Genetic Population Structure of Italy. II. Physical and Cultural Barriers to Gene Flow

Guido Barbujani*'† and Robert R. Sokal†

*Dipartimento di Biologia, Universita di Padova, Padua; and †Department of Ecology and Evolution, State University of New York, Stony Brook

Summary

Three approaches were employed to evaluate the relative importance of geographic and linguistic factors in maintaining genetic differentiation of Italian populations as shown by blood groups and erythrocyte and serum markers. Genetic distances are closer to linguistic than to geographic distances. Gene-frequency change across 12 linguistic boundaries is significantly more rapid than at random locations. The zones of sharp genetic variation correspond to physical barriers to gene flow and to boundaries between dialect families, which overlap widely. However, two linguistically differentiated populations appear genetically differentiated despite the absence of physical obstacles to gene flow around them. The Po River is associated with abrupt genetic change only in the area where it corresponds to a dialect boundary. At most loci the genetic population structure seems affected by linguistic rather than geographic factors; exceptions are the systems that were subject to malarial selection in geographically close but linguistically heterogeneous localities. Gene flow appears to homogenize gene frequencies within regions corresponding to dialect families but not between them, leading to the patchy distributions of allele frequencies that were detected in an earlier study.

Introduction

Differentiation of gene frequencies in spatially contiguous populations has been attributed to isolation by distance in many studies on humans (for a review, see Wijsman and Cavalli-Sforza 1984). The decline of genetic similarity with distance is approximately exponential within regions as large as a few hundred kilometers across, which corresponds to the predicted effect of individual dispersal and random genetic drift on gene frequencies (Sokal and Wartenberg 1983; Barbujani 1987). Only more complex biological or demographic processes, though, can account for long-distance differentiation of populations (Barbujani 1988). Some extensive gene-frequency gradients seem due to directional gene flow, which in turn has been associated with large-scale population movements inferred from archaeological evidence (Menozzi et al. 1978; Sokal and Menozzi 1982; Ammerman and Cavalli-Sforza 1984; Rendine et al. 1986). In a few other cases, evidence for balancing or differential selection has been found (Piazza et al. 1981*b;* Hedrick and Thomson 1983; Klitz et al. 1986). However, geographic variation observed at most loci seems to depend on interactions among diverse evolutionary pressures, some of which may have acted in the remote past (Piazza et al. 1988).

In a previous study of Italian populations we recognized seven distinct patterns of gene frequencies (Barbujani and Sokal, in press). Although short-range differentiation was consistent with isolation-by-distance expectations for 60 of 61 alleles analyzed, the modes of long-range differentiation were highly variable. This is not compatible with the hypothesis of a major migratory movement, which is expected to affect several gene frequencies approximately equally (Slatkin 1975, 1985). Both heterogeneous selective pressures in different areas, or factors of isolation other than distance, can cause large-scale population divergence. However, in the Italian study, selection appears unlikely to be the major process involved. Indeed, evidence for selective effects exists only for

Received July 9, 1990; revision received September 17, 1990.

Address for correspondence and reprints: Dr. Robert R. Sokal, Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794–5245.

0002-9297/91/4802-0025$02.00

two polymorphisms maintained by malaria, G6PD deficiency, and beta-thalassemia. Wide latitudinal clines, which could suggest an effect of climatic selection (Piazza et al. 1981*b*), are uncommon (only one such pattern among the 61 alleles in this study; Barbujani and Sokal, in press). It is impossible to rule out other, unidentified selective pressures that might have affected genetic variation in Italy, but it is also impossible to study their effects. Anyway, even when G6PD and beta-thalassemia are neglected, the markers studied show substantial heterogeneity in their geographic patterns of variation. It is therefore of interest to test which factors, in addition to geographic distance and malaria-related selection, can explain the observed genetic divergence of populations.

Starting with the work of Livingstone (1963), various authors have observed positive association of genetic and linguistic differentiation indices. Positive significant correlations were apparent in Oceania ($.49 < r < .57$; Friedlaender et al. 1971), Australia ($r = .44$; White and Parsons 1973), New Guinea ($r = .74$; Serjeantson et al. 1983), Brazil ($r = -.27$ in a comparison of indices of genetic dissimilarity and linguistic similarity; Salzano et al. 1977), Central America ($r = .69$; Barrantes et al. 1990), Europe (average r across 20 systems = .280; Sokal 1988) and sub-Saharan Africa ($.32 < r < .57$; Excoffier et al., in press). Although there are a few exceptions (e.g., see Livingstone 1963; Spuhler 1972), there seems to be a general parallelism of genetic and linguistic variation (Cavalli-Sforza et al. 1988; Sokal 1988; Sokal et al. 1989*b*, 1990; Barton and Jones 1990). In particular, Harding and Sokal (1988) proposed that affinities between modern gene pools reflect their geographic distance only within homogeneous language groups; additionally, language differences act as barriers that restrict gene flow, often leading to increased genetic divergence among linguistically different populations. This was called a modified gene-flow model (as opposed to a model of unconstrained gene flow) and is further supported by the fact that the zones of rapid genetic change in Europe overlap widely with the boundaries between regions where different languages are spoken (Barbujani and Sokal 1990).

The purposes of this study are to analyze the association between language and genetics within Italy, and to test whether linguistic differences can account for a fraction of genetic differences that cannot be explained by selection or isolation by distance. Italy was chosen both for the large amount of genetic data available and for the presence of a well-described linguistic structure within the country. Three complementary approaches were followed. We first tested whether genetic distances are still associated with linguistic distances when the effect of geographic distance is partialled out. Once the presence of some genetic variation due to language had been established (geography held constant), we tested whether genetic change across linguistic boundaries is more rapid than across random lines drawn on the map. Finally, we located the zones where genetic change per unit distance is highest and compared them with the distribution of potential obstacles to gene flow—physical barriers (mountain ranges, seas, rivers) and cultural (linguistic) boundaries.

Material and Methods

1. *The Data*

Allele frequencies for 20 systems in Italian populations (Table 1) were taken from a data base of European gene frequencies (see Sokal et al. 1989*a*, and references therein); the data published by Livingstone (1967), Piazza et al. (1982), and Tills et al. (1983) were incorporated, as were data resulting from a computer search of the recently published literature. In general, each system represents a different locus, with three exceptions: ABO (system 1.1 is ABO typed with anti-A and anti-B sera, system 1.2 is the same typed with anti-A1, anti-A2, and anti-B sera); MN (systems 2.5 and 2.7); and Rh (systems 4.1, 4.13, and 4.19). The numerical codes for each system come mostly from Mourant et al. (1976). One allele was omitted in each system to approximate the independence of allele frequencies. Sixty-one sets of allele frequencies, which are referred to as gene-frequency surfaces, were eventually obtained, for a total of 1,119 data points overall (detailed in Barbujani and Sokal, in press). The number of localities typed ranges from 215 (for Rh) to 19 (for Rh haplotypes [system 4.13], HLA-A and HLA-B), with an average number greater than 50. These data include most of the gene frequencies recently compiled by Piazza et al. (1989). For each system at least four localities (but generally more) were in Sardinia.

