

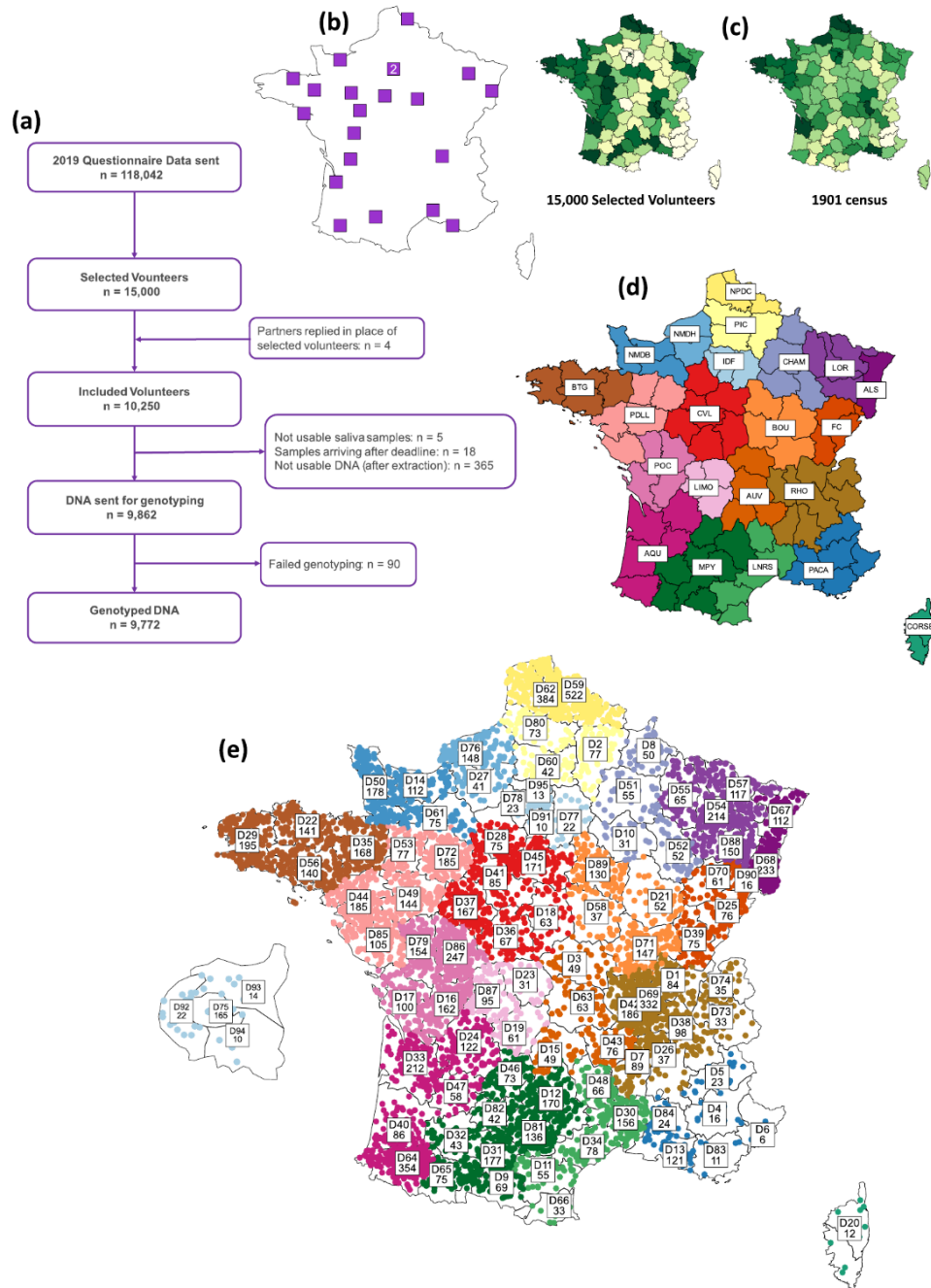


Figure S1: The POPGEN dataset and recruitment.

(a) Flowchart of the selection of 15,000 CONSTANCES participants to receive an invitation and a saliva collection kit for POPGEN. (b) Map of France with the CONSTANCES recruitment centers. (c) Heatmaps of the departments of France showing the targeted counts per department based on the 1901 census (right) and the final counts of the 15,000 selected individuals (left). (d) Colour chart of the 22 'old regions' of France with the labels used for each region throughout this work. (e) Map of France with all of the final genotyped individuals (9,772) individuals' assigned longitude and latitude coordinates. Colors refer to the regions as in (d), each department is labeled with its number with the prefix 'D' as well as the number of individuals in the plot in each department. For example, in department 64 (D64 in the south-west, there are 354 individuals in the dataset). A zoom of the Ile-de-France region (around and including Paris) is given on the left.

36. Pour chacun de vos ascendants (père, mère, grands-parents) indiquez s'ils sont nés en France ou dans un autre pays. S'ils sont nés en France, précisez les noms des départements et communes. (merci d'écrire en lettres capitales)

Sil vous ne souhaitez pas répondre, cochez cette case : ☐ 1

	Né(e) en France	Né(e) hors de France	Je ne sais pas	Si né(e) en France, précisez si possible le département et la commune :
Mère A A A A Année de naissance	☐ 1	☐ 2	☐ 3	Département (nom ou numéro) : Commune :
Grand-mère maternelle A A A A Année de naissance	☐ 1	☐ 2	☐ 3	Département (nom ou numéro) : Commune :
Grand-père maternel A A A A Année de naissance	☐ 1	☐ 2	☐ 3	Département (nom ou numéro) : Commune :
Père A A A A Année de naissance	☐ 1	☐ 2	☐ 3	Département (nom ou numéro) : Commune :
Grand-mère paternelle A A A A Année de naissance	☐ 1	☐ 2	☐ 3	Département (nom ou numéro) : Commune :
Grand-père paternel A A A A Année de naissance	☐ 1	☐ 2	☐ 3	Département (nom ou numéro) : Commune :

Figure S2: Question added to the CONSTANCES 2019 annual questionnaire to obtain the information on birthplaces (Question #36).

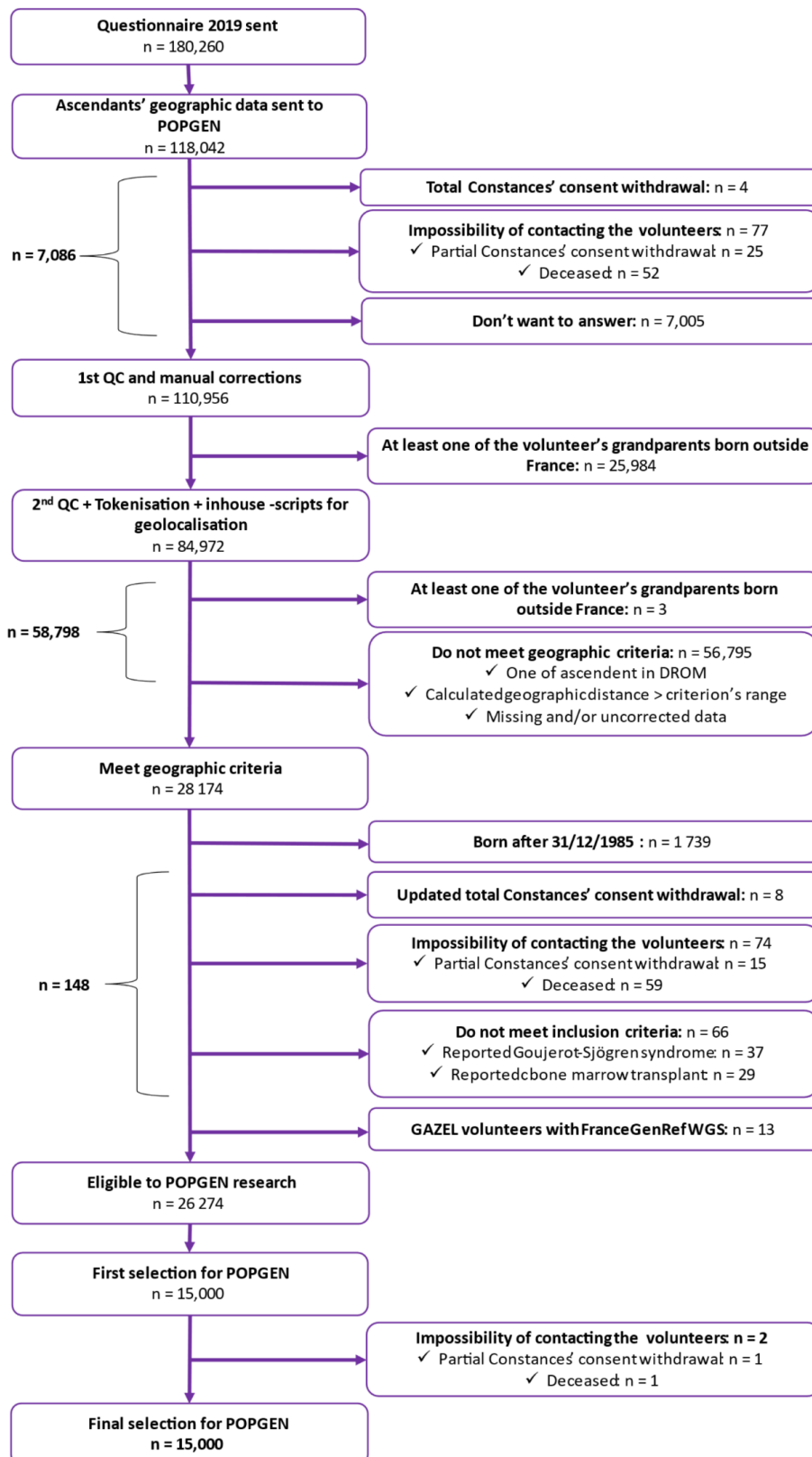


Figure S3: Flowchart of the selection of 15,000 volunteers of CONSTANCES for the POPGEN study.

