

Supplementary Materials

Building a French reference panel to study genetic variability between and within French regions: the POPGEN project

Supplementary Note 1 - Selection process of 15,000 volunteers from the Constances Cohort	2
Supplementary Note 2 - Details on identifying outliers with preliminary FineSTRUCTURE analyses	7
Supplementary Note 3 - Selection of 4000 individuals for Whole-Genome Sequencing	8
Supplementary Tables	9
Supplementary Figures	19

Supplementary Note 1 - Selection process of 15,000 volunteers from the Constances Cohort

The section below presents the different steps of the selection process performed for the POPGEN study.

- 1- Presentation of the Constances data used for POPGEN Study
- 2- Processing of geographic data for the purposes of the POPGEN Study
- 3- Selection of 15,000 volunteers for the POPGEN Study

1- Presentation of Constances data used for the POPGEN Study

In 2019, we worked with Constances coordination team to add a question (question 36) to the annual questionnaire on the birthplaces of parents and grandparents. Question 36 was “for each of your ancestors (father, mother, grandparents) indicate whether they were born in France or in another country. If they were born in France, specify the names of the departments and municipalities”. The year of birth of the ascendants was also requested as shown in **Figure S1**. Volunteers also had the option of not responding by ticking a dedicated box. Volunteers could answer by a dedicated and personalised website or by postal mail.

Additional information was also obtained from Constances that were needed for our selection:

- Places and partial dates of birth (month/year) of volunteers who agreed to have their data chained with French administrative databases. Birthplaces were in the Official Geographic Code (COG) format where each French commune, canton, district, department, region or overseas communities as well as foreign countries and territories has its own coding serving. The COGs provided allow us to find the department and the commune for individuals born in France or the country for those born outside France.
- Date (month/year) of inclusion in the cohort and the examination center where the inclusion took place
- Information concerning their follow-up in the cohort (if the volunteer no longer wishes to participate in Constances and/or other supported projects, such as POPGEN, and if the volunteer is known to be deceased).

The 2019 Questionnaire was sent to 180,260 Constances volunteers and, in December 2020 we had information about question number 36 for 118,042 volunteers. Among them we excluded:

- 7,086 individuals who either had withdrawn their consent to participate in Constances, were deceased or indicated they did not want to answer the questionnaire.
- 25,984 volunteers who had at least one of their grandparents born outside France. This was done after manual curation of the information they provided on the commune/department and on the item ‘born in France’ or ‘born outside France’ as follows
 - a. When instead of giving the information on the commune and/or department of all their ascendants, the volunteer indicated "same", we replaced it by the corresponding commune and/or department.
 - b. When information was discordant with several boxes checked for the item ‘born in France’ or ‘outside France’ or ‘I do not know’, we only kept the most consistent according to the data indicated in the municipality and/or department.
 - c. When ascendants were born in a country other than France or in former French colonies, we verified that the volunteer indicated ‘born outside France’
 - d. For volunteers with ascendants born in Alsace at a time where Alsace was part of Germany, we verified they indicated ‘born in France’.

In total, we were left with 84,972 individuals to enter the second phase of data monitoring where geographic data was processed.

2- Processing of geographic data for the POPGEN purpose

To calculate a distance between ascendants and to anchor individuals by a point (see barycentre calculation) on the map of France, we needed to transform the various geographic data into geographic coordinates (with latitude and longitude). To do it, we built a database of geographic coordinates of French metropolitan municipalities using free openstreetmap (OSM) data. This OSM database was produced from the file `france-latest.osm.pbf` dated October 29, 2020 and the file `departments-without-sea.geojson` used to extract geographical division of the departments.

(<https://github.com/gregoireddavid/france-geojson/blob/master/departements-avec-outre-mer.geojson>).

We were able to extract from these files all the metropolitan capitals, villages and hamlets and their corresponding departments using `osmium-tools` (<https://osmcode.org/osmium-tool/>). A total of 237,617 named entities were extracted. The source code and the file database are available on github. https://github.com/dridk/france_place_gps.

