AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 79:489–502 (1989)

Genetic Differences Among Language Families in Europe

ROBERT R. SOKAL, NEAL L. ODEN, PIERRE LEGENDRE, MARIE-JOSÉE FORTIN, JUNHYONG KIM, AND ALAIN VAUDOR
Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794-5245 (R.R.S., M.-J.F., J.K.); Department of Community and Preventive Medicine, Division of Epidemiology, Health Sciences Center, SUNY, Stony Brook, New York, 11794-8036 (N.L.O.); and Département de Sciences biologiques, Université de Montréal, C.P. 6128, Succursale A, Montréal, Canada H3C 3J7 (P.L., A.V.)

KEY WORDS Gene frequencies, Cranial variables, European populations, Languages

ABSTRACT We investigated whether 59 allele frequencies and 10 cranial variables differed among speakers of the 12 modern language families in Europe. Although this is a classical analysis of variance design, special techniques had to be developed for the analysis because of spatial autocorrelation of both biological and language data. The method examines pooled sums of squares within language families. These are compared with the same quantities obtained by randomly partitioning the available data points in Europe into internally cohesive subsets representing the same sample sizes for each language family as in the originally observed data. Our results suggest that for numerous genetic systems, population samples differ more among language families than they do within families. These findings are considered in relation to two contrasting models: a model of random spatial differentiation of gene frequencies unrelated to language and a model of aboriginal genetic differences among speakers of different language groups. Our observed findings suggest partial validity of both models.

Human populations in Europe differ considerably in genetic and morphometric traits. Genetic differentiation of populations may be due to one or a composite of several forces. The differentiation may be due to sampling variance from gene pools of limited size (genetic drift) (Nei, 1987; Wright, 1969) and to the limited mobility of individuals within the area of study (isolation by distance) (see Endler, 1977; Nei, 1987; Rohlf and Schnell, 1971). Several models attempt to explain the amount of differentiation in terms of distance among sampling units (Malecot, 1973; Morton, 1973, 1982). Alternative explanations interpret the differences in terms of adaptive processes (selection) (e.g., Flint et al., 1986; for review, see Endler, 1986). Finally, differences may be traced to the historical origins of the populations concerned, which in turn may be due to directed migration and settlement patterns (as distinct from the assumption of random dispersal of individuals underlying the isolation by distance model). These alternatives are being investigated in a database of European gene frequencies (and cranial measurements) assembled in our laboratory.

One way to investigate which of these processes has been at work in human populations is to study the relations between the patterns of gene-frequency differentiation and the distribution of the language families in Europe. Since a common language frequently indicates a common origin, and a related language suggests a common origin further back in time, one might expect such commonality of origin to be reflected in genetic relationships. There are, however, several complicating factors. One of these is the well-documented, repeated genetic and linguistic assimilation of disparate ethnic elements within a named ethnic group of migrants, increasing genetic heterogeneity in

Received June 23, 1988; accepted December 16, 1988.

the migrating population. A second factor is that even if immigrant populations are assumed to have been homogeneous, they rarely settled in unoccupied areas of Europe but frequently absorbed the native populations of their settlement area, the resulting admixed population adopting the language of either the natives or the immigrants. Both factors tend to diminish the correspondence between genetic and linguistic similarity. In contrast, since language differences themselves are barriers to free gene flow, they will enhance genetic differentiation, counteracting the earlier two forces to some extent.

It was of interest, therefore, to investigate whether, despite these complications stemming from the fact that language is a culturally transmitted set of traits (see Cavalli-Sforza and Feldman, 1981), it is possible to detect genetic differences among speakers of the different language families of Europe. If such differences exist, and if these can be shown to be not simply artifacts of the spatial patterning of both gene frequencies and language families, the results will permit inferences about the origins of the populations speaking the various languages.

This paper describes and tests the differences in gene frequencies for separate genetic systems (and for some cranial measurements) among the language families of Europe. The analysis of such geographic data presents complications, making it improper to apply conventional statistical tests. The geographically varying gene frequencies are spatially autocorrelated (Harding et al., 1987; Sokal et al., 1989). Most classical statistical techniques assume independent observations or, at least, independent errors. These assumptions of independence are violated when the data are autocorrelated and classical statistical analysis is no longer appropriate. For this reason, tests of the questions addressed in this paper required unconventional approaches to compensate for the inherent spatial autocorrelation of the data.

The approach taken in this study addresses the question of differences among language families and is a nonparametric, permutational analysis of variance. Other approaches investigating different aspects of the relation between genetics and language have been pursued in our laboratory. Sokal et al., (1988) examined whether allele frequencies show increased change at language-family boundaries. Sokal (1988) tested whether genetic and linguistic distances are correlated among these populations.

MATERIALS AND METHODS

Data

For purposes of our study, we consider the boundaries of Europe to lie west of the Urals and north of the Caspian and Caucasus, but to include also all but the easternmost regions of Turkey. We analyzed records of 59 allele frequencies and 10 cranial measurements at 3,466 geographic locations in Europe. The 69 different variables can be grouped into 27 systems, each, with some exceptions, corresponding to a genetic locus (we counted HLA-A and HLA-B as two separate systems, but counted the 10 cranial variables as a single system). The details of this arrangement are shown in Table 1.

To arrive at the 59 gene frequencies, we eliminated complementary allele frequencies in biallelic loci to avoid redundancy of information. Also, for each locus with multiple alleles, we eliminated highly dependent allele frequencies, i.e., those having high correlations over the sample localities with the other allele frequencies of their genetic systems. The number of localities sampled for each of the separate systems (also given in Table 1) ranges from 870 for the ABO system to 33 for the Lutheran system. Because more than one allele frequency was studied for some systems, the 3,466 locations yielded a total number of 7,556 data points to be employed in the analyses. The maximum likelihood gene-frequency estimates are based on samples collected since World War II, ranging in size from 50 to many thousands of persons. The gene frequencies were extracted from Mourant et al. (1976) and Tills et al. (1983) and from the results of an extensive computer search of the recent literature. Frequencies for 21 HLA-A and HLA-B histocompatibility alleles were obtained through the courtesy of P. Menozzi, A. Piazza, and L. Cavalli-Sforza. They were first analyzed by Menozzi et al. (1978). Immunoglobulin polymorphisms for the Gm and Km loci were taken from Steinberg and Cook (1981). Samples identified as Jews or Gypsies, listed separately in the cited sources, were omitted since the geographic locations furnished usually were not sufficiently specific. Samples for nonindigenous populations, e.g., Chinese in London, were not used. The cranial measurements are