Linguistic change is virtually continuous in Italy (Goebl 1981, and in press), and the regions where different dialects are spoken are sometimes difficult to determine, owing to the large number of gradations present. For the purpose of this study, though, an approximate clustering of populations reflecting their main dialect affiliation seemed sufficient. To this end,

Table 1

Genetic Systems Considered

Code[a]	System[b]	n[c]	y[d]	Notes
1.1	ABO	17	680	
1.2	ABO	12	480	As studied with anti-A1, -A2, and -B sera
2.5	MN	14	560	
2.7	MN	10	400	With anti-M, -N, and -S
3.1	P	10	400	
4.1	Rh	17	680	
4.13	Rh	10	400	With anti-C, -D, -E, and -c
4.19	Rh	12	480	With anti-C, -D, -E, -c, and -e
6.1	K	14	560	
8.1	Fy	11	440	
36.1	Hp	12	480	
38.1	Gc	9	360	
50.1.1	ACP	11	440	
51.1.1	G6PD	9	360	
53	PGM	12	480	
56	AK	9	360	
63	ADA	8	320	
100	HLA-A	8	320	
101	HLA-B	8	320	
203	Th	13	520	

[a] Numerical codes are from Mourant et al. (1976), except for HLA and thalassemia.

[b] K = Kell; Fy = Duffy; Hp = haptoglobin; Gc = group-specific complement; ACP = acid phosphatase; G6PD = glucose 6-phosphate dehydrogenase; PGM = phosphoglucomutase 1; AK = adenylate kinase; ADA = adenosine deaminase; Th = beta-thalassemia.

[c] Number of linguistic boundaries that could be tested for each system.

[d] Number of permutations employed in the test of the rate of gene-frequency change.

Pellegrini's (1977) classification was employed, based on linguistic characteristics mapped in the 1930s. Since that time, dialect differences have been progressively blurred by the spread of literacy and television. Language has evolved fast, showing a general trend towards greater uniformity. However, gene frequencies are unlikely to have undergone a similar process in the same period; moreover, recent immigrants were excluded from the samples this study is based upon. Therefore there is no reason to suspect that the estimated gene frequencies should differ substantially from those one could have observed in the 1930s. Thus, although most samples surely include individuals born after 1930, the data employed seem suitable for a comparison of the genetic and linguistic structure of Italian populations.

The 10 major dialect groups recognized in Italy (Pellegrini 1977) are Franco-Provenzale (FP), Gallo-Italico (GI), Veneto (VE), Friulano (FR), Toscano (TO), Mediano (MD), Meridionale Intermedio (MI), Meridionale Estremo (ME), Sassarese-Gallurese (SG), and Logudorese-Campidanese (SC). These groups are assigned to six dialect systems as follows: Franco-Provenzale (FP), Cisalpino (GI, VE), Friulano (FR), Toscano (TO), Centro-Meridionale (MD, MI, ME), and Sardo (SG, SC). In addition, French (F), German (G), Slovenian (S), and Albanian (A) are spoken by Italian populations. In figure 1, 19 numbered boundaries between different languages, dialect systems, and dialect groups are mapped.

2. *Distances between Populations*

The geographical distribution of the sampled populations is different for the different loci. To employ all the information available, we computed matrices of genetic, geographic, and linguistic distances independently for each of the 20 systems, thus obtaining 60 distance matrices overall. They will be referred to also as GEN, GEO, and LAN matrices, respectively.

Among the various available measures of genetic distance we chose Prevosti's index (Prevosti et al. 1975), for consistency with the previous study on Eu-

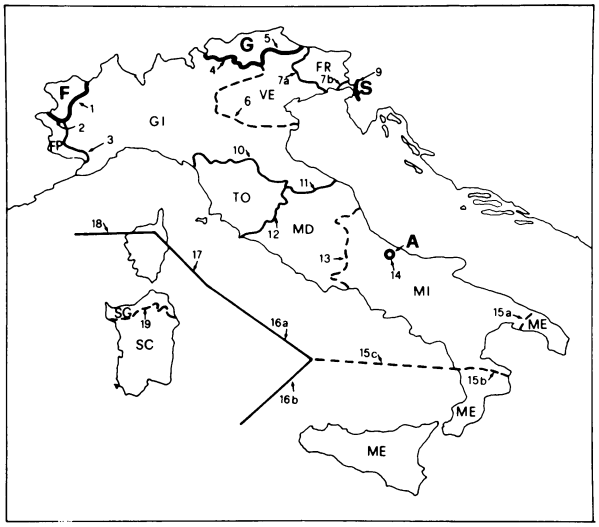

Figure 1 Languages, dialect systems, and dialect groups in Italy. Abbreviations for the dialect groups are as follows: FP = Franco-Provenzale; GI = Gallo-Italico; VE = Veneto; FR = Friulano; TO = Toscano; MD = Mediano; MI = Meridionale Intermedio; ME = Meridionale Estremo; SG = Sassarese-Gallurese; SC = Logudorese-Campidanese. Dialect systems to which these dialect groups belong are given in the text. Languages other than Italian are as follows: F = French; G = German; S = Slovenian; A = Albanian. Thick solid lines are language boundaries, thin solid lines are dialect-system boundaries, dashed lines are dialect-group boundaries. The boundaries are numbered consecutively from 1 to 19.

ropean populations (Sokal 1988). Prevosti's distance D_{jk} between two populations j and k is

$$D_{jk} = 0.5 \sum_{i=1}^{m} |P_{ij} - P_{ik}| ,$$

where p is the frequency of the ith allele in the two populations, and summation is over the m alleles at that locus. The geographic distances were computed as great-circle distances in kilometers.

Linguistic distances were calculated on the basis of Pellegrini's (1977) map of Italian dialects. Pairs of localities whose inhabitants speak dialects belonging to the same dialect group were assigned distance 0. Linguistic distance was 1 for populations belonging to different dialect groups within the same dialect system, and 2 for populations belonging to different dialect systems. Finally, localities where different languages are spoken (the samples analyzed included Albanians, as well as German, French, and Slovenian speakers) were assigned distance 3.

