Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 5836–5840, June 1995
Evolution

Genetics and the origin of European languages

ALBERTO PIAZZA*, SABINA RENDINE*, ERIC MINCH†, PAOLO MENOZZI‡, JOANNA MOUNTAIN§, AND LUIGI L. CAVALLI-SFORZA†

*Dipartimento di Genetica, Biologia e Chimica Medica e Centro Consiglio Nazionale delle Ricerche di Immunogenetica ed Oncologia Sperimentale, Università di Torino, 10126 Torino, Italy; ‡Istituto di Ecologia, Università di Parma, 43100 Parma, Italy; †Department of Genetics, Stanford University, Stanford, CA, 94305; and §Department of Integrative Biology, University of California at Berkeley, Berkeley, CA 94720

Contributed by Luigi L. Cavalli-Sforza, November 10, 1993

ABSTRACT **A new set of European genetic data has been analyzed to dissect independent patterns of geographic variation. The most important cause of European genetic variation has been confirmed to correspond to the migration of Neolithic farmers from the area of origin of agriculture in the Middle East. The next most important component of genetic variation is apparently associated with a north–south gradient possibly due to adaptation to cold climates but also to the differentiation of the Uralic and the Indo-European language-speaking people; however, the relevant correlations are not significantly different from zero after elimination of the spatial autocorrelation. The third component is highly correlated with the infiltration of the Yamna ("Kurgan") people, nomadic pastoralists who domesticated the horse and who have been claimed to have spread Indo-European languages to Europe; this association, which is statistically significant even when taking spatial autocorrelations into account, does not completely exclude the hypothesis of Indo-European as the language of Neolithic farmers. It is possible that both expansions were responsible for the spread of different subfamilies of Indo-European languages, but our genetic data cannot resolve their relative importance.**

Human geographic expansions during prehistoric and historical times played a major role in shaping the genetic geography of human populations (1). Expansions in general are caused by cultural innovations that change the economy of a whole geographical region and, therefore, its demographic equilibrium. It seems reasonable to assume that: (*i*) by the end of the Paleolithic, genetic drift due to very low population densities had produced major genetic differences among the human populations that already inhabited all parts of the world; and (*ii*) expansions that took place during that period left a genetic footprint not completely erased by later population movements (2, 3). No single gene alone can trace such processes, but combining the information from many genes by the statistical technique of "principal component analysis" (PCA) can reveal geographic patterns of genetic variation that may indicate past expansions.

Cultivated cereals and domesticated animals which spread to Europe are found first in archaeological sites of the Middle East. The spread from the center of origin, at an average rate of 1 km per year (4), is quite regular in time. The radiation beginning 10,000 years B.P. could have been cultural (the technology diffused) or demic (the farmers moved) or both (4, 5). The ^{14}C dates have been shown to be compatible with a demic spread (5–7). Extensive simulations (7, 8) have shown that traces of this migration can be detected by applying the PCA technique to contemporary population genetic data. The remarkable similarity between the archaeological map and the "synthetic" map of first principal component (PC1) values from 39 gene frequencies was given as evidence that the diffusion of agriculture was a spread of farmers rather than of the innovative technology alone (2). Sokal and collaborators (9–12) have confirmed this result. They used spatial autocorrelation analysis (9) and tested the statistical significance of the partial correlation between genetic distances and distances especially designed to represent the spread of agriculture to Europe when geographic distances are held constant (11). The problem of verifying a geographical genetic pattern by a statistical test of significance (12) challenged us to refine our methodology and also to look for possible genetic traces left in Europe by early Indo-European (IE) speakers as opposed to other language speakers.

Origin of Early IE Speakers

According to Renfrew (6), Neolithic migrants from Anatolia (Turkey), who established the first European farming communities in Greece at around 6500 B.C., spoke IE languages. From here, further population growth and expansion (2) spread their economy and language to the rest of Europe. The origin of IE languages has been the source of much discussion. While some linguists agree that proto-IE (PIE) may have originated either in Anatolia (13) or in Transcaucasia (14), more linguists accept the "Kurgan" (meaning barrow in Russian) theory by Gimbutas (15), which views the original speakers of PIE as moving sometime between 4300 and 2800 B.C. (calibrated years) from the southern steppes of Ukraine (between the Black and the Caspian Seas where the Kurgan culture has been first documented) and spreading to the extreme west and north of Europe. In comparing the reconstructed cultural vocabulary of PIE with the archaeological and environmental record, Mallory (16) reviews a series of inconsistencies with an Anatolian and Greek origin going back to 9000–10,000 years B.P. Some further criticisms based on linguistic evidence have been detailed (17).