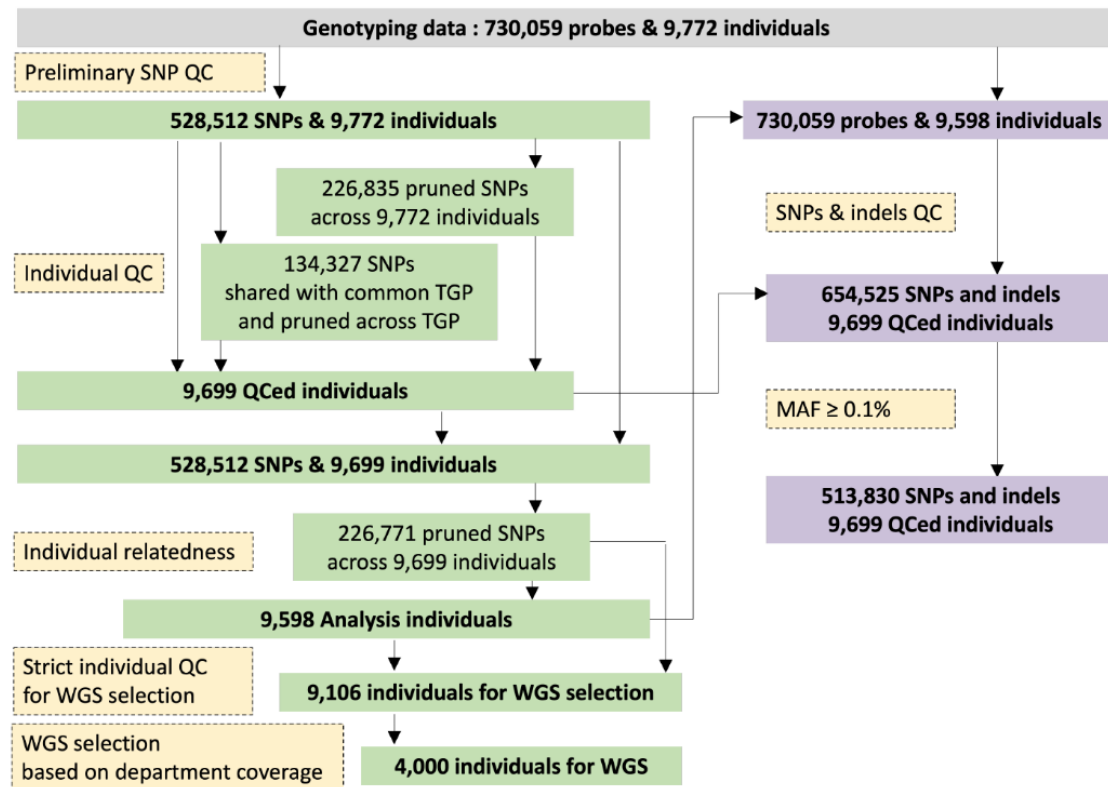


Figure S4: Overview of the quality control (QC) process. First a preliminary SNP QC was performed on which the individual QC was based. Individual QC led to 3 nested sets: a main QCed individual set, an unrelated QCed individual set, and a stringent set of individuals for WGS selection. The final dataset used for population genetic analyses was obtained through a variant QC performed on the unrelated QCed individual set, but relatives were used in some analyses. TGP: 1000 Genome Project.

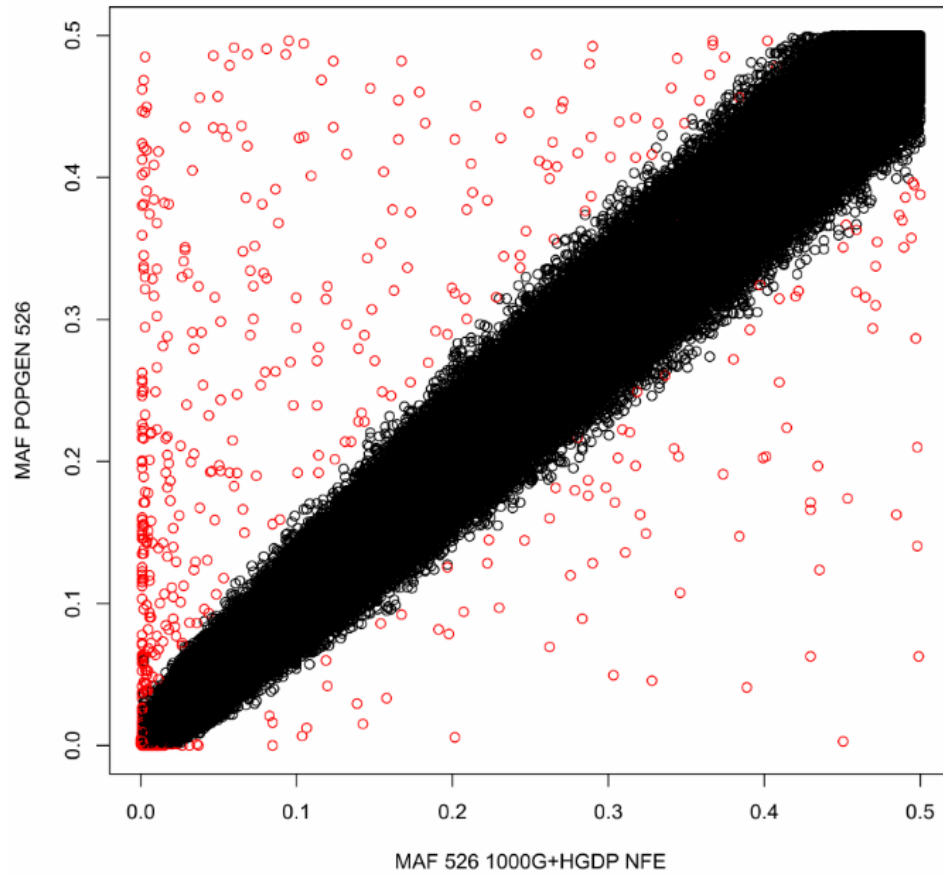


Figure S5: Comparison of Minor Allele Frequencies between POPGEN and NFE_1KG_HGDP [5] during the merge of the two datasets. A simple association study of POPGEN against NFE_TGP_HGDP was performed and 430 variants (highlighted in red) with a p-value below 1×10^{-5} were excluded.

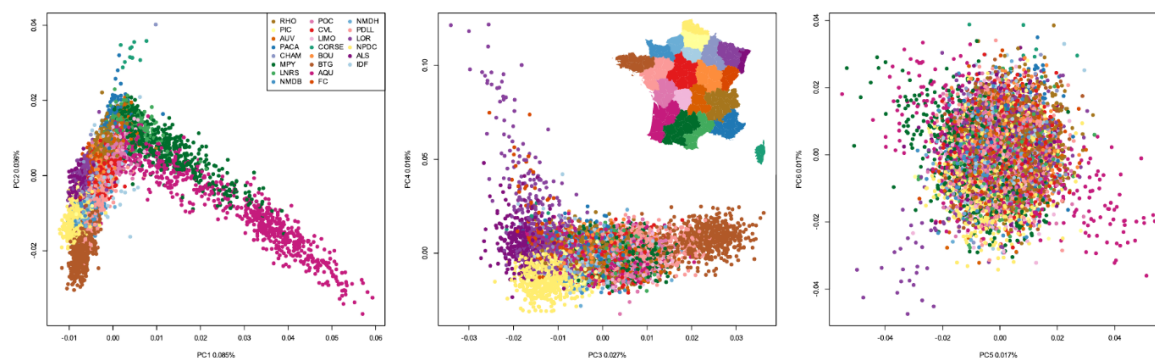


Figure S6: PCA of POPGEN9598. Components 1-6 are plotted with colours corresponding to the ‘old administrative regions’ of France (see Table S1). The plot on the left corresponds to Figure 1d in the main text.

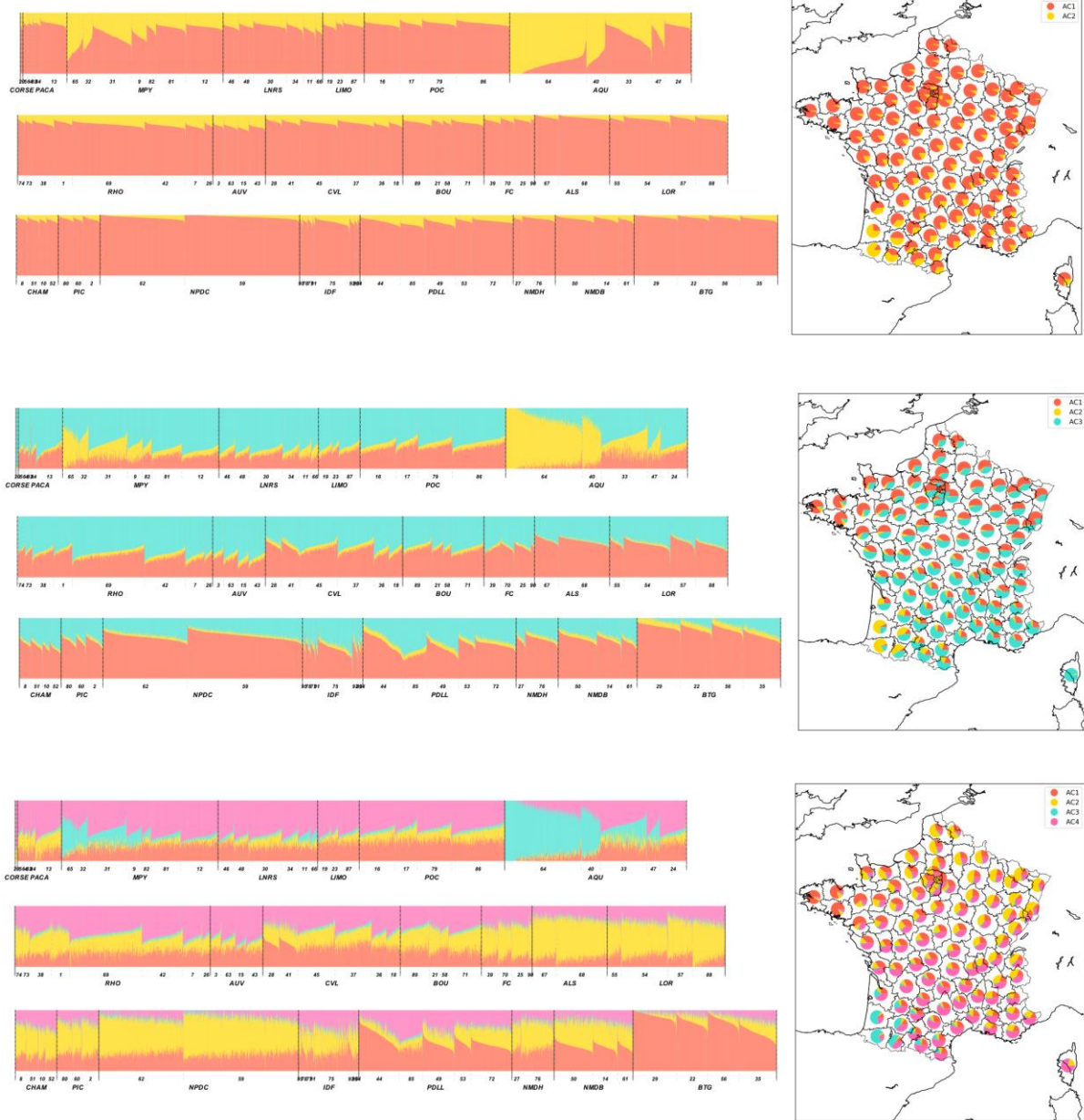


Figure S7: Unsupervised admixture plots on POPGEN9598 for different numbers of groups ($K=2$ to $K=8$ below). In the three left panels (for each value of K), the individual admixture components of all 9,598 individuals are plotted, arranged first into 22 ‘old regions’ of France and then into 95 departments. Within each department, the individuals are ordered based on the admixture component that is (on average) contributing the most for each department. On the right (for each value of K), for each department we calculate the average admixture components (ACs) across individuals and present these as pie-charts on the map of France. with the same colours as in the left. Cross-validation statistics: $K=2$ (0.32361), $K=3$ (0.32356), $K=4$ (0.32357), $K=5$ (0.32364), $K=6$ (0.32372), $K=7$ (0.32380), $K=8$ (0.32386).

