Printed in Great Britain

The frequency in Japanese of genetic variants of 22 proteins

V. Summary and comparison with data on Caucasians from the British Isles

By J. V. NEEL,* N. UEDA‡, C. SATOH,‡ R. E. FERRELL,*† R. J. TANIS*
and H. B. HAMILTON‡

During the years 1972–5 some 4000 blood samples from residents of Hiroshima and Nagasaki participating in the Adult Health Study of the Radiation Effects Research Foundation were electrophoretically screened for variants of 22 different proteins. The examinations were undertaken in the course of a study to establish the feasibility of an investigation of this type on children born to survivors of the atomic bombings. However, all of the participants in this pilot study were born prior to the atomic bombings, so that the question of induced genetic effects does not arise in considering these findings. The first four papers in this series have described the variants which were encountered (Ferrell *et al.* 1977; Ueda *et al.* 1977; Satoh *et al.* 1977; Tanis *et al.* 1978). The purpose of this paper is to compare these various systems with respect to the frequency of 'rare' variants, and to contrast these findings with essentially comparable data on the Caucasoid inhabitants of the British Isles. Preliminary data are also available from this laboratory on rare variants of a twenty-third system, Esterase D (J. Yamashita *et al.*, unpublished) and this will be included in the comparison. A total of 74799 determinations is involved.

In this treatment we shall define a rare variant as one present in less than 2 % of the population. This commonly employed definition of a 'rare' variant is of course quite arbitrary, since there is a continuum in the frequency with which individual variants occur in populations, from very rare to common. A second arbitrary element in the definition derives from the fact that for the present, identity has been defined by electrophoretic behaviour. As the evidence from the haemoglobinopathies so well demonstrates, a particular electrophoretic variant (gain of one charge, as in haemoglobin S) may be demonstrated to be due to any of some 30 different amino acid substitutions, when appropriate techniques are available. Although it is planned ultimately to apply to these variants the studies on kinetic parameters, physical characteristics, etc., which should permit further resolution of these phenotypes, this has not been possible up to now. Thus, the present analysis is clearly not definitive, but more in the nature of a beginning.

* Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109. Financial support for the study derived from Contract E(11-1)-1552 with the Energy Research and Development Agency, Washington, D.C.

† Present address: Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences at Houston, The University of Texas, Houston, Texas 77025.

‡ Radiation Effects Research Foundation, Hiroshima, Japan. The Foundation, the successor to the Atomic Bomb Casualty Commission (ABCC), was established in April 1975 as a private non-profit Japanese Foundation, under the auspices of the Ministry of Health and Welfare and Finance Ministry, supported equally by funds from the Government of Japan through the Ministry of Health and Welfare, and the Government of the United States through the National Academy of Sciences under contracts with the Energy Research and Development Administration, the National Cancer Institute and the National Heart and Lung Institute.

THE GENETIC NATURE OF THE VARIANTS

As noted in the preceding papers, limited genetic studies, sufficient to establish the presence of the variant in at least one other member of the family, were carried out where possible. However, for such reasons as lack of co-operation or unavailability of other family members, family studies were not always possible. Furthermore, in several instances all the first-degree relatives who could be studied proved to be negative for the trait. Since we propose to use these data for normative purposes, it seems important to establish in so far as possible that the numbers are not inflated by persistent artifacts. A simple test for the genetic basis of what we have reported is provided by the expectation that if only the first of the first-degree relatives to be contacted is considered, 50 % of them should show the trait if it is inherited in the usual co-dominant fashion. The occurrence of mutation should decrease the expectation below 50 %, but presumably this is such a sufficiently uncommon event that this small bias may be ignored. In the preceding four papers in this series, in presenting the family data we have indicated the initial first-degree relative contacted by an asterisk. In the total material, 121 family studies were undertaken, and of the 121 first-degree relatives through whom the family contact was initiated, 57 were affected ($\chi^2 = 0{\cdot}405$, D.F. = 1, $0{\cdot}50 < P < 0{\cdot}70$). Of the 64 instances in which the initial first-degree relative contacted did not possess the trait, in 22 cases no other family members were available for study. Where additional family members were available, the trait was subsequently demonstrated in the family in 28 instances. With respect to the remaining 14 instances, in which the trait was not observed in the family, the mean number of first-degree relatives studied was only 2·6. We shall presume that all the variants reported in the preceding four papers were genetic in nature, even where confirmatory family data are lacking.

Inspection of the tables which summarize the family studies in the four preceding papers indicates that the examined first-degree relatives of the propositi include very few parents. It should be recalled that the individuals included in this investigation, drawn from the Adult Health Study of the Radiation Effects Research Foundation, were all alive at the time of the atomic bombings. In Hiroshima their mean age at the time of the bombings was 32·1 years for males and 30.6 years for females, while for Nagasaki the corresponding figures are 25·0 and 23·6 (Belsky, Tachikawa & Jablon, 1971). Given attrition in the 30-year interval since the bombings plus the increased mortality resulting at the time of the bombings, it is clear why so few parents could be included in the family studies, but this fact renders the data of little value as a contribution to the question of what proportion of such variants can be attributed to mutation occurring in the preceding generation.