Recent findings have added important elements to this picture. On their basis, Anthony (18) identifies the PIE homeland in a region of the order of 500,000 km^2 in eastern Europe north of the Black and Caspian Seas and gives 3300 B.C. (calibrated) as the date after which dispersal and language differentiation began.

The linguistic argument for a PIE origin in the southern steppes is based mainly on the contact between PIE and Ugro-Finnic and PIE and Kartvelian. The archaeological argument is the clear presence of wheeled vehicles in the PIE homeland. The main Kurgan culture involved in this expansion (Yamna) spread later into the lower Danube and the Carpathian Basin. This, and the Corded Ware culture in parts of northern Europe, might have provided a medium through which IE languages diffused to the rest of Europe. Horseback riding and the important socioeconomic changes involved in a rapid long-distance way of moving might have also provided a mechanism of diffusion of the IE speakers out of Ukraine.

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

Abbreviations: IE, Indo-European; PIE, proto-IE; PC1, etc., first principal component, etc.

Other related Kurgan cultures are responsible for the spread of IE languages to Iran and India. The search for Indo-Iranian origins east of the Urals has in the Andronovo culture of Central Asia a good candidate.

A comprehensive genetic picture of Europe is given by the analysis of the European fraction of a new (and to our knowledge, the most complete) collection of available population gene frequency data (3). In that collection (detailed references and analysis are found in ref. 3), we considered the following genetic systems (in parentheses number of selected samples): ABO(2650), A1A2BO(337), ACP1(182), ADA(139), AK1(140), PI(85), AG(55), LPA(42), CHE1(54), CHE2(41), C3(58), FY(193), ESD(65), GPT(37), BF(35), GC(257), GLO1(54), HP(410), HLAA(132), HLAB(132), IGHG1(157), IGHG3(157), IGKM(108), KELL(315), JK(60), LE(73), LU(70), MNS(161, 577 only MN), P1(201), PTC(139), PGM1(207), PGD(94), RH(362, 1287 only D), SE(122), TF(118). To this data base, given the relevance of the region in the history of the European populations, a large body of genetic data from the Caucasus area has been added (19–21). It includes the following 15 polymorphic systems with 31 alleles: ABO(102 samples), ACP1(45), C3(42), FY(27), ESD(43), GC(42), GLO1(40), HP(50), KELL(34), LE(33), MNS(30, 85 only MN), P1(43), RH(34, 73 only D), SE(24), TF(46). The average sample size is 210 (for more details and analysis, see ref. 20).

An interpolation procedure has been applied to the above loci (with a total of 95 alleles) to build surfaces of gene frequencies that minimize distances from observed gene frequencies and cover a regular set of grid points in a geographical map of Europe. These grid points form geographical units common to all alleles. Thus, an ordinary principal component analysis can be applied to these geographical units, and the corresponding first, second, third, etc., component (PC1, PC2, PC3, etc.) scores can be plotted and contoured on a geographical map of Europe.

First Genetic Component

The map given by PC1 scores, shown in Fig. 1 *Left*, synthesizes 26% of the original genetic variation and, compared with the previous analysis with fewer data (2), agrees in more detail with the archaeological information. Because the genes for which data are available are not a random sample of our genome, and because not all genes were tested in all populations, it is important to test the robustness of this map. Therefore, standard errors were estimated for each map point by bootstrapping (22) of genes. In practice, a random sample of genes is taken, with replacement, from the data matrix (gene frequencies × grid points of the map), generating a new data matrix in which some of the genes appear only once, others two or more times, and about one-third have completely disappeared. The total number of genes after each resampling of the matrix is the same as the original number. We repeated this resampling procedure (bootstrapping) 100 times. Each resampling produces a new synthetic map (if necessary, inverted so as to be congruent with the original principal component map), and each of its points will oscillate in the 100 replicas around a mean with a given standard deviation. In Fig. 1 *Right* we show the map of the ratios of means to standard deviations; it indicates that the error in calculating the principal component scores due to sampling of genes is especially low in the relevant areas at the extremes of the gradient (e.g., in the Middle East). A similar approach has been described (23) in a different context. Our conclusion is that the gene frequency gradient associated with the spread of Neolithic farmers is robust with respect to resampling of genes.

