Evidence of pre-Roman tribal genetic structure in Basques from uniparentally inherited markers

Begoña Martínez-Cruz[1], Christine Harmant[2], Daniel E. Platt[3], Wolfgang Haak[4], Jeremy Manry[2], Eva Ramos-Luis[5], David F Soria-Hernanz[6], Frédéric Bauduer[7], Jasone Salaberria[8], Bernard Oyharçabal[8], Lluis Quintana-Murci[2], David Comas[1], the Genographic Consortium.

[1]Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, 08003 Barcelona, Spain

[2]Human Evolutionary Genetics Unit, Centre National de la Recherche Scientifique URA31012, Institut Pasteur, 75015 Paris, France

[3] Bioinformatics and Pattern Discovery, IBM, TJ Watson Research Center, Yorktown Hgts, NY 10598, USA

[4] Australian Centre for Ancient DNA, The University of Adelaide, SA-5005, Australia

[5] Grupo de Medicina Xenómica, Instituto de Ciencias Forenses Luis Concheiro, Universidade de Santiago de Compostela, CIBERER, Spain

[6] The Genographic Project, National Geographic Society, Washington, DC, USA

[7] Laboratoire MRGM, EA 4576, University of Bordeaux, Bordeaux, France

[8] Centre National de la Recherche Scientifique, CNRS UMR5478 IKER, Bayonne, France

Corresponding authors: Lluis Quintana-Murci (quintana@pasteur.fr) and David Comas (david.comas@upf.edu)

Key words: Basques, Y-chromosome, mitochondrial DNA, haplogroups, human populations, tribal genetic structure

Abstract

Basque people have received considerable attention from anthropologists, geneticists and linguists during the last century due to the singularity of their language and to other cultural and biological characteristics. Despite the multidisciplinary efforts performed to address the questions of the origin, uniqueness and heterogeneity of Basques, the genetic studies performed up to now have suffered from a weak study-design where populations are not analyzed in an adequate geographic and population context. To address the former questions and to overcome these design limitations, we have analyzed the uniparentally inherited markers (Y chromosome and mitochondrial DNA) of ~900 individuals from 18 populations, including those where Basque is currently spoken and populations from adjacent regions where Basque might have been spoken in historical times. Our results indicate that Basque-speaking populations fall within the genetic Western European gene pool and they are similar to geographically surrounding non-Basque populations, and also that their genetic uniqueness is based on a lower amount of external influences compared to other Iberians and French populations. Our data suggest that the genetic heterogeneity and structure observed in the Basque region results from pre-Roman tribal structure related to geography and might be linked to the increased complexity of emerging societies during the Bronze Age. The rough overlap of the pre-Roman tribe location and the current dialect limits supports the notion that the environmental diversity in the region has played a recurrent role in cultural differentiation and ethnogenesis at different time periods.

Introduction

Basques represent one of the European ethnic groups that have drawn the attention of anthropologists in the last century due to their cultural and biological characteristics. Basques live in the western edge of the Pyrenees, in the Atlantic area of the present Spanish-French administrative border. *Euskal-Herria* ("Basque Country" in the Basque language, i.e. *Euskera*) is made up of seven provinces: Bizkaia, Gipuzkoa, Araba, and Navarre located in Spain; and Zuberoa, Lapurdi, and Nafarroa Beherea located in France. The main peculiarity of the Basques is their language, an isolated non-Indo-European language that has no close relation with any other language spoken in the world at present (Campell 1998; Comrie, Matthews, Polinsky 2003) although some authors defend that Basque belongs to the Dene-Caucasic family (Ruhlen 2001). At present around 650,000 people speak Basque (EuskoJaurlaritza 2008) in one of the five Basque varieties that modern experts acknowledge (Zuazo 2010) (Figure 1). Scholars affirm that Basque had been spoken in the past in territories where it was lost later, even in regions outside the historical Basque Country (*Euskal Herria*) (Luchaire 1875/1877; Rohlfs 1935; Caro Baroja 1943; Michelena 1961/1962; Michelena 1976; Gorrochategui 1984; Gorrochategui 1995) as substantiated by inscriptions of the names of people, places and deities, and remnants of Basque in the languages spoken currently in those regions (Gorrochategui 1984; Gorrochategui 1995), although some scholars point to a reduced extension of the Basque territory (Villar, Prosper 2005). By the time of Roman penetration into the Iberian Peninsula along the Ebro river, classical authors like Ptolemy, Strabo and Pliny mentioned six different tribes or peoples living in the territory between the Garonne and Ebro rivers: *Aquitani*, *Vascones*, *Varduli*, *Caristi*, *Autrigones*, and *Berones*. Based on theonyms and personal names found in Aquitania (Luchaire 1875/1877; Gorrochategui 1984) and additional evidence (such as some

morphological characteristics of place names (Rohlfs 1935)) there is general agreement that the *Aquitani* spoke a form of proto-Basque language. Otherwise Trask (1997) suggested that Basque language penetrated south to the Pyrenees in post-Roman times. However, given that Aquitanian onomastic elements have been found in present-day Navarre, small parts of Gipuzkoa and neighboring zones on the East (Gorrochategui 2007) most scholars agree today that Basque was also spoken by *Vascones* in pre-Roman times. As for the other tribes there are more doubts whether they spoke Basque at this time, because there is no evidence of inscriptions or place names in Basque language.

Besides their linguistic isolation, the uniqueness of Basques has been addressed using craniometrical traits (de la Rúa 1992) and by studying their gene pool. However, studies have thus far shown no agreement regarding genetic uniqueness of Basques when compared to their neighboring populations in Europe. A large number of genetic studies, mainly *classical* polymorphisms and HLA antigens (Calafell, Bertranpetit 1994; Cavalli-Sforza, Menozzi, Piazza 1994; Lucotte, Hazout 1996; Comas et al. 1998a; Comas et al. 1998b; Bauduer, Feingold, Lacombe 2005), but also mtDNA (Garcia et al. 2011), Y-chromosome (Chikhi et al. 2002) and autosomal markers (Perez-Miranda et al. 2005; Li et al. 2008; Rodriguez-Ezpeleta et al. 2010), argue for a distinct genetic 'outlier' status. In complete contrast, other studies have shown that Basques fall well within the European genetic landscape (Bertranpetit et al. 1995; Alonso et al. 2005; Alzualde et al. 2005; Adams et al. 2008; Garagnani et al. 2009; Laayouni, Calafell, Bertranpetit 2010). Genetic distinctness of Basques has been explained by their linguistic and geographical isolation, leading to their identification as potential candidates for Palaeolithic remnant populations with little admixture with the Neolithic populations coming from the Near East (Chikhi et al. 2002; Bauduer, Feingold,

Lacombe 2005). In contrast, other studies have explained the unique genetic characteristics as a result of lower gene flow during later, i.e. post-Neolithic times than in their surrounding populations and as a consequence stronger effects of drift (Adams et al. 2008).

In addition, the genetic similarity among Basque groups as suggested by HLA antigens (Comas et al. 1998a) and autosomal SNPs (Rodriguez-Ezpeleta et al. 2010) has been challenged by the degree of genetic heterogeneity within Basques, based on *classical* and autosomal markers (Manzano et al. 2002; Iriondo, Barbero, Manzano 2003; Perez-Miranda et al. 2005; Alfonso-Sanchez et al. 2008), and does not correspond to the current administrative borders but points to some influence of the pre-Roman tribal division in the present genetic structure of the Basque population.

Despite the large number of genetic analyses that have been done focusing on the Basques, most of these bare considerable limitations. For example, a drawback of the use of *classical* markers is that they may be under natural selection thus reflecting environmental factors rather than genuine demographic processes. Additionally, in some of these studies, Basques were not analyzed in an adequate population context and/or were defined as a single population (Li et al. 2008; Rodriguez-Ezpeleta et al. 2010), thus limiting clear conclusions about their genetic relationships with the surrounding populations. To date, none of the many studies have included all present-day populations were Basque is currently spoken, nor included populations where ancient Basque is assumed to have been spoken in historical times.

Here we present a high resolution analysis of uniparentally inherited markers (Y-chromosome and mtDNA) in ~900 individuals from 18 geographical areas to address questions about the origin, uniqueness, and heterogeneity among Basques in full detail. Our study design incorporates, for the first time all, geographical regions where Basque

is currently spoken including the seven provinces of *Euskal-Herria*, as well as the surrounding regions where Basque was probably spoken in historical times. We aim to disentangle whether there is a genetic structure in the region studied, and if so, if it can be better explained by pre-Roman tribal affiliation of present populations or by their linguistic affiliation. We aim also to understand how Basques and surrounding non-Basque populations are situated in a broader Iberian and French context.