3. *The First Approach: The Analysis of Distance Matrices*

The individual distance measures thus obtained are not independent, as a set of n localities yields $n(n-1)/2$ pairwise distances. The standard statistical techniques that test for the association between pairs of variables are therefore inapplicable. Mantel's test (Mantel 1967; Sokal 1979) is a nonparametric procedure for comparing such matrices; it leads to calculation of a test statistic Z that can be transformed into a correlation coefficient but whose significance is assessed on the basis of an empirical null distribution. Let **X** and **Y** be the two distance matrices of interest with elements X_{ij} and Y_{ij}. Mantel's statistic is defined as

$$Z_{XY} = \sum_{ij} X_{ij} Y_{ij} ,$$

where summation is over all the ij pairs other than $i = j$. This statistic will be maximal for positive association of the corresponding elements of the **X** and **Y** matrices, and is compared with the distribution of the Z'_{XY} values that are calculated when there is no association between matrices (the prime indicates these are not observed values). For this purpose, the rows and columns of one matrix are permuted at random many times while the other is kept constant, and Z'_{XY} is calculated each time, so that a Monte-Carlo null distribution is eventually generated. It is then possible to compute the empirical probability of observing a value smaller than Z_{XY}: $P = \Pr(Z'_{XY} < Z_{XY})$. Mantel statistics are finally normalized so as to range between -1 and $+1$ (Smouse et al. 1986), and to resemble a correlation coefficient.

Mantel's method was applied to test for the association of GEN with GEO, LAN with GEO, and GEN with LAN. In addition, we applied an extension of Mantel's test (Smouse et al. 1986) to see whether there is an association between GEN and GEO when language is held constant, and between GEN and LAN when the effects of geographic distance are held constant. For the former comparison, residual matrices were computed from the regression of GEN on LAN, and GEO on LAN, and the Mantel test was performed on these residual matrices. The procedure for comparing GEN and LAN, with GEO held constant, was analogous. In this way we tested how well the genetic distances between populations can be predicted on the basis of either their geographical or their linguistic distances, once the correlation between the latter two variables is eliminated.

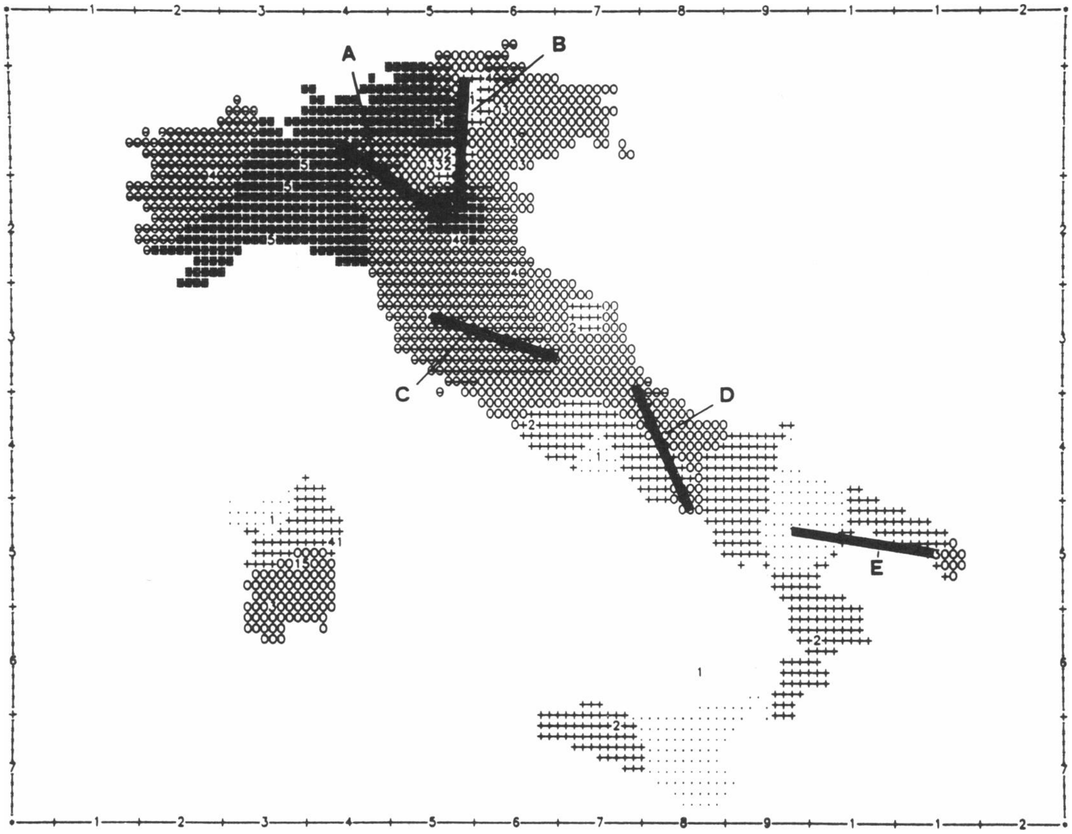

Figure 2 An example of an interpolated surface (allele *A1* of ABO). Allele frequencies have been grouped into quintiles represented by increasing darkness of shading. Sampled localities are indicated by numerals referring to the quintile into which the allele frequency of that locality falls. Lines of equal length connect arbitrary points to show that regions of both gradual and abrupt genetic change are present.

4. *Interpolation of Gene Frequencies*

For the subsequent phases of the analysis, the discontinuous gene-frequency distributions were transformed into 61 quasi-continuous surfaces by means of an interpolation algorithm. SYMAP (Dougenik and Sheehan 1979). On the basis of inverse squared distance weighting, sixty-one 74-row × 110-column lattices of hypothetical gene-frequency values were obtained (see fig. 2 for an example). The properties and drawbacks of such interpolated surfaces are discussed in Piazza et al. (1981*a*). Interpolation tends to smooth down the gene-frequency surfaces. Since the purpose of this study was to detect zones of increased rate of genetic change, the use of interpolated surfaces appeared conservative.

5. *The Second Approach: Rates of Gene-Frequency Change across Linguistic Boundaries*

This approach, developed by Sokal et al. (1988), tests whether the boundaries between populations belonging to different linguistic units correspond to zones where gene frequencies change more rapidly than at random locations. The boundaries of figure 1 were superimposed on the interpolated gene-frequency surfaces and the rate of genetic change across each linguistic boundary was estimated (boundaries 2 and 3 could not be analyzed owing to lack of data). For each straight line segment of the boundary of interest, the directional derivative of a given surface perpendicular to the boundary was calculated separately for each lattice square (pixel) crossed by the line segment. The final test statistic was the integral T of that derivative, computed along the boundary. The larger the T, the sharper the genetic change along that boundary. The significance of the observed T value was assessed by repeatedly placing a line of equal length and shape randomly on the map, each time recomputing the test statistic T'; once again, the prime denotes values calculated in a Monte-Carlo simulation test. T was then compared with the resulting null distribution. Details of the algorithm employed are given by Sokal et al. (1988).