THE NUMBER AND FREQUENCY OF VARIANTS OF THE 22 PROTEINS

Comparisons across loci within a population are of limited value for a number of reasons: (1) the proportion of variants detected by the electrophoretic approach may not be the same for all loci, (2) the variants observed in a population are a complex function of mutation, selection, migration and random loss, at least the first two of which may vary from locus to locus within a population, so that a difference in variant frequency between two proteins can reflect the action of a variety of parameters, and (3) the various proteins may differ in size and/or number of constituent polypeptides, to an extent where failure to correct for the number of nucleotide pairs in a cistron at risk of mutation may obscure basic similarities (or result in spurious differences).

There is, however, one situation wherein cross-locus comparisons have relative validity. This is in the case of structural genes which may be assumed to have arisen through duplication from a common precursor. In the present series, the genes responsible for the α, β and δ polypeptides of haemoglobins A and A_2 and those responsible for carbonic anhydrase I and II (CA I and CA II) would clearly fall into this category (Ingram, 1961; Tashian & Carter, 1976). The case is less clear for phosphoglucomutase 1 and 2 (PGM_1 and PGM_2), but the similarity in their molecular size and activity (refs. in Harris, 1975) is such that following the suggestion of Hopkinson & Harris (1969), we will consider them as products of a duplication, even though the structural genes are now located on separate chromosomes and the chemical evidence concerning homologies in the two polypeptides is not yet at hand. Even for these loci, however, one must be cautious in comparisons, since, for instance, the minor representation of haemoglobin A_2 as contrasted with A almost surely affects the relative ease of detection of variants of A_2.

Having now entered the usual caveats, we nevertheless proceed, in Table 1, to a summary comparison of the findings. Note the results are now presented in terms of allele products examined, i.e. number of determinations × 2. In our subsequent statistical treatment we shall neglect the fact that uniting gametes are not completely uncorrelated in Japan (as well as the fact there are a few siblings in our sample). For oligomeric proteins composed of non-identical subunits, we list the subunits separately. Thus, data are available for 25 polypeptides. These data, like all electrophoretic data, represent minimal estimates of the frequency of amino acid substitutions or other genetic events resulting in a charge change in the molecule. It is apparent by inspection that there are large differences between these loci in the number of different variants encountered, and the representation of each. This of course comes as no surprise, being well established in the literature. What does seem noteworthy is the difference between the results for the haemoglobin polypeptides and for CA I and II, on the one hand, as contrasted, on the other hand, to the results for the two phosphoglucomutases. Our relatively limited data indicate comparability on the part of the haemoglobin polypeptides. Although the world literature on haemoglobins, including that from Japan, contains many more reports of variants of the β than of the δ chain (summarized in Jones & Koler, 1975), it is clear that much more effort has gone into detecting variants of the former, and one cannot take this finding alone as reflecting a real difference. There are in our series four individuals with a variant of CA I as contrasted to none of CA II in a sample of 3969, but all the variants of CA I have the same electrophoretic mobility and may well trace to the same mutation. The two phosphoglucomutases, however, exhibit a large difference, which cannot readily be attributed to differences in the ease of detection of variants of the two proteins. In the course of 1892 determinations, no variants of PGM_2 were encountered, but there were 15 rare variants of PGM_1, of seven different electrophoretic types (as well as two genetic polymorphisms). Thus far other investigators have also failed to detect variants of phosphoglucomutase-2 in Japanese (summary in Blake & Omoto, 1975).

The significance of this difference has been evaluated by two different approaches, neither completely satisfactory. First, an expectation for each locus based on the pooled data has been calculated ($\bar{p}$) and then the following computed:

$$\frac{(a-N_1\bar{p})^2}{N_1\bar{p}}+\frac{(b-N_1\bar{q})^2}{N_1\bar{q}}+\frac{(c-N_2\bar{p})^2}{N_2\bar{p}}+\frac{(d-N_2\bar{p})^2}{N_2\bar{q}}, \qquad (1)$$

Table 1. *A summary and comparison of genetic variants of 22 proteins (25 polypeptides) in Japanese and British Caucasoid populations*

The data on the Japanese are from the preceding four papers in this series. Data on British are from the summary of Harris, Hopkinson & Robson (1974) except as indicated by footnotes. With respect to the entry under 'variants', the first figure is the total number and the second figure, in parentheses, is the number of electrophoretically different variants. No appropriate data have been located for ceruloplasmin, or haemoglobin A_2. In comparing the numbers of determinations listed for the Japanese material with the data given in the earlier papers, bear in mind that untypable patterns are not included in the total.