Congruence between the gene frequency gradient and archaeological dates has been evaluated by using Pearson's correlation coefficient between the 93 archaeological dates of first arrival of Neolithic farmers (5) and PC1 interpolated at the same geographical locations. As gene frequencies and dates of the first arrival of Neolithic farmers are both spatially autocorrelated processes, the statistical significance of the correlation coefficient has to be properly modified if we wish to test the hypothesis that the two processes are mutually correlated while correcting for their spatial autocorrelation. The method proposed by Clifford *et al.* (24) based on the evaluation of an "effective" sample size that takes the spatial structure into account has been applied. We obtained $r = 0.86$ with an estimated effective sample size $M = 5.288$. To test the statistical significance of the correlation against the null hypothesis, a t statistic is calculated with M-2 degrees of freedom. In our case, $t = 3.039$ with 3.288 degrees of freedom is statistically significant ($P < 0.05$).

Second Genetic Component

As noted but not tested in the earlier analysis (2), the synthetic map of the PC2 values shows a north–south gradient, possibly correlated with climate. A direct test of the hypothesis [using

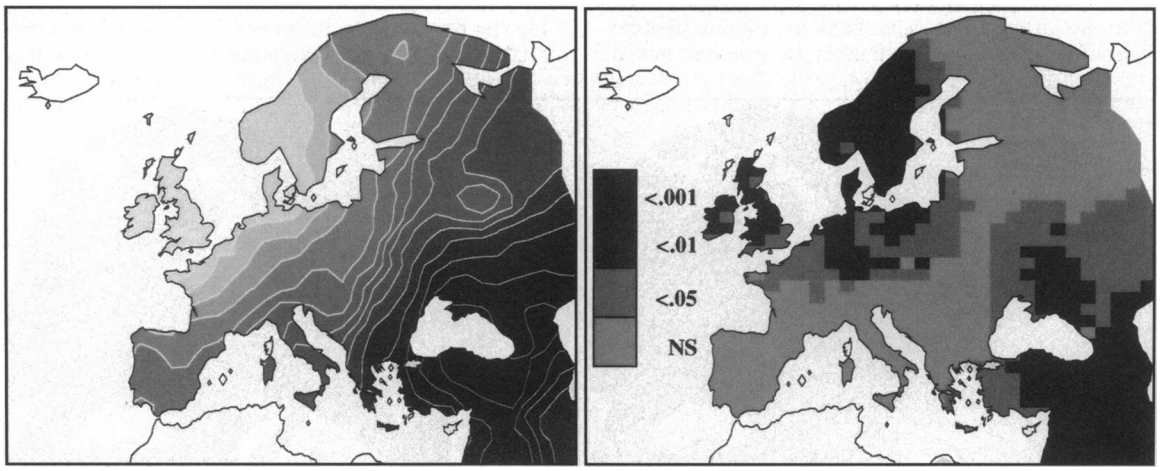

FIG. 1. (*Left*) Synthetic map of the PC1 values calculated from 95 gene frequencies in Europe. The map is based on the genetic systems listed in the text and conveys 26% of the total genetic variation. (*Right*) Error plot of the synthetic map. It tests the robustness of the synthetic map with respect to the sampling of genes and is obtained by resampling with replacement (bootstrap) of the 95 gene frequencies of the original data 100 times. Displayed is the map of the ratios (mean)/(standard deviation) in each grid square unit of the map. The four levels of grey are chosen to reflect different probability levels of these ratios. They correspond (from grey to black) to the probability intervals >0.05 = not significant (NS), 0.05–0.01, 0.01–0.001, and <0.001.

six climate variables collected from 297 meteorological stations (25) whose altitude is below 2000 feet] shows that, allowing for the effect of autocorrelation by the Clifford test (24), climatic variables are not significantly correlated with PC2.

