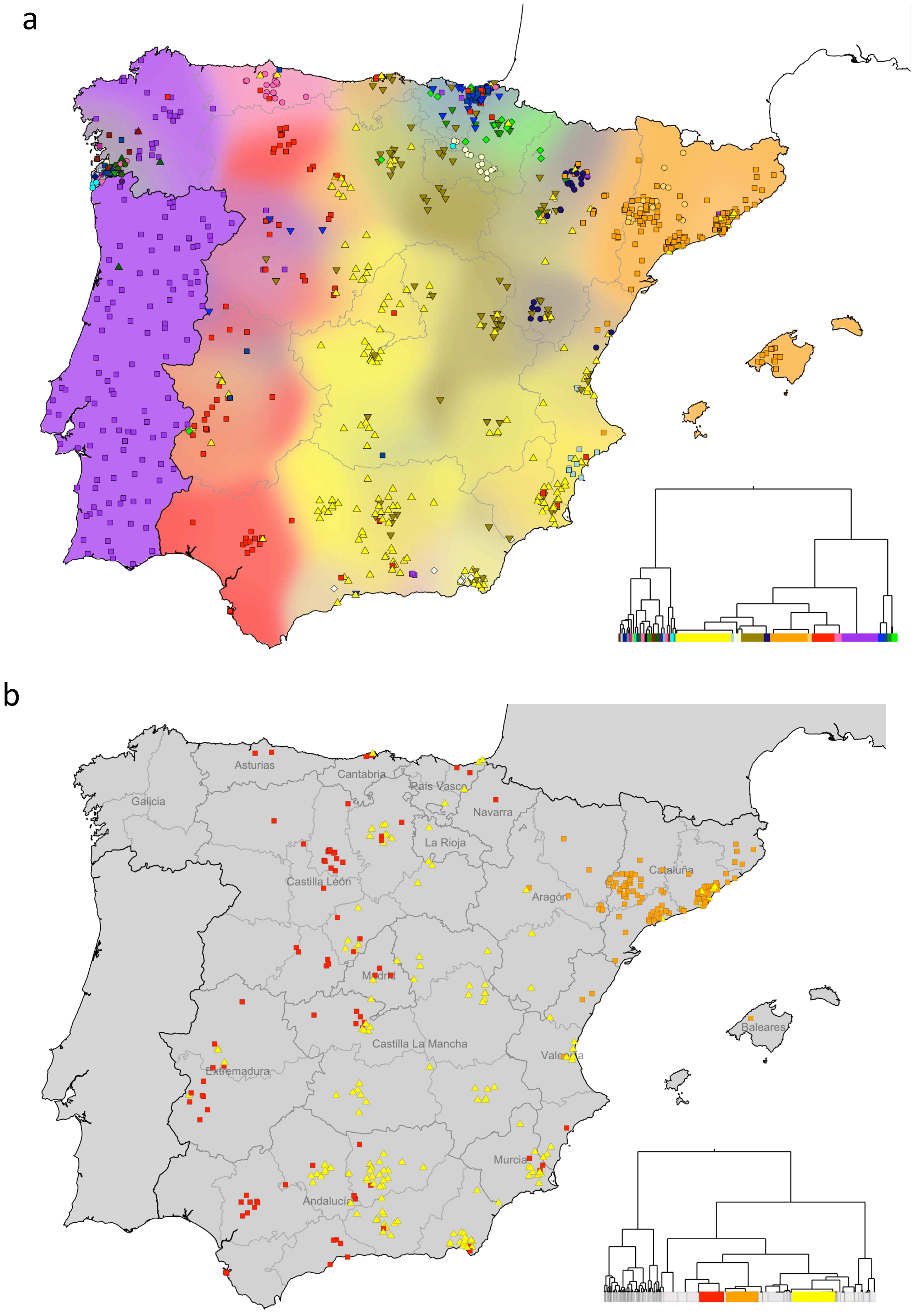
Supplementary Figures

Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula

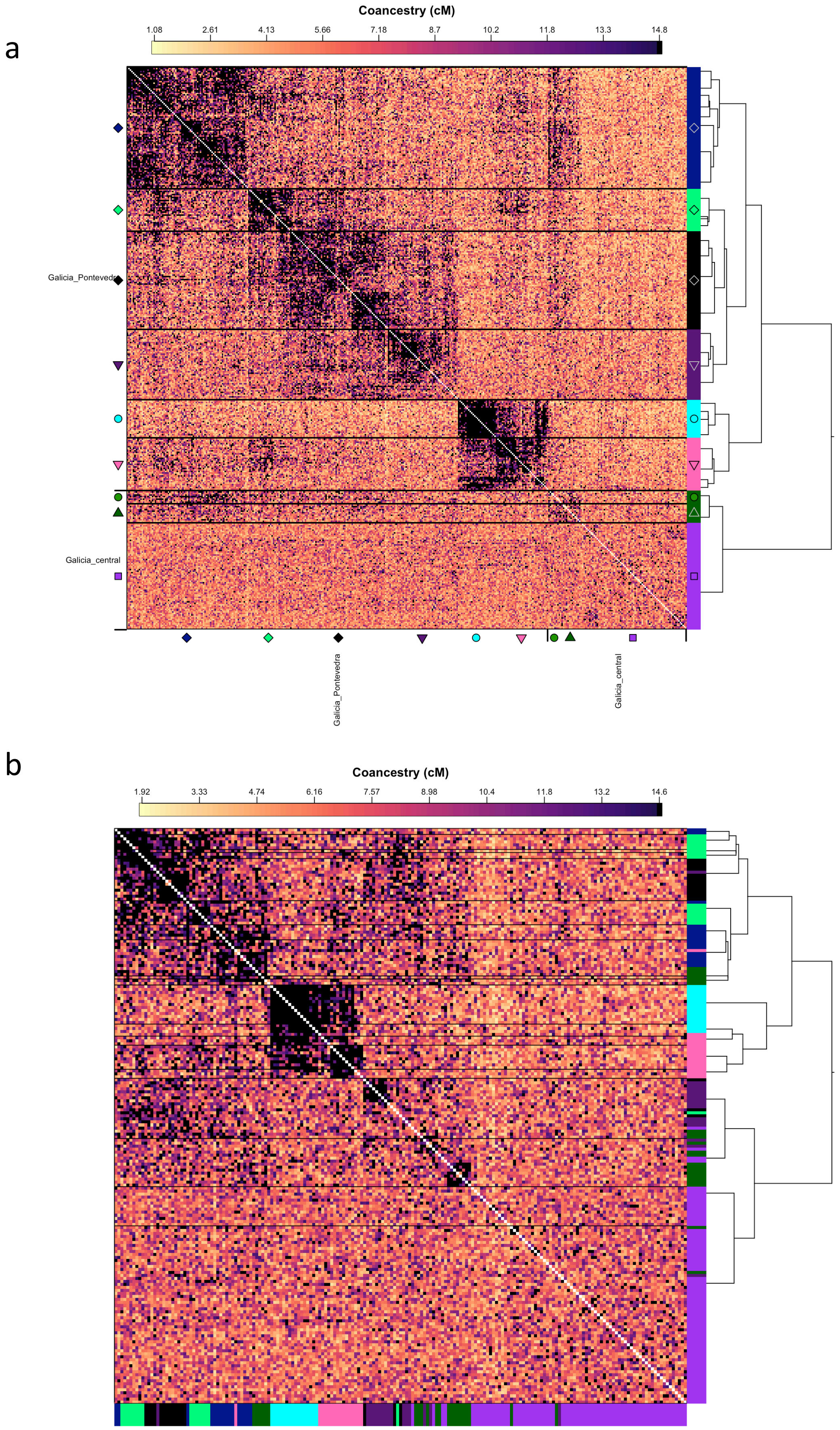
Clare Bycroft[1], Ceres Fernandez-Rozadilla[2], Clara Ruiz-Ponte[2], Inés Quintela-García[2,3], Ángel Carracedo[2,3], Peter Donnelly[1,4†], Simon Myers[4,1†‡]

1) Wellcome Centre for Human Genetics, University of Oxford