TABLE 1. Systems analyzed and numbers of localities

Mourant system no.[1]	System name	Allele (or variable) names				No. of surfaces per system	No. of localities per system
1.1	ABO	I^A	I^B	I^{O**}		2	870
1.2	ABO with anti-A, -A1, and -B	I^{A1}	I^{A2}	I^B	I^{O**}	3	157
2.5	MN	*M*	*N**			1	194
2.7	MN with anti-M, -N, and -S	*MS*	*Ms***	*NS*	*Ns***	2	68
3.1	P	*P1*	*P2+p**			1	102
4.1	Rhesus	*D*	*d**			1	568
4.13	Rhesus with anti-C, -D, -E, and -c	*CDE***	*CDe***	*CdE**	*Cde*	5	82
		cDE	*cDe*	*cdE*	*cde*		
4.19	Rhesus with anti-C, -D, -E, -c, and -e	*CDE*	*CDe***	*CdE**	*Cde*	6	76
		cDE	*cDe*	*cdE*	*cde*		
5.1	Lutheran	*Lu-a*	*Lu-b**			1	33
6.1	Kell	*K*	*k**			1	116
6.3	Kell with anti-K and -k	*K*	*k**			1	39
7.1	ABH secretion	*Se*	*se**			1	53
8.1	Duffy	*Fy-a*	*Fy-b+Fy**			1	108
36.1	Haptoglobin	*Hp-1*	*Hp-2**			1	175
37.1	Transferrin	*Tf-C*	*Tf-B***	*Tf-D*		2	38
38.1	Gc	*Gc-1*	*Gc-2**			1	112
50.1.1	Acid phosphatase	*P-a*	*P-b***	*P-c*		2	72
52	6-Phosphogluconate dehydrogenase	*PGD-A*	*PGD-C***	*PGD-R*		2	42
53	Phosphoglucomutase PGM1	*PGM1-1*	*PGM1-2***	*PGM1-R*		2	70
56	Adenylate kinase	*AK-1*	*AK-2**	*AK-3**		1	64
63	Adenosine deaminase	*ADA-1*	*ADA-2**			1	53
65	Phenylthiocarbamide tasting	*T*	*t**			1	62
100	HLA-A	*A1*	*A2*	*A3*	*A9*	7	66
		A10	*A11*	*A28*			
101–102	HLA-B	*B5*	*B7***	*B8*	*B12*	10	66
		B13	*B14*	*B18***	*B27*		
		BW15	*BW17*	*BW21***	*BW22*		
		BW35	*BW40***				
200	Gm 1,2,5	*Gm5***	*Gm1*	*GM1-2*	*GM1-5**	2	45
201	Inv (Km)	*Km1*	*Km2**			1	38
901–910	Cranial variables†	*1*, (maximum) cranium length; *8*, (maximum) cranium breadth; *9*, (minimum) frontal breadth; *17*, basion-bregma height; *45*, bizygomatic breadth; *48*, Facial height; *51*, Orbital breadth; *52*, Orbital height; *54*, Nasal breadth; *55*, Nasal height				10	97
Totals						69	3,466

[1]The Mourant system numbers are those given by Mourant *et al.* (1976). Mourant system numbers were invented for systems 100–910.
*Alleles that are strictly complementary or have fewer than 10% nonzero values. These alleles were eliminated from all analyses.
**Alleles that are strongly correlated with other alleles in the same system and were also eliminated from the analyses.
†The individual cranial variables are preceded by their code numbers as assigned by Martin and Saller (1957).

means based on sample sizes of at least 25 skulls from populations dated between 1500 AD and the present (Schwidetzky and Rösing, 1984). The samples range widely over the continent of Europe. Their distribution, pooled for the different systems, is illustrated in an earlier publication in this journal (Sokal et al., 1988, Fig. 2).

We recognize 12 language families in our study area (see Ruhlen, 1987). The families, preceded by their phyla, in capitals, are INDO-EUROPEAN: Albanian, Baltic, Celtic, Germanic, Greek, Romance, Slavic: FINNO-UGRIC: Finnic, Ugric (Hungarian); ALTAIC: Turkic; AFRO-ASIATIC: Semitic (Maltese); LANGUAGE ISOLATES: Basque. The language family areas and boundaries were obtained from a number of sources (Cowgill, 1976; Harms, 1976; Ivanov, 1976; Meillet and Cohen, 1952a,b;

Mather et al., 1975; Moulton et al., 1976; Posner, 1976; von Czoernig, 1984). The areas are shown in Figure 1. Samples located close to language boundaries were investigated carefully to ascertain the language actually spoken by the persons in the sample. This was relatively easy when there were sharp language transitions, as along much of the boundary between the Germanic and Romance languages through Belgium or in regions of Switzerland. In other situations, e.g., in the Germanic–Romance boundary coursing through the Alto Adige region of northern Italy, the exact location of a language boundary is difficult to place. In such cases a sample was assigned to the language family of the majority of speakers. There are, however, very few samples in the data base for which there is any doubt about the language spoken by the persons sampled.

Method

We wish to test whether mean gene frequencies of samples belonging to different language families differ significantly from each other. This question corresponds to a

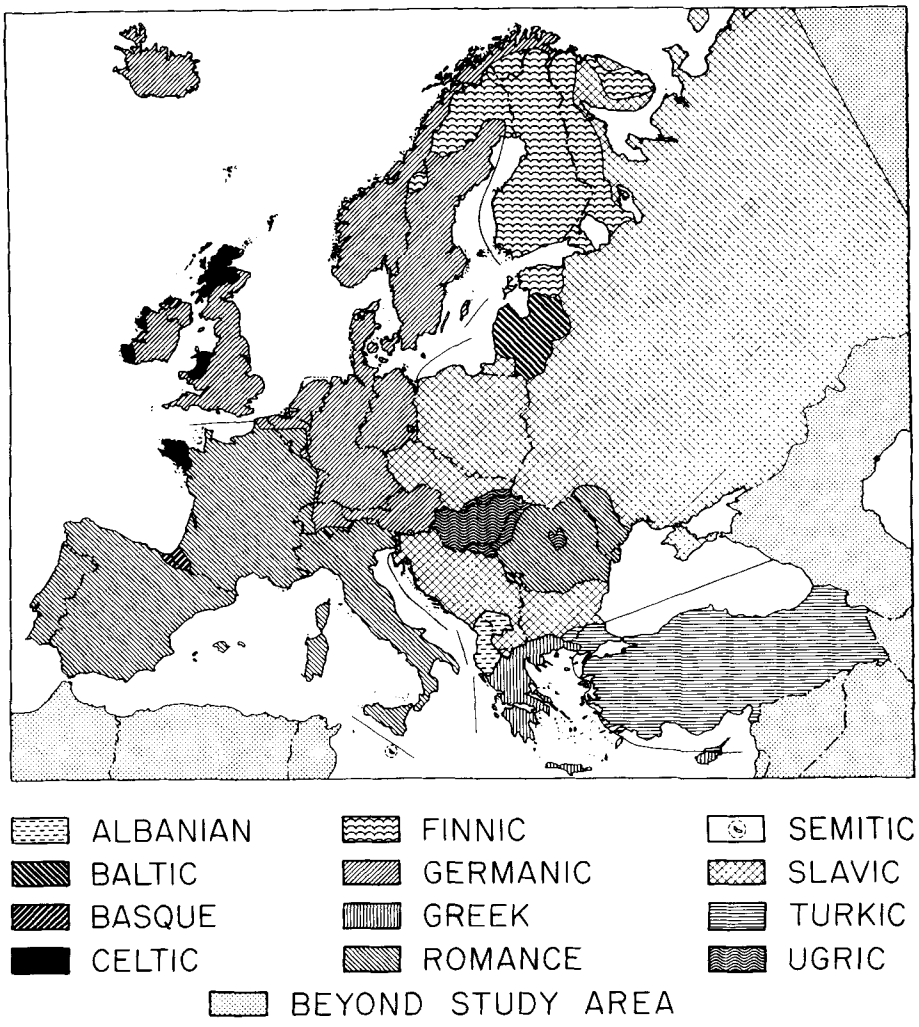

Fig. 1. Language family areas in Europe. Only the major areas are shown. Areas for some non-Slavic populations in the Soviet Union have been omitted, since we lack the data to test them.

