

High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations

Naiara Rodríguez-Ezpeleta · Jon Álvarez-Busto · Liher Imaz · María Regueiro ·
María Nerea Azcárate · Roberto Bilbao · Mikel Iriondo · Ana Gil · Andone Estonba ·
Ana María Aransay

Received: 24 March 2010 / Accepted: 23 April 2010
© Springer-Verlag 2010

Abstract A recent study reported that Basques do not constitute a genetically distinct population, and that Basques from Spanish and French provinces do not show significant genetic similarity. These conclusions disagree with numerous previous studies, and are not consistent with the historical and linguistic evidence that supports the distinctiveness of Basques. In order to further investigate this

controversy, we have genotyped 83 Spanish Basque individuals and used these data to infer population structure based on more than 60,000 single nucleotide polymorphisms of several European populations. Here, we present the first high-throughput analysis including Basques from Spanish and French provinces, and show that all Basques constitute a homogeneous group that can be clearly differentiated from other European populations.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-010-0833-4) contains supplementary material, which is available to authorized users.

Basques are a linguistically isolated population who speak the only non-Indo-European language in Western Europe. They live in the West Pyrenees, across the French-Spanish border (Fig. S1), and are considered one of the European genetic outliers. The distinctiveness of Basques has been deduced from the analysis of classical markers (i.e., HLA, Rh and ABO) (Aguirre et al. 1991; Calafell and Bertranpetit 1994; Cavalli-Sforza et al. 1994; Comas et al. 1998; Manzano et al. 1996), of microsatellites data (Belle et al. 2006; Iriondo et al. 1997, 2003; Zlojutro et al. 2006), of mitochondrial DNA (mtDNA) sequences (Achilli et al. 2004; Corte-Real et al. 1996) and Y-chromosome polymorphisms (Alonso et al. 2005; Hurler et al. 1999; Rosser et al. 2000). Yet, complementary analyses based on *Alu* insertions as well as on mtDNA variation were not able to detect differences between Basques and other European populations (Bertranpetit et al. 1995; Comas et al. 2000).

N. Rodríguez-Ezpeleta (✉) · J. Álvarez-Busto · L. Imaz ·
M. Regueiro · A. M. Aransay (✉)
Genome Analysis Platform, Functional Genomics Unit,
CIC bioGUNE, Bizkaia Technology Park,
Building 502, 48160 Derio, Spain
e-mail: nrodriguez@cicbiogune.es

A. M. Aransay
e-mail: amaransay@cicbiogune.es

M. N. Azcárate · R. Bilbao
Basque Foundation for Health Innovation and Research,
Plaza Asua 1, 48150 Sondika, Spain

M. Iriondo · A. Gil · A. Estonba
Department of Genetics, Physical Anthropology and Animal
Physiology, Faculty of Science and Technology,
University of the Basque Country, 48940 Leioa, Spain

Present Address:

M. Regueiro
Department of Molecular and Human Genetics,
College of Medicine, Florida International University,
Miami, FL 33199, USA

Present Address:

M. N. Azcárate
Centro Vasco de Transfusión y Tejidos Humanos,
Barrio Labeaga s/n, 48960 Galdakao, Spain

Livening up the debate on the genetic distinctiveness of Basques, the respective analysis of 123 single nucleotide polymorphisms (SNPs) located in a region of chromosome 22 typed individually, and of 280,862 SNPs typed on pools of 30 individuals shows no difference between Basques from Spain and other Spanish populations (Garagnani et al. 2009; Laayouni et al. 2010). Furthermore, a principal component analysis (PCA) based on 109 SNPs typed individually reveals two clusters: one grouping Spanish Basques with other Spanish regions and the other grouping French

Basques with the remaining European populations included in the study (Laayouni et al. 2010). These results lead the authors to conclude that there is a “clear separation between Basques from Spain and Basques from France” (Laayouni et al. 2010), and to suggest that the assumed genetic distinctiveness of Basques is due to the action of pathogen-mediated natural selection on some classical polymorphisms (Garagnani et al. 2009; Laayouni et al. 2010). Besides conflicting with some of the existing genetic data, these findings strongly contradict historical and linguistic evidence that supports the distinctiveness of Basques; therefore, the necessity of testing these hypotheses with a large number of SNPs typed individually becomes evident.