2) Fundación Pública Galega de Medicina Xenómica- CIBERER-IDIS, Santiago de Compostela
3) Grupo de Medicina Xenómica, Centro Nacional de Genotipado (CEGEN-PRB2-ISCIII). Universidade de Santiago de Compostela
4) Department of Statistics, University of Oxford

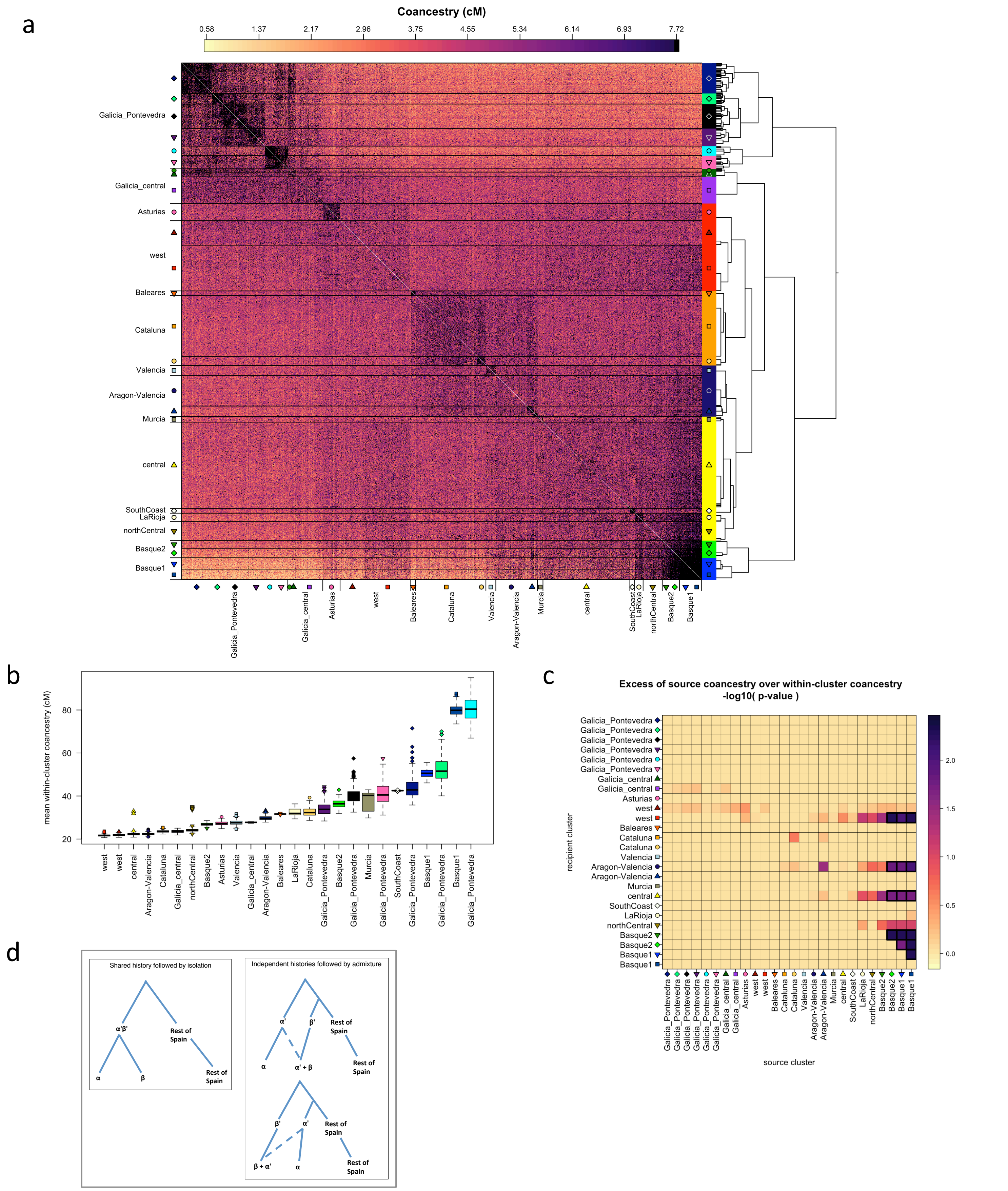
† These authors jointly directed this work.