classical analysis of variance design among language families. However, it cannot be analyzed by analysis of variance because of the spatial autocorrelation of the data known to exist for virtually all of the gene frequencies tested (Harding et al., 1987; Sokal et al., 1989). As Cliff and Ord (1981) have shown, such a property of the underlying variable violates the assumptions of analysis of variance. Because the individuals constituting a single locality sample are generally considered to have come from a single point, spatial autocorrelation of individuals *within* a locality is usually ignored. However, when, as in the present study, the separate locality samples are grouped into classes, with the criterion of classification having a spatial aspect, the positive spatial autocorrelation of the replicate localities in each class will inflate the nominal significance of the differences among classes. In this study the locality samples are grouped by language families. The samples within one family tend to be neighbors, hence will be spatially autocorrelated. Thus by standard analysis of variance we might obtain differences among the language families that are due to the autocorrelation of individual allele frequency samples rather than reflecting inherent differences in mean gene frequency among the language families.

To overcome this problem we decided to employ a nonparametric permutational approach. For each allele frequency tested, we investigate whether the partition of the sample points into groups corresponding to the observed language families yields a pooled within-group sum of squares lower than that obtained when the data set is randomly partitioned into comparable geographically cohesive groups. The basic idea of this permutational approach is to randomize the allocation of localities to language-family areas while keeping constant the spatial dependencies inherent in the data. This permutation with spatial constraints can be contrasted with an ordinary permutational approach that mimics a conventional analysis of variance. We could simply permute the observed values of the gene frequencies at random over the map, group them into language families by the established language-family boundaries, and compare the observed partition to those obtained by repeated random permutations. Such random permutations would destroy any spatial autocorrelation inherent in the observations. Should the results show the observed partition to be significantly more differentiated among language families than the random partitions (equivalent to a significant analysis of variance), it would not be clear whether this finding was due to actual gene frequency differences among the language families or to the spatial autocorrelation within these.

To carry out a random permutation of localities successfully while retaining their spatial dependencies requires that we make the following mimicking assumption: The shapes of the original language-family areas are reasonably approximated by a computer algorithm. The exact algorithms for carrying out the rather involved computations are described in a technical companion paper (Legendre, et al., 1989), and only a rough outline needs to be presented here. Of the two algorithms described in that paper only the first, the ring algorithm, was employed for the present study mainly because the second, the random-tree algorithm, had not yet been developed when the computations reported here were carried out.

To meet the mimicking assumption, these random partitions of samples had to obey two requirements. The first is intended to reflect the geographical compactness of the original language families: Each randomly formed group constituting a pseudofamily must also be geographically compact. Because geographical contiguity is difficult to represent numerically, we used the idea of connectedness and connected the sample localities with a Delaunay triangulation (Brassel and Reif, 1979). This is a connection network tending to link close neighbors. A different triangulation had to be computed for each genetic system, since the number of localities differed for each system. Figure 2 shows an example of a Delaunay triangulation for one of the smaller systems employed in the study. By limiting randomly generated areas to sets of connected localities (connected subgraphs of the triangulation), the pseudofamilies form relatively cohesive areas. Links between two localities greater than 1,500 km were removed to avoid the creation of exceedingly elongated, spidery areas. The second requirement is that each pseudofamily correspond in sample size (number of locality samples observed) to that of the observed language family.

In the ring algorithm, seed points for the pseudoareas are chosen at random among

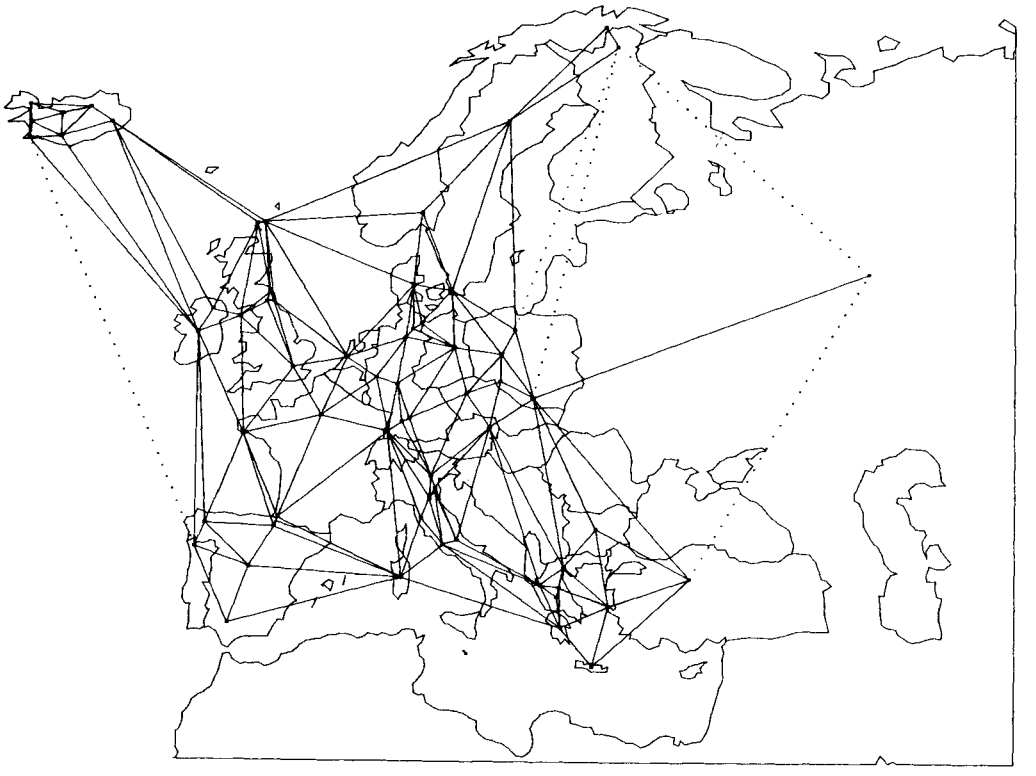

Fig. 2. A Delaunay triangulation for a representative allele frequency, adenylate kinase. Links between any two localities greater than 1,500 km are indicated by dotted lines and were not included in our computations.

the localities on the map. Then each group is grown in steps, by attaching concentric rings of points around the seed locality (hence the name of the algorithm), following the connecting graph. When pseudoareas meet, growth is no longer possible along their common border, and each one has to grow in different directions as available points permit. When ring growth is no longer possible, another procedure takes over that forces the incomplete groups to grow at the expense of their neighbors. This goes on until all pseudogroups have reached the required number of localities, i.e., the same number as in the observed geographic areas they mimic. If this turns out to be too tedious, the incomplete pseudomap is abandoned and the procedure is restarted from the beginning.

