Proc. Natl. Acad. Sci. USA
Vol. 78, No. 4, pp. 2638–2642, April 1981
Population Biology

Synthetic gene frequency maps of man and selective effects of climate

(population genetics/evolution/migration/multivariate analysis/ecology)

A. PIAZZA*, P. MENOZZI†, AND L. L. CAVALLI-SFORZA

Department of Genetics, Stanford University School of Medicine, Stanford, California 94305

Contributed by L. L. Cavalli-Sforza, January 9, 1981

ABSTRACT **The world distribution of 39 independent gene frequencies in human populations is analyzed by multivariate techniques and synthetic geographic maps. Most genetic variation is associated with longitude, with South Asia showing a tendency to be central. Also latitude and, more particularly, distance from the equator play a significant role in a way that suggests that climatic factors exercise selective pressures, especially for certain genes.**

The study of evolutionary factors that affect all genes jointly, such as drift, migration, and admixture processes, will obviously gain from the simultaneous consideration of many genes and, ideally, of all alleles known at all loci. We have used multivariate techniques to generate synthetic variables from gene frequencies (such as principal component and discriminant function scores), which were then plotted as contoured geographic maps. We first applied such techniques to frequencies of 34 genes in Europe and the Near East (1, 2).

The main aim was that of testing the hypothesis that the spread of early farming from the Near East in the Neolithic period was largely determined by the spread of farmers (3, 4). The geographic distribution of the first principal component in Europe was found to be in agreement with the idea that a similar process of gene flow is the major contributor to the geography of European genes. The next two principal components also could be explained as independent migratory fluxes. The second principal component probably represents prehistoric and historic migrations from North and Central Asia into Europe. It was suggested to the authors that the third principal component might represent the spread of Indo-European speaking people beginning some 5000 years ago (5). It may be confounded with the much later barbarian invasions at the end of the Roman Empire which had a somewhat similar geographic origin.

As with most reconstructions of historical events, these associations are inevitably tentative, at least at this stage. To qualify as a candidate, a migratory flux must involve a reasonably large ratio of immigrants to prior occupants of the area. This ratio must be greater the longer the time elapsed, because local migratory exchange after the major initial immigration will tend to even out existing gradients and will wipe them out entirely if given sufficient time. A kinematic analysis of the spread of early farmers in Europe in Neolithic times is given in ref. 6.

In this paper we extend the approach to aboriginal populations of the whole world and also present evidence for possible effects of natural selection. Statistical analysis of part of the same material used here from another point of view is in ref. 7.

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

Building world geographic maps of principal component scores

Using techniques described elsewhere (2), we have prepared geographic maps of 39 independent alleles from 10 loci. These include blood groups ABO (alleles A_1, B, O), MN (MS, Ms, Ns), Rh (CDe, Cde, cDE, cDe, cdE, cde), Lewis (Le^a), Duffy (Fy^a); haptoglobin (Hp1), acid phosphatase (P^a, P^b), and phosphoglucomutase (PGM1); and HLA-A (A1, A2, A3, A9, A10, A11, A28) and HLA-B (B5, B7, B8, B12, B13, B14, B15, B17, B18, B21, B22, B27, B35, B40). The non-HLA data were obtained from Mourant *et al.* (8). The HLA data from Europe (2) have been incorporated in a world data bank,‡ and a total of 116 samples have been used for the present analysis. Only aboriginal populations were considered.

The observed gene frequency data available vary considerably in the number and locations of populations studied for each genetic locus. For computing synthetic variables, it is necessary to have a full matrix of locations and gene frequency for all alleles (at all loci). To this aim we interpolated from the observed data the values of gene frequencies at 300 points (nodes) of a grid with steps of 12° latitude (between −56° N and 76° N) and of 16° longitude. The interpolation algorithm has been described (2). The choice of steps was dictated by graphical considerations. Of the 300 nodes, 159 fell on or near land. Principal component scores were computed from these 159 interpolated gene frequencies and displayed on a finer grid of points, whose values were calculated by using splines in tension. The first, second, and third principal components explained, respectively, 29%, 17%, and 10% of the total variation between nodes. In Table 1 we represent the principal component scores by continent (or part of it), with values from 1 to 10 that correspond to 10 classes of equal width spanning the whole range of actual values. A trichromic geographic map that blends the three component scores, each corresponding to a different color (first, green; second, blue; third, red), gives a synthetic representation of the world variation in human gene frequencies (Fig. 1).

