Genetic origins, singularity, and heterogeneity of Basques

André Flores-Bello, Frédéric Bauduer, Jasone Salaberria, Bernard Beñat, B. Oyharçabal, Francesc Calafell, Jaume Bertranpetit, Lluis Quintana-Murci, David Comas

Genetic origins, singularity and heterogeneity of Basques

André Flores-Bello,[1] Frédéric Bauduer,[2] Jasone Salaberria,[3] Bernard Oyharçabal,[3] Francesc Calafell,[1] Jaume Bertranpetit,[1] Lluis Quintana-Murci,[4,5] David Comas[1]*

[1]Departament de Ciències de la Salut i de la Vida, Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, 08003, Spain; [2]Laboratoire PACEA UMR 5199 CNRS, Université de Bordeaux, Pessac, 33615, France; [3]Centre de Recherche sur la Langue et les Textes Basques, Centre National de la Recherche Scientifique, UMR5478 IKER, 64100, Bayonne, France; [4]Human Evolutionary Genetics Unit, Institut Pasteur, Centre National de la Recherche Scientifique, UMR 2000, Paris, 75015, France; [5]Human Genomics and Evolution, Collège de France, Paris, France.

*Correspondence: david.comas@upf.edu

Summary

Basques have historically lived along the Western Pyrenees, in the Franco-Cantabrian region, straddling the current Spanish and French territories. Over the last decades, they have been the focus of intense research due to their singular cultural and biological traits that, with high controversy, placed them as a heterogeneous, isolated, and unique population. Their non-Indo-European language, Euskara, is thought to be a major factor shaping the genetic landscape of the Basques. Yet, there is still a lively debate about their history and assumed singularity due to the limitations of previous studies. Here, we analyze genome-wide data of Basque and surrounding groups that do not speak Euskara at a micro-geographical level. A total of ~629,000 genome-wide variants were analyzed in 1,970 modern and ancient samples, including 190 new individuals from 18 sampling locations in the Basque area. For the first time, local- and wide-scale analyses from genome-wide data have been performed covering the whole Franco-Cantabrian region, combining allele frequency and haplotype-based methods. Our results show a clear differentiation of Basques from the surrounding populations, with the non-Euskara-speaking Franco-Cantabrians located in an intermediate position. Moreover, a sharp genetic heterogeneity within Basques is observed with significant correlation with geography. Finally,

the detected Basque differentiation cannot be attributed to an external origin compared to other Iberian and surrounding populations. Instead, we show that such differentiation results from genetic continuity since the Iron Age, characterized by periods of isolation and lack of recent gene flow that might have been reinforced by the language barrier.

Keywords

Basques, Euskara, Franco-Cantabrian, population structure, isolated population, demographic history, genome-wide data

Introduction

The Franco-Cantabrian region, which includes the western part of the actual border of Spain and France through the Pyrenees, has drawn the attention of several disciplines due to its relevant role in European human history. This region was one of the most densely populated glacial refugia in Europe during the Last Glacial Maximum (LGM), and it is related to pivotal archaeological discoveries, especially the oldest known European cave paintings[1]. One of the most interesting features of the region is the presence of the Basques. They have been historically distributed along the western edge of the Pyrenees, spanning across the present Spanish and the French territories currently organized in seven provinces: Gipuzkoa, Bizkaia, Araba and Nafarroa in the southern side of the Pyrenees; and Zuberoa, Lapurdi, and Nafarroa Beherea located in the northern side. Basques have presumably stood out due to their historical, anthropological, and biological traits that define their singularity and isolation within the European context. A remarkable feature is Euskara, with its five main dialects (Figures 1, S1 and Table 1), which is a non-Indo-European language isolate with no close relationship to any other extant language[2,3]. Beyond its current distribution, Euskara was historically spoken in the seven present provinces, before its geographic regression due to the pressures of Romance languages[4]. Moreover, an archaic Euskara-related language is suggested to have been spoken in a much wider area in earlier periods. This area would include the neighboring Northern Spanish areas and the Southern half of Aquitaine in France[5] (Figure S1). Whereas Euskara has been pointed as a possible cultural barrier between Basques and neighboring populations, its

dialects might have also acted as an internal barrier[6]. They show reduced interintelligibility and a standard language (*batua* in Euskara) was not established until 1968 and was widely used only since the 1980s.

Although numerous studies have focused on the genetics of Basques, a lively debate on their population history is still ongoing (see, for instance, Laayouni et al.[7] and Rodríguez-Ezpeleta et al.[8]). Such an interest in the genetics of Basques started with the remarkable observation of a high frequency of the Rh-negative blood group[9], a genetic variant associated with the hemolytic disease of the newborn, which was confirmed in following studies[10]. Subsequent studies using more genetic markers revealed that Basques were particularly differentiated within the European genetic context[11,12]. These results, together with archaeological, cultural, and linguistic data, were interpreted as Basques descending from an ancient population that remained isolated in the region[13]. However, other studies have not provided evidence of a genetic distinctiveness for the Basque population, suggesting genetic homogeneity through Europe[14].

The origins of Basques have also been controversial. Some studies based on uniparentally-inherited markers have proposed that Basques represent isolated Pre-Neolithic European groups that stayed in the region after the resettlement of Europe from the glacial refugia in the post-LGM periods[15,16]. Conversely, other studies have shown the influence of the Neolithic migrations in the Basque area, refuting the Paleolithic genetic continuity[17]. Exhaustive analysis of uniparental genomes showed a continuity since Pre-Neolithic times and a pre-Roman genetic structure in Basques[6,18]. In support of this, ancient DNA data has suggested that Basques can actually be explained as a common Iberian Iron Age population, with an important genetic influence of the post-Neolithic Steppe pastoralists ancestries, but lacking admixture from subsequent incoming populations such as Romans or North Africans[19].

The controversies about the Basques have not only been focused on their distinctiveness and their origins, but also on their internal genetic heterogeneity. Genome-wide data in Basque groups have shown contradictory results, with some studies suggesting that French Basques

are markedly different from Spanish Basques, being the latter similar to other Iberian populations [7]; whereas other data were interpreted as showing internal homogeneity within Basques and marked genetic differentiation from non-Basque groups[8].

These remarkable contradictory results might be explained by a limited methodology and resolution. The low number of samples used in these analyses to represent the Basque groups and their neighboring areas has supposed the major limiting factor. Furthermore, they have been based on the allele frequencies of classical genetic markers, the phylogeny of uniparental genomes, or a reduced number of sample and markers[13]. To overcome the limitations in the previous studies, we adopt a robust genome-wide study design that enable us to unravel the controversial population history of Basques, including their uniqueness, their origins, and their genetic structure. The unique and exhaustive dataset of the whole Franco-Cantabrian region presented here (Figure 1, S1 and Table 1) limit any possible sampling biases affecting the previous studies and affords an exhaustive analysis at micro-geographical and wide-scale level. Moreover, the ethno-linguistic information considered in our sampling allowed us to interpret the results beyond the genetic data, considering how cultural factors could be involved in shaping the genetics of the Basques and surrounding populations. Finally, we leverage the more precise haplotype-based methods to uncover fine-grained genetic structure and admixture patterns[20,21].

Results

Basques display marked differentiation in the European/Mediterranean landscape

We first studied Basques in a wide-scale context to assess their genetic variability within a large and diverse population panel, which includes a complete dataset of West Eurasia and North African samples. In a Principal Component Analysis (PCA) the Basque samples fall in the opposite edge of the North African samples and in the periphery of Europe, similarly to Sardinians, with the Peri-Basque groups (surrounding traditionally Gascon- and Spanish-speaking areas; see sample collection in STAR★METHODS) being in an intermediate position (Figure 2A). Using admixture analyses, considering the global genetic ancestry components of these populations (Figure 2B), a differential genetic pattern is observed in Basques. In K=6,

Basques present mainly two components: a major component (green), which is also present in European samples and it is found at low frequencies in the Middle East/Caucasus and North Africa; plus, a minor component (pink) found at high frequencies in Central/Eastern Europe. The other components found in the rest of European groups are not present in Basques (frequencies <1%). The Peri-Basque samples show a similar pattern to Basques but with low frequencies of other external components absent in Basques. From K=7 onwards a new specific component appears, maximized in Basques and with frequencies over 50% in Peri-Basques. This component is also observed in Spanish and French samples, while it is virtually absent in the other external European samples.