The number of randomizations employed to generate the null T' distributions varied among systems; this was because the number of boundaries that could be tested differed among the loci, since each was based on a different-sized subset of samples. To be able to achieve overall significance at the 5% level when the probabilities for n individual boundaries were eventually combined, we generated $n/0.025$ values of T'. This number ranged from a minimum of 320 (for ADA and the HLA systems, where eight boundaries could be tested) to a maximum of 680 (for ABO and Rh [systems 4.1] with 17 boundaries; table 1). In all, 226 of the 380 possible tests (19 boundaries × 20 systems) were performed.

6. *The Third Approach: Wombling*

The zones of highest rate of genetic change (genetic boundaries: Barbujani and Sokal 1990) were investigated following an approach first put forward by Womble (1951), and recently developed for application to two-dimensional surfaces (Barbujani et al. 1989). We computed the partial derivatives of each interpolated surface at each lattice point with respect to a longitudinal and a latitudinal axis, recording both their magnitude and their direction. Magnitude and direction values were averaged across the 61 surfaces according to a procedure described in Barbujani et al. (1989), and the lattice points that showed average gene-frequency derivatives whose magnitude fell in the upper 5% of the distribution of magnitudes were plotted (except those elements of the lattice that were not adjacent to at least one another element belonging to the highest 5% of the distribution or connected to such an element by a "bridge" element from the next 5% of the distribution). In this way, single locations where the rate of genetic change was high were eliminated, and only strings of significant values appeared in the final plot.

In addition to this procedure, which will be referred to as "overall 'wombling'," we plotted the results of separate analyses of the 61 surfaces ("individual wombling"). The occurrences of each boundary in different surfaces were counted, and the following criterion was applied. The genetic boundaries are the zones of an individual surface where the magnitude of the gene-frequency derivative belongs to the upper 5% of its distribution. Every element of the lattice, i.e., every locality, has a probability $P = .05$ of belonging to a boundary in a given surface. The probability of a locality's belonging to a certain genetic boundary in a surfaces out of b surfaces tested is binomial:

$$\Pr(a|b) = C(b,a)\ .05^{a}\ .95^{b-a}$$

where $C(b,a)$ is the number of combinations that can be formed by taking b items at a time. We chose to consider significant only the boundaries whose probability is less than 5% in the joint analysis of the surfaces studied by individual wombling.

Results

1. *Genetic, Geographic, and Linguistic Distances*

Genetic distances correlate significantly ($P \leqslant .032$) with spatial distances for eight systems (table 2), including G6PD and Th. The significance is assessed by one-tailed tests, since only positive correlations were expected. To obtain an overall probability for rejection of the null hypothesis of no significant correlation, we applied Fisher's test for combining probabilities (Sokal and Rohlf 1981) to the probabilities associated with the correlation coefficients of table 2. First we reduced the number of systems from 20 to 15 by computing Bonferroni probabilities (Sokal and Rohlf 1987) for pooled estimates of the replicated systems ABO, MN, Rh, and HLA. These systems are, obviously, not independent. For the correlation between genetics and geography the overall probability is $P < .001$. As expected, geographic and linguistic distances correlate significantly ($P = .004$) for all the sets of localities where genetic markers were typed (overall $P << .001$). The correlation between GEN and LAN matrices is significant ($P \leqslant .040$) for 11 systems, including six of those whose genetic distances are significantly correlated with spatial distances (the systems significant for GEN-GEO but not for GEN-LAN are Hp and AK). The overall probability is $P < .001$. The matrix correlations are generally higher between GEN and LAN than between GEN and GEO (true for all systems where the GEN-LAN correlation is significant but three, Rh [system 4.19], G6PD, and Th).

When the effects of language variation are kept constant (table 3), the association between genetic and geographic distances remains significant ($P \leqslant .040$) for 6 systems, one of which (ADA) did not show significant association of GEN with either GEO or LAN in

Table 2

Mantel Tests of Matrix Association

Code	System	GEN-GEO[a]		LAN-GEO[b]		GEN-LAN[c]	
		r[d]	P[e]	r[d]	P[e]	r[d]	P[e]
1.1	ABO	.128	**.996**	.471	**.996**	.144	**.996**
1.2	ABO	.087	.932	.192	**.996**	.150	**.992**
2.5	MN	.163	**.996**	.725	**.996**	.223	**.996**
2.7	MN	.135	.912	.644	**.996**	.277	**.996**
3.1	P	−.027	.438	.741	**.996**	.072	.827
4.1	Rh	.139	**.996**	.472	**.996**	.161	**.996**
4.13	Rh	−.099	.301	.543	**.996**	−.177	.028
4.19	Rh	.427	**.996**	.649	**.996**	.421	**.996**
6.1	K	−.092	.048	.547	**.996**	−.049	.084
8.1	Fy	−.039	.313	.563	**.996**	−.015	.426
36.1	Hp	.101	**.968**	.639	**.996**	−.005	.518
38.1	Gc	−.005	.526	.443	**.996**	.267	**.960**
50.1.1	ACP	.043	.715	.760	**.996**	−.031	.386
51.1.1	G6PD	.271	**.996**	.940	**.996**	.238	**.996**
53	PGM	−.046	.229	.732	**.996**	−.133	.076
56	AK	.147	**.972**	.796	**.996**	.111	.940
63	ADA	.069	.775	.879	**.996**	−.011	.482
100	HLA-A	.165	.888	.698	**.996**	.407	**.996**
101	HLA-B	.221	.940	.698	**.996**	.390	**.996**
203	Th	.464	**.996**	.859	**.996**	.401	**.996**

[a] Genetic versus geographic distances.

[b] Linguistic versus geographic distances.

[c] Genetic versus linguistic distances.

[d] Mantel statistics normalized to correlation coefficients.

[e] Left-tail cumulative probability, based on 249 permutations; tests which are significant at $P \leqslant .05$ are in boldface type.

the previous pairwise tests (overall, $P < .001$). ADA was also the only system that did not show significant geographic variation in the autocorrelation analysis of gene frequencies in Italy (Barbujani and Sokal, in press). Conversely, nine systems show significant correlation ($P < .044$) of GEN and LAN, when geography is kept constant: they are the 11 systems showing positive GEN-LEN association, except for G6PD and Th. The overall probability is $P < .025$.

2. *Rates of Gene-Frequency Change Across Linguistic Boundaries*

At 11 linguistic boundaries the rate of allele-frequency change is significantly increased for one or several systems (fig. 3). The overall rate of change (tested by Fisher's method for combining probabilities) was significant across boundary 13 between MD and MI speakers, and across boundaries 17 and 18, north of Sardinia. Four, two, and seven systems, respectively, are individually significant at these boundaries. In addition, the rate of genetic change displayed by one or two systems at eight other boundaries is significantly higher than across lines of equal length and shape, randomly placed on the map.

