AK	240	AK1*1/475 AK1*2/1	0	—	—
ALB	206	ALB*1/412	0	—	—
APOA1	68	APOA1*1/134	1	2	2
CA1	73	CA1*1/144	1	2	2
CA2	93	CA2*1/177 CA2*2/9	0	—	—
CRPL	231	CRPL*B/430 CRPL*A/18 CRPL*M1/6 CRPL*CB/5	1	3	3
DIA	43	DIA*1/86	0	—	—
ENO1	86	ENO1*1/172	0	—	—
ESA1	240	ESA1*1/479	1	1	1
ESA2	240	ESA1*1/480	0	—	—
ESA3	240	ESA3*1/480	0	—	—
ESAC	240	ESAC*1/480	0	—	—
ESB	129	ESB*1/258	0	—	—
ESD	240	ESD*1/439 ESD*2/41	0	—	—
FUM	70	FUM*1/140	0	—	—
GALT	234	GALT*1/454 GALT*D/14	0	—	—
G6PD	179	G6PD*B/205 G6PD*A/50	0	—	—
GPI	238	GPI*1/476	0	—	—
GOTS	240	GOTS*1/477	3	3	1, 1, 1[c]
GPTS	88	GPTS*1/128 GPTS*2/47	1	1	1
GLO1	123	GLO1*1/177 GLO1*2/69	0	—	—
HGBA1	238	HGBA1*1/473	1	3	3
HGBA2	238	HGBA2*1/476	0	—	—
HGBB	237	HGBB*1/473	1	1	1[c]
HGBGA	238	HGBGA*1/476	0	—	—
HGBGG	238	HGBGG*1/476	0	—	—
HK1	233	HK1*1/466	0	—	—
HK2	239	HK2*1/478	0	—	—
ICDS	227	ICDS*1/454	0	—	—
ITPA	86	ITPA*1/172	0	—	—
LDHA	238	LDHA*1/472	1	4	4
LDHB	238	LDHB*1/476	0	—	—
MDH	238	MDH*1/476	0	—	—
NP	241	NP*1/482	0	—	—
PEPA	241	PEPA*1/445 PEPA*2/37	0	—	—
PEPB	242	PEPB*1/483	1	1	1
PEPC	221	PEPC*1/441	1	1	1
PEPD	223	PEPD*/433 PEPD*2A/8 PEPD*2B/5	0	0	0
PGD	237	PGD*A/461 PGD*C/13	0	—	—
PGM1	242	PGM1*1/396 PGM1*2/88	0	—	—
PGM2	242	PGM2*1/482	2	2	1[c], 1[c]
PGK	143	PGK*1/206	0	—	—
PP	72	PP*1/144	0	—	—
PALB	49	PALB*1/98	0	—	—
BG7S	41	BG7S*1/82	0	—	—
TF	231	TF*C/450 TF*D1/12	0	—	—
TPI	238	TPI*1/476	0	—	—
URO	177	URO*1/354	0	—	—

[a] Footnotes are the same as in Table 1*a*.

Table 2. *Frequency distribution of classes of variants/locus*

	Number of occurrences					
	Total alleles[a]			Rare alleles		
Variant classes/ locus	CA[b]	BL[c]	Pooled[d] total	CA[b]	BL[c]	Pooled[d] total
0	—	—	—	21	37	18
1	17	25	14	10	11	10
2	12	20	13	12	2	9
3	9	3	8	5	1	4
4	4	1	4	0	0	8
5	6	2	4	3	0	1
6	2	0	4	0	0	0
7	1	0	3	0	0	1
8	0	0	1	0	0	0

[a] Rare plus polymorphic alleles.
[b] CA = Caucasian.
[c] BL = Black.
[d] Pooled total includes all first borns, irrespective of ethnicity.

rather than subdivisions identified during subsequent studies to characterize some of the variants. This characterization provides insight into the level of microheterogeneity and is similar to the results from many similar studies designed to subdivide variants with apparently similar electrophoretic mobilities. Note that the number of variants identified differs from the data of Neel *et al.* (1987) because only unrelated (non-sibling) newborns have been included in the current sample.