The procedure described above will preserve areas as well as sample size only if the observed samples are located on a regular lattice imbedded in the geographic area under investigation. In studies of human variation such conditions can rarely, if ever, be met. For this reason the validity of the mimicking assumption must be investigated in instances of irregularly distributed sampling stations. Legendre et al. (1989) in describing the methodology in detail give an example from one of the genetic systems (ABO system 1.1) of the present study. They test the validity of the mimicking assumptions in detail by several simulations and conclude that the various simulation results build confidence for the application of the technique to the human gene frequency data analyzed in this study.

For each system in Table 1 the partitioning was carried out over all available language families simultaneously, such that each permutation produced a pseudolanguage-area map of Europe with all the origi-

nal language families represented. The permutations were repeated a sufficient number of times to result in the production of 249 random partitions for each of the systems. (Not all partitioning attempts were successful.) The test statistic employed was the pooled sum of squares within language families for each putatively independent allele frequency within a system. We also recorded the separate sums of squares within each of the language families. Following the procedure developed by Hope (1968) we added the single observed value to those obtained from the 249 random partitions, yielding a distribution of 250 computed test statistics. Sums of squares observed for the actual language families were compared with distributions of sums of squares based on the simulated language families to obtain an estimate of the significance of the observed genetic homogeneity of the language families.

For several language families the spatial distributions of the populations belonging to the families are disjunct. Specifically, Romance speakers in Romania are separated from the main Romance area by speakers of Slavic. South Slavs are separated from West and East Slavs by Hungarian and Romanian speakers, respectively. The modern distribution of Celtic speakers is quite disjunct. Other minor disjunct areas exist for Finnic, Germanic, Slavic, Turkic, and Ugric speakers. Our simulation approach cannot exactly reproduce this situation, since we constrain the pseudofamily areas to be compact and connected. A program to simulate disjunct areas would have involved additional complexity. Also, many of the small disjunct areas, e.g., among the Celtic speakers, would contain at most one observation, making it impossible to calculate sums of squares. Therefore our simulations cannot exactly mimic the actual situation, since, strictly speaking, the observed outcomes are not contained in the family of all possible partitions that can be generated by the simulation program. However, this consideration does not invalidate our approach for the following reason. It is quite improbable that language families represented by sets of disjunct areas are more homogeneous than those contained in a single contiguous area. Dismissing that possibility, we are left with the alternative that disjunct areas are equal in homogeneity or are more heterogeneous than contiguous ones. If we still find homogeneity within language family areas by our methods, then we can be certain that the homogeneity is real. Our test is therefore a conservative one, and any significance that we find is likely to be well established and would surely have been significant also had we carried out randomizations that could have resulted in disjunct language areas.

RESULTS

The 69 (unweighted) mean allele frequencies and cranial measurements for the 12 language families are shown in Table 2. Substantial apparent differences among language families are evident for some of the variables studied. It is the aim of our study to test whether such apparent differences are statistically significant.

Since positive spatial autocorrelation of the gene frequency surfaces diminishes the pooled within-language-family variance and hence increases the nominal significance of an analysis of variance, one can be reasonably certain that gene frequency differences that are not significant by a conventional analysis of variance would not become significant by the techniques applied in our study. For this reason we first carried out regular analyses of variance for each of the 69 variables, testing for differences among language families. Only in four systems (MN, Lutheran, ABH secretion, and adenylate kinase) did none of the allele frequency surfaces prove significant by analysis of variance. For this reason the four allele frequencies representing these four systems were excluded from further analyses, and we conclude that they show no differences among language families.

The results of the permutations for the 65 remaining allele frequencies and cranial variables are shown in column 1 of Table 2, expressed as probabilities $\times$ 100. Of these 65 probabilities, 23 (or 0.354) show significantly greater homogeneity ($P \leq 0.05$) within the observed language families than by chance. This is substantially higher than the conventional type I error rate. We condensed these results by calculating experimentwise error rate (Bonferroni) probabilities (Sokal and Rohlf, 1987) over all variables for each system. For instance, in system 1 (ABO), the allele with the smallest probability yielded $P = 0.004$. The Bonferroni adjustment consisted of multiplying this probability by the number of independent tests performed simultaneously: 0.004

TABLE 2. Summary of significance tests and unweighted means of 59 allele frequencies and 10 cranial variables for the language families of Europe