‡ To whom correspondence should be addressed: myers@stats.ox.ac.uk



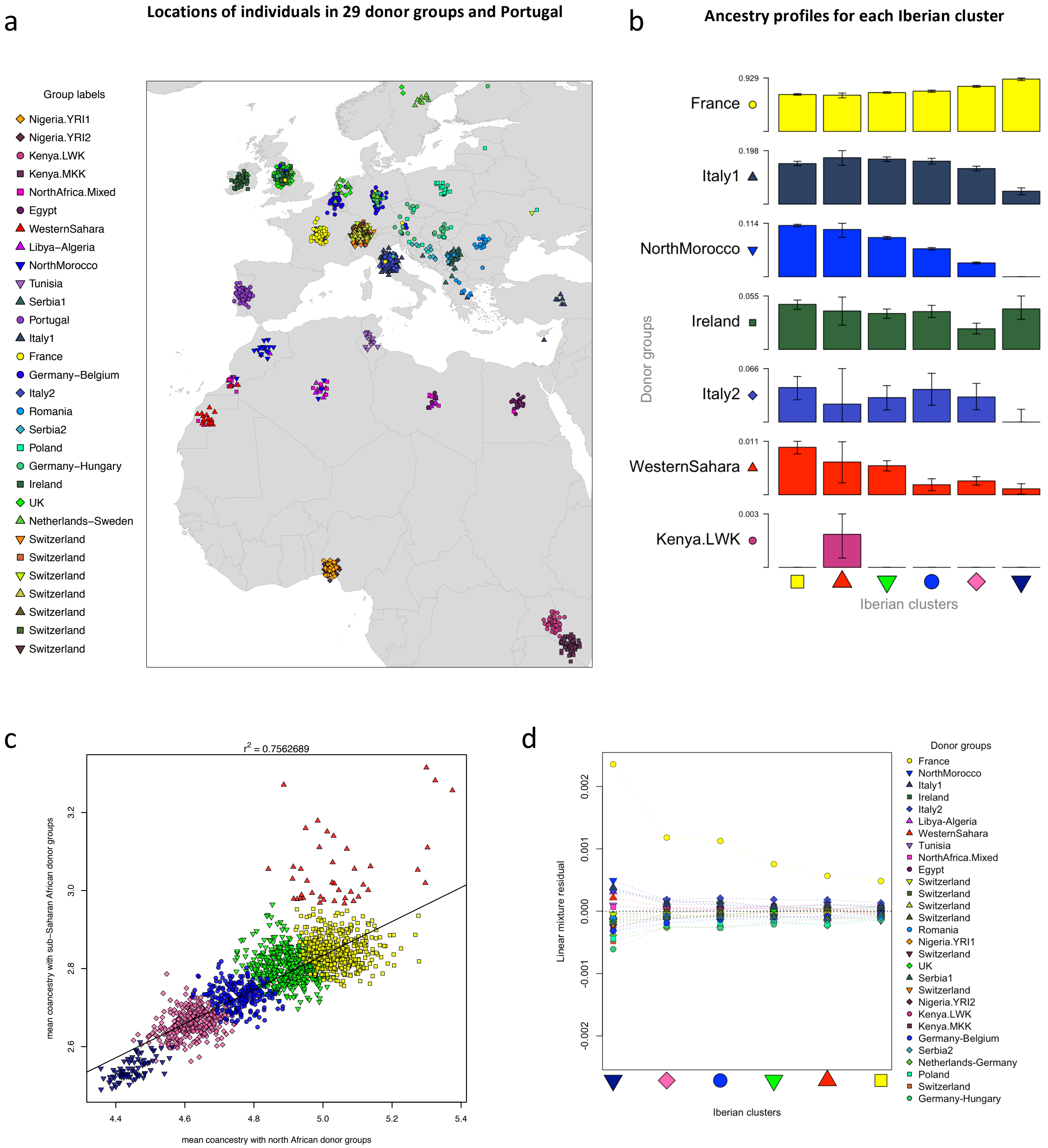
Supplementary Figure 1 | Clustering analysis including Portuguese individuals and large clusters at the bottom of the tree. (a) This map and tree show the final set of clusters inferred in the fineSTRUCTURE analysis (B) that included data from Portuguese individuals but using a smaller set of SNPs (Methods). Positions of points and background colours are determined using the same procedure as for **Figure 1b**, with the exception of Portugal. There was no fine-scale geographic information available for the Portuguese individuals (sourced from POPRES) so they are randomly assigned a position within the borders of Portugal. The background colour within Portugal is determined by assuming all individuals contribute the same weight to each grid-point. **(b)** This map shows geographic spread of the three large clusters that remain at the bottom of the tree inferred in the Spain-only fineSTRUCTURE analysis (see main text; **Figure 1a**). These clusters contains more than 100 individuals out of the full set of 1,413. The accompanying tree highlights the three clusters within the full tree structure. The width of the coloured rectangles is proportional to the number of individuals belonging to each cluster.