The distribution of classes of variants, both total, rare plus polymorphic, and rare variants per locus is in Table 2. In the Caucasian sample, 17 of the 51 loci are monomorphic, that is, neither rare nor polymorphic variants were observed, while for 4 additional loci, *ADA*, *CAII*, *ESD* and *GLOI*, polymorphic but not rare mobility variants were detected. Rare variants were thus not detected at 30 of 51 loci. The average number of determinations at the 21 loci where rare variants were not detected was somewhat less ($\bar{x} = 1772$) than at the other 30 loci ($\bar{x} = 2573$). The maximum number of alleles (rare plus polymorphic) at any locus was 7 for *TF* in the Caucasian sample. In the combined sample, *GOT* exhibited 7 different rare alleles.

Most of the rare variants were recovered as singletons. Forty-one of the 64 (64%) different rare allele products detected in the Caucasian sample occurred as a single copy while two copies were detected for another 13 (20%) allele products (Table 3). In the total data set, 46 of the 79 (58%) rare variants occurred only once and another 17 (21%) twice.

The distribution of allele frequencies for the 133 gene products, including the most common variant, at each locus identified in the Caucasian sample is presented in Table 4. Sixteen monomorphic loci are included in the tabulation. Thirty-eight alleles have frequencies of less than 0·0002 which corresponds to a single copy in approximately 2700 determinations per locus. Allele frequencies of 0·0003–0·0005 generally reflect two copies of an allele in 2700 determinations. At the other extreme, nineteen alleles, at nine loci, were encountered in polymorphic frequencies (0·01–0·99). It should be noted that for 31 of the 51 loci studied, more than 2700 determinations were completed while 2000–2700 determinations were completed for

Table 3. *Frequency distribution of copies of rare alleles per variant class*

Number of copies/rare allele	Number of occurrences	
	Caucasian	Pooled[a] total
1	41	46
2	13	17
3	2	6
4	2	2
5	1	1
6	1	2
7	1	2
8	1	1
9	0	0
11	1	1
13	0	0
36	1	1

[a] Includes all first borns, irrespective of ethnicity.

five other loci. For only 12 loci were less than 1400 individuals studied and the minimum sample size for any protein studied in the Caucasian series was 632 individuals. The units of this tabulation are of course sample size dependent and most such tabulations are based on smaller sample sizes, thus curtailing the lower limit rather sharply. Even within this sample, the varying number of determinations per locus is a complication. Thus, we have tabulated separately the 12 loci where less than 1400 unrelated individuals were studied. The number of rare variant classes does not differ significantly between the two groups, being 0·4 rare variant classes per 1000 determinations for the < 1400 group and 0·5 for the > 1400 group, but the sample-size-imposed constraint on the observed variant frequencies is obvious.

Several different comparisons of allelic frequencies are in Table 5. Among the proteins studied for electrophoretic mobility variants, 20 are monomers, (ALDH, ACP1, ADA, AK1, ALB, APOA1, CA1, CA2, CRPL, DIA, ESB, HK1, HK2, PEPB, PEPC, PGM1, PGM2, PGK, TF and URO) of which 13 are monomorphic and 7 are polymorphic. Thirty-one proteins occur as homo- or heteropolymers, of which 23 are monomorphic and 8 are polymorphic. Rare variants were not detected at 20 loci of which 14 (of 31) were polymeric proteins and 6 (of 20) were monomers. The greater variability of the monomers is further reflected in a higher number of both total alleles and rare alleles per 1000 alleles and also more total copies of rare alleles than for proteins which exist as polymers. Although a pattern seems to exist, none of the differences are significant at a $P < 0{\cdot}20$.