	Japanese			Caucasian (British Isles)			
System	Variants	Determinations × 2	Variants/1000 persons	Variants	Determinations × 2	Variants/1000 persons	χ^2
Albumin	10 (1)	8058	2·48	0	1496	0*	1·86
Ceruloplasmin	2 (2)	8052	0·50	—	—	—	—
Haptoglobin	2 (2)	6894	0·58	1	794	2·52†	—
Transferrin	84 (8)	8040	20·90	2 (2)‡	276	14·39	0·27
Carbonic anhydrase I	4 (1)	7938	1·00	3 (3)§	20,230	0·30	2·90
Carbonic anhydrase II	0	7938	0	0	868	0‖	—
Lactate dehydrogenase:							
A subunit	0	8056	0	2 (1)	2030	1·97	—
B subunit	1 (1)	8056	0·25	0	2030	0	—
Malate dehydrogenase	1 (1)	8058	0·25	1 (1)	1032	1·94	—
Nucleoside phosphorylase	0	1476	0	2 (2)	3084	1·30	—
Triose phosphate isomerase	0	1476	0	2 (2)	3410	1·17	—
Haemoglobin:							
α-chain	0	8058	0	4 (3)	21·942	0·36	1·47
β-chain	2 (2)	8058	0·50	10 (7)¶	21,942	0·91	0·63
δ-chain	0	8058	0	—	—	—	—
Phosphoglucomutase-1	15 (7)‡‡	3784	7·40	12 (5)	20,666	1·16	33·20***
Phosphoglucomutase-2	0	3784	0	7 (3)	20,666	0·68	1·28
6-Phosphogluconate dehydrogenase	3 (3)	8030	0·75	1 (1)	9878	0·20	1·47
Adenylate kinase	0	4500	0	1 (1)	13,520	0·15	—
Adenosine deaminase	0	8042	0	2 (2)	9596	0·42	—
Esterase D	0	1562	0	0	908	0	—
Phosphohexose isomerase	35 (5)	8054	8·69	1 (1)	3100	0·65	11·26***
NADP-isocitrate dehydrogenase	1 (1)	7986	0·25	4 (2)	1436	5·57	—
Peptidase A	6 (1)	8018	1·50	8 (6)	17,596	0·91	0·87
Peptidase B	8 (2)	8056	1·99	15 (3)	14,082	2·13	0·03
Acid phosphatase	0	5580	0	0	15,774	0	—

* Cooke, Cleghorn & Lockey (1961) report no variants in 'over 12,000 sera', but the method is not described. Cohen (1965) found no variants in 748 sera using, like ourselves, starch-gel electrophoresis.

† Data of Allison, Blumberg & ap Rees (1958) and Harris, Robson & Siniscalco (1959). The single variant reported by Harris *et al.*, was of the Carlsberg type. Variants of this type have unusual features which suggest they may be due to somatic mosaicism.

‡ Data of Harris (1959). A later paper (Harris *et al.* 1960) reports an additional variant found in a total of 500 determinations but does not update the series.

§ Data of Carter *et al.* 1972; Carter *et al.* 1973.

‖ Data of Hopkinson *et al.* 1974. ¶ Data compiled by Harris (1975).

‡‡ PGM_1 phenotypes 2 and 7 are excluded from this comparison; an additional electrophoretic variant seen in screening employing a different buffer system (histidine, pH 7·0) is not included in this tabulation (cf. Satoh *et al.*, in manuscript).

where a = non-variants of PGM_1, b = variants of PGM_1, c = non-variants of PGM_2, d = variants of PGM_2, $N_1 = a+b$, $N_2 = c+d$, and $\bar{q} = 1-\bar{p}$.

The sum is considered an asymptotic approximation to χ^2 with 1 degree of freedom, and a significance value assigned on this basis. This test in essence assumes the variants to be of independent origin. Since this is unlikely to be the case, this test exaggerates the significance of the difference between the two loci. For the PGM$_1$ contrast, $\chi^2 = 15{\cdot}030$, D.F. = 1, $P < 0{\cdot}001$.

Ewens (1972) has developed the basis for a second contrast. Under the assumptions that (1) the variants detected are neutral in their phenotypic effects, (2) mutation is non-repetitive, (3) each electrophoretic class defines only one variant, and (4) the population is in equilibrium as regards mutation (i.e. no recent increase or decrease in mutation rates) he defines a parameter, θ, equivalent to $4N_e\mu$ where N_e is effective population number and μ is rate of mutation/locus/generation. The parameter θ can be estimated by iteration from the relationship

$$E(k) = \frac{\theta}{\theta}+\frac{\theta}{\theta+1}+\frac{\theta}{\theta+2}+\dots+\frac{\theta}{\theta+2n-1}, \tag{2}$$

where $E(k)$ is taken to be the *total* number of different alleles detected in a population and n is the number of individuals sampled; note the $2n$ of the final term so that here, as for the χ^2, the presentation involves the number of allele products tested rather than the number of individuals.

Table 2 presents the θ values for all the loci included in this survey. Since k includes all the alleles at each locus, k will of course not agree with the number of rare variants at the locus given in Table 1. It is apparent from (2) that θ cannot be computed for a monomorphic locus (at which neither rare variants or genetic polymorphisms exist). Ewens (1972) has also provided a method for computing a confidence interval for θ, and 95 % confidence intervals have been computed for all these loci (including the monomorphic).

It is most unlikely that all the assumptions which underlie the calculation of θ are met by the Japanese population. However, one would expect that sister (duplication) loci would be affected sufficiently equally by any departures from the assumptions that a contrast between such loci is in general valid. In the case of the two PGM's, there is no overlap between the upper estimate of θ for PGM$_2$ and the lower estimate for PGM$_1$. In a problem of this nature we feel that conservative test criteria are indicated, especially since with so many tests of differences by chance some may exceed conventional levels of significance, but at this point the difference between PGM$_1$ and PGM$_2$ as regards genetic variants seems secure.

The contrast employing θ is conservative because of the fact that one electromorph may, as the study of haemoglobin variants so well demonstrates, be the manifestation of a number of different alleles. Thus, the number of mutant alleles in the Japanese population for PGM_1 may well be higher than the number of variant electromorphs, whereas for PGM_2, since there are no variant electromorphs, the basis for concealed variability does not exist (we neglect for both loci the concealed variability within the normal electromorph). We conclude that the true level of significance of the difference between the two loci is to be found somewhere between the results of these two tests.

