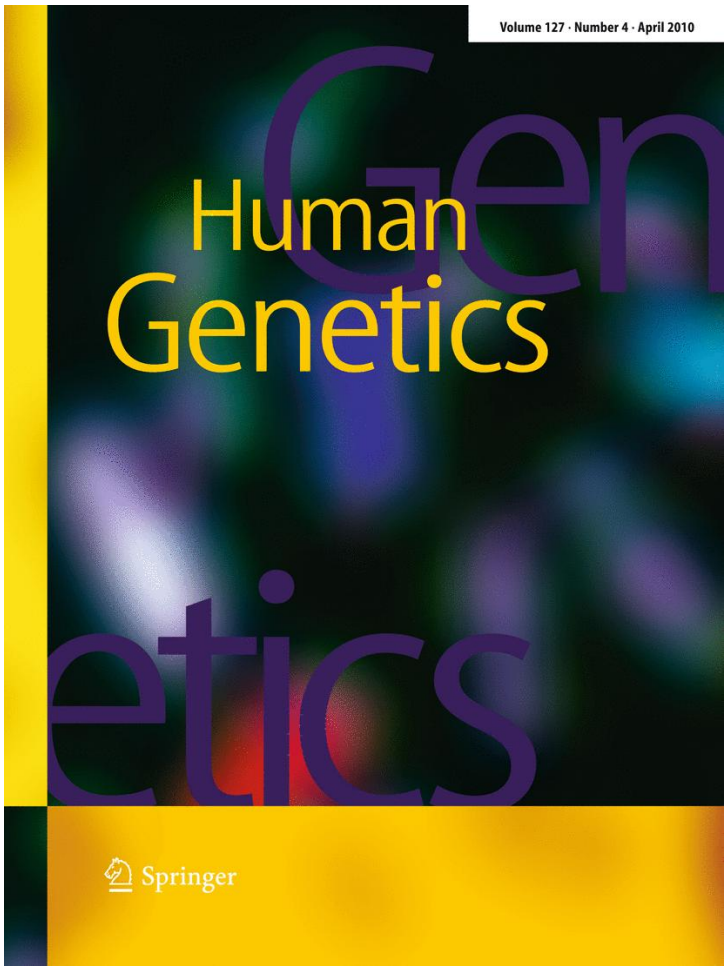
ISSN 0340-6717, Volume 127, Number 4



Hum Genet (2010) 127:455–458
DOI 10.1007/s00439-010-0798-3

SHORT REPORT

A genome-wide survey does not show the genetic distinctiveness of Basques

Hafid Laayouni · Francesc Calafell · Jaume Bertranpetit

Received: 27 October 2009 / Accepted: 1 February 2010 / Published online: 16 February 2010

Abstract Basques are a cultural isolate, and, according to mainly allele frequencies of classical polymorphisms, also a genetic isolate. We investigated the differentiation of Spanish Basques from the rest of Iberian populations by means of a dense, genome-wide SNP array. We found that F_{ST} distances between Spanish Basques and other populations were similar to those between pairs of non-Basque populations. The same result is found in a PCA of individuals, showing a general distinction between Iberians and other South Europeans independently of being Basques. Pathogen-mediated natural selection may be responsible for the high differentiation previously reported for Basques at very specific genes such as *ABO*, *RH*, and *HLA*. Thus, Basques cannot be considered a genetic outlier under a general genome scope and interpretations on their origin may have to be revised.

Introduction

The genetic distinctiveness of Basques has been assumed since the classical seminal work of Mourant (Chalmers et al. 1949). When analyzing a large set of classical genetic markers (Bertranpetit and Cavalli-Sforza 1991; Calafell and Bertranpetit 1994a, b) their distinctiveness with surrounding populations was reported and they were shown as a main population outlier in Western Europe. Results using STR markers were not that clear as Basques were not found to be differentiated from their surrounding populations, even if they were interpreted in the same way (Zlojutro et al. 2006). With the advancement of mtDNA and Y-chromosome analysis, results showed a much lower (if any) differentiation of Basques in relation to neighboring populations. In the case of mtDNA (Bertranpetit et al. 1995; Salas et al. 1998), differences are small, even if some of them are quite conspicuous, such as a higher frequency of haplogroup H and a possible origin of haplogroup H3 (Achilli et al. 2004). On the Y-chromosome, the most interesting pattern is the lack of haplogroup E3b2, of North African origin (Adams et al. 2008). A recent analysis using a genome scan of 650,000 SNPs seemed to slightly differentiate French Basques from other Western European populations (Li et al. 2008); however, the worldwide samples analyzed in that publication do not allow to place the Basques in a narrower geographic and genetic context. Garagnani et al. (2009) analyzed 144 SNPs in Basque samples from France and Spain, as well as in samples from Northern and Southern Spain, and North Africa, and did not find genetic differences between Basques and non-Basques. To date, this is the largest set of presumably neutral DNA markers analyzed in Basques and in non-Basque surrounding populations.