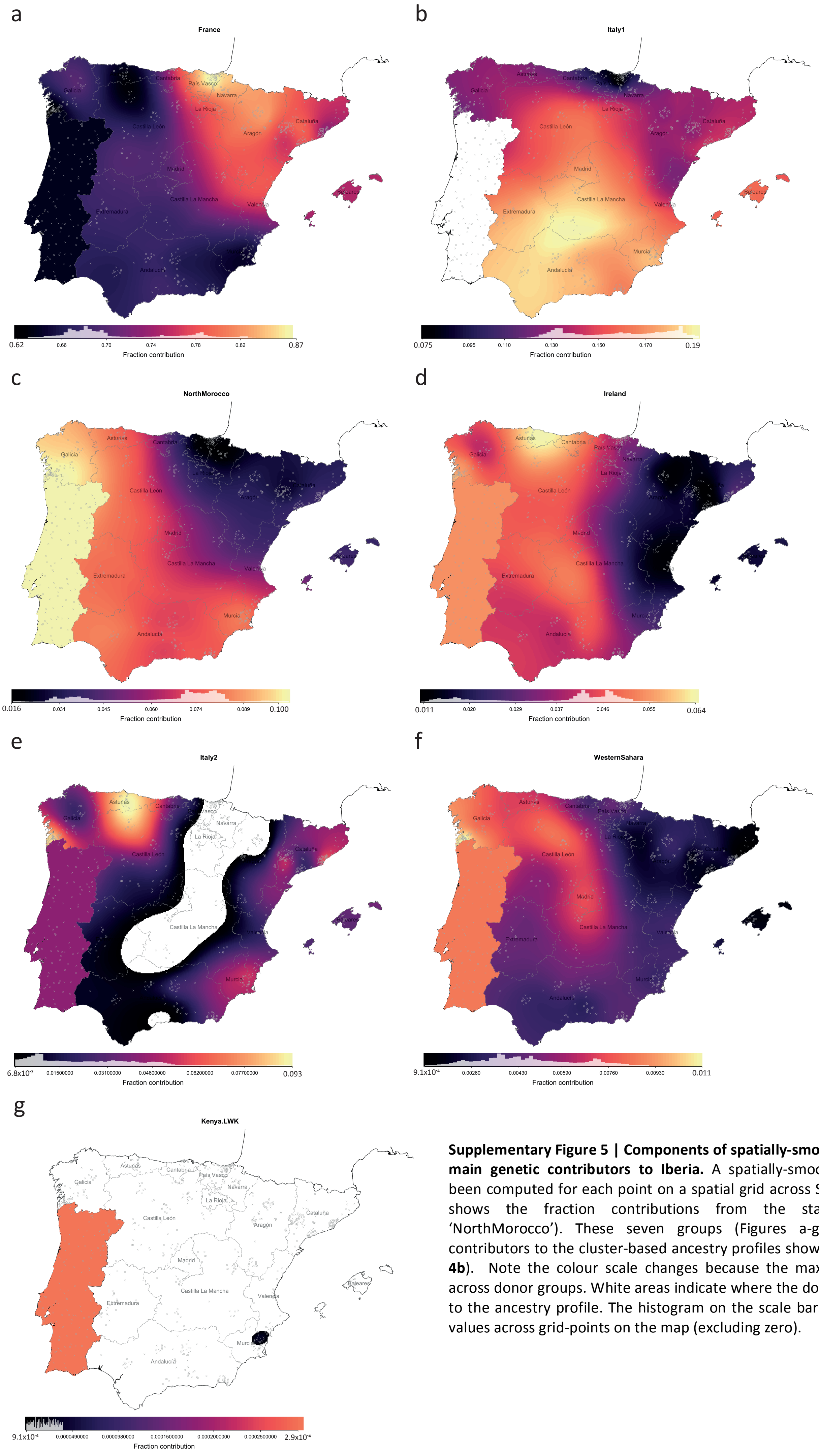
Supplementary Figure 2 | Effect of sub-sampling on fine-scale structure in Galicia. We tested the effect of high density sampling in the region of south-west Galicia by conducting a fineSTRUCTURE analysis on a subset of individuals such that the number of individuals in the set of clusters located in south-west Galicia, labelled in a) as 'Galicia_Pontevedra', was the same as the number of individuals in the clusters located in inland Galicia (see Methods). **(a)** Section of the coancestry matrix shown in **Supplementary Figure 3a** that involves the clusters located primarily in Galicia. **(b)** Coancestry matrix and fineSTRUCTURE tree inferred after sub-sampling (see Methods). The colours in the axes indicate which of the clusters each individual belongs to in the original analysis, as shown in (a).



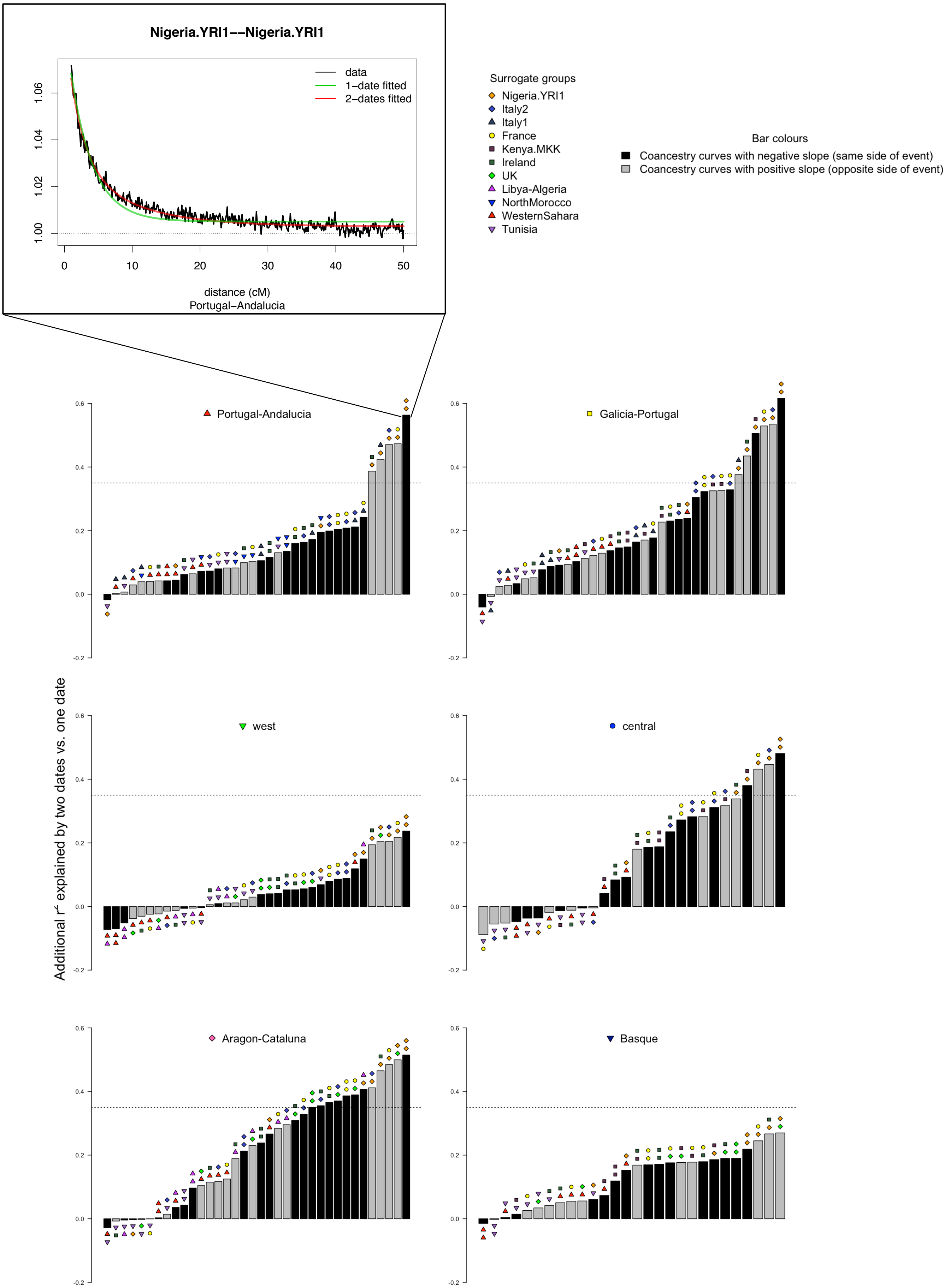
Supplementary Figure 3 | Estimates of shared ancestry between each Spanish individuals and across fineSTRUCTURE clusters. (a) Matrix of coancestry values used in cluster inference. Each individual is represented as a row, where each element is the coancestry (in cM) shared with each of the other individuals (see Methods for the definition of coancestry). In order to visualise the bulk of the variation, values equal to or above the 90th percentile (7.7 cM) are coloured black. The tree is as shown in **Figure 1a**, and the horizontal black lines demarcate the clusters at the lower level of the tree, and labelled with points. **(b)** The distribution of the mean coancestry between individuals in the same cluster in 200 bootstrap resamples. Clusters are ordered by their median value, and coloured/labelled according to those shown in (a) and **Figure 1a**. One cluster (part of the clade labelled 'Galicia_central') was excluded from this analysis as it only contains 9 individuals. **(c)** Each row of this matrix is a cluster