In the case of a within-population difference such as is exhibited by the two PGM's, the most likely explanations are differences in mutation rates or selective pressure at the two loci; the other factors mentioned earlier as producing differences between loci should bear equally on loci originating through duplication, such as these. There are important differences in the

Table 2. *Sample size* (N), *numbers of different alleles* (*electromorphs*) *recovered* (k), *estimated θ-values, and* 95% *confidence intervals* (θ_L, θ_U) *for* 25 *polypeptides in Japanese and English populations*

	Japan					England				
Genetic locus	N	k	θ_L	θ	θ_U	N	k	θ_L	θ	θ_U
Albumin	4029	2	0·0256	0·1063	0·3969	748	1	0·0000	—	0·3935
Ceruloplasmin	4026	3	0·0661	0·2159	0·6119	—†	—	—	—	—
Haptoglobin	3447	4	0·1199	0·3345	0·8220	397	3	0·0887	0·2917	0·8363
Transferrin	4020	9	0·4685	0·9284	1·7139	139	3	0·1053	0·3483	1·0082
Carbonic anhydrase I	3969	2	0·0256	0·1064	0·3975	10,115	4	0·1069	0·2974	0·7283
Carbonic anhydrase II	3969	1	0·0000	—	0·3212	434	1	0·0000	—	0·4249
Lactate dehydrogenase:										
A subunit	4028	1	0·0000	—	0·4259	1015	2	0·0300	0·1249	0·4684
B subunit	4028	2	0·0256	0·1063	0·3969	1015	1	0·0000	—	0·4684
Malate dehydrogenase	4029	2	0·0256	0·1063	0·3969	516	2	0·0328	0·1368	0·5143
Nucleoside phosphorylase	738	1	0·0000	—	0·3942	1542	3	0·0739	0·2418	0·6877
Triose phosphate isomerase	738	1	0·0000	—	0·3942	1705	3	0·0730	0·2388	0·6789
Haemoglobin:										
α-chain	4029	1	0·0000	—	0·3207	10,971	4	0·1060	0·2950	0·7222
β-chain	4029	3	0·0661	0·2159	0·6118	10,971	8	0·3478	0·7161	1·3626
δ-chain	4029	1	0·0000	—	0·3207	—*	—	—	—	—
Phosphoglucomutase-1	1892	10	0·6076	1·1668	2·1005	10,333	7	0·2829	0·6123	1·2141
Phosphoglucomutase-2	1892	1	0·0000	—	0·3496	10,333	4	0·1066	0·2968	0·7267
6-Phosphogluconate dehydrogenase	4015	4	0·1179	0·3287	0·8072	4939	2	0·0250	0·1040	0·3881
Adenylate kinase	2250	1	0·0000	—	0·3425	6760	2	0·0242	0·1007	0·3754
Adenosine deaminase	4021	2	0·0256	1·1063	0·3969	4798	3	0·0649	0·2118	0·5998
Esterase D	781	2	0·0310	0·1293	0·4852	454	2	0·0334	0·1393	0·5240
Phosphohexose isomerase	4027	6	0·2435	0·5617	1·1766	1550	2	0·0285	0·1185	0·4438
NADP-isocitrate dehydrogenase	3993	2	0·0256	0·1064	0·3972	718	3	0·0815	0·2675	0·7639
Peptidase A	4009	2	0·0256	0·1063	0·3971	8798	7	0·2878	0·6231	1·2362
Peptidase B	4028	3	0·0661	0·2159	0·6118	7041	4	0·1109	0·3089	0·7573
Acid phosphatase	2790	2	0·0266	0·1107	0·4137	7887	2	0·0239	0·0991	0·3695

† Not tested.

catalytic properties of PGM_1 and PGM_2 (Quick, Fisher & Harris, 1974). It is thus quite possible that the products associated with these two loci are subject to different selective pressures. However, while a role for selection in producing this difference cannot be excluded (in which case the calculation of θ is invalid), recourse to this explanation in the face of the present findings literally requires that all heterozygous variants of PGM_2 are quite deleterious by contrast with those of PGM_1; this seems most unlikely. Thus, the primary explanation would seem to be a difference in mutation rate at the two loci. Since $\theta = 4N_e\mu$, then the ratio $\theta_{PGM1}:\theta_{PGM2}$ is an expression of the *relative* rate of mutation at the two loci. Without a value for θ_{PGM2} in Japanese we cannot actually derive this ratio, but, employing θ_U for PGM_2 as our estimate, the data are consistent with at least a three- to fourfold difference.

Harris, Hopkinson & Robson (1974) observed in their review of rare variants in the British population an apparent bimodality in the distribution of heterozygosities for the 43 loci concerned, and pointed out the alternative explanations, namely that there are two groups of loci

Table 3. *A comparison between Hiroshima and Nagasaki with respect to the type and frequency of rare variants of three proteins*

	PGM			PGM_1		TF				
City	PHI^1	$PHI^{4\ \mathrm{HIR}\ 1}$	PHI^{Other}	$PGM_1{}^{1\ 2,7}$	$PGM_1{}^{\mathrm{Other}}$	TF^C	$TF^{\mathrm{D\ Chi}}$	$TF^{\mathrm{B\ HIR\ 2}}$	$TF^{\mathrm{B\ 3}}$	TF^{Other}
Hiroshima	5288	11	3	1638	2	5224	36	15	6	7
Nagasaki	2731	15	6	2031	13	2732	11	4	3	2
Total	8019	26	9	3669	15	7956	47	19	9	9
	$\chi^2 = 10{\cdot}67$, D.F. = 2 $P < 0{\cdot}005$			$\chi^2 = 4{\cdot}73$, D.F. = 1 $0{\cdot}025 < P < 0{\cdot}05$		$\chi^2 = 4{\cdot}53$, D.F. = 3 $0{\cdot}10 < P < 0{\cdot}25$				

as regards the intensity of the selection to which the phenotypes associated with alleles at these loci were subjected, or that the loci fell into two groups with respect to mutation rates. We did not observe such a bimodality in a much smaller body of data from Amerindians (Neel, 1978), nor is such a bimodality suggested when the present Japanese data are plotted in the same fashion. Thus, while we cannot support the findings which led them to their original alternative suggestions, we do support on a different basis the argument for locus differences in mutability.