At loci classified as polymorphic (at least two alleles exist at a frequency greater than 0·01), the number of rare variant classes is higher than at monomorphic loci ($\chi^2 = 2{\cdot}5$, D.F. = 1, $P < 0{\cdot}15$). A problem exists in a similar comparison of the number of copies of rare alleles as 36 copies of a single transferrin variant were identified. This frequency is almost three times the number of variants encountered for the next most numerous rare variant. The total number of rare variant copies is higher at polymorphic loci if all of the data are utilized but the number of copies of rare alleles is not increased if the one transferrin variant is excluded from the data base. The number of classes of rare variants detected among the serum proteins was higher than for either erythrocyte proteins or erythrocyte enzymes ($\chi^2 = 3{\cdot}2$, D.F. = 1, $P < 0{\cdot}10$).

Table 4. *Distribution of allele frequencies in the Caucasian sample of Ann Arbor*

	Number of occurrences	
Allele frequency[a]	< 1400[b]	> 1400[b]
⩽ 0·0002	0	38
0·0003–0·0005	1	12
0·0006–0·0010	2	4
0·0011–0·0020	3	6
0·0021–0·0050	0	0
0·0051–0·0100	0	4
0·011–0·050	0	4
0·051–0·100	0	1
0·11–0·20	0	—
0·21–0·50	2	2
0·51–0·75	2	1
0·76–0·999	3	28
1·000[c]	8	9

[a] Note that intervals are not equal.
[b] Sample size.
[c] These loci were monomorphic.

Table 5. *Various comparisons of allele frequency in the Caucasian sample*

A.	Total variants – classes/1000 alleles	
	Monomers	0·67
	Polymers	0·49
	Rare variants – classes/1000 alleles	
	Monomers	0·33
	Polymers	0·25
	Rare variants – copies/1000 alleles	
	Monomers	1·08
	Polymers	0·44
B.	Rare variants – classes/1000 alleles	
	Monomorphic	0·25
	Polymorphic	0·36
	Rare variants – copies/1000 alleles	
	Monomorphic	0·25
	Polymorphic	1·26
C.	Total variants – classes/1000 alleles	
	Serum proteins	0·82
	Erythrocyte proteins	0·59
	Erythrocyte enzymes	0·64
	Rare variants – classes/1000 alleles	
	Serum proteins	0·48
	Erythrocyte proteins	0·26

The relative electrophoretic mobility of the variant classes is symmetrically distributed about the primary isozyme. The charge gain or loss of the variants was estimated from their mobility relative to the separation of the primary isozyme and other variants in the same system and also from the mobility of the degradation products/secondary isozymes relative to the parent molecule. The assignment of relative mobilities is a subjective estimate and is not meant to imply that the actual amino acid composition of the variants is known. Twenty-four allelic subunits have a mobility of plus one charge unit, while 23 are minus one relative to the most common allele. The number of additional alleles which differ by other than one charge unit are: +2/13, +3/7, +4/3 and −2/14, −3/5, −4/3. Nine other variant subunits had

Table 6. *Relationship of subunit size and variant frequency in the Caucasian sample expressed as classes of variants per 100 amino acid residues*

		Intercept	Slope	Correlation coefficient
A.	Rare variants			
	Monomeric proteins	1·01	0·12	0·32
	Polymeric proteins	−0·44	0·57	0·60
B.	Rare variants			
	Monomorphic loci	0·60	0·18	0·34
	Polymorphic loci	0·50	0·30	0·67
C.	Rare variants	0·58	0·22	0·42
D.	Total variants	1·48	0·30	0·44

intermediate mobilities. Again they were distributed symmetrically about the primary isozyme.

The relationship between subunit size and the number of different variants at a locus is outlined in Table 6. At all loci where no rare variants had been detected, it was assumed that this reflected sample size limitations rather than an absolute estimate of the frequency. Therefore, a 'calculated' rare variant frequency was obtained for these nonvariant loci, by assuming a numerator of 0·4 variants and a denominator equal to the number of alleles actually studied, a technique similar to that previously employed by Harris *et al.* (1977). In the few cases for which the subunit size of the various proteins, for the human, was not available from the literature, data from the most closely related species were used. The regression of classes of rare variants (following the adjustment described above) on number of amino acids per peptide was calculated assuming a linear relationship. The data are consistent with the expectation of more variant classes being associated with larger gene products, the scale being approximately 0·2 variant classes for each 100 amino acid residue increase in mass. The slope is steeper, 0·3 variant classes per 100 residues, when the relationship between total variant classes per locus (rare plus polymorphic allele classes) and subunit size is the basis for the calculation. The effect of subunit size on the number of variant classes was more pronounced for polymers than monomers and also for polymorphic over monomorphic proteins.

