J. biosoc. Sci. (2002) **34**, 289–301

SURNAME ANALYSIS OF THE CORSICAN POPULATION REVEALS AN AGREEMENT WITH GEOGRAPHICAL AND LINGUISTIC STRUCTURE

LAURA MORELLI*, GIORGIO PAOLI† AND PAOLO FRANCALACCI*

**Dipartimento di Zoologia e Antropologia Biologica, Università di Sassari, Sassari, Italy and †Dipartimento di Ecologia, Etologia e Evoluzione, Università di Pisa, Pisa, Italy*

Summary. The surname is a cultural trait that is extremely useful for historical and linguistic studies and can effectively be used as a genetic marker. In many human populations the surname is inherited in the paternal lineage, and can therefore be considered a marker for the Y chromosome. In this study, surnames were recorded from the white pages of telephone directories in current use in Corsica in 1993. All surnames present in thirteen villages scattered over the whole island and covering the main historical regions were transcribed. Surname variability was found to be higher in coastal villages, and lower in more isolated communities. The isonymy detected among the thirteen villages allowed the calculation of kinship values, visualized in a tree showing two main clusters, one referring to the northern villages and one encompassing the villages of the south. The pattern reflects the administrative division of the island, with the exception of Vico, which belongs to the southern administrative region but is geographically close to the northern villages, and Ghisoni, which belongs to the northern district but is more similar to the village of Bastelica in the southern district. The data presented here show a structure in the surname distribution that is in substantial agreement with the geographical patterns. The kinship values are consistent with a moderated gene flow among villages producing a surname structure according to the geographic features of the territory.

Introduction

The use of surname analysis in studies of population genetics has a number of advantages, such as simple and inexpensive data collection, the high number of individuals sampled (up to whole populations) and the possibility of a diachronic approach whenever historical archives are available (Lasker, 1985). However, the methodology has some disadvantages that should be considered if data are to be

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

Table 1. Number of individuals, number of surnames and altitude of villages analysed

Town	Abbreviation	No. of individuals	No. of surnames	Altitude (m.a.s.l.)
Bastelica	BS	298	138	885
Bonifacio	BN	1045	610	80
Calacuccia	CC	206	65	812
Calenzana	CZ	702	379	273
Ghisoni	GH	229	100	635
Lecci	LC	467	402	80
Levie	LV	383	170	112
Luri	LR	332	193	112
Morosaglia	MR	175	117	805
Oletta	OL	346	232	235
Sartene	SR	1252	576	421
Venaco	VN	309	159	580
Vico	VC	385	225	491

interpreted correctly, mainly due to the fact that the surname is a cultural trait and not a real genetic marker, and the linkage is not valid in all cases. In addition, its recent origin (starting from about 1600 in European cultures) prevents the study of ancient peopling events, and the polyphyletic origin of some surnames can bias phylogenetic analysis.

In spite of these limitations, surname analysis has been used to evaluate the dynamics of populations (estimated by the persistence, entrance or extinction of surnames), thereby reflecting the demographic history of the peoples carrying them (Paoli, Franceschi & Lasker, 1999; Pettener, 1990; Sanna *et al.*, 1999). Features of population genetics such as endogamy, inbreeding, genetic drift, similarity or differentiation among populations and migratory patterns can be estimated by surname analysis (Biondi *et al.*, 1996; Lucchetti & Soliani, 1989; Martuzzi Veronesi, Gueresi & Pettner, 1996; Paoli, Franceschi & Taglioli, 1996; Vona *et al.*, 1996). In particular, kinship estimated from surnames can be correlated with different biological, cultural or environmental parameters. For example, the degree of inbreeding, calculated using data reported on marriage dispensations, increases significantly with altitude (Cavalli-Sforza & Bodmer, 1971; Fuster *et al.*, 1996), and the same correlation has been shown with values of local kinship (Franceschi & Paoli, 1994).

This study examined the population genetics of the Mediterranean island of Corsica (France) through surnames analysis and estimated the relation between isonymy structure and geographical and linguistic patterns. The Corsican region was chosen because: (1) Corsica, as an island, is an interesting and relatively restricted experimental model for studying internal processes of differentiation; (2) it has a particular orography that determines relative geographical and cultural barriers inside the territory; (3) it has not been studied before using this approach, and its genetic location in the Mediterranean context is still a matter of controversy.