inferred in the fineSTRUCTURE analysis as labelled and demarcated in (a). For each recipient cluster (rows) we tested whether the mean coancestry among individuals within the recipient cluster is smaller than their mean coancestry with individuals in each of the other clusters (columns). *p*-values are based on 200 bootstrap resamples with even sample sizes across each recipient cluster (Methods). Dark borders indicate source-recipient pairs with a *p*-value < 0.02. **(d)** Demographic scenarios leading to high coancestry between different clusters. The symbols α and β represent groups of individuals today, and α' and β' represent their ancestral populations. Solid lines represent drift over time, and dashed lines represent admixture from one group into another (see Supplementary Information for discussion).



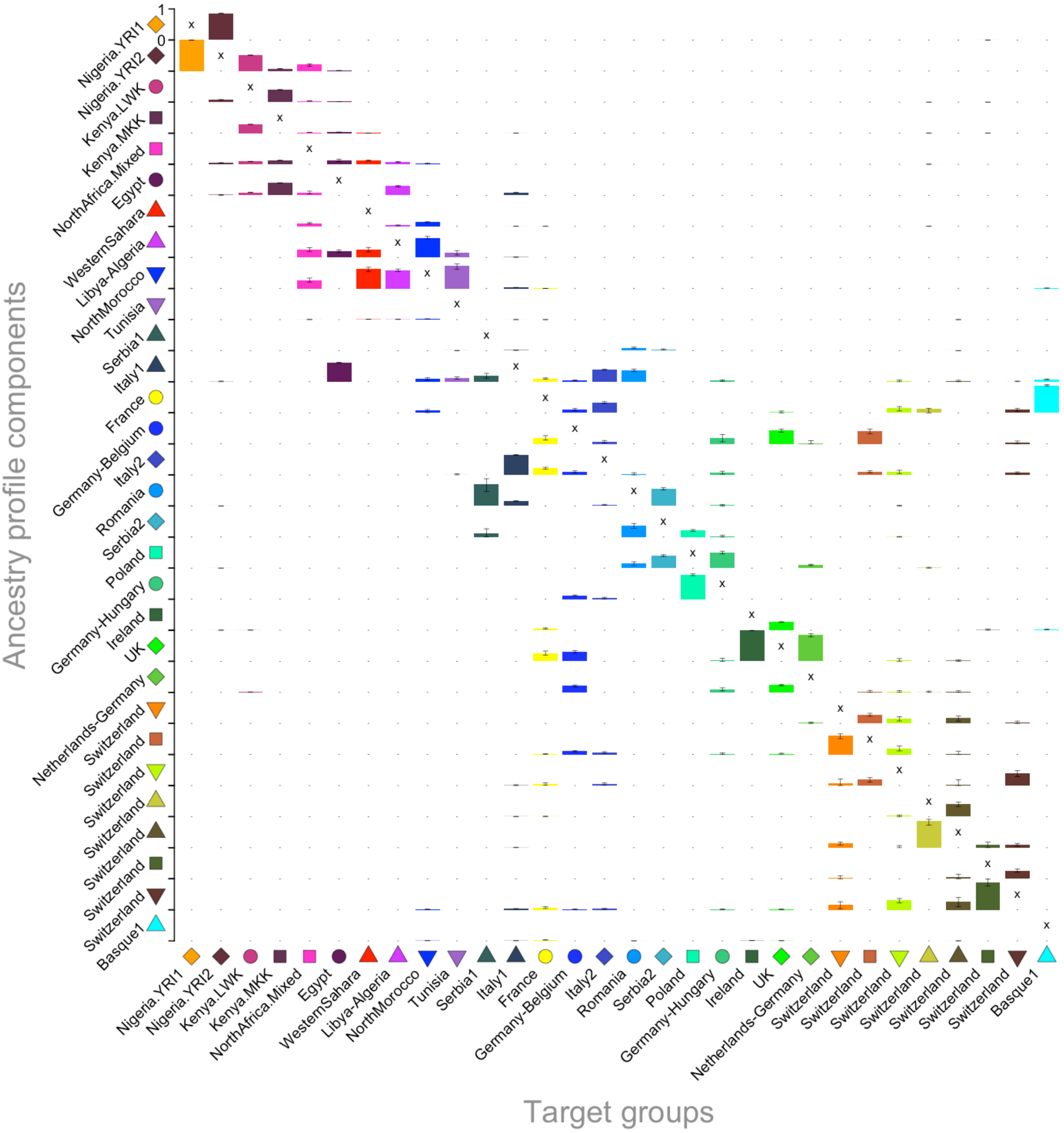
Supplementary Figure 4 | Locations of donor groups and ancestry profiles of Iberian clusters. (a) Locations of individuals within 29 non-Spanish genetic groups inferred using fineSTRUCTURE (Methods). Each point represents an individual, placed at their country-level location of origin, and coloured according to their inferred genetic group. Individuals from the same location (country) have been randomly jittered for visual clarity. Names are assigned to clusters based on where the majority of the individuals in the clusters are located. Where a cluster was split more evenly across two regions, a double-barrel name is used. All groups shown here, except ‘Portugal’, were used as donor groups in the analyses of Iberia. **(b)** Each column shows the ancestry profile of each of the inferred clusters shown in **Figure 3a**. The heights of the bars show the proportion of each cluster’s ancestry which is best represented by that of the labelled non-Iberian donor group (Methods). Note that each row has a different y-axis range for visibility of the smaller components. Error