We have undertaken a study of the genetic differentiation of populations within Spain, including Basques, based on an overall and ample view of the genome, in order to clarify the genetic position of the Basques in relation to surrounding populations. The study has a two-step procedure. We first genotyped, using the 300K Illumina array, pools of DNA of 30 individuals for each of the 10 populations, including some replicates (regions within Spain, not

Electronic supplementary material The online version of this article (doi:10.1007/s00439-010-0798-3) contains supplementary material, which is available to authorized users.

H. Laayouni (✉) · F. Calafell · J. Bertranpetit
IBE, Institute of Evolutionary Biology (UPF-CSIC),
CEXS-UPF-PRBB, Doctor Aiguader 88, 08003 Barcelona,
Catalonia, Spain
e-mail: hafid.laayouni@upf.edu

H. Laayouni · F. Calafell · J. Bertranpetit
Centro de Investigación Biomédica En Red,
Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

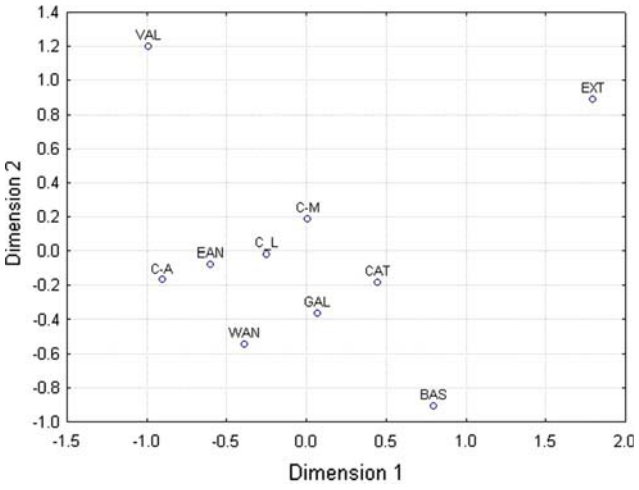


Fig. 1 Multidimensional scaling plots from F_{st} distance matrix of ten Spanish populations. Stress is smaller than 0.10

considering the Canary and Balearic Islands; see Fig. 1 in Supplementary information for more detail); a total of 280,862 SNPs passed quality control criteria. For the second step, we selected and genotyped 120 SNPs with the highest contribution to PC scores in the first step. These highly informative SNPs were selected using vectors (scores) for each factor (SNP) obtained from the principal component analysis (pcaMethods Package, implemented on the R program; Stacklies et al. 2007). The SNP contributions to the principal component factors were estimated using the single value determinant (SVD). All nine factors obtained were used as the amount of variation (r^2) explained by these factors ranging from 14% for the first PC to 6% for the ninth. The SNPlex^{TM} genotyping system was used and 238 individuals (all of them included in the first-step pools) were successfully genotyped for the 109 SNPs and passed the quality control criteria (genotyping failure rate <0.05 and Hardy–Weinberg equilibrium with $p > 0.05$). Information on the selected SNPs is shown in Table S3.

For the pooled samples, allele frequencies were estimated and normalized using the *k*-correction method described by Pearson et al. (2007). We measured the level of genetic differentiation between the populations using Wright's *F* statistic (Weir and Cockerham 1984). The F_{ST} values obtained had a low mean (0.026), with 95% of values <0.052 and a skewed distribution: only 0.15% of the SNPs have an $F_{ST} > 0.1$. Overall, this can be interpreted as a measure of high genetic homogeneity among Spanish populations.

We computed the F_{ST} distances between each pair of populations using all the available SNPs. The mean pairwise F_{ST} values (Table S1) ranged from 0.012 (Catalonia vs. Galicia) to 0.018 (Cantabria_Asturias vs. Extremadura). The mean F_{ST} value separating Basques from the rest of populations is similar to that separating other Spanish non-insular populations and thus using a large, genome-wide set of SNPs, Basques appear genetically as an ordinary Iberian population.