Twenty-three out of 226 tests performed yielded significance at the .05 level. This is more than 10%, i.e., large enough to infer that at least some linguistic boundaries are associated with substantial increase of the slope of gene-frequency surfaces.

To compare the observed rates of change across different boundaries (which generally had to be tested for different combinations of systems), the absolute rates of change were also expressed in terms of percentiles for each surface. After summing these percentiles, we calculated the expected sum of percentiles for each boundary on a randomness hypothesis (table 4). The standardized deviations between observed and expected sums were finally compared with 1.645, that is, the 5% critical value for the hypothesis of equality of genetic change across all boundaries (this test was

Table 3

Smouse-Long-Sokal Tests of Partial Matrix Association

		GEN-GEO.LAN[a]		GEN-LAN.GEO[b]	
Code	System	r^c	P^d	r^c	P^d
1.1	ABO	.069	.892	.096	**.988**
1.2	ABO	.060	.867	.136	**.988**
2.5	MN	.002	.526	.155	**.984**
2.7	MN	−.060	.386	.251	**.984**
3.1	P	−.119	.137	.136	.884
4.1	Rh	.072	**.960**	.110	**.988**
4.13	Rh	−.004	.530	−.147	.205
4.19	Rh	.223	**.980**	.209	**.980**
6.1	K	−.078	.116	.002	.498
8.1	Fy	−.037	.386	.009	.570
36.1	Hp	.136	**.980**	−.091	.141
38.1	Gc	−.143	.129	.300	**.956**
50.1.1	ACP	.102	.896	−.097	.108
51.1.1	G6PD	.144	**.996**	−.053	.153
53	PGM	.076	.739	−.146	.092
56	AK	.097	.948	−.010	.466
63	ADA	.165	**.968**	−.151	.080
100	HLA-A	−.183	.125	.414	**.996**
101	HLA-B	−.078	.353	.337	**.996**
203	Th	.254	**.996**	.005	.578

[a] Geographic versus genetic distances, language kept constant.

[b] Genetic versus linguistic distances, geography kept constant.

[c] Mantel statistics normalized to partial correlation coefficients.

[d] Left-tail cumulative probability based on 249 permutations; tests significant at $P \leqslant .05$ are in boldface type.

one-tailed). The four boundaries around Sardinia, and language boundary 5, indicated higher-than-average gene-frequency change (double dotted lines in fig. 3).

By and large, gene-frequency variation appears more rapid than expected by chance alone (1) between Sardinia and the rest of Italy, including Sicily, (2) between populations of MD speakers and their northern and southern neighbors, and (3) across three language and dialect boundaries of northeastern Italy. Significant change at three boundaries (11, 12, 13) is caused by sharp differences at the K locus between the central Italian and surrounding populations, with high frequency of the *k* allele in central Italy; boundary 12 is reinforced by a large difference between the Florence and Rome populations in the allele frequencies at the P locus. Boundary 15a, significant for Rh (system

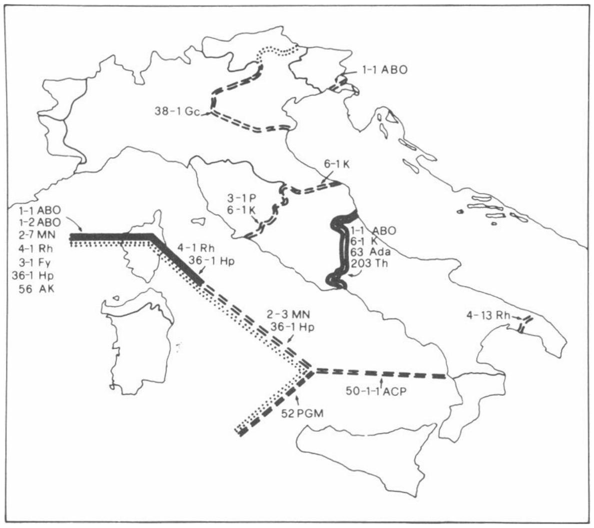

Figure 3 Results of significance tests of gene-frequency change across linguistic Boundaries. *Double solid lines,* Boundaries where the overall rate of genetic change is significant. *Double dashed lines,* Boundaries where the average change is significant for at least one locus. *Double dotted lines,* Boundaries consistently showing high-ranking rates of change (table 4, last col.). *Solid lines,* Other boundaries. The systems contributing to the significance of each boundary are given. Some boundaries are arbitrarily located in the sea.

Table 4

Test of the Overall Probability of the Observed Rate of Gene-Frequency Change across Linguistic Boundaries

Boundary[a]	Observed Sum	Expected Sum	N[b]	SAP[c]
1	5.51	5.50	11	.013
4	6.88	5.50	11	1.338
5	7.30	5.50	11	1.746*
6	24.04	30.50	61	−2.586
7a..................	9.27	9.00	18	.205
7b..................	9.48	8.50	17	.754
10..................	6.75	14.00	28	−4.343
11..................	21.79	30.50	61	−3.489
12..................	15.59	14.00	28	.954
13..................	24.58	30.50	61	−2.371
14..................	.00	3.00	6	−3.869
15a	7.17	14.00	28	−4.106
15b	22.47	23.00	46	−.244
15c	27.32	29.50	59	−.888
16a	32.97	26.50	53	2.781**
16b	33.44	25.50	51	3.481***
17..................	13.88	10.00	20	2.767**
18..................	43.30	26.50	53	7.261***
19..................	4.14	4.50	9	−.385

[a] Codes as in fig. 1.

[b] Number of surfaces considered.

[c] Deviation, expressed in terms of standardized average percentiles.

* $.01 < P \leq 0.05$.

** $.001 < P \leq .01$.

*** $P \leq .001$.

4.13), is determined by the high frequency of the *cde* haplotype in the extreme south of Puglia (eastern tip of the peninsula). Boundaries 16a, 17, and 18 all reflect population differences in Hp allele frequencies between eastern Sardinia and the Ligurian-Tyrrhenian coast of mainland Italy.

3. *Genetic Boundaries, Geographic and Linguistic Barriers*

Seven zones of rapid genetic change (abstracted as solid lines in fig. 4) are apparent in the overall wombling analysis of the 61 surfaces. Sixteen further genetic boundaries are significant when analyzed individually (dashed lines). All but one of the boundaries recognized in the overall analysis were significant when the surfaces were individually considered.

There is a pattern in the geographic distribution of genetic boundaries. They surround two wide regions, one corresponding to the northeastern corner of the country and one to Sardinia. All dialect boundaries in these regions were found significantly associated with high genetic variation by the second approach; high rates of genetic change are apparent within them as well. In addition, significant genetic variation is observed along the Apuan Alps and the central section of the Apennine mountains, and across the peninsula in approximate correspondence with the highly significant boundary 13 detected through the second approach. Sharp genetic differences are not detected between Sicily and continental Italy, but a genetic boundary in the south of Sicily is significant in the individual analysis.