The average rare variant frequency for all loci is 1·9/1000 person determinations. To prevent any one system from dominating the results simply because a large number of determinations have been made, we use an unweighted average. While this average is a useful index figure, clearly, because of the large differences between loci, it must be used with caution.

DIFFERENCES BETWEEN HIROSHIMA AND NAGASAKI IN TYPES AND FREQUENCY OF RARE VARIANTS

For three systems (TF, PHI and PGM_1) rare variants occur in sufficient numbers that a statistical contrast between Hiroshima and Nagasaki has seemed appropriate. The pertinent data are presented in Table 3, where the contrast is again based on an actual gene count. Because of the small numbers in some categories, certain gene counts have been combined, as indicated. For two of the systems, PHI and PGM_1, the city differences are significant. Although not apparent from the table, a noteworthy feature with respect to the PGM_1 data was that whereas six persons with a 3-type variant were encountered in Nagasaki, none was encountered in Hiroshima. In addition, whereas 18 persons with the $PHI4_{\mathrm{HIR1}}$ variant were observed in Nagasaki, there were only 11 in Hiroshima in a sample approximately twice as large.

Kirk (1975) has recently emphasized the potential usefulness of rare variants in tracing rather major population movements. Obviously, as the characterization of specific populations with respect to the rare variants is intensified, they may also be useful in unravelling minor population movements. Some of the rare variants which have a widespread but uneven distribution in specific civilized populations, and which are restricted to these populations may, during the tribal period of that population's history, have existed as 'private' polymorphisms, comparable to the case of Albumin Yanomama-2 amongst the Yanomama Amerindians (Tanis *et al.* 1974). Such traits would then be of greater value in reconstructing relatively localized population movements, as within Japan, than the common polymorphisms of world-wide distribution, whose origin may be assumed to be ancient. With respect to the transferrin variants, this line of reasoning suggests that the $TF^{\mathrm{D\ Chi}}$, the TF^{B3} and the $TF^{\mathrm{B\ HIR\ 1}}$ alleles are rather 'old' variants (certainly true for the former in view of its widespread occurrence in Mongoloid populations). The same reasoning applies to the PGM_1^7 allele, a marginal polymorphism with equal frequencies in Hiroshima and Nagasaki. This deduction, like that based on $TF^{\mathrm{D\ Chi}}$, is confirmed by the occurrence of this allele elsewhere in the Western Pacific area. On the other hand, on the basis of the present limited data, we may speculate that the $PHI^{4\ \mathrm{HIR\ 1}}$ allele arose somewhat later than the abovementioned TF variants, and the PGM_1^3 allele, much later. Further studies on the distribution of these alleles in Japan will test this speculation.

A COMPARISON OF JAPANESE AND CAUCASIANS

Although comparisons across non-homologous proteins within a population are fraught with many difficulties, comparisons between the same protein across carefully chosen populations appear to be more valid. In Table 1 we also present for Caucasoid inhabitants of the British Isles sampled in roughly the same fashion as the Japanese, data on as many of the same proteins studied in Japanese as possible. The data on enzymes are for the most part drawn from the summary paper of Harris *et al.* (1974), while the source of the data on serum proteins is indicated by the appropriate footnotes. By and large the isozyme techniques employed in Japan were patterned after those developed by Harris & colleagues. Harris *et al.* (1974) defined a rare variant as one involving less than 1 % of the population, pointing out at the same time how arbitrary any definition is. Since the usual (equally arbitrary) definition of a genetic polymorphism is a trait involving more than 2 % of the population, their definition leaves in limbo those traits affecting between 1 and 2 % of the population. In this discussion we have defined a rare variant as one affecting 2 % or less of the population, but in the preceding papers in this series the data have been presented in such a way that others can pursue their own definitions. It so happens this difference in definition will not affect our comparison. We will thus compare the findings in samples drawn in cities from two insular populations of roughly the same total population size and mobility, temperateness of climate, and geographic area. Furthermore, in the two countries the transition from a tribal society through feudalism to a mobile, highly industrialized society has occurred on a time span which from the genetic standpoint must be regarded as quite comparable. Unlike the earlier, within-population comparisons, now the amount of admixture with the surrounding populations is critical. The relative rates of immigration into the two sets of islands cannot be rigorously compared, but clearly the 'original' inhabitants of both sets of islands have experienced major waves of invasion.