Using the haplotype-based analyses implemented in fineSTRUCTURE, we detected marked levels of differentiation in Basques (Figures 3, S2 and Table S1). First, the Basque groups cluster together within the large European branch, but in a differentiated external cluster (Figures 3A and S2). This result points to low haplotype sharing between this cluster and the rest of European groups, showing a clear internal and specific genetic profile of Basques. Second, the Peri-Basque groups also exhibit a differentiation from the other external populations, clustering internally with Europeans, but forming a specific branch with the exception of the Cantabrian samples (gCAN), which cluster with Spanish samples (Figures 3A and S2). To discard putative artifacts due to the overrepresentation of Franco-Cantabrian samples in the fineSTRUCTURE analysis, we performed a random sampling of the Franco-Cantabrian region (50 samples including Basques and Peri-Basques), and obtained similar results (data not shown). Furthermore, the ancestry profile calculated in the Non-negative least squares (NNLS) analysis mirrored the results above (Figure 3B). Basques share haplotypes exclusively with the internal groups in the Franco-Cantabrian region. Peri-Basques mainly share haplotypes internally with the groups in the region but also with the non-Franco-Cantabrian Spanish and French groups, acting as a buffer zone between Basques and the surrounding external populations. The intermediate ancestry profile observed in Peri-Basques suggests gene flow between the Franco-Cantabrian region and the external groups. Therefore, potential admixture events were tested in Peri-Basques by using GLOBETROTTER, considering Basques and all the external clusters inferred by fineSTRUCTURE as surrogates (Figures S2D,

S2E and S2F). Single admixture events involving two sources were detected for all Peri-Basque targets, with close dates between the 11th and 16th centuries. Similar sources were described in each target cluster: a major source represented by mainly Basque and Spanish ancestries; and a minor source that is predominantly represented by Spanish ancestry. Moreover, the confidence intervals for the dates estimated from bootstrapping resulted in overlapped and quite large ranges in each Peri-Basque target, especially in Bigorre (gBIG) (Figures S2E and S2F). This might evoke a single, but continuous, pulse of admixture in the Peri-Basques, and a general large-scale demographic event that affected the area in recent historical times, between the 11th and 16th centuries.

To explore further the genetic differentiation of Basques, we performed an analysis of runs of homozygosity (ROHs). Basques show the overall highest total number (NROH) and total length (SROH) of ROHs, even higher than Sardinians, which are reported to carry long ROHs[22] and show ROH values slightly above the European average[23], and followed by the Peri-Basque groups (Figures 3C and S3A). In the intermediate ROH categories, the total proportion of samples represented in the external populations is very small (Figure S3A). This shows that these categories are more common in the isolated groups, Basques and Sardinians, as well as in the Peri-Basques, while in the external groups the values observed could be more related to cryptically inbred outliers[24]. These results are in agreement with the exploration of the proportion of identity-by-descent sharing between samples (PI_HAT) within the groups (Figure S3B). Moreover, the estimation of the effective population size (*Ne*) over time showed Basques with low and stable values, whereas external groups (i.e., Spanish and French) show a dramatic increase around one thousand generations ago (Figure S3C). These results suggest a pattern of isolation together with inbreeding in Basques, and to a lower extent in Peri-Basques. Such patterns of isolation, which by all other evidence is much more recent (see Discussion), seems to have been deep enough to erase the traces of an apparent *Ne* Paleolithic growth in the surrounding populations.

Post-Iron Age demographic processes explain the genetic uniqueness of Basques

Then, analyses including ancient samples were performed to throw light on the origin of the genetic singularity of Basques. The PCA projection of ancient samples shows Basques closer to Pre-Neolithic hunter-gatherers and European Neolithic farmers, but also to some post-Neolithic Steppe herders associated to the Pontic-Caspian Steppe Yamnaya ancestry (Figure S4A). ADMIXTURE analysis (K=4, lowest cross-validation error, Figure S4B) show Basques and Peri-Basques with the lowest proportions of a Levant- and Iran-related Neolithic components, together with a slightly higher proportion of the Anatolian/European farmer component compared to other European populations. When testing for the shared drift with ancient samples, outgroup f3-statistics show high shared drift between Basques and the three major ancient components in Europe (i.e., Paleolithic hunter-gatherers, Neolithic farmers, and post-Neolithic Steppe herders related to the Yamnaya ancestry) (Figure S4C). We then modeled Franco-Cantabrian groups and other European populations with these ancient samples by using qpGraph (Figure S4D). The model fit each tested European population, with Z-scores close to 0 for all the 100 permutations performed. The inferred admixture proportions for the three ancient components do not show differences of Basques compared to the general European context (Figure S4E), following the Northern-Southern European expected cline of these ancient components[25]. Moreover, no internal differences regarding the proportions of these ancient components were observed when modeling the Franco-Cantabrian groups individually (Figure S4F). This suggests that the genetic singularity of Basques might rely on demographic processes after the Iron Age, such as recent historical influences in Roman and Islamic periods. Therefore, qpAdm analysis was performed to formally test plausible post-Iron Age admixture models that might explain the singularity of Basques. This analysis shows that Basques can be mostly explained with the Iron Age samples from the Iberian Peninsula[19], with very limited influence of Roman Empire samples[26] in some groups next to Peri-Basques (Figure 4 and Table S2). These Roman-related proportions increase as farther the target populations are from the Franco-Cantabrian region. Moreover, no significant results were observed for any model considering North African samples[27] in the region (Figure 4 and Table S2). Overall, these results suggest that Basques might have received limited gene flow during recent historical events in the Iberian Peninsula.

Franco-Cantabrians show a heterogeneous genetic landscape correlated to geography

Since one of the major controversies regarding Basques is their internal genetic heterogeneity, we focused our analyses on the Franco-Cantabrian region to explore the internal genetic diversity in the region. A first PC separates all Franco-Cantabrian groups through a genetic cline, with all Basques in one extreme, the Spanish and French non-Franco-Cantabrians in the opposite one, and the Peri-Basques in an intermediate position (Figure 5A). The second PC, in turn, separates the regions from the western and eastern areas of the Franco-Cantabrian region. The PCA shows a remarkable micro-geographical genetic structure and clustering of the Basque groups. To obtain an external reference of similar size and sampling density at a micro-geographical scale such as the present study, the Catalan samples from Biagini et al.[28] were compared to our dataset. The PCA of Catalans does not show any geographical structure comparable to that of Basques and Peri-Basques (Figure S5), which may imply that the clustering observed in Basques is specific to them and unrelated to the sample strategy. To quantify and compare the genetic differentiation among the Franco-Cantabrian and Catalan groups, F_{ST} distances for each pairwise combination were estimated and shown in a MDS plot (Figure S6). Again, Franco-Cantabrians show a clear internal differentiation with distances in a range of 10^{-2}, whereas Catalans showed no evidence of genetic structure or extreme internal differentiation with distances in the range of 10^{-3}. Moreover, heterogeneity was tested within the region by performing an analysis of molecular variance (AMOVA) analyses at different strata in the geography. Though the explained genetic variance was small, all the results were statistically significant, pointing to an internal differentiation of the region and especially in the Basque groups (Figure S7). In fact, the same analysis in Catalonia shows lower explained variance in all comparisons (Figure S7D).

The analysis of genetic components performed in the region with ADMIXTURE mirrors the results described above (Figure 5B). At K=2 (best K with the lowest cross-validation error), Basques present a main component (green) that is also present at substantial proportions in the Peri-Basque groups and marginally found in the external samples. At K=3 and K=4, internal different components appear within Basques. In K=3, the Basque-related component splits in two specific components: a Western component (blue) and an Eastern component (green).

These components are barely presented in the non-Franco-Cantabrian samples. Finally, at K=4 another component arises, maximized in Araba and the surrounding groups (pink). Thus, these four components could be summarized in: non-Franco-Cantabrian (orange), Eastern-related Basque (green), central-related Basque (blue), and Western-related Basque (pink) components. The distribution of these components among the samples evidences the correlation between genetics and geography in the region. To formally test for this correlation, we performed an isolation by distance (IBD) analysis. A Mantel test was applied between the F_{ST} values and the geographic distances, resulting in a positive and clear statistically significant result (R^2 = 0.242, p-value = 0.0163) (Figure S5B). Then, the spatially explicit statistical method, EEMS, showed a well-defined internal pattern of barriers (Figure S5A), both between the Basque and Peri-Basque area, as well as within them. In fact, the pattern of the corridors with higher migration rate mirrors the observed relationships in the ADMIXTURE analysis and the PCA, between groups with overlapping standard deviations (Figures 5 and S5C). The same analyses were performed for the Catalan samples and despite being a geographical region of similar size, the results show a non-significant and negative trend for the Mantel test (R^2 = -0.151, p-value = 0.749), and the absence of barriers in the EEMS analysis (Figures S5A and S5B).

Finally, to refine the relationships between the populations within the area, we applied haplotype-based methods, and similar patterns of internal heterogeneity were observed (Figures 3, 6 and Table S1). Besides the differentiation of the Basque cluster in the fineSTRUCTURE dendrogram (Figures 3A, 3B and S2), several internal clusters in the region can be defined mostly related to geography and language. On the one hand, three clusters are shown in the Basque branch (Figure 6A right): one that encompasses the Central Basques, plus an Eastern Basque and a Western Basque cluster. On the other hand, two clusters are shown in the Peri-Basque branch (Figure 6A left): a Western and an Eastern Peri-Basque clusters, besides the Cantabrian samples (gCAN) that fall within the external Spanish cluster. Similar clusters are observed when performing the analysis reducing the dataset (Figures S2B and S2C). The differences found in Basques are also shown in the ancestry profiles calculated in the NNLS analysis (Figures 3B and 6B). Basque clusters are formed by exclusively Basque or Peri-Basque components, whereas Peri-Basques show also external ancestries. Focusing on

Basques, the Central group is defined exclusively by Basque ancestry, whereas Eastern and Western Basques present ~25% of Peri-Basque ancestry. Despite some traces of Spanish and French ancestries in the Western Basque group, other external ancestries are absent in Basques, suggesting haplotype sharing within the region without external contributions.