Comments on the map

Genetic Centrality of Central and South Asia. Extreme values of the three principal components are found in Africa, Australia, or America but not in the southern (central or eastern) part of Asia, which always has intermediate values of the first

Abbreviation: HLA, histocompatibility antigen complex.

* Permanent address: Instituto di Genetica Medica CNR, Centro per l'Immunogenetica e l'Istocompatibilità, University of Torino, Italy.
† Permanent address: Instituto di Ecologia, University of Parma, Italy.

‡ The list of 116 references and data employed will be supplied upon request.

Table 1. Principal component scores grouped in classes on an arbitrary scale from 1 to 10

	First	Second	Third
Africa	8–10	5–8	1–5
Europe	7–8	4–6	6–7
Near East	7–8	5–6	3–6
Central and South Asia	4–7	5–7	7–8
Southeast Asia	2–4	5–7	5–7
Far East	3–4	5–7	6–7
Northeast Asia and Alaska	3–4	2–3	8–9
North America	3–5	1–4	4–8
Central America	5–7	2	2–4
South America	3–5	1–3	2–5
Australia and New Guinea	1–3	8–10	3–5

First, second, and third principal component scores explain, respectively, 29%, 17%, 10% of the total variation.

and second principal components. This also is true with two nonrandom subsets of the data, HLA (21 independent alleles) and non-HLA (18 independent alleles). The genetic centrality of Asia shown by synthetic variables can be understood from historical and geographic considerations. Major migrations to America and Australia must have started from Asia, respectively from the northeastern and southeastern parts. There must also have been considerable exchanges between Asia, Europe, and Africa at various times and especially in the last periods of formation of modern man. Among these, (*i*) the replacement of Neanderthal man with modern man, probably of Asiatic origin; (*ii*) the radiation of Neolithic man from the Near East 10,000–5,000 years ago; and (*iii*) the other later radiations mentioned above are examples of gene flow from Asia into the rest of the Old World. Thus, in the last tens or hundreds of thousands of years, Asia may have become increasingly the center of both the Old and the New World, and the analysis of genetic data agrees with this picture.

Neolithic Radiations. The trichromic map in Fig. 1, summarizing almost 60% of the total information from 39 independent alleles, shows in a direct way the overall similarities between aboriginal populations. Although a study of special regions would benefit more from a computation of principal components limited to local data, this global representation seems to have preserved some information of interest also for local phenomena. Thus, the map indicates that the Neolithic radiation connected with the diffusion of farmers from the Near East (1) seems to have spread almost concentrically not only to Europe, but also to northern Africa, to the Arabian Peninsula, and eastwards to Iran and India. The ellipses with major axis directed N–S and centered in East Asia may be hints of another radiation, perhaps connected with agricultural development,

