Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 12728–12731, November 1997
Population Biology

Ethnohistory, genetics, and cancer mortality in Europeans

ROBERT R. SOKAL*†, NEAL L. ODEN‡, MICHAEL S. ROSENBERG*, AND DONNA DIGIOVANNI*

*Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245; and ‡The EMMES Corporation, 11325 Seven Locks Road, Suite 214, Potomac, MD 20854

Contributed by Robert R. Sokal, September 8, 1997

ABSTRACT **Geographic variation in cancer rates is thought to be the result of two major factors: environmental agents varying spatially and the attributes, genetic or cultural, of the populations inhabiting the areas studied. These attributes in turn result from the history of the populations in question. We had previously constructed an ethnohistorical database for Europe since 2200 B.C., permitting estimates of the ethnic composition of modern European populations. We were able to show that these estimates correlate with genetic distances. In this study, we wanted to see whether they also correlate with cancer rates. We employed two data sets of cancer mortalities from 42 types of cancer for the European Economic Community and for Central Europe. We subjected spatial differences in cancer mortalities, genetic, ethnohistorical, and geographic distances to matrix permutation tests to determine the magnitude and significance of their association. Our findings are that distances in cancer mortalities are correlated more with ethnohistorical distances than with genetic distances. Possibly the cancer rates may be affected by loci other than the genetic systems available to us, and/or by cultural factors mediated by the ethnohistorical differences. We find it remarkable that patterns of frequently ancient ethnic admixture are still reflected in modern cancer mortalities. Partial correlations with geography suggest that local environmental factors affect the mortalities as well.**

Cancer rates vary geographically and differ among diverse ethnic units (1, 2). In trying to elucidate the causes for such differences, epidemiologists distinguish between genetic, cultural, and environmental factors. Genetic factors comprise specific cancer-causing or predisposing genes, as well as others whose frequencies may serve only to estimate genetic distances between populations, as in this study. Cultural factors known to affect cancer rates include dietary habits, sexual practices, occupational practices, etc. Ethnic differences encompass both genetic and cultural components. Geographic differences, insofar as they do not reflect the populations that inhabit the areas being compared, may represent environmental contrasts.

Here we shall test whether measures of ethnohistorical differences between populations are correlated with their differences in mortality rates due to various cancers. Next, we shall try to hold constant the effects of genetic differences among these populations, and also of geography. We shall also compare the effects of ethnohistory with those of genetics.

MATERIALS AND METHODS

We used three European databases: ethnohistory (3), genetics (4), and cancer mortalities (5, 6) to determine the effects of the first two variables and geography on the mortalities. We computed interlocality distances for all four variables because theory in population genetics (7), genetic epidemiology (8), and anthropology (9) is frequently expressed in terms of distances. In the case of the ethnohistorical distances, they also permitted considerable data compression over the original values. We assembled these distances as matrices and tested the significance of their association by means of Mantel matrix permutation tests (9–12). Ethnohistory (3), genetics (4), and cancer mortalities (unpublished work) are strongly spatially autocorrelated (13). This would result in overly liberal conventional significance tests of their association. We consequently also tested the distance matrix correlations as partial correlations, controlling for geography by using a multiple matrix extension (9, 14) of the Mantel test.

