Differences in the Relative Distribution of Human Gene Diversity between Electrophoretic and Red and White Cell Antigen Loci

Nils Ryman\textsuperscript{a}, Ranajit Chakraborty\textsuperscript{b}, Masatoshi Nei\textsuperscript{b}

\textsuperscript{a}Department of Genetics, University of Stockholm; \textsuperscript{b}Center for Demographic and Population Genetics, Health Science Center at Houston, The University of Texas, Houston, Tex., USA

Key Words. Human genetics \cdot Gene diversity \cdot Blood groups \cdot Electrophoresis

Abstract. Gene frequency data for 25 loci (2 HLA loci, 9 blood group loci, and 14 electrophoretically detectable loci) were collected from the literature of 18 human populations from all over the world. The data were subjected to a hierarchical gene diversity analysis to provide an estimate of the relative distribution of genetic variation between and within populations and population groups for different types of loci. Two different ways of grouping the populations, i.e., according to anthropological criteria and to a cluster analysis based on gene frequency data, gave essentially the same results. For all loci combined approximately 86\% of total gene diversity was found within populations, 3\% was associated with differences between populations within groups, and 11\% related to group differences. These results are very similar to those obtained in previous studies based on fewer loci and different sets of populations. The distribution of genetic variation is different for different types of loci. The HLA loci give a picture very similar to that of the electrophoretic loci while the blood group loci have a substantially larger fraction of the total gene diversity distributed between populations or population groups.

Introduction

The development of efficient electrophoretic and other biochemical techniques has dramatically increased the potential for genetic analysis of natural populations during the last decade. Recent electrophoretic studies have revealed the existence of previously unknown groupings in a variety of organisms such as polychaetes [Grassle and Grassle, 1976], fishes [e.g., Shaklee and Tamaru, 1977; Ryman et al., 1979; Ferguson and Mason, 1981] and large mammals [Ryman et al., 1980]. Recently, attention has been focused on the patterns of the distribution of genetic variability at various systematic levels in very different disciplines of theoretical and ap-
plied biology. Detection of new genetic structurings calls for continuous revision of previous estimates of the distribution of genetic variability. With regard to the relative distribution of genetic variability among taxa there appears to be considerable differences between different groups of organisms [e.g., Avise et al., 1980], and also to some extent between loci coding for different kinds of proteins, e.g., between electrophoretically detectable protein loci and blood group loci in man [Nei and Roychoudhury, 1974, 1982]. Historically, different kinds of loci were studied in different organisms like blood groups in man and electrophoretically detectable soluble proteins in natural populations of other organisms like Drosophila and fishes, potentially resulting in a biased bias for interspecific comparisons. It is crucial for a correct understanding of evolutionary relationships across a wide range of taxa that comparisons of the distribution of genetic variability are based on comparable sets of data.

In the present paper we report the result of a hierarchical analysis of gene diversity between and within 18 human populations representing different races from all over the world. The study deals with the variation at 25 electrophoretically detectable protein and blood group loci (including the HLA A and B loci). The analysis is essentially based on the results and the data presented by Nei and Roychoudhury [1982] in their extensive study of genetic relationships and evolution of human races.

There are three major purposes of the present study. The first is to compare the variability pattern for electrophoretic loci (which are most frequently analyzed in organisms other than man) with that obtained for the blood group and HLA loci. In particular, we are interested in knowing whether the pattern for the extremely variable HLA loci is more similar to that of the blood group loci (which can be detected only if they are polymorphic) than to that of the electrophoretic loci, which do not require polymorphism for detection. Second, we want to test how a regrouping of the populations, based on the genetic relationships, affect the estimates of distribution of genetic variation at different hierarchical levels. Third, we would like to set a standard for the distribution of the relative amount of genetic variability at different hierarchical levels within a well-known species to serve as a basis for comparisons with data obtained from other species.