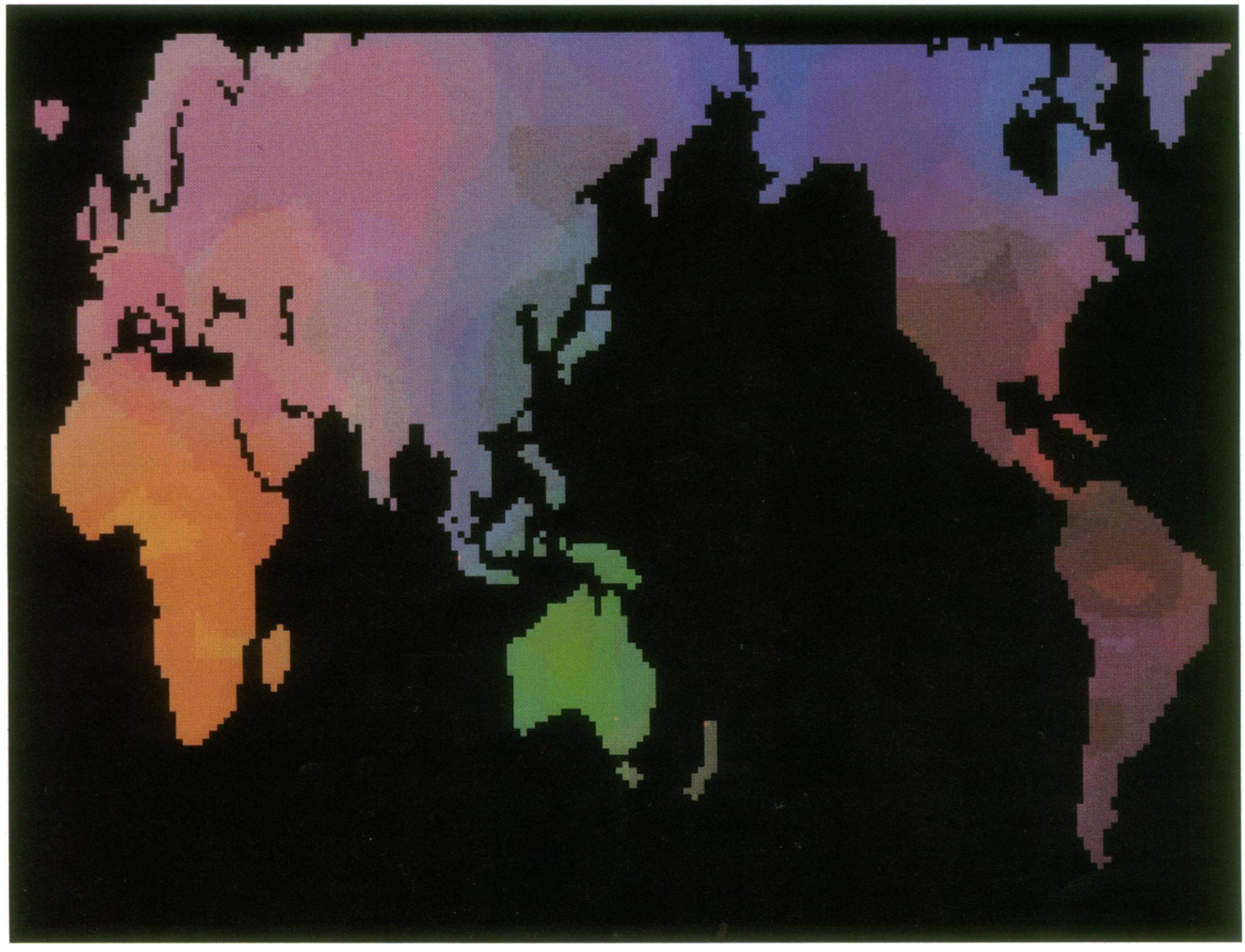

FIG. 1. Trichromic synthetic map of the first three principal components of gene frequencies, computed from 39 independent alleles at the human loci: ABO, Rh, MN, Lewis, Duffy, Hp, acid phosphatase, PGM1, HLA-A, and HLA-B. Shades indicate different intensities of the principal components and colors (on arbitrary scales) represent different blendings of the first (green), the second (blue), and the third (red) component. Gene frequencies are from aborigines only. Tasmania and Cuba, for which there are no surviving aborigines, borrowed their colors from neighbors.

because there was one (and probably more) major center of origin of such cultures in this area (9). There is also some support for expansion of Bantu-speaking agriculturalists from Nigeria/Cameroon (10) and for the centers of origin of neolithic spread in South America postulated by Lathrap (11). All of these are merely suggestions, however, and indicate the need of further local analyses. Unfortunately, the genetic and archeological knowledge in these regions is not as detailed as in Europe.

Variation of distance: comparison of latitude and longitude effects

The principal component scores interpolated at the 159 locations were subjected to a two-way analysis of variance to show the relative effects of longitude and latitude. To alleviate the problems generated by analysis of variance with an unequal number of subclasses, the 24 longitudes were grouped into 12 adjacent pairs. Results of the analysis are shown in Table 2; all F values are significant at $P = 0.001$. The term "longitudes" is the variance among "lunes," a lune being the area bounded by two meridians (here 32° apart); for the purposes of analysis, the value of a lune was computed by averaging the principal component scores at the nodes within the lune. There were 12 lunes, with the last lune smaller because 360° is not a multiple of 32°. Similarly, the term "latitudes" is the variance among the 12 means of the scores at nodes lying on a parallel, starting with parallel 76° N and ending at 56° S, with step 12° latitude.

