Ann. Hum. Genet. (1998), **62**, 123–132
Printed in Great Britain

HLA evidence for the lack of genetic heterogeneity in Basques

D. COMAS, F. CALAFELL*, E. MATEU, A. PÉREZ-LEZAUN AND J. BERTRANPETIT†
Unitat d'Antropologia, Departament de Biologia Animal, Facultat de Biologia, Universitat de Barcelona, Avinguda Diagonal 645, 08028 Barcelona, Spain

(Received 3.10.97. Accepted 6.3.98)

SUMMARY

To examine the possible internal heterogeneity within the Basque population, nine samples typed for several HLA loci were compiled and HLA-A, B, C and DR loci were analysed. First, the shared features of HLA in Basques were analysed by principal component analysis and genetic distances. Two major Basque dialect groups ('French' and 'Spanish') were considered. F_{ST} statistics were computed and corrected for sampling intensity. The dialectal and political division did not seem to differentiate these two groups genetically. Analysis of Molecular Variance also failed to show consistently significant genetic variance components between French and Spanish Basques. Thus, in this particular example, linguistic diversity does not seem to correlate with a genetic stratification.

INTRODUCTION

The genetic uniqueness of the Basque population has long been recognized (Boyd & Boyd, 1937; Mourant, 1947, 1983). The Basque language is also unique, with no clear relationship to any other language or linguistic family (Ruhlen, 1991), and with a profound internal subdivision into seven major dialects (de Yrizar, 1981), grouped in the northern and southern dialects, which are spoken in the French and Spanish Basque Country, respectively. The correlation between genetic and linguistic differentiation has been observed at a global scale (Cavalli-Sforza *et al.* 1988), within continents (Barbujani & Sokal, 1990) and within isolated regions (Cappello *et al.* 1996; Stenico *et al.* 1996). Restricted migration across linguistic and cultural boundaries can preserve the genetic differentiation generated by random drift in small populations (Cavalli-Sforza *et al.* 1994). This can also be the case for the observed genetic differentiation in the Basques: the genetic differentiation created by drift before the Neolithic (probably in the Upper Paleolithic) might have been spared by the Neolithic wave of advance (Ammerman & Cavalli-Sforza, 1984) for ecological reasons (Calafell & Bertranpetit, 1993) and preserved until recent times because of the cultural and linguistic distinctiveness of the Basques (Bertranpetit & Cavalli-Sforza, 1991; Calafell & Bertranpetit, 1994; Bertranpetit *et al.* 1995). Hypothetically, that mechanism could also have operated to maintain genetic differentiation among Basque subpopulations speaking different dialects.

Several research groups have produced in the last two decades a wealth of genetic data, mainly on the so-called classical polymorphisms (i.e. blood groups, HLA antigens and protein electromorphs), at a very detailed geographical level. Populations from small regions or even single valleys were sampled and analysed separately. The results of those microgeographic studies were compared to results from external populations which often covered entire countries. Thus, Aguirre *et al.* (1991) showed that the levels

* Present address: Department of Genetics, School of Medicine, Yale University, USA.

† Correspondence: Unitat d'Antropologia, Departament de Biologia Animal, Facultat de Biologia, Universitat de Barcelona, Avinguda Diagonal 645, 08028 Barcelona, Spain. Tel.: +34-93-4021461; Fax: +34-93-4110887.

E-mail: jaumeb@porthos.bio.ub.es

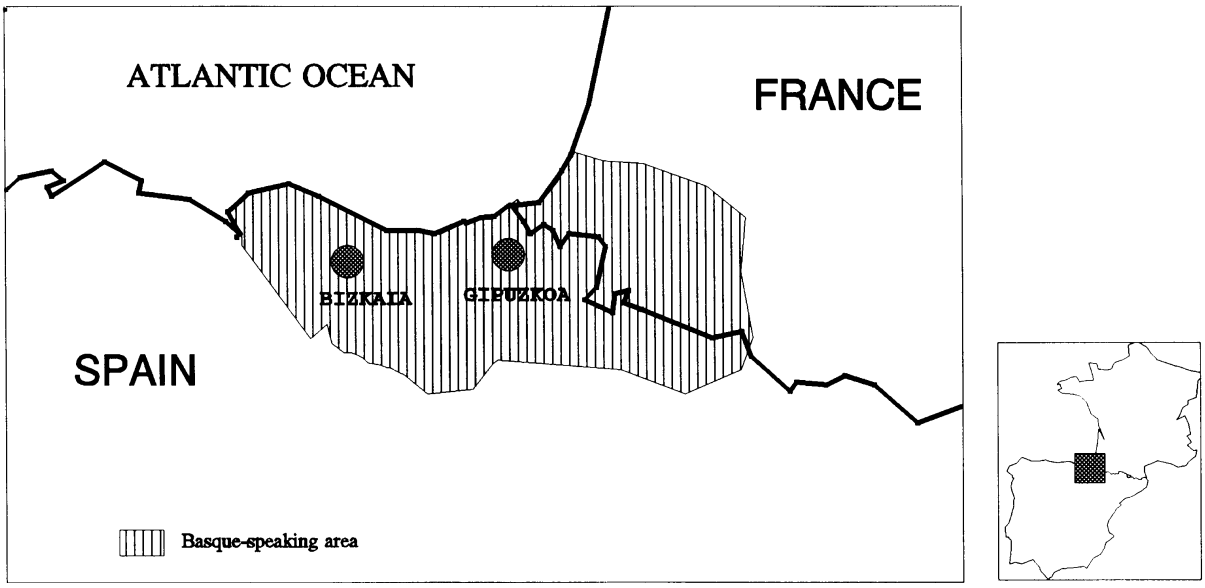

Fig. 1. Map of the Basque-speaking areas.