Discussion

Our results show a clear genetic distinctiveness of the Basques within the European landscape in all the analyses performed, with evidence of continuous inbreeding reflected in their small *Ne* values, the large number and length of ROHs and PI_HAT values (Figures 3C and S3), which suggests a pattern of genetic isolation in Basques in their recent history. In contrast, Peri-Basques show a transition between the Basques and the external Spanish and French populations (Figures 3, 5, 6 and S6A), as an example of intermediate area between open and isolated populations[29]. Indeed, Basques do not show evidence of recent admixture events with non-Franco-Cantabrian groups, whereas Peri-Basques show gene flow at least since Medieval times (Figures 3B, 6B, S2E and S2F). The inferred admixture events in these groups, involving potential Franco-Cantabrian-related and Spanish-related sources, mainly dated back to the so-called period of the Reconquista and the immediate next centuries. This was a long period characterized by the conflicts between the Islamic troops and the Christian kingdoms to maintain their territories in the Iberian Peninsula, expanding from the alleged Battle of Covadonga in 718 until the end of the Iberian Islamic rule with the fall of the Nasrid Kingdom of Granada in 1492[30]. It led to a complex political and administrative situation in the history of the Iberian Peninsula, so these results might be echoing the consequences of the recurrent reorganization of the territory during these centuries along the Franco-Cantabrian region and specially the Peri-Basque area.

Nevertheless, our analyses support the notion that the genetic uniqueness of Basques cannot be attributed to a different origin relative to other Iberian populations, but instead to a reduced and irregular external gene flow since the Iron Age as suggested by Olalde et al.[19]. The observed clines of post-Iron Age gene flow in the region suggest that the specific genetic profile of Basques might be explained by the lack of recent gene flow received. Our analyses confirm

297 that Basques were influenced by the major migration waves in Europe until the Iron Age, in a
298 similar pattern as their surrounding populations. At that time, Basques experienced a process of
299 isolation, characterized by an extremely low admixture with the posterior population movements
300 that affected the Iberian Peninsula, such as the Romanization or the Islamic rule, as observed in
301 the present genetic landscape (Figures 3B, 4, S4 and Table S2). This does not exclude
302 plausible previous periods of isolation as attested by the presence of short ROHs and small *Ne*
303 values (Figure S3) that support signals of ancient inbreeding in the region, even higher than in
304 Sardinia, which is suggested to be isolated after Neolithic times. Thus, the increase of the Ne
305 observed only in the external groups about 1000 generations ago might be potentially linked to
306 the role of the Franco-Cantabrian region as glacial refugium during Last Glacial Maximum
307 periods and the subsequent expansion [31]. Although our results support the genetic continuity
308 from the Iron Age[19] in most of the present-day Basques, those located in the periphery of the
309 Basque core area show signals of contacts compatible with the Roman Empire presence in the
310 Iberian Peninsula (Figure 4 and Table S2). These results are in agreement with archaeological
311 and historical records. An important presence of the Roman Empire has been reported in the
312 whole Franco-Cantabrian region, but the scholars suggest a much higher impact in the
313 peripheral areas of the southern side, specially Nafarroa and Araba[32]. Otherwise, North African
314 influence only fit the models where southern and northwestern Iberians are included (Figure 4
315 and Table S2). This confirms the reduced gene flow between the eastern and northern areas of
316 the Iberian Peninsula with the North African incomers during the Islamic rule, as already
317 reported by using uniparental markers[33] and more recently through genome-wide data and
318 haplotype-based methods[34].
319
320 Language might play a major role in the demographic processes of the populations, and in the
321 present study, given the extraordinary traits of Euskara, the ethno-linguistic scenario of the
322 region should be considered in the interpretation of the analyses performed. Our results are
323 compatible with the Euskara as one of the main factors preventing major gene flow after the
324 Iron Age and shaping the genetic panorama of the Basque region. The genetic continuity of
325 Basques since the Iron Age also supports the hypothesis that the expansion of the Steppe
326 ancestry did not completely erase Pre-Indo-European languages in Western Europe, as

previously suggested in other studies[19]. Before the arrival of the Romans to the Iberian Peninsula, the Euskara coexisted with other pre-Indo-European languages, such as Iberian. The contact with the Romans, and thus with Latin, was earlier and stronger in the Mediterranean watershed of the Iberian Peninsula followed by their expansion towards the Atlantic coast, with a later and lower impact in the Franco-Cantabrian area[35] as confirmed in our analyses (Figure 4 and Table S2). Thus, it is expected that Latin had a deeper effect in the areas with stronger Roman rule, speeding the language replacement, whereas Euskara was scarcely affected. Once Latin was established in most of the Iberian Peninsula, Euskara might have acted as a cultural barrier to gene flow, leading to a genetic differentiation of the Basques and a low influence of the linguistic Romance substratum in Euskara [4,5].

Together with a positive strong correlation with geography (Figure S5), our results confirm a clear internal heterogeneity in Basques and Peri-Basques, where East-West genetic clusters are evident along the Franco-Cantabrian region, showing a genetic cline from the core to the external areas with higher gene flow among the closest groups. This genetic substructure is more complex than the Northern-Southern orographic and administrative limits between the present-day Spanish and French territories, separated by the Pyrenees Mountains. Instead, haplotype-based methods enabled to accurately define a central Basque cluster, plus a Western and an Eastern Basque and Peri-Basque genetic clusters (Figures 3B, 5 and 6); clarifying previous results that barely suggested this pattern in the Franco-Cantabrian region by using classical markers[36,37]. Whereas the Eastern and the Central Basques do not show relevant gene flow with external sources, the groups from the Western Basque cluster present small levels of gene flow with them (Figures 3B, 4, 5 and 6). This genetic substructure reflects the historical and linguistic context between the Basques and the neighboring areas. Those from the Central region are the most differentiated, due to the smaller influence from external sources along history regarding its farthest location[37,38]. Moreover, the linguistic, political and administrative situation in the Central and the Eastern areas have been quite stable along history[39]. However, the Western areas are characterized by a complex scenario, related to the reorganization of the administrative limits and the regression of Euskara due to the strong influence of surrounding Romance languages and mainly Spanish[32,39,40].

Finally, the role of the Euskara dialects on the present-day Basque genetic substructure is more difficult to assess. Most linguist scholars agree on a Western-Eastern dialectal discontinuity with small distances among the closest dialects[41,42], resembling the genetic clusters observed in the present analysis. The most accepted hypothesis about their origin is that they emerged during the medieval period[4,42,43]. Yet, our results reveal a clearly defined genetic structure within Basques that might have been formed in earlier times. Indeed, a pre-Roman genetic substructure has been already suggested based on uniparental markers[6]. Therefore, the diversification of the dialects and the genetic heterogeneity within the Basques might have in common their correlation to geography and it might be worth to reconsider an earlier origin of the Euskara dialects in linguistic studies.

In this study, we disentangle long-standing controversies about the genetic uniqueness of Basques from a new perspective, providing finer and more reliable conclusions that help to resolve the debate on the demographic history of Basques. This contributes to demystify the so often assumed genetic characteristics of Basques, even though little clear evidence was present to support their position as some sort of Proto-Europeans. Moreover, some new lines of evidence are presented to the Archaeology and Anthropology fields that can be further pursued. The genetic results presented here give support to the anthropological and chronological connections with the ancient remains, as well as the historical and linguistic relationships around the Euskara and its dialects. It is pivotal to emphasize the importance of a multidisciplinary approach when studying the evolution and demographic history of populations to contextualize the study, test hypotheses and interpret the results. In this sense, it is crucial to encourage the integration and cooperative research among the different areas of knowledge with the aim of boosting complete, contrasted, and reliable studies.

Acknowledgements

The authors acknowledge all the volunteers who kindly accepted to participate in the study. This study was supported by the Spanish Ministry of Science, Innovation and Universities (MCIU) and the Agencia Estatal de Investigación (AEI) grant number PID2019-106485GB-

I00/AEI/10.13039/501100011033, and “Unidad de Excelencia María de Maeztu” (AEI,CEX2018-000792-M). A.F-B was supported with a FI fellowship from the Generalitat de Catalunya (2016FI_B 00446). The sampling 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 Centre National de la Recherche Scientifique (CNRS) (interdisciplinary programme), OHLL (Origine de l’Homme, des Langues et du Langage), and Association Sang 64”. We are indebted to all the people contributing samples to this study. Especially to Estibaliz Montoya, David Basterot, Tristan Carrère, Mònica Vallés, and Stéphanie Plaza.

