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
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 surname 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.

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

<table>
<thead>
<tr>
<th>Town</th>
<th>Abbreviation</th>
<th>N. of individuals</th>
<th>N. of surnames</th>
<th>Altitude (m.a.s.l.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bastelica</td>
<td>BS</td>
<td>298</td>
<td>138</td>
<td>885</td>
</tr>
<tr>
<td>Bonifacio</td>
<td>BN</td>
<td>1045</td>
<td>610</td>
<td>80</td>
</tr>
<tr>
<td>Calacuccia</td>
<td>CC</td>
<td>206</td>
<td>65</td>
<td>812</td>
</tr>
<tr>
<td>Calenzana</td>
<td>CZ</td>
<td>702</td>
<td>379</td>
<td>273</td>
</tr>
<tr>
<td>Ghisoni</td>
<td>GH</td>
<td>229</td>
<td>100</td>
<td>635</td>
</tr>
<tr>
<td>Lecci</td>
<td>LC</td>
<td>467</td>
<td>402</td>
<td>80</td>
</tr>
<tr>
<td>Levie</td>
<td>LV</td>
<td>383</td>
<td>170</td>
<td>112</td>
</tr>
<tr>
<td>Luri</td>
<td>LR</td>
<td>332</td>
<td>193</td>
<td>112</td>
</tr>
<tr>
<td>Morosaglia</td>
<td>MR</td>
<td>175</td>
<td>117</td>
<td>805</td>
</tr>
<tr>
<td>Oletta</td>
<td>OL</td>
<td>346</td>
<td>232</td>
<td>235</td>
</tr>
<tr>
<td>Sartene</td>
<td>SR</td>
<td>1252</td>
<td>576</td>
<td>421</td>
</tr>
<tr>
<td>Venaco</td>
<td>VN</td>
<td>309</td>
<td>159</td>
<td>580</td>
</tr>
<tr>
<td>Vico</td>
<td>VC</td>
<td>385</td>
<td>225</td>
<td>491</td>
</tr>
</tbody>
</table>
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.

Fig. 1. Positions of studied villages. The continuous line indicates the administrative subdivision, while the dotted one indicates the main linguistic division.
The island of Corsica (western Mediterranean) has a population of about 250,000 in a territory of 8680 km². 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,
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.
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.

![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.](image)
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).

**Fig. 4.** Spatial distribution of the first and second principal components. The percentages of variance are given in parentheses.
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.
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).
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 ($r=0.61$, $p<0.05$) 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 Sartène, 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).

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

<table>
<thead>
<tr>
<th>Matrices compared</th>
<th>Correlation ($r$)</th>
<th>Significance $p$ (Mantel’s test)</th>
<th>Proportion of the variance explained (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIN × GEO</td>
<td>0.5940</td>
<td>0.0001</td>
<td>35.3</td>
</tr>
<tr>
<td>KIN × LIN</td>
<td>0.5848</td>
<td>0.0001</td>
<td>34.2</td>
</tr>
<tr>
<td>GEO × LIN</td>
<td>0.5623</td>
<td>0.0000</td>
<td>31.6</td>
</tr>
<tr>
<td>Partial correlations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIN × GEO (LIN)</td>
<td>0.3150</td>
<td>0.0013</td>
<td>9.9</td>
</tr>
<tr>
<td>KIN × LIN (GEO)</td>
<td>0.2470</td>
<td>0.0037</td>
<td>6.1</td>
</tr>
</tbody>
</table>
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
grants from the Italian National Council of Researches CNR 97.00712PF36 (L. Terrenato) and M U R S T 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