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

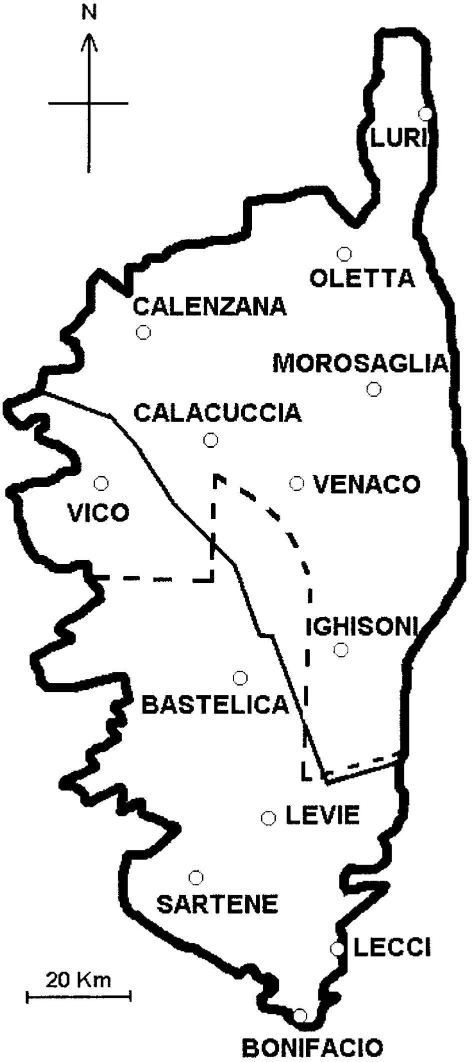

Fig. 1. Positions of studied villages. The continuous line indicates the administrative subdivision, while the dotted one indicates the main linguistic division.

The environmental features of the island should affect the degree of isolation of its different communities, and the surname distribution should reflect this. The correlation between altitude, local kinship and population size was assessed with the aim of determining which factor is most important for causing isolation. In addition, the reasons were evaluated for similarity between villages following internal geographical barriers or administrative and linguistic divisions of the island by comparing the matrix of kinship with the geographic and linguistic distance matrices. Available genetic data and historical knowledge were used to develop predictions of surname distribution in the Corsican population.

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

Table 2. Matrix of local (diagonal) kinship (r_{ii}) and between-population (below the diagonal) kinship (r_{ij}) values and genetic distance values (above the diagonal). Reported values were multiplied by 10,000. The names are abbreviated as in Table 1

	BS	BN	CC	CZ	GH	LC	LV	LR	MR	OL	SR	VN	VC
BS	**30·39**	36·56	154·68	48·34	87·51	31·83	83·63	43·39	50·98	40·04	46·52	49·61	42·80
BN	0·26	**6·69**	129·70	23·86	67·69	7·63	56·57	19·49	27·90	16·98	21·64	28·67	21·40
CC	1·79	2·43	**127·87**	127·62	181·35	128·59	180·89	133·43	134·32	130·34	143·08	140·51	119·66
CZ	0·29	0·68	9·39	**18·53**	76·97	19·53	73·13	29·81	36·48	27·94	35·54	40·63	25·54
GH	2·42	0·48	4·24	1·76	**61·96**	63·28	114·84	71·06	81·99	70·15	78·21	80·48	75·87
LC	0·50	0·75	0·86	0·72	0·56	**2·44**	53·00	16·26	25·11	12·91	18·47	25·32	17·99
LV	1·25	2·93	1·36	0·57	1·43	2·59	**55·74**	68·14	77·67	66·41	58·61	76·22	69·91
LR	0·78	0·88	4·50	1·64	2·73	0·37	1·08	**14·56**	31·93	22·89	31·25	33·58	27·27
MR	1·53	1·22	8·60	2·85	1·81	0·49	0·86	3·14	**23·65**	27·94	39·98	25·69	34·90
OL	1·04	0·72	4·63	1·16	1·77	0·63	0·53	1·70	3·72	**11·73**	28·74	28·95	25·04
SR	1·13	1·72	1·59	0·69	1·07	1·18	7·76	0·85	1·03	0·69	**18·39**	40·11	33·40
VN	2·09	0·71	5·38	0·65	2·44	0·26	1·46	2·19	10·68	3·09	0·84	**23·40**	34·63
VC	2·20	1·05	12·51	4·90	1·45	0·63	1·32	2·05	2·78	1·75	0·90	2·79	**16·81**