of genetic differentiation between seven small regions in the Basque province of Bizkaia were similar to, or even higher than, those found between European countries. More recently, Manzano *et al.* (1996) typed nine classical polymorphisms in the Araba province, which is geographically related to the Ebro valley region inhabited mainly by non-Basque Spanish speakers. They concluded that the main genetic differentiation in the Basque population was between 'Atlantic' and 'Mediterranean' Basques, and suggested that this was due to differential levels of external admixture.

However, Cavalli-Sforza & Feldman (1990) showed that genetic differentiation, as measured by F_{ST}, and under a stepping-stone model of migration, is a function of the degree of subdivision of the populations being compared. Thus, F_{ST} increases when the populations being compared are split into smaller units; conversely, F_{ST} decreases when the units being compared comprise distinct subpopulations. Calafell & Bertranpetit (1994) compiled data for 31 classical markers (62 independent alleles) and, with correction for sampling intensity, found that the genetic differentiation within the Basque Country was less than that observed in the non-Basque populations of the Iberian Peninsula.

The degree of differentiation among Basque populations, as explained above, is a debated subject that could illustrate the more general interplay of genetic, linguistic and cultural diversities. In order to address this issue, we have compiled literature and our own data from Basque samples on the most polymorphic nuclear DNA region commonly analysed, HLA. First, we described the shared features of HLA in Basques, shown by means of principal component analysis and genetic trees. Next, we investigated the Basque internal genetic subdivision by computing F_{ST} statistics (Wright, 1951) and AMOVA tests (Excoffier *et al.* 1992). We divided the samples according to the first linguistic split: between those speaking the Bizkaian, Gipuzkoan, Araban and High Nafarroan Basque dialects (i.e. the Basques residing in the Spanish Basque Country, who we labelled 'Spanish Basques') and those speaking the Lapurdean, Zuberoan and Low Nafarroan dialects ('French Basques' residing in the French Basque country). It should be stressed that 'Spanish' and 'French' are, in fact, recent political labels, and that Basques on both sides of the border speak primarily Basque (Fig. 1). The political border between France and Spain is superimposed on the dialectal division. The present line of the border was established in

1659, and its effects on migration, and ultimately in preserving genetic differences, are expected to be weaker than those of the much older linguistic division; in other parts of the Pyrenees it has been demonstrated that the real separator effect of the political boundary between France and Spain has been reinforced only in very recent times (Salvat *et al.* 1997).

MATERIAL AND METHODS

All the available data on HLA typing in the Basque Country was compiled. Given the large number of non-Basque immigrants living in the Basque Country, we selected for further analysis those studies based exclusively on autochthonous samples. Autochthony has been established through place of birth (of subject and ancestors), surnames (which are often remembered back to the third generation) and language. The origin of the samples corresponds to either the French or Spanish parts of the Basque Country, or to smaller regions, such as the provinces of Bizkaia or Gipuzkoa. The samples analysed in the present study are shown in Table 1; a total of nine samples from eight references were used. The original studies span more than 20 years, in which techniques have evolved and new loci have been discovered; therefore, the loci typed range from HLA-A and HLA-B with a few specificities, to HLA-A, B, C, DR, DQ and DP typed at the DNA level according to the 12th International Histocompatibility Workshop protocols. In order to analyse homogeneously the information from the different studies and render them comparable, we pooled some splits into their original specificities, such as HLA-A*23 and HLA-A*24 into HLA-A9, and the DNA-typed HLA-DRB1 sequences into their serologic HLA-DR equivalents. In some studies, only phenotype frequencies were given, and those were converted into allele frequencies using the equation $g = 1-\sqrt{(1-p)}$, where p stands for phenotype frequency and g for allele frequency. Our analysis comprises nine HLA-typed Basque samples, five from Spanish Basques and four from French Basques, and takes into account four HLA loci and a total of 41 different alleles (11 alleles for HLA-A, 17 alleles for HLA-B, 5 alleles for HLA-C and 8 alleles for HLA-DR). Allele frequencies were compared between Basque samples, and between Basques and other European populations.

Principal component (PC) analysis was performed on the correlation matrix of the allelic frequencies. Only HLA-A, B, C and DR alleles typed in at least five different samples were taken into account. The allele frequencies of the few loci not typed in a sample were replaced with the mean allele frequency in Basques. That is the case of HLA-C in three populations (FRE1, FRE2 and SPA1) and HLA-DR in four (FRE1, FRE2, BIZ and GIP1) of the populations analysed. The bias introduced by this procedure will be discussed. Several European populations collected for the 11th International Histocompatibility Workshop were entered into the PC analysis for comparison.