To assign the geographical coordinates of the communes of Constances volunteers via the OSM database, the first step was to tokenize the 2 databases with the same methods. The tokenisation involved turning all character strings such as 'xxx, yyy aaa' into the form XXX-YYYY-AAA; i.e. capitalising all letters, replacing accented characters with non-accented equivalents, removing punctuation and turning white-space between string fragments into dashes (-). Furthermore, some very common string fragments in French commune and department names were homogenised, for examples 'SAINT-' became 'ST-' and 'HAUTE-' became 'HTE-', in order to facilitate string comparisons to known databases. Information concerning the ascendants could be in digital format too: department number and/or five-digit postcodes for the communes. This data was also tokenised. The tokenization was carried out on the 6 ascendants' data of 84,972 volunteers. Naturally, the equivalent tokenisation was applied to all department and commune names in the OSM database.

For the assignment of geographical coordinate, we used the R-function 'stringdist' [1] to ascertain the similarity between what was written in the questionnaire and in our database. We were looking for a perfect match between the names, as spelled by the volunteer, of the commune and the department, with a single commune in the same department referenced in the OSM database. For example: raw data of '64' and 'PAU' would easily be matched to the city of Pau in department 64 or *Pyrénées Atlantiques*. For all these communes, the geographical coordinates (latitude and longitude) of the OSM base were assigned to them. However, we would often be confronted by imperfect matches and/or multiple matches. In the case of an imperfect match, we did allow for the character string supplied by the volunteer to be corrected. For example, the raw data might have been tokenised to '44' and 'ST-PHILBERT-DE-GRAND-LIEU' which has a string distance on just 1 to the town of 'ST-PHILBERT-DE-GRAND-LIEU' (spelt with just one 'L') and correctly in the department 44 (*Loire-Atlantique*). In cases like this, we wished to automatically make such a correction as the chances that the volunteer was referring to a different commune were clearly very small. For all these communes, the geographical coordinates (latitude and longitude) of the OSM base were assigned to them too.

For all communes that could not be matched in this way (unfound or multiple matches in OSM, more than 1 string distance correction, etc.) a more complex version of our algorithm was applied. This algorithm allows a greater string distance for longer character strings, being strict of changes of the first letter of the character string, and deletions, insertions, substitutions and/or transpositions of characters according data available in OSM database. In cases where ambiguity or significant

uncertainty remained, we set either department or commune fields as 'UNCORRECTED' to differentiate from actual missing data. No results from this more complex script were used in this study without manual verification.

Across the 84,972 questionnaires, there were a potential of 509,832 pairs of longitude and latitude coordinates to assign for the six ancestors of each volunteer. After the curation described above, commune data was missing for 249,123 (48.9%) ancestors, and longitude and latitude was successfully assigned for 237,502 (46.6%) ancestors with our script, leaving data for 23,207 (4.6%) without a successful correction. We set out to manually assign longitude and latitude to these 23,207 ancestors by manually inspecting their questionnaire data and checking the propositions of our more complex script, but finally this was only done for roughly half of these data entries. This was because with the successful corrections, initial exploration showed that some departments were already "well-represented" (departments in *Bretagne* and *Nord-Pas-de-Calais*, *Pyrénées-Atlantiques*, etc.) while others had too few potential candidates for selection. It was decided to prioritize these "under-represented" departments/regions (PACA, RHO, IDF, CVL, LNRS, LIMO, MPY, BOU, CHAM, CORSE, BFC and PIC; see Table S1 for definitions) during manual curation. We reviewed over 10,000 data entries by hand, we considered the raw data (for paper questionnaires the automatic proofreading could transform an "o" into "0", for data entered online a "z" instead of an "a" - key neighbour on 'azerty' keyboard, etc.) as well as the data of other descendants. Some volunteers who did not know the commune or the department indicated administrative regions (*Aquitaine*, *Franche-Comté*, *Île-de-France*, etc.) and/or historical regions (*Béarn*, *Anjou*, *Maine*, etc.). An inhouse-dictionary was therefore created listing each region and the associated departments. For the particular case of the Ile-de-France departments, several correspondences have been added to the dictionary. The current department 75 corresponds to the city of Paris; so, if Paris was indicated as a commune and/or department, the result is Paris - 75. Before the 1960s, department 75 corresponded to the "Seine", 78 to "Seine-et-Oise". The modification of departments 75, corresponding to "Seine", and 78 for "Seine-et-Oise" at the end of the 1960s, impacted the departmental division of the current *Île-de-France* with the creation of new departments (91, 92, etc.). The Corsica region has been divided into 2 departments *Corse-du-Sud* and *Haute-Corse* since 1975 and their "number" is respectively 2A and 2B (initially 20). To facilitate the treatment of departments, it was chosen to consider all of Corsica as the same department, bearing the number 20.