The island of Corsica (western Mediterranean) has a population of about 250,000 in a territory of 8680 km^2. The demographic development of the Corsican population has been hampered by several bottleneck events. Between the IVth and Vth centuries the coastal towns became progressively extinct, destroyed by decaying trade, malaria, and incursion by pirates. In this period, the island population was reduced to 50,000 inhabitants. During the Middle Ages, Corsica was ruled by Pisa (from 1078), which gave the Neolatin language of the island its main ‘footprint’. About three centuries later, Genoa substituted the Pisan government, but its linguistic influence was limited to a few fortified towns and harbours (Bonifacio, Calvi, etc.). In 1348 the islanders were affected by a plague epidemic which reduced the population of the island dramatically. The resulting economical and social depression led to a significant emigration to Pisa. France obtained Corsica after the Versailles Treaty of 1768, and divided the island into two administrative regions: a southern (*Corse du Sud*) and a northern one (*Haute Corse*). Corsica can also be divided into many historical regions that cluster into two main geographical areas: *Banda di Fuori* in the south-west and *Banda di Dentro* in the north-east, which are divided by a mountain range.

The Corsican language belongs to the Tuscan dialectal area of the Italian language system. Even though intelligibility is complete within the island (about 80% of overall lexical similarity; Grimes, 1996), two main dialectal areas can be recognized: the *Cismontano* in the north, with stronger Tuscan influences, and the *Oltremontano* in the south, which shares many affinities with the Gallurese dialect spoken in northern Sardinia. Genoan dialectal traits are maintained in Bonifacio and Calvi. Recent loans from French are also present in the modern Corsican language.

The genetics of Corsica are poorly understood, especially in comparison with neighbouring Sardinia. Qualitative and quantitative dermatoglyphic characteristics,

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

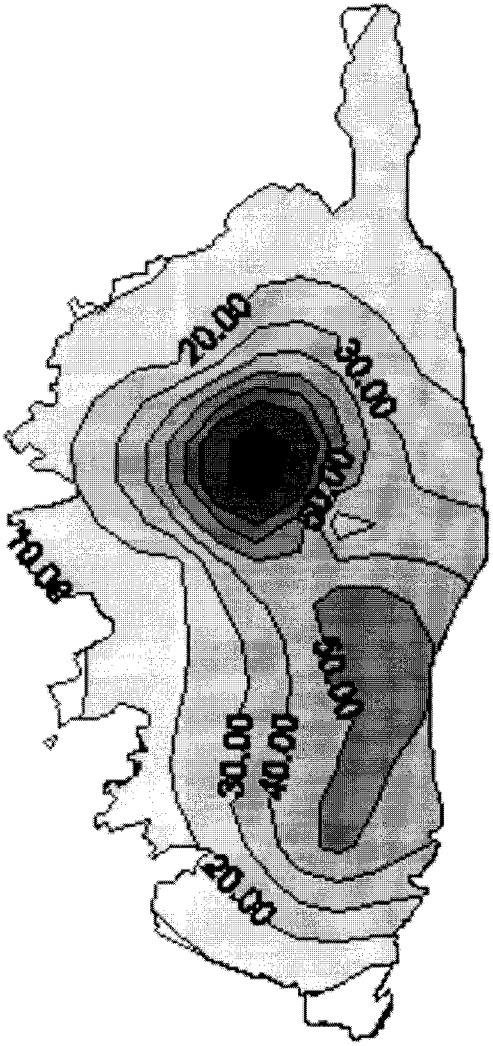

Fig. 2. Map of local kinship.

sensitive to genetic and environmental factors, place the Corsican population in an intermediate position between north Sardinian and central Italian populations (Morelli *et al.*, 1999). A previous paper on classical genetic markers, such as blood groups and serum proteins (Calafell *et al.*, 1996), has shown a large genetic distance of Corsicans from Sardinia and Tuscany, leading to the conclusion that the historically known relationships between Corsica and these two regions are mainly cultural, excluding a significant gene flow. Other studies (Vona *et al.*, 1995; Varesi *et al.*, 1996) on fourteen genetic markers have pointed to a similarity with Sardinia, and larger differences from Tuscanians and Ligurians, who apparently left almost no trace in the genes of Corsicans. However, these contradictory results can be explained