The significance of the observed differences between the two populations has been evaluated by the same type of χ^2 contrast employed earlier with respect to $\mathrm{PGM_1}$ and $\mathrm{PGM_2}$ and the result is presented in the final column of Table 1. Wherever expectation of number of variants is less than 1 for either the Japanese or the British populations, no χ^2 has been calculated. There are nine such contrasts, which pending the acquisition of further data, we will set aside. In addition, where possible θ and its 95 % confidence interval have been calculated for the British data; these results are presented in parallel with the Japanese in Table 2. Of the 11 possible χ^2 contrasts, 2 are quite significant. In both cases ($\mathrm{PGM_1}$ and PHI), variant frequency is higher in Japanese. As emphasized previously the $\mathrm{PGM_1}$ system exceeds all others dealt with in these papers as regards apparent dependence on subtle factors for the demonstration of variants

(Satoh *et al.* 1977); the possibility that technical factors play some role in this apparent difference must be kept in mind. That PGM_1 might exhibit greater variability in Japanese than in British was first suggested by Shinoda & Matsunaga (1970), but the possibility of the other difference has not been previously recognized. With respect to the contrast based on θ, however, there is no instance among the 23 possible contrasts in which the upper confidence limit for θ in one of these two populations fails to overlap with the lower confidence limit for the other. The conservative course is to regard the differences between the PGM_1 and the *PHI* loci in the two populations as not yet firmly established, but a possibility to be pursued.

DISCUSSION

The primary purpose of this series of five papers has been to develop a data base for the frequency of 'rare' variants (those not achieving the frequency of a genetic polymorphism) in the Japanese population, for a series of 25 polypeptides. It has been shown that the average frequency is 1·9/1000 determinations. For 23 of these polypeptides, comparable data are available for British Caucasoids (cf. Harris *et al.* 1974). For these, the frequency of rare variants is 1·6/1000 in the British population and 2·0 in the Japanese.

Two specific questions concerning this variation have been raised in the present paper: (1) within the Japanese population, do three sets of proteins, the structural genes for which presumably arose through a genetic duplication, exhibit comparable patterns of rare variants, and (2) in a locus-by-locus contrast of Japanese and British with respect to rare variants in 23 polypeptides, are there significant differences between the populations? In either case, if differences do exist, how are they to be interpreted? Unfortunately, even with the large body of data available, it is clear that the present analyses are more notable for the questions they raise than for the answers they provide. On the other hand, there is every reason to expect the accumulation of substantially more data in the near future. Not only will studies projected by ourselves and others contribute to data of the type presently available but technical developments will greatly facilitate the necessary task of subdividing electrophoretic classes of variants into subtypes based on kinetics and physical properties.

With respect to the first of these questions, the most noteworthy observation is of a greater frequency of variants of PGM_1 (7·4/1000) than PGM_2 (0/1000) in the Japanese, an observation most readily explained by a difference in mutation rate at the two loci. By contrast with data on British Caucasoids, PGM_1 in Japanese shows a greater frequency of all variants and relatively more different variants, whereas PGM_2 exhibits a lower frequency of variants (none, to be exact). Assuming this difference to be primarily mutational in origin, the data do not yet permit one to conclude whether the difference is due to a relatively low mutation rate at the PGM_2 locus, a relatively high rate at the PGM_1 locus, or some combination of these possibilities.

With respect to the second question, a locus-by-locus contrast across populations for the 11 proteins where a χ^2-type contrast is possible, suggests that with respect to total number of variants, PGM_1 and PHI variants are significantly more common in Japanese. However, when the contrast is based on number of electrophoretically distinguishable variants, by use of the θ statistic of Ewens (1972), now among the 23 possible contrasts there is no single difference for which the 95 % confidence intervals do not overlap. It is noteworthy that the somewhat greater average frequency of variants in Japanese than in British is entirely accounted for by the PGM_1

and PHI systems; if these are eliminated from consideration, then the average frequency for the 21 polypeptides for which data are available from the two populations is 1·5/1000 determinations in Japanese and 1·7 in British.

Earlier we have pointed out the broad historical parallelism between the inhabitants of the Japanese Islands and the British Isles, and we shall take the position that the likelihood of random loss of a mutation has been very similar in these two populations from prehistoric times. This is tantamount to the postulate that N_e has been quite similar in the two populations. The probability of introduction of mutants from the outside may be assumed to be about the same for the two populations. With respect to selection, given the generally low frequency of these variants, selection would have to be exercised against the heterozygote. Imperfect though our knowledge of selection is, it is difficult to visualize important differences between these two populations as regards selection against heterozygotes for enzyme variants. If these postulates are accepted, and if we also assume that the departures from equilibrium within these two populations have affected variant frequency equally, then, by the reasoning developed earlier, a contrast of number of variants or of θ across these two populations can be seen as a very approximate test of similarity in mutation rate at corresponding loci.

Admittedly, each of the two tests employed (χ^2, θ) is biased, but the biases are in opposite directions. With respect to comparative mutation rates, the question we are attempting to answer is whether the number of different variants at a given locus is the same for the two populations. The χ^2 test in this context carries the implicit assumption that each of the variants encountered is independent in its origin from the others, almost certainly untrue. On the other hand, the θ contrast assumes that each electromorph is a homogeneous entity, and this too is almost certainly untrue. Although the theory of population genetics permits a calculation of the number of different alleles with the same electrophoretic behaviour (Chakraborty & Nei, 1976; Nei & Chakraborty, 1976), these calculations demand a knowledge of N_e and the assumption of genetic equilibrium. For civilized populations such as we are considering here, the assumption is clearly violated, in part because of the relatively recent fusion of tribal populations, which latter development in any event makes an estimate of present-day N_e of little value. Further refinement in contrasts of this type calls for studies on the thermostability and kinetic properties of the independently ascertained variants comprising a given electromorph, as a means of detecting heterogeneity. For certain proteins in the present battery – haemoglobin, albumin, carbonic anhydrase – it should also be feasible to proceed to 'fingerprinting' and amino acid sequence studies where indicated. These refinements in characterization are projected for the Japanese material. Pending the completion of such studies, we suggest only that whereas there is no reason to suspect differences in mutation rates between those corresponding loci in the two populations where neither the χ^2 or θ contrast yields significance, for two loci – PGM_1 and PHI – we may be suspicious that differences exist.