Proposed corrections from our more complex script were checked and in cases where ambiguity or too much uncertainty remained, (duplicates, several possible spellings, etc.), we set either department or commune fields as 'UNCORRECTED' to differentiate from actual missing data.

All manual corrections were carried out via OSM database and/or on internet-based research (Google, Wikipedia, INSEE, etc.). Our manual corrections allowed for an additional 9,336 (1.83%) pairs of longitude and latitude co-ordinates to be assigned.

For the birthplaces of the volunteers, that is to say the COG data, it was necessary to reassign them the commune (name in full) in order to be able to find the corresponding geographical coordinates of the OSM database. COG-commune correspondence is made available by INSEE on its website. Each year, INSEE updates the different correspondence tables between the COG (municipality, canton, country, etc.) and the names in full, containing each update (for example merger of 2 communes under a new name on a specific date). As this data was not necessary for the selection process, they were processed more recently using the files provided by INSEE in 2024 were used (<https://www.insee.fr/fr/information/7766585>).

With the longitudes and latitudes of each commune, a minimum radius was calculated (using spherical geography), so as to obtain the smallest circle containing the municipalities of the 4 grandparents and classified the individuals into 3 levels of precision based on this radius: radius <15km, 15-30km, 30-50km. Initial exploratory tests quickly showed that there would not be 15,000 candidates across the entire metropolitan France, even with a radius of 50km. We therefore chose to broaden our selection criteria while integrating a prioritization of these criteria, Table S3 details the eventual 18 different selection criteria that were used and Table S4 details the number of individuals in each category throughout the course of the POPGEN recruitment. The 18 criteria are split into three larger classes, 'Excellent', 'Very Good' and 'Good'.

The selection of volunteers followed the order of priority indicated in the 2nd column of the Table S3. For criteria 10, 14, 15, 16, 17, the radius of the circle integrating the communes of birth of the two parents was calculated in the same way as for the grandparents for other criteria.

To be able to transcribe the origin of eligible volunteers, via the data of their ascendants, a department, as well as geographical coordinates in this department, were assigned to them with the following rules.

- ✓ When the majority of ascendants were born in the same department, this department was attributed to volunteers.
For example, 3 grandparents born in department A and one in B: the assigned volunteer's department will be A.
- ✓ When ascendants were born in different departments with no majority, one of these departments was randomly drawn and assigned to the volunteer.
First example, 1 grandparent born in A, other in B, other in C and last in D: the assigned volunteer's department could be either A, B, C or D.
Second example, 2 grandparents born in A and 2 others in B: the assigned volunteer's department could be either A or B.
- ✓ For the REGION criteria, a department was occasionally chosen at random among the concerned departments of the region in question. For example, if a volunteer simply indicated that all four grandparents were born in PACA (see Table S1), one of the six departments from PACA (04, 05, 06, 13, 83, 84) was randomly assigned.

Once departments were assigned to all individuals, the geographical coordinates for each eligible volunteer, with latitude and longitude, was assigned via the following rules: we select the sub-set of ancestors who have geographical coordinates falling within the department assigned to the individual, and take the average (barycentre) of those coordinates. In the case of criteria such as 12 or 18, where no commune data was available but only department data, we assigned geographical coordinates by randomly selecting a commune from the OSM database restricting to the assigned department of the individual.