DISCUSSION

Only two other studies have examined rare variant frequencies at such a range of loci in similar large, nontribal population surveys. The rare variant frequency in the Japanese study was 1·9/1000 determinations, an allele frequency of 0·00095 (Neel *et al.* 1978). The variant frequency in similar data from England, following exclusion of the very high frequency for placental alkaline phosphatase variants, is 1·1 heterozyotes/1000 individuals, an allele frequency of 0·00057 (Harris *et al.* 1974). These figures contrast with estimates of 1·3 rare variant heterozygotes per 1000 determinations (allele frequency of 0·00068) for Ann Arbor Caucasians and 2·8 rare variant heterozygotes per 1000 determinations (allele frequency of 0·00141) for Ann Arbor Blacks. If, however, the 'adjustment' for possible admixture effects described earlier is omitted, the variant frequency for Ann Arbor Caucasians becomes 2·2 variants/1000 determinations while the variant frequency for Blacks remains about the same at 2·9 variants/1000 determinations. All 25 of the loci included in the Japanese data base and 23 of the 43 loci in the English study are included in the 51 loci examined in this study (most of the other loci in the English survey are expressed only in tissues other than erythrocytes).

The frequency of different rare variant classes in the English sample is 0·35 different rare variants/1000 determinations (Harris *et al.* 1974). The comparable figure for Japan is 0·43 rare variant classes/1000 determinations (Neel *et al.* 1978). The frequency of unique rare variants in the Ann Arbor Caucasian sample is similar, 0·55 classes/1000 determinations, but the frequency of 1·87 rare variant classes/1000 determinations in the Black sample is much higher than the other three estimates. The difference in frequency of rare variant classes in Ann Arbor Blacks and Caucasians is increased slightly if the 'adjustment' for admixture is removed.

In all of the groups, approximately 50% of the rare variants are identified in only a single individual. In the English study, no rare variants were detected at 21 of 43 loci, while five of the remaining loci exhibit only polymorphic variation. Eleven of 25 loci in the Japanese study were without rare variants, while four of the 25 loci exhibited only polymorphic variants. No rare variants were detected at 20 of 51 loci in the Ann Arbor Caucasian sample, although four of the remaining 31 loci did exhibit polymorphic variation. Thus, in each study, over 40% of the loci did not yield a rare variant even though the number of individuals studied generally ranged from 1000–10000.

As only four of the rare variants in the Ann Arbor data base (two variants at the PGM2 locus and one each for HGBB and GOTS, see Table 1*a*, *b*) were apparently in common to the two ethnic groups, it would seem unlikely that the higher incidence of rare variants in the Black sample reflects only admixture of the Black gene pool with Caucasian genes. To what extent the increased frequency of variant classes in the Black population is an artifact reflecting the smaller sample size is unclear, although the apparently higher frequency of rare variants is observed in other data where the Black and Caucasian samples are of similar size (Silver, 1982; Gershowitz, pers. commun.).

Previous data, summarized from several different human populations, have indicated that 56% of monomeric enzymes are polymorphic, while only 30% of the polymeric proteins are polymorphic (Harris *et al.* 1977; Ward, 1977). Within the Ann Arbor sample, 34% of the monomers are polymorphic while 24% of the polymers are polymorphic. These numbers are not significantly different from the data collected by Harris *et al.* (1974) in their study of European populations. The reduced proportion of polymorphic loci in the latter two studies probably reflects the fact that the loci included in these studies were selected because of the availability of an electrophoretic technique rather than other selective criteria.