In the present analysis, with a larger set of genetic markers and populations, the updated map, which conveys 20.6% of the total genetic variation in Europe, is shown in Fig. 2 *Left* and its bootstrap test of robustness in Fig. 2 *Right*. An interesting point is that the synthetic map of Fig. 2 *Left* can also be associated with a partition of Europe into two linguistic areas, (*i*) IE and (*ii*) Uralic, spoken almost exclusively in the north by Lapps in northern Scandinavia, Finns, Estonians, and several populations in northern Russia and in a southern isolated pocket inhabited by Hungarians who owe their Uralic language to the invasion by Magyars in the ninth and tenth centuries A.D. The upper extreme in the PC2 map is among Lapps, some of whom have Mongolian-like traits such as darker hair and skin. Genetic analysis (26) has shown that Lapps have up to 48% genetic admixture with Uralic people further east, while Finns have 10% and Hungarians 12%. To test the association of PC2 and Uralic languages, we calculated the mean of the second principal axis scores in the area occupied by Uralic-speaking people and that occupied by IE speakers (27). This would be a classical analysis-of-variance, were it not for the fact that the PC2 values are spatially autocorrelated in an unknown way, and therefore the critical assumption of independence of the error terms is not satisfied. One solution to the problem has been proposed by Legendre *et al.* (28), and we used their permutational method to test the statistical significance of the difference between the PC2 means in the two linguistic (Uralic, IE) areas. The method does not allow broken areas to be taken into account, so we did not include the small area occupied by Hungarians within the Uralic dominium. The results indicate a difference that, however, does not reach statistical significance.

Even though our test does not reach the significance level, it seems likely, in view of the historical evidence, that neither association (genes–languages, genes–climatic factors) is spurious. Further analysis is necessary to confirm this.

One plausible hypothesis is that people speaking Uralic languages spread westward along the Arctic coast from an unknown area of origin in northern Siberia. Note, by way of analogy, that other Arctic populations (e.g., Eskimos) have always remained at low density and spread mostly or only along the coast. Today Uralic-speaking Samoyeds, possibly the population ancestral to Lapps (Saame), live not far from the Arctic Ocean east of the Urals. The Uralic speakers, who we assume migrated west of the Urals, remained in Arctic areas but mixed largely with the presumably more numerous speakers of IE languages who migrated from northern Russia. While the original Uralic language survived, the original genes of the western Uralic speakers may have been highly diluted in the process. The Uralic-speaking Finnish population is a case in point. We have here a clear-cut example of a discrepancy between the language (Uralic) and the genes, which are much more similar to those of IE-speaking populations further South and show only small traces of genes similar to those of other Uralic-speaking populations (26). Hästbacka *et al.* (29) elaborate on earlier demographic and epidemiological observations (30, 31) adding new genetic evidence and note that the present Finnish population originated from a small number of individuals (they suggest 1000) who settled around 2000 years B.P. in a southwestern area of Finland. This region had already been occupied by other people for at least 3000 years. No further immigration apparently took place.

Third Genetic Component and the Spread of Pastoral Nomads from the Kurgan Region

The map of Fig. 3 *Left* shows the contour plot of the gene frequency PC3 scores, which convey 8.8% of the total genetic variation in Europe; Fig. 3 *Right* is the associated map of robustness. Fig. 4 shows the origin and the diffusion of the Kurgan culture developed by pastoral nomads of the Eurasian steppes starting around 4300 B.C. (calibrated years). Given the suggested connection between this culture and migrations of IE speakers (15), it is worth testing the similarity of the significant gradient displayed in Fig. 3 *Left* with the map in Fig. 4.

Statistical significance testing for the association between PC3 and the Kurgan expansion proposed by Gimbutas is not easy to perform, mainly because the dating of the spatial diffusion by the Kurgan culture is not as detailed as that collected for early Neolithic farmers in Europe. Gimbutas' work reviewed in ref. 15 identifies the area of origin of Kurgan people and the extension of their three waves of diffusion in Europe, but the dynamics of the process with its precise dating are confused by the infiltration of other cultures in the same geographical areas. Therefore, we are limited to testing how different geographic areas (that of the Kurgan people's claimed origin and the regions of Europe that received none, one, two, or three of the Kurgan waves described by Gimbutas; see Fig. 4) are correlated to the PC3 scores of Fig. 3 *Left*. We calculated the mean of the PC3 in the Kurgan area of origin ("ORIGIN" in Fig. 4), and we tested whether such a mean is statistically different from the PC3 mean over Europe excluding the Kurgan area of origin. This again would be a classical analysis of variance, were it not for the spatial autocorrelation