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

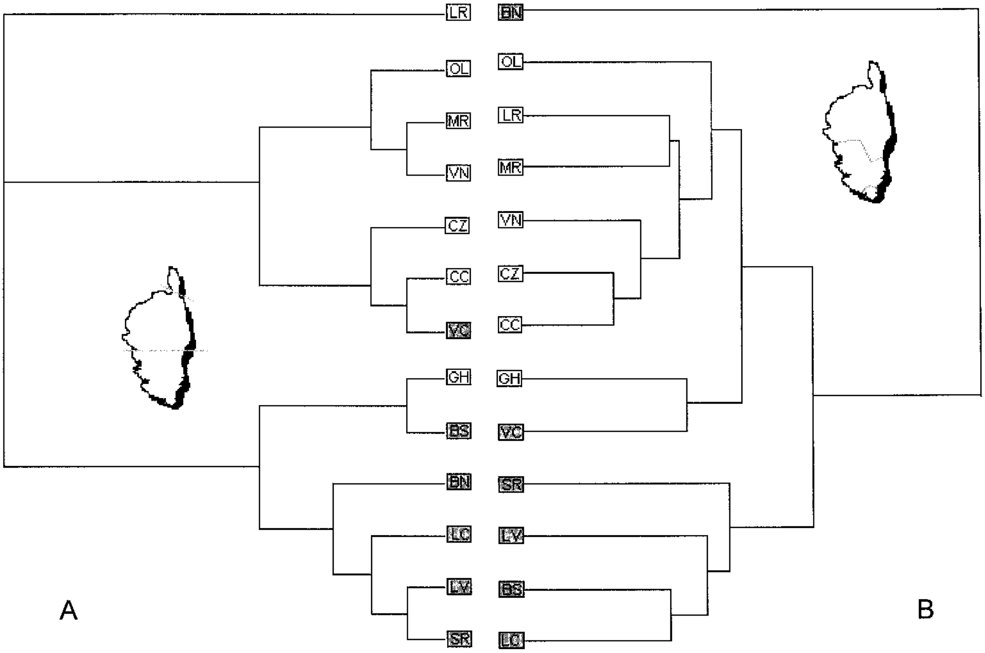

Fig. 3. Trees of genetic (A) and linguistic (B) distances. Shaded boxes indicate villages belonging to the southern administrative district. The names are abbreviated as in Table 1.

in terms of drift, which is particularly important in an island environment and affects the various studied genetic markers dramatically and differently. In fact, a study on the mtDNA haplogroups (Morelli *et al.*, 2000), which is less sensitive to recent micro-differentiation, pointed to a remarkable similarity of the Corsican haplogroup distribution with those of Tuscany and northern Sardinia, and a strong discontinuity with central Sardinia and Catalonia. Similar results were also obtained using classical markers to analyse the towns of Corte and Bastia (Moral *et al.*, 1996).

Materials and methods

Surnames were obtained from the white pages of telephone directories in current use in Corsica in 1993. All surnames present in thirteen villages scattered over the whole island (Table 1) and covering the main historical regions were transcribed (Fig. 1). Villages were selected on the basis of their geographical position and population size, as small villages would presumably have maintained a more conservative identity than the main towns and tourist settlements where recent immigration has had a major influence. However, the coastal resort of Lecci was included in the study as a control. Names of corporations, companies and other institutions were excluded from the analysis. The use of telephone directories can introduce a bias in surname analysis