Within the Ann Arbor Caucasian sample, the number of variant classes, both total and rare, and the frequency of rare variant heterozygotes, are higher for monomeric than for polymeric proteins, but the differences are not statistically significant. Similar observations had been noted in previous summaries of collections of data from various studies (Ward, 1977; Harris *et al.* 1977). Likewise more classes of rare variants and more copies of rare variants are observed at polymorphic loci than at monomorphic loci. This increase is still seen when the loci are further subdivided into monomers and polymers.

Within a subset of this newborn sample, other techniques have been utilized to identify enzyme deficiency and thermostability variants. Enzyme deficiency variants exist at higher frequency than do electrophoretic mobility variants in both populations, enzyme deficiency alleles being detected with a frequency of 3·1 variants/1000 determinations in Caucasians and 19·2 variants/1000 determinations in Blacks (Mohrenweiser, 1981; 1983; in preparation). The frequency for the Black population is reduced to 9·1 deficiency variants if the polymorphic TPI*1° and G6PD*A$^-$ variants are excluded. The allele frequency for thermostability variants not characterized by an altered electrophoretic mobility is 4 variants/1000 determinations in the Caucasian sample (Mohrenweiser & Neel, 1981). (Data are not available for estimating the frequency of thermostability variants in the Black sample.) The total detectable variation associated with rare variants in the Caucasian population becomes 8–10 variants/1000 determinations if the data from the three approaches, which detect classes of genetic variation which do not overlap significantly, are combined.

The frequency distribution of the relative electrophoretic mobilities of the products of the variant alleles is consistent with previous observations and with models of step-wise mutation (Fuerst & Ferrell, 1980; Haldorson & King, 1976; Marshall & Brown, 1975). It is intriguing to note that 12 of the 18 variant classes which differ from the most common allele by 3 or 4 charge units exist at polymorphic loci, although the polymorphic loci constitute only 15 of the 51 loci studied and account for only 30 of the 79 rare variant classes identified. Therefore, the variant frequency at the polymorphic loci would not be the only explanation for this distribution. It is interesting to speculate that intragenic recombinational events as well as recurrent mutation at an already polymorphic allele, may be responsible for generating some of these variants with mobilities that are quite different from the common alleles (Bowman & Kurosky, 1982; Carter *et al.* 1979; Takahashi *et al.* 1982).

A positive relationship between subunit molecular mass and both the number of classes and number of copies of rare variants was detected in this study, the increase being of the order of 0·3 variants for each 100 amino acid residue increase in peptide length and similar to other estimates (Ward, 1978). The relationship is stronger for polymeric proteins than for monomers, as was previously observed (Eanes & Koehn 1978; Harris *et al.* 1977; Koehn & Eanes, 1977). This would be consistent with the larger proteins (genes) presenting a larger target for mutation. It should also be noted that many gene products are identified, following electrophoretic separations, by functional characteristics, thus amino acid substitutions which alter protein function or stability may not be detected as electrophoretic mobility variants. Therefore, the size/variation relationship could also be explained if larger proteins have less constrained conformations and therefore are more capable of tolerating a range of different amino acid substitutions. Also, mutations giving rise to variants with impaired function would often be associated with negative selection and therefore have a greater probability of not being transmitted to future generations. The real explanation for the relationship between subunit size and variant frequency is undoubtly an interplay of both factors but is probably dominated by the conformational requirements for catalytic function and protein stability, as observed with recent studies using site directed mutagenesis techniques and genetic variants (e.g. Daar *et al.* 1986, Straus *et al.* 1985).

This work was supported by Department of Energy contract DE-AC02-82ER-60089. Many individuals have provided a tremendous amount of assistance, which is most gratefully acknowledged, during the generation of this data base, the management of the data and the writing of the manuscript.

REFERENCES

ASAKAWA, J. & MOHRENWEISER, H. W. (1982). Characterization of two new electrophoretic variants of human triosephosphate isomerase: stability, kinetic and immunologic properties. *Biochem. Genet.* **20**, 59–76.