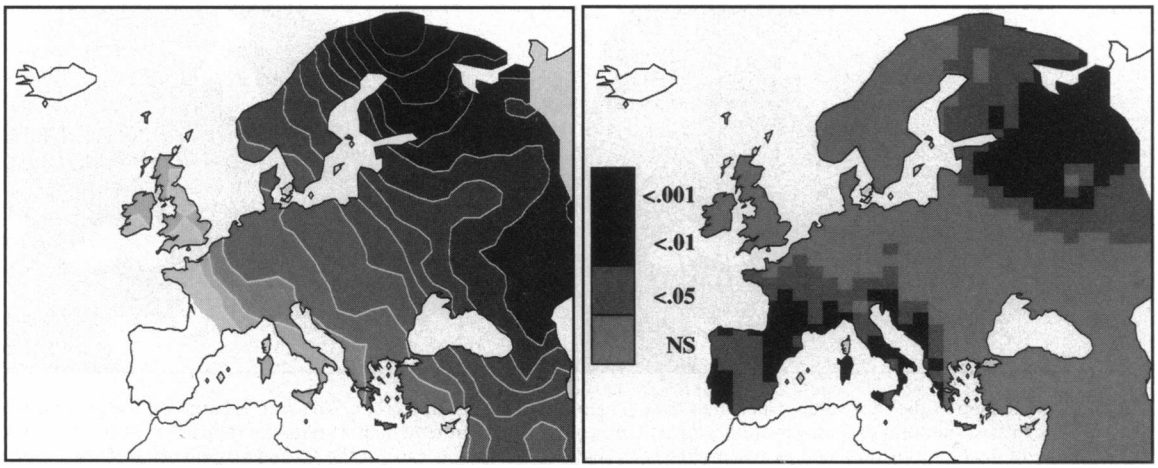

FIG. 2. Synthetic map of the PC2 values calculated from 95 gene frequencies in Europe. The map (*Left*) is based on the genetic systems listed in the text and conveys 20.6% of the total genetic variation. The plot of its error (*Right*) is built by using the same technique as for Fig. 1 *Right*.

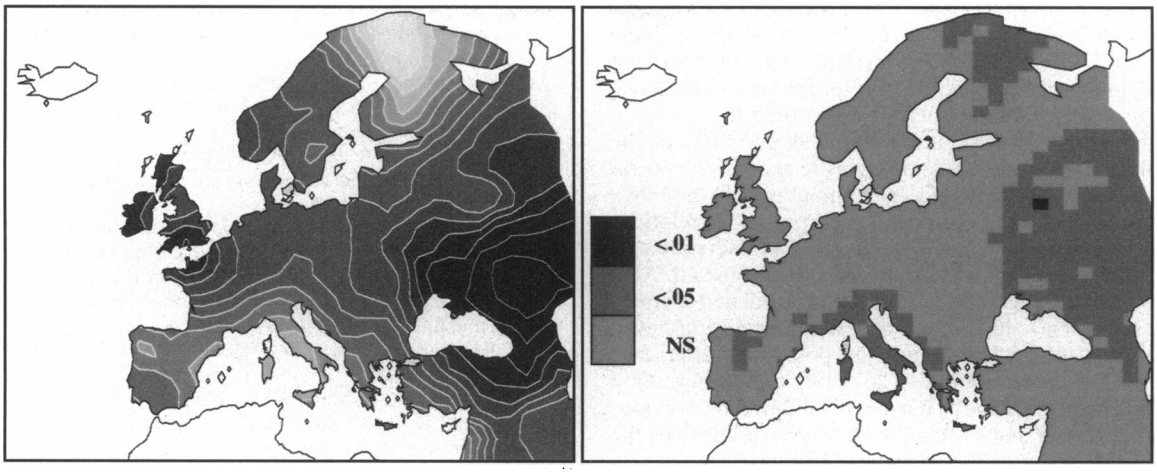

FIG. 3. Synthetic map of the PC3 values calculated from 95 gene frequencies in Europe. The map (*Left*) is based on the genetic systems listed in the text and conveys 8.8% of the total genetic variation. The plot of its error (*Right*) is built by using the same technique as for Fig. 1 *Right*.