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

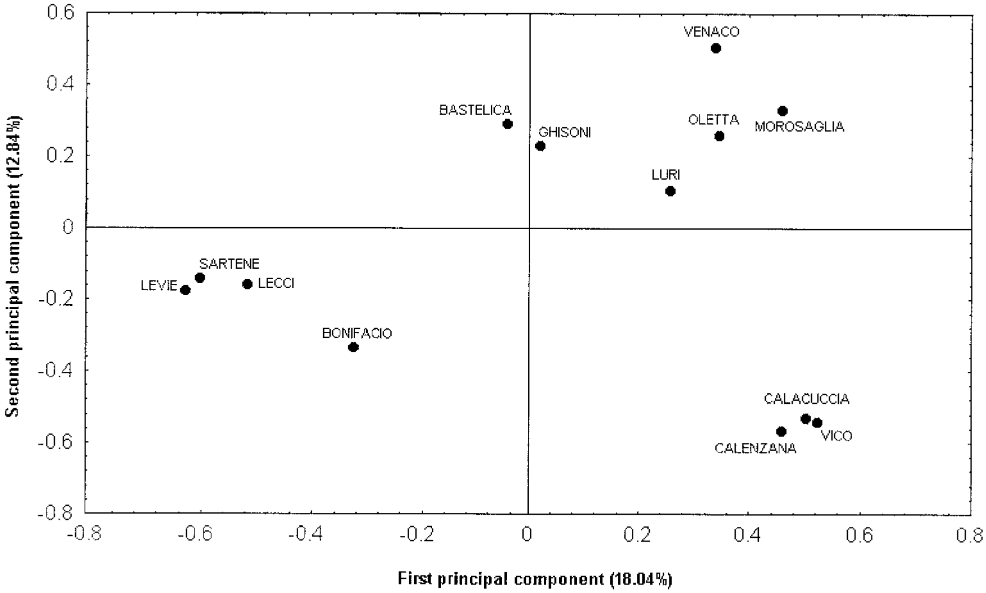

Fig. 4. Spatial distribution of the first and second principal components. The percentages of variance are given in parentheses.

where there is an uneven diffusion of telephones in the population due to socioeconomic factors, such as urbanization, social status, age classes, etc. However, this is not the case in a technologically developed country like France, where the use of the telephone is so widespread that directories report nearly all resident families, and thus represent a random sample of the whole population.

The kinship values by surnames within (r_{ii}) and between (r_{ij}) villages were evaluated according to the method of Relethford (1988) as one-quarter of the random isonymy coefficient. The local kinship values (r_{ii}) calculated from the within-group random isonymy are essentially a measure of the genetic isolation of each subdivision. A high degree of isolation, due in the present study to the altitude of the village, should be reflected by high values of r_{ii}. Local kinship values were interpolated into a grid delimited by parallels 41 °N and 43 °N and by meridians 8 °E and 10 °E and grouped by contours in order to obtain different patterns of local kinship's coverage of the map's surface.

The 'kinship between populations' matrix was used for the Principal Components analysis using the package 'Statistica'. Geographic maps for the first two principal components extracted were drawn as described above. Genetic distances (D^2) were derived from the kinship values, according to the formula reported in Relethford (1988), and were used to form a phylogenetic tree using the UPGMA method of clustering of the Phylogeny Inference Package, Phylip version 3·2 (Felsenstein, 1989). This tree was compared with a linguistic one obtained using linguistic similarities (Calafell *et al.*, 1996).

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

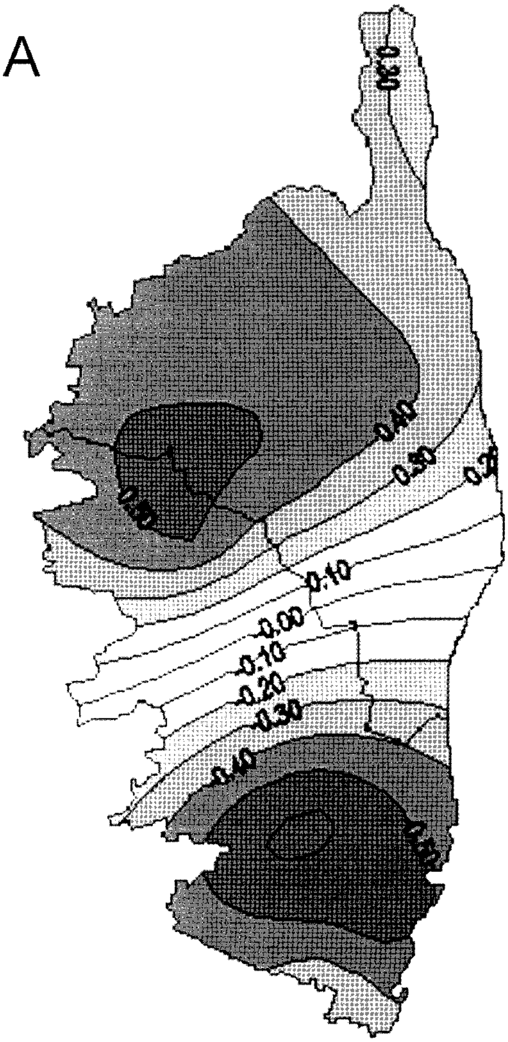

Fig. 5A.