Author Contributions

A.F-B. and D.C. designed and performed research; F.B., J.S., B.O. collected and provided the biological and ethno-linguistic data; A.F-B. processed and analyzed the data; A.F-B., F.C., J.B., L. Q-M. and D.C. contributed to the interpretation of results. A.F-B. and D.C. wrote the paper; all authors revised the final version of the manuscript.

Declaration of Interests

The authors declare no competing interests.

Figures and Table Legends

Table 1. Information of the Franco-Cantabrian samples reported in this study.

Our dataset includes a total of 190 samples from 18 areas in the present-day Spanish and French territories of the Franco-Cantabrian region. Both Euskara and surrounding Romance speaking groups were considered. All four grandparents of the individuals were born in the same geographical region of the collected sample. N represents the number of individuals genotyped for each region. See also Figures 1 and S1.

Figure 1. Geographic distribution of the Franco-Cantabrian region and the areas included in the study.

ALA, Araba; BBA, Bizkaia; BOC, Western Bizkaia; GUI, Central Gipuzkoa; GSO, Southwestern Gipuzkoa; NNO, Northwestern Nafarroa; NCO, Central Western Nafarroa; ZMX, Lapurdi/Baztan; NLA, Lapurdi Nafarroa; SOU, Zuberoa; RON, Roncal; CAN, Cantabria; BUR, Northern Burgos; RIO, Rioja; NAR, Northern Aragon; CHA, Chalosse; BEA, Béarn; BIG, Bigorre. The dark shaded area represents the current extension of the main Euskara-speaking zone. The colors represent the ethno-linguistic information of the sample set: in green, the sampled areas where the Euskara was spoken in historical times and at least until the 19th century; here referred as Basques. In pink and blue, the sampled areas where Spanish and Gascon/French are spoken today, respectively; referred as Spanish (the former) and French (the latter) Peri-Basques in this study. See sample collection in STAR★METHODS for detailed information. Km, Kilometers. See also Table 1 and Figure S1.

Figure 2. Contextualizing the Franco-Cantabrian region.

(A) Principal component analysis including all modern samples analyzed in the present study. (B) ADMIXTURE results with similar and lowest cross-validation errors (K=6 and K=7). ★, Spanish Basque and Peri-Basque; ●, French Basque and Peri-Basque. In the large Caucasus/Middle East section, the two populations with a maximized component are Bedouin (light green) and Druze (gold). In the North African section, the purple component is maximized in Mozabite. See also Figure S2.

Figure 3. Haplotype-based analyses.

(A) European branch of the fineSTRUCTURE dendrogram inferred for the whole modern dataset. The complete dendrogram is shown in Figure S2A. The Franco-Cantabrian, other Spanish, and other French clusters are highlighted in color. (B) Inferred proportions of shared ancestry among genetic clusters using the NNLS method from GLOBETROTTER. Ancestries that are neither Spanish nor French were pooled and labeled “external”. (C) Whisker plots representing the individual information of number (NROH, left) and length (SROH, right) of Runs of Homozygosity for Basques, Peri-Basques, French and Spanish. Sardinian, Central Western

Europeans (CEU), and sub-Saharan Africans (YRI) were also included as reference populations with well-known demographic histories. Mb, Megabases. For the sake of clarity, the name of those genetically homogeneous clusters inferred from fineSTRUCTURE are referred with “g”. See also Figures S2, S3 and Table S1.

Figure 4. Modelling potential post-Iron Age gene flow in the Iberian Peninsula.

(A) Geographic distribution of the inferred proportions in the map. Shadings for proportions are scaled according to the maximum and minimum proportions of each source. (B) Representation of the admixture proportions in each target population based on the statistically significant models obtained with qpAdm. See also Figure S4 and Table S2. Km, Kilometers.

Figure 5. Population stratification of the Franco-Cantabrian region.

(A) PCA using the samples from the Iberian Peninsula and France. The PCA average and standard deviation values of the different geographic groups were plotted. (B) ADMIXTURE results from K=2 to K=4. The lowest cross-validation error in the analysis was K=2. The asterisks represent “Spanish_Pais_Vasco_IBS” and “French_Bearn”. PB, Peri-Basque. See also Figures S5, S6 and S7.

Figure 6. Haplotype distribution and sharing within the Franco-Cantabrian region.

(A) Internal Franco-Cantabrian clusters inferred with fineSTRUCTURE. Left and right dendrograms represent a zoom of the Peri-Basque and Basque specific branches, respectively, in the complete dendrogram from Figure 3. The central map shows the geographic distribution of the haplotypes inferred in the fineSTRUCTURE dendrogram. Each pie chart represents the proportion of the clusters in the corresponding region. (B) Inferred proportions of shared ancestry among the Franco-Cantabrian groups. This plot is in continuity with the Figure 3B. Here, the internal proportions of haplotype sharing in the Franco-Cantabrian region are dissected. Wavy lines represent external proportions. See also Figures S2, S5, S6, S7, and Table S1.

STAR★METHODS

KEY RESOURCES TABLE

RESOURCES AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, David Comas (david.comas@upf.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The original dataset generated during this study is available for download at Figshare: https://figshare.com/s/61a54472b63fd0101859

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample Collection

A total of 190 individual samples were collected from 18 microgeographical areas in the present-day Spanish and French territories of the Franco-Cantabrian region, including Basque and Romance (Spanish, French, Aragonese and Gascon) speaking groups (Table 1 and Figure 1). Written informed consent was obtained from the participants, who were interviewed to determine their local dialect and to confirm that their four grandparents were born in the same microgeographical area (Figures 1, S1 and Table 1). The collection procedures were approved by the CCPPRB (Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale d'Aquitaine) and the IRBs at Universitat Pompeu Fabra, Institut Pasteur and Université Michel de Montaigne Bordeaux 3. The dataset includes seven areas where Euskara is spoken at present[6,18] (Central Gipuzkoa, Southwestern Gipuzkoa, Bizkaia, and Northwestern Nafarroa in the southern side of the Pyrenees; the zone of Lapurdi/Baztan, Lapurdi/Nafarroa Beherea, and Zuberoa in the northern side), besides three regions where it was spoken at least until the end of the 19th century[39,44] (Central Western Nafarroa, Roncal, and Araba). For the sake of clarity, the previous ten regions where Euskara was spoken in historical times will be

503 referred as “Basque” groups. The dataset also includes i) five surrounding Spanish-speaking
504 areas (Cantabria, Northern Burgos, La Rioja, Northern Aragon and Western Bizkaia; in the case
505 of Northern Aragon, a relict local language, Aragonese, is still spoken) and ii) three traditionally
506 Gascon-speaking areas (Bigorre, Béarn and Chalosse)[4,39–41]. In the latter, influences of
507 Euskara-related archaic languages have been suggested[5] (Figure S1), but records of spoken
508 Euskara are absent. Moreover, Gascon has been replaced extensively by French, with less
509 than half of the population in Béarn being fluent in Gascon. We refer to these two groups of
510 populations as Spanish and French Peri-Basques, respectively.
511
512 **METHOD DETAILS**
513 ***Sample Genotyping***
514 DNA was extracted from blood samples using standard methods, and genotyped with the
515 Axiom™ Genome-Wide Human Origins Array (~629,443 variants)[45]. Genotype calling was
516 performed by using the software Axiom™ Analysis Suite 3.1.51.0 through the Affymetrix Best
517 Practices Workflow. All samples passed the genotype calling process with an average quality
518 control call rate of 99.8%. After exporting those variants that properly passed the recommended
519 thresholds, 597,638 single-nucleotide polymorphisms (SNPs) remained in the dataset.
520
521 ***Data Quality Control***
522 PLINK 1.9[46] was used to apply the quality control filters detailed below, after excluding
523 uniparental markers and the X chromosome. A filter for more than 10% missing SNPs per
524 sample resulted in no sample exclusion. A total of 434,664 SNPs remained after removing
525 SNPs that were missing in more than 5% of the samples, with an extreme deviation from Hardy-
526 Weinberg equilibrium ($p < 10^{-5}$), and a minor allele-frequency (MAF) below 0.05. For those
527 analyses that required linkage equilibrium between SNPs, linkage disequilibrium (LD) pruning
528 was performed using a window size of 200 SNPs, shifting by 25 SNPs and a maximum pairwise
529 LD threshold (r^2) of 0.5. A total of 171,275 SNPs remained after LD pruning. Two individual
530 samples (one from the Lapurdi/Baztan zone and one from Zuberoa) were discarded from the
531 dataset due to their high genetic relatedness to other samples in the study (PI_HAT>0.125).
532

Merging Data

Our Franco-Cantabrian samples were merged with public available data genotyped with the same array: samples from Catalonia, Valencia, and Balearic Islands[28]; plus modern and ancient North African and Western Eurasian samples[19,26,27,47]. Modern Basque samples from Lazaridis et al.[27] were included in the analyses in the Basque category as general "Spanish" and "French" Basques due to the lack of information about their microgeographical origin. Samples from Eivissa (within the Balearic Islands dataset) were removed from most of the analyses since it has been reported to be an isolated population differentiated from the Iberian mainland context[28]. Thus, the final dataset contained 1,970 samples, as well as 434,664 SNPs in the dataset used for haplotype-based methods (modern data) and 171,275 SNPs after linkage disequilibrium pruning in the dataset used for allele frequency methods (both modern and ancient data).

