

# SCIENTIFIC REPORTS

OPEN

## Spatially explicit analysis reveals complex human genetic gradients in the Iberian Peninsula

João Pimenta<sup>1,2,3,4,9</sup>, Alexandra M. Lopes<sup>1,2</sup>, Angel Carracedo<sup>5,6</sup>, Miguel Arenas<sup>1,2,7,8</sup>, António Amorim<sup>1,2,4</sup> & David Comas<sup>1,3</sup>

The Iberian Peninsula is a well-delimited geographic region with a rich and complex human history. However, the causes of its genetic structure and past migratory dynamics are not yet fully understood. In order to shed light on them, here we evaluated the gene flow and genetic structure throughout the Iberian Peninsula with spatially explicit modelling applied to a georeferenced genetic dataset composed of genome-wide SNPs from 746 individuals belonging to 17 different regions of the Peninsula. We found contrasting patterns of genetic structure throughout Iberia. In particular, we identified strong patterns of genetic differentiation caused by relevant barriers to gene flow in northern regions and, on the other hand, a large genetic similarity in central and southern regions. In addition, our results showed a preferential north to south migratory dynamics and suggest a sex-biased dispersal in Mediterranean and southern regions. The estimated genetic patterns did not fit with the geographical relief of the Iberian landscape and they rather seem to follow political and linguistic territorial boundaries.

Recent approaches combining geographic information with genetic data provide robust analyses of the population structure across a landscape<sup>1–6</sup>. Essentially, these methods allow the study of patterns of spatially-explicit genetic differentiation by modulating genetic distances between samples or populations as a function of the geographic distance<sup>1,2</sup>.

Studies based on genome-wide autosomal data showed that genetic differentiation among European populations is, in general, highly correlated with geography<sup>7,8</sup>. However, genetic differentiation has also been associated with political-cultural boundaries that limit the gene flow between populations<sup>8–10</sup>. Previous studies have shown that major geographic features often lead to strong genetic differentiation at a global scale<sup>11</sup>, while at a regional/local scale cultural barriers usually modulate the genetic structure of populations<sup>11–14</sup>.

The Iberian Peninsula is separated from the rest of Europe by a range of mountains (the Pyrenees) and from Africa by a small stretch of water (the Strait of Gibraltar). The history of the Iberian Peninsula is characterized by multiple migrations and settlements from diverse population groups. The region was a refuge for populations fleeing from glaciers advance during the Last Glacial Maximum (LGM)<sup>15,16</sup> and was probably a reservoir for the repopulation of Europe after the end of the LGM<sup>17</sup>. Recently, several studies based on ancient DNA shed new light into the genomic history of the Iberian Peninsula. Olalde *et al.*<sup>18</sup> reported the presence of genetic structure in Palaeolithic hunter-gatherer populations. While the adoption of a sedentary lifestyle influenced the genetic variability of Iberians<sup>19–21</sup>, in the Neolithic individuals still carried a strong hunter-gatherer component due to admixture with hunter-gatherer populations<sup>19,22,23</sup>. Post Neolithic migrations also affected the genetic landscape of the region. A large replacement of male ancestry during the Bronze Age was inferred<sup>24</sup> and it was recently associated with the massive immigration of steppe populations at ~4,500 years ago<sup>25,26</sup>. Additionally, two recent studies based on ancient DNA documented gene flow from North Africa toward the Iberian Peninsula at least

<sup>1</sup>Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal. <sup>2</sup>Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal. <sup>3</sup>Institute of Evolutionary Biology (CSIC-UPF). Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain. <sup>4</sup>Faculty of Sciences, University of Porto, Porto, Portugal. <sup>5</sup>Instituto de Ciencias Forenses, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. <sup>6</sup>Grupo de Medicina Xenómica, CIBERER, Santiago de Compostela, Spain. <sup>7</sup>Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain. <sup>8</sup>Biomedical Research Center (CINBIO), University of Vigo, 36310, Vigo, Spain. <sup>9</sup>Present address: Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Universidade do Porto, Vairão, Portugal. Correspondence and requests for materials should be addressed to D.C. (email: [david.comas@upf.edu](mailto:david.comas@upf.edu))

Population	Sample size	Reference
Basque Country, Spain	57	50,51
La Rioja	10	51
Navarre	12	51
Catalonia	93	51
Aragon	24	51
Valencia	16	51
Murcia	13	51
Andalusia	68	30,51
Extremadura	14	51
Castile La Mancha	34	51
Madrid	2	51
Castile and Leon	36	51
Cantabria	10	51
Asturias	11	51
Galicia	177	30,51
Porto	124	49
Lisbon	45	49
<b>Europe</b>		
Finland (FIN)	79	53
Northern and Western Europeans from Utah (CEU)	87	53
Tuscany, Italy (TSI)	96	53
Basque Country, France	24	52
<b>North Africa</b>		
Morocco South	16	30
Morocco North	18	30
Algeria	19	30
Tunisia	18	30
<b>Sub-Saharan Africa</b>		
Yoruba, Nigeria (YRI)	101	53