Material and Methods

Sample collection

A total of 886 unrelated individuals from 18 geographical areas from the Basque country and surrounding Spanish and French speaking regions were collected. For all subjects, written informed consent was obtained, and Ethics Committees at Universitat Pompeu Fabra, Institut Pasteur, Université Michel de Montaigne Bordeaux 3 and the CCPPRB (Comité Consultative de Protection des Personnes dans la Recherche Biomédicale d'Aquitaine) approved all procedures. All individuals were interviewed in order to assess the geographical origin of their grandparents and their speaking dialect. DNA was extracted from fresh blood by standard phenol-chloroform methods.

The geographical area surveyed (Figure 1 and Table S1) includes seven regions where Basque dialects are currently spoken (Zuazo 1998) (Lapurdi/Baztan, Lapurdi Navarre, Zuberoa, Northwestern Navarre, Gipuzkoa, Soutwestern Gipuzkoa, Bizkaia) and three regions where it was spoken up to the last century (Roncal, Central Western Navarre, and parts of Araba). In addition, we collected samples from three French-speaking regions where Gascon was spoken in historical times (Bigorre, Bearn, Chalosse) and five Spanish-speaking regions (Western Bizkaia, Cantabria, Northern Burgos, La Rioja, Northern Aragon) where Basque languages are suspected to have been spoken prior to medieval times (Zuazo 2010 and references therein). Based on the geographical location, we assigned these 18 populations to the six historical pre-Roman tribes as follow: *Aquitani* (Bigorre, Béarn, Chalosse, Lapurdi Navarre, Zuberoa), *Vascones* (Roncal, Central Western Navarre, Northwestern Navarre), *Varduli* (Gipuzkoa, Southwestern Gipuzkoa), *Caristi* (Bizkaia), *Autrigones* (Western Bizkaia, Cantabria) and *Berones* (La Rioja) (see Figure 1 and Table S1). Although the precise geographic boundaries are not known, descriptions from Greco-latin historians situate

grosso modo the *Vascones* inhabiting the present Navarre, the *Aquitani* the present Aquitania, the *Varduli* the present Gipuzkoa and East Araba, the *Carisiti* the eastern part of Bizkaia and west Araba, the *Berones* the present La Rioja, and the *Autrigones* the North of present Burgos. Populations from Lapurdi/Baztan, Araba, Burgos and North Aragon could not be assigned to a single pre-Roman tribe because their modern-day geographical ranges were historically shared between two or three tribes, and thus were excluded from all the analyses involving pre-Roman tribes.

In order to integrate the analyzed populations under study in a broader Iberian and French context, we generated a Y-chromosome database of fourteen additional Iberian and seven French populations typed for biallelic markers and ten common STRs (Adams et al. 2008; Ramos-Luis et al. 2009) and a mitochondrial database with additional 15 Iberian (Corte-Real et al. 1996; Salas et al. 1998; López-Soto, Sanz 2000; Martinez-Jarreta et al. 2000; Pereira, Prata, Amorim 2000; Larruga et al. 2001; Maca-Meyer et al. 2003; Plaza et al. 2003; Picornell et al. 2005; Alfonso-Sanchez et al. 2008) and 9 French populations (Dubut et al. 2004; Richard et al. 2007) which encompassed 275 bp of the mitochondrial control region (positions 16090 to 16365 according to Andrews et al. (1999)). These two databases are hereafter referred to as Y extended and mtDNA extended databases (see Table S2).

Y-chromosome genotyping

A total of 835 samples were genotyped for the non-recombining region of the Y chromosome with the TaqMan technology in a hierarchical manner for a set containing a variable number of SNP markers (54 in total) and for six indels amplified in a single multiplex named Multiplex-2 (see Supplementary Material for a complete description of the markers used and the methodology of typing employed). In order to have an external

control, some individuals were also genotyped using the OpenArray technology as described previously (Martinez-Cruz et al. 2011). Nomenclature of the haplogroups (Supplementary Table S3a) is in accordance with the Y-Chromosome Consortium (Karafet et al. 2008).

All the individuals were typed for a set of 19 STRs using two different multiplexes: 17 STRs were amplified with the commercial AmpF*l*STR® Yfiler kit™ (Applied Biosystems) and two additional STRs, DYS426 and DYS388, were amplified within the Multiplex 2 (see Supplementary Material). DYS385 was excluded from all the analyses performed as the Yfiler kit amplifies DYS385a/b simultaneously rendering the distinction of each of the two alleles (a or b) impossible.

Mitochondrial DNA genotyping

A total of 881 samples were sequenced for both Hyper Variable Segments I and II (HVS-I and HVS-II) of the control region (Behar et al. 2007) and typed using a 22 coding region SNPs multiplex SNaPshot assay (GenoCoRe22), diagnostic of the major branches of mtDNA phylogeny (Haak et al. 2010). Variable positions throughout the control region were determined from positions 16,024–573. Sequences were deposited in the Genebank nucleotide database under accession numbers JQ737242 - JQ738121.

As mitochondrial hg H is the most frequent haplogroup, reaching up to 50%, in Western European populations, we designed a specific multiplex (termed HPLEX17) in order to resolve the most common 17 distinct sub-haplogroups of hg H (as described in the Supplementary Material), and utilize their genetic information value.

Based on combined HVS and coding region SNP data individuals were assigned to the major haplogroups of the mtDNA phylogeny (Supplementary Table S3b). Due to

phylogenetic uncertainty, indels at nucleotide positions 309, 315, and 16193 were not taken into account.

Statistical Analyses

Diversity parameters such as haplogroup diversitiy, number of observed STR haplotypes, pairwise F_{ST} (for haplogroups), and pairwise R_{ST} (for STRs), sequence diversity values, and mean pairwise differences were calculated with Arlequin 3.4 (Excoffier, Laval, Schneider 2005) and are shown in Supplementary Table S4. T-test were performed in order to compare differences in haplogroup diversity and mean number of pairwise differences between Basque and non-Basque speakers. For reasons of compatibility with published data, mtDNA F_{ST} were restricted to HVRI nucleotide positions 16,090-16,365. In all cases F_{ST} and R_{ST} results were corrected using the Bonferroni correction for multiple comparisons ($p < 0.05$).

Principal component analyses (PCA) based on haplogroup frequencies for both Y chromosomal and mitochondrial markers were performed using the software package STATISTICA 7 (http://www.statsoft.com). For the Y-chromosome, the analyses were first performed with our set of populations and subsequently for the Y extended database. For the latter, the haplogroup definition had to be adjusted to the haplogroup resolution of the samples presented in Adams et al. (2008) and in Ramos-Luis et al. (2009). For mtDNA, PCA was first performed with the present sample set of populations for both HVRI and HVRII. Subsequently, the analyses were performed with the mtDNA extended database taking into account only HVRI region since data for HVRII was not available for all samples in the literature.

In order to investigate the role of language, current administrative borders, and assignment to pre-Roman tribal groups in relation with the genetic diversity of our

samples, a hierarchical analyses of molecular variance (AMOVA) pooling populations into ethnic groups (French, Basque, and Spanish), current administrative borders (France and Spain) or pre-Roman tribal groups (*Aquitani*, *Vascones*, *Varduli*, *Berones*, *Caristi*, and *Autrigones*) was performed (Table 1). Testing was performed using frequencies for the haplogroup data, R_{ST} distances for the Y STRs, and number of pairwise differences for the HVRI-HVRII mtDNA sequence data. The correlation between geographical *vs.* genetic distances was investigated with a Mantel test using Genepop (Raymond, Rousset 1995).

Time to the most common recent ancestor (TMRCA) for the most common Y haplogroups was estimated by calculating the mean STR variance as proposed by Morral et al. (2001) using a mean STR mutation rate of 0.00069 per generation of 25 years (Zhivotovsky et al. 2004). TMRCA for the most common mtDNA haplogroups was estimated based on the average number of mutations accumulated from an ancestral sequence as a linear function of time and mutation rate using Network 4.5.0.0 (http://www.fluxus-engineering.com). The ages were estimated based on the mutation rate corrected for purifying selection for the mtDNA control region (one mutation every 9,058 years) as described by Soares et al. (2009).

TMRCA and splitting time between French-Basques-Spanish and among the different tribal groups was estimated with BATWING (Wilson, Weale, Balding 2003) using the 17 STRs under a model of exponential growth and splitting from a constant-size ancestral population. In the preliminary analyses, subsampling had been performed noting that UEPs (unique event polymorphisms) that were under-represented in any of the populations tended to produce disproportionate impacts on population split tree topologies. Consequently, we retained fourteen UEPs insensitive to sub-sampling variations, derived from the markers M170, M253, P215, P37.2, M26, M9, M45, M173,

L23, P311, U106, P312, M153, and L21. Mutation rate priors were used as proposed by Xue et al.(2006) based on Zhivotovsky et al. (2004). Priors and further information about the BATWING procedure are shown in the Supplementary Material.