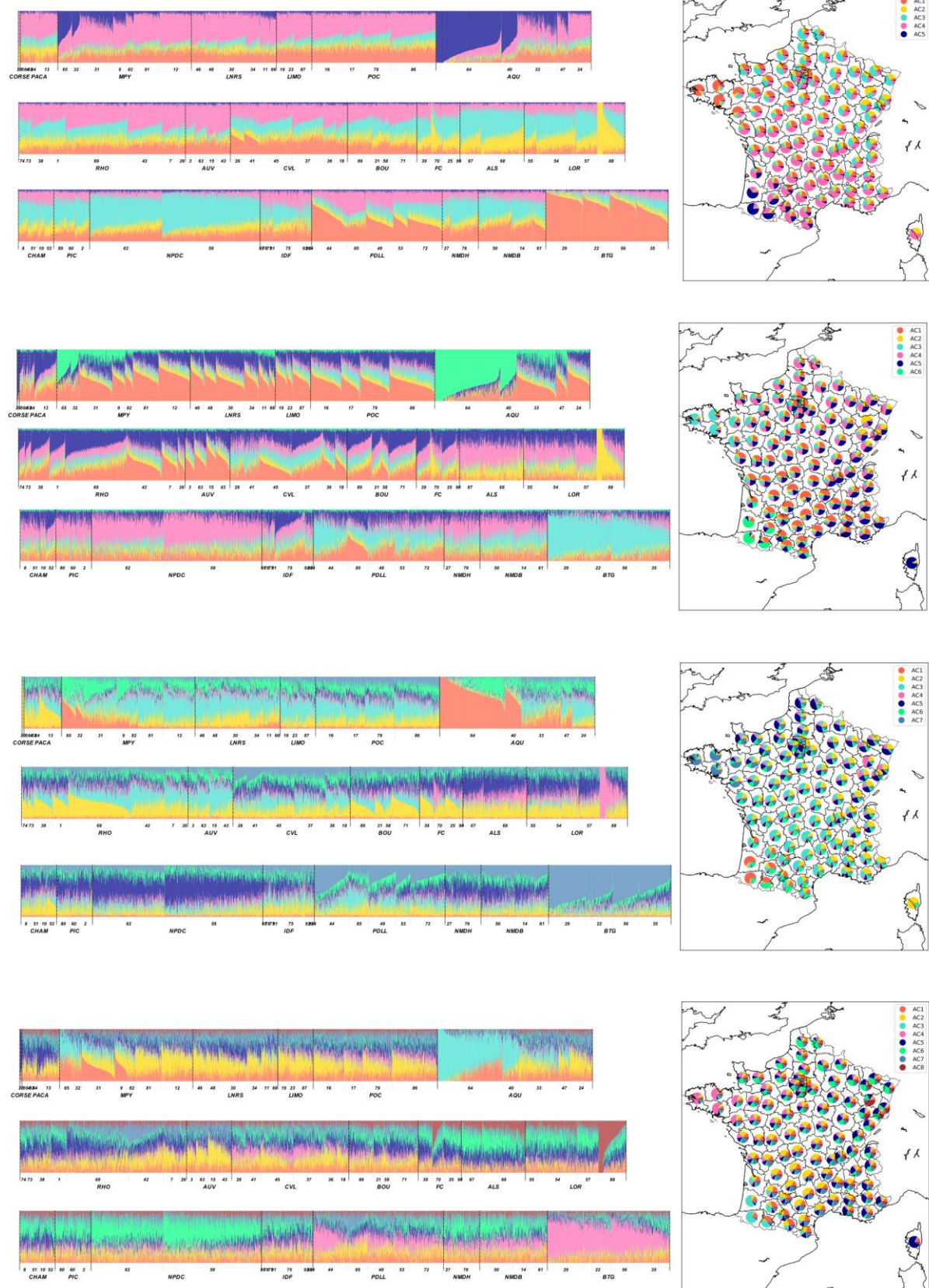


Figure S7 continued

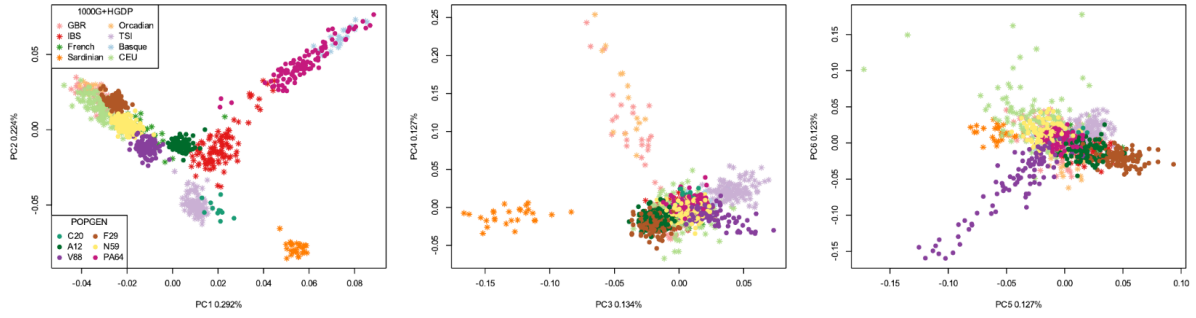


Figure S8: PCA analysis on the combined data of 526 individuals from 6 specific departments of POPGEN with 526 individuals from NFE_1KG_HGDP.

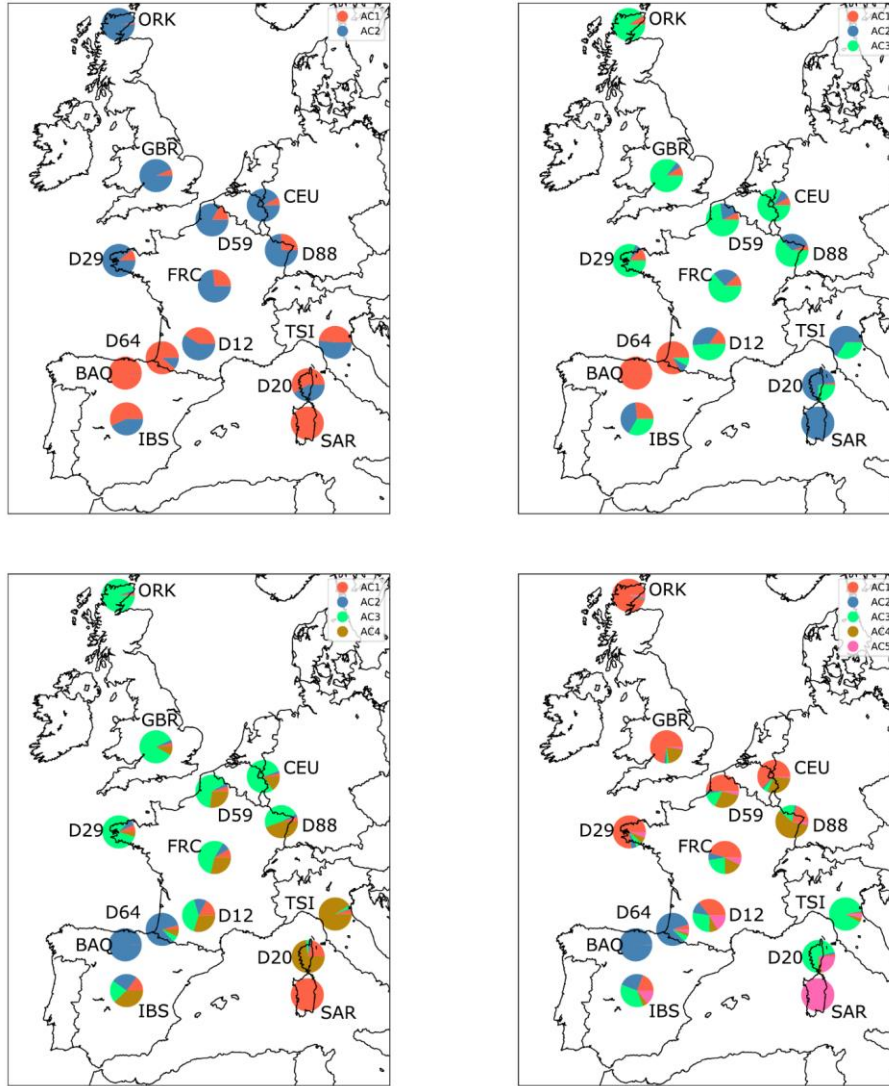


Figure S9: Unsupervised Admixture plots on the combined data of 526 individuals from 6 specific departments of POPGEN and 526 individuals from NFE_1KG_HGDP for different number of groups (K=2 to K=11 in subsequent plots).

Mean Admixture Components (ACs) for each group are calculated and presented as pie-charts with approximate geographical positions for the different European groups. Cross validation error statistics: K=2 (0.32438), K=3 (0.32467), K=4 (0.32625), K=5 (0.32772), K=6 (0.32927), K=7 (0.33034), K=8 (0.33154), K=9 (0.33324), K=10 (0.33474), K=11 (0.33692).