Discussion

In the earlier study, Barbujani and Sokal (in press) showed patterns of gene frequencies that can be partly accounted for by mechanisms of isolation by distance (Sokal and Wartenberg 1983; Barbujani 1987). Under this model, the genetic relationships between near populations depend on past and current gene flow (Cavalli-Sforza 1984), "near" meaning localities that keep exchanging individuals also in the absence of

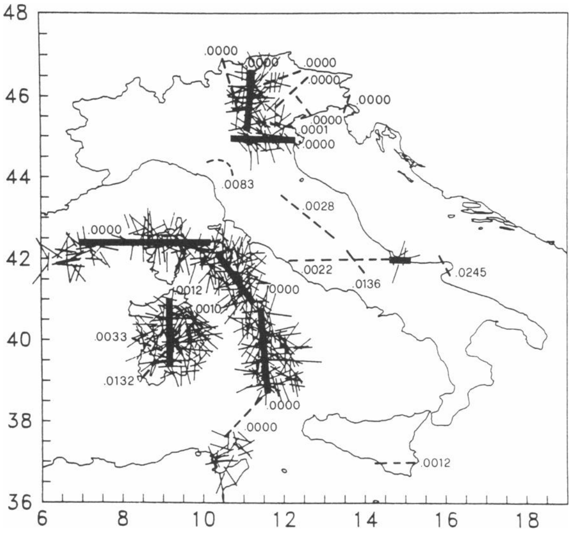

Figure 4 Genetic boundaries determined by the wombling procedure. The figure is a plot of the highest 5% of the average derivatives of the gene-frequency surfaces. The length of the rods is proportional to the overall magnitude of gene-frequency change, and their orientation is the direction of overall maximum slope of the gene-frequency gradients. The genetic boundaries observed in the overall analysis and those yielding significant results in the individual analyses are abstracted as thick solid and dashed lines, respectively; numbers next to the boundaries are probabilities indicating their significance.

major directed migrations. The rates of gene flow surely varied with time and place; on the average, the patches of homogeneous allele frequencies thus generated span from 100 to several hundred kilometers. Large-scale processes must be hypothesized to explain genetic differences beyond this range.

The Mantel tests carried out show that a fraction of genetic variation correlates more with linguistic, rather than geographic, distance. This is the case for nine markers (6 loci) mapping on different regions of the genome (1p36.2-p34 for Rh; 4q12 for Gc; 4q28-31 for MN; 6p21.3 for HLA-A and HLA-B; 9q34 for ABO; McKusick 1988), which rules out linkage as a major confounding factor. There are six systems showing association with geography when the effects of linguistic variation are kept constant. In only one of them is the correlation GEN-LAN.GEO higher than GEN-GEO.LAN.

There are five systems whose variation is related more to geographic than to linguistic factors, but only three of them show significant GEN-LAN association as well, namely Rh (system 4.19), G6PD, and Th. G6PD and Th were protected polymorphisms in malarial regions until 45 years ago (Livingstone 1971, 1983). Although there is no apparent heterozygote advantage today (Canella et al. 1988), these areas still show characteristically high frequencies of the rarer alleles, resulting in association of genetic with geographic distances. On the other hand, three formerly malarial areas (Po delta, Puglia, Calabria) are crossed by dialect-group and dialect-family boundaries, that is, they are not linguistic units, and so the genetic distances for G6PD and Th do not correlate with linguistic distances when the effect of geography is held constant. There is no obvious adaptive explanation for the association of Rh (system 4.19) allele frequencies with geography, especially in view of the fact that the other two systems describing variation at the Rh locus behave differently. However, genetic distances for Rh (system 4.19) were expected to correlate with geographic distances, since the allele frequencies at this system form a north-sough cline across all of Italy (Barbujani and Sokal, in press). Apart from the three exceptions listed above, the observed correlation between genetic and linguistic distances is not due to the spatial clustering of linguistically homogeneous groups. On the contrary, it indicates that linguistic differences have an effect on the genetic structure of populations, either directly or because they are associated with other social and cultural differences, which in turn influence mating and/or dispersal of individuals. On a European scale, Sokal (1988) demonstrated that language does contribute to genetic differentiation of populations, but exerts a smaller effect than that caused by geographic distances. Within Italy, conversely, the relative weight of linguistic and geographic differences seems reversed, although the former are mostly differences in the dialect, not the language, spoken. A likely (but not necessarily the only) explanation for that may be that at a continental scale the linguistic differences add little to the levels of genetic differentiation caused by the large physical distances between many pairs of populations. This may not be the case at a smaller scale, i.e., within a single country, even in the presence of comparatively minor linguistic barriers.

Spatial autocorrelation indices (Sokal and Oden 1978) and similar statistics assume stationarity (Matheron 1970) of the gene frequency distributions they are applied to; the rate of gene-frequency change is assumed to be constant over the area considered. On the whole, this is a reasonable assumption, but there are zones where this does not hold true. Figure 2 shows some examples in an allele-frequency surface. In this study we mapped these zones and looked for the evolutionary factors that may have caused them.

The second approach employed here revealed that the highest average rates of genetic change are between Sardinians and speakers of GI dialects, and between MD and MI speakers. Equally important from a population genetic viewpoint is the finding that various linguistic boundaries (11 out of 19 that could be tested) are associated with sharper-than-expected genetic change. Wombling showed that two areas are surrounded by genetic boundaries, and demonstrated additional zones of rapid gene-frequency change. The genetic boundaries recognized within the northeastern region separate VE from FR and G speakers, with further variation occurring around the Po river (discussed below). However, within Sardinia the genetic boundaries do not coincide with the line separating the two dialect groups. Two more findings appear worth stressing, namely, the lack of genetic differentiation of Sicily and the significant variation detected by overall wombling in the Molise region.

What then are the factors accounting for large-scale genetic differentiation of population units in Italy? Areas of abrupt genetic change may result either from abrupt ecologic change causing an adaptive response, or from scarce exchange of genes across a reproductive barrier (Endler 1977). Although massive movements of populations in Italian prehistory are considered unlikely (Adams et al. 1978; Barker 1981), few anthropologists would agree that the population characteristics we can study now result from adaptations to local environmental conditions. Strictly speaking, differential selection cannot be rejected based only on genetic evidence (Felsenstein 1982), but the view that the extant structure of populations reflects mainly demographic events is much more widespread (see, e.g., Cavalli-Sforza 1966; Cavalli-Sforza and Edwards 1967; Wijsman and Cavalli-Sforza 1984; Piazza et al. 1987; Slatkin 1989). Comparison of figure 4 with a physical map of Italy shows that most genetic boundaries correspond not only to linguistic barriers, but also to physical obstacles to population movements. In general, therefore, it is reasonable to conclude that zones of abrupt genetic change (1) exist, (2) result from anisotropies in the migration patterns, and (3) are associated with obstacles to migration; however, it is not easy to discriminate between the effects of geographic and linguistic barriers, as they largely overlap. As a consequence, it is often impossible to say whether differences in the dialect spoken reinforce the isolating effect of geographic barriers, or result from such an isolating effect.