In an attempt to resolve the controversy surrounding the genetic differentiation of Basques within the European populations and the relationships between Spanish and French Basques, we generated a genome-wide high-density dataset. Since no individual genotyping data was available for Spanish Basques, we sampled 83 individuals from four provinces (22 from Alava, 22 from Biscay, 22 from Guipuscoa and 17 from Navarre) with the following criteria: (1) they are healthy on the basis of haematological and physical examination, (2) they, their parents, grandparents and great grandparents are born in the same province (genealogies were studied based on parish and civil registries data), (3) they are unrelated to each other by at least three generations and (4) they are between 18 and 45 years old to avoid generation overlapping. All participants gave informed consent before any genealogy or blood extraction was done, and the ethics committee of the Galdakao Hospital approved the sampling procedures. The samples are available from the Basque Biobank for Research (<http://www.biobancovasco.org>). Genotyping was carried out using three different technologies (Table S1), and SNPs were selected according to several quality control criteria (Table S2). Additionally, individual genotypes from eight European populations (including French Basques) were obtained from the Human Genome Diversity Project (<http://www.cephb.fr/en/hgdp/diversity.php/>). After combining our samples with the unrelated, non-atypical and unique individuals described in Rosenberg (2006) for the eight European populations (Table S3), there are 61,504 polymorphic SNPs in common that have no >10% of missing data.

To determine whether our geography-based assignment of samples to populations reflects the underlying genetic structure, a Bayesian clustering approach (Falush et al. 2003) was used to examine the genetic ancestry of each of the 240 studied European individuals (Fig. 1). This analysis measures the contribution of each of the K potential ancestral populations to each individual without any prior assignment of individuals to populations. Already at $K = 3$, a clear differentiation of Basques, Sardinians and the rest of the