**Table 1.** Location, sample size and references of the datasets analysed in the present study.

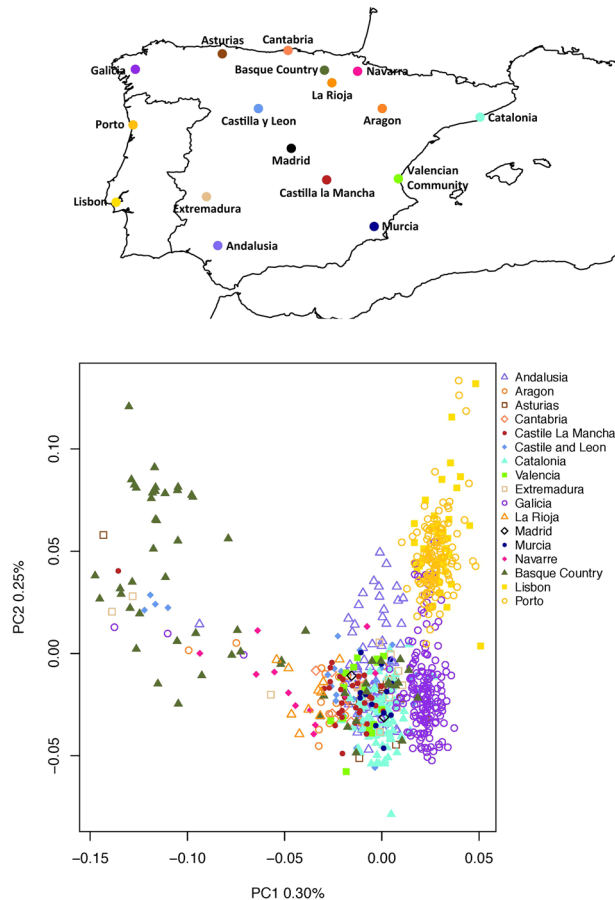
since the Late Neolithic<sup>18,27</sup>. In the last two millennia, the Iberian Peninsula was occupied by Phoenicians, Greeks, Romans, German tribes and, more recently, most of the Peninsula was under Islamic rule (from the beginning of the 8<sup>th</sup> until the end of the 15<sup>th</sup> century) that was revoked with the so-called *Reconquista* where Catholic kingdoms recolonized the territory<sup>28,29</sup>. Current Iberian populations reflect this complex admixture of cultures and genetic backgrounds<sup>9,30</sup>, being one of the regions in Europe with the highest genetic diversity<sup>31</sup>. On the other hand, results based on autosomal markers showed a general homogeneity among Iberian populations<sup>32</sup> with some local differentiation identified with mitochondrial DNA and Y-chromosome data<sup>33–36</sup>. Interestingly, previous studies showed that genetic variation correlates with the geographic distance in the Iberian Peninsula<sup>37,38</sup>. However, some of these studies were based on limited genetic information (for instance, only HLA genes). Altogether, the patterns of genetic variation across the Iberian Peninsula are still not totally clear and an analysis of comprehensive genetic data is required to address this issue.

Here we extended previous works by investigating genetic structure, genetic gradients and migratory dynamics of humans in the Iberian Peninsula at a fine-scale level through the analysis of a genome-wide dataset of 1,204 individuals belonging to 26 populations, based on current state of the art spatially-explicit models. Our results show that the genetic structure in northern Iberia agrees with the political frontiers established during the first centuries of the Reconquista, while the genetic landscape of central and southern regions do not show this association and present large migration corridors especially throughout the Mediterranean coast.

## Results

We compiled genome-wide SNP data of individuals belonging to 26 populations from publicly available resources to generate two datasets: a global dataset with all the compiled populations (Table 1), and a second dataset (hereafter, Iberian dataset) that includes the 17 Iberian populations (Fig. 1). The global dataset was used for an exploratory analysis of the patterns of genetic variation and ancestral components in the Iberian Peninsula at the continental level, while the Iberian dataset was used for a fine-scale analysis of the genetic structure and heterogeneity in the Iberian Peninsula.

**High genetic similarity and mainly European ancestry in the Iberian Peninsula.** In order to explore the presence of population stratification in global and Iberian datasets, we performed a principal component analysis (PCA). PCA results obtained from the global dataset showed a clear genetic differentiation between

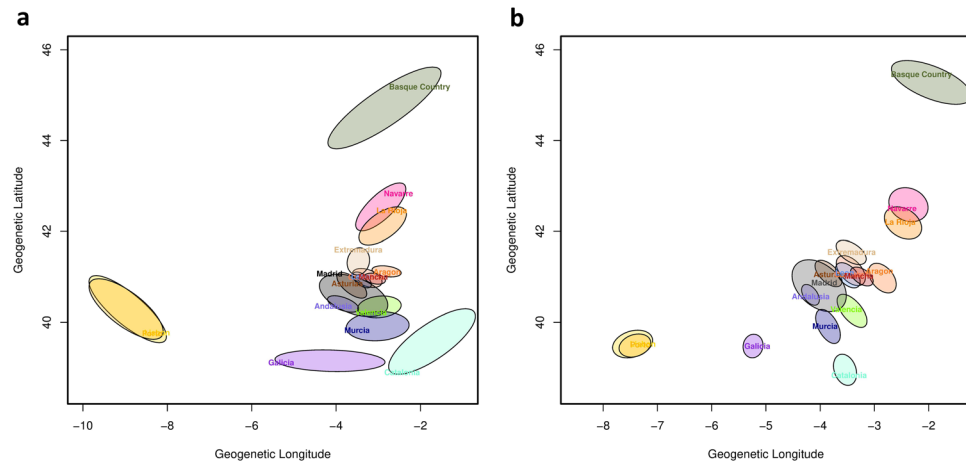


**Figure 1.** Geographic location of the 17 Iberian populations studied and population structure of the Iberian Peninsula. Geographic locations of the analysed data are shown above. The first and second principal components of the Iberian dataset are shown below.