F_{ST}-related genetic distances were computed between pairs of populations according to Reynolds *et al.* (1993). A neighbour-joining tree was built from the genetic distance matrix by means of the PHYLIP 3.5c package (Felsenstein, 1989), and tree robustness was assessed through 1000 bootstrap iterations (Felsenstein, 1985).

In order to estimate inter-population differentiation, F_{ST} statistics (Wright, 1951) were estimated as $F_{ST} = V/(p(1-p))$, where V is the variance of the frequency of an allele across populations and p is the mean allele frequency. F_{ST} was computed for HLA-A, B, C and DR alleles typed in at least five different samples. Genetic variance in Basques was hierarchically apportioned through the analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992), performed by the Arlequin package (Schneider *et al.* 1995). Each allele of each locus was compared with the other alleles of the same sample, with the alleles of the other samples within the same group (Spanish or French Basques) and with all the alleles from the rest of the samples. The significance of the estimated variance components was tested by 1000 bootstrap iterations, in each of which individuals were randomly

assigned to the populations and the variance recalculated, in order to obtain an empirical distribution under the null hypothesis (i.e. absence of genetic differentiation).

RESULTS

The sources and abbreviations for the nine Basque samples considered in this paper are listed in Table 1. Class II loci HLA-DQ and DP have been typed in very few Basque samples (Table 1) and excluded from further analyses, which are centered on HLA-A, B, C and DR. There are some specific features in the frequency of the alleles of the HLA system in the Basque samples that distinguish them from the rest of the European populations. Basques are characterised by low values of A28 (lower than 2%), save for the two populations from the study by Calderón *et al.* (1993) (BIZ, 5%; GIP1, 4%). However, A30/A31 are found at very high frequencies, around 10% (save for BIZ, 5%; GIP1, 3%), surpassed in Europe only by Sardinians (Tsuji *et al.* 1992; Cavalli-Sforza *et al.* 1994). Another distinctive Basque feature is their high frequency of A29, above 10%, the highest found in Europe (Tsuji *et al.* 1992); but Calderón *et al.* (1993) (BIZ, 7%; GIP1, 6%) and Cambon-de Mouzon *et al.* (1982) (FRE3, 5.13%) found surprisingly low frequencies of this allele. The highest worldwide frequency of the B12 allele (B44 and B45 combined) is found in Basques (Tsuji *et al.* 1992; Cavalli-Sforza *et al.* 1994). This high frequency is basically due to the B44 allele because B45 has low frequencies (around 1%). The mean frequency of B12 in the present samples is 22.05%; all of them have frequencies above 20%, except GIP1 (13%). B18 is also found at high frequencies in Basques; and again Sardinians are the only European population with higher frequencies (Tsuji *et al.* 1992). On the contrary, B22, B37 and B41 are almost absent in Basques. The most frequent HLA-C allele in Basques is Cw7, but unfortunately it only has been typed in a few Basque samples and therefore it has not been included in our study. Other characteristic features in Basques are the low frequency of the Cw2 allele and the high frequency of Cw5. Basques have the highest worldwide frequencies of DR7 (Tsuji *et al.* 1992), and the second highest frequency of DR3 in Europe after Sardinians (Tsuji *et al.* 1992), whereas DR5 is found at much lower frequencies than in other European populations. The exceptionally high frequencies of some HLA alleles are a result of the high frequencies of some common haplotypes found in Basques, such as A29-Cw16-B44-DR7 and A30-Cw5-B18-DR3 (Tsuji *et al.* 1992; Martínez-Laso *et al.* 1995; Comas *et al.* 1998, and references therein).

Principal component analysis (PC) was performed to characterise the extent of the differences in HLA allele frequencies between the Basques and other European populations (data from Tsuji *et al.* 1992). A total of 37 independent alleles for a set of 11 European populations and nine Basque samples were used in the PC analysis. The two-dimensional plot of the first two PC axes, which account for 47.2% of the variance observed, is shown in Fig. 2. Three major clusters are defined by the first two principal components: Basques, Sardinians and the rest of the European populations. The first principal component, which encompasses 29.6% of the total variance, separates the Basques from the other European populations. The singularity of this outlier population, confirmed by other classical markers, is defined (with an absolute correlation higher than 0.7) by high frequencies of A29, B12, Cw5, DR3 and DR7 and low frequencies of A28, B37, B41, B22, Cw2, DR5 and DR8. Two Basque samples (BIZ and GIP1) have first PC scores intermediate between those of the other Basque samples and the European populations. Moreover, allele frequencies for the HLA-DR locus were not available for those samples, and were filled in with the average frequency of the other Basque samples; this procedure could have resulted in BIZ and GIP1 being close to the centroid of the Basque samples in the PC graph, and, therefore, makes their outlier location even more surprising. The second principal component, which explains 17.6% of the total variance, separates Sardinians, another

Table 1. *Samples used in the present study*

Sample	N[a]	Loci
French Basques (FRE1) (Dausset *et al.* 1972)	88	A(13), B(14)
French Basques (FRE2) (de Mouzon *et al.* 1980)	198	A(10), B(15)
Spanish Basques (SPA1) (García de Masdevall *et al.* 1980)	90	A(15), B(18), DR(8)
French Basques (FRE3) (Cambon-de Mouzon *et al.* 1982)	50	A(13), B(16), C(6), DR(7)
French Basques (FRE4) (Tsuji *et al.* 1992)	60	A(13), B(21), C(8), DR(12), DQ(5)
Bizkaia (BIZ) (Calderón *et al.* 1993)	110	A(13), B(19), C(5)
Gipuzkoa (GIP1) (Calderón *et al.* 1993)	50	A(13), B(18), C(5)
Spanish Basques (SPA2) (Martínez-Laso *et al.* 1995)	82	A(14), B(20), C(6), DRB1(17), DRB3(3), DRB4(1), DQA(7), DQB(13)
Gipuzkoa (GIP2) (Comas *et al.* 1998)	100	A(16), B(20), C(16), DRB1(14), DQA(7), DPA(3), DPB(15)

[a] N, number of individuals analysed in each study. Numbers in parentheses are the number of different alleles given for each locus with an allele frequency higher than zero.