BARRANTES, R., SMOUSE, P. E., NEEL, J. V., MOHRENWEISER, H. W. & GERSHOWITZ, H. (1982). Migration and genetic infrastructure of the Central American Guaymi and their affinities with other tribal groups. *Am. J. Phys. Anthropol.* **58**, 201–214.

BOWMAN, B. H. & KUROSKY, A. (1982). Haptoglobin: the evolutionary product of duplication, unequal crossing over, and point mutation. *Adv. Hum. Genet.* **12**, 189–261.

CARTER, N. D., WEST, C. M., EMES, E., PARKIN, B. & MARSHALL, W. H. (1979). Phosphoglucomutase polymorphism detected by isoelectric focusing: gene frequencies, evolution and linkage. *Ann. Hum. Biol.* **6**, 221–230.

DAAR, I. O., ARTYMIUK, P. J., PHILLIPS, D. C. & MAQUAT, L. E. (1986). Human triosephosphate isomerase deficiency: A single amino acid substitution results in a thermolabile enzyme. *Proc. Natl. Acad. Sci. USA* **83**, 7903–7907.

EANES, W. F. & KOEHN, R. K. (1978). Relationship between subunit size and number of rare electrophoretic alleles in human enzymes. *Biochem. Genet.* **16**, 971–985.

FUERST, P. A. & FERRELL, R. E. (1980). The stepwise mutation model: an experimental evaluation utilizing hemoglobin variants. *Genetics* **94**, 185–201.

FUJITA, M., SATOH, C., ASAKAWA, J., NAGAHATA, Y., TANAKA, Y., HAZAMA, R. & KRASTEFF, T.(1985). Electrophoretic variants of blood proteins in Japanese. VI. Transferrin. *Jap. J. Hum. Genet.* **30**, 191–200.

HALDORSON, L. & KING, J. H. L. (1976). Unimodality, symmetry and step – state hypothesis of electrophoretic variation in natural populations. *J. Mol. Evol.* **8**, 351–356.

HARRIS, H. & HOPKINSON, D. A. (1976). *Handbook of Enzyme Electrophoresis in Human Genetics.* New York: American Elsevier.

HARRIS, H., HOPKINSON, D. A. & EDWARDS, Y. H. (1977). Polymorphism and the subunit structure of enzymes: a contribution to the neutralist–selectionist controversy. *Proc. Natl. Acad. Sci. USA* **74**, 698–701.

HARRIS, H., HOPKINSON, D. A. & ROBSON, E. B. (1974). The incidence of rare alleles determining electrophoretic variants: Data on 43 enzyme loci in man. *Ann. Hum. Genet.* **37**, 237–253.

HAUPT, H. & BÖHM, H. (1977) Isolierung und Charakterisierung eines 7S–beta Globulins aus menschlichen Erythrozyten. *Blut* **35**, 229–239.

KOEHN, R. K. & EANES, W. F. (1977). Subunit size and genetic variation of enzymes in natural populations of *Drosophila. Theoret. Popul. Biol.* **11**, 330–341.

LONG, J. C., NAIDU, J. M., MOHRENWEISER, H. W., GERSHOWITZ, H., JOHNSON, P. L., WOOD, J. W. & SMOUSE, P. E. (1986). Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New-Guinea. *Am. J. Phys. Anthropol.* **70**, 75–96.

MARSHALL, D. R. & BROWN, A. H. D. (1975). The charge state model of protein polymorphism in natural populations. *J. Mol. Evol.* **6**, 149–163.

MOHRENWEISER, H. W. & DECKER, R. S. (1982). Identification of several electrophoretic variants of human ceruloplasmin including CP* Michigan, a new polymorphism. *Hum. Hered.* **32**, 369–373.

MOHRENWEISER, H. W. & FIELEK, S. Identification and characterization of a series of rare G6PD variants in a human newborn series. *Hum. Genet.* (submitted).