Results

Y chromosome perspective

A total of 31 Y-chromosome haplogroups were observed in the present sample set, defined by the 54 binary markers genotyped (Supplementary Table S3a). With the exception of four individuals, all the samples within the R haplogroup belonged to the branch defined by M269 (R1b1b2), which is the most frequent in Western Europe (see (Francalacci, Sanna 2008; Francalacci et al. 2010) for a review). When disentangling the variation concealed in the highly frequent R1b1b2 haplogroup, the most frequent lineage was R1b1b2b* (defined by P312) in all the populations analyzed with the exception of SOU (with higher frequencies of R1b1b2b, defined by M153). The other most frequent sub-haplogroups within R1b1b2-M269 in the present sample set were R1b1b2b5-L21, R1b1b2b3-SRY2627 and R1b1b2b-M153. Besides the high frequency of R1b1b2-M269 and its sub-haplogroups, the frequency of the haplogroup I2a1-M26 is noteworthy, and is consistent with what has been reported for other regions in Spain (Francalacci, Sanna 2008). I2a1-M26 is present at high frequencies in Sardinia (35-37%, (Passarino et al. 2001; Semino et al. 2004; Capelli et al. 2006; Contu et al. 2008) but is very rare in other western European populations and even absent in the rest of Europe (Francalacci, Sanna 2008). The low frequency of haplogroups E1b1b1b-M81 should be also highlighted, which is otherwise more frequent in the Iberian Peninsula and whose presence has been suggested to be the result of north African influence (see Adams et al. 2008 and references therein).

Due to the refined resolution of haplogroup R1b1b2-M269 of our study, we observed higher levels of haplogroup diversity than in previous works (Alonso et al. 2005; Adams et al. 2008). However, three of our seven Basque-speaking populations, GUI, GSO and BBA, displayed lower levels than the rest (Supplementary Table S4).

Consequently, when we grouped our populations according to speaking a Basque or a non-Basque language, the haplogroup diversity was lower for Basque speakers (haplogroups p=0.0078, t=3.04, df=16; mean number of pairwise differences p=0.0023, t=3.62, df=16) (Supplementary Table S4). F_{ST} genetic distances based on haplogroups and R_{ST} haplotype distances were not significant with minor exceptions (Tables S5a and S5b). There was a significant correlation between geographic and genetic distances for both Y chromosome haplogroups (p=0.007) and haplotypes (p=0.034). PCA based on haplogroup frequencies grouped all samples except BUR, RIO and CAN, with ALA slightly differentiated from the cluster (Figure 2a). The three French populations (BIG, BEA and CHA) and the Spanish NAR and BOC populations cluster together with the Basques, suggesting a similar haplogroup distribution in samples of the Western Pyrenees. This picture is similar when all the historical regions of the Iberian Peninsula and the additional seven French populations are incorporated (Figure 2b). Interestingly, BRI, the French Bretagne population, clusters in the middle of the Basque group.

Despite the population clustering shown in the PCA, when samples were grouped by ethnic affiliation, French-Basques-Spanish, and tested through the analyses of molecular variance (AMOVA), some level of differentiation among groups of populations was shown in some of the analyses (Table 1). No genetic differences were found when the administrative border (present Spanish and French border) was taken into account (Table 1).

BATWING was used to assess the slight structure within ethnic groups identified by AMOVA, to examine whether the structure in pre-Roman tribes was sex-biased, and to estimate the timing of population splits into different tribes. In the case of ethnic affiliation (French-Spanish-Basques), times of split included systematically the zero thus indicating no clear structure upon this criterion (data not shown). In the case

of the affiliation to pre-Roman tribes BATWING presented three relatively strong candidate trees representing 0.405 of the total number of reported samples. Among these three trees, the modal tree represented 0.389 of the samples, the second represented 0.326, and the weakest signal showed 0.285. In all trees, Aquitani and Berones where the first to split from the rest of the populations. The topology of the modal tree and the TMRCA estimates and their 95% confidence intervals (CI) obtained for pre-Roman tribes are depicted in Supplementary Figure 1. While the modal tree was represented as (((3,((6,5),4))2)1) (following the nomenclature in Supplementary Figure 1), the next two in order showed (((((3,6),5),4)2)1), and ((((3,4),(6,5))2)1), indicating that the Vascones and Autrigones splits likely should be considered polytomous. The time-order specific notation BATWING employs makes construction of mixed state time estimates in polytomous situations difficult. BATWING estimated the effective size of the ancestral population to be 12,700 (95%CI 7,600 – 24,400). BATWING modeled population growth starting at 10,900 years ago (95%CI: 6,500 – 17,800), with a rate of 0.000278 (95%CI 0.00016 – 0.00046) per year. Our analyses suggest that a common ancestral population started to split around 4,520 years ago (95%CI: 3,060 – 6,940), proceeding in successive splits into the six different tribes described by Romans upon their arrival into the region 2,000 years ago. These splits were computed to be 3,730 (2,350 – 5,810), 2,950 (1,860 – 4,580), 2,470 (1,420 - 4,000), and 1,620 (650 – 2,900) years ago for the modal tree. Since BATWING does not take into account migration, thus, some admixture would likely bias separation time estimates towards more recent dates.

STR variation within the most common haplogroups in the geographical area studied suggests that the variation found in the studied populations date back to Neolithic times, after the Younger Dryas, suggesting that current diversity in this region

must descend from few common ancestors from this time period (Table 2). Since the STR mutation rate used (Zhivotovsky et al. 2004) is more conservative than other higher mutation rate estimates (Goedbloed et al. 2009), the dating of demographic increase could indeed be more recent.

Mitochondrial DNA perspective.

The 881 individuals analyzed for mtDNA were classified into 48 different haplogroups and sub-haplogroups. Haplogroup frequencies are given in Supplementary Table S3b and diversity indices are shown in Supplementary Table S4. For the main haplogroups, the frequencies found in Basques, Spanish and French are consistent with previous works either for Basques or other regions in Europe (Achilli et al. 2004; Alfonso-Sanchez et al. 2008) except for J (Bertranpetit et al. 1995; Corte-Real et al. 1996; Garcia et al. 2011).

The frequency of haplogroup U8a in French and Basques, although low, is noteworthy because it is not present in any of the Spanish samples. This rare haplogroup is considered a remnant of the Upper Palaeolithic (Gonzalez et al. 2006), and reaches its maximum (8%) in the French region of Var (Dubut et al. 2004). It has previously been observed in Basques (Gonzalez et al. 2006), however it was absent in other studies on the Basque population (e.g. Alfonso-Sanchez et al. 2008). Given its frequency in the populations studied, the detection of this lineage must be related to random sampling bias.

In general, slightly lower levels of diversity were observed among Basque speaking populations compared to French or Spanish speakers in agreement with our Y chromosome results, with some exceptions (Supplementary Table S4). The Basque ALA shows high levels whereas Spanish NAR and BOC show levels similar to the

Basques. However, haplogroup diversities among French, Basques and Spanish were not significantly different (p=0.141, F=2.42, df=2). F_{ST} genetic distances between populations based on haplogroups were not significant with minor exceptions (Supplementary Tables S5a and S5b). In contrast with what has been found among the Franco-Cantabrian populations, including Basques, in a recent study (Garcia et al. 2011), F_{ST} pairwise distances after Bonferroni correction among the studied populations did not show any significant micro-structuring, again, with few exceptions. The correlation of geographic and genetic distances was significant in the case of mtDNA haplogroups (p=0.042), but not with regards to distances based in the mean number of pairwise differences between sequences (p=0.237). Similar to our Y-chromosome data, PCA based on haplogroup frequencies show that BUR, RIO and CAN Spanish populations appear in a peripheral position (Figure 3a). Additionally BIG, NAR and NLA are also peripheral for mtDNA, whereas ALA is located among the Basques in contrast with what is seen in the Y-chromosome. When additional information from Spanish and French samples were included in the PCA, Basques cluster together including some Spanish and French samples, whereas the rest of the French and Spanish samples seem to be slightly differentiated from the Basque group (Figure 3b).

In agreement with the Y-chromosome results, the AMOVAs performed on the mtDNA control region in the present sample set show some level of differentiation among groups of populations when grouping by ethnic affiliation, French-Basques-Spanish, although the administrative borders do not explain the distribution of the mtDNA diversity. Interestingly, and in agreement with Y-chromosome coalescence-based results, when populations were grouped according to their pre-Roman tribal affiliation, a significant percentage of the mtDNA haplogroup variance could be observed among samples.