It is apparent immediately that the variation by longitude is much more important than that by latitude for the first and third components but somewhat less important for the second component. Both latitude and longitude are expected to show effects because of isolation by distance, and longitude is expected to show the greater effect, as there is a wider range in the E–W direction than in the N–S direction. The variance due to latitude, however, also includes effects due to climate, which are partially confounded with those of distance. Therefore, the lack of a major effect of latitudes in this analysis gives a first qualitative indication that climate does not play a *major* role in determining the total genetic variation, because climate is associated with latitude rather than with longitude. By contrast, an analysis of Howells' craniometric data (12), obtained on 17 ethnic groups, showed that the first canonical variate was associated very significantly with climate (13), in particular temperature measurements, and this association explained a large fraction of it. However, the lower importance of latitude compared to longitude in the present case does not entirely exclude an effect of climate on gene frequencies also. Therefore, the search for effects of latitude was extended by more direct techniques.

Discriminant Analysis. Multiple discriminant analyses were carried out separately for longitude and latitude. In the discriminant analysis for longitude, the gene frequencies for the 39 alleles interpolated at the 159 nodes were grouped by 24 meridians, 16° apart. The discriminant was computed so as to maximize the ratio of the variance of the means of gene frequencies among meridians to that within meridians. The two leading discriminant functions accounted, respectively, for 43% and 20% of the total variation among longitudes. The mean values of the discriminants at the different longitudes are plotted in Fig. 2. This plot tends to reproduce in an approximate way the geographic map that is expected when there is sufficiently high correlation between geographic distance and distance in the traits being examined (genetic distance in the discriminant space in our case).

The discriminant plot of Fig. 2 shows the similarities of human ethnic groups when compared on the basis of longitude only. The order of populations is the same as in the geographic maps, with distortions in the metric plot. A major one is the reduction in size of the Pacific Ocean with respect to the continent of Asia, indicating that it has been easier for man to cross the seas than land. We also know that the passage from northeast Asia to North America has been made repeatedly in the last 15,000 years. Another distortion in Fig. 2 shows the Pacific Ocean as being smaller than the Atlantic Ocean. In fact, one might expect even greater distortion, because crossings of the Atlantic prior to 1492 were extremely limited. In other words, one might expect an even greater distance between American and European aborigines across the Atlantic Ocean than Fig. 2 shows. It is possible that, in the discriminant plot, one point (−31° long) is largely responsible for this effect, as it represents almost exclusively Greenland Eskimos, some of whom show Caucasian admixture, most probably of late origin (14). Other admixtures of American natives with Caucasians and Africans may also have an influence and will be further analyzed elsewhere.

Discrimination among 12 latitudes are carried out by grouping the 159 nodes according to latitudes. The first two discriminants explain 45% and 30%, respectively, of the variation among means at the 12 latitudes. The means of the first two discriminants at the 12 latitudes are plotted in Fig. 3, and neighboring latitudes are joined with segments. We would expect the figures thus generated to form a closed pattern only if there were an important selective effect on very cold climates, so that people living in the coldest northern regions would tend to have gene frequencies very similar to those of populations living in the coldest regions of the southern hemisphere. Even though

Table 2. Variation by longitude and latitude of the three leading principal components summarizing world distributions of 39 allele frequencies

Source of variation	df	Principal component: First Mean square	First F	Second Mean square	Second F	Third Mean square	Third F
Longitudes	11	4.431	227	1.090	35	2.115	53
Latitudes	11	0.201	10	2.011	65	0.583	15
Interaction	78	0.099	5	0.100	3	0.134	3
Residual	58	0.020	—	0.031	—	0.040	—

df, Degrees of freedom. All F values are significant at $P = 0.1\%$.