Material and Methods

Except for the HLA data our study is entirely based on gene frequency data that were collected by Drs. M. Nei and A. Roychoudhury. These investigators gathered a large body of gene frequency data from the literature for an extensive study of genetic relationships among human races [Nei and Roychoudhury, 1982], and the collection procedures and the origin of data are described in detail in their paper. The populations chosen for the present study were 18 representative races from all over the world selected on the basis of availability of gene frequency data, i.e., Lapp, English, Italian, Iranian, northern Indian, Malay, Chinese, Japanese, Polynesian, Micronesian, South American Indian, Eskimo, North American Indian, Australian Aborigine, Papuan, Nigerian, Bantu, and Bushman (table 11 in Nei and Roychoudhury, 1982). The gene frequency data at the 23 loci collected by Nei and Roychoudhury referred to nine blood group loci, i.e., A, ABO, MNs, Rh, Duffy, Diego, Kell, Kidd, P, secretor and 14 electrophoretic loci, i.e., HbA, Alb, Hp, Tf, Cg, ACP, AK, ADA.
Gene (haplotype) frequency data at the HLA A and B loci for the above populations (or closely related ones) were also collected from the literature. As the data were from several different laboratories and some of the reports were relatively old we chose a rather conservative strategy when identifying alleles from different studies: several alleles had to be lumped into the 'blank' category to permit comparisons across populations. The haplotypes were classified into eight 'allelic' groups at the HLA A locus (A1, A2, A3, A9, A10, A11, A28, and blanks) and into 13 ones at HLA B (B5, B7, B8, B12, B13, B14, B27, BW35, BW40, BW15, BW17, BW22, and blank).

In a few cases we were not able to find HLA data for exactly those populations analyzed for the other loci; in those cases the HLA data were taken from populations judged to be as closely related as possible. In particular, the HLA data for North American Indians were from the Chippewa tribe [Noves et al., 1975] while the other gene frequencies were from the Athabaskans. For the rest of the populations the data sources are as follows (population specification and reference in parenthesis): Lapp [Ttilkinen et al., 1973], English [Dick et al., 1973], Italian [Bari, Laurentian and Bisquard, 1976], data submitted to the HLA and Disease Registry - ref. 41 of Ryder et al., 1978], Iranian [Zoroastrians, Mohagheghpour et al., 1981], northern Indian [Singal, 1973], Malay [Jorsey et al., 1973], Chinese [Panze et al., 1973a], Japanese [Mizutani et al., 1973], Polynesian [Samoa, Kirk, 1979], Micronesian [Panze et al., 1973b], South American Indian [Cayapo: Black et al., 1980], Eskimo [MacKenzie Delta, Dossier et al., 1973], Australian Aborigine [Bulmer et al., 1973], Papuan [Morris et al., 1973], Nigerian [Twareg: Colombo et al., 1973], Bantu [Tonga: Fertigstein et al., 1973], and Bushman [Sun: Bush et al., 1973].

The populations were grouped in two different ways on the basis of anthropological criteria and a dendrogram clustering based on gene frequency data, respectively. There is no unambiguous anthropological classification of human populations [e.g. Bond, 1963; Chon, 1965; Nei and Ranchoudhury, 1982], for the purpose of the present analysis we chose six different groups for the anthropological classification (fig. 1); i.e., Caucasoid (Lapp, English, Italian, Iranian, northern Indian), Mongoloid (Chinese, Japanese, Malay), Negroid (Bantu, Bushman, Nigerian), Amerind (Eskimo, North American Indian, South American Indian, Oceanian (Micronesian, Polynesian), and Austrailian (Australian Aborigine, Papuan). The genetic classification was based on the dendrogram constructed from genetic distance values [of Nei, 1975] on the basis of the variation at the 25 loci analyzed in the present study using the unweighted pair group method (UPGMA) of Sneath and Sokal, 1973]. The cut-off point for dividing the populations into groups (the dotted line in figure 1 resulting in seven groups) was somewhat arbitrarily chosen with the double purpose of (i) locating it in the approximate center of the dendrogram (or D scale), and (ii) not to provide an unnaturally coarse grouping, e.g., through putting the Caucasoids and Mongoloids in the same major group.