Let us then examine the modes of genetic variation in zones that are associated with either physical or cultural isolating factors, but not with both. Rapid genetic change was observed in central Italy across mountain ranges that do not separate different dialects (fig. 4). On the other hand, although physically an island, Sicily does not appear to be genetically isolated, as already observed by Beretta et al. (1986). As for the effects of linguistic factors alone, the speakers of at least one dialect, FR, are sharply differentiated genetically even in the absence of geographical constraints to gene flow around them; so much so that a genetic boundary significant at $P < .0001$ occurs even at the short linguistic boundary 7b that separates FR speakers from the eastern group of VE speakers. A second example of abrupt genetic variation that can be related only to linguistic variation is the southern boundary of the VE-speaker population. This boundary is marked by the lower course of the Po river, which surely constrained gene flow in the past. However, there is no significant genetic variation across the western section of the same river, flowing through a linguistically homogeneous area. The width of the river does not change enough to justify such a difference in its migration-preventing effect. The third observation supporting language differences as an isolating factor is the genetic boundary, detected by wombling, around Albanian speakers in Molise. Thus, although generally associated with geographic factors, linguistic differences by themselves appear sufficient to prevent population admixture at various locations.

In synthesis, we have seen that, (1) apart from malaria-protected polymorphisms, the association between genetic and linguistic distances in Italy is stronger than that between genetic and geographic distances, (2) several linguistic boundaries display larger-than-expected genetic change, (3) the zones of rapid genetic change are associated with both physical and linguistic barriers but (4) some genetic boundaries overlap with linguistic, not geographic, barriers, and (5) the Po river is associated with abrupt genetic change only where it separates two different dialect families.

Piazza et al. (1988) related the spatial distributions of synthetic variables calculated from gene frequencies to the locations of three ancient peoples inhabiting Italy. They recognized three major components of variation, and associated these with genetic substrates due to Greek, Etruscan, and Ligurian populations. Since their data form a subset of the data employed by us, we would expect to find the three components if we applied their methodology to our data. Various population movements along certain directions, followed by population admixture, are expected to result in clinal or partly clinal distributions of allele frequencies such as those detected in the study by Piazza et al. Our approaches 2 and 3 emphasized sharp changes and discontinuities in the gene-frequency surfaces and would not necessarily demonstrate the continuous change reported by Piazza et al. (1988). Therefore, this study cannot confirm or reject the hypotheses put forth by these authors. Also, the hypotheses we tested were somewhat different, although there is no doubt that extant linguistic differences are the product of past events, including migration-admixture processes. What we can say, though, is that sharp genetic change is apparent around two areas where high scores of the synthetic variables were observed by Piazza et al. (1988), suggesting centers of origin for alleles that later diffused in the population. These areas are southern Calabria-Sicily and Central Italy (associated, respectively, with Piazza et al's Greek and Etruscan components). If Piazza's hypotheses are correct, such abrupt genetic change linked to dialect differences supports cultural (linguistic) isolation as a factor slowing down population admixture.

All this appears consistent with the modified gene-flow model (Harding and Sokal 1988) outlined in a previous section. Differential selection has affected a limited number of genetic systems, whose variation correlates with the distribution of the areas where they maintained their polymorphism. However, most gene frequencies have been homogenized by gene flow. This effect is generally apparent within, but not always among, linguistically homogeneous regions. As a consequence, the distributions of most gene frequencies are patchy. This causes departures at great distances from the gene-frequency patterns predicted by models of isolation by distance. It is often impossible to assess the origin of the association between genetic and linguistic variation, but there are cases in which linguistic differentiation appears to maintain, and not only to be correlated with, genetic differentiation.

Acknowledgments

This is contribution 771 in Ecology and Evolution from the State University of New York at Stony Brook. We are indebted to Donna DiGiovanni, Marie-Josée Fortin, Barbara Thomson, and Chester Wilson for technical assistance in carrying out this research. Claudio Friso prepared the illustrations. Prof. Carlo Paganotto provided valuable suggestions on the treatment of linguistic variation. This research was carried out with support of grant GM 28262 from the National Institutes of Health to RRS.

References

Adams WY, Van Gerven DP, Levy RS (1978) The retreat from migrationism. Annu Rev Anthropol 7:483–532

Ammerman AJ, Cavalli-Sforza LL (1984) The Neolithic transition and the genetics of populations in Europe. Princeton University Press, Princeton, NJ

Barbujani G (1987) Autocorrelation of gene frequencies under isolation by distance. Genetics 117:777–782

——— (1988) Diversity of some gene frequencies in European and Asian populations. IV. Genetic population structure assessed by the variogram. Ann Hum Genet 52: 215–225

Barbujani G, Oden NL, Sokal RR (1989) Detecting areas of abrupt change in maps of biological variables. Syst Zool 38:376–389

Barbujani G, Sokal RR (1990) The zones of sharp genetic change in Europe are also language boundaries. Proc Natl Acad Sci USA 87:1816–1819

——— Genetic population structure of Italy. I. Geographical patterns of gene frequencies. Hum Biol (in press)

Barker G (1981) Landscape and society: prehistorical central Italy. Academic Press, London

Barrantes R, Smouse PE, Mohrenweiser HW, Gershowitz H, Azofeifa J, Arias TD, Neel JV (1990) Microevolution in lower Central America: genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a consensus taxonomy based on genetic and linguistic affinity. Am J Hum Genet 46:63–84

Barton NH, Jones JS (1990) The language of genes. Nature 346:415–416

Beretta M, Mazzetti P, Frosina G, Schiliro G, Russo A, Russo G, Barrai I (1986) Population structure of eastern Sicily. Hum Hered 36:379–387

Canella R, Barbujani G, Cucchi P, Siniscalco M, Vullo C, Barrai I (1988) Biological performance in beta-thal heterozygotes and normals: results of a longitudinal comparison in a former malarial environment. Ann Hum Genet 51:337–343

Cavalli-Sforza LL (1966) Population structure and human evolution. Proc R Soc Lond [B] 164:362–379

——— (1984) Isolation by distance. In: Chakravarti A (ed) Human population genetics: the Pittsburgh symposium. Van Nostrand-Reinhold, New York, pp 229–247

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet 19:233–257

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 85:6002–6006

Dougenik JA, Sheehan DE (1979) SYMAP user's reference manual, version 5.20. Camera Stat., Bedford, MA

Endler JA (1977) Geographic variation, speciation, and clines. Princeton University Press, Princeton, NJ