Variables	Columns 1	Columns 2	Language families G	I	S	F	C	H	K	Q	T	B	Y	E
1.1 I^A	15	2	263	268	268	296	301	254	203	251	299	244	241	—
1.1 I^B	0		072	069	142	120	139	095	066	037	121	151	121	—
1.2 I^{A1}	53		188	211	236	196	222	226	239	206	221	—	—	186
1.2 I^{A2}	12		075	056	053	124	072	101	056	017	074	—	—	071
1.2 I^B	7		075	079	144	104	156	116	050	013	119	—	—	052
2.5 *M*	—	—	569	560	584	544	572	662	582	526	571	—	—	—
2.7 *MS*	75	52	226	246	249	245	265	227	309	—	255	—	—	267
2.7 *NS*	26		064	092	079	089	103	103	087	—	088	—	—	035
3.1 *P1*	10	10	500	491	497	442	395	472	518	514	—	—	—	290
4.1 *D*	4	4	588	628	598	662	597	678	570	438	677	—	667	—
4.13 *Cde*	13	2	008	018	013	010	051	021	005	018	—	—	—	—
4.13 *cDE*	0		154	115	143	126	167	075	114	061	—	—	—	—
4.13 *cDe*	16		023	035	023	033	022	061	012	040	—	—	—	—
4.13 *cdE*	42		009	005	005	005	007	000	011	008	—	—	—	—
4.13 *cde*	14		381	370	378	312	412	328	411	435	—	—	—	—
4.19 *CDE*	46	19	003	004	007	—	004	000	—	000	000	—	—	000
4.19 *Cde*	38		013	021	026	—	008	018	—	012	014	—	—	000
4.19 *cDE*	3		123	106	136	—	205	136	—	064	171	—	—	151
4.19 *cDe*	21		022	030	030	—	026	062	—	015	013	—	—	036
4.19 *cdE*	33		011	004	004	—	001	000	—	006	000	—	—	000
4.19 *cde*	20		389	302	394	—	394	317	—	490	320	—	—	308
5.1 *Lu-a*	—	—	026	016	025	011	—	020	004	041	009	—	—	017
6.1 *K*	10	17	049	038	056	004	046	039	066	043	036	—	—	070
6.3 *K*	8		048	039	053	020	039	129	043	—	—	—	—	—
7.1 *Se*	—	—	504	542	552	533	572	—	—	409	—	—	—	—
8.1 *Fy-a*	10	10	439	395	469	463	438	428	388	301	458	—	—	327
36.1 *Hp-1*	16	16	391	370	394	381	363	326	—	419	336	338	—	—
37.1 *Tf-C*	19	18	995	992	992	987	—	998	—	983	—	—	—	—
37.1 *Tf-D*	9		001	002	000	011	—	001	—	000	—	—	—	—
38.1 *Gc-1*	2	2	731	689	716	803	700	744	—	665	743	718	—	—
50.1.1 *P-a*	8	7	355	302	304	393	359	328	383	248	292	—	—	—
50.1.1 *P-c*	4		057	066	068	048	045	032	030	017	034	—	—	—
52 *PGD-A*	10	20	979	981	988	968	—	961	985	988	999	—	—	—
52 *PGD-R*	36		000	000	000	000	—	004	000	000	000	—	—	—
53 *PGM1-1*	1	2	776	722	790	595	—	700	757	734	687	—	—	—
53 *PGM1-R*	86		001	000	000	000	—	000	000	000	001	—	—	—
56 *AK-1*	—	—	958	967	964	984	—	956	962	973	958	—	—	—
63 *ADA-1*	0	0	942	931	927	875	—	824	—	972	904	—	—	—
65 *T*	0	0	448	623	431	532	425	—	480	338	553	—	—	—
100 *A1*	1	3	158	129	131	073	152	105	224	102	090	—	—	—
100 *A2*	14		309	262	279	308	278	258	283	277	238	—	—	—
100 *A3*	0		154	118	126	263	140	083	156	096	094	—	—	—
100 *A9*	2		104	139	114	160	116	145	056	115	166	—	—	—
100 *A10*	0		048	061	099	040	089	069	030	089	080	—	—	—
100 *A11*	16		054	062	064	049	070	030	051	065	061	—	—	—
100 *A28*	29		041	031	031	054	036	043	033	012	044	—	—	—
101 *B5*	4	4	059	100	086	046	070	181	022	108	143	—	—	—
101 *B8*	3		116	065	076	066	092	038	162	077	036	—	—	—
101 *B12*	6		134	114	105	063	106	067	193	216	086	—	—	—
101 *B13*	6		022	033	047	024	050	038	012	000	029	—	—	—
101 *B14*	48		027	039	020	000	038	015	034	017	019	—	—	—
101 *B27*	0		048	027	055	094	053	027	068	052	020	—	—	—
102 *BW15*	1		077	038	051	150	046	012	068	047	024	—	—	—
102 *BW17*	24		037	045	036	027	042	022	037	077	044	—	—	—
102 *BW22*	0		018	022	018	034	024	048	013	006	046	—	—	—
102 *BW35*	3		077	139	103	098	101	200	052	052	125	—	—	—
200 *Gm1*	2	5	191	211	171	254	144	163	196	252	—	—	—	—
200 *Gm1-2*	16		113	070	054	116	052	017	089	118	—	—	—	—
201 *Km1*	1	1	074	096	060	034	074	106	—	—	074	—	—	—
901 Cranlen	36	4	181	177	178	179	—	182	178	180	179	183	—	180
902 Cranbred	58		147	146	143	144	—	148	145	143	145	143	—	139
903 Forehead	27		98.9	98.1	97.3	96.5	—	99.3	95.5	96.4	96.8	98.0	—	96.6
904 Basionbr	6		131	131	134	132	—	138	127	129	132	133	—	136
905 Zygomat	20		133	132	132	134	—	139	128	129	135	133	—	132
906 Facehgt	0		71.2	68.9	68.9	69.9	—	73.4	67.5	68.8	72.2	68.9	—	68.1
907 Orbbred	95		40.6	40.6	40.9	41.8	—	43.2	42.0	41.4	42.1	41.6	—	41.2
908 Orbhgt	10		33.7	33.3	32.7	33.2	—	33.9	34.0	34.0	33.3	32.5	—	33.3
909 Naslbred	82		24.3	24.5	24.5	24.7	—	26.1	24.0	24.0	25.2	24.6	—	24.4
910 Naslhgt	2		51.2	51.9	50.7	51.3	—	54.2	54.5	53.2	52.4	51.5	—	51.3

Notes: The allele frequencies are identified by system number and the allelic symbol. They can be looked up in detail in Table 1. Names of the cranial variables are abbreviated here but are written out in full in Table 1. Column 1: Results of significance test per allele frequency or cranial variable expressed as probabilities × 100. Column 2: Results of Bonferroni tests per system are shown always in the first line for each system and are expressed as probabilities × 100. The four allele frequencies marked with dashes were not tested because their analyses of variance yielded nonsignificant results. Language families: B, Baltic; C, Ugric (Hungarian); E, Semitic (Maltese); F, Finnic (North Finnic speakers); G, Germanic; H, Greek; I, Romance; K, Celtic; Q, Basque; S, Slavic; T, Turkic; Y, Albanian. All allele frequency values have been multiplied by 1,000. The cranial variables are in millimeters.

$\times 5 = 0.02$. The results of these computations are shown in column 2 of Table 2. Twelve out of 21 (57.1%) of the probabilities are ≤ 0.05, indicating considerably greater homogeneity within observed language families than in randomly generated ones. This is far in excess of expected type I error. Thus there is little doubt that there is significant genetic homogeneity within the language families of Europe. We also employed Fisher's method of combining probabilities (Sokal and Rohlf, 1981) on the values reported in columns 1 and 2 of Table 2. The four systems that had been omitted from the study following their nonsignificance by analysis of variance were allowed for in the computations so as not to bias our results in favor of significance. The results yield highly significant values ($P \leq 0.00005$) for both columns.

The distribution of the significant departures from random expectations in the direction of increased homogeneity within single language families (data not shown) appears nonrandom. The Germanic language family has five systems significantly more homogeneous than by chance expectations, whereas only three other families (Romance, Slavic, and Ugric) each have two systems significant. No other language family has any significant system. Germanic is also the only language family to yield a significant P value (≤ 0.0005) when Fisher's test for combining probabilities over all systems is applied. To test the nonrandomness of the distribution of significant values across the language families, we performed a goodness of fit test using the Pearson statistic X^2 (Sokal and Rohlf, 1981, p. 702) for comparing the observed number of significant values in each language family to its expected number. That expected number was calculated on the basis of the total number of systems available for computation in each language family. The value of Pearson's X^2 was 36.34. However, we did not test its significance by comparing this quantity to chi squared, since the relevant assumptions are not met. Rather, we performed a randomization test in which the total number of significant values was randomly assigned 1,000 times to the available systems and language families in Table 2. The probability obtained was a highly significant 0.0070. In summary, then, our results suggest that, for numerous genetic systems, population samples overall differ more among language families than they do within families and that, conservatively viewed, only Germanic among the language families of Europe exhibits significant homogeneity for some systems by comparison with randomly placed areas.

Since the overall results suggest that at least some of the language families differ among themselves with respect to their gene frequencies and cranial variables, it was of interest to examine which of the mean values for the variables in Table 2 are significantly different from each other. Two problems accompany a systematic statistical test designed to answer this question. The first is the well-known problem of multiple comparisons testing (Sokal and Rohlf, 1981), which requires an experimentwise approach to evaluating the probabilities of individual tests. The second problem relates to the spatial autocorrelation of the individual samples for each variable, which engendered the unconventional approaches employed in this paper in the first place. Keeping these considerations in mind, we settled upon the following approach.