MOHRENWEISER, H. W., WADE, P. T. & WURZINGER, K. H. (1987). Characterization of a series of electrophoretic and enzyme activity variants of human glucose phosphate isomerase. *Hum. Genet.* **75**, 28–31.

MOHRENWEISER, H. W. (1981) Frequency of enzyme deficiency variants in erythrocytes from newborn infants. *Proc. Natl. Acad. Sci. USA* **78**, 5046–5050.

MOHRENWEISER, H. W. (1983) Enzyme deficiency variants: Frequency and potential significance in human populations. *Isozymes: Curr. Topics in Biol. Med. Res.* **10**, 51–68.

MOHRENWEISER, H. W. Functional hemizygosity in the human genome: Direct estimate from twelve erythrocyte enzyme loci. *Hum. Genet.* (In the Press).

MOHRENWEISER, H. W. & NEEL, J. V. (1981) Frequency of thermostability variants and estimation of the total "rare" variant frequency in human populations. *Proc. Natl. Acad. Sci. USA* **78**, 5729–5733.

MOURANT, A. E., KOPEĆ, A. C. & DOMANIEWSKA-SOBCZAK, K. (1976). *The distribution of human blood groups and other polymorphisms.* Oxford Monographs on Medical Genetics. Oxford, London: Oxford University Press.

NAIDU, J. M., MOHRENWEISER, H. W. & NEEL, J. V. (1985). A sero-biochemical genetic study of Jalari and Brahmin caste populations of Andhra Pradesh, India. *Hum. Hered.* **35**, 148–156.

NEEL, J. V., MOHRENWEISER, H. W. & GERSHOWITZ, H. (1987). A pilot study of the use of placental cord blood samples in monitoring for mutational events. *Mutat. Research* (In the Press).

NEEL, J. V., MOHRENWEISER, H. W. & MEISLER, M. M. (1980). Rate of spontaneous mutation at human loci encoding protein structure. *Proc. Natl. Acad. Sci. USA* **77**, 6037–6041.

NEEL, J. V., UEDA, N., SATOH, C., FERRELL, R. E., TANIS, R. J. & HAMILTON, H. B. (1978). The frequency in Japanese of genetic variants of 22 proteins. V. Summary and comparison with data on Caucasians from the British Isles. *Ann. Hum. Genet.* **41**, 429–441.

PEARSE, A. G. E. (1972). *Histochemistry – theoretical and applied.* 3rd ed. vol. 2. Baltimore, MD: Williams & Wilkins Co.

REED, T. E. (1969). Caucasian genes in American Negroes. *Science* **165**, 762–768.

SILVER, H. (1982). *Probability of inclusion in paternity testing.* Arlington, VA: American Association of Blood Banks.

STRAUS, D., RAINES, R., KAWASHIMA, E., KNOWLES, J. R. & GILBERT, W. (1985). Active site of triosephosphate isomerase: *In vitro* mutagenesis and characterization of an altered enzyme. *Proc. Natl. Acad. Sci. USA* **82**, 2272–2276.

TAKAHASHI, N., NEEL, J. V., SATOH, C., NISKIZAKI, J. & MASUNARI, N. (1982). A phylogeny for the principal alleles of the human phosphoglucomutase-1 locus. *Proc. Natl. Acad. Sci. USA* **79**, 6636–6640.

TILLS, D., KOPEĆ, A. C. & TILLS, R. E. (1983). *The distribution of human blood groups and other polymorphisms: a supplement.* Oxford Monographs on Medical Genetics. Oxford, London: Oxford University Press.

WARD, R. D. (1977). Relationship between enzyme heterozygosity and quaternary structure. *Biochem. Genet.* **15**, 123–135.

WARD, R. D. (1978). Subunit size of enzymes and genetic heterozygosity in vertebrates. *Biochem. Genet.* **16**, 799–810.

WURZINGER, K. H. & MOHRENWEISER, H. W. (1982). Studies on the genetic and non-genetic (physiological) variation of human erythrocyte glutamic oxaloacetic transaminase. *Ann. Hum. Genet.* **46**, 191–201.