The TMRCA estimates for some of the most common haplogroups were 12.8 kya for H1 and 14.2 kya for H3 (Table 2b). Western Eurasian age estimates for both haplogroups oscillate between 13-14 kya (non-corrected mutation rate), and 9 kya and 11 kya (corrected mutation rate) in the case of H1 and between 9 kya and 11.8 kya (for both non-corrected and corrected mutation rates) in the case of H3 (Achilli et al. 2004; Pereira et al. 2005; Soares et al. 2009; Soares et al. 2010). Our older estimates for H1 and H3 would support the hypothesis that those lineages increase their frequency in the Franco Cantabrian refugium and that they subsequently expanded during the Magdalenian period that began around 15 kya (Soares et al. 2010).

Discussion

The present analysis is the first comprehensive genetic study performed at a micro-geographical and dialectal level for all administrative regions of the present Basque country, including the surrounding areas where Basque was previously spoken. Our previous genetic knowledge of Basques at a molecular level, was limited by studies based on small numbers of individuals, ethno-linguistically poorly defined "Basque" samples, or single Basque populations not accounting for neighboring populations (Bertranpetit, Cavallisforza 1991; Comas et al. 1998b; Manzano et al. 2002; Iriondo, Barbero, Manzano 2003; Laayouni, Calafell, Bertranpetit 2010; Rodriguez-Ezpeleta et al. 2010). The results obtained here clearly highlight the importance of the comprehensive sampling coverage in order to understand the patterns and processes of the distribution of diversity in the geographic area under study.

The hypothesis of genetic distinctiveness of Basques within Western Europe has been based on contradictory results from classical markers (Chalmers, Wikin, Mourant 1949; Bertranpetit, Cavallisforza 1991; Calafell, Bertranpetit 1994) uniparental lineages (Bertranpetit et al. 1995; Achilli et al. 2004; Alonso et al. 2005; Adams et al. 2008) and autosomal SNPs (Li et al. 2008; Laayouni, Calafell, Bertranpetit 2010; Rodriguez-Ezpeleta et al. 2010). Our data show that Basques present uniparental lineage gene pools similar to other Western European populations although they show slight differences in their frequencies (Supplementary Table S3). When the Basque haplogroup diversity is placed in the framework of the surrounding populations, the PCA obtained (Figures 2a, 3a) together with previous knowledge of the haplogroup distribution in Western Europe and North Africa, and the detailed knowledge of the recent history in the Iberian Peninsula suggest that all populations share a common ancestral genetic pool, but that populations have been affected by external influences to

different extents (Plaza et al. 2003; Alonso et al. 2005; Adams et al. 2008)). This external influence can be seen in the case of non-Basque BUR, RIO and CAN samples for both uniparental genomes and additionally for NAR, BIG and the Basque NLA in the case of mtDNA. Interestingly, regions that are not considered Basque-speaking, such as BOC and NAR, cluster with the Basques as do the three French populations, indicating that they have not experienced considerable external influence. When other Iberian and French samples are compared to our present sample set (Figures 2b and 3b), Basque samples cluster with other surrounding non-Basque speaking populations, which suggests a genetic distinctiveness, not exclusive to Basque speakers, of the populations inhabiting this geographical area. Moreover the geographically distant population from the French Bretagne (BRI), which shows no North African haplogroups and very little Neolithic influences, falls within our Basque populations for the Y-chromosome data, whereas geographically closer French populations do not. Bretons speak a Celtic language with roots in the British Isles and that has no relation with Basque. This suggests that other geographically and ethnically separated Western European populations might exhibit the genetic composition similar to the Basques and some surrounding populations but that this peculiarity is not linked to the fact of having a Basque culture.

Basque-speaking populations show lower levels of diversity in their Y-chromosome (haplogroup diversity and mean number of pairwise differences) compared to their surrounding populations. Likewise, low levels of diversity have been observed in some populations along the Pyrenees (Lopez-Parra et al. 2009). Levels of consanguinity in the Basque country, especially in Gipuzkoa, have been shown to be very high, particularly in rural areas (Alfonso-Sanchez et al. 2005). This could have contributed to the low levels of diversity in this area in addition to the reduced external

gene flow, and is in accordance with demographic isolation proposed for these and other Basque populations due to geographic or cultural reasons (e.g. Lopez-Parra et al. 2009). This suggested demographic isolation might have yielded some genetic heterogeneity among Basque samples with respect to non-Basques as shown in the Y chromosome but not in mtDNA lineages (as seen in networks of haplotypes, data not shown), which might suggest a patrilocal pattern in the area. These opposite results in paternal and maternal lineages would explain, at least partially, the contradictory results supporting (Manzano et al. 2002; Iriondo, Barbero, Manzano 2003; Perez-Miranda et al. 2005; Alfonso-Sanchez et al. 2008) and rejecting (Comas et al. 1998a; Rodriguez-Ezpeleta et al. 2010) the genetic heterogeneity in Basques.

Our results show that there is no significant genetic micro-structure related to all the sub-dialects currently spoken in *Euskal Herria* (but see below). However, some intrinsic level of genetic structure is present within the Basques that may be the consequence of different cultural, geographical and historical factors. Our results suggest that, in addition to the sex-biased influence of geography in the distribution of genetic diversity, the factor that seems to best explain the genetic structure in the region is the affiliation to pre-Roman tribes. In contrast to the lack of a clear genetic structure shown in the PCA, MDS, and lineage networks (data not shown), our coalescent based results indicate the existence of barriers to male gene flow in historical times, which seems to be consistent with the territories of the different pre-Roman tribes. Given BATWING behavior in the case of recent admixture (see Supplementary Material) and that the confidence intervals obtained do not include the present; our results reveal some degree of isolation for an important period of time. Our analyses suggest that a common population started to split some 4.4 kya ago, with the *Aquitani* forming the deepest split. The formation of the tribes during the Bronze Age and establishment in the Iron Age is

in agreement with previous archaeological and anthropological research (Almagro Gorbea 2005). This time depth might be taken as an underestimation meaning that real estimates would be older given admixture, as it has been shown that even modest migration among populations can reduce the divergence times estimates (Haber et al. 2010). Although some authors consider it unlikely that structure could have resulted from drift during the Neolithic demographic expansion (Iriondo, Barbero, Manzano 2003), it is plausible that the formation of the tribes indeed could be a response to the demographic and economic growth in the Chalcolithic (6-4 ky BP) associated with an increase in social complexity (Almagro Gorbea 2005). Consequently, high levels of endogamy within tribes could have been one of the main factors driving drift.

How could have this signal of ancestral structure been maintained up to present? To some extent, there is a correlation between the geographical localization of pre-Roman tribal groups and the five dialects described by Zuazo (1998; 2010), although there is not total agreement on this point (see Zuazo 2010). This correlation was interpreted as an indicator that dialects dated back from pre-Roman times (Caro Baroja 1943), but other hypotheses have been put forward, e.g., the distribution of dialects is more related to the belonging to ecclesiastic dioceses than to historical tribal ascription (Mujika 1914-1917; Caro Baroja 1943; Caro Baroja 1945). At present there is a strong agreement among Basque specialists that historical dialects are issued from a common Basque language, that probably existed 14 or 15 centuries ago and thus both hypotheses have been abandoned (Michelena 1985; Trask 1997; Zuazo 2010). However, the similar structure for both tribal genetic heterogeneity and Basque dialect distribution indicates similar barriers to flow and would suggest that the same processes are resulting in these boundaries. Despite the geographical proximity, some environmental diversity may have played a strong role in cultural differentiation and ethnogenesis (Almagro Gorbea

2005) and could be a source of genetic structure even in the present.

Conclusions

The genetic similarity of uniparental genomes of Basque to some surrounding and even distant non-Basque populations suggests that the so-called genetic uniqueness of the Basques is the result of a lesser external influence and low gene flow from recent migrations with respect to other surrounding populations. Still, this lower external genetic influence in Basque populations could make them good candidates to represent the genetic profile of the older European populations, although this might also be the case for other non-Basque populations, such as the French Bretons. Studies based on Y-chromosome sequences, entire mtDNA genomes or even whole human genomes may unmask genetic particularities of the Basque population that distinguish them from their western European neighbors. Interestingly, our genetic results clearly correlated with a pre-Roman tribal genetic structure in this area that may be related to environmental diversity and which is still retrievable in the current population. Some authors think that this structure also correlates with the dialectal structure that subsequently appeared. It has been suggested that ancient patterns of organization around the main rivers and their tributaries could be causing the genetic structure and that the persistence of it may be related to a determined type of organization in small population units (Iriondo, Barbero, Manzano 2003). This would suggest that both genetic and dialectal structures might have been driven by the same environmental factors that are still acting at present.