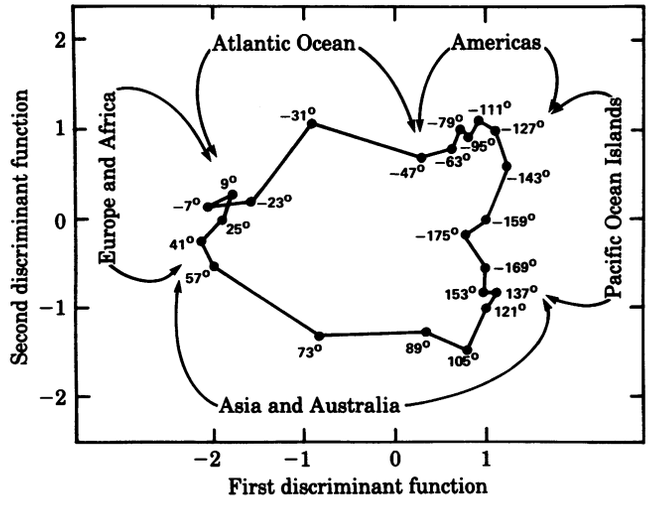

FIG. 2. Discriminant analysis among longitudes: the means (across latitudes) of the first and the second discriminant functions are plotted for 23 different longitudes. The values corresponding to neighboring longitudes are joined by segments. Negative and positive numbers represent western and eastern longitudes, respectively.

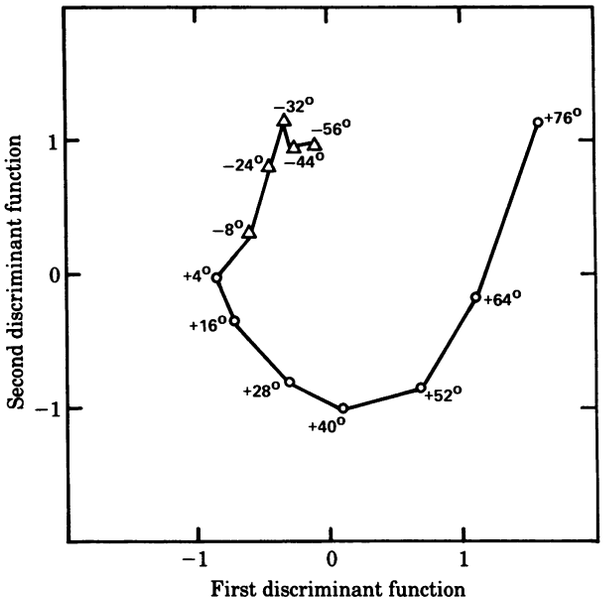

FIG. 3. Discriminant analysis among latitudes: the means (across longitudes) of the first and the second discriminant functions are plotted for 12 different latitudes. The values corresponding to neighboring latitudes are joined by segments. ○, Northern latitudes; △, southern latitudes.

a fully closed pattern is not found in Fig. 3, one notes a tendency towards it, suggesting some effect of climate. This is more clearly seen by plotting the first discriminant versus distance from the equator. In both northern and southern hemispheres the mean discriminant increases with latitude, as can be seen in Fig. 3 and, more directly, in Fig. 4. The second and third discriminants did not show such a correlation. If we remember that discrimination has been carried out among latitudes *irrespective of whether they were northern or southern* and, therefore, irrespective of distance from the equator, the association of the first discriminant with distance from the equator and, therefore, presumably with climate is truly striking.

The analysis with two subsets of the alleles studied, those of the HLA system (15) and all non-HLA alleles, give conclusions qualitatively similar to those of the total set, both for the discrimination of latitudes and of longitudes.

Genes showing correlation with distance from the equator

An indication of which genes are most sensitive to the effect of climate can be obtained by correlating directly the frequencies of each allele with the distance from the equator of the corresponding populations. We computed three sets of correlation

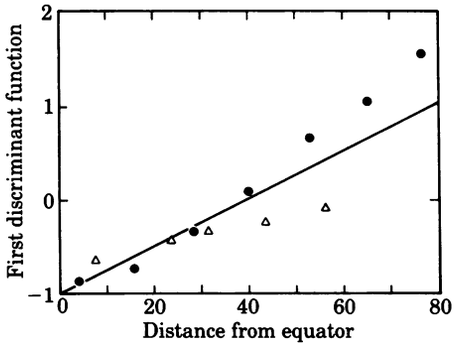

FIG. 4. The means of the discriminant function between latitudes are plotted on the ordinate and the distance from the equator (in degrees) on the abscissa. Northern (○) and southern (△) latitudes are fitted with a common regression line.

coefficients between distance from the equator (taken as the absolute value of latitude) and gene frequencies for each allele.

(*i*) The first set of correlation coefficients with distance from the equator, r_A, was computed using gene frequencies of the original populations. The r_A values can be read on the ordinate of the upper graph in Fig. 5.