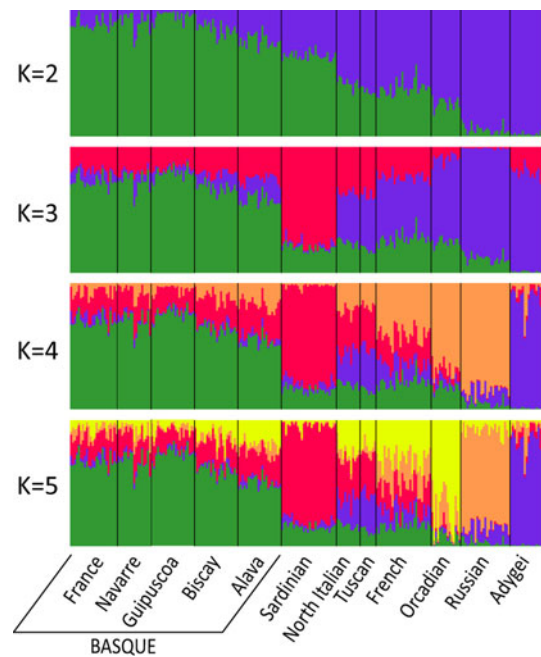


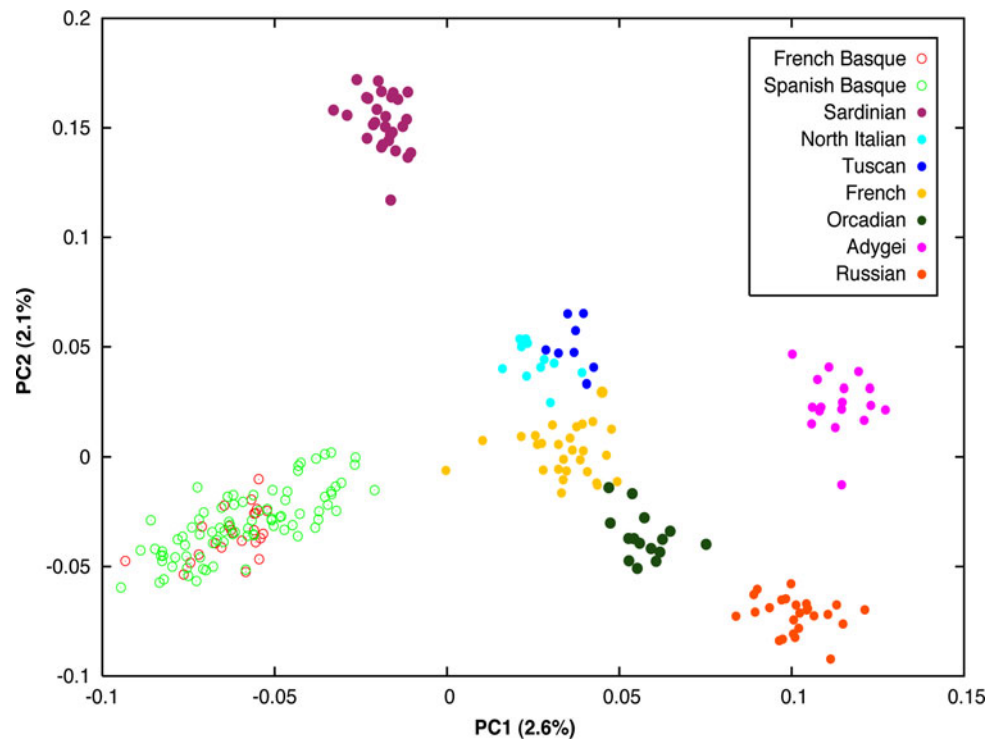
Fig. 1 Population structure inferred by Bayesian clustering. Graphical representation of the individual ancestry where each bar represents an individual and each color, its inferred membership in each of the K genetic clusters. Clustering was inferred with STRUCTURE v 2.2 (Falush et al. 2003) and based on 38,228 SNPs (owing to computing limitations, only the SNPs that have 0% missing data and a $MAF > 0.05$ were used). Analysis was performed without any prior population assignment, based on the admixture model and a burn-in period of 20,000 iterations followed by 40,000 iterations from which estimates were obtained. For each value of K , ten replicate analyses were performed and analyzed with CLUMPP (Jakobsson and Rosenberg 2007) to identify common modes as in Jakobsson et al. (2008). Results were plotted using DISTRUCT (Rosenberg 2004)

populations is depicted, a result that agrees with the identification of Basques and Sardinians as European genetic outliers using other polymorphisms (Cavalli-Sforza et al. 1994). Differences among the other European populations become more noticeable at $K = 5$, when Russians, Adygei and Orcadians are segregated from the rest.

PCA was performed to capture a major proportion of the genetic variability. In the PCA, the eight European populations can be clearly separated (Fig. 2) and depict the exact same distribution obtained by Li et al. (2008) using 10 times more SNPs. Notably, the Spanish and French Basques are undistinguishable from each other, further supporting the observations from the ancestry analysis and strongly contradicting the results obtained by Laayouni et al. (2010).

In order to put our Basque samples in the context of other Iberian populations, and to test the substantial genetic similarity of the Basques to other Spanish populations found by Laayouni et al., we took advantage of 193 Basque resident individuals available in our laboratory (L. I., unpublished data), whose only sampling criteria are that

Fig. 2 Population stratification of European populations. Graphical representation of the two principal components obtained using EIGENSOFT (Price et al. 2006) and based on 61,504 autosomal SNPs. The first and second principal components explain the 2.6 and 2.1% of the variation, respectively. ‘Spanish Basque’ label (*red open circles*) designates individuals from Alava, Biscay, Guipuscoa and Navarre and ‘French Basque’ label (*green open circles*), individuals from the French Basque region



they are healthy and reside in Biscay. Given the immigration from other Spanish provinces to the Basque Country during the second half of the twentieth century, this resident population is expected to be composed of individuals of Basque, Spanish and mixed ancestry. As shown in Fig. S2, the Basque resident individuals cover a spectrum that goes from the Basque population to the center of the plot, where continental Western European populations are located (i.e., French and Italians). This suggests that the genetic location of the Spanish population is close to other Western European populations, as it has already been shown with other genetic markers (Calafell and Bertranpetit 1994; Cavalli-Sforza et al. 1994).