QUANTIFICATION AND STATISTICAL ANALYSIS

Allele frequency methods

Principal Component Analysis (PCA) eigenvectors were computed by using SmartPCA v13050 from EIGENSOFT v6.0.1[48]. For the PCA including ancient samples, these were projected to the PCA space obtained with the modern samples.

In order to explore the population structure in the region, Weir and Cockerham's pairwise F_{ST} indexes were calculated between the different groups by using the R package stAMPP[49], with confidence intervals after 1,000 bootstrap repetitions, and represented in a multidimensional scaling (MDS) plot. Then, an AMOVA was performed with the R package poppr[50]. For that, 1,000 permutations as described by Excoffier et al.[51] were carried out with ade4 R package[52], in order to obtain an empirical distribution under the null hypothesis.

To test the correlation between the genetic and geographic distances among groups, an isolation by distance (IBD) analysis was performed. Geographic Distance Matrix Generator v1.2.3 software[53] was used to calculate a geographic distance matrix between the groups

included in the analysis. Then, a Mantel test was performed between the genetic and geographic distance matrices with the mantel.rtest function in the ade4 R package[52].

Effective migration and diversity rates were modeled in the Franco-Cantabrian region dataset using the estimated effective migration surfaces (EEMS) software[54]. The number of nDemes was set to 300, and four simultaneous EEMS analyses were run with four different random seeds. Next, the MCMC chain that converged with the highest final log-posterior value was continued with four new runs. This process was repeated three more times, taking the best chain after the previous set of runs in order to be confident that the MCMC chain had converged with the optimal log-posterior value. For every set of four runs, a total of 10 million iterations each were performed, with 5 million discarded as burn-in and sampling every 50,000 iterations.

To explore patterns of population structure, Model-based individual ancestries were defined by using ADMIXTURE v1.3.0[55]. The unsupervised method was run for 10 independent iterations, both for the present-day and ancient samples analyses, and with K ancestral clusters ranging from 2 to 12. Then, PONG v1.4.7[56] was used to combine the result of the 10 different iterations, obtain the major modes of the ADMIXTURE results and plot them in a consensus graph.

AdmixTools package[45] was used to study the relationship between modern and ancient samples. Outgroup f3-statistics were performed with the qp3Pop program to calculate the shared drift between each potential ancient group and all the modern groups in the dataset. The analysis was performed in the form f3(Mbuti; Ancient, Modern), with Mbuti set as outgroup. Then, the qpGraph program was used to model the modern European samples to estimate the three major ancestral proportions in Europe as in Haak et al[25]: Western Pre-Neolithic hunter-gatherer, European Neolithic farmer, and Yamnaya from Post-Neolithic Pontic-Caspian Steppe. Based on previous studies[25,27,57] and our results, WHG, Europe_EN and Steppe_EMBA samples from Lazaridis et al.[27] were used as proxies, respectively. A total of 100 replicates were performed for each target modern population, and the average and the 95% confidence intervals of the replicates were calculated. Furthermore, the qpAdm program was run to explicitly test potential admixture models including representatives of the main post-Iron Age

migrations that might have affected the Iberian Peninsula, i.e., the influence of the Roman Empire and North African expansions. 1-way, 2-way and 3-way admixture models were tested by considering Iberian Iron Age[19], Imperial/Late Antiquity Roman[26] and North African[27] samples as proxies (See Table S2 for further details about the modeling process).

Haplotype-based analyses

Runs of Homozygosity (ROHs) were estimated with PLINK 1.9[46], including some external populations, in order to test for inbreeding. The ROH analysis was performed in sliding windows of 50 SNPs across the genome, allowing for 1 heterozygous call and 5 missing calls per window. The minimum hit rate of all scanning windows containing the SNP to be included in a ROH was set to 0.05. Only ROHs with a minimum length of 500 kilobases (kb) and containing at least 50 SNPs were included in the analysis. One SNP per 50 kb was set as the minimum density threshold and a maximum gap of 100 kb was allowed between two consecutive SNPs in a tract. Due to their isolated nature, the Canary and Balearic Islands[27] samples were removed from the dataset for this analysis. In addition, internal genetic relationships were explored within populations. To that, the PI_HAT values (i.e., proportion of segments shared between two individuals due to identity-by-descent) were inferred with PLINK 1.9[46].

SHAPEIT v2[58] was used to phase the data, with the population-averaged genetic map from HapMap phase III[59] and the available 1000 genomes data as reference panel[60]. The data was first aligned to the reference and the mismatched SNPs were removed, then the proper phase inference was performed.

Haplotype sharing between individual modern samples was estimated with ChromoPainter[61]. It depicts individually the haplotypes of each “recipient” sample, without population specification, as a haplotype combination of all the other samples, treated as “donor”. Thus, it estimates the total number and length of haplotype fragments (chunks) in the recipient’s genome that shares the most common recent ancestor with every donor sample. First, ChromoPainter was run to estimate the global mutation probability and the switch rate parameters. Thus, the expected-maximization (EM) algorithm implemented in ChromoPainter was used over chromosomes 1, 4,

17 and 20 for all individuals, with 10 iterations and parameters -in -iM. Then, the inferred values were averaged across the four chromosomes and individuals, weighting by the number of SNPs per chromosome. It resulted in 0.000661 and 222.54421 for the global mutation (M) and switch rates (n), respectively. Finally, ChromoPainter was run for all chromosomes and individuals fixing the previously estimated parameters to obtain the final count and length sharing coancestry matrices. Then ChromoCombine was used to sum the matrices across chromosomes and obtain the copying profile for each individual, as well as the C parameter (C=0.281698702) needed for running fineSTRUCTURE.

FineSTRUCTURE v2.1.0[61] was run to cluster the data provided by ChromoPainter into homogeneous genetic groups following Leslie et al.[62]. The chunkcount coancestry matrix of the total number of chunks copied among individuals was used as input. A total of 2 million MCMC iterations were performed, with 1 million burn-in iterations and sampling values from the posterior probability every 10,000 iterations. The FineSTRUCTURE dendrogram was built with the default parameters -m T. The analysis was repeated for three different seeds to check consistency among the dendrograms. Major differences were not observed among them, except the relocation of a few individuals to other genetically close clusters. When confusion between sampling and genetic definitions of the groups might exist, the homogeneous genetic clusters inferred with fineSTRUCTURE were explicitly named by adding a “g” to the group label (as show in Table S1).

GLOBETROTTER[63] was used following the recommended protocol to estimate the ancestry profile of all the modern samples in the dataset, and test for admixture events in the target populations that showed interesting patterns of admixture. To perform this analysis, ChromoPainter was run first, but classifying the individuals based on the previous clustering results from fineSTRUCTURE. All of them were included in the analysis as donors and recipients. Next, the non-negative least squares (NNLS) method included in GLOBETROTTER was applied to infer the haplotype sharing proportions between the different clusters. Thus, prop.ind parameter was set to 1, null.ind to 0 and num.mixing.iterations to 0. A different NNLS analysis was performed for each cluster as recipient and the others as donors, disallowing self-

copying from the recipient itself. Then, GLOBETROTTER was run to check if some admixture events could be detected in our target clusters, using the other clusters as surrogates. The copying vectors and the painting samples from the second run of ChromoPainter were used to run this analysis. As recommended, null.ind parameter was set to 1 when testing for plausible admixture events. This accounts for uncommon patterns in the linkage disequilibrium decay that could show false signals of admixture. Then, null.ind parameter was set to 0 when estimating admixture proportions, dates, and sources. The confidence intervals for the date estimates were inferred by performing 100 bootstrap iterations, for one-date and two-date admixture models.

Finally, in order to show how population dynamics can explain the genetic diversity observed in our analyses, effective population size (*Ne*) was estimated for the Basques, Peri-Basques, Spanish and French groups. The *NeON* R package[64] was used to calculate the effective population size across generations in the target groups, by considering the relationship between *Ne* and the population-specific patterns of linkage disequilibrium from genome-wide SNP data. The recommended guidelines were followed to perform the analysis and interpret the results[64]. First, linkage was estimated between markers considering sliding windows of 500 kb and a maximum of 10,000 confrontations in each by using the *NeLD* function. Then, Ne was estimated with the function *Nestimate,* setting a minimum and maximum r^2 values to 0.001 and 0.999 in each comparison, respectively. Quantiles of the distribution and median for the long-term *Ne* (5^{th}, 50^{th} and 95^{th} percentiles) were estimated by using the functions *Ne_CI* and *Ne_Med*.