European, North African and sub-Saharan African populations for the first PC (PC1), while PC2 could distinguish between northern, central and southern Europeans (Supplementary Fig. S1a). In both PCs, Iberian populations cluster together. Interestingly, PC3 shows a cluster mainly composed by Iberian Basques and French Basques, both separated from the rest of the Iberian Peninsula (see Supplementary Fig. S1a). On the other hand, PCA results obtained from the Iberian dataset presented only a subtle genetic differentiation throughout the Iberian Peninsula (Fig. 1). In particular, PC1 separated most Basque samples from the rest of the populations and also, Porto, Lisbon, and some Andalusian samples appear separated from other Iberian populations. PC2 separated Portuguese, Galician and Andalusian samples from the rest of the Iberian samples (Fig. 1). Combining both PCs allows to obtain a global picture of genetic differentiation among the Iberian populations. PC3, PC4, and PC5 showed the inner diversity of Iberian populations while highlighting its global homogeneity (see Supplementary Fig. S1b).

We extended the analysis of population stratification by applying an unsupervised clustering algorithm (see Material and Methods) on the global dataset. Results showed that most ancestry of Iberians is shared with other European samples, followed by contributions from Africa (see Supplementary Fig. S2a). When applying the model with the lowest cross-validation error ( $K = 4$ ) most Iberian individuals presented three main ancestral components (see Supplementary Fig. S2b). Two components are associated with European ancestry and one is associated with North African ancestry. It is noteworthy that for Basque individuals the North African ancestral component, which is present in the other Iberian samples, is only vestigial (see Supplementary Fig. S2a).

**Subtle genetic structure in the Iberian Peninsula based on the spatially explicit analysis.** We modelled the geographic structure of the Iberian Peninsula with the Bayesian framework included in the package SpaceMix (see Material and Methods). We found that isolation by distance models considering migration and migration with admixture presented a better fitting with the observed data, compared with models based on pure isolation by distance with and without admixture (see Supplementary Fig. S3). Geogenetic maps inferred under the best fitting models presented remarkable similarities (Fig. 2), which suggest that long-distance migrations followed by admixture events within the region were not a major contributor to the observed genetic structure in the Iberian Peninsula. Nevertheless, the 95% confidence interval ellipses inferred under the model of isolation by distance with migration and admixture (Fig. 2b) are smaller than the ellipses inferred under the model of isolation by distance with migration (Fig. 2a), meaning that allowing for long-range admixture more

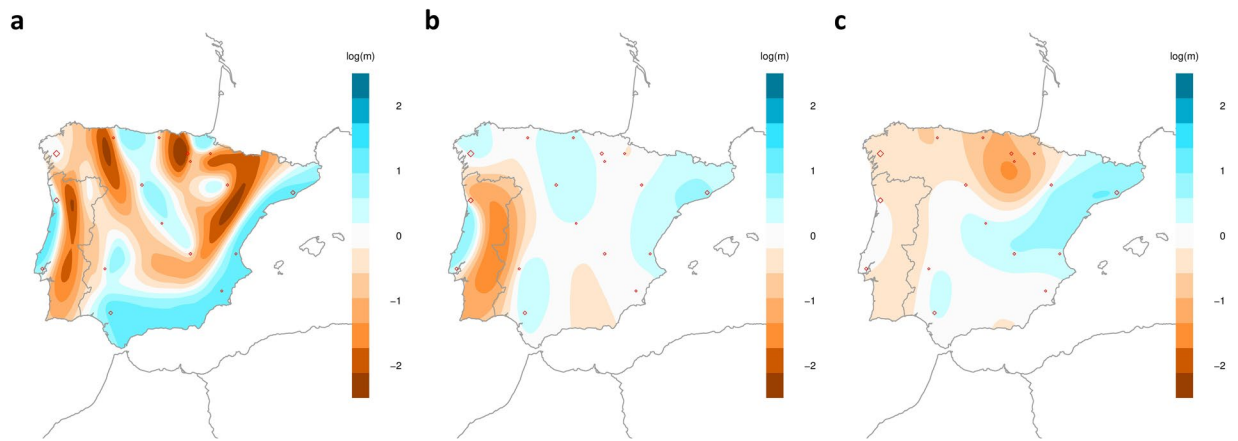


**Figure 2.** Geogenetic locations of the Iberian populations. The maps were inferred under (a) the model of isolation by distance with migration and (b) the model with migration and admixture estimated with SpaceMix. The plot shows the 95% confidence surfaces of the estimated geogenetic maps. Some names were abbreviated as Leon (Castile and Leon) and Mancha (Castile La Mancha).