For each allele frequency and cranial variable, we carried out all possible pairwise conventional t tests for the difference of means between all pairs of language families for which means are available. It should be clearly understood that the t values obtained as a result of these difference-of-means tests are simply statistics expressing the degree of difference between the sample means, are not to be construed as sample statistics from a Student's t distribution, and thus cannot be tested for significance in the conventional manner. For each allele frequency, there are at most 66 comparisons between pairs of the 12 language families. To allow for the differing numbers of observed language family pairs for any one variable, we expressed the absolute t values as percentiles of their distribution for that variable. After doing this for all allele frequencies and cranial variables and summing the percentiles for each pairwise comparison between two language families, we computed deviations of these percentile sums from their expectations and standardized them. The expectations were computed under the hypothesis that any permutation of the observed percentiles within a variable was equally likely. Assuming these standardized deviations to be normally distributed, and using conservative Dunn-Sidak 5% experimentwise error probabilities (Sokal and Rohlf, 1981), we found

only one of the comparisons to be significant at $P \leq 0.05$. It is the difference between Germanic and Romance.

DISCUSSION

By the methodology employed here, the incidence of significant heterogeneity in gene frequency among (or homogeneity within) language families has been reduced considerably over that indicated by nominal significance of an ordinary (uncorrected) analysis of variance. The latter indicated that 21 systems had one or more significantly heterogeneous allele frequencies. Our methods can confirm these conclusions for only 12 systems. Of the remaining nine systems all but one were found to be spatially autocorrelated by Sokal et al. (1989), which confirms expectations that nominally heterogeneous language families might prove not to be heterogeneous after the spatial autocorrelation is accounted for. The nonautocorrelated exception is PGD. However, this system has data for only four language families and may exhibit an atypical response because of this. Among the four systems (MN, Lu, ABH, and Ak) that lacked significance even by a conventional analysis of variance and therefore had been excluded from further tests, two lacked spatial autocorrelation (Lu and Ak). Thus spatial autocorrelation does not necessarily ensure significant differences among language families.

Implications for population structure

We find it useful to think of our results in terms of two idealized models of gene-frequency differentiation. We call these two models the geographic and the linguistic models. The geographic model assumes 1) that the variance of values in an area of specified size on the map of Europe will be the same regardless of where the area is placed on the gene-frequency surface and 2) that the variance of an area increases with its size, because gene-frequency differentiation is a function only of geographic differentiation. 3) By implication, language families as such do not differ in mean gene frequencies. A similar model, but not in the context of genetic population structure, was investigated by Whittle (1956). By contrast, the linguistic model assumes that 1) the observations from within one language family are a random sample from the same distribution, and their geographic position within the language family provides no information whatsoever. 2) The variance within one language-family area is independent of its size. 3) However, language families can differ in their mean gene frequencies.

What are the possible population-biological processes that gave rise to these models? A process by which a continent has at one time been randomly settled by a homogeneous population, with gene frequencies as well as languages differentiating subsequently and independently because of the limited mobility of the populations, will lead to a pure geographic model. In such a case no significant allele frequency differences among language families are expected. By contrast, a situation wherein genetically distinct speakers of different language families arrived en masse in their current locations, followed by the absence of gene flow between language families but with no barriers to gene flow within a language family, would give rise to the linguistic model. This in turn would lead to significant differences among language families. Clearly, neither model can fully correspond to the situation in these European gene frequencies. These models are similar to, but not identical with, models correspondingly named by Watson et al. (1961).

The results of our analyses of the actual data offer contradictory evidence concerning an overall model. Of the 21 systems tested, 12 show significant differences among language families, and, of the remaining nine nonsignificant systems, all but system MNS show relatively low probabilities of the null hypothesis (no differences) being true (see Table 2). This is reflected in the very low probability resulting from Fisher's test of combining probabilities over all systems. The evidence is clear that for the majority of systems there are genetic differences among language families. Yet in view of the nonsignificant results obtained for 13 systems (the nine mentioned above plus the four eliminated from testing because of their nonsignificant ANOVAs), we cannot dismiss the geographic model out of hand. Can these contradictory findings be reconciled?

There are three possible explanations. First, the two models are only idealized extreme situations. In reality the situation may well be intermediate in that relatively genetically homogeneous populations of speakers of a language family entered Europe, settled, and then expanded and differ-

entiated geographically. Second, it is quite possible that some of the European populations differentiated geographically, hence are closer to the geographic model, whereas others, perhaps later arrivals, exhibit more genetic homogeneity and hence are closer to the linguistic model. Third, the incorporation of smaller, genetically differing groups by major language-family groups could create the appearance of geographic differentiation within language families.

Some direct evidence concerning these points can be obtained. If the geographic model holds, there should be a correlation between variance and language area. Separately for each allele frequency or cranial variable, we computed Spearman's rank correlation between the variance within each language family and its area as well as the great circle distance between its most distant sample localities. The latter is a measure of linear rather than two-dimensional differentiation. In the correlation with area, the resulting correlation coefficients range from −0.43 to 0.93. Of the 69 coefficients, 16, all positive, are statistically significant. The correlations with great-circle distance range from −0.11 to 0.98, and 31 of the 69 coefficients, again all positive, are statistically significant. Thus there is evidence for differentiation along the lines of the geographic model in these populations, but at the same time the significant differences we found among language families by randomization (Table 2) suggest a component ascribable to the linguistic model as well.

The effects of language family can be tested independently of the effects of geographic differences. Sokal (1988), using a partial correlation approach applied to genetic, linguistic, and geographic distances for these data, found significant partial correlations between language and genetics, geography kept constant, for 11 of 27 systems. It would appear, therefore, that a mixture of the two proposed models obtains in the actual relation between language family and genetic diversities in Europe.

Single language families

A brief consideration of the results for the single language families is called for. We carried out a series of simulation experiments to test random permutation without spatial constraints, the geographic, and the linguistic model described above. Under the random permutation and geographic models, no more language families than expected from type I error were significantly more homogeneous than pseudofamilies. Thus, individual language families are not significant under the geographic model, which is as expected, since the randomized pseudoareas (of approximately the same size and shape) should have the same variance as the actual language-family areas. However, this expectation would be true only if density of localities in the various language families were proportional to area. Since this is not true in our case, it may be no accident that the only language family with a suspiciously high rejection rate of the null hypothesis is Ugric. For the system analyzed in the simulation, sampling localities of the Hungarian samples are quite dense with respect to the area from which they are sampled. When Ugric pseudoareas are generated, these generally will take up a larger area, making them more variable and tending to make the observations significant for no reason other than the uneven density.

The linguistic model in our simulations was implemented by employing the sampling localities but not the actual gene frequencies for the 175 haptoglobin samples in our data. We designated two of the nine available language families to differ in their means from the others. These two groups were assigned values from normal populations of mean 1.0 and mean 3.0, respectively, whereas the rest of the populations were given values from normal populations of mean 0. In all cases the variance remained at 1. Two linguistic models were set up. In one, the different means were assigned to the two language families with the highest sample sizes (Germanic and Romance), and in the second they were assigned to two intermediate-sized language families (Finnic and Greek). Fifty separate realizations of the data sets were carried out.