3- Selection of 15,000 volunteers for the POPGEN Study

Applying the geographic criteria listed in Table 1 to the data, 3 more individuals were removed from the selection process because their grandparent's place of birth was actually out-of-France. At the end 28,174 volunteers were eligible for the next selection step. Our next selection criterion was to select only individuals born before 1985 so that the distribution of grandparent's birthplaces of our volunteers would be well represented by the 1901 census in France. This removed 1,739 candidates.

For the remaining individuals we asked the Constances coordination for further information to check the eligibility of these candidates. Non-inclusion criteria involved self-reported Goujerot-Sjögren

syndrome, which induce lack of salivation and therefore impacts the salivary self-sampling, having history of a bone marrow transplant, which may lead to a risk of mixed DNA with that of the donor, death of the individual, or recent withdrawal of consent, which make any contact with volunteers impossible. Finally, as for several years, the Constances and GAZEL cohort share the same annual questionnaire, we removed from the process selection, GAZEL volunteers who were already sequenced in the framework of FranceGenRef project. Across all these selection criteria, a further 161 candidates were removed from consideration for POPGEN, leaving 26,274 candidates.

The next step was to decide on a strategy to select the 15,000 individuals among the 26,274 in order to best cover the different geographic areas of Metropolitan France, while maintaining a balanced sex ratio.

For that, we derived target sample size by administrative departments based on the 1901 population census. The choice of department was based on previous studies in France that showed that their boundaries best reflect historical distinctions between cultural groups and fine-scale population structure [7,8]. Because most of the volunteers grandparents were born at the beginning of the twentieth century, we chose the 1901 census as reference. Unfortunately, the 26,274 candidates were not distributed in a manner where 15,000 individuals could be selected whilst respecting the desired department-wise distribution based on the 1901 census. For any department where too few candidates could be found, substitutes were therefore selected in an ad-hoc manner based on proximity between departments and the larger 'old' administrative regions of France. For example, rather many individuals were selected from department 13 (*Bouches-du-Rhône*) as other departments such as 84 (*Vaucluse*) or 83 (*Var*) in the region *Provence-Alpes-Côte d'Azur* had insufficient numbers of candidates.

We also privileged substitution with higher ranking selection criteria. In many cases this would come down to having to select a random group of substitutes from a large pool of potential substitutes and in this case the only departure from a completely random selection was to try to take slightly more male candidates given we had observed slightly more female candidates in general across the study. The target recruitment totals for each department were calculated from the 1901 census of France. Departments 67 and 68 were not part of France in 1901 and so not included in the census data so we set their target for recruitment at 167 individuals (the mean of all other departments). Departments 91, 92, 93, 94, and 95 also did not exist in 1901, they were at that time incorporated into departments 75, 77, and 78 to make up the Ile de France (IDF) region. As globally we had far less candidates from IDF than the target from the 1901 census, we took all possible candidates from departments 91, 92, 93, 94, and 95.

A list of 15,000 volunteers were sent to Constances to plan the sending of inclusion letters. We were then informed that 2 volunteers were no longer eligible (one had died and the other had withdrawn their consent - see point 1). These 2 volunteers were replaced and the distribution of the final list of 15,000 volunteers selected for POPGEN is detailed in Table S4 and S5. The volunteer selection process is summarized in Figure 1 in the main text and in Figure S2.

Supplementary Note 2 - Details on identifying outliers with preliminary FineSTRUCTURE analyses