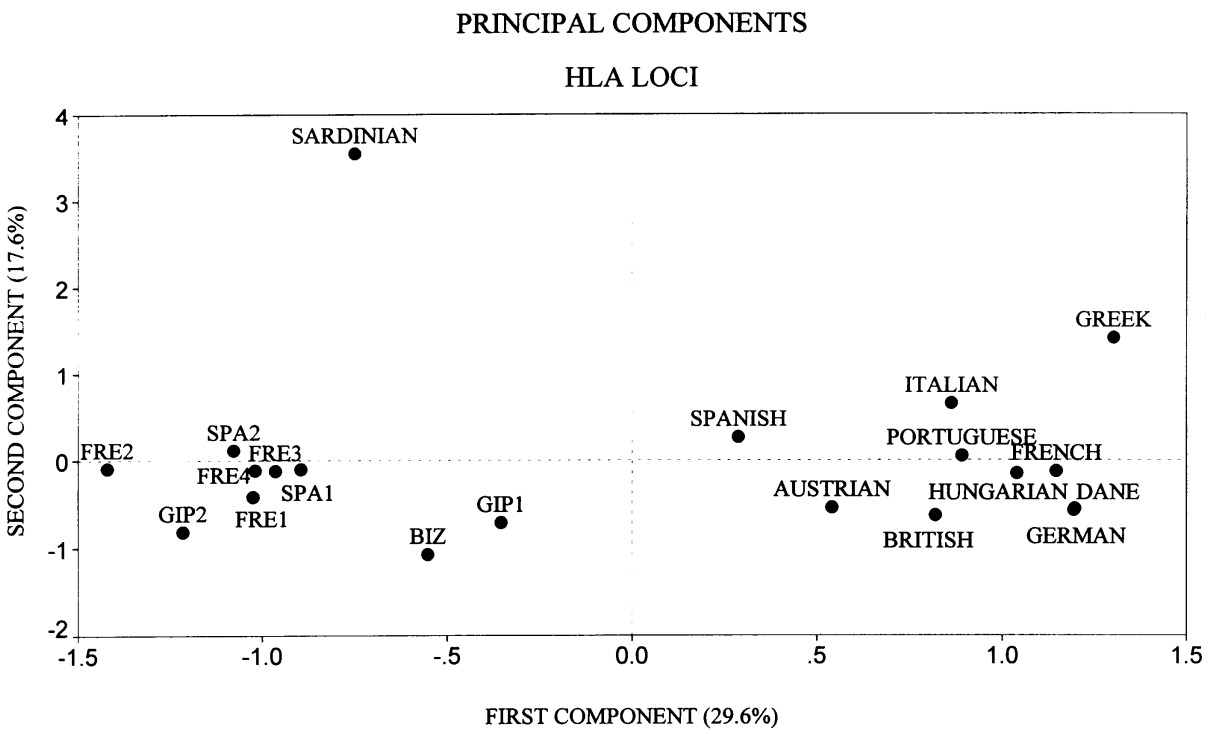

Fig. 2. Two-dimensional plot of the two first principal components (PC) axes in nine different Basque samples and some European populations for four HLA loci. The first axis encompasses 29.6% of the variance observed, and the second axis encompasses 17.6%.

outlier population in Europe, from other European populations. The same PC approach was performed, considering only those four Basque samples (FRE3, FRE4, SPA2, and GIP2) for which the four HLA loci were typed, i.e. no mean values were added, and an identical two-dimensional plot of the first two PC axes was found (not shown).

Genetic distances were computed among Basques and other European populations (11th