We obtained an excess of significant results for five language families (Germanic, Romance, Slavic, Finnic, and Ugric) in the first case with Germanic and Romance differentiated and a similar excess only for Germanic and Romance in the second case with Finnic and Greek differentiated. This outcome can again be explained by the uneven densities of the language families. In the first linguistic model there are effectively only three groups of values: those for the 57 Germanic localities assigned 1.0, those for the 56 Romance localities assigned

3.0, and those for the 59 connected localities which represent most of the rest of the area, assigned 0. It is quite improbable that Germanic or Romance pseudoareas, randomly placed, will fit exactly into the remaining area. Therefore, they will most often cross the boundaries between two of the three groups with different means and hence manifest variances higher than in the observed area. Thus Germanic, Romance, and, in fact, some of the other denser language families are significant, but this is not true of the sparser language families, which easily fit into a numerically homogeneous area. By contrast, in the second language model the three different values are assigned to Finnic (with 15 localities), to Greek (with 11 localities), and to the connected remaining 149 localities. It is quite possible for Finnic or Greek to be randomized into these remaining localities and exhibit no greater variance than in its actual observed area. Hence, again only the largest areas have difficulty fitting and in consequence will straddle two language areas that differ in their mean. However, it must be noted that without gene frequency differences among at least some of the language families, i.e., unless the linguistic model holds, uneven densities cannot engender significance for individual families.

The same consideration applies to the actually observed data in which Germanic is significant for five systems and Romance, Slavic, and Ugric are each significant for two systems. These findings confirm that there are differences among the means of the language families, since, as we have seen, without at least partial validity of the linguistic model such significance could not be achieved. The model is now somewhat more complicated than in the simulations, since in the latter the variances within each of the language families were the same, which is unlikely to be true for the real gene frequencies analyzed. However, these findings say nothing about the relative homogeneity of an individual language family analyzed in comparison to that of others.

It is pertinent to examine briefly the means for each language family for the separate allele frequencies and cranial variables recorded in Table 2. To be conservative, we consider only those variables that show significant differentiation by a value of $P \leq 0.05$ in column 1 and also belong to a system similarly significant in column 2. There are 21 such variables. Inspection of the means for each variable reveals that the outlying means most frequently are for language families located at the periphery of the continent. Greek, Basque, Finnic, and Celtic occur 38 times as outliers; the three language families occupying the largest regions on the continent, Germanic, Romance, and Slavic, occur only 10 times altogether. Is the peripheral location of these genetically differing language families an accident of history, or is it their location that determines their genetic differences? We have not reached a firm conclusion on this issue, but there are some intriguing considerations that invite further study. Of the 21 variables considered, 10 show clear clinal patterns by consensus based on a variety of conservative criteria applied by Sokal et al., (1989). An example is the east-west cline for allele frequency I^B (system 1.1), which shows a low for Basque and highs for Baltic, Slavic, and Ugric speakers. Similarly, a northwest-southeast cline for HLA allele frequency *B8* is lowest for Greek speakers and highest for Celtic speakers. Thus it would appear that we should add a third model to the two considered above: clinal patterns, caused by either gene flow or adaptation, resulting in the differentiation of language families located at the extremes of the cline.

The ethnohistory of the four most frequently extreme language families differs considerably. The Basque speakers form a single ethnic unit whose origins are generally believed to predate the arrival of the Indo-European speakers surrounding them. They differ sharply from surrounding populations (Sokal et al., 1988) in contrast to the claims of Thoma (1985, p. 216). The Finnic speakers include Finns, Estonians, Karelians, and other populations located in the Soviet Union but also include the Lapps who, because of their inherent interest to investigators, are disproportionately highly represented in our data base. The Lapps are known to have gene frequencies that differ substantially from those of other European populations and hence tend to differentiate Finnic speakers from others. The Greek and Celtic speakers both find themselves located on various peripheral areas in Europe (largely on peninsulas) and represent anciently arrived Indo-European stocks that, however, must have undergone considerable admixture, given their histories. The northwesterly location of the Celtic speakers and

the southeasterly location of the Greek speakers readily differentiate these populations with respect to any northwest-southeast clines, in contrast with possibly aboriginal differences for the Basque and the Finnic speakers (largely the Lapps who are considered biologically quite diverse from other North Finnic speakers, but who acquired a North Finnic language prior to settling in the area they occupy at present). These views are advanced only tentatively at this time. Further evidence is needed to substantiate this interpretation.

The single pairwise comparison shown significant by the t statistic and others with relatively high t values mostly do not involve the extreme means belonging to the peripherally located populations just noted. This may seem surprising. However, the t method gives weight to populations that differ consistently over all allele frequencies, whereas the immediately preceding discussion focused on extremes for some individual allele frequencies. Just as a student doing well, but not superbly so, in all subjects may rise to the top of his or her class, consistent although not necessarily extreme differences emerge by employment of the t statistic.

A substantial number of studies have concerned themselves with the relation between language and gene frequencies (Dow and Cheverud, 1985; Dow et al., 1987; Jorde, 1980; Parsons and White, 1973; Rösing, 1984–85; Smouse and Wood, 1987; Sokal et al., 1986, 1987; Sokal and Winkler, 1987; Vecchi and Passarello, 1977–79). These have been carried out on varying spatial scales. However, workers have mostly contented themselves with establishing the relation between genetics and language rather than specifically testing for differences between populations classified by language. In fact, in an extensive review of the language-genetics literature, we have been unable to discover any such tests, whether carried out improperly by conventional statistical techniques or correctly by tests allowing for spatial autocorrelation. The problem addressed here is, however, a general one, whether the classification of gene frequency samples is by language, by culture, or by time. For this reason we feel that the methods presented here should be of general interest to workers in a variety of fields beyond human biology. Such applications are especially numerous in ecological research (see Legendre et al., 1989).

ACKNOWLEDGMENTS

This is contribution No. 694 in Ecology and Evolution from the State University of New York at Stony Brook. We thank Barbara Thomson and Donna DiGiovanni for technical assistance, Joyce Schirmer for art work, and Cheryl Daly for word processing. Supercomputer time for the randomization procedures was furnished by the Cornell Theory Center. The methodological portion of this research was supported by grant BSR 8614384 from the National Science Foundation and the application to gene frequencies and language areas in Europe by grant GM28262-07 from the National Institutes of Health, both to R.R.S.

LITERATURE CITED

Brassel KE, and Reif D (1979) A procedure to generate Thiessen polygons. Geogr. Anal. *11*:289–303.

Cavalli-Sforza LL, and Feldman MW (1981) Cultural Transmission and Evolution: A Quantitative Approach. Princeton, NJ: Princeton University Press.

Cliff AD, and Ord JK (1981) Spatial Processes. London: Pion.

Cowgill W (1976) Indo-European Languages. The New Encyclopaedia Brittanica. Chicago: Macropaedia, Vol. 9, pp 431–438.