Acknowledgements

We are indebted to all the people contributing samples to this study. We are very grateful to David Basterot, Tristan Carrère, Mònica Vallés, Roger Anglada, and Stéphanie Plaza for technical support and to Isabel Mendizabal, Urko M Marigorta, Laura Rodriguez-Botigué, Michael Ducore, Koldo Zuazo and Francisco Villar for fruitful discussion. Francesc Calafell kindly provided the mitochondrial DNA database from the literature. We thank CESGA (Supercomputational Centre of Galicia) where part of the computational analyses were performed and to Oleg Balanovsky for the translation of mutational data from literature into sequences done with the MURKA software. This work was supported by the HIPVAL (*Histoire des populations et variation linguistique dans les Pyrénées de l'Ouest*) project. The HIPVAL project was made possible by grants from the *Conseil Régional d'Aquitaine*, the *Conseil Général des Pyrénées-Atlantiques*, the *Conseil des Elus du Pays-Basque*, the CNRS (interdisciplinary programme OHLL (*Origine de l'Homme, des Langues et du Langage*), and Association *Sang 64*. We thank all the volunteers who kindly accepted to participate in the study. Authors are also grateful to many contributors for their valuable help with the recruitment of individuals especially Estibaliz Montoya.

The Genographic Consortium includes: Janet S. Ziegle (Applied Biosystems, Foster City, California, United States); Li Jin & Shilin Li (Fudan University, Shanghai, China); Pandikumar Swamikrishnan (IBM, Somers, New York, United States); Asif Javed, Laxmi Parida & Ajay K. Royyuru (IBM, Yorktown Heights, New York, United States); R. John Mitchell (La Trobe University, Melbourne, Victoria, Australia); Pierre A. Zalloua (Lebanese American University, Chouran, Beirut, Lebanon); Syama Adhikarla, ArunKumar GaneshPrasad, Ramasamy Pitchappan & Arun Varatharajan Santhakumari (Madurai Kamaraj University, Madurai, Tamil Nadu, India); R. Spencer Wells

(National Geographic Society, Washington, District of Columbia, United States); Angela Hobbs & Himla Soodyall (National Health Laboratory Service, Johannesburg, South Africa); Elena Balanovska & Oleg Balanovsky (Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia); Chris Tyler-Smith (The Wellcome Trust Sanger Institute, Hinxton, United Kingdom); Daniela R. Lacerda & Fabrício R. Santos (Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil); Pedro Paulo Vieira (Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil); Jaume Bertranpetit, Marc Haber & Marta Melé (Universitat Pompeu Fabra, Barcelona, Spain); Christina J. Adler, Alan Cooper, Clio S. I. Der Sarkissian & Wolfgang Haak (University of Adelaide, South Australia, Australia); Matthew E. Kaplan & Nirav C. Merchant (University of Arizona, Tucson, Arizona, United States); Colin Renfrew (University of Cambridge, Cambridge, United Kingdom); Andrew C. Clarke & Elizabeth A. Matisoo-Smith (University of Otago, Dunedin, New Zealand); Matthew C. Dulik, Jill B. Gaieski, Amanda C. Owings, Theodore G. Schurr & Miguel G. Vilar (University of Pennsylvania, Philadelphia, Pennsylvania, United States).

Tables

Table 1. Analyses of the molecular variance (AMOVAs). Apportionment of the variance in %

Populations		Y-chromosome		Mitochondrial DNA	
		STR	**Haplogroup**	**Sequence**	**Haplogroup**
French / Spanish / Basques	Among groups	0.86 *	0.24	0.46	0.53*
	Within groups	0.84 *	2.48***	1.29 ***	1.11***
	Within populations	98.31 ***	97.29***	98.25 ***	98.53***
Administrative border France-Spain	Among groups	-0.09	0.43	0.06	0.38
	Within groups	1.36 ***	2.39***	1.51 ***	1.21***
	Within populations	98.73 ***	97.18***	98.43 ***	98.42***
Pre-roman Tribes	Among groups	0.87	0.86	0.58	0.97**
	Within groups	0.58	2.26 **	0.92 *	0.55
	Within populations	98.55 **	96.88 ***	98.50 ***	98.49 ***

All values have been corrected for multiple testing (STR/sequences and haplogroups) using Bonferroni; *p<0.025, **p<0.005, ***p<0.0005

Table 2. Ages (in Ky) of STR variation of the most common haplogroups in the present sample set.

Y chromosome haplogroup	**Age ± SE (N)**
I-M26	7.6 ± 1.8 (57)
R1b1b2-M269	10.7 ± 2.4 (677)
R1b1b2-MP312	10.4 ± 1.8 (337)
R1b1b2-M153	6.1 ± 1.3 (109)
R1b1b2-SRY2627	8.4 ± 1.7 (67)
R1b1b2-L21	7.5 ± 1.7 (115)
MtDNA haplogroup	**Age ± SE (N)**
H	15.0 ± 0.13 (409)
H1	12.8 ± 0.23 (215)
H3	14.2 ± 0.45 (77)
U5b	29.9 ± 1.14 (121)
V	16.39 ± 0.65 (58)

SE standard error; N number of individuals

Figure legends

Figure 1. Map showing the location of the populations used in the present study. Color of the circles indicates the linguistic affiliation (Red- Spanish, Green- Basque, Blue- French) and color shades represent the five main Basque dialects spoken at present (Zuazo, 2010). Dotted gray line delimitates the area of the pre-roman tribes indicated in gray letters. BIG, Bigorre; BEA, Béarn; CHA, Chalosse; ZMI, Lapurdi/Baztan; NLA, Lapurdi Nafarroa; SOU, Zuberoa; RON, Roncal and Salazar valleys; NCO, Central Western Nafarroa; NNO, North Western Nafarroa; GUI, Gipuzkoa; GSO, South Western Gipuzkoa; ALA, Araba; BBA, Bizkaia; BOC, Western Bizkaia; CAN, Cantabria; BUR, Burgos; RIO, La Rioja; NAR, North Aragon.

Figure 2. Principal Component Analysis based on the Y-haplogroup frequency data in A) 18 populations typed in the present study. Names are as in Figure 1. B) 18 populations from the present study plus 15 Spanish populations from (Adams et al. 2008) and seven populations from Ramos-Luis et al. (Ramos-Luis et al. 2009).
Populations have been colored according to ethnic affiliation. Populations from the present study in bold; populations from the literature in italics. Names of the populations from the present study are as in Figure 1. AST, Asturias; ARA, Aragón; GAL, Galicia; GAS, Gascony; BAS, Basques; CAT, Catalonia; NEC, North East Castille; NWC, North West Castille; WAN, Western Andalusia; EAN, Eastern Andalusia; EXT, Extremadura; CLM, Castille La Mancha; VAL, Valencia; NPO, North Portugal; SPO, South Portugal; ALS, Alsace; AUV, Auvergne; BRI, Bretagne; NPC, Nord-Pas-de-Calais; IDF, Île de France; MPY, Midi-Pyrennés; PAS, Provence-Alpes-Côte d'Azur.

Figure 3. Principal Component Analysis based on the mtDNA-haplogroup frequency data in A) 18 populations typed in the present study. B) 18 populations from the present study plus 12 Iberian and 10 French additional populations.

Populations have been colored according to ethnic affiliation. Populations from the present study in bold; populations from the literature in italics. Names of the populations from the present study are as in Figure 1. GAL, Galice; LEO, León; CANT, Cantabria; BAS, Basques; ARA, Aragón; CAT, Catalonia; VAL, Valencia; CSP, Central Spain; AND, Andalusia; NPO, North Portugal; CPO, Central Portugal; SPO, South Portugal; BEAR, Béarn; POI, Poitou; LAN, Languedoc; VAR, Var; LYO, Lyon; PER, Périgord; BRI, Brittany; MAI, Maine; NOR, Normandie; PIC, Picardie.

References

Achilli A, Rengo C, Magri C, et al. (13 co-authors). 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, et al. (20 co-authors). 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.

Alfonso-Sanchez MA, Aresti U, Pena JA, Calderon R. 2005. Inbreeding levels and consanguinity structure in the Basque province of Guipuzcoa (1862-1980). Am. J. Phys. Anthropol. 127:240-252.

Alfonso-Sanchez MA, Cardoso S, Martinez-Bouzas C, Pena JA, Herrera RJ, Castro A, Fernandez-Fernandez I, De Pancorbo MM. 2008. Mitochondrial DNA haplogroup diversity in Basques: A reassessment based on HVI and HVII polymorphisms. Am. J. Hum. Biol. 20:154-164.

Almagro Gorbea M. 2005. Etnogénesis del País Vasco: de los antiguos mitos a la investigación actual. Munibe: Antropologia - Arkeologia 57:345-364.

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. European Journal of Human Genetics 13:1293-1302.