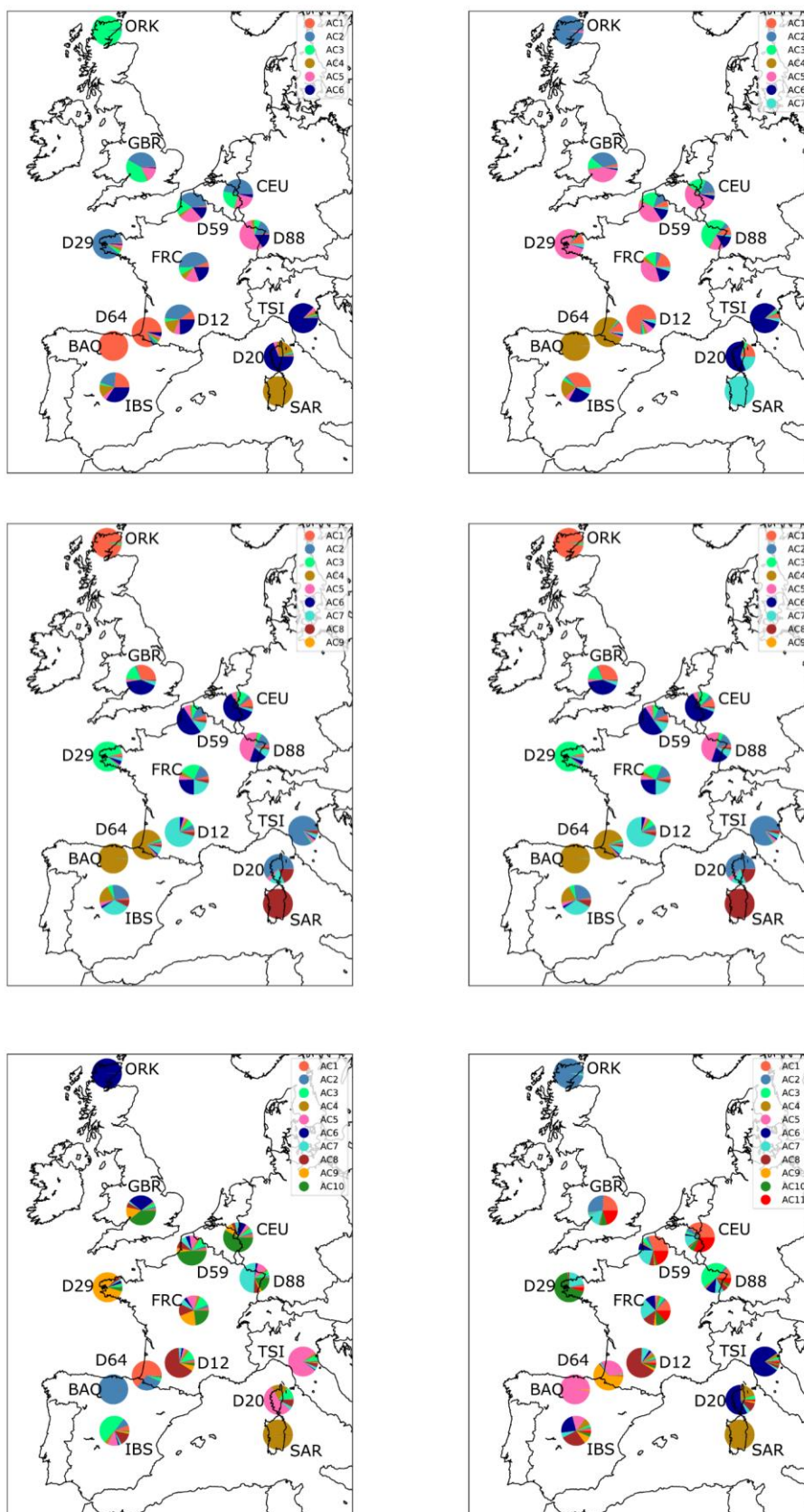


Figure S9 continued

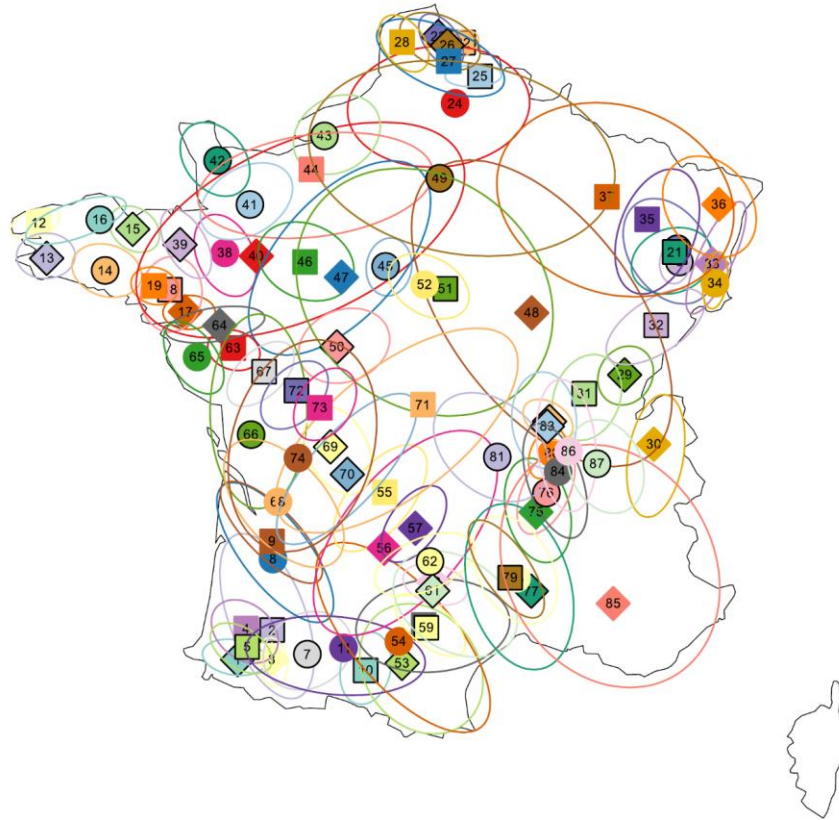


Figure S10: For each of the 87 FineSTRUCTURE clusters presented in Figure 4 of the main text, the centre-of-mass (in longitude and latitude) is plotted onto the map of France along with minimal ellipses containing 90% of the individuals in each cluster.

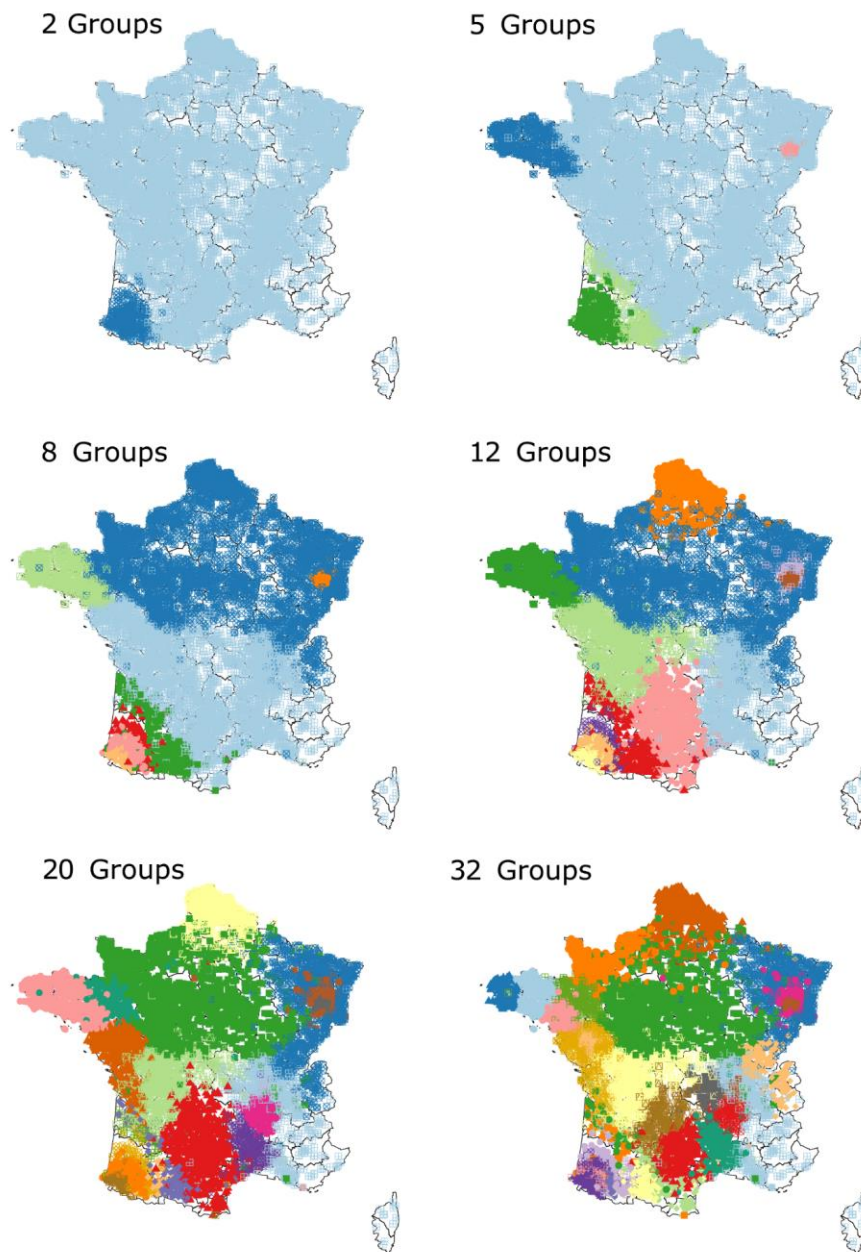


Figure S11: Cuts of the final FineSTRUCTURE TVD tree (see Figure 4d in the main text) at 2, 5, 8, 12, 20, and 32 that correspond to the analysis presented in Figure 4 in the main text.

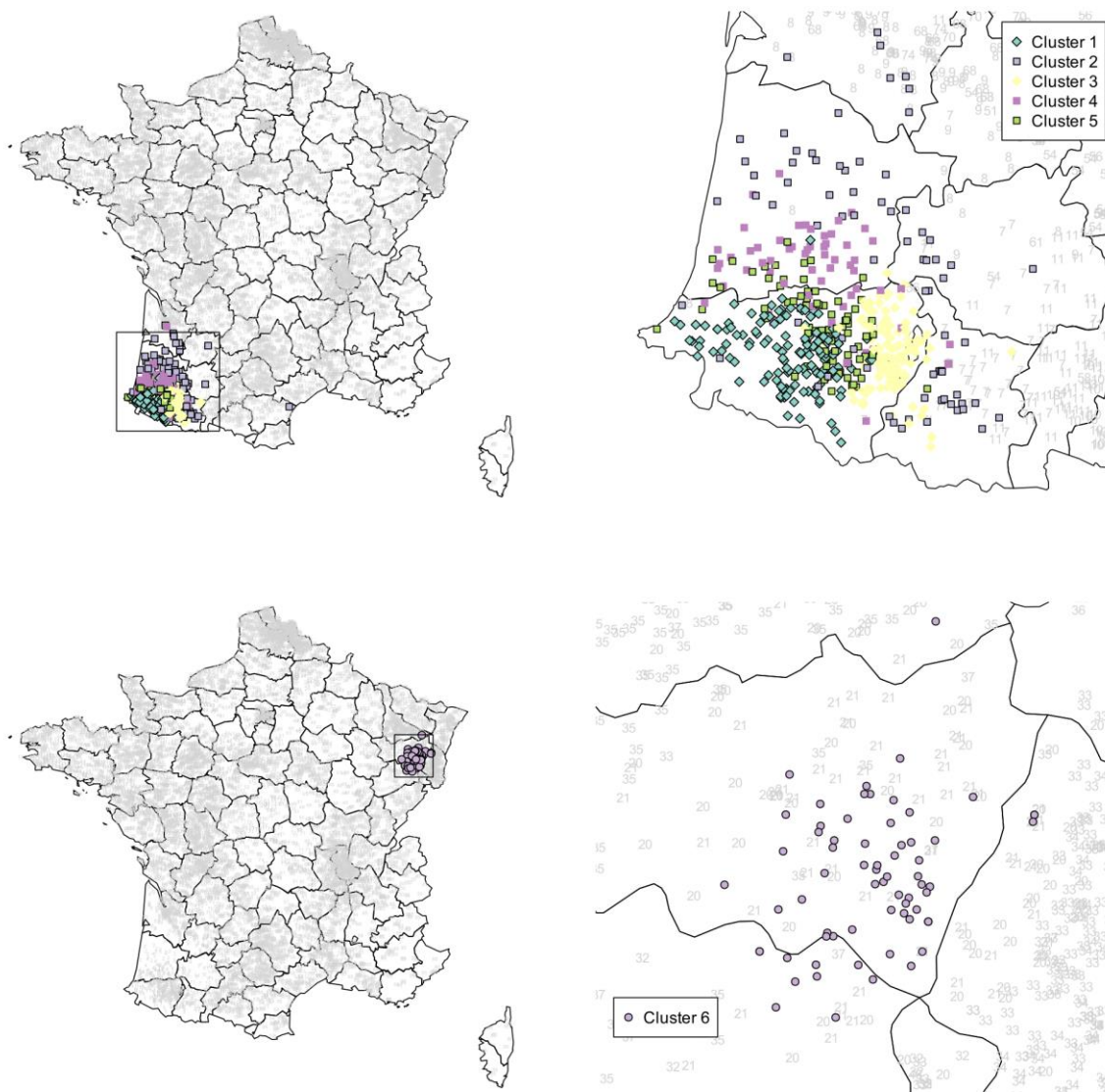


Figure S12: The final FineSTRUCTURE TVD tree (see Figure 4d in the main text) is divided into 16 different sections, and more detailed plots of the results are given in succession.

In each plot the map of France with all individuals of POPGEN is presented, individuals within the section of the FineSTRUCTURE TVD tree being highlighted are plotted with their cluster specific points (colour, shape, and outline), all other individuals are plotted in grey with their cluster number. A zoom of the region in question is also presented to aid the visualisation. In the first pairs of sub-figures, individuals from clusters 1 to 5 are plotted (above), who are all located in the South-West of France. These clusters form the first large branch that splits off the FS TVD tree (on the rightmost of Figure 4d in the main text). In subsequent plots (below), we go round the FS TVD tree clockwise, in the 2nd pair of sub-figures only cluster 6 is plotted, in the third pair clusters 7 to 11 are plotted, and so on. The 87 clusters were split into 16 groups for these plots simply to aid the overall visualisation of the FineSTRUCTURE results, given the complexity of the full results in Figure 4c in the main text.