The FS analyses used for the detection of outliers during the selection do not correspond precisely to the three FS runs is given in the main text, the FS analyses here were of a preliminary nature given that it was necessary to complete the selected list of 4000 individuals for sequencing as soon as possible in order to respect the timeline of the POPGEN project, specifically to begin the lengthy process of preparing the 4000 DNA samples for whole-genome sequencing and the delivery of said samples to the sequencing team at the CEA in Evry, France. Hence FS was run on phased data only using SHAPEIT4 and it was also completed on an earlier data freeze before the exact quality control measures for the genotyping data were decided. FS does not produce the same results between different runs. As described in the main text, we ran FS three times, creating a hierarchical clustering tree each time as described in Leslie et al [2]. For each run we then sort to cut the tree such that we had the largest number of groups while ensuring that each group contained at least 40 individuals. This was achieved in the same manner to the FineSTRUCTURE results presented in the main text. Then for each cluster identified in any of the three final hierarchical trees, we calculated the distance in kilometres between each individual and all other individuals in the cluster. We designated an outlier if for an individual any of the 10 closest neighbours within the cluster was found to be over 200 kilometres away. An example is given in Figure S14. The motivation of these steps was to try as best as possible to avoid sequencing individuals whose geographical anchorage did not correspond very well to their genomic similarity to their peers within the sample, bearing in mind that the sequencing data's primary objective is to produce estimates of allele-frequencies that can be stratified using the geographical information of the sample.

Supplementary Note 3 - Selection of 4000 individuals for Whole-Genome Sequencing

The following criteria were used to exclude individuals from the selection for WGS (see Figure S15):

- Lack of precise department data: 2 individuals removed (pertaining to the 4 individuals who replied in the place of their partners, see Table S4).
- Poor/low quality of DNA sample observed during genotyping, suggesting that WGS might fail: 81 individuals were removed.
- Individuals not agreeing to general re-use of their data: 185 individuals. Given that the eventual panel of 4,000 WGS samples will be used for various different purposes, we desired not to lose too many individuals for re-use. Of course, individuals may still withdraw consent from the study at a later date.
- Individuals excluded from the quality control of the genotyping data: 174 individuals.
- Outliers based on initial FineSTRUCTURE analyses: 126 individuals (see Supplementary Note 2). (1 was already excluded for the “no data re-use” criteria).
- Individuals with a high inbreeding coefficient: 48 in total, 9 were already FS outliers, and 1 was also already excluded for the “no data re-use” criteria. Inbreeding coefficient was estimated with FSuite v1.0.3 [3,4] and individuals with an estimated inbreeding coefficient above 3% were excluded.
- One individual of each pair (or set) of individuals with a Pi-Hat coefficient greater than 0.06: 60 individuals.

After filtering, we were left with a list of 9,106 eligible individuals. Our aim was to achieve a uniform recruitment across all departments and between men and women. This would mean selecting 21 men and 21 women from each department. We also wanted to have individuals with the most precise anchorage on the map and so we privileged individuals with the ‘best’ selection criteria (see Table S3). To further provide a uniform coverage of individuals across the map we also tried to avoid selecting individuals close to one another and leaving empty gaps on the map.

The following procedure was applied per department: Start by counting how many men and women are eligible for selection and had previously a geographical selection criterion between 1 and 6 (see Table S3). If sufficient numbers of individuals were available, we randomly selected one individual, then alternated between men and women, and we always took the individual with the highest minimum distance in kilometres to the individuals already selected, terminating with 21 men and 21 women selected.

For many departments, less than 42 candidates were available in this manner, leading us to first pass on to lower geographical selection criteria, first by searching through criteria 7-10 and then for 11-18. If there were still too few candidates, we allowed for some sex-imbalance per department. For some departments there were still too few candidates, in this manner we were only able to select 3514 candidates (1760 men and 1808 women). We decided to fill in the remaining places randomly from departments where there were still individuals with geographic selection criteria between 1 and 6 available, in an even manner across such departments. This led to many departments contributing 55 individuals in total, others contributing 42, and others contributing less than 42. In the departments with 55 individuals selected, we chose 29 men and 26 women in order to reach a final balance of 2000 men and 2000 women.