Although mortalities are less reliable cancer rates than incidences, we chose cancer mortalities over other statistics because, for Europe, they are by far the most comprehensive data. Maps and data tables are available for 355 registration areas in the quondam European Economic Community (EEC) (5) and for 194 areas in Central Europe (CE) (6). At the time of reporting the mortalities (1970s), the EEC comprised Belgium, Denmark, Eire, France, Italy, Luxembourg, Netherlands, United Kingdom, and West Germany. The CE data are for 1983–1987 and include Austria, Bulgaria (1986–1987), Czechoslovakia, East and West Germany (the West German data are for a later time span than that of their EEC counterparts), Hungary, Poland, Romania, and Yugoslavia. The EEC cancer mortalities are a balanced data set broken down into 40 site- and sex-specific rates. Thirty-six such rates are furnished for CE. However, the 194 areas of CE include only 32 cancer sites; the remaining 4 sites are not recorded for Romania and are limited to 153 areas. The EEC and CE overlap in West Germany and share 34 site- and sex-specific rates. We decided to analyze the two data sets separately, comparing the findings for the two, rather than pool the data, because the mortalities were for different time periods and partially nonoverlapping cancer types. Furthermore, each data set was sufficiently large to furnish adequate power for the statistical tests undertaken. The mortality rates in refs. 5 and 6 are stated as age-standardized deaths per 100,000 population size per annum. For the EEC and CE, separately, the mortality distances for any site- and sex-specific cancer rate were computed as absolute differences in rates between all pairs of the 355 or 194 (or 153) areas.

The ethnohistorical distances are based on an ethnohistorical database (3) for Europe, compiled in our laboratory and consisting of 3,460 records of ethnic locations and movements from 2200 B.C. to 1970 A.D. There are 1,750 active movement and 1,710 passive location or assimilation records. Each record lists the name of a population unit (e.g., tribe, people) and the language family spoken by them, when known; reports the dates; and defines the areas of movement and location. The ethnohistorical database can be found on the World Wide Web at http://life.bio.sunysb.edu/ee/msr/ethno.html. The pro-

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: EEC, European Economic Community; CE, Central Europe; CAN, cancer mortality differences; ETH, ethnohistorical distances; GEN, genetic distances; GEO, geographic distances.

†To whom reprint requests should be addressed. e-mail: sokal@life.bio.sunysb.edu.

gram ETHNO estimates the admixture of populations from specific language families, following an updating algorithm given in ref. 3. The optimal weights for each type of movement, furnished in that reference, were employed. At the completion of the program there are vectors of estimated proportions of contribution by 17 language families and 2 unknown groups to the population mix at each of 2,216 land-based $1° \times 1°$ quadrats in Europe. Most quadrats receive input from numerous other quadrats (26 on average). From these vectors arc distances (15) are computed between all pairs of quadrats. The distances are estimates of the (dis)similarity between the ethnic mixes of each quadrat pair. Sensitivity experiments showed that ethnohistorical–genetic correlations were remarkably robust against reasonable perturbations in time of movement, location, ethnic (language-family) designation, and completeness of the database (3). To assemble ethnohistorical distance matrices, we chose the set of quadrats that matched the genetic and mortality data locations.

Details of our genetic database for Europe are furnished in ref. 4. It comprises 26 genetic systems with 93 allele or haplotype frequencies and is based on 3,481 samples. Genetic distances were computed separately for each genetic system, because the systems differed in the number and location of sampling points. Systems with fewer than 30 genetic sampling points per region were omitted. This reduced the number of genetic systems to 17 in the EEC data set and to 7 in the CE data. For each genetic sampling point a computer program found the closest cancer registration area in the region being worked on to form a matching pair of gene-frequency and mortality values. If the closest area was more than 100 km from the genetic sampling point, the point was omitted from the study. We computed Prevosti distances (16) between gene-frequency samples and assembled them into genetic distance matrices. The correlations for separate genetic systems (and in some cases for the separate cancers) were averaged to yield the coefficients given in the text and Tables 1 and 2. Geographic distances were calculated as great-circle distances (in km) between all pairs of cancer registration areas within the EEC or CE.