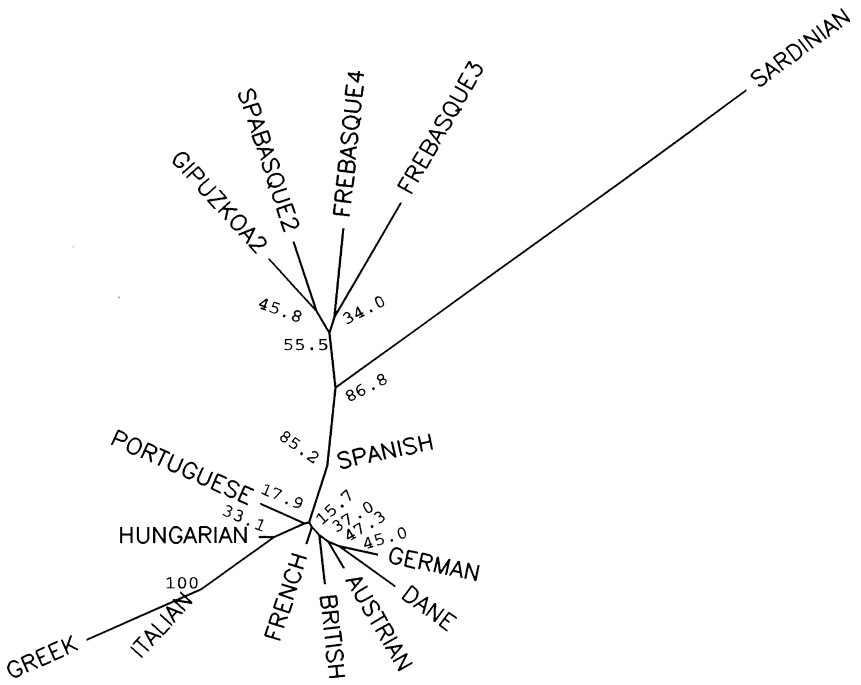

Fig. 3. Neighbour-joining tree linking four Basque samples and some European populations. Numbers in the nodes represent the bootstrap suports after 1000 iterations.

IHWC) by using allelic frequencies from the A, B, C and DR loci. Unlike PC analysis, the genetic distance used assumes a specific genetic process of differentiation among populations. In this case, we took the restrictive approach of allowing into the analysis only the four studies (FRE3, FRE4, SPA2 and GIP2) where all four loci were typed. Basque samples showed their shortest distances to each other. The populations closest to them were the Spanish, although the Spanish had shorter distances to other European populations (except for Sardinians and Greeks) than to the Basques. The neighbour-joining tree (Fig. 3) constructed from this distance matrix has three separate clusters: Sardinians, Basques and the rest of the European populations. The robustness of the tree was assessed by 1000 bootstrap iterations; in 55.5% of those, all four Basque samples clustered together, and in 86.8% they clustered with Sardinians. Within the Basque samples, genetic distance was not correlated to geographic distance ($r = 0.096$, $p = 0.287$, Mantel test with 10000 iterations).

F_{ST} values for the different alleles of the loci analysed in the Basque samples are shown in Table 2. Mean F_{ST} for the HLA-A, B, C and DR loci is, respectively, 0.0119, 0.0108, 0.0197 and 0.0133. The average F_{ST} for those four HLA loci is 0.0139. In a previous study (Calafell & Bertranpetit, 1994) the mean value of F_{ST} for 26 different classical polymorphic systems (excluding HLA) in Basques was 0.0098. F_{ST} for HLA appears to be slightly higher than for other classical markers ($p = 0.044$, Mann–Whitney's U test), but such comparison can be heavily biased by the different number and origin of samples in both datasets. The highest mean F_{ST} value is presented by the HLA-C locus. Of all the loci considered, HLA-C has the highest frequency of 'blank' alleles, even though the detection techniques have greatly improved: the number of detectable alleles with DNA techniques is increasing, reaching 42 alleles in a recent review (Bodmer *et al.* 1997), whereas there were only six 15 years ago. The resolution with which HLA-C was typed is certainly not the same across all the references we analysed, but the F_{ST} value we found for HLA-C was based on the five alleles that were typed on all samples, and the frequency of which does not appear to depend on technical reasons. The differences in F_{ST} across loci are not statistically significant ($p = 0.331$, Kruskal–

Table 2. *HLA alleles analysed with their means, minimum and maximum values, and* F_{ST}[a]

Allele	Mean	Min	Max	F_{ST}
A1	0.1284	0.1017	0.1800	0.0042
A2	0.2672	0.2460	0.2929	0.0012
A3	0.1133	0.0700	0.2000	0.0178
A9 (A23, A24)	0.1017	0.0355	0.1500	0.0114
A10 (A25, A26)	0.0461	0.0200	0.0900	0.0170
A11	0.0607	0.0300	0.1056	0.0084
A28	0.0178	0.0000	0.0500	0.0154
A29	0.1022	0.0513	0.1560	0.0134
A30/A31	0.1050	0.0300	0.1800	0.0263
A32	0.0163	0.0050	0.0513	0.0144
A33	0.0082	0.0050	0.0110	0.0009
B5 (B51, B52)	0.0843	0.0556	0.1200	0.0069
B7	0.1163	0.0920	0.1400	0.0039
B8	0.0852	0.0305	0.1600	0.0204
B12 (B44, B45)	0.2205	0.1300	0.2634	0.0085
B13	0.0089	0.0000	0.0250	0.0062
B14	0.0249	0.0100	0.0410	0.0039
B15	0.0365	0.0045	0.0750	0.0172
B16 (B38, B39)	0.0315	0.0050	0.0682	0.0232
B17 (B57, B58)	0.0624	0.0100	0.1282	0.0263
B18	0.1107	0.0673	0.1770	0.0149
B21 (B49,B50)	0.0375	0.0151	0.0710	0.0125
B22 (B54, B55, B56)	0.0107	0.0020	0.0253	0.0060
B27	0.0369	0.0100	0.0670	0.0089
B35	0.0509	0.0356	0.0620	0.0020
B37	0.0026	0.0000	0.0160	0.0139
B40 (B60, B61)	0.0337	0.0080	0.0505	0.0052
B41	0.0023	0.0000	0.0060	0.0036
Cw1	0.0335	0.0000	0.0590	0.0169
Cw2	0.0240	0.0051	0.0430	0.0077
Cw3	0.0761	0.0300	0.1258	0.0141
Cw4	0.1132	0.0356	0.1900	0.0327
Cw5	0.1148	0.0680	0.1876	0.0271
DR1	0.1157	0.0730	0.1515	0.0085
DR2	0.1324	0.1050	0.1754	0.0076
DR3	0.1919	0.1282	0.2260	0.0096
DR4	0.0770	0.0714	0.0835	0.0003
DR5	0.0503	0.0000	0.0970	0.0256
DR6	0.0906	0.0410	0.1713	0.0343
DR7	0.2621	0.1938	0.3367	0.0153
DR8	0.0234	0.0060	0.0350	0.0055