precisely delimited population location on the geogenetic map. Despite the improvement provided by considering long-distance admixture events, the estimated proportions of admixture in all populations were very low ( $<1\%$ ; see Supplementary Fig. S4). The geogenetic map inferred with the model of isolation by distance with migration and admixture presented 5 distinct population groups based on the 95% confidence surfaces (Fig. 2). The largest genetic divergence was observed between Portuguese and Basque Country populations. Indeed, our results highlighted some genetic isolation of the Galician population with respect to the other populations of the Iberian Peninsula, as well as a genetic isolation of populations of the Basque region (Basque Country, La Rioja and Navarre). The remaining Spanish populations (Northern, Central and Mediterranean) presented close geogenetic proximity, suggesting a high genetic similarity among them (Fig. 2).

The effective migration surface (EEMS) estimated from the autosomes presented several barriers to gene flow (regions of low effective migration rate) splitting the northern regions. It also showed corridors of genetic similarity (areas of high effective migration rate) connecting northern, central and southern regions (Fig. 3a). Populations from the Basque region appeared almost genetically isolated from the rest of the Peninsula. Also, Portuguese populations (southwest) were separated from central Iberian populations by a barrier to gene flow (see Supplementary Fig. S5). To a lesser extent, the region of Galicia presented some isolation from the rest of the Peninsula through barriers with Asturias and with the north of Portugal (see Supplementary Fig. S5). Another barrier was detected separating north and northeast regions (this is, separating Aragon and Catalonia) (see Supplementary Fig. S5). Concerning the opposite pattern (high genetic similarity), we inferred some regions with a high effective migration rate. One of them is a corridor throughout the Mediterranean coast, from the northeast to the south of the Iberian Peninsula. Another corridor presenting genetic similarity was identified connecting the central and northern coast of Portugal. A final migration corridor connected northern regions (Asturias and Cantabria) with central regions of the Iberian Peninsula (Castile and Leon, Madrid, Castile La Mancha) (Fig. 3a). The correlation between the estimated and observed genetic dissimilarities between and within demes ( $R^2$  coefficients of 0.80 and 0.95, respectively) suggested that the EEMS model was robust to describe the observed data (see Supplementary Fig. S6a, b). Indeed, the lack of correlation between geographic distance and genetic distance reveals that a model of isolation by distance cannot explain the population structure observed in the Iberian Peninsula (see Supplementary Fig. S6c).

Concerning the detection of putative sex-biased population structure in the Iberian Peninsula, we compared EEMS results from the autosomes with EEMS results from the X chromosome. EEMS results from the X chromosome revealed some different features concerning the genetic structure in this region, comparison with autosomes (Fig. 3b). Northern Iberian populations showed a stronger genetic structure (lower migration rates) than central and southern Spanish populations (higher migration rates). In agreement with the findings derived from autosomes, the analyses based on the X chromosome also indicated that the area with the highest genetic differentiation was the Basque region (see Supplementary Fig. S7). In particular, a strong barrier to gene flow surrounded the Basque Country and neighbouring populations. In contrast, a large region with high effective migration rate clusters populations from Eastern and Central regions of Iberia (Fig. 3b). Another region with high effective migration rate was identified, connecting the southern regions Andalusia and Extremadura (Fig. 3b). As for the autosomes, analyses based on the X chromosome showed that Portuguese populations are clustered together and are isolated from Spanish populations (Fig. 3b). Diagnostic plots for model fitting (see Supplementary Fig. S8a, b) showed that EEMS results present an excellent fitting with the data concerning dissimilarities within demes ( $R^2 = 1.00$ ) but the fitting was weak when considering dissimilarities between demes ( $R^2 = 0.373$ ). Similarly to the what was found with the autosomes, a significant deviation from an isolation by distance model ( $R^2$  coefficient = 0.01) was found when the observed pairwise genetic distances between populations were related with geographic distances (see Supplementary Fig. S8c). To test the robustness of the differences between X chromosome



**Figure 3.** Effective Migration Maps estimated for the Iberian dataset. The maps were inferred from autosomes (a), X chromosome (b) and chromosome 7 (c). The plots were estimated with EEMS under a  $\log_{10}$  scale and after mean centering. Blue regions indicate areas with effective migration rate higher than average (high genetic similarity), while brown colours indicate regions with a lower effective migration (compared to the average) between demes (high genetic differentiation).