The four types of distance matrices for each of the two regions were designated CAN, ETH, GEN, and GEO for cancer mortality, ethnohistory, genetics, and geography, respectively. We computed zero-order matrix correlations, as well as partial correlations (17), between the distance matrices as follows: *r*(CAN,ETH), *r*(CAN,GEN), *r*(CAN,GEO); *r*(CAN,ETH.GEO), *r*(CAN,GEN.GEO); *r*(CAN,ETH.GEN, GEO), *r*(CAN,GEN.ETH,GEO), and *r*(CAN,GEO. GEN,ETH). These computations were carried out for each cancer site by sex combination and for each genetic system. To compute the zero-order correlations, we employed the Mantel (10, 11) test, with the matrix elements scaled to yield a correlation coefficient as the Mantel product. Partial correlations were obtained from the appropriate residual distance matrices by subjecting them to the Mantel test (9, 12, 14, 18). The significance of each matrix correlation coefficient was assayed by 999 row–column permutations (12). When needed, the resulting probabilities over all cancers or genetic systems were calculated by Fisher's method of combining probabilities (17).

RESULTS

To conserve space we do not feature zero-order and first-order partial correlations for individual cancers but report only average correlations over all cancers in Table 1. These average correlations should be interpreted as measures of central tendency of the individual correlations, and not as a measure of the overall response of "cancer" to ethnohistory or genetics. The latter interpretation is fallacious also because not every cancer is individually significantly correlated with the putative causal factors ethnohistory and genetics. Even the ones that are significant may differ with regard to the ethnohistoric components or the genetic loci that are associated with a specific cancer type. The overall significance tests reported also are not blanket statements for all cancers. Rather they tell us whether the null hypothesis of no correlation of cancer mortality with ethnohistorical distance can be upheld or whether we should conclude that some (or all) cancers are so correlated.

The average correlations in Table 1 seem low by conventional criteria. This is characteristic of correlations between distance matrices (19), which are usually far lower than those of the variables on which they are based. Furthermore, the average coefficients reported here include nonsignificant as well as significant *r* values, which are appreciably higher. Corresponding average correlations are within the same order of magnitude for the EEC and CE. There are more than enough highly significant coefficients for individual cancers so that overall results associated with each average exhibit overall significance ($P < 0.000005$) by Fisher's method. Because CAN, ETH, and GEN are spatially autocorrelated, this would tend to spuriously increase the significance of the correlations. To correct for this we hold constant geographic distances and obtain averages shown in Table 1 in the first-order line. Although the average partial correlations of CAN with GEN in CE are negative, the relatively large and significant individual partial correlations are all positive. Table 1 also shows, for zero-order, first-order, and second-order partial correlations, the number of individual cancers for which correlation of mortality with ethnohistory exceeds that with genetics. Clearly, cancer mortality differences covary with both ethnohistorical and genetic distances, with the former having a much greater effect than the latter.

The second-order partial correlations are shown in Table 2. In the EEC, *r*(CAN,ETH.GEN,GEO) is significant (most at $P < 0.000005$) in 30 of 40 cancers; in CE in 30 of 36. Consequently, the combined probabilities for each region are once more vanishingly small. The values of *r*(CAN,GEN.ETH, GEO) are significant in 16 of 40 cancers in the EEC (each significant, $P \leq 0.02620$); in 7 of 36 in CE (each significant, $P \leq 0.04138$). The combined *P* values are again close to 0. Finally,

Table 1. A summary of zero-, first-, and second-order partial correlations of cancer mortality distances (CAN) with ethnohistorical (ETH) and genetic (GEN) distances in Europe

	EEC			CE		
Order	ETH	GEN	*n* (ETH > GEN)	ETH	GEN	*n* (ETH > GEN)
Zero	0.1632	0.0733	32	0.1585	0.0208	24
First	0.0598	0.0268	23	0.0693	−0.0101	21
Second	0.0891	0.0028	29	0.1445	−0.0133	32

Values in columns 1, 2, 4, and 5 are averages of partial matrix correlation coefficients (9, 14), as follows: zero order in the ETH columns stand for *r*(CAN,ETH), first order is *r*(CAN,ETH.GEO), and second order is *r*(CAN,ETH.GEN,GEO), where GEO stands for geographic distances. For the GEN column just interchange GEN with ETH. In columns 3 and 6, headed *n* (ETH > GEN), we furnish counts of the number of cancers for which correlation of mortality with ethnohistory exceeds that with genetics. EEC, 40 cancers; CE, 36 cancers.