It should be noted that the original data are highly clustered in space because some areas have been tested much more extensively than others and, thus, are more dense with data points, each corresponding to a sampled population. The den-

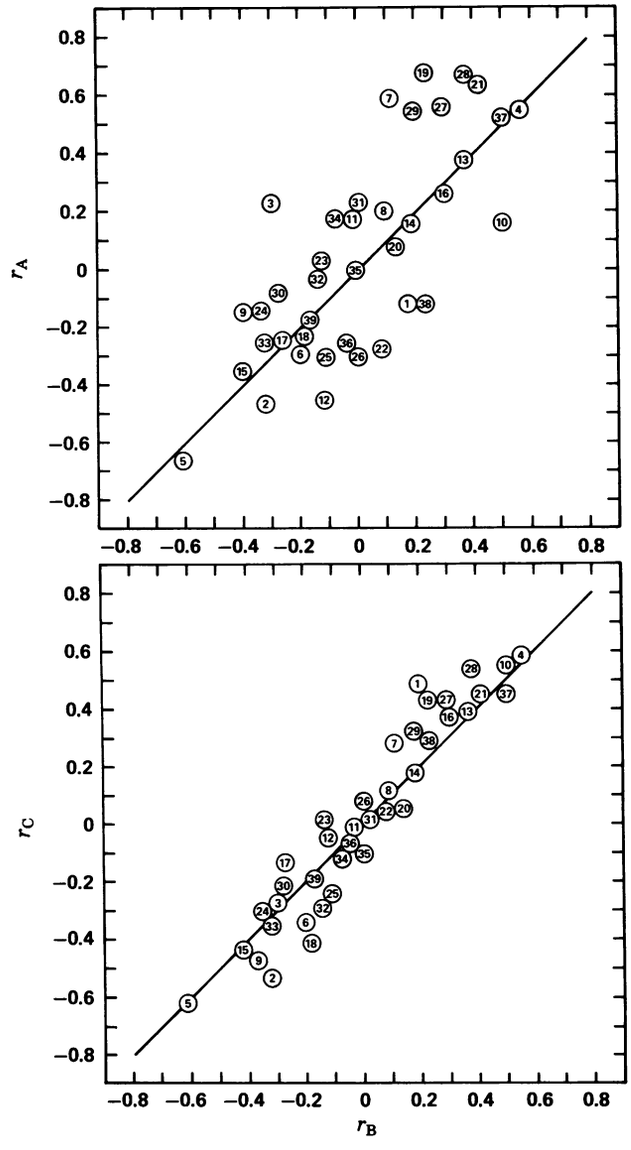

FIG. 5. Each point in the above diagram is a correlation coefficient between distance from the equator and a gene frequency for a given allele. The alleles studied are indicated with numbers from 1 to 39: 1, Fy^a; 2, Hp1; 3, Le (a^+); 4, P^a; 5, P^b; 6, PGM1 1; 7, Rh (cde); 8, Rh (cdE); 9, Rh (cDe); 10, Rh (cDE); 11, Rh (Cde); 12, Rh (CDe); 13, MS; 14, Ms; 15, Ns; 16, A_1; 17, B; 18, O; 19, HLA–A1; 20, HLA–A2; 21, HLA–A3; 22, HLA–A9; 23, HLA–A10; 24, HLA–A11; 25, HLA–A28; 26, HLA–B5; 27, HLA–B7; 28, HLA–B8; 29, HLA–B12; 30, HLA–B13; 31, HLA–B14; 32, HLA–B15; 33, HLA–B17; 34, HLA–B18; 35, HLA–B21; 36, HLA–B22; 37, HLA–B27; 38, HLA–B35; 39, HLA–B40. Correlation coefficents have been computed in three ways: (*i*) r_A values (ordinate of upper graph) are obtained using gene frequencies from the original population data; (*ii*) r_B values (abscissa of both graphs) are obtained from gene frequencies interpolated in our maps at 159 nodes specified in the text; (*iii*) r_C values (ordinate of lower graph) are computed as the r_B values, but within the lunes defined in the text.

sity is highest in Europe and is likely to generate spurious correlations with geographic variables such as latitude.

(*ii*) To eliminate error due to this source, we computed a second set of correlation coefficients (abscissa r_B of Fig. 5) by using gene frequency values interpolated at the 159 nodes.