Alzualde A, Izagirre N, Alonso S, Alonso A, de la Rua C. 2005. Temporal mitochondrial DNA variation in the Basque Country: Influence of post-Neolithic events. Ann. Hum. Genet. 69:665-679.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 23:147.

Bauduer F, Feingold J, Lacombe D. 2005. The Basques: Review of population genetics and Mendelian disorders. Hum. Biol. 77:619-637.

Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, Comas D, Mitchell RJ, Quintana-Murci L, Tyler-Smith C, Wells RS. 2007. The genographic project public participation mitochondrial DNA database. Plos Genet. 3:1083-1095.

Bertranpetit J, Cavallisforza 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.

Calafell F, Bertranpetit J. 1994. Principal Component Analysis of Gene-Frequencies and the Origin of Basques. Am. J. Phys. Anthropol. 93:201-215.

Campell GL. 1998. Concise compendium of the world's language. London and New York: Routledge.

Capelli C, Redhead N, Romano V, et al. 2006. Population structure in the Mediterranean basin: A Y chromosome perspective. Ann. Hum. Genet. 70:207-225.

Caro Baroja J. 1943. Los pueblos del norte de la península Ibérica. Madrid. Reprinted in Caro Baroja 1985, 121-240: Instituto Bernardino de Sahagún.

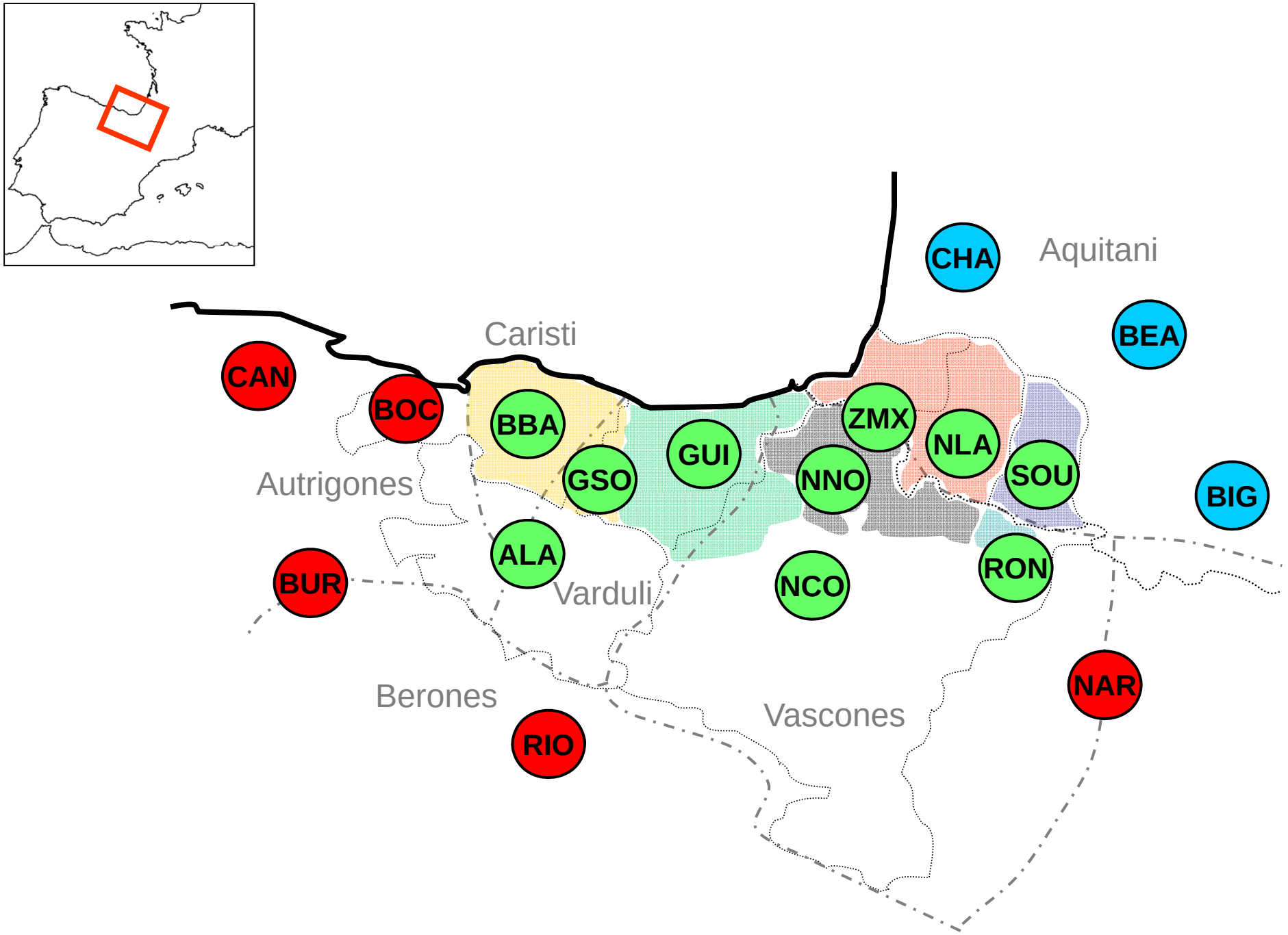
Caro Baroja J. 1945. Materiales para una historia de la lengua vasca en su relación con la latina. Salamanca: Universidad de Salamanca.