as discussed above. In this case we again used the permutational method (28) for testing the statistical significance of the difference between the PC3 means in the two (Kurgan vs. non-Kurgan) areas. To carry out the test, 1000 geographical permutations were performed in which two regions of Europe with the same area and the same shape of the Kurgan and the non-Kurgan regions were chosen, and the difference between their mean scores on the third principal component was calculated. The result indicates a clearly significant difference: only 2 (or 23, depending on the permutation algorithm; see ref. 28) of the 1000 permutations show a difference greater than or equal to that observed between the Kurgan and the non-Kurgan regions. This excludes a random association between the area of origin of the Kurgan culture and one extreme (a possible center of genetic diffusion) in the synthetic map of the third principal component scores summarizing the information from 95 gene frequencies in Europe. A demic expansion from the steppes north of the Black and the Caspian Seas is therefore suggested.

Renfrew (6) asked what cultural advantage might have allowed the invaders and their descendants to establish their language over such a wide area. Recent archaeological evidence for horseback riding at the Sredny Stog site of Dereivka in Ukraine around 4000 B.C. (32) suggests that the spread of Yamna people might have found an initial boost not in the process of riding alone but rather in the addition of riding to preexisting agriculture and herding (33). The invention of the wheel [archaeological records of wagon transport in Kurgan graves in the steppes west of the Urals have been radiocarbon-dated from 3000 B.C. or earlier to around 2200 B.C.; archaeological evidence has been documented in the Sintashta–Petrovka complex in the steppes east of the Ural mountains within a radiocarbon–calibrated time interval (2σ) of 2137–1938 B.C. for horse draft and spoked wheel chariots (see ref. 34)] probably provided further economic and military advantages that accelerated the expansion of Yamna people into most of Europe.

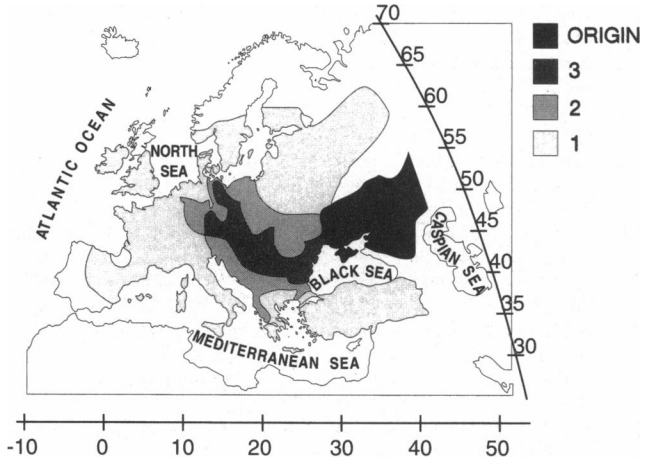

FIG. 4. Different shadings show the origin of Kurgan influence in Lower Volga–Don steppes (full black) and the distribution of archaeological sites of Kurgan waves numbers 1, 2, and 3 as described in ref. 15 and as drawn in figure 2B of reference 37. Shades numbered 1, 2, and 3 outline regions that received one, two, or three of the Kurgan waves. Thus, area 1 received wave 1 or wave 2 or wave 3; area 2 received waves 1 and 2 or 1 and 3 or 2 and 3; area 3 received all waves 1, 2, and 3.

Discussion

Most archaeologists since midcentury have reacted strongly to the earlier trend of considering local change of artifacts as signs of migratory movements of large groups of people. Thus, the hypothesis of the migration of farmers was not accepted by some (35). Renfrew (6, 36) has accepted, on the basis of theoretical considerations, our hypothesis that agriculture was spread from the Near East by people, the farmers themselves, rather than as a technology, and he used this conceptual framework to propose that Neolithic farmers spoke IE languages, which they spread to Europe. Simulations showed that later migrations do not easily erase the pattern generated by earlier major migrations such as the spread of farmers (8). An important factor in determining the degree to which genetic gradients that are correlated to major expansions can be observed is the ratio of population saturation density allowed by two economies: that of migrants, in this case the farmers, and that of earlier settlers, in this case hunter-gatherers. Our present study confirms earlier results (2) through the addition of data and a test of statistical robustness and gives further evidence that the Neolithic farmers' migration is the most important factor in determining the genetic geography of Europe.