Excoffier L, Harding RM, Sokal RR, Pellegrini B, Sanchez-Mazas A. Spatial differentiation of rhesus and Gm haplotype frequencies in sub-Saharan Africa and its relation to linguistic affinities. Hum Biol (in press)

Felsenstein J (1982) How can we infer geography and history from gene frequencies? J Theor Biol 96:9–20

Friedlaender JS, Sgaramella-Zonta LA, Kidd KK, Lai LYC, Clark P, Walsh RJ (1971) Biological divergences in south-central Bougainville: an analysis of blood polymorphism gene frequencies and anthropometric measurements utilizing tree models, and a comparison of these variables with linguistic, geographic and migrational "distances." Am J Hum Genet 23:253–270

Goebl H (1981) Éléments d'analyse dialectométrique (avec application a l'AIS). Rev Linguistique Romane 45:349–420

——— L'espace géolinguistique en perspective dialectométrique. In: Piazza A, Cavalli-Sforza LL (eds) Language change and biological evolution. Stanford University Press, Stanford, CA (in press)

Harding RM, Sokal RR (1988) Classification of European language families by genetic distance. Proc Natl Acad Sci USA 85:9370–9372

Hedrick PW, Thomson G (1983) Evidence for balancing selection at HLA. Genetics 104:449–456

Klitz W, Thomson G, Baur MP (1986) Contrasting evolutionary histories among tightly linked HLA loci. Am J Hum Genet 39:340–349

Livingstone FB (1963) Blood groups and ancestry: a test case from the New Guinea highlands. Curr Anthropol 4:541–542

——— (1967) Abnormal hemoglobins in human populations. Aldine, Chicago

——— (1971) Malaria and human polymorphisms. Annu Rev Genet 5:33–64

——— (1983) The malaria hypothesis. In: Bowman JE (ed) Distribution and evolution of the hemoglobin and globin loci. Elsevier, New York, pp 15–44

McKusick VA (1988) Mendelian inheritance in man, 8th ed. Johns Hopkins University Press, Baltimore and London

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Mathéron G (1970) La théorie des variables regionalisées, et ses applications. Centre de Morphologie Mathématique, Fontainebleau

Menozzi P, Piazza A, Cavalli-Sforza LL (1978) Synthetic maps of human gene frequencies in Europeans. Science 201:786–792

Mourant AE, Kopéc AC, Domaniewska-Sobczak K (1976) The distribution of human blood groups. Oxford University Press, Oxford

Pellegrini GD (1977) Carta dei dialetti d'Italia. Pacini, Pisa

Piazza A, Cappello N, Olivetti E, Rendine S (1988) A genetic history of Italy. Ann Hum Genet 52:203–213

Piazza A, Menozzi P, Cavalli-Sforza LL (1981*a*) The making and testing of geographic gene frequency maps. Biometrics 37:635–659

——— (1981*b*) Synthetic gene frequency maps of man and selective effects of climate. Proc Natl Acad Sci USA 78: 2638–2642

Piazza A, Olivetti E, Barbanti N, Reali G, Domenici R, Giari A, Benciolini P, et al (1989) The distribution of some polymorphisms in Italy. Gene Geogr 3:69–139

Piazza A, Olivetti E, Carbonara O, Bargagna M, Pecori F, Benciolini P, Cortivo P, et al (1982) La distribuzione di alcuni polimorfismi genetici in Italia. Il Ponte, Milano

Piazza A, Rendine S, Zei G, Moroni A, Cavalli-Sforza LL (1987) Migration rates of human populations from surname distributions. Nature 329:714–716

Prevosti A, Ocana J, Alonso G (1975) Distances between populations of *Drosophila suboscura* based on chromosome arrangement frequencies. Theor Appl Genet 45: 231–241

Rendine S, Piazza A, Cavalli-Sforza LL (1986) Simulation and separation by principal components of multiple demic expansions in Europe. Am Nat 128:681–706

Salzano FM, Neel JV, Gershowitz H, Migliazza EC (1977) Intra and intertribal genetic variation within a linguistic group: the Ge-speaking Indians of Brazil. Am J Phys Anthropol 47:337–348

Serjeantson SW, Kirk RL, Booth PB (1983) Linguistic and genetic differentiation in New Guinea. J Hum Evol 12: 77–92

Slatkin M (1975) Gene flow and selection in a two-locus system. Genetics 81:787–802

——— (1985) Gene flow in natural populations. Annu Rev Ecol Syst 16:393–430

——— (1989) Population structure and evolutionary progress. Genome 31:196–202

Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Zool 35:627–632

Sokal RR (1979) Testing statistical significance of geographic variation patterns. Syst Zool 28:227–231

——— (1988) Genetic, geographic, and linguistic distances in Europe. Proc Natl Acad Sci USA 85:1722–1726

Sokal RR, Harding RM, Oden NL (1989*a*) Spatial patterns of human gene frequencies in Europe. Am J Phys Anthropol 80:267–294

Sokal RR, Menozzi P (1982) Spatial autocorrelation of HLA frequencies in Europe support demic diffusion of early farmers. Am Nat 119:1–17

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology. I. Methodology. Biol J Linnean Soc 10:199–228

Sokal RR, Oden NL, Legendre P, Fortin M-J, Kim J, Vaudor A (1989*b*) Genetic differences among language families in Europe. Am J Phys Anthropol 79:489–502

Sokal RR, Oden NL, Legendre P, Fortin M-J, Kim J, Vaudor A, Harding RM, Barbujani G (1990) Genetics and language in European populations. Am Nat 135:157–175

Sokal RR, Oden NL, Thomson BA (1988) Genetic changes across language boundaries in Europe. Am J Phys Anthropol 76:337–361

Sokal RR, Rohlf FJ (1981) Biometry, 2d ed. WH Freeman, San Francisco

——— (1987) Introduction to biostatistics, 2d ed. WH Freeman, New York

Sokal RR, Wartenberg DE (1983) A test of spatial autocorrelation analysis using an isolation-by-distance model. Genetics 105:219–237

Spuhler JN (1972) Genetic, linguistic, and geographical distances in native North America. In: Weiner JS, Huizinga J (eds) The assessment of population affinities in man. Clarendon Press, Oxford, pp 72–95

Tills D, Kopéc AC, Tills RE (1983) The distribution of the human blood groups and other polymorphisms, suppl 1. Oxford University Press, London

White NG, Parsons PA (1973) Genetic and socio-cultural differentiation in the aborigines of Arnhem Land, Australia. Am J Phys Anthropol 38:5–14

Wijsman EM, Cavalli-Sforza LL (1984) Migration and genetic population structure with special reference to humans. Annu Rev Ecol Syst 15:279–301

Womble WH (1951) Differential systematics. Science 114: 315–322