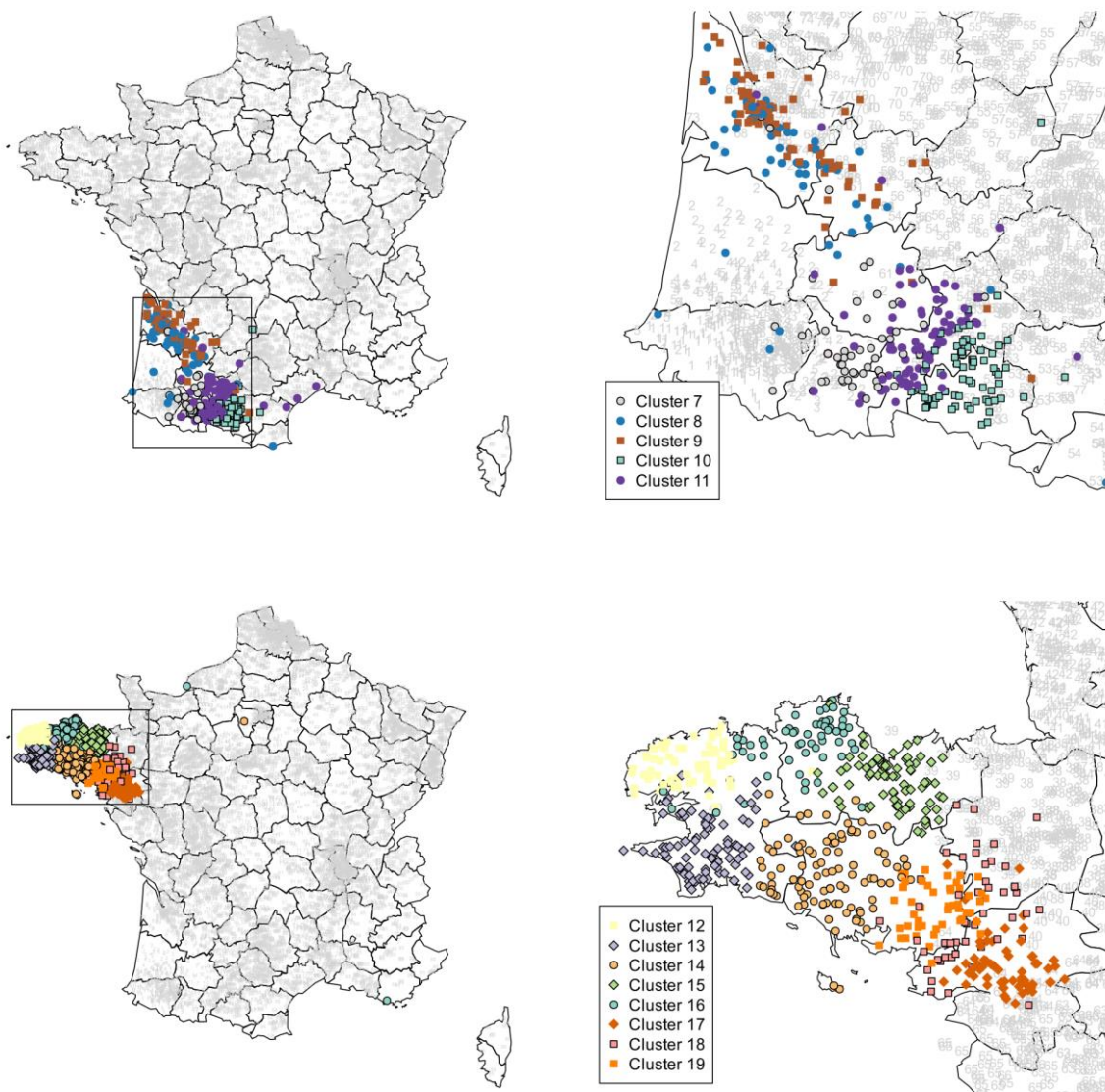


Figure S12 continued

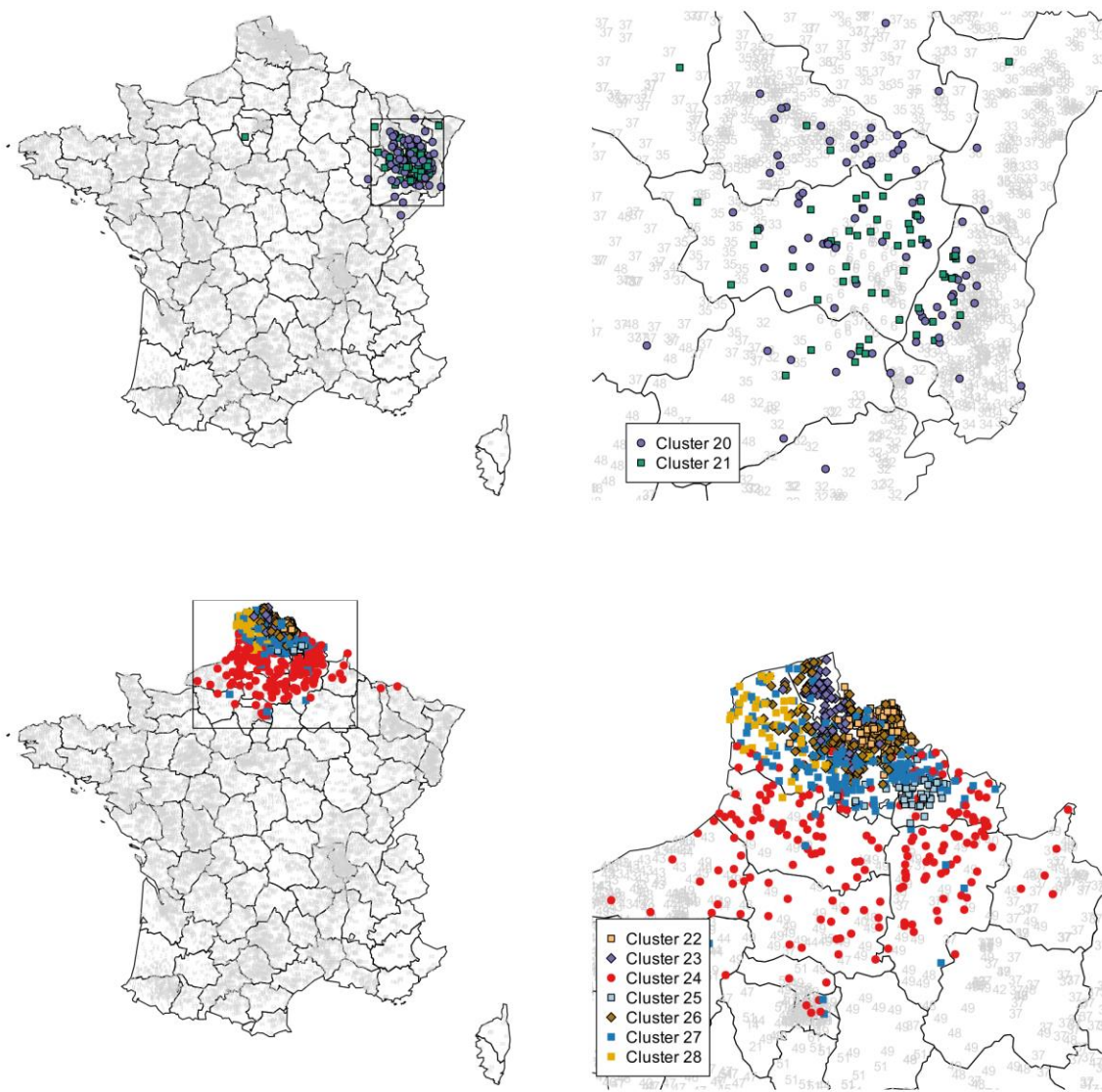


Figure S12 continued

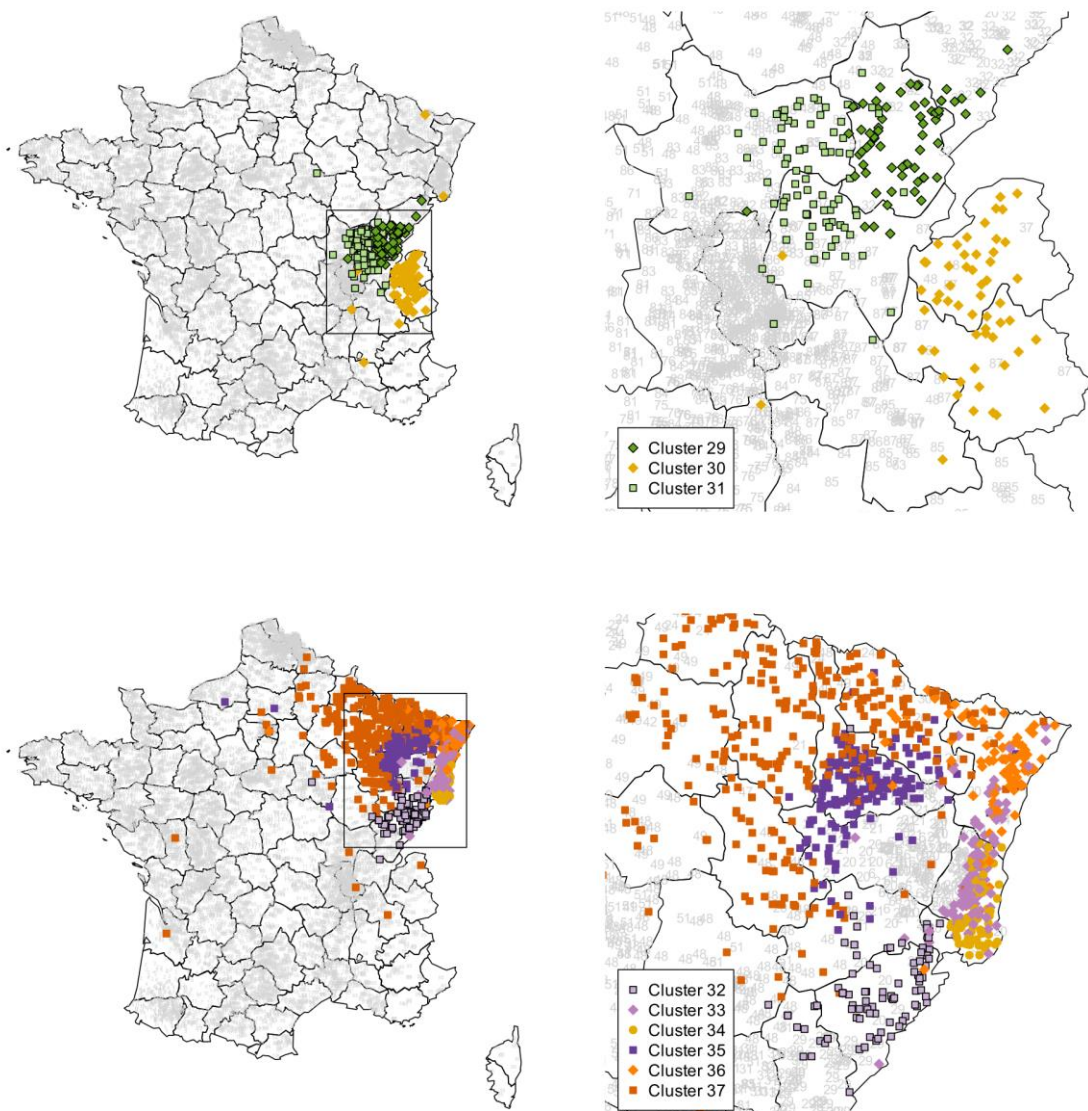


Figure S12 continued

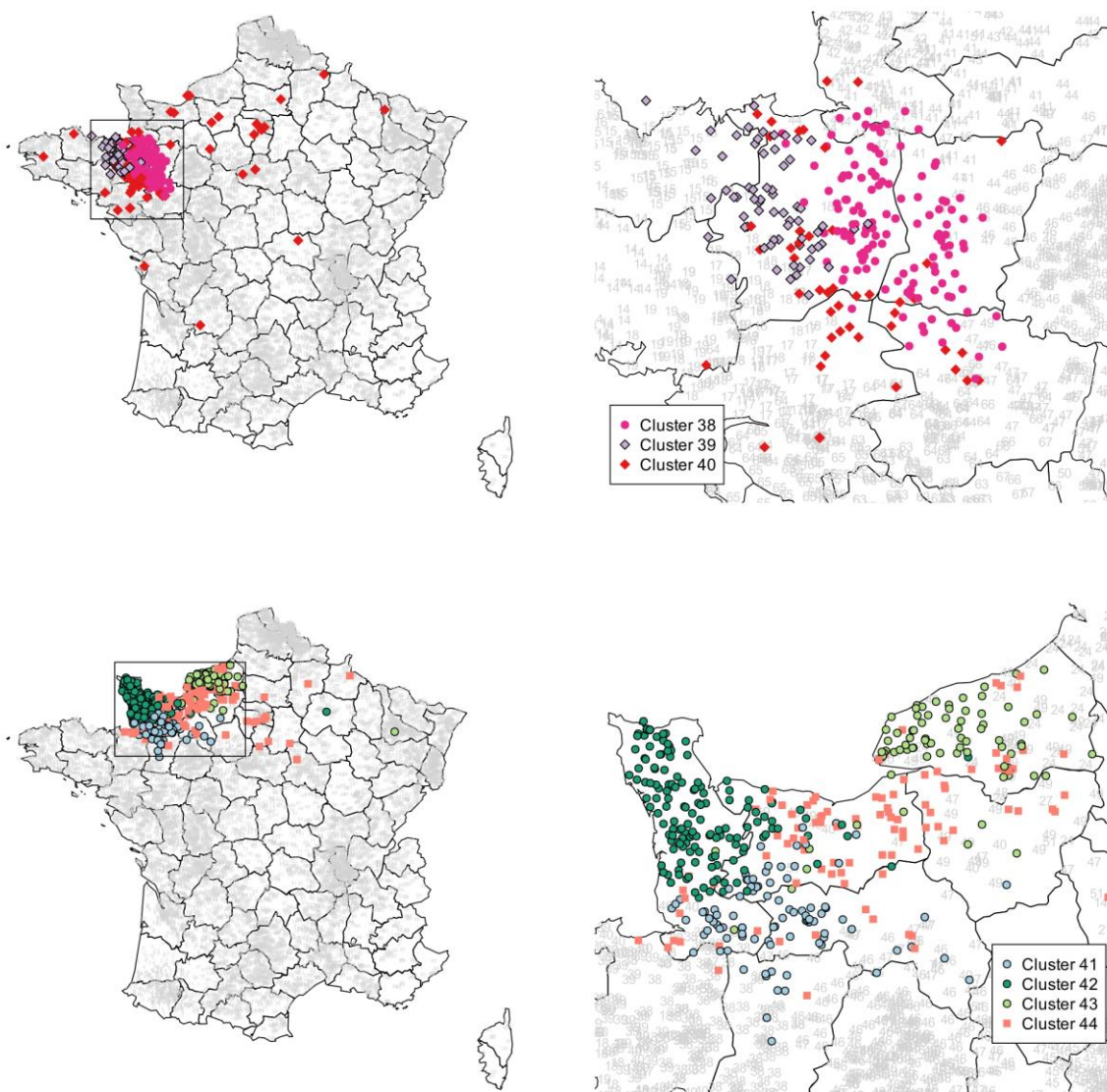