References

1. Pike, A.W.G., Hoffmann, D.L., Garcia-Diez, M., Pettitt, P.B., Alcolea, J., De Balbin, R., Gonzalez-Sainz, C., de las Heras, C., Lasheras, J.A., Montes, R., et al. (2012). U-Series Dating of Paleolithic Art in 11 Caves in Spain. Science (80-.). *336*, 1409–1413.
2. Eberhard, M., D., Simons, G.F., and Fennig, C.D. eds. (2020). Ethnologue: Languages of the World 23rd ed. (SIL International).
3. Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2020). Glottolog 4.3 (Max Planck Institute for the Science of Human History).
4. Zuazo, K. (2010). El euskera y sus dialectos (Alberdania S. L.).

5. Núñez Astrain, L. (2007). El euskera arcaico: extensión y parentescos (Txalaparta).
6. Martínez-Cruz, B., Harmant, C., Platt, D.E., Haak, W., Manry, J., Ramos-Luis, E., Soria-Hernanz, D.F., Bauduer, F., Salaberria, J., Oyharçabal, B., et al. (2012). Evidence of pre-roman tribal genetic structure in basques from uniparentally inherited markers. Mol. Biol. Evol. *29*, 2211–2222.
7. Laayouni, H., Calafell, F., and Bertranpetit, J. (2010). A genome-wide survey does not show the genetic distinctiveness of Basques. Hum. Genet. *127*, 455–458.
8. Rodríguez-Ezpeleta, N., Álvarez-Busto, J., Imaz, L., Regueiro, M., Azcárate, M.N., Bilbao, R., Iriondo, M., Gil, A., Estonba, A., and Aransay, A.M. (2010). High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations. Hum. Genet. *128*, 113–117.
9. Etcheverry, M.A. (1945). El factor rhesus: Su genética y su importancia clínica. Dia Med. *17*, 1237–1259.
10. Flores-Bello, A., Mas-Ponte, D., Rosu, M.E.M.E., Bosch, E., Calafell, F., and Comas, D. (2018). Sequence diversity of the Rh blood group system in Basques. Eur. J. Hum. Genet. *26*, 1859–1866.
11. Bertranpetit, J., and Cavalli-Sforza, L.L. (1991). A genetic reconstruction of the history of the population of the Iberian Peninsula. Ann. Hum. Genet. *55*, 51–67.
12. Calafell, F., and Bertranpetit, J. (1994). Principal component analysis of gene frequencies and the origin of Basques. Am. J. Phys. Anthropol. *93*, 201–215.
13. López, S., and Alonso, S. (2013). Genetics and the History of the Basque People. In eLS (John Wiley & Sons, Ltd).
14. Pérez-Lezaun, A., Calafell, F., Mateu, E., Comas, D., Ruiz-Pacheco, R., and Bertranpetit, J. (1996). Microsatellite variation and the differentiation of modern humans. Hum. Genet. *99*, 1–7.
15. Izagirre, N., and de la Rúa, C. (1999). An mtDNA analysis in ancient Basque populations: implications for haplogroup V as a marker for a major paleolithic expansion from southwestern europe. Am. J. Hum. Genet. *65*, 199–207.
16. Lopez-Parra, A.M., Gusmão, L., Tavares, L., Baeza, C., Amorim, A., Mesa, M.S., Prata, M.J., and Arroyo-Pardo, E. (2009). In search of the Pre- and Post-Neolithic Genetic

712 Substrates in Iberia: Evidence from Y-Chromosome in Pyrenean Populations. Ann. Hum.
713 Genet. *73*, 42–53.
714 17. Günther, T., Valdiosera, C., Malmström, H., Ureña, I., Rodriguez-Varela, R.,
715 Sverrisdóttir, Ó.O., Daskalaki, E.A., Skoglund, P., Naidoo, T., Svensson, E.M., et al.
716 (2015). Ancient genomes link early farmers from Atapuerca in Spain to modern-day
717 Basques. Proc. Natl. Acad. Sci. *112*, 11917–11922.
718 18. Behar, D.M., Harmant, C., Manry, J., Van Oven, M., Haak, W., Martinez-Cruz, B.,
719 Salaberria, J., Oyharabal, B., Bauduer, F., Comas, D., et al. (2012). The Basque
720 paradigm: Genetic evidence of a maternal continuity in the Franco-Cantabrian region
721 since pre-neolithic times. Am. J. Hum. Genet. *90*, 486–493.
722 19. Olalde, I., Mallick, S., Patterson, N., Rohland, N., Villalba-Mouco, V., Silva, M., Dulias,
723 K., Edwards, C.J., Gandini, F., Pala, M., et al. (2019). The genomic history of the Iberian
724 Peninsula over the past 8000 years. Science (80-.). *363*, 1230–1234.
725 20. Gattepaille, L.M., and Jakobsson, M. (2012). Combining markers into haplotypes can
726 improve population structure inference. Genetics *190*, 159–174.
727 21. Wangkumhang, P., and Hellenthal, G. (2018). Statistical methods for detecting
728 admixture. Curr. Opin. Genet. Dev. *53*, 121–127.
729 22. Di Gaetano, C., Fiorito, G., Ortu, M.F., Rosa, F., Guarrera, S., Pardini, B., CCusi, D.,
730 Frau, F., Barlassina, C., Troffa, C., et al. (2014). Sardinians genetic background
731 explained by runs of homozygosity and genomic regions under positive selection. PLoS
732 One *9*.
733 23. Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, J.Z.
734 (2012). Genomic patterns of homozygosity in worldwide human populations. Am. J.
735 Hum. Genet. *91*, 275–292.
736 24. Kirin, M., McQuillan, R., Franklin, C.S., Campbell, H., McKeigue, P.M., and Wilson, J.F.
737 (2010). Genomic Runs of Homozygosity Record Population History and Consanguinity.
738 PLoS One *5*, e13996.
739 25. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G.,
740 Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the
741 steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211.

26. Antonio, M.L., Gao, Z., Moots, H.M., Lucci, M., Candilio, F., Sawyer, S., Oberreiter, V., Calderon, D., Devitofranceschi, K., Aikens, R.C., et al. (2019). Ancient Rome: A genetic crossroads of Europe and the Mediterranean. Science (80-.). *366*, 708–714.
27. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature *536*, 419–424.
28. Biagini, S.A., Solé-Morata, N., Matisoo-Smith, E., Zalloua, P., Comas, D., and Calafell, F. (2019). People from Ibiza: an unexpected isolate in the Western Mediterranean. Eur. J. Hum. Genet. *27*, 941–951.
29. Anagnostou, P., Dominici, V., Battaggia, C., Pagani, L., Vilar, M., Wells, R.S., Pettener, D., Sarno, S., Boattini, A., Francalacci, P., et al. (2017). Overcoming the dichotomy between open and isolated populations using genomic data from a large European dataset. Sci. Rep. *7*, 41614.
30. Melo Carrasco, D., and Francisco, V.C. (2012). A 1300 años de la conquista de al-Andalus (711-2011): Historia, cultura y legado del Islam en la Península Ibérica. (Centro Mohammed VI para el diálogo de civilizaciones).
31. Gómez-Carballa, A., Olivieri, A., Behar, D.M., Achilli, A., Torroni, A., and Salas, A. (2012). Genetic continuity in the franco-cantabrian region: New clues from autochthonous mitogenomes. PLoS One *7*.
32. Montero, M. (2008). Historia general del País Vasco (Txertoa).
33. Bosch, E., Calafell, F., Comas, D., Oefner, P.J., Underhill, P.A., and 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.
34. Bycroft, C., Fernandez-Rozadilla, C., Ruiz-Ponte, C., Quintela, I., Carracedo, A., Donnelly, P., and Myers, S. (2019). Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. Nat. Commun. *10*, 551.
35. Garcia Sinner, A., and Velaza, J. eds. (2019). Palaeohispanic Languages and Epigraphies (Oxford University Press).
36. Calafell, F., and Bertranpetit, J. (1994). Mountains and genes: population history of the

Pyrenees. Hum. Biol. *66*, 823–42.

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

38. Manzano, C., Aguirre, A.I., Iriondo, M., Martín, M., Osaba, L., and de la Rúa, C. (1996). Genetic polymorphisms of the Basques from Gipuzkoa: genetic heterogeneity of the Basque population. Ann. Hum. Biol. *23*, 285–296.

39. Abasolo, C.C. (2002). Las fronteras de la lengua vasca a lo largo de la historia. Rev. Filol. Románica *19*, 15–36.

40. Borrás, H.K. (2008). El euskera en tierras del romance: Rioja alavesa, La Rioja, Burgos, Encartaciones. Huarte San Juan . Filol. y Didáctica la Leng. *10*, 163–171.

41. Abaitua, J. (2018). Patrones geolingüísticos, áreas dialectales y cronologías absolutas del EHHA. Fontes linguae Vascon. Stud. Doc. *50*, 283–322.

42. Igartua, I., and Zabaltza, X. (2016). A Brief History of the Basque Language (Etxepare Euskal Institutua).

43. Mitxelena Elissalt, L. (1981). Lengua común y dialectos vascos. Int. J. Basqu. Linguist. Philol. *15*, 289–313.

44. Zuazo, K. (1996). The Basque Country and the Basque language: an overview of the external history of the basque language. In Towards a History of the Basque Language, J. I. Hualde, J. A. Lakarra, and R. L. Trask, eds. (John Benjamins Publishing Company), p. 5.

45. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–93.

46. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

47. van de Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E.H., El Hajraoui, M.A., et al. (2018). Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations.

Science (80-.). *360*, 548–552.

48. Patterson, N., Price, A.L., and Reich, D. (2006). Population Structure and Eigenanalysis. PLoS Genet. *2*, e190.

49. Pembleton, L.W., Cogan, N.O.I., and Forster, J.W. (2013). StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Mol. Ecol. Resour. *13*, 946–952.

50. Kamvar, Z.N., Tabima, J.F., and Grünwald, N.J. (2014). Poppr : an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ *2*, e281.

51. Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992). Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. Genetics *131*, 479–491.

52. Dray, S., and Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. J. Stat. Softw. *22*, 1–20.

53. Ersts, P.J. (2014). Geographic Distance Matrix Generator (version 1.2.3). American Museum of Natural History, Center for Biodiversity and Conservation.

54. Petkova, D., Novembre, J., and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. Nat. Genet. *48*, 94–100.

55. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

56. Behr, A.A., Liu, K.Z., Liu-Fang, G., Nakka, P., and Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics *32*, 2817–2823.

57. Chiang, C.W.K., Marcus, J.H., Sidore, C., Biddanda, A., Al-Asadi, H., Zoledziewska, M., Pitzalis, M., Busonero, F., Maschio, A., Pistis, G., et al. (2018). Genomic history of the Sardinian population. Nat. Genet. *50*, 1426–1434.

58. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. *10*, e1004234.

59. The International HapMap Consortium (2003). The International HapMap Project. Nature

426, 789–796.

60. Consortium, T. 1000 G.P. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

61. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. PLoS Genet. *8*, e1002453.

62. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al. (2015). The fine-scale genetic structure of the British population. Nature *519*, 309–314.

63. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science *343*, 747–751.

64. Mezzavilla, M., and Ghirotto, S. (2015). Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS. J. Comput. Sci. Syst. Biol. *8*, 37–44.

Table 1

Label	Location	Territory	Main language/dialect	N
BIG	Bigorre	France	French	10
BEA	Béarn	France	French	10
CHA	Chalosse	France	French	10
ZMX	Lapurdi/Baztan	France	Euskara - Upper Navarrese/Lapurdian	11
NLA	Lapurdi Nafarroa	France	Euskara - Lower Navarrese	11
SOU	Zuberoa	France	Euskara - Zuberoan	11
RON	Roncal	Spain	Euskara - Roncalese (now Spanish)	11
NCO	Central Western Nafarroa	Spain	Euskara - Upper Navarrese (now Spanish)	11
NNO	Northwestern Nafarroa	Spain	Euskara - Upper Navarrese	11
GUI	Central Gipuzkoa	Spain	Euskara - Gipuzkoan	11
GSO	Southwestern Gipuzkoa	Spain	Euskara - Biscayan	11
ALA	Araba	Spain	Euskara - Biscayan (now Spanish)	11
BBA	Bizkaia	Spain	Euskara - Biscayan	11
BOC	Western Bizkaia	Spain	Spanish	10
CAN	Cantabria	Spain	Spanish	10
BUR	Northern Burgos	Spain	Spanish	10
RIO	Rioja	Spain	Spanish	10
NAR	Northern Aragon	Spain	Spanish	10

Figure 1

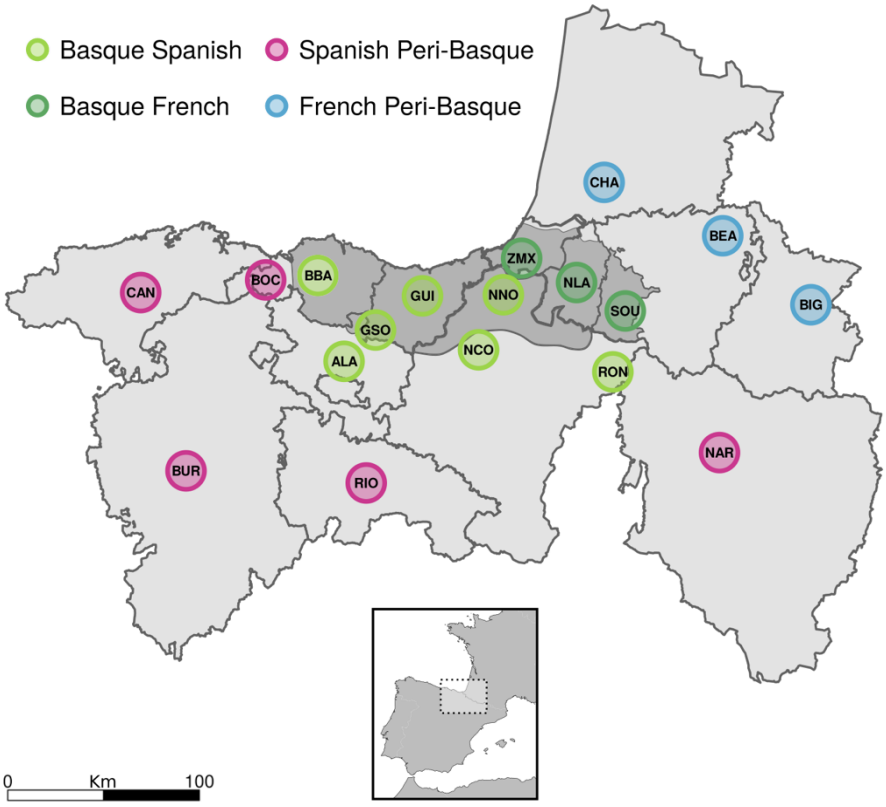
Basque Spanish
Spanish Peri-Basque
Basque French
French Peri-Basque
CHA
BEA
ZMX
NLA
BIG
SOU
BOC
BBA
GUI
NNO
CAN
GSO
NCO
ALA
RON
NAR
BUR
RIO
0
Km
100

Figure 2

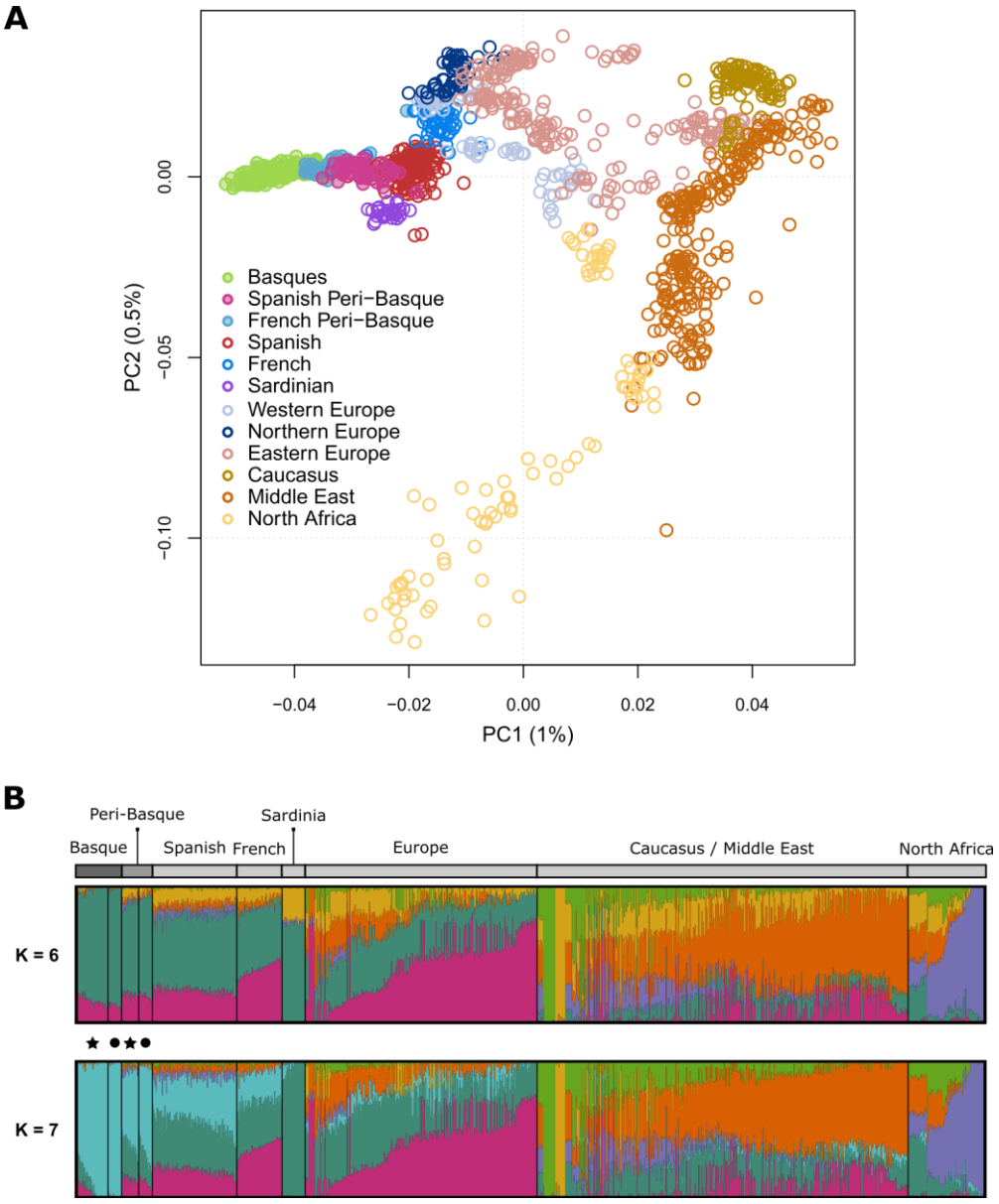
A
PC2 (0.5%)
0.00
−0.05
−0.10
−0.04
−0.02
0.00
0.02
0.04
PC1 (1%)
Basques
Spanish Peri−Basque
French Peri−Basque
Spanish
French
Sardinian
Western Europe
Northern Europe
Eastern Europe
Caucasus
Middle East
North Africa
B
Peri-Basque
Sardinia
Basque
Spanish
French
Europe
Caucasus / Middle East
North Africa
K = 6
K = 7