The hierarchical gene diversity analysis follows the logic of Nei, 1973, 1975. The total gene diversity.
\( H_T \) is partitioned according to the causal components so that:

\[
H_T = H_p + D_m + D_{1:T}
\]

where \( H_p \) is average gene diversity within populations, \( D_m \) is gene diversity between populations within groups, and \( D_{1:T} \) corresponds to diversity between groups. The relative importance of the various components are expressed in terms of \( G \) values (e.g., \( G_{m} = D_m / H_p \) is the relative contribution to the total gene diversity arising from differences between populations within groups). The computations were performed using a Fortran IV computer program [Chakraborty et al., 1982] that is based on the algorithm presented by Chakraborty [1980].

**Results and Conclusions**

The results are summarized in table I. In accordance with previous studies on genetic variability among human populations [e.g., Lewontin, 1972; Nei and Roychoudhury, 1972, 1974, 1982; Latter, 1980] the major part of the total variability was found within populations and only a small fraction, i.e., 10-20%, between populations or population groups. In the recent and most extensive study on this topic Nei and Roychoudhury [1982] presented a formal gene diversity analysis of genetic variation between and within the three major races of man (Caucasoid, Mongoloid, and Negroid) based on 85 blood group and electrophoretic loci, 59 of which were polymorphic. They found that, on the average, 10-11% of the total gene diversity was attributable to differences between the major races and approximately 90% to variation within races. It is interesting how that estimate coincides with the present one for all loci despite the fact that a much finer grouping of populations was used in the present study.

In agreement with the above observation the two different methods for grouping populations employed in the present study give very similar results. In all cases the regrouping results in a rearrangement of the relative importance of the two sources of variation that is less than 2%. Of course, the major reason for this is that our primary (although somewhat arbitrary) anthropological grouping to a great extent coincides with the dendrogram grouping. It is an interesting observation, though, that the splitting of for instance the genetically very distinct Australian Aborigine and Papuan populations into two different groups has such a small effect on the relative distribution of gene diversity between and within groups of populations. At any rate, as the two groupings give so similar results we shall confine the subsequent discussion to the dendrogram grouping which we consider the most appropriate basis for the analysis of gene diversity.

The most relevant studies for comparison with our present results are those by Lewontin [1972] and Latter [1980]. Using the Shannon information index Lewontin analyzed the distribution of genetic variation at 17 protein and blood group loci among major racial groups, each of which consisted of a number of populations. It has been pointed out [Nei, 1973, 1975] that the Shannon information index does not relate to any defined biological quantity. Latter [1980] performed a similar analysis but used another statistic and calculated the proportion of genes expected to be different in randomly chosen pairs of individuals at various levels of relationship; the estimates were based on 18 protein and blood group loci. Nevertheless, Lewontin's
Table I. Absolute ($H_r$) and relative ($G$) gene diversity for 18 human populations grouped on the basis of anthropological and genetic evidence, respectively; the estimates are based on HLA A and B, 9 blood group loci, and 14 protein loci (standard errors in parenthesis)