Table 2. Partial correlations of cancer mortality distances with ethnohistorical, genetic, and geographic distances in Europe

		r (CAN,ETH.GEN,GEO)		*r* (CAN,GEN.ETH,GEO)		*r* (CAN,GEO.ETH,GEN)	
Cancer	Sex	EEC	CE	EEC	CE	EEC	CE
Bladder	M	0.1240***	0.2690***	0.0217**	−0.0853	0.1668***	0.1186***
	F	0.1131***	0.0509**	−0.0542	−0.0484	0.2064***	0.1719***
Brain	M	0.2226***		0.0149*		0.0969***	
	F	0.1710***		−0.0321		0.1335***	
Breast	F	0.0040	0.2358***	−0.0714	−0.0572	0.3777***	0.2075***
Cervix	F	0.1591***	0.0624***	−0.0402	0.0282	0.4225***	0.2528***
Colon/rectum or	M	0.3207***	0.2934***	0.0617***	−0.0478	0.1334***	0.0253*
large bowel	F	0.0908***	0.2365***	−0.0035	−0.0628	0.2902***	0.1686***
Esophagus	M	−0.0704	0.0957***	0.0169	−0.0216	0.1021***	0.1383***
	F	0.1549***	0.0008	0.0120*	−0.0079	0.4154***	0.1602***
Gall bladder	M	−0.0999	0.1083**	−0.0146	0.0261	0.1565***	0.0962**
	F	−0.0205	0.1400***	−0.0287	0.0899**	0.0949***	0.1978***
Hodgkin's disease	M	0.1804***	0.0669**	0.0828***	−0.0830	0.0006*	−0.0031
	F	0.1555***	0.1622***	0.0641**	−0.0161	0.0469***	−0.0617
Kidney	M		0.4457***		−0.0743		0.1574***
	F		0.3662***		−0.0499		0.0519***
Larynx	M	−0.0890	0.0814**	0.0294*	0.0664*	0.2445***	0.1121***
	F	0.0498*	0.1045***	0.0309***	0.0408	0.0576***	−0.0187
Leukemia	M	0.0174**	0.1944***	0.0210**	0.0203	0.0609***	−0.0282
	F	0.1999***	0.1825***	0.0174**	−0.0161	0.0545***	−0.0442
Lung	M	0.0243*	0.1782***	−0.0042	0.0424	0.0844***	0.0189
	F	0.0565**	0.1621***	−0.0363	−0.0565	0.3632***	−0.0605
Lymphoma	M	0.0953***	0.0751*	−0.0359	−0.0398	0.0985***	0.1468***
	F	0.1167***	0.2228***	−0.0386	−0.0612	0.0942***	0.1198***
Malignant	M	0.1994***	0.1357***	−0.0209	−0.0898	0.2588***	0.0643*
melanoma	F	0.1620***	0.0801***	−0.0464	−0.0469	0.2924***	0.0974***
Multiple	M	0.0006		−0.0078		0.1597***	
myeloma	F	0.0328		0.0074		0.1204***	
Oral	M	−0.0956	0.0894***	0.0016	0.0263*	0.0941***	−0.0669
	F	−0.0235	0.0267	−0.0150	0.0200	0.0625***	−0.0551
Ovary	F	0.0828***	0.2732***	−0.0278	−0.0469	0.4788***	0.1727***
Pancreas	M	0.0883***	0.2159***	0.0635**	−0.0566	0.3171***	0.0438**
	F	−0.0049	0.2681***	0.0192	−0.0424	0.4398***	0.1156***
Prostate	M	0.1864***	0.1997***	0.0414***	−0.0011	0.1426***	0.2518***
Stomach	M	0.0359*	−0.0186	−0.0056	0.0156*	−0.0363	0.1385***
	F	0.0500*	0.0025	−0.0052	−0.0137	−0.0102	0.1378***
Testis	M	0.2522***	0.1284***	0.0648***	−0.0252	0.1702***	−0.0242
Thyroid	M	0.0883***	−0.0231	0.0711***	0.1358***	−0.0087	0.1549***
	F	0.0263**	0.1010***	0.0468**	0.1232***	0.0478***	0.0823***
Urinary tract	M	0.0883***		0.0015*		0.1982***	
	F	0.2056***		−0.0496		0.3581***	
Uterus	F	0.2127***	−0.0264	−0.0408	0.0550**	0.1357***	0.3098***
Mean *r*		0.0891	0.1445	0.0028	−0.0100	0.1731	0.0907
Combined *P*		0.00000	0.00000	0.00000	0.00002	0.00000	0.00000