[a] For HLA-A and B, nine Basque populations have been considered. For HLA-C, six populations and for HLA-DR five Basque populations have been considered.

Wallis' H test), even when only the two extreme values (HLA-B and HLA-C) are compared ($p = 0.055$, Mann–Whitney's U test).

There are some specific alleles that present high F_{ST} values. For the mentioned HLA-C locus, the high F_{ST} value is mainly due to the divergences in frequencies in alleles Cw4 and Cw5. A30 and A31 (which were considered together in the present work) have an F_{ST} value of 0.263 as a result of an extremely low frequency in two samples of the same study (GIP1 and BIZ) that is not found in other samples coming from the same areas. A similar pattern is shown by the allele B8 in these two populations, with very high values. Other alleles that show high F_{ST} values are B16 (B38 and B39), B17 (B57 and B58), DR6 and DR5.

The hierarchical apportioning of the genetic variance for HLA in Basques was investigated through the AMOVA test, which was performed separately for the four HLA loci analysed. As some of the loci were not determined in all the populations, in some loci the AMOVA analysis was performed with a reduced number of samples. That is the case of HLA-C, only typed in two French samples (FRE3 and FRE4) and four Spanish samples (BIZ, GIP1, SPA2 and GIP2); and the locus HLA-DR, typed in two French samples (FRE3 and FRE4) and three Spanish samples (SPA1, SPA2 and GIP2). The variance partition of the Basques samples is shown in Table 3*a*. Most of the variation detected is due to differences within the samples (an average of 98.38%), whereas the proportion of the variance due to differences among Basque samples is very low (1.62% on average), and only one of the HLA loci (HLA-C) shows significantly different values among samples, probably due to spurious reasons as discussed below.

As an approach to the hierarchical structure, samples were split into French Basques (four samples) and Spanish Basques (five samples), as described above. We have, thus, a three-level situation, in which there are a high number of individuals (a total of 828) from nine samples belonging to two different regions. Variance is partitioned within samples, among samples within the two groups, and among the two groups. The percentages of the variance components and their significance are shown in Table 3*b*.

Most of the variation detected is due to differences within the samples, which account for an average of 98.53% of the total variance of the HLA system, whereas only 1.72% and −0.25% of the variance is the result of the variation

Table 3. *Analysis of molecular variance* (*AMOVA*) *for four HLA loci in Basque samples.* (*a*) *All the Basque samples considered together,* (*b*) *Basque samples divided in two groups*: *Spanish and French Basques*

(*a*)

	Variance components % (p)	
	Within samples	Among samples
HLA-A	99.95	0.05 ($p = 0.388$)
HLA-B	100.02	−0.02 ($p = 0.509$)
HLA-C	93.69	6.31 ($p < 0.001$)*
HLA-DR	99.84	0.16 ($p = 0.325$)
Mean	98.38	1.62

(*b*)

	Variance components % (p)		
	Within samples	Among samples within groups	Among groups (Spanish Basques-French Basques)
HLA-A	99.85 ($p = 0.070$)	−0.25 ($p = 0.716$)	0.40 ($p = 0.511$)
HLA-B	99.68 ($p = 0.369$)	−0.31 ($p = 0.277$)	0.63 ($p = 0.012$)*
HLA-C	94.87 ($p < 0.001$)*	7.40 ($p = 0.026$)*	−2.27 ($p = 0.784$)
HLA-DR	99.72 ($p = 0.301$)	0.03 ($p = 0.900$)	0.25 ($p = 0.320$)
Mean	98.53	1.72	−0.25

* Significant values, $p < 0.05$.