Figure 3

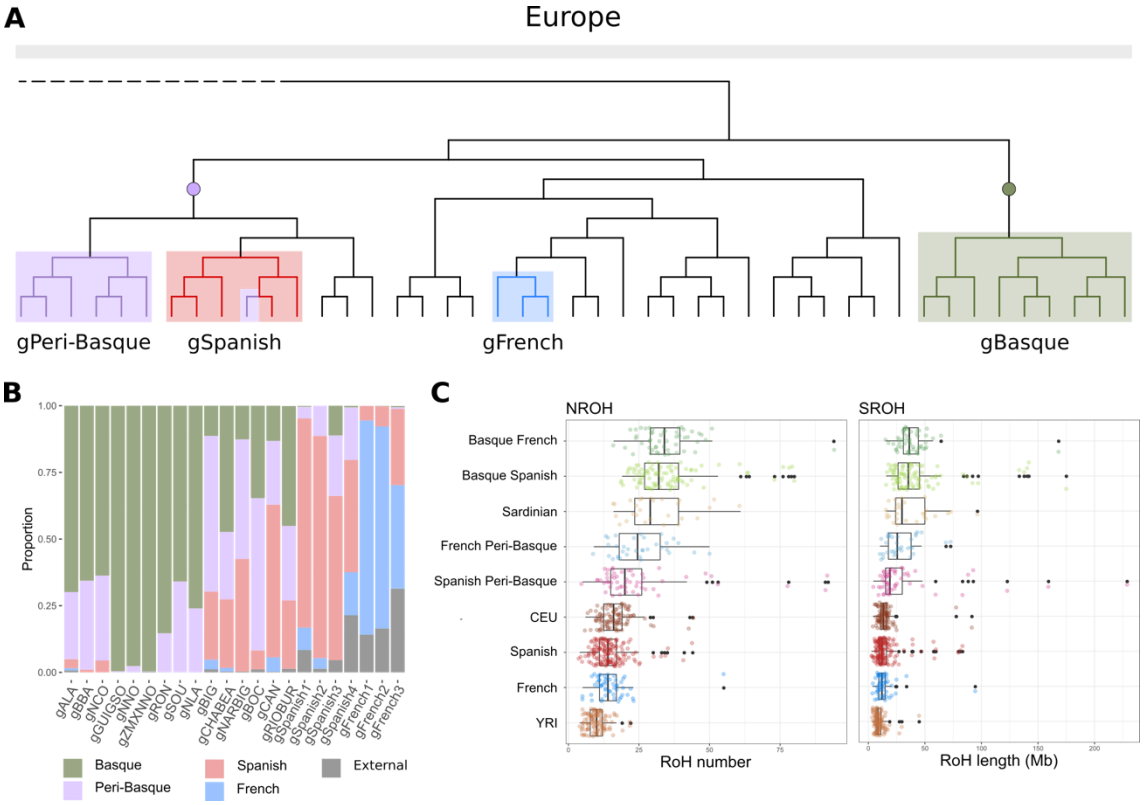

Figure 4

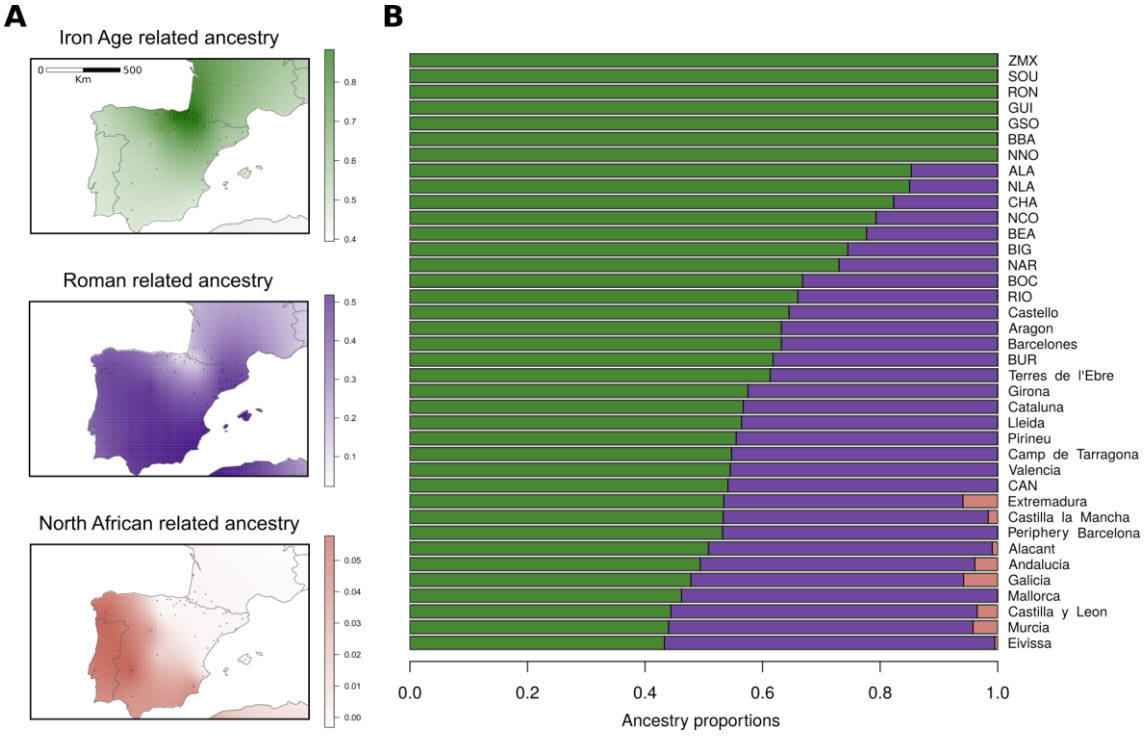

Figure 5

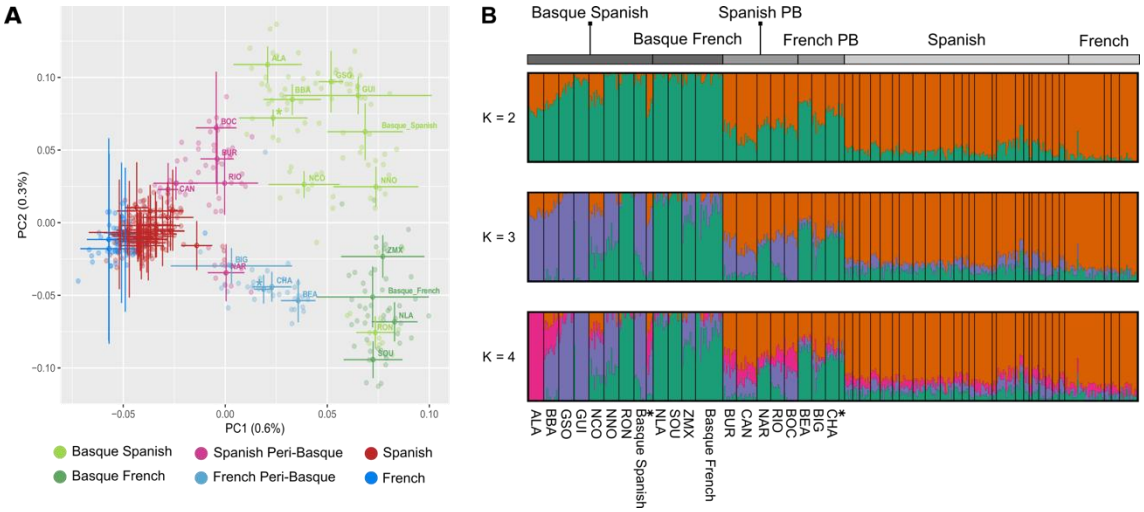

Figure 6