Values are partial matrix correlation coefficients (9, 14), as described by the column headings; M, males; F, females. *, $0.05 \geq P > 0.01$; **, $0.01 \geq P > 0.001$; ***, $0.001 \geq P$.

the partial correlations *r*(CAN,GEO.ETH,GEN) are highly significant in 36 of 40 cancers in the EEC; in 22 of 36 in CE. The combined *P* values are vanishingly small. A correlation test (not detailed here) failed to find significant associations in partial correlations between the EEC and CE.

To support these findings we carried out a spatial randomization of the ethnohistory movement records, keeping their chronological sequence constant, but randomizing the location of the target quadrat while retaining the original direction and distance of the movement. Because the randomized records were constrained to have both source and target areas fully land-based, some records (especially in the EEC) can relocate only in a few limited locations. This randomization test was described in ref. 3, where it was employed to test the significance of the observed GEN,ETH correlations. In the present supporting analysis, we recorded the position of the observed correlations with respect to the distribution of 100 randomized samples. In the EEC the correlation of observed ETH with CAN is higher than any correlation of randomized ETH with CAN in 29 of 40 cancers, and it is in second to fifth place in an additional 5 cancers. Twenty-four of these 34 cancers remain significant for the CAN,ETH.GEO partial correlation. In the CE data fewer cancers are significant (13 and 9 of 32, respectively), although that too is far more than expected. Finally, if the average zero-order and first-order correlations are computed, we find that the observed average correlation is higher than any of the randomized correlations. These results strongly support the significance of the cancer correlations with ethnohistory.

DISCUSSION

How are we to interpret these findings? Clearly, these cancer mortality differences are not likely to have brought about either ethnohistorical or genetic differences between popula-

tions. It seems more likely that causal paths course from genetics and ethnohistory to mortality. We know from previous work (3) that our ethnohistorical distances predict modern genetic distances. Even casual inspection of Tables 1 and 2 reveals that the great majority of individual correlations as well as all the average correlations show higher values with ethnohistory than with genetics. It is quite difficult to test for the significance of this effect by conventional methods because we cannot assume independence of the cancers. However, in our opinion, dependence between cancers is not strong enough to explain away this finding.

Most genetic systems employed in this study probably have little effect on cancer rates. Other loci, which we have not studied, may well differ with ethnicity and may actually mediate the effects of ethnohistory on the mortalities. It is therefore not legitimate to look for cancers known to have high mortalities in response to the presence of some carcinogenic alleles and to expect that, in consequence, the correlation with our genetic data should be especially high. Our genetic distances are estimates of the overall distances between pairs of sampling stations. They may or may not successfully predict that a given cancer is affected by alleles at a given locus. Our findings are compatible with a model in which cancers are partially determined by additive effects of genes. Such cancers would yield substantial partial correlations with our genetic data, but cancers driven by single, rare allelomorphs that are not represented in our data would not be correlated with them.