within groups and among groups, respectively (negative variance components should be regarded as zero). For the HLA-A locus, although French Basques show higher frequencies of A30/A31 and lower frequencies of A28 than Spanish Basques, no significant differences were found either among the samples within each group or between French Basques and Spanish Basques. For the HLA-B locus, significant differences were found between French Basques and Spanish Basques, probably because the first group has higher frequencies of B18 and B17 and lower frequencies of B8 than the latter. The differences in the HLA-C locus are extraordinary: 7.4% of the total genetic variance was found within populations of the same group, while Barbujani *et al.* (1997), who analysed groups from all over the world, found that genetic variance within populations of the same continent was between 3.9% (79 RFLP loci) and 5.5% (30 microsatellites). Given their anomalous PC scores, we suspected that the GIP1 and BIZ samples (Calderón *et al.* 1993) could be responsible for that high genetic variance. To test that hypothesis, we divided the samples in two groups: one with GIP1 and BIZ and the other with the remaining populations. With this new partition, the variance within groups dropped to 1.85% ($p < 0.001$), while the variance between groups became 7.6% ($p < 0.001$), i.e. three-quarters of the mean variance between continents (10.8%, Barbujani *et al.* 1997). Clearly, GIP1 and BIZ have outstanding allele frequencies at HLA-C, even though GIP2 and SPA2 were sampled in the same province as GIP1. Excluding GIP1 and BIZ from the analysis, 2.21% ($p = 0.0182$) of the genetic variance for HLA-C was found within the French or Spanish Basques and only 0.08% ($p = 0.670$) of the genetic variance was found between French and Spanish Basques. Thus, with the exception of HLA-B, the political and linguistic division between French and Spanish Basques has not led to a significant amount of genetic variance between the two groups.

Discussion

The detection of genetic heterogeneity within a given population is not a simple endeavour.

The descriptive analysis of data, like principal component analysis or genetic distances, does not give a straightforward response unless the appropriate samples, in terms of the size and clustering of smaller populations units, are used. Such reference populations are not usually as readily available as country-wide samples are. The pitfall of many studies analysing a new, small population is that, in order to understand its genetic composition, authors compare it with much wider and more general populations. Thus, genetic trees or principal component plots should be interpreted with extreme caution if heterogeneous populations (in terms of size or dispersion) are analysed at the same time or if several samples from the same population are included. Other factors may further confuse attempts to ascertain genetic heterogeneity: technical differences, particularly in a field such as HLA that has been continuously evolving for the last 20 years, may hinder direct comparison among samples; the same effect can be produced by different sampling strategies, i.e. the extent of the area and the choice of individuals being sampled.

HLA is a genetic system with such a great amount of genetic variation that makes it especially useful in the detection of micro-geographic differentiation. Haplotype frequencies are highly variable among populations (Tsuji *et al*. 1992 for a global overview) even at a low geographic scale; thus, for instance, the genetic divergence of Sardinians from other European populations is clearly ascertained through the HLA analysis.

The AMOVA analysis seems the most appropriate for recognising the existence of stratification and giving a real idea of statistical significance. It is a non-parametric, model-free procedure of estimating the significance of the genetic variance components in a hierarchy of samples. With only one exception, we have found that genetic variance at HLA loci in Basques is not significantly different from zero between samples, and between the Spanish and French Basques.

Besides ascertaining genetic heterogeneity, we can try to discern how genetic heterogeneity is spatially organized. A spatial pattern that is coherent with external evidence (e.g. cultural differences or geographical obstacles) is slightly more likely than non-patterned genetic heterogeneity. In this case, we have found that genetic distance is not correlated to geographic distance, in concordance with what might be expected within a single and panmictic population. A main linguistic and political barrier, the one separating Spanish and French Basques, does not seem to have contributed significantly to the genetic differentiation between the two groups. Thus, there does not seem to be a discernible pattern of genetic heterogeneity among the available Basque samples for the highly polymorphic HLA system.

This research was supported by Dirección General de Investigación Científica y Técnica (Spain) grant PB95-0267-C02-01, by the Direcció General de Recerca, Generalitat de Catalunya (1995SGR00205 and 1996SGR00041) and by the Institut d'Estudis Catalans. The study was also made possible by the fellowships to D.C. (FI/93-1151; Comissionat per a Universitats i Recerca, Catalan Autonomous Government), to E.M. (Vicerectorat de Recerca, Universitat de Barcelona) and to A.P.-L. (FP93-38110903; Spanish Ministry of Education and Science). F.C. is the recipient of a postdoctoral fellowship from CIRIT, Catalan Autonomous Government.

REFERENCES

Aguirre, A., Vicario, A., Mazón, L. I., Estomba, A., Martínez de Pancorbo, M., Arrieta Picó, V., Pérez Elortondo, F. & Lostao, C. M. (1991). Are the Basques a single and a unique population? *Am. J. Hum. Genet.* **49**, 450–458.

Ammerman, A. J. & Cavalli-Sforza, L. L. (1984). *The Neolithic transition and the genetics of populations in Europe*. Princeton, NJ: Princeton University Press.

Barbujani, G., Magagni, A., Minch, E. & Cavalli-Sforza, L. L. (1997). An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. USA* **94**, 4516–4519.

Barbujani, G. & Sokal, R. R. (1990). Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl. Acad. Sci. USA* **87**, 1816–1819.

Bertranpetit, J. & Cavalli-Sforza, L. L. (1991). A genetic reconstruction of the history of the Iberian Peninsula. *Ann. Hum. Genet.* **55**, 51–67.

Bertranpetit, J., Sala, J., Calafell, F., Underhill, P. A., Moral, P. & Comas, D. (1995). Human mitochondrial DNA variation and the origin of the Basques. *Ann. Hum. Genet.* **59**, 63–81.

Bodmer, J. G., Marsh, S. G. E., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Charron, D., Dupont, B., Erlich, H. A., Fauchet, R., Mach, B., Mayr, W. R.,

Parham, P., Sasazuki, T., Schreuder, G. M. T., Strominger, J. L., Svejgaard, A. & Terasaki, P. I. (1997). Nomenclature for factors of the HLA system, 1996. *Tissue Antigens* **49**, 297–321.