Figure S12 continued

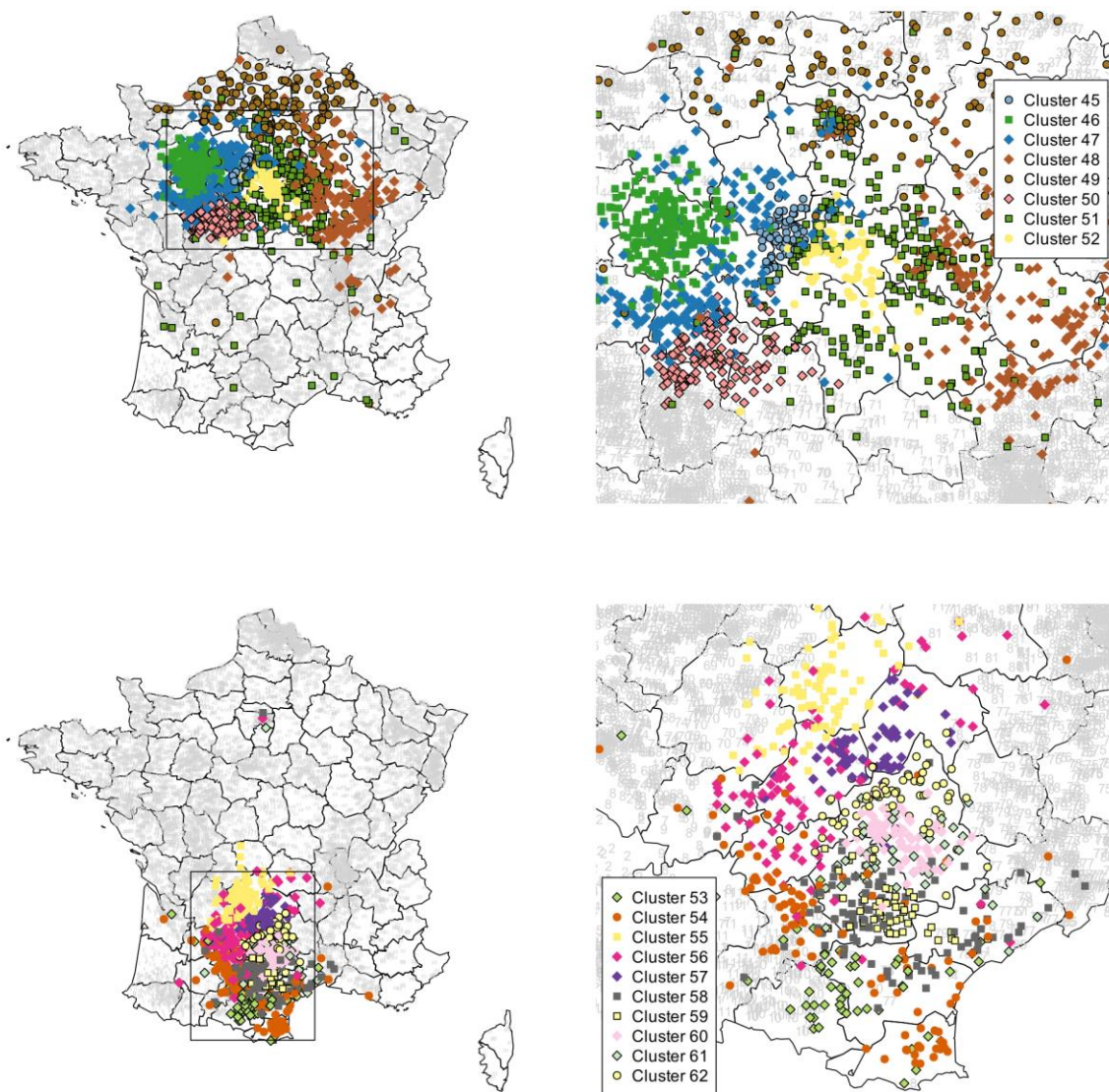


Figure S12 continued

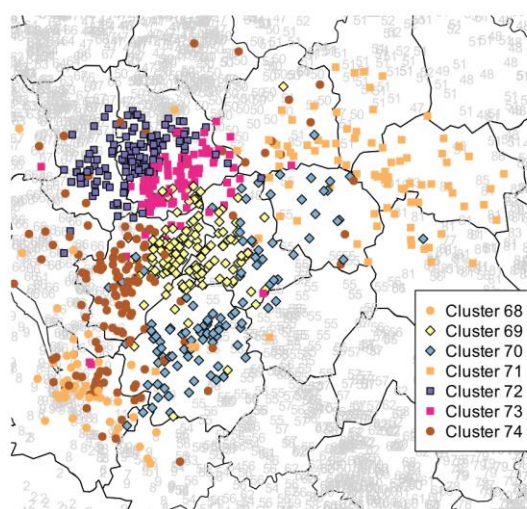
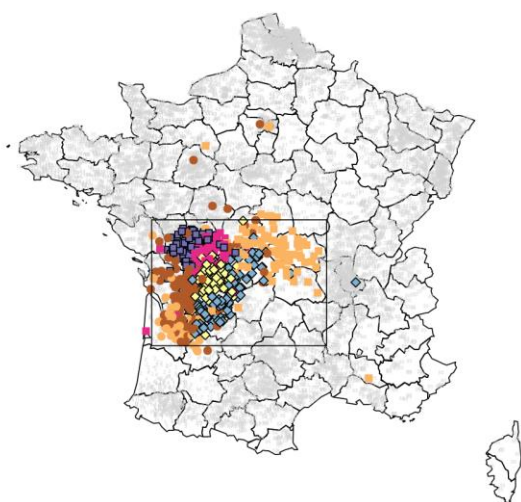
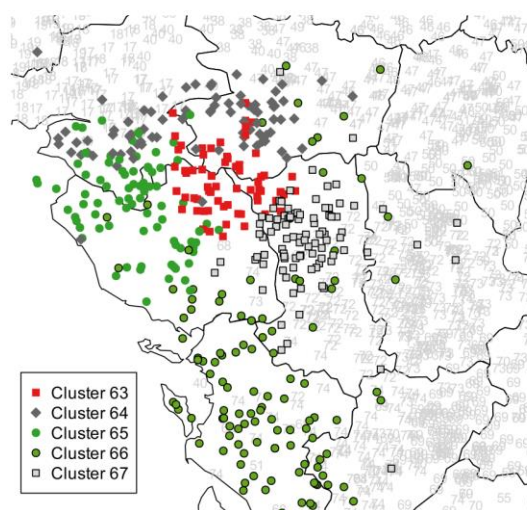
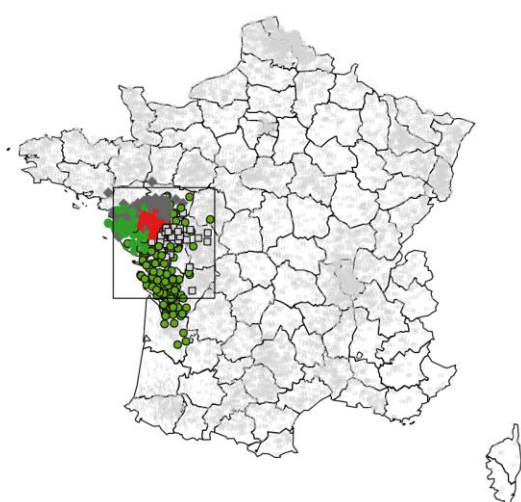