and autosomal analysis, we compared the differentiation patterns estimated for chromosome 7 with those from all the autosomes. Our results showed similarities in the patterns of genetic differentiation, even despite the poorer resolution due to the smaller amount of available data (Fig. 3c). Considering chromosome 7, Portuguese populations are genetically similar and isolated from the rest of Iberia (see Supplementary Fig. S9), and northern and central Iberian populations presented a high genetic similarity (Fig. 3c). Moreover, populations from the Basque region also showed higher degrees of genetic differentiation with respect to surrounding regions (Fig. 3c), in agreement with the analyses of all the autosomes (Fig. 3a). The most relevant difference between the results derived from chromosome 7 and from the autosomes is that chromosome 7 showed a barrier to gene flow separating Extremadura and Andalusia regions from other Mediterranean regions (Fig. 3c) which was not found in the analyses based on all the autosomes (Fig. 3 and see Supplementary Fig. S5). Diagnostic plots based on chromosome 7 showed that EEMS results present a reasonable fitting with the data in terms of dissimilarities within demes ( $R^2 = 0.547$ ) (see Supplementary Fig. S10a, b), but a weak fitting when exploring dissimilarities between demes ( $R^2 = 0.107$ ). Additionally, a deviation from an isolation by distance model ( $R^2 = 0.074$ ) was found when the observed pairwise genetic distances between populations were correlated with geographic distances (see Supplementary Fig. S10c). Taking into account the results from chromosome 7, we believe that our findings based on the X chromosome are not biased due to the smaller sequence length analysed but they should be interpreted with caution since the patterns found for the autosomes could only be partially replicated by the chromosome 7 analysis.

To further explore sex-biased patterns of genetic differentiation we applied the SpaceMix framework on the X chromosome data. However, none of the models implemented in this framework could accurately describe the pattern of decay of genetic covariance present in the observed data (results not shown).

## Discussion

The presence of genetic structure in the Iberian Peninsula has been described based on the Y chromosome and mtDNA at a regional level<sup>9,33,36,39</sup>. Here we extended those studies by considering a more comprehensive dataset of genome-wide genetic information, analysing autosomes and, for the first time, the X chromosome. We found that the characterization of the genetic landscape of the Iberian Peninsula, with spatially explicit approaches, presents subtle structure features. Patterns of genetic differentiation were largely observed along a longitudinal orientation (Fig. 3), in agreement with findings from other genetic markers<sup>38</sup>. Indeed, strong genetic differentiation was observed in northern regions of the Iberian Peninsula (Figs 2 and 3), while corridors of genetic similarity mainly appeared along a latitudinal orientation, along the Mediterranean and Atlantic shores (Fig. 3). Remarkably, we did not find an agreement between our results and the geographic relief of the Iberian landscape (see Supplementary Fig. S11), which suggests that geographic features did not have a major influence on the genetic patterns observed in Iberia. However, the patterns of genetic structure found in northern regions of the Iberian Peninsula are compatible with political and linguistic boundaries associated with the Catholic kingdoms formed during the first two centuries of the *Reconquista* (Fig. 3). This interpretation is in agreement with previous results based on haplotype data of Spanish samples<sup>40</sup>. However, the findings for northern regions contrasted with those for central and southern regions which presented a much more homogeneous genetic structure. In particular, we found that central and Mediterranean populations are genetically similar (Fig. 2) with only a barrier to gene flow separating Mediterranean and Extremadura populations from all the other Iberian populations (Fig. 3a). Interestingly, our results showed that Mediterranean populations present a high genetic similarity (see Supplementary Fig. S5), which is agreement with recent results by Olalde *et al.*<sup>18</sup> showing that during the Roman period and onwards southern and Mediterranean populations had an influx of genes from southern Europe and North Africa, most likely reflecting the mobility by land and sea during the Roman empire<sup>41</sup>, the commercial

trade across the Mediterranean Sea<sup>29,42</sup> and the Islamic occupation of the Iberian Peninsula during 8 centuries<sup>28</sup>. In addition, our results for the X chromosome suggest that the current genetic structure in the Iberian Peninsula was influenced by sex-biased migrations, given that different genetic structures were found when analysing separately the X chromosome and the autosomes (e.g., absence of genetic differentiation between the regions of Catalonia and Aragon and lack of structure in Central and Southern regions of the Peninsula). However, as previously indicated, these results should be interpreted with caution since the patterns found for the autosomes could only be partially replicated when analysing only chromosome 7 markers, and further studies will be necessary to clarify this issue.

Populations from the Basque region (Basque Country, La Rioja and Navarre) showed a genetic distinctiveness from the other Iberian populations for both autosomes (Figs 1, 2a and S5) and X chromosome (Fig. 3b). This genetic differentiation could be caused by cultural aspects since Basques are characterized by their unique non-Indo-European language and limited gene flow from other populations outside Iberia such as north Africans, as shown in our analyses and also previously reported<sup>43,44</sup>. Additionally, the genetic similarity found between Spanish and French Basques using both PCA (Supplementary Fig. S1a) and ancestry profiles (Supplementary Fig. S2) could be explained by shared cultural traditions between those regions.

The northwest region of the Peninsula (Galicia) presented a higher than average genetic differentiation when compared to other Spanish populations (Figs 2 and 3). As for Basque populations, cultural and linguistic differences could account for this genetic divergence<sup>45</sup>. Moreover, a study on marital behaviour showed a high proportion of inbreeding that could lead to genetic differentiation from other Iberian populations<sup>46</sup>. Interestingly, Portugal and Galicia may share their ancestral history<sup>40</sup> and we also found results supporting this hypothesis. The geogenetic map (Fig. 2) shows that Galicia is, genetically, the Spanish region closest to Portugal, despite the larger geogenetic distance estimated between Galicia and Portugal with respect to the distance between Galicia and the central regions of Spain. Additionally, the estimated effective migration surface for autosomes (Fig. 3) suggests a small Atlantic coastal corridor of gene flow connecting Galicia and northern Portugal, which can be attributed to the long historical relationship between these regions. Indeed, before the Islamic invasion in the 8<sup>th</sup> century, both regions belonged to the Roman province of Gallaecia and later on to the kingdom of the Suebi (405 CE and 585 CE), before the annexation by the Visigoths<sup>47,48</sup>. Portugal became politically independent in 1143 and expanded rapidly toward the south (Portuguese *Reconquista* ended by 1249). The establishment of a political border could have promoted some cultural divergences but still important relationships were kept between Galicia and Portugal because of their geographic proximity, similar language and sociological factors. A recent study based on genome data but applying other approaches also showed remarkable genetic similarities between these regions<sup>40</sup>, in agreement with our findings.