Supplementary Tables

Department Number	Department Name	Modern Regions (since 2015)	Old Regions (until 2015)
01	Ain	Auvergne-Rhône-Alpes (ARA)	Rhône-Alpes (RHO)
02	Aisne	Hauts-de-France (HDF)	Picardie (PIC)
03	Allier	Auvergne-Rhône-Alpes (ARA)	Auvergne (AUV)
04	Alpes-de-Haute-Provence	Provence-Alpes-Côte d'Azur (PACA)	Provence-Alpes-Côte d'Azur (PACA)
05	Hautes-Alpes	Provence-Alpes-Côte d'Azur (PACA)	Provence-Alpes-Côte d'Azur (PACA)
06	Alpes-Maritimes	Provence-Alpes-Côte d'Azur (PACA)	Provence-Alpes-Côte d'Azur (PACA)
07	Ardèche	Auvergne-Rhône-Alpes (ARA)	Rhône-Alpes (RHO)
08	Ardennes	Grand Est (GEST)	Champagne-Ardenne (CHAM)
09	Ariège	Occitanie (OCC)	Midi-Pyrénées (MPY)
10	Aube	Grand Est (GEST)	Champagne-Ardenne (CHAM)
11	Aude	Occitanie (OCC)	Languedoc-Roussillon (LNRS)
12	Aveyron	Occitanie (OCC)	Midi-Pyrénées (MPY)
13	Bouches-du-Rhône	Provence-Alpes-Côte d'Azur (PACA)	Provence-Alpes-Côte d'Azur (PACA)
14	Calvados	Normandie (NMD)	Basse-Normandie (NMDB)
15	Cantal	Auvergne-Rhône-Alpes (ARA)	Auvergne (AUV)
16	Charente	Nouvelle-Aquitaine (NAQ)	Poitou-Charentes (POC)
17	Charente-Maritime	Nouvelle-Aquitaine (NAQ)	Poitou-Charentes (POC)
18	Cher	Centre-Val de Loire (CVL)	Centre (CVL)
19	Corrèze	Nouvelle-Aquitaine (NAQ)	Limousin (LIMO)
20	Corse	Corse (CORSE)	Corse (CORSE)
21	Côte-d'Or	Bourgogne-Franche-Comté (BFC)	Bourgogne (BOU)
22	Côtes-d'Armor	Bretagne (BTG)	Bretagne (BTG)
23	Creuse	Nouvelle-Aquitaine (NAQ)	Limousin (LIMO)
24	Dordogne	Nouvelle-Aquitaine (NAQ)	Aquitaine (AQU)
25	Doubs	Bourgogne-Franche-Comté (BFC)	Franche-Comté (FC)
26	Drome	Auvergne-Rhône-Alpes (ARA)	Rhône-Alpes (RHO)
27	Eure	Normandie (NMD)	Haute-Normandie (NMDH)
28	Eure-et-Loir	Centre-Val de Loire (CVL)	Centre (CVL)
29	Finistère	Bretagne (BTG)	Bretagne (BTG)
30	Gard	Occitanie (OCC)	Languedoc-Roussillon (LNRS)
31	Haute-Garonne	Occitanie (OCC)	Midi-Pyrénées (MPY)
32	Gers	Occitanie (OCC)	Midi-Pyrénées (MPY)
33	Gironde	Nouvelle-Aquitaine (NAQ)	Aquitaine (AQU)
34	Hérault	Occitanie (OCC)	Languedoc-Roussillon (LNRS)
35	Ille-et-Vilaine	Bretagne (BTG)	Bretagne (BTG)
36	Indre	Centre-Val de Loire (CVL)	Centre (CVL)
37	Indre-et-Loire	Centre-Val de Loire (CVL)	Centre (CVL)