To analyse the relationships between genetic similarity (*R* matrix values) and geographic and linguistic data, Spearman's *rho* correlation and partial correlation coefficients were computed. This method was preferred to Pearson's product moment correlation '*r*' since it assumes only a monotonic relationship, not a linear one, between the variables (Pollard, 1977). Since the *R* matrix values measure similarity, whereas the geographic distance matrix measures dissimilarity between population subsamples, these last values were transformed through a sign inversion (Holloway & Sofaer, 1989). Probabilities of correlation between matrices were derived using a directional hypothesis as required by basic models of genetic population structure, i.e.

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

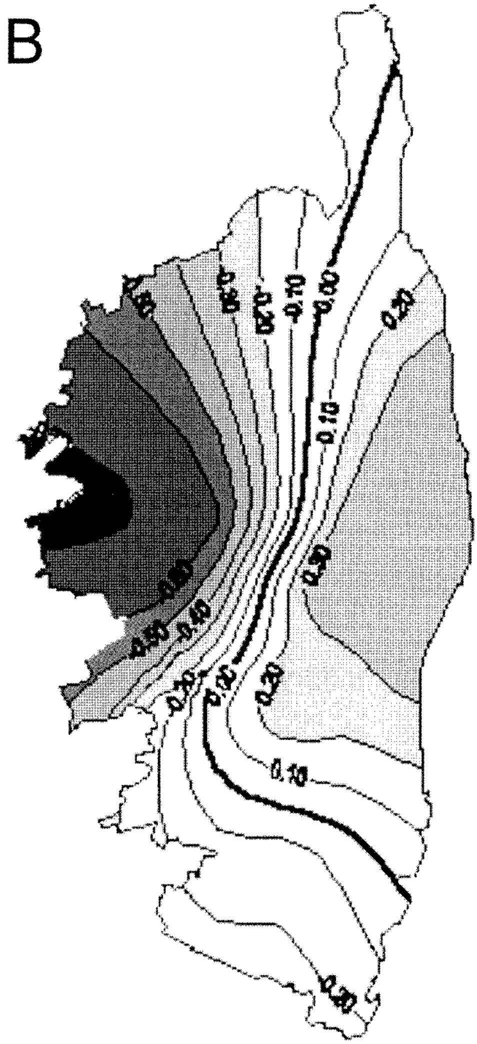

Fig. 5B.

Fig. 5. First (A) and second (B) principal component maps of Corsica.

that the correlation between genetic, linguistic and geographic affinity should be positive. Probability values were assessed using the Mantel test (Mantel, 1967; Relethford, 1990); the number of permutations was fixed at 10,000 to minimize fluctuation of probability values (Jackson & Somers, 1989). Linguistic similarities were obtained from the amount of identity in a set of common words (Calafell *et al.*, 1996). Geographic distances were calculated according to the minimum road distance. Partial correlations were evaluated using the software developed by De Vries, Netto & Hanegraaf (1993).

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

Table 3. Spearman's rank correlation for kinship (KIN), and geographic (GEO) and linguistic (LIN) matrices

Matrices compared	Correlation (*r*)	Significance *p* (Mantel's test)	Proportion of the variance explained (%)
Correlations			
KIN × GEO	0·5940	0·0001	35·3
KIN × LIN	0·5848	0·0001	34·2
GEO × LIN	0·5623	0·0000	31·6
Partial correlations			
KIN × GEO (LIN)	0·3150	0·0013	9·9
KIN × LIN (GEO)	0·2470	0·0037	6·1

Results and discussion

The isonymy detected among the thirteen villages allowed the calculation of a kinship matrix (Table 2). The local kinship map (Fig. 2) illustrates that the highest kinship values are in the internal region, corresponding to the mountainous areas of the island. A significant positive correlation ($r=0.61$, $p<0{\cdot}05$) was found between altitude and local kinship values, while population size was independent of altitude and local kinship. For this reason altitude seems to be the most important factor causing isolation in Corsica.

The inter-population kinship data reported in Table 2 were used to infer genetic distances, and these can be visualized in the tree shown in Fig. 3A. The villages are subdivided into three main clusters, one referring to the northern villages and one encompassing the villages of the south. The third is the village of Luri. The pattern corresponds with the administrative division of the island, with the exception of Vico, which belongs to the southern administrative region, but is geographically close to the northern villages. Vico is in a subgroup of the northern cluster including the villages of the western area (Vico, Calacuccia and Calenzana). Another exception is the village of Ghisoni, which belongs to the *Haute Corse* but clusters with the nearest village Bastelica, in the *Corse du Sud* group.