<table>
<thead>
<tr>
<th></th>
<th>Total gene diversity ($H_r$)</th>
<th>Relative distribution, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>within populations ($G_p$)</td>
</tr>
<tr>
<td>Blood group loci (9 loci)</td>
<td>0.4170 (0.0772)</td>
<td>82.7 (3.9)</td>
</tr>
<tr>
<td>Anthropological</td>
<td></td>
<td>82.7 (3.9)</td>
</tr>
<tr>
<td>Genetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electrophoretic loci (14 loci)</td>
<td>0.1683 (0.0478)</td>
<td>90.3 (1.5)</td>
</tr>
<tr>
<td>Anthropological</td>
<td></td>
<td>90.3 (1.5)</td>
</tr>
<tr>
<td>Genetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA (2 loci)</td>
<td>0.8497 (0.0357)</td>
<td>87.4 (1.9)</td>
</tr>
<tr>
<td>Anthropological</td>
<td></td>
<td>87.4 (1.9)</td>
</tr>
<tr>
<td>Genetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All loci except HLA (23 loci)</td>
<td>0.2656 (0.0484)</td>
<td>85.6 (2.6)</td>
</tr>
<tr>
<td>Anthropological</td>
<td></td>
<td>85.6 (2.6)</td>
</tr>
<tr>
<td>Genetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All loci (25 loci)</td>
<td>0.3124 (0.0550)</td>
<td>86.0 (2.1)</td>
</tr>
<tr>
<td>Anthropological</td>
<td></td>
<td>86.0 (2.1)</td>
</tr>
<tr>
<td>Genetic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and Latter's results both agree almost perfectly with the present ones which are based on a larger number of loci and a different set of populations. They estimated approximately 85% of the total genetic variation to be attributable to differences within populations and only 15% was explained by differentiation between populations and racial groups. This is almost exactly the same result as obtained in the present study when all loci are considered. The only difference refers to the relative importance of the component associated with differences between racial groups ($G_{GT}$) and between populations within groups ($G_{PG}$). In our present study approximately 10–12% of the variation is found between groups and only 2–4% between populations within groups, while Lewontin [1972] and Latter [1980] both estimated these two sources of variation to be of approximately equal importance.

The most important result of the present study refers to the different variability patterns obtained for different types of loci. As compared to the electrophoretic and HLA loci, the blood group loci have a conspicuously larger fraction of the total
gene diversity distributed among population groups while the relative importance of the two other components is correspondingly smaller. The difference is not statistically significant for the \( G_P \) and \( G_{GT} \) components, but there is a significant heterogeneity among the \( G_{PG} \) values obtained for blood group loci, electrophoretic loci, and the HLA loci (\( \chi^2 = 11.0 \) with \( p < 0.01 \) and \( \chi^2 = 32.6 \) with \( p < 0.001 \) for the anthropological and dendrogram groupings, respectively, using the test procedure from Rao [1973], p. 323). Further, there are no indications that this difference is caused by the variability pattern at any particular blood group locus. For instance, excluding the Duffy system results in no substantial reduction of the \( G_{PG} \) and \( G_{GT} \) blood group components.

It is an open question to what degree the variability pattern at various types of loci reflects that of the whole genome [e.g., Gillespie and Kojima, 1968; Kojima et al., 1970; King and Wilson, 1975; Nei, 1975; Ward, 1977; Nevo, 1978; Walton et al., 1979; Brown and Langley, 1979]. It has been argued [Nei and Roychoudhury, 1982] that for the analysis of gene diversity it is more appropriate to study electrophoretically detectable proteins than red or white cell antigens. There is a direct relation between nucleotide differences at the codon level and electrophoretically detectable protein variation while the relationship between codon differences and antigen variation is more complicated and not well understood. Further, loci determining red and white cell antigens constitute a selected part of the genome as they can be detected only if they are polymorphic while electrophoresis provides a tool for assaying polymorphic as well as monomorphic loci. The present results indicate that there is no clear-cut difference between loci coding for electrophoretically detectable proteins on the one hand and antigens on the other. First, there appears to be no obvious association between the amount of variability, measured as total gene diversity (\( H_T \)) or number of alleles, and the relative distribution of genetic variation. The loci in the HLA series are more variable than any other locus described in man. Further, the relation between codon differences and biochemically detectable variation is by no means better understood for HLA than for the red cell antigens. Nevertheless, the distribution pattern at HLA appears much more similar to that of the electrophoretic loci than to the blood group loci.