With reference to the material on British Caucasoids, it is of interest to examine with respect to total number of variants and θ the three pairs of homologous proteins previously contrasted for the Japanese. A comparison of HGB A and A_2 is not possible because of the lack of data on A_2. The data on CA II are so scanty as to render a contrast with CA I of little value. However, there are abundant data for a contrast of PGM_1 and PGM_2. Although the overall frequency of variants and the value of θ is again larger for PGM_1 than PGM_2, the pronounced difference present in the Japanese is not observed.

While analyses of the type just presented may ultimately permit tentative conclusions concerning relative mutation rates, they will not contribute in any important way to our knowledge of absolute rates. For the indirect approach to the latter in our material – which would be necessary in this instance because of the inability to examine both parents of affected propositi (see above) – it would be necessary to estimate the number of individuals in the population originally supporting the variants, the average numbers of variants per locus, and mean mutant survival time, and in addition reach some conclusions concerning the number of populations whose amalgamation resulted in the present population. While the necessary estimates of the first-mentioned parameters are possible for tribal populations, and have been utilized for indirect estimates of mutation rates of traits of this type (cf. Neel, 1973), such data are completely lacking for the populations under discussion. Furthermore, in view of the fact that many of the variants being detected have arisen many generations in the past, even direct estimates based on the current generations cannot with certainty be applied to the past, so that failure to detect differences between Japanese and Caucasoid populations in present mutation rates would not invalidate the thesis that the findings reflect past differences. Unfortunately, the collection of data for direct estimates of mutation at the protein level (the more desirable proceduie) is sufficiently laborious that it will be many years before estimates for individual loci become available; in the meantime considerations such as the foregoing will remain pertinent to our efforts to understand locus differences in mutability.

SUMMARY

The frequencies in Hiroshima and Nagasaki of rare variants (represented in less than 2% of the individuals surveyed) is summarized for a series of 22 proteins (25 polypeptides). The average number of persons examined for each protein was 3312. There are three pairs of homologous proteins in the series: PGM_1 and PGM_2, CA I and CA II, and HGB A and A_2. Only for the first pair is there a significant difference between the two in the total frequency and number of different kinds of variants; it is suggested this may reflect differences in the mutation rates of the corresponding structural genes. For 23 of these polypeptides, comparable data are available for British Caucasians. The average frequency of variants for loci in common in the two series is 2·0/1000 person determinations for Japanese and 1·6/1000 for Caucasoids. At two loci (PGM_1 and *PHI*) there were significantly more variants in Japanese than in British; these two loci account for the greater average frequency of variants in Japanese. However, a conservative comparison of number of *different* variants (electromorphs) encountered, using the θ statistic of Ewens (1972), yields no significant difference for any of the 22 possible contrasts. The potential usefulness of data of this type in reaching conclusions regarding comparability of mutation rates in two populations is discussed. For the present, the fact that one electromorph may shelter multiple different amino acid substitutions in a protein limits the inferences to be drawn from such contrasts.

It is a pleasure to acknowledge our indebtedness to Dr Peter Smouse and Dr Warren Ewens for statistical consultations, but responsibility for the tentative interpretations is entirely our own.

REFERENCES

ALLISON, A. C., BLUMBERG, B. S. & AP REES, W. (1958). Haptoglobin types in British, Spanish Basque and Nigerian African populations. *Nature, Lond.* **181**, 824.

BELSKY, J., TACHIKAWA, K. & JABLON, S. (1971). ABCC-JNIH Adult Health Study Report 5: Results of the first five examination cycles, 1958–1968. Atomic Bomb Casualty Commission Technical Report 9–71.

BLAKE, N. M. & OMOTO, K. (1975). Phosphoglucomutase types in the Asian-Pacific area: a critical review including new phenotypes. *Ann. Hum. Genet., Lond.* **38**, 251.

CARTER, N. D., TASHIAN, R. E., HUNTSMAN, R. G. & SACKER, L. (1972). Characterization of two new variants of red cell carbonic anhydrase in the British population: CA Ie Portsmouth and CA Ie Hull. *Am. J. Hum. Genet.* **24**, 330.

CARTER, N. D., TANIS, R. J., TASHIAN, R. E. & FERRELL, R. E. (1973). Characterization of a new variant of human red cell carbonic anhydrase I, CA If London (Glu-102 → Lys). *Biochem. Genet.* **10**, 399.

CHAKRABORTY, R. & NEI, M. (1976). Hidden genetic variability within electromorphs in finite populations. *Genetics* **84**, 385.

COHEN, B. L. (1965). Paucity of albumin variants. *Nature, Lond.* **207**, 1109.

COOKE, K. B., CLEGHORN, T. E. & LOCKEY, E. (1961). Two new families with bisalbuminaemia: An example of possible links with other genetically controlled variants. *Biochem. J.* **81**, 39*P*.

EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **3**, 87.