Our findings entirely fit with those of Li et al. (2008), but strongly disagree with the results obtained by Laayouni et al. (2010), who concluded that there is a lack of geographic structure for genetic variation in Spain and, in particular, that Basques are not specially differentiated. This may be due to the fact that their PCA was based on 280,862 SNPs genotyped for one pool of DNA per population, and not for each individual. Although genotyping of pooled samples has proved to be efficient for assessing the genetic ancestry of a population, this method only performs well for SNPs displaying large allele differences across populations, and genome-wide genotyping of pooled DNA needs to be followed by strict quality control filtering to remove poorly performing SNPs (Chiang et al. 2010). Additionally, in another analysis based on 109 SNPs, Laayouni et al. found a clear separation between Spanish Basques and

French Basques, the first ones clustering with other Spanish populations, and the second, with the remaining European individuals. Except for the Spanish Basques, the European individuals used by Laayouni et al. are the same as the ones used here and in Li et al. For that reason, the surprising fact that these 109 SNPs were not able to find the clearly delineated population structure found by us and by Li et al. suggests that the SNP choice of Laayouni et al. may be biased. Indeed, the 109 SNPs used to detect population structure within Europe were selected on the basis of their contribution to the principal component factors in the initial analysis of the pooled genotypes, which may be introducing a bias since a panel of informative markers for a particular question may not be effective for differentiating other populations (Chiang et al. 2010).

Finally, to quantify genetic variation between populations, we calculated the F_{ST} distances among the eight populations from the allele frequencies across all studied SNPs (Table 1, see also Table S4). Among the populations studied, French and Italians are the most similar to Basques, with F_{ST} values ranging from 0.005 to 0.007. The values obtained here for European populations (ranging from 0.002 to 0.020) are similar to the ones obtained from the analysis of 377 microsatellites on a similar set of populations (Belle et al. 2006) and to the ones obtained using classical polymorphisms (Cavalli-Sforza et al. 1994). Strikingly, the F_{ST} values obtained by Laayouni et al. (2010) for the Spanish populations are much bigger on average (ranging from 0.012 to 0.018) and comparable to

Table 1 Pairwise F_{ST} values across European populations

	B	A	R	S	O	T	NI
Adygei	0.016						
Russian	0.013	0.012					
Sardinian	0.011	0.019	0.020				
Orcadian	0.010	0.014	0.008	0.017			
Tuscan	0.007	0.007	0.009	0.007	0.008		
N. Italian	0.006	0.009	0.010	0.007	0.007	0.002	
French	0.005	0.009	0.005	0.009	0.004	0.002	0.002

In bold, distances between Basques and the other populations

B Basque (including Spanish and French), A Adygei, R Russian, S Sardinian, O Orcadian, T Tuscan, NI North Italian

the values we obtain for distant European populations. Again, this may be an indicator of the inaccuracies a flawed genotyping strategy may introduce.

To summarize, our genetic analysis using a large number of SNPs in European populations confirms that Basques are differentiated from other European populations and shows that Spanish and French Basques are not genetically distinct from each other. These results are consistent with historical and linguistic data of European populations and settle the question of the genetic distinctiveness of Basques. The inconsistencies found between our study and the one of Laayouni et al. point out important marker selection and methodological issues and suggest that deciphering the structure of Spanish populations requires more in-depth analysis of a large number of SNPs typed individually. Finally, our results have an important impact in future epidemiological and evolutionary studies concerning Basques. Indeed, the genetic isolation of this population calls for additional analyses determining the haplotype block structure and exploring the portability of the European tag SNPs to the Basque population, which shows some epidemiological particularities such as the highest worldwide prevalence of certain diseases (e.g. the limb-girdle muscular dystrophy (Urtasun et al. 1998)).