Comparison of the genetic and linguistic trees (Fig. 3A and B) points out the linguistic isolation of Bonifacio from other Corsican villages where intelligibility is greater. Genetic distances, however, do not seem to have been affected by this difference. A differentiation between the northern and southern villages is also apparent in this tree.

The main north–south differentiation is also apparent in the first principal component (Fig. 4), which explains 18·04% of the total variance. Northern populations show positive values of the relative factor coefficients, with a peak in the north-west, while negative values can be found in the south with a maximum in the area of Sarténe, as can be seen in the geographical representations reported in Fig. 5A. The second principal component, explaining 12·84% of the variance, shows a differentiation between western and eastern populations (Fig. 5B).

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

Correlation and partial correlation coefficients between genetic, geographic and linguistic matrices are reported in Table 3. The correlation between any two given matrices showed the expected sign and was highly significant, indicating that both geography and language are important determinants of the kinship by isonymy among the population subsamples. As reported in Table 3, 35·3% and 34·2% of the genetic variance is explained, respectively, by geography and linguistic matrix values. Furthermore, when either the language or geography was kept constant, there was a significant association between kinship by isonymy and, respectively, language and geography.

These results could be explained by moderated gene flow, since either complete endogamy or panmixia would have produced no correlation with the geographical pattern.

Conclusions

The data presented here show a structure in surname distribution that is in substantial agreement with geographical and linguistic patterns, and which does not necessarily reflect the administrative division of the island. The main north–south dialectal subdivision (*Cismontano–Oltremontano*) is reflected by the kinship values calculated for surnames. The geographical west–east division (*Banda di Fuori–Banda di Dentro*) is represented by the second principal component. It is interesting to note that the relatively recent administrative division of the island has not produced changes in the surname structure. In fact, the village of Vico belongs to the southern district, and Calacuccia and Calenzana to the northern one, but they are strictly associated in terms of kinship, and they are included in the *Cismontano* dialectal area. Analogously, the central communities of Bastelica and Ghisoni, although administratively divided, are geographically close and are in an intermediate position between the northern and southern clusters in the principal component map (Fig. 4). However, a major contribution to this differentiation pattern has been made by environmental factors, such as geographical barriers, which hamper gene flow, and altitude, which is greatly involved in internal isonymy.

In conclusion, the kinship values obtained in thirteen Corsican villages are consistent with a moderated gene flow among villages, sensitive to the geographical and linguistic structure of the territory. Surnames provide a short-term scale tool for human genetics analyses, since surname attribution goes back several centuries. Considering the demographic changes that have affected the island of Corsica in modern times, the surname approach, which maps recent microevolutionary events, was particularly reliable.

Acknowledgments

The authors would like to thank Laurent Varesi for study materials and useful information, Francesc Calafell for supplying the linguistic similarity matrix, Luca Taglioli for technical help and Giuseppe Vona for helpful discussion. L. M. was awarded a research contract (art. 51, L. 27.12.1997). The research was supported by

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

grants from the Italian National Council of Researches CNR 97.00712PF36 (L. Terrenato) and MURST 60% to P. F.; 40% Cofin. 1999 to P. F. and G. P.; Young Researchers' Fund of the University of Sassari to L. M.

References

BIONDI, G., RASPE, P., MASCIE-TAYLOR, C. G. N. & LASKER, G. W. (1996) Repetition of the same pair of surnames in marriage in Albanian Italians, Greek Italians, and the Italian population of Campobasso province. *Hum. Biol.* **68**, 573–583.

CALAFELL, F., BERTRANPETIT, J., RENDINE, S., CAPPELLO, N., MERCIER, P., AMOROS, J. P. & PIAZZA, A. (1996) Population history of Corsica: a linguistic and genetic analysis. *Ann. hum. Biol.* **23**, 237–251.

CAVALLI SFORZA, L. L. & BODMER, W. F. (1971) *The Genetics of Human Populations*. Freeman, San Francisco.

DE VRIES, H., NETTO, W. J. & HANEGRAAF, P. L. H. (1993) Matman: a program for the analysis of sociometric matrices and behavioral transition matrices. *Behavior* **125**, 157–175.

FELSESTEIN, J. (1989) PHYLIP – Phylogeny Inference Package (Version 3·2). *Cladistics* **5**, 164–166.