Differences between types of loci with regard to the distribution of genetic variability is an indicator of different selective forces operating on those types. Bodmer et al. [1973] discussed how disruptive selection would tend to increase the \( G_{PG} \) and \( G_{GT} \) components while stabilizing selection would result in a corresponding decrease as compared to a group of selectively neutral loci. The present results do not indicate any clear-cut difference between electrophoretic loci on the one hand and antigen loci on the other. Of the loci analyzed in the present study HLA A and B are those which are most commonly cited as an example of a selectively controlled polymorphism in man [e.g., Cavalli-Sforza and Bodmer, 1971; Bodmer et al., 1973; Vogel and Motulsky, 1979]. There is considerably more controversy about the determinants of red cell antigen variability [Yamazaki and Maruyama, 1974; Nei, 1975], while the neutral model appears
most appropriate for the majority of electrophoretically detectable loci (Fuersi et al., 1977; Chakraborty et al., 1978, 1980). Of course, there are no indications that possible selective forces operating at different loci would all be either disruptive or stabilizing, but it is an interesting observation that the HLA loci give approximately the same pattern as the electrophoretic loci while the blood group loci are the outliers.

Another contrast that should be mentioned in the present context concerns the differences with regard to dominance between blood group loci on the one hand and electrophoretic and HLA loci on the other. Most loci of the former group show dominance relationship whereas alleles at loci of the latter group are generally codominant. This difference may reflect dissimilar patterns for production of new alleles in the two groups. The difference with regard to the relative distribution of gene diversity could be due to such a dissimilarity.

It has been suggested that the variation of the highly polymorphic HLA system would contain so much information, e.g., with regard to phylogenetic relationships (Piazza and Viganotti, 1973), that it might be as well suited for evolutionary studies as a combination of several other systems. The present data do not support that concept. As illustrated in figure 2, the dendrogram constructed from the HLA A and B loci alone gives a very skewed picture of the generally accepted evolutionary pathways connecting the present populations. A dendrogram should be constructed from a large number of loci (Tateno, 1978; Nei, 1981) to provide an accurate picture of the evolutionary relationships, and the large number of alleles at the HLA system does not change this basic requirement.

In conclusion, it is clear that different types of loci appear to picture the distribution of genetic variation somewhat differently. Particularly, the blood groups differ from the electrophoretic and HLA loci, and this is an important consideration when comparing results from different groups of organisms. With the exception of man, the vast majority of loci examined in population genetic studies of various organisms have up till now been electro-
phoretically detectable loci. As mentioned above the great applicability of multiple locus studies is becoming increasingly acknowledged in very diverse biological disciplines such as systematics, conservation biology, and fish and wildlife management. Evaluation of data includes the comparison with results obtained in similar studies in other organisms. It is obviously imperative that such comparisons refer to similar sets of loci as to not severely bias the conclusions. We suggest that human data of the present kind should be used as a standard for comparisons among species whenever possible, as man will most likely continue to be the most well-known organism studied in great detail at large numbers of different classes of loci over a wide range of ethnical groups and environmental conditions.

Acknowledgments

This work was performed during leaves when 2 of the authors (R.C. and N.R.) were on mutual visits to the other’s home department. The investigation was supported by grants from the Swedish Natural Science Research Council. Support was also obtained from the US Public Health Service grants CA 19311 and GM 20239 from the National Institutes of Health, USA.

References


Cavalli-Sforza, L.L.; Bodmer, W.F.: The genetics of human populations (Freeman, San Francisco 1971).


Dick, H.M.; Crichton, W.B.; Ferguson-Smith, M.A.; Izatt, M.M.: Study of the HL-A and other poly-


Nei, M.: Molecular population genetics and evolution (North Holland, Amsterdam 1975).


Tateno, Y.: Statistical studies on the evolutionary changes of macromolecules; thesis (University of Texas at Houston, Houston 1978).


Nils Ryman,
Department of Genetics,
University of Stockholm,
S-106 91 Stockholm (Sweden)