Figure S12 continued

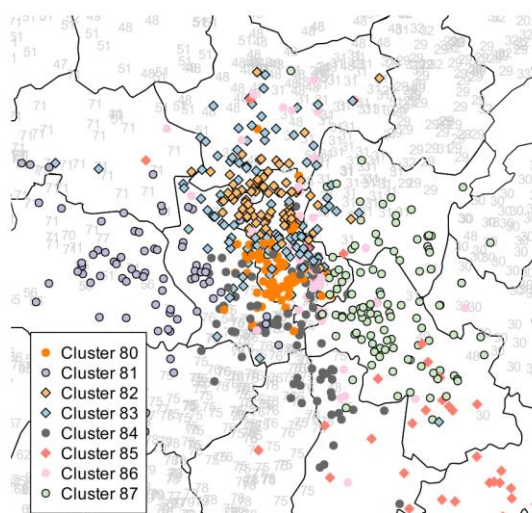
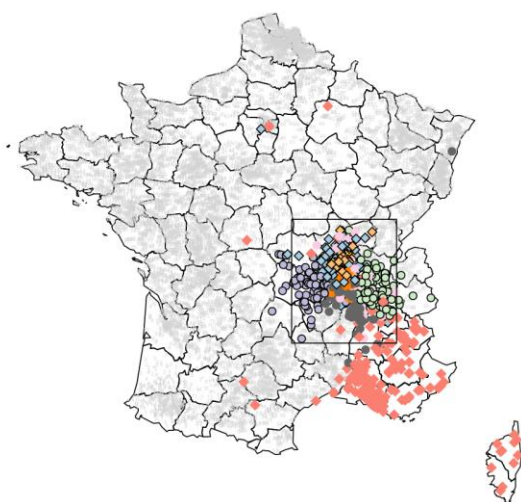
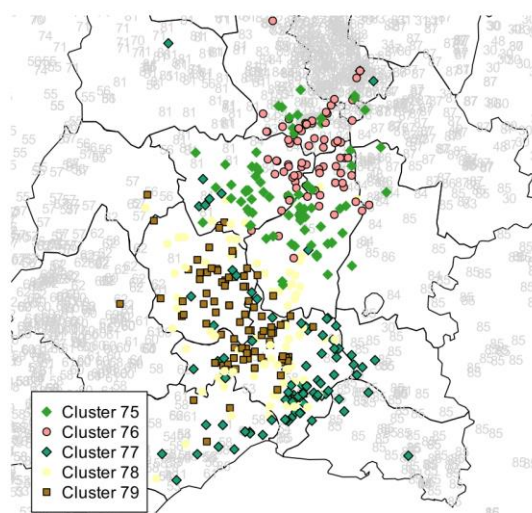
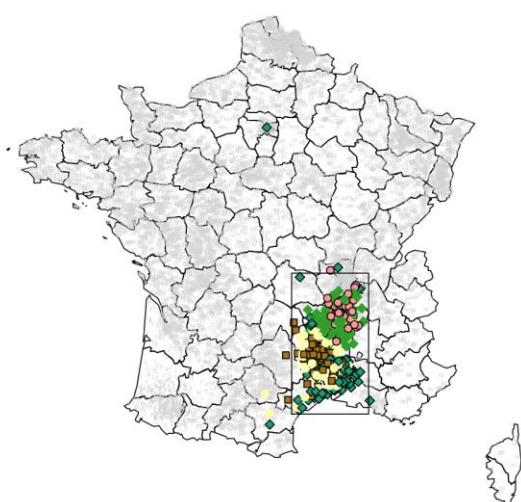


Figure S12 continued

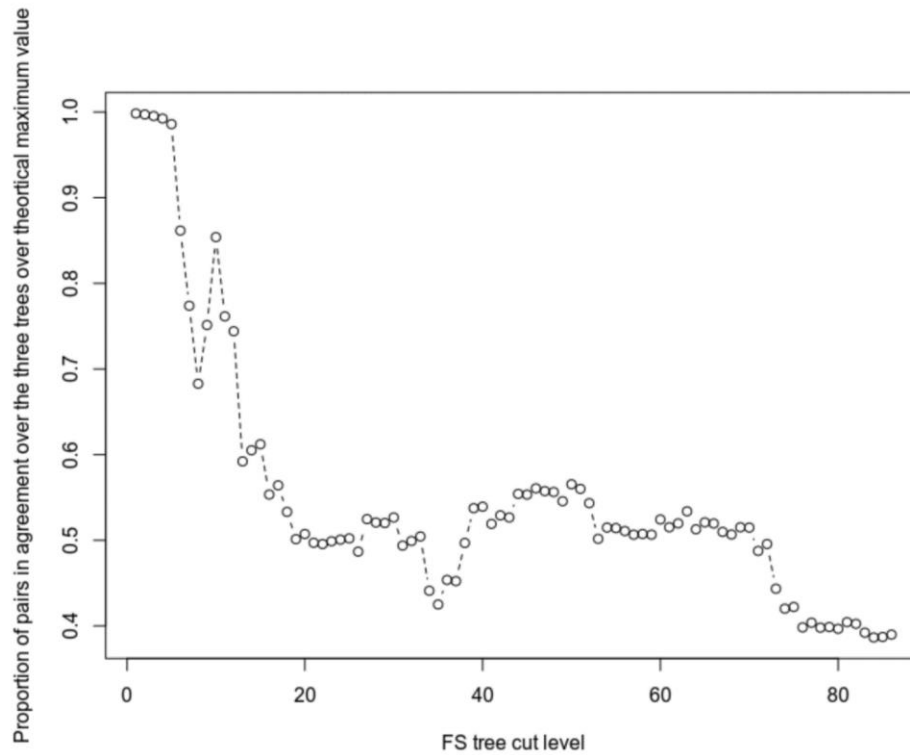


Figure S13: Consistency of FineSTRUCTURE results across three runs. At different cuts of the three trees, we calculated the number of pairs of individuals who are in the same cluster as each other across the three FineSTRUCTURE runs. We then divided this by its maximum possible value which is the minimum of the numbers of pairs of individuals found in the same cluster in each of the three trees separately. This ratio will equal 1 when all three trees are exactly in agreement and will decrease as the three trees diverge. As cited in the main text, at the end of the trees, the first run has an agreement on 53.1% of its pairs who are clustered together with run 2 and 56.0% with run 3.

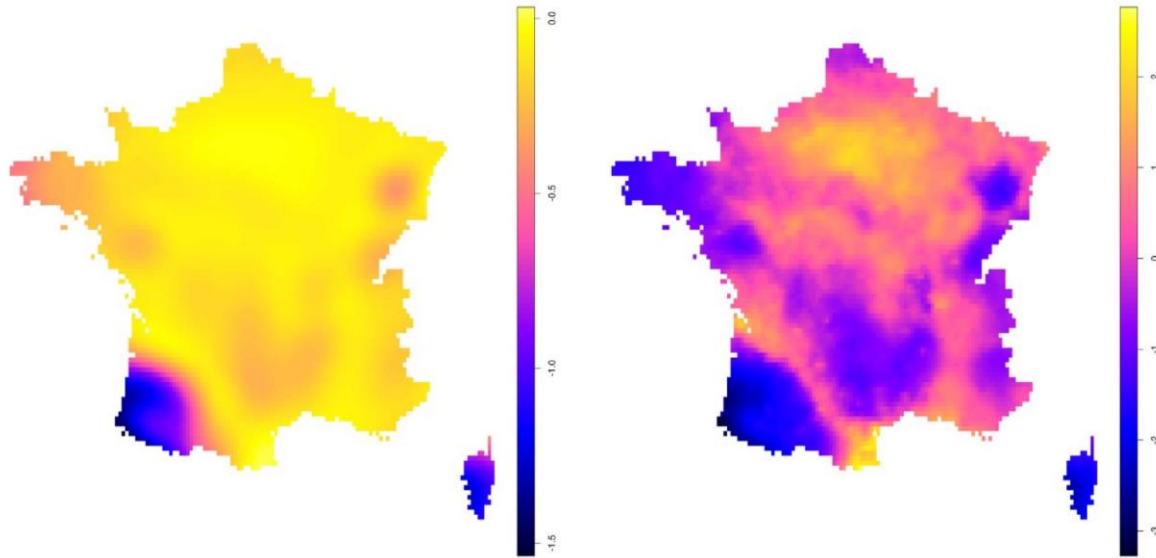


Figure S14: Universal Kriging analysis of the slope coefficient from 9598 regression models that compare physical distance with IBD sharing (both on the log scale) as described in the main text in the analysis pertaining to Supplementary Table S5. On the left, the raw slope coefficients are presented, with the South West and Corsica standing out. On the right, the slope coefficients are re-scaled using a rank-inverse normalisation to better visualise the geographical patterns relating to local strength of isolation-by-distance.



Figure S15: Example of two clusters identified by FineSTRUCTURE with 3 outliers (2 on the left panel, 1 on the right panel identified). The logic behind the exclusion was to avoid individuals whose anchorage in longitude and latitude on the map of France was less well correlated with their genetic data. The definition of the outliers is given in Supplementary Note 3.

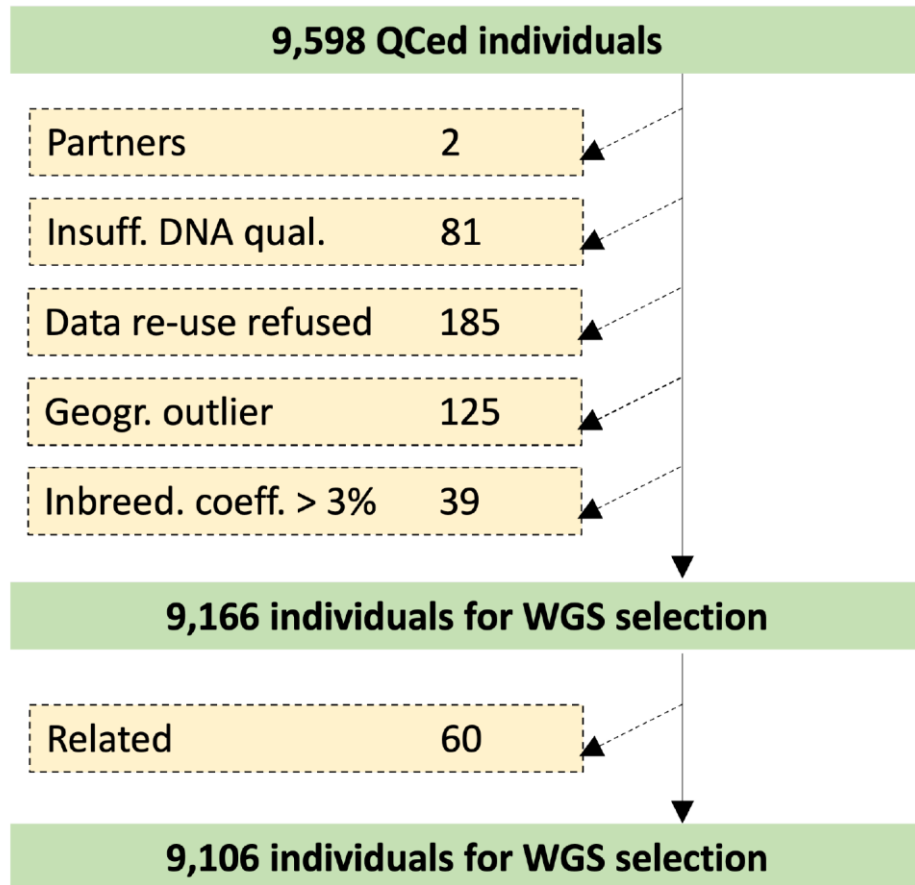


Figure S16: Flow chart detailing the exclusions that lead to a list of 9,106 individuals from which the 4,000 WGS individuals were selected.