The F_{ST} matrix was represented by means of multidimensional scaling (MDS, Fig. 1) and does not apparently show any special separation of Basques or any other expected genetic or geographic pattern. A Mantel test between F_{ST} and geographical distances showed a non-significant negative correlation ($r = -0.107$, $p = 0.314$). A barrier analysis (Bosch et al. 1997) identified first the zone of sharpest genetic change as that around Extremadura; next came a barrier circling Valencia, and a genetic boundary around the Basque Country appeared only in third position. These tests show a general lack of geographic structure for genetic variation in Spain, and, in particular, that Basques are not specially differentiated. Results for principal components analysis on estimated allele frequencies are very similar; nine factors were identified, with a proportion of variation (r^2) explained ranging from 14% for the first PC to 6% for the ninth.

With the second approach (109 highly informative SNPs typed in 238 individuals from the first approach), to detect population structure between samples and individuals, we performed a principal components analysis as implemented in the Eigensoft package (Patterson et al. 2006). ANOVA statistics for population differences along the eigenvectors (first eigenvector $p = 0.16$, second eigenvector $p = 0.167$), as well as visual inspection of the distribution of the individuals along the first two eigenvectors, revealed no detectable separation between Basque and the rest of Spanish individuals. Moreover, PCA analysis taking into account European populations from HGDP panel shows a clear separation between Basques from Spain and Basques from France (Fig. 2). Analogous figures were reported in Novembre et al. (2008) and Lao et al. (2008), with individuals from Iberia separated from other European populations.

The most distinctive feature of these results is the substantial genetic similarity of the Basque population in relation to other Spanish Peninsular populations, observed both at the population and at the individual levels. In the genetic literature, Basques have been described as the most differentiated population in continental Western Europe based on classical genetic polymorphisms. Calafell and Bertranpetit (1994a, b) compiled and analyzed the data available on allele frequencies in the Iberian Peninsula and France. Their results showed a sharp peak in the first PC in the Basque area, which remained even when the geographic scope was widened to include Western Europe. The genetic polymorphisms used in that review included blood groups, genes of the HLA system and some proteins and enzymes, all of them biologically functional, implying that they may have been targets for natural selection. Although they were

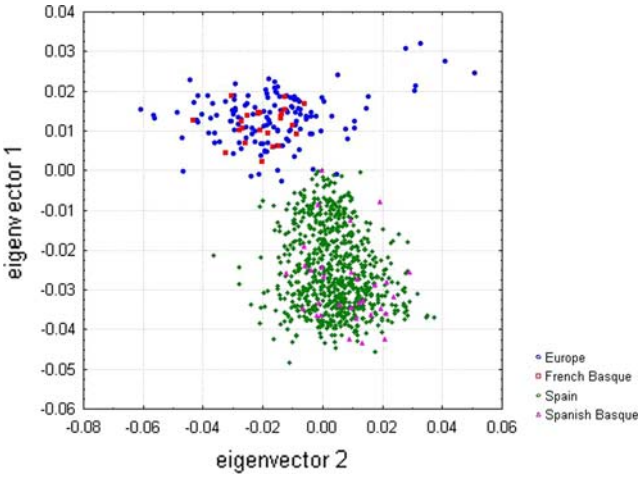


Fig. 2 Spanish and European populations from HGDP samples plotted for the first two principal components obtained by PCA analysis using 109 highly informative SNPs from genotyping data. (Spain: Spanish non-Basque; Europe comprise individuals in HGDP from various locations: French, Sardinian, North Italian, Orcadian, Adygei and Russian)

the only genetic markers available at that time, they are not the ideal neutral markers for tracing population history. Eighteen out of 29 genes included in Calafell and Bertranpetit (1994b) are represented in our dataset (Table S2) by at least one SNP within the gene, even if none is presumably functionally relevant. When F_{ST} values are computed considering Basques versus non-Basques, the mean F_{ST} value obtained is slightly higher in this subset of genes than in the rest (0.011 vs. 0.009; permutation test $p = 0.082$). This marginal increase may explain, at least partially, why this subset of markers shows more differentiation of the Basques versus non-Basques in Calafell and Bertranpetit (1994b) and in other classical genetic markers.

The study of mtDNA sequences and Y-chromosome polymorphisms in the Basque population did not show such relatively intense differences, and located them at the extreme of European-wide gradients in lineage frequencies (Bertranpetit et al. 1995; Bosch et al. 2001; Rosser et al. 2000). Some Mendelian diseases were also found to have increased or decreased incidences in Basques (Bauduer et al. 2005). It is possible that the results were overinterpreted in some of these studies to emphasize the Basque differentiation in light of the previous results on classical genetic markers.