bars show the range of the inner 95% of 1,000 bootstrap re-samples (Methods), and donor groups are only shown if at least one cluster has a range not including zero and a point estimate greater than 0.001. The exact values shown in this plot are tabulated in Supplementary Information Table S1. **(c)** Correlation in Iberians’ ancestry sharing with north African and sub-Saharan African individuals. Each point represents an Iberian individual, with colours and symbols corresponding to the Iberian clusters shown in (b). The *x* and *y*-axes show the mean coancestry with north African and sub-Saharan African individuals, respectively. We defined ‘sub-Saharan African’ as donor groups Kenya.LWK, Kenya.MKK, Nigeria.YRI1 and Nigeria.YRI2; ‘north African’ as the donor groups NorthAfrica.Mixed, WesternSahara, NorthMorocco, Tunisia, Libya-Algeria, and Egypt. **(d)** Residuals for each component of the ancestry profiles shown in (b). Each point represents the residual for a donor group indicated by a colour/symbol. Positive values on the y-axis indicate that the observed coancestry component is larger than the fitted component.



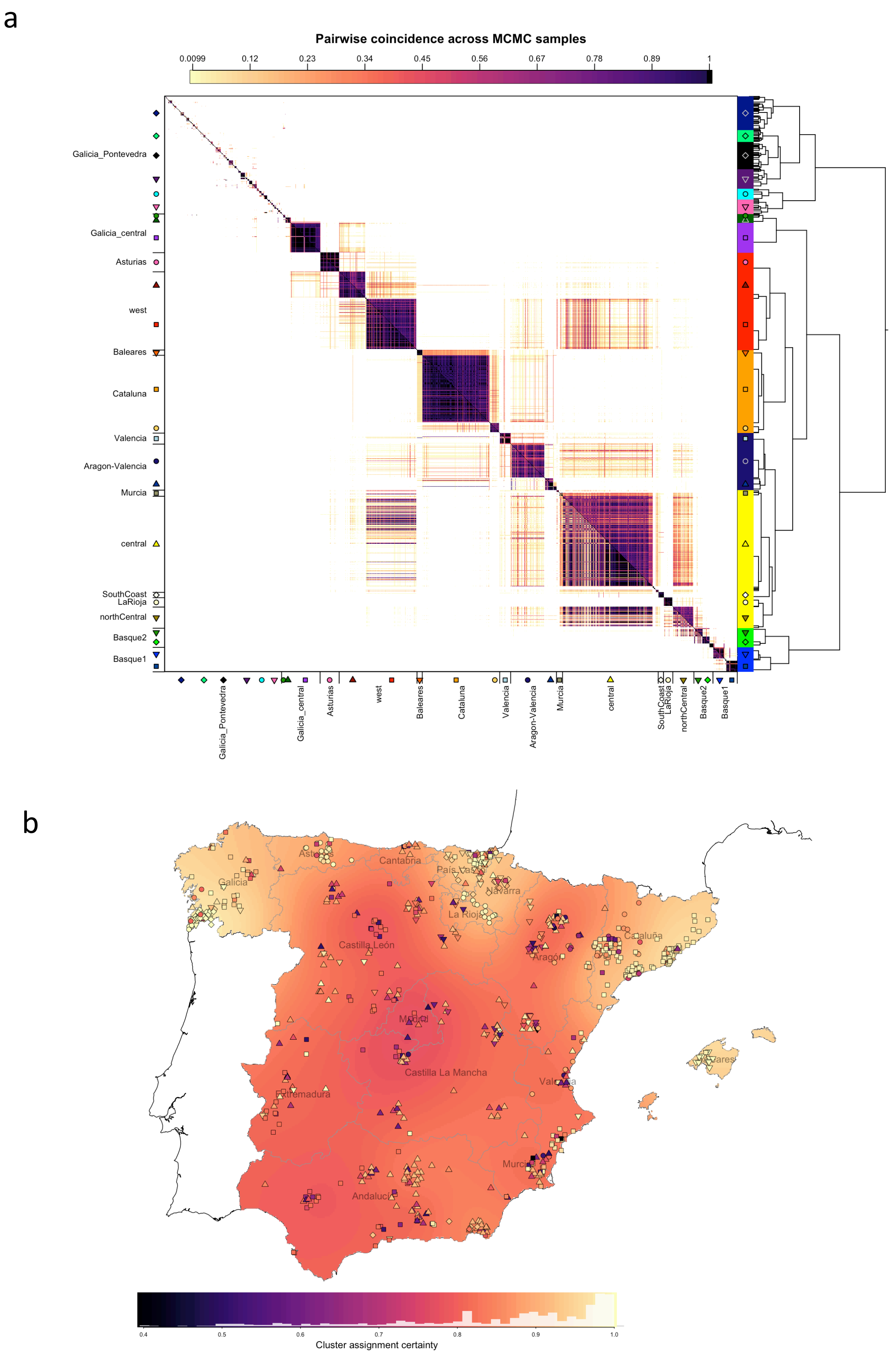
Supplementary Figure 5 | Components of spatially-smoothed ancestry profiles for main genetic contributors to Iberia. A spatially-smoothed ancestry profile has been computed for each point on a spatial grid across Spain (Methods). Each map shows the fraction contributions from the stated donor group (e.g. 'NorthMorocco'). These seven groups (Figures a-g) are exactly the same contributors to the cluster-based ancestry profiles shown in **Supplementary Figure 4b**). Note the colour scale changes because the maximum contribution differs across donor groups. White areas indicate where the donor group contributed zero to the ancestry profile. The histogram on the scale bars shows the distribution of values across grid-points on the map (excluding zero).