In conclusion, we found that the genetic landscape across the Iberian Peninsula is complex, with contrasting patterns of remarkable genetic dissimilarity in the North and genetic homogeneity in the South. We interpret our findings considering that the geography is not the main factor shaping the genetic landscape of the Iberian Peninsula. Instead the major genetic dissimilarities estimated from our data better fitted with historical, political and cultural barriers that influenced migratory patterns and the relationships between populations.

## Materials and Methods

**Data and genotyping.** We examined a genome-wide SNP dataset genotyped on the Affymetrix 6.0 chip, composed of 26 populations reported in several previous studies: 17 Iberian populations (2 Portuguese populations retrieved from Lopes *et al.*<sup>49</sup> and 15 Spanish populations retrieved from Botigue *et al.*<sup>30</sup>, Henn *et al.*<sup>50</sup> and Fernandez-Rozadilla *et al.*<sup>51</sup>), and 4 European populations (French Basques from the Human Genome Diversity Panel<sup>52</sup>, northern Europeans (CEU), Tuscans (TSI) and Finns (FIN) obtained from the 1000 Genomes Project<sup>53</sup>). Out of Europe we included 4 North African populations (Algeria, Morocco North, Morocco South and Tunisia) retrieved from Henn *et al.*<sup>50</sup> and one sub-Saharan (Yoruba) from the 1000 Genomes Project<sup>53</sup>.

A quality control filter was applied using PLINK 1.9<sup>54</sup>. For each population, we excluded SNPs with missing genotype rate >10%, and those that failed Hardy-Weinberg equilibrium under a threshold of 0.05. We also excluded individuals with a missing rate >10% and those individuals that shared an identity-by-state >85%. In addition, after merging all populations, SNPs with a minor allele frequency (MAF) <0.05 were excluded. After applying quality control filters, the global dataset and the Iberian dataset presented a total of 1,204 and 746 individuals, respectively (Table 1). Additionally, SNPs were pruned using PLINK 1.9 with a sliding window of 50 kb, a shift step of 5 SNPs and a LD threshold of 0.5. We finally obtained a total of 64,302 and 174,001 SNPs for the global and the Iberian datasets, respectively (Supplementary Table S1). For the X chromosome, we excluded SNPs for both pseudoautosomal regions (PAR1 and PAR2) and SNPs heterozygous in the X specific region, keeping a total of 4,792 SNPs (Supplementary Table S1).

**Analysis of the population structure in the Iberian Peninsula.** We performed the PCA with the *smartPCA* algorithm implemented in EIGENSOFT v5.0.1<sup>55</sup> software. Additionally, we applied ADMIXTURE v1.3.0<sup>56</sup> under unsupervised mode, testing from  $K=2$  to  $K=10$  ancestral clusters. We ran 10 replicates with different random seeds and merged the outputs from different runs. Results were then depicted with Distruct1.1<sup>57</sup>.

**Analysis of the spatial structure in the Iberian Peninsula.** We investigated patterns of isolation by distance and gene flow throughout the Iberian Peninsula for the autosomes in the Iberian dataset with the Bayesian framework SpaceMix<sup>1</sup>. This analysis provides genetic relationships between populations through a geogenetic map in which geographical distances between populations are related with genetic distances. SpaceMix implements four different types of models: (1) a pure isolation by distance model, in which populations are stationary (absence of migration) and do not present admixture; (2) a model of isolation by distance with admixture, in which populations are stationary (absence of migration) and can present admixture; (3) a model of isolation by

distance with migration, in which population locations are inferred (allowing migration) and cannot present admixture; (4) a model of isolation by distance with migration and admixture, in which population locations are inferred (allowing migration) and can present admixture. For each model we ran 10 independent short chains of  $10^6$  iterations each, followed by a long chain of  $10^7$  iterations based on the estimates of the last iteration of the short chain with the highest posterior probability. A sample was taken every  $10^4$  iterations leading to a total of 1,000 points to estimate the posterior distribution of each parameter. Initial population locations were randomly taken from a uniform distribution of  $-180$  to  $180$  and  $-90$  to  $90$  for longitude and latitude, respectively.