Boyd, W. C. & Boyd, L. G. (1937). New data on blood groups and other inhereted factors in Europe and Egypt. *Am. J. Phys. Anthropol.* **23**, 49–70.

Calafell, F. & Bertranpetit, J. (1993). The genetic history of the Iberian Peninsula: a simulation. *Current Anthropology* **34**, 735–745.

Calafell, F. & Bertranpetit, J. (1994). Principal component analysis of gene frequencies and the origin of Basques. *Am. J. Phys. Anthropol.* **93**, 201–215.

Calderón, R., Wentzel, J. & Roberts, D. F. (1993). HLA frequencies in Basques in Spain and in neighbouring populations. *Ann. Hum. Biol.* **20**, 109–120.

Cambon-de Mouzon, A., Ohayon, E., Hauptmann, G., Sevin, A., Abbal, M., Sommer, E., Vergnes, H. & Ducos, J. (1982). HLA-A, B, C, DR antigens, Bf, C4 and glyoxalase I (GLO) polymorphism in French Basques with insulin-dependent diabetes mellitus (IDDM). *Tissue Antigens* **19**, 366–379.

Cappello, N., Rendine, S., Griffo, R., Mameli, G. E., Succa, V., Vona, G. & Piazza, A. (1996). Genetic analysis of Sardinia: I. Data on 12 polymorphisms in 21 linguistic domains. *Ann. Hum. Genet.* **60**, 125–141.

Cavalli-Sforza, L. L. & Feldman, M. W. (1990). Spatial subdivision of populations and estimates of genetic variation. *Theor. Popul. Biol.* **37**, 3–25.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994). *The history and geography of human genes*. Princeton, NJ: Princeton University Press.

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. (1988). Reconstruction of human evolution; bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. USA* **85**, 6002–6006.

Comas, D., MAteu, E., Calafell, F., Pérez-Lezaun, A., Bosch, E., Martínez-Arias, R. & Bertranpetit, J. (1998). HLA class I and class II DNA typing and the origin of Basques. *Tissue Antigens* **51**, 30–40.

Dausset, J., Legrand, L., Levine, M. H., Quilici, J. C., Colombani, M. & Ruffié, J. (1972) Genetic structure and distribution of HL-A antigens in a Basque village. In *Histocompatibility testing* (eds. J. Dausset & J. Colombani J), pp 99–105. Copenhagen: Munksgaard.

Excoffier, L., Smouse, P. E. & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

Felsenstein, J. (1989). PHYLIP- phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166.

García de Masdevall, M. D., Ercilla, M. G., Arrieta, A. & Vives, J. (1982). Estudio genético del sistema HLA en la poblacion vasca. *Sangre* **27**, 182–189.

Manzano, C., Orue, J. M. & de la Rúa, C. (1996). The 'Basqueness' of the Basques of Alava: a reappraisal from a multidisciplinary perspective. *Am. J. Phys. Anthropol.* **99**, 249–258.

Martínez-Laso, J., de Juan, D., Martínez-Quilez, N., Gómez-Casado, E., Cuadrado, E. & Arnaiz-Villena, A. (1995). The contribution of the HLA-A, -B, -C and -DR, -DQ DNA typing to the study of the origins of Spaniards and Basques. *Tissue Antigens* **45**, 237–245.

Mourant, A. E. (1947). The blood groups of the Basques. *Nature* **160**, 505.

Mourant, A. E. (1983). *Blood relations. Blood groups and Anthropology*. London, Oxford: University Press.

de Mouzon, A., Constans, J., Sommer, E., Sevin, A., Quilici, J. C., Fernet, P., Ducos, J. & Ohayon, E. (1982). HLA-A, B typing in Basque and other Pyrenean populations. *Tissue Antigens* **15**, 11–18.

Reynolds, J., Weir, B. S. & Cockerham, C. C. (1993). Estimation of the coancestry coefficient: basis for a short term genetic distance. *Genetics* **105**, 767–779.

Ruhlen, M. (1991). *A guide to the world's languages*. Stanford: Stanford University Press.

Salvat, M., Vigo, M., Macbeth, H. & Bertranpetit, J. (1997). Seasonality of marriages in Spanish and French parishes in the Cerdanya valley, Eastern Pyrenees. *J. Biosoc. Sci.* **29**, 51–62.

Schneider, S., Kueffer, J. M., Roessli, D. & Excoffier, L. (1996). *Arlequin (ver. 1.0): a software enviroment for the analysis of population genetics data*. Geneva: Genetics and Biometry Lab.

Stenico, M., Nigro, L., Bertorelle, G., Calafell, F., Capitanio, M. A., Corrain, C. & Barbujani, G. (1996). High mitochondrial sequence diversity in two linguistic isolates of the Alps. *Am. J. Hum. Genet.* **59**, 1363–1375.

Tsuji, K., Aizawa, M. & Sasazuki, T. (1992). *HLA 1991*. Oxford: Oxford University Press.

de Yrizar, P. (1981). *Contribución a la dialectología de la lengua vasca*. Donostia: Caja de Ahorros Provincial de Guipúzcoa.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.* **15**, 223–354.