We should also keep in mind that differences in ethnic composition may encompass cultural differences that lead to diverging cancer rates. Although some of the ethnic admixtures in our model are quite ancient, some cultural traits that affect cancer mortalities may have persisted in the modern admixed populations. In any case, whether mediated by genetics or by culture, it is clear that the ethnohistorical affinities contribute to differences in cancer mortalities.

Although the average correlations and combined probabilities at the bottom of Table 2 permit general statements about the relations of ethnohistorical and genetic distances to the set of individual cancer mortalities, we note that for some cancers (e.g., multiple myeloma M + F, oral F, and stomach F), the mortality differences are not affected by either factor. If genetic factors affect these, they presumably are not in our genetic database. It is also possible that they are subject to environmental influences for which our analysis did not test.

The partial correlations of cancer with geography shown in Table 2 are generally higher than those with either ethnohistory or genetics. This implies that in the absence of ethnohistory and genetics, greater geographic distances are associated with greater differences in cancer mortality. Such an effect is most likely due to strong environmental differences with increasing distance, and it emphasizes the important role of the environment in cancer causation.

We are indebted to Dr. D. M. Parkin of the International Agency for Research on Cancer for a critical reading of the manuscript. This work was supported by Grant SBR 9419349 from the National Science Foundation to R.R.S. This is Contribution No. 1003 in Ecology and Evolution from the State University of New York at Stony Brook.

1. Higginson, J., Muir, C. S. & Muñoz, N. (1992) *Human Cancer: Epidemiology and Environmental Causes* (Cambridge Univ. Press, Cambridge, U.K.).
2. Polednak, A. P. (1989) *Racial and Ethnic Differences in Diseases* (Oxford Univ. Press, New York).
3. Sokal, R. R., Oden, N. L., Walker, J., DiGiovanni, D. & Thomson, B. A. (1996) *Hum. Biol.* **68,** 873–898.
4. Sokal, R. R., Harding, R. M. & Oden, N. L. (1989) *Am. J. Phys. Anthropol.* **80,** 267–294.
5. Smans, M., Muir, C. S. & Boyle, P., eds. (1992) *Atlas of Cancer Mortality in the European Economic Community* (International Agency for Research on Cancer, Lyon, France) IARC Scientific Publ. No. 107.
6. Zatonski, W., Smans, M., Tyczynski, J. & Boyle P., eds. (1996) *Atlas of Cancer Mortality in Central Europe* (International Agency for Research on Cancer, Lyon, France) IARC Scientific Publ. No. 134.
7. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
8. Morton, N. E. (1982) *Outline of Genetic Epidemiology* (Karger, Basel).
9. Smouse, P. E. & Long, J. C. (1992) *Yearbk. Phys. Anthropol.* **35,** 187–213.
10. Mantel, N. (1967) *Cancer Res.* **27,** 209–220.
11. Sokal, R. R. (1979) *Syst. Zool.* **28,** 227–231.
12. Hubert, L. J. (1987) *Assignment Methods in Combinatorial Data Analysis* (Dekker, New York).
13. Cliff, A. D. & Ord, J. K. (1981) *Spatial Processes: Models and Applications* (Pion, London).
14. Smouse, P. E., Long, J. C. & Sokal, R. R. (1986) *Syst. Zool.* **35,** 627–632.
15. Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967) *Evolution* **21,** 550–570.
16. Prevosti, A., Ocaña, J. & Alonso, G. (1975) *Theor. Appl. Genet.* **45,** 231–241.
17. Sokal, R. R. & Rohlf, F. J. (1995) *Biometry* (Freeman, New York), 3rd Ed.
18. Oden, N. L. & Sokal, R. R. (1992) *J. Class.* **9,** 275–290.
19. Sokal, R. R., Oden, N. L., Rosenberg, M. S. & DiGiovanni, D. (1997) *Am. J. Hum. Biol.* **9,** 391–404.