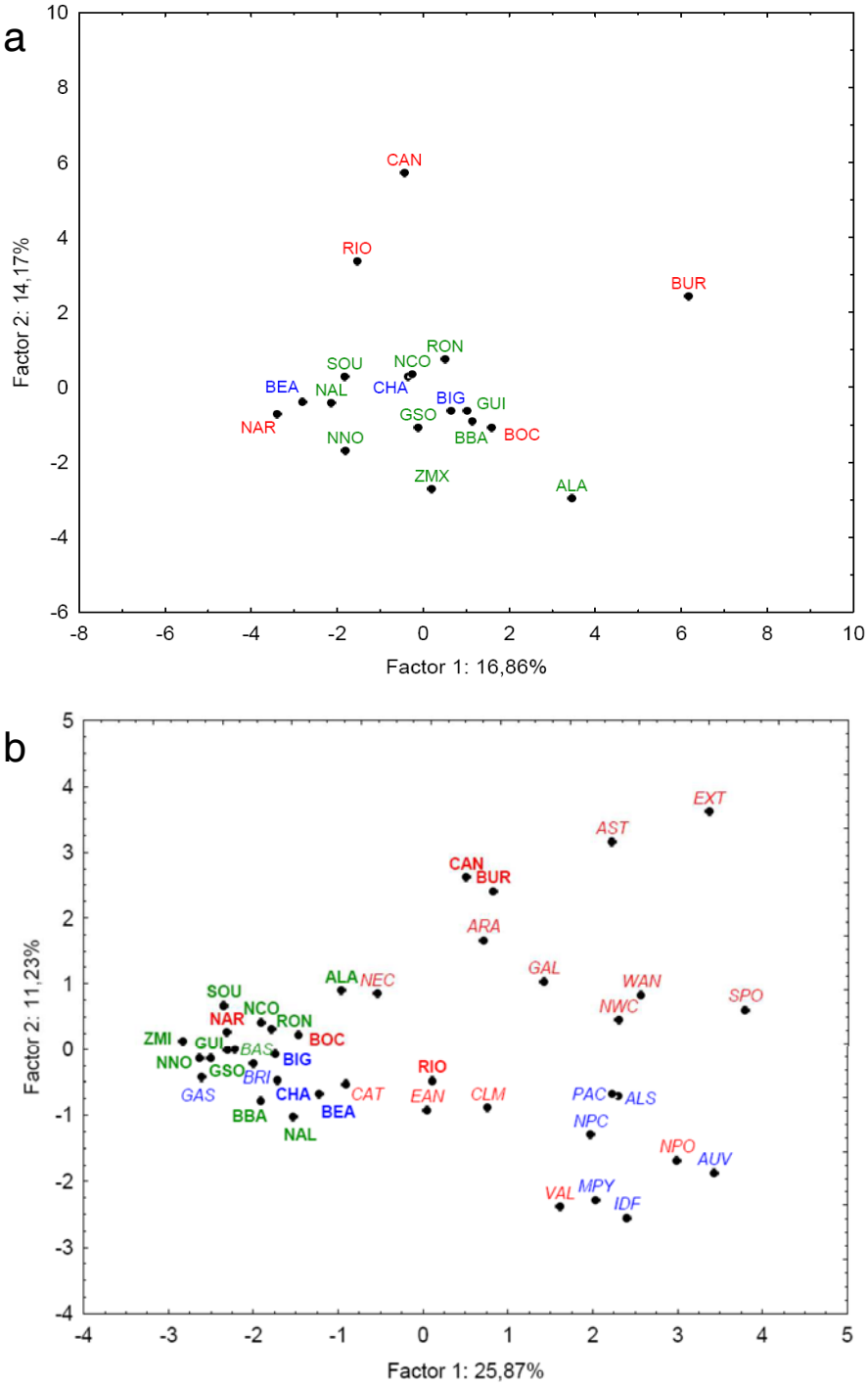
Cavalli-Sforza L, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton: Princeton University Press.
Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bertranpetit J. 1998a. HLA evidence for the lack of genetic heterogeneity in Basques. Ann. Hum. Genet. 62:123-132.
Comas D, Mateu E, Calafell F, Perez-Lezaun A, Bosch E, Martinez-Arias R, Bertranpetit J. 1998b. HLA class I and class II DNA typing and the origin of Basques. Tissue Antigens 51:30-40.
Comrie B, Matthews S, Polinsky M. 2003. The Atlas of Languages. The origin and development of languages throughout the world. New York: Facts on File.
Contu D, Morelli L, Santoni F, Foster JW, Francalacci P, Cucca F. 2008. Y-Chromosome Based Evidence for Pre-Neolithic Origin of the Genetically Homogeneous but Diverse Sardinian Population: Inference for Association Scans. Plos One 3.
Corte-Real HBSM, Macaulay VA, Richards MB, Hariti G, Issad MS, CambonThomsen 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.
Chalmers JNM, Wikin E, Mourant AE. 1949. The ABO, MN and Rh blood groups of the Basque people. Am. J. Phys. Anthropol. 7:545-548.
Chikhi L, Nichols RA, Barbujani G, Beaumont MA. 2002. Y genetic data support the Neolithic demic diffusion model. P. Natl. Acad. Sci. USA 99:11008-11013.
de la Rúa C. 1992. Craniofacial factors in the Basque skull. A comparative study. Homo 43:135-161.
Dubut V, Chollet L, Murail P, Cartault F, Beraud-Colomb E, Serre M, Mogentale-Profizi N. 2004. mtDNA polymorphisms in five French groups: importance of regional sampling. European Journal of Human Genetics 12:293-300.
EuskoJaurlaritza. 2008. Fourth Sociolinguistic Survey 2006. Vitoria-Gasteiz, http://www.euskara.euskadi.net/r59738/en/contenidos/informacion/inkesta_sozi olinguistikoa2006/en_survey/adjuntos/IV_incuesta_en.pdf: Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia.
Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evol. Bioinformatics 1:47-50.
Francalacci P, Morelli L, Useli A, Sanna D. 2010. The History and Geography of the Y Chromosome SNPs in Europe: an update. J. Anthropol. Sci. 88:207-214.
Francalacci P, Sanna D. 2008. History and geography of human Y-chromosome in Europe: a SNP perspective. J. Anthropol. Sci. 86:59-89.
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. European Journal of Human Genetics 17:1490-1494.
Garcia O, Fregel R, Larruga JM, Alvarez V, Yurrebaso I, Cabrera VM, Gonzalez AM. 2011. Using mitochondrial DNA to test the hypothesis of a European post-glacial human recolonization from the Franco-Cantabrian refuge. Heredity 106:37-45.
Goedbloed M, Vermeulen M, Fang RXN, et al. 2009. Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR (R) Yfiler (R) PCR amplification kit. Int. J. Legal Med. 123:471-482.
Gonzalez AM, Garcia O, Larruga JM, Cabrera VM. 2006. The mitochondrial lineage U8a reveals a Paleolithic settlement in the Basque country. Bmc Genomics 7.

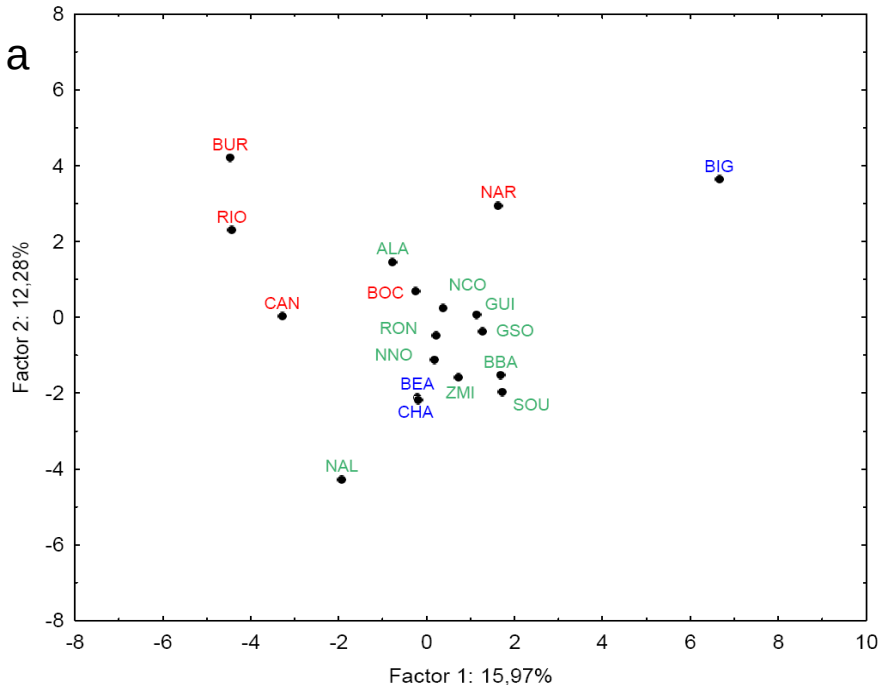
Gorrochategui J. 1984. Onomástica indígena de Aquitania. Bilbao: Euskal Herriko Unibertsitatea-Universitad del País Vasco.
Gorrochategui J. 1995. Basque and its neighbors in Antiquity. In: JI Hualde, JA Lakarra, RL Trask, editors. Towards a History of the Basque Language. Amsterdam/Philadelphia: John Benjamins. p. 31-63.
Gorrochategui J. 2007. Onosmástica de origen vasco-aquitano en Hispania y el Imperio Romano. In: M Mayer, G Baratta, A Guzmán, editors. Acta XII Congressus Internationalis Epigraphiae Graecae et latinae. Barcelona. p. 629-634.
Haak W, Balanovsky O, Sanchez JJ, et al. (17 co-authors). 2010. Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. Plos Biol. 8:e1000536.
Haber M, Platt DE, Badro DA, et al. (13 co-authors). 2010. Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon. European Journal of Human Genetics 19:334-340.
Iriondo M, Barbero MC, Manzano C. 2003. DNA polymorphisms detect ancient barriers to gene flow in Basques. Am. J. Phys. Anthropol. 122:73-84.
Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. 18:830-838.
Laayouni H, Calafell F, Bertranpetit J. 2010. A genome-wide survey does not show the genetic distinctiveness of Basques. Human Genetics 127:455-458.
Larruga JM, Diez F, Pinto FM, Flores C, Gonzalez AM. 2001. Mitochondrial DNA characterisation of European isolates: The Maragatos from Spain. European Journal of Human Genetics 9:708-716.
Li JZ, Absher DM, Tang H, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.
Lopez-Parra AM, Gusmao L, Tavares L, Baeza C, Amorim A, Mesa MS, Prata MJ, Arroyo-Pardo E. 2009. In search of the Pre- and Post-Neolithic Genetic Substrates in Iberia: Evidence from Y-Chromosome in Pyrenean Populations. Ann. Hum. Genet. 73:42-53.
López-Soto M, Sanz P. 2000. Polimorfismos de ADN mitocondrial en individuos residentes en Andalucía y Extremadura. Cuadernos de Medicina Forense 20:17-24.
Lucotte G, Hazout S. 1996. Y-chromosome DNA haplotypes in Basques. Journal of Molecular Evolution 42:472-475.
Luchaire A. 1875/1877. Les origines linguistiques de l'Aquitaine. Bulletin de la Société des Sciences, Lettres et Arts. Pau. p. 349-423.
Maca-Meyer N, Sanchez-Velasco P, Flores C, Larruga JM, Gonzalez AM, Oterino A, Leyva-Cobian F. 2003. Y chromosome and mitochondrial DNA characterization of Pasiegos, a human isolate from Cantabria (Spain). Ann. Hum. Genet. 67:329-339.
Manzano C, de la Rua C, Iriondo M, Mazon LI, Vicario A, Aguirre A. 2002. Structuring the genetic heterogeneity of the Basque population: A view from classical polymorphisms. Hum. Biol. 74:51-74.
Martinez-Cruz B, Ziegle J, Sanz P, Sotelo G, Anglada R, Plaza S, Comas D. 2011. Multiplex screening of the human Y chromosome using TaqMan probes. Investigative Genet. 2.
Martinez-Jarreta B, Prades A, Calafell F, Budowle B. 2000. Mitochondrial DNA HVI and HVII variation in a North-East Spanish population. J. Forensic Sci. 45:1162-1163.