FERRELL, R. E., UEDA, N., SATOH, C., TANIS, R. J., NEEL, J. V., HAMILTON, H. & BABA, K. (1977). The frequency in Japanese of genetic variants of 22 proteins. I. Albumin, ceruloplasmin, haptoglobin and transferrin. *Ann. Hum. Genet., Lond.* **40**, 407.

HARRIS, H. (1959). *Human Biochemical Genetics.* Cambridge University Press.

HARRIS, H. (1975). *The Principles of Human Biochemical Genetics*, 2nd ed. Amsterdam: North-Holland.

HARRIS, H., PENINGTON, D. C., ROBSON, E. B. & SCRIVER, C. R. (1960). A further genetically determined transferrin variant in man. *Ann. Hum. Genet., Lond.* **24**, 327.

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., Lond.* **37**, 237.

HARRIS, H., ROBSON, E. B. & SINISCALCO, M. (1959). Genetics of the plasma protein variants. In *Biochemistry of Human Genetics* (ed. G. E. W. Wolstenholme and C. M. O'Connor), p. 151. London: J. and A. Churchill.

HOPKINSON, D. A. & HARRIS, H. (1969). Red cell acid phosphatase, phosphoglucomutase, and adenylate kinase. In *Biochemical Methods in Red Cell Genetics* (ed. J. J. Yunis), pp. 337–375. New York: Academic Press.

HOPKINSON, D. A., COPPOCK, J. S., MÜHLEMANN, M. F. & EDWARDS, Y. H. (1974). The detection and differentiation of the products of the human carbonic anhydrase loci, CA I and CA II, using fluorogenic substrates. *Ann. Hum. Genet., Lond.* **38**, 155.

INGRAM, V. M. (1961). Gene evolution and the haemoglobins. *Nature, Lond.* **189**, 704.

JONES, R. T. & KOLER, R. D. (1975). A proposal for reporting and recording of studies of abnormal hemoglobins. In *Abnormal hemoglobins and thalassaemia* (R. M. Schmidt, ed.). pp. 311–66.

KIMURA, M. & OHTA, T. (1971). *Theoretical Aspects of Population Genetics.* Princeton: Princeton University Press.

KIRK, R. L. (1975). Isozyme variants as markers of population movement in man. In *Isozymes.* Vol. IV. *Genetics and Evolution* (ed. C. L. Markert), pp. 169–180. New York: Academic Press.

NEEL, J. V. (1973). 'Private' genetic variants and the frequency of mutation among South American Indians. *Proc. Natn. Acad. Sci. U.S.A.* **70**, 3311.

NEEL, J. V. (197–). The circumstances of human evolution. *Bull. Johns Hopkins Hosp.* (in the Press.)

NEI, M. & CHAKRABORTY, R. (1976). Electrophoretically silent alleles in a finite population. *J. Mol. Evol.* **8**, 381.

QUICK, C. B., FISHER, R. A. & HARRIS, H. (1974). A kinetic study of the isozymes determined by the three human phosphoglucomutase loci, PGM_1, PGM_2, and PGM_3. *Europ. J. Biochem.* **42**, 511.

SATOH, C., FERRELL, R. E., TANIS, R. J., UEDA, N., KISHIMOTO, S., NEEL, J. V., HAMILTON, H. B. & BABA, K. (1977). The frequency in Japanese of genetic variants of 22 proteins. III. Phosphoglucomutase-1, phosphoglucomutase-2, 6-phosphogluconate dehydrogenase, adenylate kinase, adenosine deaminase. *Ann. Hum. Genet., Lond.* **41**, 169.

SHINODA, T. & MATSUNAGA, E. (1970). Studies on polymorphic types of several red cell enzymes in a Japanese population. *Jap. J. Hum. Genet.* **15**, 133.

TANIS, R., FERRELL, R. E., NEEL, J. V. & MORROW, M. (1974). Albumin Yanomama-2, a 'private' polymorphism of serum albumin. *Ann. Hum. Genet., Lond.* **38**, 179.

TANIS, R. J., UEDA, N., SATOH, C., FERRELL, R. E., KISHIMOTO, S., NEEL, J. V., HAMILTON, H. B. & OHNO, N. (1978). The frequency in Japanese of genetic variants of 22 proteins. IV. Acid phosphatase, NADP-isocitrate dehydrogenase, peptidase A, peptidase B, and phosphohexose isomerase. *Ann. Hum. Genet., Lond.* **41**, 419.

TASHIAN, R. E. (1969). The esterase and carbonic anhydrases of human erythrocytes. In *Biochemical Methods in Red Cell Genetics* (ed. J. J. YUNIS), pp. 307. New York: Academic Press.

TASHIAN, R. E. & CARTER, N. D. (1976). Biochemical genetics of carbonic anhydrase. In *Advances in Human Genetics* (ed. K. Hirschhorn and H. Harris). New York: Plenum Press. pp. 1-55.

UEDA, N., SATOH, C., TANIS, R., FERRELL, R., KISHIMOTO, S., NEEL, J., HAMILTON, H. & BABA, K. The frequency in Japanese of genetic variants of 22 proteins. II. Carbonic anhydrase I and II, lactate dehydrogenase, malate dehydrogenase, nucleoside phosphorylase, triose phosphate isomerase, haemoglobin A_1 and haemoglobin A_2. *Ann. Hum. Genet., Lond.* **41**, 43-52.

YAMASHITA, J., KIMURA, Y. & SATOH, C. Polymorphism of esterase D among residents of Hiroshima and Nagasaki (personal communication).