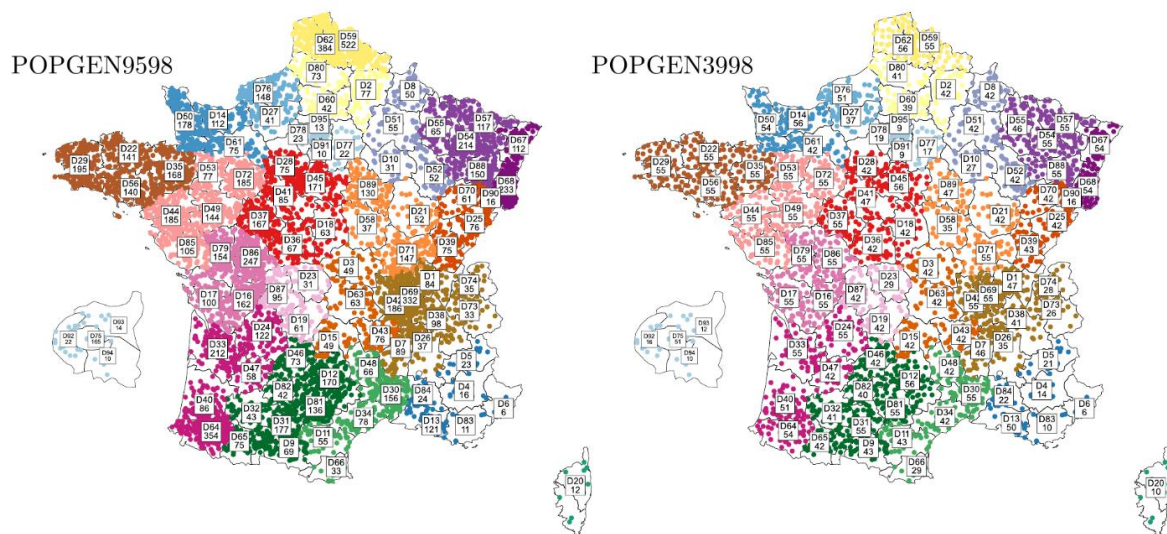


Figure S17: Assigned positions of all individuals in POPGEN9598 (left) and the subset of individuals in POPGEN3998 selected for WGS (right) with colours corresponding to administrative regions in France and department numbers (top) and total number of individuals (bottom) indicated for each department.

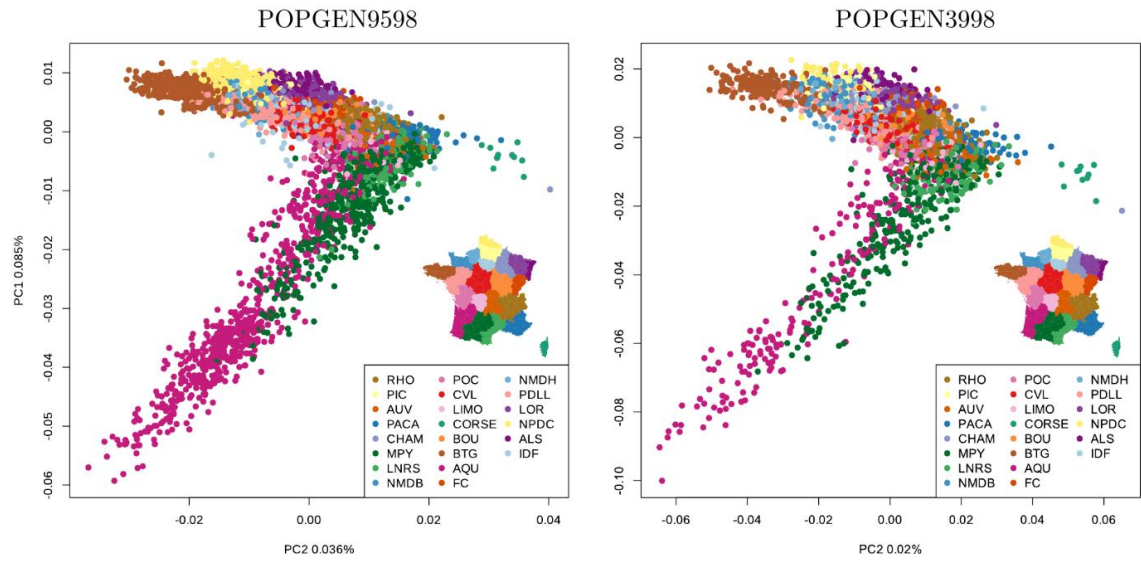


Figure S18: Principal components analysis of POPGEN9598 (left, a direct copy of Figure 1d in main text) compared to the same analysis recalculated on only the POPGEN3998 individuals using the same set of genetic variants (right). Colours correspond to administrative regions in France. The squared-correlation between projections on to the first principal components is 0.995, and is -0.970 for the 2nd principal component (we multiplied the 2nd PC projection by -1 for the purposes of the plot).

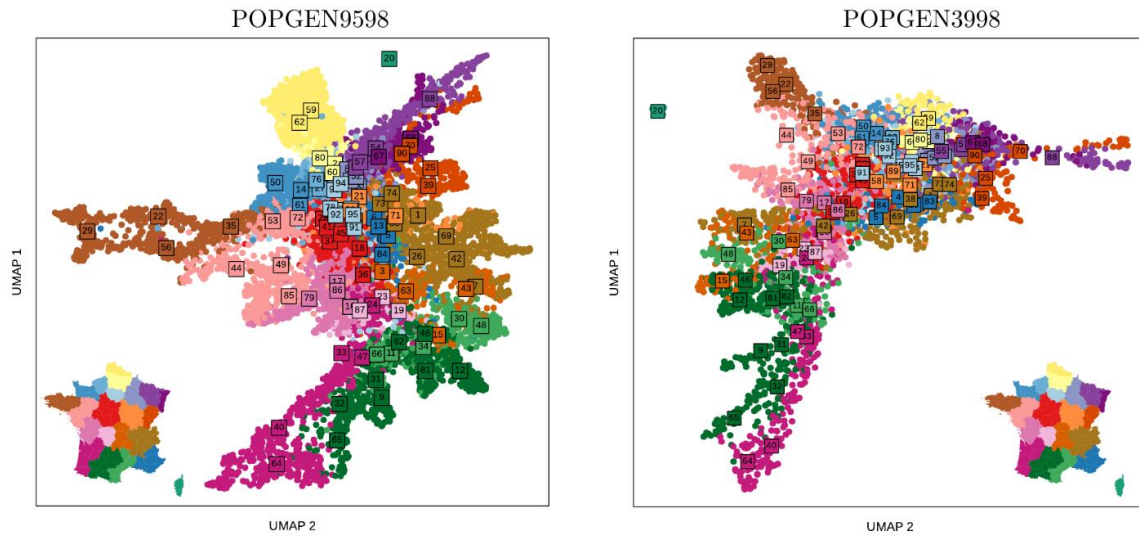


Figure S19: Visual comparison of the UMAP projection of the PBWT-paint matrix of the POPGEN9598 sample (left) and the POPGEN3998 sample (right). Both show separation of individuals at the department level.

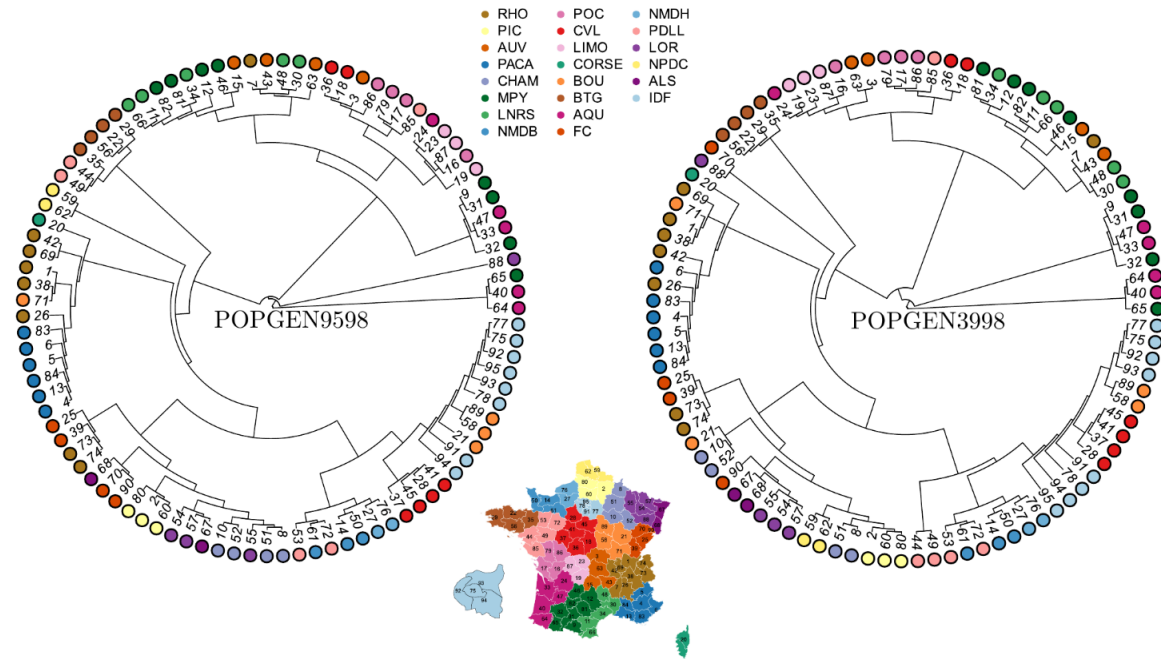


Figure S20: Comparison of the TVD tree of departments calculated using PBWT-paint similarity matrices for the POPGEN9598 sample (left) and the POPGEN3998 sample (right). The normalised Robinson-Foulds distance between the two trees is 0.543 indicating that roughly half of the exact splits are shared between the two trees. Starting on the far right of both trees at departments 64, 40, and 65 which is the first split in both, and then working around counter-clockwise, the similarity between the two trees becomes apparent even if the order of some splits is slightly different. If the distance between all pairs of departments is measured on the two plots in this way (number of counter clockwise steps), the two sets of distances across all department pairs have a correlation of 0.810. A few noticeable differences include the behaviour of the Vosges cluster that stands out more in POPGEN9598 whereas in POPGEN3998 it splits off with department 70 but the two still form a branch of their own. A few other departments such as 21, 44, and 49 also behave differently. In POPGEN3998 departments 21, 10, and 52 now form their own group earlier, these departments have relatively few individuals when considering the POPGEN9598 set and so potentially the more even sampling in POPGEN3998 aids in resolving this part of the tree. Departments 44 and 49 contain individuals on both sides of the Loire and hence these departments may be challenging to place in this analysis. On the left they group with the other departments from Brittany, on the right they group with other departments from Pays de la Loire and Normandie. In both cases the department 85 which is geographically very close is placed very far away on the tree along with other departments from Poitou Charente.