There is reasonable similarity between the two sets of correlation coefficients, r_A and r_B, as can be seen in Fig. 5 *Upper*, but there are also important discrepancies. For instance, the correlation between distance from the equator and frequency of the Rh negative gene (which is very high in Europe and low elsewhere) is $r_A = +0.59$ and, therefore, highly significant when using the raw gene frequencies but is only $r_B = +0.11$ and, thus, practically nil when using the 159 nodes, which give approximately equal weight to all inhabited parts of the world. Other alleles for which striking discrepancies are noticeable are cDE, Lewis, and four HLA alleles: A1, A3, B7, B8. Note that A1, A3 are known to be in linkage disequilibrium with B8, B7, respectively. Differences like these throw doubt, in our view, on the validity of correlations of gene frequencies with environmental parameters, when these are computed without eliminating possible effects of clustering, as happened in the r_A set. Support for the results obtained with the r_B method came from an independent analysis in which a weighted correlation was computed, the weight of each point being inversely proportional to the local density of data. However, the method used for computing the r_B does not eliminate the effect of major historical accidents in the geographic distribution of gene frequencies, which need not reflect selective effects of climate. This led us to try the following third method.

(*iii*) As a majority of the variation in gene frequencies is associated with longitude, we reestimated correlations of distance from the equator with gene frequency values at the 159 nodes *within longitudinal strips* ("lunes"), each lune having a width of 32° longitude. This generated the r_C set of correlations, and we conjecture that this set (ordinate of Fig. 5 *Lower*) is the most reliable. There is no great difference, however, between the r_B and r_C sets. Alleles showing the most important changes, all of which are in the direction of increase in absolute values of the correlation coefficients when replacing r_B with r_C, are Duffy (from 0.20 to 0.49), Hp (from −0.32 to −0.53), O (from −0.18 to −0.41), HLA-A1 (from 0.22 to 0.44), and B8 (from 0.37 to 0.53).

The highest correlation with distance from the equator is for the erythrocyte acid phosphatase polymorphism ($r_C = 0.59$ for P^a, $r_C = -0.62$ for P^b) and is high for all three r sets. This finding agrees with the results by Ananthakrishnan and Walter (16), who showed a significant negative correlation ($r = -0.71$) for the P^a allele frequency with the mean annual temperature of the places where the populations sampled were living. Acid phosphatase alleles P_a and P_c showed higher frequency among Sardinian males with favism (17), but conflicting data have been reported from other populations (18).

Other non-HLA alleles showing correlations with latitude above 0.4 are Fy^a, haptoglobin, Rh (cDE, cDe), O, Ns. It has recently been shown that the Fy antigen has receptor-like properties for malarial parasites, which could explain the correlations with climate (19). The positive (about equally high) correlations of the HLA alleles (A1, A3, B8, B7) with distance from the equator support the notion (20) that the linkage disequilibrium observed between the pairs of alleles (A1, B8), (A3, B7) at loci HLA-A and HLA-B is not due to genetic drift and migration or the latter only. It is worth noting that the allelic combination A1-B8-Dw3 is highly associated with susceptibility to celiac disease (21), whereas A3-B7-Dw2 is associated with multiple sclerosis (22), whose prevalence seems to increase in colder North European climates (23). The remaining HLA allele correlating more than 0.4 with distance from the equator is B27 ($r_C = 0.45$), whose association with rheumatic diseases (mostly ankylosing spondylitis) is known (24, 25).

The sample of our genes is biased towards blood groups, and 10 out of 35 showed a correlation above 0.4. Data on many enzymes and proteins are not yet abundant enough for this type of study. It is noteworthy, however, that two out of three proteins or enzymes considered here show significant associations with climate.

We thank R. Lyons and J. Hwang for help in the preparation of the display graphics; D. Carmelli, R. Lewontin, R. Sokal, and D. E. Wartenberg for comments on the manuscript; and W. Goodenough for advice. Work was supported by National Institutes of Health Grants GM 20467 and 10452 and Department of Energy Contract DEAS-03-326PA79E.

1. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. (1978) *Science* **201**, 786–792.
2. Piazza, A., Menozzi, P. & Cavalli-Sforza, L. (1980) *Biometrics*, in press.
3. Ammerman, A. & Cavalli-Sforza, L. L. (1971) *Man* **6**, 674–688.
4. Ammerman, A. J. & Cavalli-Sforza, L. L. (1973) in *The Explanation of Culture Change*, ed. Renfrew, C. (Duckworth, London), pp. 343–357.
5. Goodenough, W. H. (1970) in *Indo-European and Indo-Europeans*, eds. Cardona, G., Hoenigswald, H. M. & Senn, A. (Univ. Pennsylvania Press, Philadelphia), pp. 253–265.
6. Sgaramella-Zonta, L. & Cavalli-Sforza, L. L. (1973) in *Genetic Structure of Populations*, Population Genetics Monograph (Univ. Press Hawaii, Honolulu, HI), Vol. 8, pp. 128–135.
7. Latter, B. D. H. (1980) *Am. Nat.* **116**, 220–237.
8. Mourant, A. E., Kopec, A. C. & Domanyiewska-Sobczak, K. (1976) *The Distribution of the Human Blood Groups* (Oxford Univ. Press, Oxford, England). Tables 1.2, 1.2.1, 2.7, 2.11.1, 4.13, 4.13.1, 4.19.1, 7.3, 7.7.1, 8.1.1, 8.3.1, 8.3.2, 36.1, 50.1.1, 53.
9. Chang, K. C. (1978) *J. Am. Oriental Soc.* **98**, 85–91.
10. Hiernaux, J. (1974) *The People of Africa* (Scribners, New York).
11. Lathrap, D. W. (1977) in *Origins of Agriculture*, ed. Reed, C. A. (Mouton, The Hague, The Netherlands), pp. 713–751.
12. Howells, W. W. (1973) in *Peabody Museum Papers* (Harvard Univ. Press, Cambridge, MA), Vol. 67.
13. Guglielmino-Matessi, C. R., Gluckman, P. & Cavalli-Sforza, L. (1979) *Am. J. Phys. Anthro.* **50**, 549–564.
14. Kissmeyer-Nielsen, F., Kyerbye, K. E., Lamm, L. U., Jorgensen, J., Petersen, G. B. & Gurtler, H. (1973) in *Histocompatibility Testing 1972*, eds. Dausset, J. & Colombani, J. (Williams and Wilkins, Baltimore), pp. 317–324.
15. Piazza, A., Menozzi, P. & Cavalli-Sforza, L. L. (1981) *Hum. Immunol.*, in press.
16. Ananthakrishnan, R. & Walter, H. (1972) *Humangenetik* **15**, 177–181.
17. Bottini, E., Lucarelli, P., Agostino, R., Palmarino, R., Businco, L. & Antogroni, G. (1971) *Science* **171**, 409–411.
18. Hartz, J., El Maghrabi, R. El., Namen, A., Gabor, M., Bowman, J., Carson, P., Ajmar, F. & Kamel, K. (1973) *Clin. Chim. Acta* **48**, 117–126.
19. Miller, H. L., Mason, S. J., Clyde, D. F. & McGinniss, M. H. (1976) *N. Engl. J. Med.* **295**, 302–304.
20. Bodmer, W. F., Cann, H. & Piazza, A. (1973) in *Histocompatibility Testing 1972*, eds. Dausset, J. & Colombani, J. (Williams and Wilkins, Baltimore), pp. 753–768.
21. Keuning, J. J., Pena, A. S., van Leeuwen, A., van Hooff, J. P. & van Rood, J. J. (1976) *Lancet* **i**, 506–511.
22. Fog, T., Schuller, E., Jersild, C., Engelfriet, C. P. & Bertrams, J. (1977) in *HLA and Disease*, eds. Dausset, J. & Svejgaard, A. (Williams and Wilkins, Baltimore, MD), pp. 108–125.
23. Kurtzke, J. F. (1977) *J. Neurol.* **215**, 1–26.
24. Ryder, L. P. & Svejgaard, A. (1978) in *Blood Groups and Diseases: A Study of Associations of Diseases with Blood Groups and Other Polymorphisms*, ed. Mourant, A. E. (Oxford Univ. Press, Oxford, England), Appendix, pp. 275–298.
25. Daneo, V., Migone, N., Modena, V., Bianchi, S. D., Alfieri, G., Carbonara, A. O. & Piazza, A. (1977) *J. Rheumatol.* **4**, Suppl. 3, 5–10.