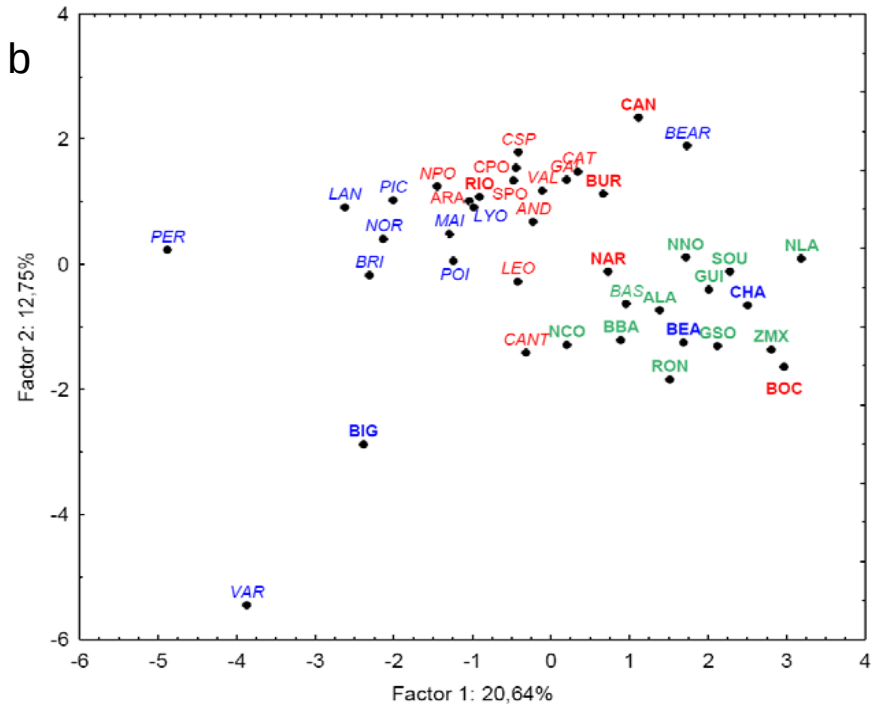
Michelena L. 1961/1962. Los nombres indígenas de la inscripción hispano-romana de Lerga. Principe de Viana. p. 65-74.
Michelena L. 1976. Lenguas indígenas y lengua clásica en Hispania. Travaux du VIè Congrès International d'Etudes Classqiues. Bucarest-Paris. p. 41-51.
Michelena L. 1985. Lengua e Historia. Madrid: Paraninfo.
Morral N, Llevadot R, Casals T, Gasparini P, Macek M, Dork T, Estivill X. 1994. Independent origins of cystic-fibrosis mutation R334W, R347P, R1162X, and 3849+10BC-]T provide evidence of mutation recurrence in the CFTR gene. Am. J. Hum. Geneti 55: 890-898.
Mujika S. 1914-1917. El obispado de Bayona con relación a los pueblos de Guipúzcoa adscritos a dicha diócesis. Reprinted in Revista internacional de estudios vascos, 8:198-229. Bilbao: La gran enciclopedia vasca 1969.
Passarino G, Underhill PA, Cavalli-Sforza LL, et al. (12 co-authors). 2001. Y chromosome binary markers to study the high prevalence of males in Sardinian centenarians and the genetic structure of the Sardinian population. Hum. Hered. 52:136-139.
Pereira L, Prata MJ, Amorim A. 2000. Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. Ann. Hum. Genet. 64:491-506.
Pereira L, Richards M, Goios A, et al. (13 co-authors). 2005. High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res. 15:19-24.
Perez-Miranda AM, Alfonso-Sanchez MA, Kalantar A, Garcia-Obregon S, de Pancorbo MM, Pena JA, Herrera RJ. 2005. Microsatellite data support subpopulation structuring among Basques. Journal of Human Genetics 50:403-414.
Picornell A, Gomez-Barbeito L, Tomas C, Castro JA, Ramon MM. 2005. Mitochondrial DNA HVRI variation in Balearic populations. Am. J. Phys. Anthropol. 128:119-130.
Plaza S, Calafell F, Helal A, Bouzerna N, Lefranc G, Bertranpetit J, Comas D. 2003. Joining the pillars of hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. Ann. Hum. Genet. 67:312-328.
Ramos-Luis E, Blanco-Verea A, Brion M, Van Huffel V, Carracedo A, Sanchez-Diz P. 2009. Phylogeography of French male lineages. Forensic Sci. Int.l: Genetics Supplement Series 2:439-441.
Raymond M, Rousset F. 1995. Genepop version 1.2: population genetics software for exact tests and ecumenicism. J. Hered. 86:248-249.
Richard C, Pennarun E, Kivisild T, et al. (19 co-authors). 2007. An mtDNA perspective of French genetic variation. Ann. Hum. Biol. 34:68-79.
Rodriguez-Ezpeleta N, Alvarez-Busto J, Imaz L, Regueiro M, Azcarate MN, Bilbao R, Iriondo M, Gil A, Estonba A, Aransay AM. 2010. High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations. Human Genetics 128:113-117.
Rohlfs G. 1935. Le Gascon: études de philologie pyrénéenne. Zeitschrift für Romanische Philologie Beiheft Tübingen and Pau.
Ruhlen M. 2001. Dene-Caucasian: a new linguistic family. Pluriverso 2:76-85.
Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo A. 1998. MtDNA analysis of the Galician population: A genetic edge of European variation. European Journal of Human Genetics 6:365-375.
Semino O, Magri C, Benuzzi G, et al. (16 co-authors). 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: Inferences on the

neolithization of Europe and later migratory events in the Mediterranean area. Am. J. Hum. Genet. 74:1023-1034.
Soares P, Achilli A, Semino O, Davies W, Macaulays V, Bandelt HJ, Torroni A, Richards MB. 2010. The Archaeogenetics of Europe. Curr. Biol. 20:R174-R183.
Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. Am. J. Hum. Genet. 84:740-759.
Trask RL. 1997. History of Basque. London and New York: Routledge.
Villar F, Prosper BM. 2005. Vascos, Celtas e Indoeuropeos: genes y lenguas. Salamanca: Ediciones Universidad de Salamanca.
Wilson IJ, Weale ME, Balding DJ. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. Roy. Stat. Soc. A Sta. 166:155-188.
Xue YL, Zejal T, Bao WD, et al. (12 co-authors). 2006. Male demography in East Asia: A north-south contrast in human population expansion times. Genetics 172:2431-2439.
Zhivotovsky LA, Underhill PA, Cinnioglu C, et al. (17 co-authors). 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am. J. Hum. Genet. 74:50-61.
Zuazo K. 1998. Euskalkiak, gaur. Fontes Linguae Vasconum. Pamplona: Institución Príncipe de Viana. p. 191-233.
Zuazo K. 2010. El Euskera y sus dialectos. Irún: Alberdania.


CHA
Aquitani
BEA
Caristi
CAN
BOC
BBA
GUI
ZMX
NLA
SOU
GSO
NNO
Autrigones
BIG
ALA
BUR
Varduli
NCO
RON
NAR
Berones
Vascones
RIO


a
Factor 2: 14,17%
Factor 1: 16,86%
CAN
RIO
BUR
RON
NCO
SOU
BEA
NAL
CHA
BIG
GUI
GSO
NAR
BBA
BOC
NNO
ZMX
ALA
b
Factor 2: 11,23%
Factor 1: 25,87%
EXT
AST
CAN
BUR
ARA
GAL
WAN
SPO
ALA
NEC
SOU
NCO
NAR
RON
NWC
ZMI
GUI
BAS
BOC
NNO
BIG
GSO
BRI
RIO
GAS
CHA
CAT
EAN
CLM
PAC
ALS
BBA
BEA
NPC
NAL
NPO
AUV
VAL
MPY
IDF


a
BUR
RIO
CAN
NAR
BIG
ALA
BOC
NCO
GUI
RON
GSO
NNO
BBA
BEA
ZMI
CHA
SOU
NAL
Factor 2: 12,28%
Factor 1: 15,97%


b
CAN
BEAR
CSP
CAT
GAR
NPO
CPO
VAL
BUR
RIO
PIC
LAN
ARA
SPO
AND
LYO
MAI
NOR
PER
BRI
POI
LEO
NAR
NNO
SOU
NLA
GUI
BAS
ALA
CHA
CANT
NCO
BBA
BEA
GSO
ZMX
RON
BOC
BIG
VAR
Factor 2: 12,75%
Factor 1: 20,64%