Acknowledgments We wish to thank all participating individuals for blood donation, Lore Erriondo (Udako Euskal Unibertsitatea) for assistance in sample collection, Miguel Angel Vesga for technical support, and Ewa Gubb, Ane Fullaondo and Michael P. Cummings for useful comments on the manuscript. This project was supported by the Department of Industry, Tourism and Trade of the Government of the Autonomous Community of the Basque Country (Ertortek Research Programs 2005/2006/2007) and from the Innovation Technology Department of the County of Biscay.

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, Torrioni 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

Aguirre A, Vicario A, Mazon LI, Estomba A, Martinez de Pancorbo M, Arrieta Pico V, Perez Elortondo F, Lostao CM (1991) Are the Basques a single and a unique population? *Am J Hum Genet* 49:450–458

Alonso S, Flores C, Cabrera V, Alonso A, Martin P, Albarran C, Izagirre N, de la Rua C, Garcia O (2005) The place of the Basques in the European Y-chromosome diversity landscape. *Eur J Hum Genet* 13:1293–1302

Belle EM, Landry PA, Barbujani G (2006) Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. *Proc Biol Sci* 273:1595–1602

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

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

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. University Press, Princeton

Chiang CWK, Gajdos SKZ, Korn JM, Kuruvilla FG, Butler JL, Hackett R, Guiducci C, Nguyen TT, Wilks R, Forrester T, Haiman C, Henderson K, Le Marchand L, Henderson B, Palmert M, McKenzie C, Lyon H, Cooper R, Zhu X, Hirshhorn J (2010) Rapid assessment of genetic ancestry in populations of unknown origin by genome-wide genotyping of pooled samples. *PLoS Genet* 6:1–11

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

Comas D, Calafell F, Benchemsi N, Helal A, Lefranc G, Stoneking M, Batzer MA, Bertranpetit J, Sajantila A (2000) Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Hum Genet* 107:312–319

Corte-Real HB, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, Bertranpetit J, Sykes BC (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* 60:331–350

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587

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

Hurles ME, Veitia R, Arroyo E, Armenteros M, Bertranpetit J, Perez-Lezaun A, Bosch E, Shlumukova M, Cambon-Thomsen A, McElreavey K, Lopez De Munain A, Rohl A, Wilson JJ, Singh L, Pandya A, Santos FR, Tyler-Smith C, Jobling MA (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am J Hum Genet* 65:1437–1448

Iriondo M, Barbero MC, Izagirre N, Manzano C (1997) Data on six short-tandem repeat polymorphisms in an autochthonous Basque population. *Hum Hered* 47:131–137

Iriondo M, Barbero MC, Manzano C (2003) DNA polymorphisms detect ancient barriers to gene flow in Basques. *Am J Phys Anthropol* 122:73–84

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–1806

- Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003
- Laayouni H, Calafell F, Bertranpetit J (2010) A genome-wide survey does not show the genetic distinctiveness of Basques. *Hum Genet* 127:455–458
- 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
- Manzano C, Aguirre AI, Iriondo M, Martin M, Osaba L, de la Rua C (1996) Genetic polymorphisms of the Basques from Gipuzkoa: genetic heterogeneity of the Basque population. *Ann Hum Biol* 23:285–296
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
- Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Mol Ecol Notes* 4:137–138
- Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet* 70:841–847
- Rosser ZH, Zerjal T, Hurler 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
- Urtasun M, Saenz A, Roudaut C, Poza JJ, Urtizberea JA, Cobo AM, Richard I, Garcia Bragado F, Leturcq F, Kaplan JC, Marti Masso JF, Beckmann JS, Lopez de Munain A (1998) Limb-girdle muscular dystrophy in Guipuzcoa (Basque Country, Spain). *Brain* 121(Pt 9):1735–1747
- Zlojutro M, Roy R, Palikij J, Crawford MH (2006) Autosomal STR variation in a Basque population: Vizcaya Province. *Hum Biol* 78:599–618