It is well known that hypotheses on the geographic origin of a language are difficult to test. The problem is complicated by the possibility that IE language speakers might have had several expansions at different times and places. The assumption of a PIE speakers' homeland in the Kurgan region is really not incompatible with an earlier Anatolian origin. Both may be correct, the Kurgan culture being substantially later than, and as a consequence of, the spread of agriculture to the steppes. Both demic expansions left their genetic traces in Europe: unfortunately synthetic genetic maps are inherently undated.

The question whether the maps of both PC1 and PC3 or only the map of PC3 was relevant to the history of IE-speaking peoples cannot be settled with our currently available tools, but it is in principle likely that the order of importance of principal components tends to follow their order in time (8).

Recently Sokal *et al.* (37) attempted a direct study of the correlations between genetic and linguistic expansions postulated by Renfrew (6, 36) and by Gimbutas (15) and had negative results for both. They calculated partial correlations between frequencies of single genes and possible routes of expansions with geographic distances held constant. Our method is entirely different and relies on interpolated surfaces of gene frequencies; it reaches different conclusions even though we also took into account the possible confounding effect of spatial autocorrelation. The ability of our method to separate different expansions has been validated by previous simulations (8). The reason for the discrepancy between the two approaches requires further investigation.

It is interesting that the map of PC2 may also give some information on the spread of a language, in this case Uralic. As the finding at the moment is not statistically significant (nor is the effect of climatic factors), these correlations are only suggestions worth further testing. In any case, the spread of Uralic-speaking people and the association with climate may not generate separate genetic patterns if Uralic speakers lived long enough in the northern climate to show genetic adaptation to it.

Barbujani and Sokal (38) found a correlation between linguistic and genetic boundaries in Europe. In the majority of cases, 22 out of 33, there were also physical barriers that may be the cause of both genetic and linguistic boundaries. In 9 cases there were only linguistic and genetic boundaries but not physical ones: 3 of them (northern Finland vs. Sweden, Finland vs. Kola peninsula, and Hungary vs. Austria) separate Uralic from IE languages. It remains to be established in these cases (or in some of them) if linguistic boundaries have generated or enhanced genetic boundaries or if both are the consequence of political, cultural, and social boundaries (as in the case of Lapps and non-Lapps) that have played a role similar to that of physical barriers.

Our results suggest how important events of the demographic history and prehistory of Europe can be clarified by studying its genetics. This knowledge may contribute to archaeology, history, and linguistics, and the joint study from all these perspectives will be—we think—especially effective in a time when modern molecular techniques are bringing analysis of genetic variation to an unprecedented degree of resolution.

Prof. R. R. Sokal read a very preliminary version of this paper, made useful suggestions on testing the statistical significance of correlations, and kindly volunteered to do some of the calculations. We also used computer programs provided by Dr. Clifford and Dr. Legendre. We are grateful to all of them as well as to Prof. B. Efron for stimulating discussions. The generosity of Dr. I. S. Nasidze in giving us the gene frequency data he collected in the Caucasus before they were published is specially acknowledged. This work was supported by the National Institutes of Health (GMS20467), by the Ministero Università Ricerca Scientifica Tecnologica 60% (Italy), and by Consiglio Nazionale delle Richerche Target Projects "Genetic Engineering" and "Biotechnology and Bioinstrumentation." Computer time and graphics were granted by Consorzio Sistema Informativo-Piemonte (Torino, Italy).

1. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1993) *Science* **259,** 639–646.
2. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. L. (1978) *Science* **201,** 786–792.
3. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994) *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ).
4. Ammerman, A. J. & Cavalli-Sforza, L. L. (1973) in *The Explanation of Culture Change*, ed. Renfrew, C. (Duckworth, London), pp. 343–357.
5. Ammerman, A. J. & Cavalli-Sforza, L. L. (1984) *Neolithic Transition and the Genetics of Populations in Europe* (Princeton Univ. Press, Princeton, NJ).
6. Renfrew, C. (1987) *Archaeology and Language: The Puzzle of Indo-European Origins* (Cambridge Univ. Press, New York).
7. Sgaramella-Zonta, L. & Cavalli-Sforza, L. L. (1973) in *Genetic Structure of Populations*, ed. Morton, N. E. (Univ. Hawaii Press, Honolulu), pp. 128–135.
8. Rendine, S., Piazza, A. & Cavalli-Sforza, L. L. (1986) *Am. Nat.* **128,** 681–706.
9. Sokal, R. R. & Menozzi, P. (1982) *Am. Nat.* **119,** 1–17.
10. Sokal, R. R., Oden, N. L., Legendre, P., Fortin, J. J., Kim, J., Thomson, B. A., Vaudor, A., Harding, R. M. & Barbujani, G. (1990) *Am. Nat.* **135,** 157–175.
11. Sokal, R. R., Oden, N. L. & Wilson, C. (1991) *Nature (London)* **351,** 143–145.
12. Sokal, R. R., Oden, N. L. & Wilson, C. (1992) *Nature (London)* **355,** 214.
13. Dolgopolvsky, A. B. (1988) in *Mediterranean Language Review*, eds. Borg, A. & Wexler, P. (Harrassowitz, Wiesbaden, Germany) Vol. **3,** pp. 7–31.
14. Gamkrelidze, T. V. & Ivanov, V. V. (1990) *Sci. Am.* **262** (3), 82–89.
15. Gimbutas, M. (1991) *The Civilization of the Goddess: The World of Old Europe* (Harper, San Francisco), Chap. 10.
16. Mallory, J. P. (1989) *In Search of the Indo-Europeans: Language, Archaeology and Myth* (Thames & Hudson, London).
17. Jasanoff, J. H. (1988) *Language* **64,** 800–802.
18. Anthony, D. W. (1994) in *Die Indogermanen und das Pferd*, eds. Hänsel, B. & Zimmer, S. (Archaeolingua, Budapest), pp. 185–197.
19. Nasidze, I. S. (1992) *Gene Geogr.* **6,** 85–88.
20. Barbujani, G., Nasidze, I. S. & Whitehead, G. N. (1994) *Hum. Biol.* **66,** 639–668.
21. Barbujani, G., Whitehead, G. N., Bertorelle, G. & Nasidze, I. S. (1994) *Hum. Biol.* **66,** 843–863.
22. Efron, B. & Tibshirani, R. J. (1993) *An Introduction to the Bootstrap* (Chapman & Hall, New York).
23. Diaconis, P. & Efron, B. (1983) *Sci. Am.* **248** (5) 96–108.
24. Clifford, P., Richardson, S. & Hémon, D. (1989) *Biometrics* **45,** 123–134.
25. Great Britain Meteorological Office (1964) *Tables of Temperatures, Relative Humidities and Precipitation of the World* Report 617A (Air Ministry Meteorological Office, London).
26. Guglielmino-Matessi, C. R., Piazza, A., Menozzi, P. & Cavalli-Sforza, L. L. (1990) *Am. J. Phys. Anthropol.* **83,** 57–68.
27. Ruhlen, M. (1991) *A Guide to the World's Languages* (Stanford Univ. Press, Stanford, CA).
28. Legendre, P., Oden, N. L., Sokal, R. R., Vaudor, A. & Kim, J. (1990) *J. Classif.* **7,** 53–75.
29. Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A. & Lander, E. (1992) *Nat. Genet.* **2,** 204–211.
30. Nevanlinna, H. R. (1972) *Hereditas* **71,** 195–236.
31. Norio, R., Nevanlinna, H. R. & Perheentupa, J. (1973) *Ann. Clin. Res.* **5,** 109–141.
32. Anthony, D. W. & Brown, D. R. (1991) *Antiquity* **65,** 22–38.
33. Diamond, J. M. (1991) *Nature (London)* **350,** 275–276.
34. Anthony, D. W. & Vinogradov, N. B. (1995) *Archaeology* **28** (2), 36–41.
35. Zvelebil, M. & Zvelebil, K. V. (1988) *Antiquity* **62,** 574–583.
36. Renfrew, C. (1992) *Man* **27,** 445–478.
37. Sokal, R. R., Oden, N. L. & Thomson, B. A. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7669–7673.
38. Barbujani, G. & Sokal, R. R. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 1816–1819.