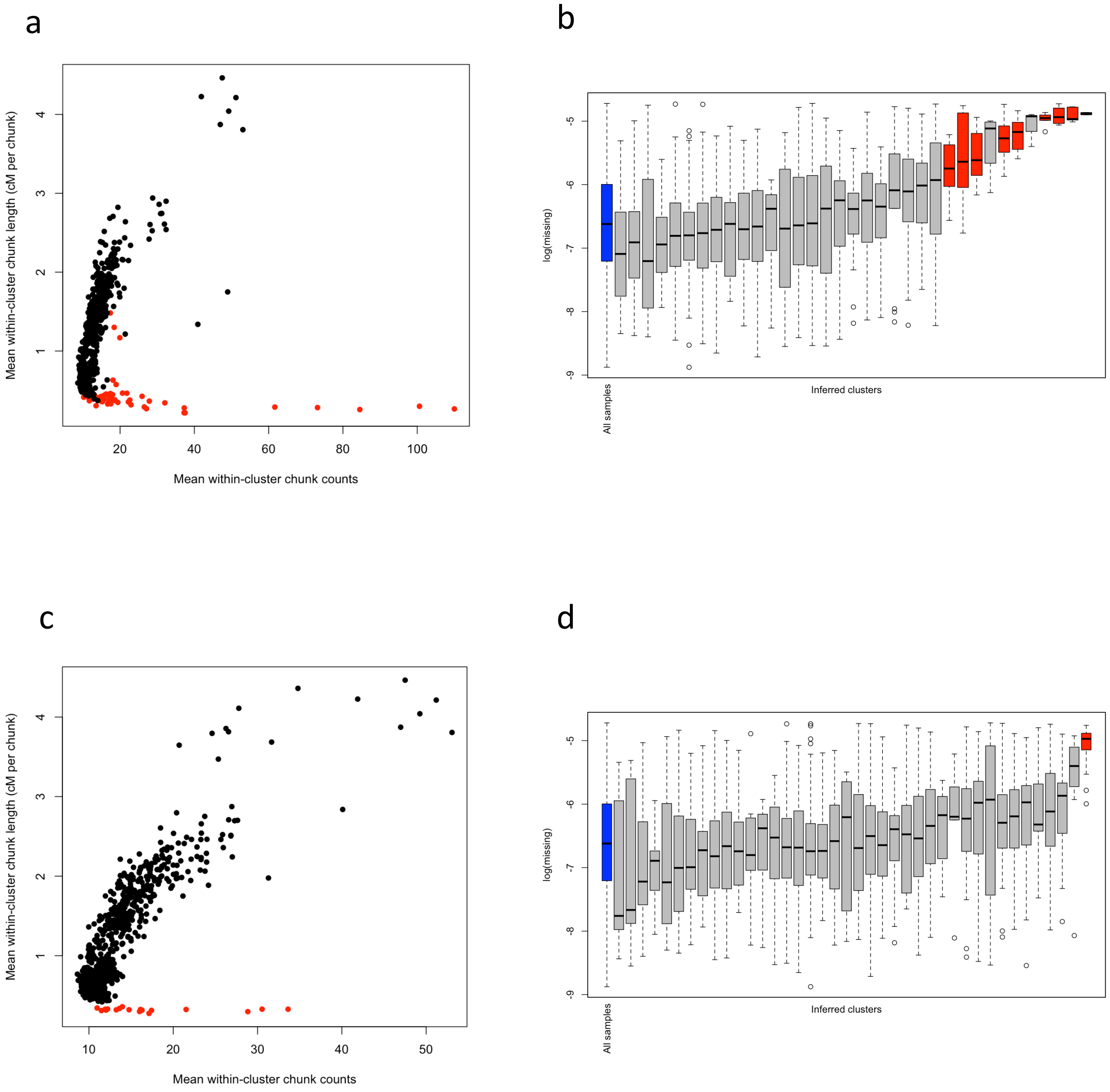
Supplementary Figure 6 | GLOBETROTTER model fit statistics for one-date verses two-date admixture events for each Iberian cluster. All plots refer to results for GLOBETROTTER analysis gtA, allowing all donor groups to be surrogates (Methods; Supplementary Table S2). Barplots for each target Iberian group show the fraction of additional R^2 explained by a two-date admixture model compared to a one-date model (Methods). Negative values can occur when the R^2 for a two-date model is lower than for a one-date model, and the dotted line (0.35) is the value above which there is evidence for a two-date admixture event, as recommended by the authors of GLOBETROTTER. Pairs of surrogate groups are indicated by colors/symbols above the bars; the color of the bars indicates which pairs have a coancestry curve with a negative (black) or positive (grey) slope, which indicate pairs on the same and opposite side of an admixture event, respectively. The inset curve shows the coancestry curves for the target group 'Portugal-Andalucia' for a sub-Saharan African-like surrogate group (YRI), and the fits for one-date and two-dates admixture models.