FRANCESCHI, M. G. & PAOLI, G. (1994) Isolation factors and kinship by isonymy in a group of parishes in northern Tuscany (Italy): influence of within-parish similarity level on between-parish similarity pattern. *Hum. Biol.* **66**, 905–916.

FUSTER, V., MORALES, B., MESA, M. S. & MARTIN, J. (1996) Inbreeding patterns in the Gredos mountain range (Spain). *Hum. Biol.* **68**, 75–93.

GRIMES, B. F. (1996) *Ethnologue: Languages of the World*, 13th Edition. SIL International, Dallas.

HOLLOWAY, S. M. & SOFAER, J. A. (1989) Coefficients of relationship by isonymy within and between the regions of Scotland. *Hum. Biol.* **61**, 87–97.

JACKSON, D. A. & SOMERS, K. M. (1989) Are probability estimates from the permutation model of Mantel's test stable? *Can. J. Zool.* **67**, 766–769.

LASKER, G. W. (1985) *Surnames and Genetic Structure*. Cambridge University Press, Cambridge.

LUCCHETTI, L. & SOLIANI, L. (1989) Similarità tra popolazioni esaminate mediante i cognomi. *Riv. Antrop.* **67**, 181–198.

MANTEL, N. (1967) The detection of disease clustering and generalized regression approach. *Cancer Res.* **27**, 209–220.

MARTUZZI VERONESI, F., GUERESI, G. & PETTENER, D. (1996) Biodemographic analysis of Italian alpine communities (Upper Sole Valley, 1725–1923). *Riv. Antrop.* **74**, 55–75.

MORAL, P., MEMMI, M., VARESI, L., MAMELI, G. E., SUCCA, V., GUTIERREZ, B., LUTKEN, N. & VONA, G. (1996) Study on the variability of seven genetic serum protein markers in Corsica (France). *Anthrop. Anz.* **54**, 97–107.

MORELLI, L., GROSSO, M. G., VONA, G., VARESI, L., TORRONI, A. & FRANCALACCI, P. (2000) Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. *Hum. Biol.* **72**, 585–595.

MORELLI, L., VONA, G., VARESI, L., MEMMI, M., AUTUORI, L. & CALÒ, C. M. (1999) Finger dermatoglyphics in the Corsican population (France). *Anthrop. Anz.* **57**, 339–347.

PAOLI, G., FRANCESCHI, M. G. & LASKER, G. W. (1999) Changes over 100 years in degree of isolation of 21 parishes of the Lima Valley, Italy, assessed by surname isonymy. *Hum. Biol.* **71**, 123–133.

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press

Paoli, G., Franceschi, M. G. & Taglioli, L. (1996) Kinship by isonymy and by gene frequencies: a comparison of population structures at different hierarchical population levels. *Am. J. hum. Biol.* **8**, 445–455.

Pettener, D. (1990) Temporal trends in marital structure and isonymy in S. Paolo Albanese, Italy. *Hum. Biol.* **62**, 837–851.

Pollard, J. H. (1977) *Numerical and Statistical Techniques.* Cambridge University Press, London.

Relethford, J. H. (1988) Estimation of kinship and genetic distance from surnames. *Hum. Biol.* **60**, 475–492.

Relethford, J. H. (1990) *Mantel: A Microcomputer Program for Computing the Mantel Probability between Distance Matrix Elements.* Department of Anthropology, State University of New York, College at Oneonta, NY.

Sanna, E., Villecco, M., Melis, M. & Floris, G. (1999) Endogamia, esogamia, consanguineità e coefficiente di relazione dall'isonimia: studio condotto su dieci comuni della Sardegna dal 1800 al 1974. *Riv. Antrop.* **77**, 129–150.

Varesi, L., Memmi, M., Moral, P., Mameli, G. E., Succa, V. & Vona, G. (1996) La distribution de quatorze marqueurs génétiques dans la population de l'île de Corse (France). *Bull. Mém. Soc. Anthropol. Paris* **8**, 5–14.

Vona, G., Francalacci, P., Paoli, G., Latini, V. & Salis, M. (1996) Study of the matrimonial structure of the population of central Sardinia (Italy). *Anthrop. Anz.* **54**, 317–329.

Vona, G., Memmi, M. R., Varesi, L., Mameli, G. E. & Succa, V. (1995) A study of several genetic markers in the Corsican population (France). *Anthrop. Anz.* **53**, 152–132.

https://doi.org/10.1017/S0021932002002894 Published online by Cambridge University Press