We also identified patterns of spatial structure and genetic heterogeneity within the Iberian Peninsula applying the framework EEMS<sup>2</sup> to the Iberian dataset. We estimated migration rate surfaces allowing the visualization of corridors and barriers to gene flow. Basically, EEMS considers the stepping-stone migration model to infer migration rates through a Bayesian approach<sup>2</sup>. The method applies a dense triangular grid that fills the entire landscape and assigns each individual to the geographic neighbour deme of each population to finally provide a map quantifying genetic dissimilarities. We also estimated a matrix of genetic dissimilarities between all 746 individuals with the *bed2cliffs* method implemented in the EEMS package. For all the analyses we specified a total of 1,000 demes and we performed 5 independent runs with  $1.1 \times 10^7$  iterations, a thinning interval of 1,000 iterations and  $10^6$  iterations as burn-in. Finally, we adjusted migration and diversity parameters to model acceptance rates between 10–40% following the software documentation.

**Sex-biased population structure.** In order to test the presence of sex-biased migration in the Iberian Peninsula, we compared the patterns of genetic differentiation of the X chromosome with the autosomes applying the EEMS framework. Additionally, we analysed the patterns of genetic differentiation only on chromosome 7, which presents a size similar to the X chromosome (around 150 million bases), to ensure that the higher linkage on the X chromosome or the lower number of SNPs did not affect the results. Therefore we trimmed chromosome 7 to have a similar SNP density to the X chromosome (4,792 SNPs for the X chromosome and 4,755 SNPs for chromosome 7) by applying a LD threshold of 0.205 with PLINK 1.9 (Supplementary Table S1). We ran EEMS using the settings previously used to analyse the autosomes (see above).

## Data Availability

The data from North African populations and some Iberian populations can be found at <http://bhusers.upf.edu/dcomas/>. The data from Portuguese and Spanish populations is available upon request to [alopes@ipatimup.pt](mailto:alopes@ipatimup.pt) and [angel.carracedo@usc.es](mailto:angel.carracedo@usc.es), respectively.

## References

- Bradburd, G. S., Ralph, P. L. & Coop, G. M. A Spatial Framework for Understanding Population Structure and Admixture. *PLoS Genet.* **12**, e1005703 (2016).
- Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
- Duforet-Frebourg, N. & Blum, M. G. B. Nonstationary patterns of isolation-by-distance: Inferring measures of local genetic differentiation with bayesian kriging. *Evolution (N. Y.)* **68**, 1110–1123 (2014).
- Hanks, E. M. & Hooten, M. B. Circuit theory and model-based inference for landscape connectivity. *J. Am. Stat. Assoc.* **108**, 22–33 (2013).
- Arenas, M., François, O., Currat, M., Ray, N. & Excoffier, L. Influence of admixture and paleolithic range contractions on current European diversity gradients. *Mol. Biol. Evol.* **30**, 57–61 (2013).
- Branco, C. & Arenas, M. Selecting among Alternative Scenarios of Human Evolution by Simulated Genetic Gradients. *Genes (Basel)* **9**, (506 (2018)).
- Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
- Lao, O. *et al.* Correlation between Genetic and Geographic Structure in Europe. *Curr. Biol.* **18**, 1241–1248 (2008).
- Adams, S. M. *et al.* 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 (2008).
- Barbujani, G. & Sokal, R. R. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl. Acad. Sci. USA* **87**, 1816–1819 (1990).
- Peter, B. M., Petkova, D. & Novembre, J. Genetic landscapes reveal how human genetic diversity aligns with geography. Preprint at, <https://www.biorxiv.org/content/10.1101/233486v2> (2018).
- Messina, F. *et al.* Spatially explicit models to investigate geographic patterns in the distribution of forensic STRs: Application to the north-eastern mediterranean. *PLoS One* **11** (2016).
- Uren, C. *et al.* Fine-scale human population structure in Southern Africa reflects ecogeographic boundaries. *Genetics* **204**, 303–314 (2016).
- Jeong, C. *et al.* A longitudinal cline characterizes the genetic structure of human populations in the Tibetan plateau. *PLoS One* **12**, e0175885 (2017).
- Achilli, A. *et al.* 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 (2004).
- Torroni, A. *et al.* A Signal, from Human mtDNA, of Postglacial Recolonization in Europe. *Am. J. Hum. Genet.* **69**, 844–852 (2001).
- Pereira, L. *et al.* High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res.* **15**, 19–24 (2005).
- Olalde, I. *et al.* The genomic history of the Iberian Peninsula over the past 8000 years. *Science (80-.)* **363**, 1230 LP–1234 (2019).
- Szécényi-Nagy, A. *et al.* The maternal genetic make-up of the Iberian Peninsula between the Neolithic and the Early Bronze Age. *Sci. Rep.* (2017).
- Olalde, I. *et al.* A common genetic origin for early farmers from mediterranean cardial and central european LBK cultures. *Mol. Biol. Evol.* **32**, 3132–3142 (2015).
- Günther, T. *et al.* Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci.* **112**, 11917 LP–11922 (2015).
- Villalba-Mouco, V. *et al.* Survival of Late Pleistocene Hunter-Gatherer Ancestry in the Iberian Peninsula. *Curr. Biol.* (2019).
- Valdiosera, C. *et al.* Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. *Proc. Natl. Acad. Sci.* **115**, 3428 LP–3433 (2018).
- Olalde, I. *et al.* The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**, 190 (2018).

25. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167 (2015).
26. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207 (2015).
27. González-Fortes, G. *et al.* A western route of prehistoric human migration from Africa into the Iberian Peninsula. *Proc. R. Soc. B Biol. Sci.* **286** (2019).
28. Carr, R. *Spain: a history*. (Oxford University Press, 2000).
29. Prag, J. R. W. & Quinn, J. C. *The Hellenistic west: Rethinking the ancient mediterranean. The Hellenistic West: Rethinking The Ancient Mediterranean* (Cambridge University Press, 2011).
30. Botigue, L. R. *et al.* Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc. Natl. Acad. Sci.* **110**, 11791–11796 (2013).
31. Wang, C., Zöllner, S. & Rosenberg, N. A. A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS Genet.* **8**, e1002886 (2012).
32. Gayán, J. *et al.* Genetic Structure of the Spanish Population. *BMC Genomics* **11**, 326 (2010).
33. Pardiñas, A. F., Roca, A., García-Vázquez, E. & López, B. Assessing the Genetic Influence of Ancient Sociopolitical Structure: Micro-differentiation Patterns in the Population of Asturias (Northern Spain). *PLoS One* **7**, e50206 (2012).
34. López-Parra, A. M. *et al.* 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 (2009).
35. Brion, M. *et al.* Micro-geographical differentiation in Northern Iberia revealed by Y-chromosomal DNA analysis. *Gene* **329**, 17–25 (2004).
36. Alvarez, L. *et al.* Mitochondrial DNA patterns in the Iberian Northern plateau: Population dynamics and substructure of the Zamora province. *Am. J. Phys. Anthropol.* **142**, 531–539 (2010).
37. Baran, Y., Quintela, I., Carracedo, Á., Pasaniuc, B. & Halperin, E. Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am. J. Hum. Genet.* **92**, 882–894 (2013).
38. Romón, I. *et al.* Mapping the HLA diversity of the Iberian Peninsula. *Hum. Immunol.* **77**, 832–840 (2016).
39. Hernández, C. L. *et al.* The distribution of mitochondrial DNA haplogroup H in southern Iberia indicates ancient human genetic exchanges along the western edge of the Mediterranean. *BMC Genet.* **18** (2017).
40. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat. Commun.* **10**, 551 (2019).
41. Kolb, A. In *Oxford Handbook of Roman Epigraphy* (eds Bruun, C. & Edmondson, J.) 649–670 (Oxford University Press, 2014).
42. Bierling, M. R. & Gitin, S. *The Phoenicians in Spain: an archaeological review of the eighth-sixth centuries B.C.E.: a collection of articles translated from Spanish*. (Eisenbrauns, 2002).
43. Martínez-Cruz, B. *et al.* Evidence of pre-roman tribal genetic structure in basques from uniparentally inherited markers. *Mol. Biol. Evol.* **29**, 2211–2222 (2012).
44. Behar, D. M. *et al.* The Basque paradigm: Genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-neolithic times. *Am. J. Hum. Genet.* **90**, 486–493 (2012).
45. Mackenzie, D. *Encyclopedia of the Languages of Europe*. (Blackwell Publishing, 2000).
46. Varela, T. A., Aínsua, R. L. & Fariña, J. Consanguinity in the Bishopric of Ourense (Galicia, Spain) from 1900 to 1979. *Ann. Hum. Biol.* **30**, 419–433 (2003).
47. Barreiro Fernández, X. R., Diaz-Fierros, F. & Fabra Barreiro, G. *Los Gallegos*. (Istmo, 1984).
48. Mattoso, J. *História de Portugal - Antes de Portugal - Vol. I* (1994).
49. Lopes, A. M. *et al.* Human Spermatogenic Failure Purges Deleterious Mutation Load from the Autosomes and Both Sex Chromosomes, including the Gene DMRT1. *PLoS Genet.* **9**, e1003349 (2013).
50. Henn, B. M. *et al.* Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* **8**, e1002397 (2012).
51. Fernandez-Rozadilla, C. *et al.* A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. *BMC Genomics* **14**, 55 (2013).
52. Li, J. Z. *et al.* Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* (80-). **319**, 1100–1104 (2008).
53. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
54. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
55. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).
56. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
57. Rosenberg, N. A. DISTRUCT: a program for the graphical display of population structure. *Mol Ecol Notes* **4**, 137–138 (2004).

## Acknowledgements

IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT in the framework of the project “Institute for Research and Innovation in Health Sciences” (POCI-01-0145-FEDER-007274). J.P. and A.M.L. are funded by the Portuguese Government through the FCT fellowship SFRH/BD/97200/2013 and the research contract IF/01262/2014, respectively. M.A. was supported by the “Ramón y Cajal” grant RYC-2015-18241 from the Spanish Government. D.C. was supported by the Spanish grant CGL2016-75389-P (AEI, MINEICO/FEDER, UE), and “Unidad Maria de Maeztu” funded by the MINECO (MDM-2014-0370).

## Author Contributions

J.P. and D.C. designed the study. A.L., A.C. and D.C. collected the data. J.P. performed the experiments. J.P., M.A., A.A. and D.C. analysed the data. J.P., M.A., A.A. and D.C. contributed to the interpretation of the results. J.P., M.A., A.A. and D.C. wrote the manuscript. All authors reviewed the manuscript and accepted it for publication.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-019-44121-6>.

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019