Our analysis showed that, when a genome-wide perspective is applied, Basques are not particularly differentiated from other Iberian populations. The contradiction with previous reports that depicted Basques as genetic outliers can be resolved if we consider that the polymorphisms accounting for most of this differentiation lie in genes such as *ABO*, *RH*, and the *HLA* complex that are, given their involvement in host–pathogen interactions, obvious targets for natural selection in the ancestral populations even at a microgeographic scale. This is yet another example of the sound insights in population genetics that can be achieved with a dense map of genome-wide SNPs, even if only the simplest statistical descriptor, namely, allele frequencies, is pressed into service. Future data with hundreds of thousands of SNPs typed individually in large samples will have to confirm the present findings.

Acknowledgments This research was supported by Genoma España (Proyecto Piloto CeGen), Dirección General de Investigación, Ministerio de Ciencia y Tecnología, Spain (grants SAF2007-63171, BFU2007-63657) and Direcció General de Recerca, Generalitat de Catalunya (2005SGR00608). SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (CEGEN; http://www.cegen.org). Bioinformatic services were kindly provided by the Genomic Diversity node, Spanish Bioinformatic Institute (http://www.inab.org).

References

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogvali EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75:910–918

Adams SM, Bosch E, Balaresque PL, Ballereau SJ, Lee AC, Arroyo E, López-Parra AM, Aler M, Grifo MS, Brion M, Carracedo A, Lavinha J, Martínez-Jarreta B, Quintana-Murci L, Picornell A, Ramon M, Skorecki K, Behar DM, Calafell F, Jobling MA (2008) The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews and Muslims in the Iberian Peninsula. Am J Hum Genet 83:725–736

Bauduer F, Feingold J, Lacombe D (2005) The Basques: review of population genetics and Mendelian disorders. Hum Biol 77(5):619–637

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

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

Bosch E, Calafell F, Pérez-Lezaun A, Comas D, Mateu E, Bertranpetit J (1997) Population history of north Africa: evidence from classical genetic markers. Hum Biol 69:295–311

Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68:1019–1029

Calafell F, Bertranpetit J (1994a) Mountains and genes: population history of the Pyrenees. Hum Biol 66:823–842

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

Chalmers JN, Ikin EW, Mourant AE (1949) The ABO, MN and Rh blood groups of the Basque people. Am J Phys Anthropol 7:529–544

Garagnani P, Laayouni H, Gonzalez-Neira A, Sikora M, Luiselli D, Bertranpetit J, Calafell F (2009) Isolated populations as treasure troves in genetic epidemiology: the case of the Basques. Eur J Hum Genet 17:1490–1494

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18:1241–1248

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456:98–101

Patterson N, Price AL, Reich D (2006) Population structure and eigen analysis. PLoS Genet 2:e190

Pearson JV, Huentelman MJ, Halperin RF, Tembe WD, Melquist S, Homer N, Brun M, Szelinger S, Coon KD, Zismann VL, Webster JA, Beach T, Sando SB, Aasly JO, Heun R, Jessen F, Kolsch H, Tsolaki M, Daniilidou M, Reiman EM, Papassotiropoulos A, Hutton ML, Stephan DA, Craig DW (2007) Identification of the genetic basis for complex disorders by use of pooling-based genome wide single-nucleotide-polymorphism association studies. Am J Hum Genet 80:126–139

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, Beckman G, Beckman L, Bertranpetit J, Bosch E, Bradley DG, Brede G, Cooper G, Corte-Real HB, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Golge M, Hill EW, Jeziorowska A, Kalaydjieva L, Kayser M, Kivisild T, Kravchenko SA, Krumina A, Kucinskas V, Lavinha J, Livshits LA, Malaspina P, Maria S, McElreavey K, Meitinger TA, Mikelsaar AV, Mitchell RJ, Nafa K, Nicholson J, Norby S, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, Previdere C, Roewer L, Rootsi S, Rubinsztein DC, Saillard J, Santos FR, Stefanescu G, Sykes BC, Tolun A, Villems R, Tyler-Smith C, Jobling MA (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67:1526–1543

Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo A (1998) mtDNA analysis of the Galician population: a genetic edge of European variation. Eur J Hum Genet 6:365–375

Stacklies W, Redestig H, Scholz M, Walther D, Selbig J (2007) pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23:1164–1167

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Zlojutro M, Roy R, Palikij J, Crawford MH (2006) Autosomal STR variation in a Basque population: Vizcaya Province. Hum Biol 78:599–618