Dow MM, and Cheverud JM (1985) Comparison of distance matrices in studies of population structure and genetic microdifferentiation: Quadratic assignment. Am. J. Phys. Anthropol. *68*:367–373.

Dow MM, Cheverud JM, and Friedlaender JS (1987) Partial correlation of distance matrices in studies of population structure. Am. J. Phys. Anthropol. *72*:343–353.

Endler JA (1977) Geographic Variation, Speciation, and Clines. Princeton, NJ: Princeton University Press.

Endler JA (1986) Natural Selection in the Wild. Princeton, NJ: Princeton University Press.

Flint J, Hill AVS, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, Bana-Koiri J, Bhatsu K, Alpers MP, Boyce AJ, Weatherall DJ, and Clegg JB (1986) High frequencies of alpha thallassemia are the result of natural selection by malaria. Nature *321*:744–750.

Harding RM, Sokal RR, and Oden NL (1987) An examination of spatial patterns of gene frequencies in Europe. Am. J. Phys. Anthropol. *72*:207–208.

Harms RT (1976) Uralic Languages. The New Encyclopaedia Brittanica. Chicago: Macropaedia, Vol. 18, pp. 1022–1032.

Hope ACA (1968) A simplified Monte Carlo significance test procedure. J. R. Stat. Soc. Ser. B *30*:582–598.

Ivanov VV (1976) Slavic Languages. The New Encyclopaedia Britannica. Chicago: Macropaedia, Vol. 16, pp. 866–874.

Jorde LB (1980) The genetic structure of subdivided human populations. A review. In JH Mielke and MH Crawford (eds.): Current Developments in Anthropological Genetics. New York: Plenum Press, Vol. 1, pp. 135–208.

Legendre P, Oden NL, Sokal RR, Vaudor A, and Kim J (1989) Approximate analysis of variance of spatially autocorrelated regional data. J. Classif. (In press).

Malecot G (1973) Isolation by distance. In NE Morton (ed.): Genetic Structure of Populations. Honolulu: University of Hawaii Press, pp. 72–75.

Martin R, and Saller K (1957): Lehrbuch der Anthropologie. Stuttgart: G Fischer-Verlag, 4 Vol.

Mather JY, Speitel HH, and Leslie GW (eds.) (1975) The Linguistic Atlas of Scotland. Hamden, CT: Archon Books.

Meillet A, and Cohen M (eds.) (1952a) Langues Indo-Européennes d'Europe (Carte). Les Langues du Monde. Paris: Centre National de la Recherche Scientifique.

Meillet A, and Cohen M (eds). (1952b) Langues Ouraliennes (Carte). Les Langues du Monde. Paris: Centre National de la Recherche Scientifique.

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

Morton NE (1973) Isolation by distance. In NE Morton (ed.): Genetic Structure of Populations. Honolulu: University of Hawaii Press, pp. 76–79.

Morton NE (1982) Estimation of demographic parameters from isolation by distance. Hum. Hered. *32:*37–41.

Moulton WG, Haugen E, and Herzog MI (1976) Germanic Languages. The New Encyclopaedia Britannica. Chicago: Macropaedia, Vol. 8, pp. 19–31.

Mourant AE, Kopéc AC, and Domaniewska-Sobczak K (1976) The Distributions of the Human Blood Groups. London: Oxford University Press.

Nei M (1987) Molecular Evolutionary Genetics. New York: Columbia University Press.

Parsons PA, and White NG (1973) Genetic differentiation among Australian aborigines with special reference to dermatoglyphics and other anthropological traits. In RL Kirk (ed.): The Human Biology of Aborigines in Cape York. Australian Aboriginal Studies No. 44. Canberra: Australian Institute of Aboriginal Studies, pp. 81–94.

Posner R (1976) Romance Languages. The New Encyclopaedia Britannica. Chicago: Macropaedia, Vol. 15, pp. 1025–1045.

Rohlf FJ, and Schnell GD (1971) An investigation of the isolation by distance model. Am. Nat. *105:*295–324.

Rösing FW (1984–85) Relations between morphometric distances and linguistic classification in the populations of sub-Saharan Africa. Riv. Antropol. *63:*259–262.

Ruhlen M (1987) A Guide to the World's Languages. Vol. 1: Classification. Stanford: Stanford University Press.

Schwidetzky I, and Rösing FW (1984) Vergleichend-statistische Untersuchungen zur Anthropologie der Neuzeit (nach 1500). Homo *35:*1–49.

Smouse PE, and Wood JW (1987) The genetic demography of the Gainj of Papua New Guinea: Functional models of migration and their genetic implications. In BD Chepko-Sade and Z Halpin, (eds.): Mammalian Dispersal Patterns: The Effects of Social Structure on Population Genetics. Chicago: University of Chicago Press, pp 211–224.

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

Sokal RR, Harding RM, and Oden NL (1989) Spatial patterns of human gene frequencies in Europe. Am. J. Phys. Anthropol. (In press).

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

Sokal RR, and Rohlf FJ (1981) Biometry, 2nd ed. San Francisco: W.H. Freeman.

Sokal RR, and Rohlf FJ (1987) Introduction to Biostatistics, 2nd. ed. New York: W.H. Freeman.

Sokal RR, Smouse PE, and Neel JV (1986) The genetic structure of a tribal population, the Yanomama Indians. XV. Patterns inferred by autocorrelation analysis. Genetics *114:*259–281.

Sokal RR, Uytterschaut H, Rösing FW, and Schwidetzky I (1987) A classification of European skulls from three time periods. Am. J. Phys. Anthropol. *74:*1–20.

Sokal RR, and Winkler E-M (1987) Spatial variation among Kenyan tribes and subtribes. Hum. Biol. *59:*147–164.

Steinberg AG, and Cook CD (1981) The Distribution of the Human Immunoglobulin Allotypes. London: Oxford University Press.

Thoma A (1985) Eléments de Paléoanthropologie. Louvain-la-Neuve, Belgium: Institut Supérieur d'Archéologie et d'Histoire de l'Art.

Tills D, Kopéc AC, and Tills RE (1983) The Distribution of the Human Blood Groups and Other Polymorphisms. Supplement 1. London: Oxford University Press.

Vecchi F, and Passarello P (1977–79) Relazioni tra distanza morfologica e caratteristiche linguistiche nelle popolazioni dell'Africa Sub-sahariana. Riv. Anthropol. *60:*187–194.

von Czoernig C (1984) Ethnographische Karte der oesterreichischen Monarchie. In H Kühnel, E Vavra, and G Stanger (eds.): Von der Revolution zur Gründerzeit. Katalog des niederösterreichischen Landesmuseums, N.F. Nr. 147. Wien: Amt der niederösterreichischen Landesregierung, p. 141ff.

Watson JB, Zigas V, Kooptzoff O, and Walsh RJ (1961) The blood groups of natives in Kainantu, New Guinea. Hum. Biol. *33:*25–41.

Whittle P (1956) On the variation of yield variance with plot size. Biometrika *43:*337–343.

Wright S (1969) Evolution and the Genetics of Populations, Vol. 2: The Theory of Gene Frequencies. Chicago: University of Chicago Press.