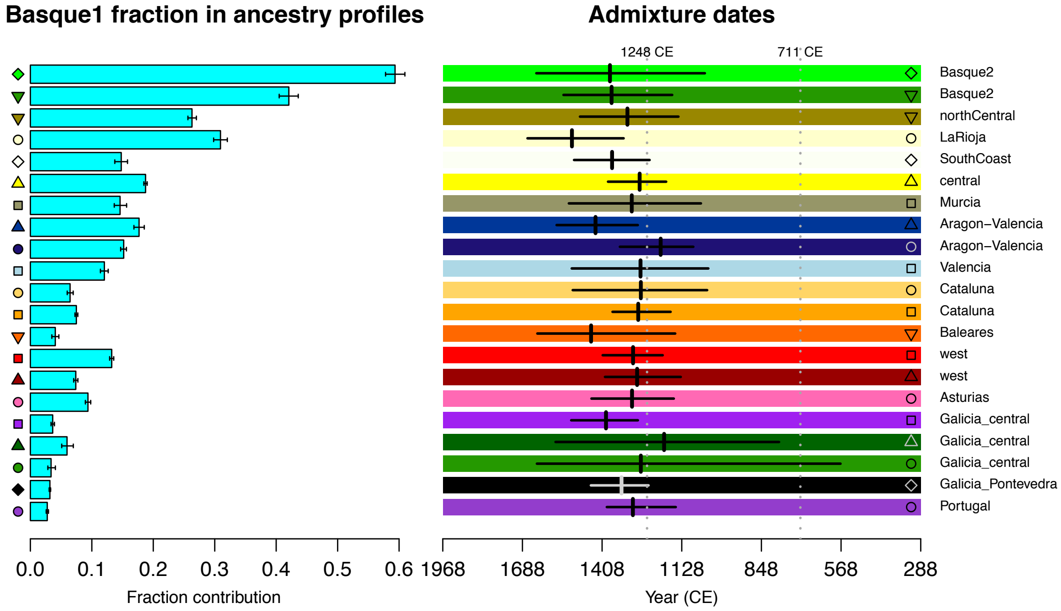
Supplementary Figure 7 | Ancestry profiles of non-Iberian groups and Basque cluster. (a) Each column shows the estimated ancestry profile (see Methods) for each of the non-Iberian donor groups plus a Basque cluster (labelled 'Basque1' in **Figure 1a**). The groups are ordered based on the fineSTRUCTURE analyses we used to define the donor groups (see Methods). The heights of the bars within each column sum to one; donor groups are labelled based on the locations of most individuals in each group (see **Supplementary Figure 4a**). Error bars show the range of the inner 95% of 1000 bootstrap re-samples (Methods), and are only shown if the point estimate was greater than zero. Within-group copying was not allowed under the model, indicated by an 'x' on the diagonal entries.



Supplementary Figure 8 | Convergence of fineSTRUCTURE MCMC and cluster assignment certainty (a) Pairwise coincidence of cluster assignments for two independent fineSTRUCTURE runs. Each element of this matrix shows the fraction of times that a pair of individuals (row and column) are assigned to the same cluster across the set of MCMC samples used to construct the final set of clusters (Methods). White (0) indicates no coassignment across all MCMC samples, and black (1) indicates perfect co-assignment across all MCMC samples. The upper triangle shows results for analysis (A), as discussed in this chapter, and the lower triangle shows results for an independent run of fineSTRUCTURE using exactly the same input data and parameters, but with different a random seed. The sample ordering and tree are from fineSTRUCTURE analysis (A), as shown in **Figure 1a**. The clear similarity between the two runs indicates convergence of the MCMC samples to the posterior distribution. **(b)** Cluster assignment certainty for analysis of fine-scale structure. For the finer level of the tree shown as points in **Figure 1b** (27 clusters), we computed a measure of cluster assignment certainty which measures the co-clustering of individuals over multiple MCMC samples, and can take on values between 0 and 1, where 1 is high certainty (Methods). Points have been coloured according to this certainty measure computed for each individual, and the symbols match those shown **Figure 1b** to distinguish between different clusters. The histogram shows the distribution of the certainty measure for the individuals shown on the map. The background colour has been determined by applying a spatial smoothing algorithm to the same data (Methods).



Supplementary Figure 9 | Effect of using total lengths verses chunk counts as coancestry measure in fineSTRUCTURE algorithm. We ran fineSTRUCTURE for the Spanish cohort using two different coancestry measures, and compared their robustness to genotype quality (Methods). Plots **(a)** and **(b)** show results from a fineSTRUCTURE analysis using chunk counts as the coancestry measure. Plots **(c)** and **(d)** show results from a fineSTRUCTURE analysis using total amount of genome copied as the coancestry measure. For both runs we show these metrics for the level of the hierarchical tree with 35 clusters. The left-hand plots (a) and (c) show, for each individual, the mean chunk lengths (i.e. the average length of copied chunks), and mean number of chunks copied from other individuals inferred to be part of the same cluster. The right-hand plots (b) and (d) show the distribution of genotype missing rates (on the log scale) for the samples in each of the inferred clusters. In all plots, the clusters with significantly higher missing rates from the overall cohort are shown in red ($p < 0.001$, one-sided *t*-test on log-transformed sample missing rates).



Supplementary Figure 10 | Variation and timing of Basque-like genetic contributions in Iberia. Fraction contributions from the Basque-like donor group in ancestry profiles, and Basque-like admixture dates (GLOBETROTTER), for each cluster inferred in the Spain-only analysis (as shown in **Figure 1a**), plus Portugal. The clade labelled 'Galicia_Pontevedra' in **Figure 1a**. was combined into one group for this analysis. The admixture dates are for a two-way admixture event involving a Basque-like side and a European-like side, and shown with 95% bootstrap intervals (Methods). The dates shown assume a 28-year generation time, and a 'now' date of 1940. Detailed results of this GLOBETROTTER analysis are tabulated in Supplementary Table S3b.