38	Isère	Auvergne-Rhône-Alpes (ARA)	Rhône-Alpes (RHO)
39	Jura	Bourgogne-Franche-Comté (BFC)	Franche-Comté (FC)
40	Landes	Nouvelle-Aquitaine (NAQ)	Aquitaine (AQU)
41	Loir-et-Cher	Centre-Val de Loire (CVL)	Centre (CVL)
42	Loire	Auvergne-Rhône-Alpes (ARA)	Rhône-Alpes (RHO)
43	Haute-Loire	Auvergne-Rhône-Alpes (ARA)	Auvergne (AUV)
44	Loire-Atlantique	Pays de la Loire (PDLL)	Pays de la Loire (PDLL)
45	Loiret	Centre-Val de Loire (CVL)	Centre (CVL)
46	Lot	Occitanie (OCC)	Midi-Pyrénées (MPY)
47	Lot-et-Garonne	Nouvelle-Aquitaine (NAQ)	Aquitaine (AQU)
48	Lozère	Occitanie (OCC)	Languedoc-Roussillon (LNRS)
49	Maine-et-Loire	Pays de la Loire (PDLL)	Pays de la Loire (PDLL)
50	Manche	Normandie (NMD)	Basse-Normandie (NMDB)
51	Marne	Grand Est (GEST)	Champagne-Ardenne (CHAM)
52	Haute-Marne	Grand Est (GEST)	Champagne-Ardenne (CHAM)
53	Mayenne	Pays de la Loire (PDLL)	Pays de la Loire (PDLL)
54	Meurthe-et-Moselle	Grand Est (GEST)	Lorraine (LOR)
55	Meuse	Grand Est (GEST)	Lorraine (LOR)
56	Morbihan	Bretagne (BTG)	Bretagne (BTG)
57	Moselle	Grand Est (GEST)	Lorraine (LOR)
58	Nièvre	Bourgogne-Franche-Comté (BFC)	Bourgogne (BOU)
59	Nord	Hauts-de-France (HDF)	Nord-Pas-de-Calais (NPDC)
60	Oise	Hauts-de-France (HDF)	Picardie (PIC)
61	Orne	Normandie (NMD)	Basse-Normandie (NMDB)
62	Pas-de-Calais	Hauts-de-France (HDF)	Nord-Pas de Calais (NPDC)
63	Puy-de-Dôme	Auvergne-Rhône-Alpes (ARA)	Auvergne (AUV)
64	Pyrénées-Atlantiques	Nouvelle-Aquitaine (NAQ)	Aquitaine (AQU)
65	Hautes-Pyrénées	Occitanie (OCC)	Midi-Pyrénées (MPY)
66	Pyrénées-Orientales	Occitanie (OCC)	Languedoc-Roussillon (LNRS)
67	Bas-Rhin	Grand Est (GEST)	Alsace (ALS)
68	Haut-Rhin	Grand Est (GEST)	Alsace (ALS)
69	Rhône	Auvergne-Rhône-Alpes (ARA)	Rhône-Alpes (RHO)
70	Haute-Saône	Bourgogne-Franche-Comté (BFC)	Franche-Comté (FC)
71	Saône-et-Loire	Bourgogne-Franche-Comté (BFC)	Bourgogne (BOU)
72	Sarthe	Pays de la Loire (PDLL)	Pays de la Loire (PDLL)
73	Savoie	Auvergne-Rhône-Alpes (ARA)	Rhône-Alpes (RHO)
74	Haute-Savoie	Auvergne-Rhône-Alpes (ARA)	Rhône-Alpes (RHO)
75	Paris	Île-de-France (IDF)	Île-de-France (IDF)
76	Seine-Maritime	Normandie (NMD)	Haute-Normandie (NMDH)
77	Seine-et-Marne	Île-de-France (IDF)	Île-de-France (IDF)
78	Yvelines	Île-de-France (IDF)	Île-de-France (IDF)

79	Deux-Sèvres	Nouvelle-Aquitaine (NAQ)	Poitou-Charentes (POC)
80	Somme	Hauts-de-France (HDF)	Picardie (PIC)
81	Tarn	Occitanie (OCC)	Midi-Pyrénées (MPY)
82	Tarn-et-Garonne	Occitanie (OCC)	Midi-Pyrénées (MPY)
83	Var	Provence-Alpes-Côte d'Azur (PACA)	Provence-Alpes-Côte d'Azur (PACA)
84	Vaucluse	Provence-Alpes-Côte d'Azur (PACA)	Provence-Alpes-Côte d'Azur (PACA)
85	Vendée	Pays de la Loire (PDLL)	Pays de la Loire (PDLL)
86	Vienne	Nouvelle-Aquitaine (NAQ)	Poitou-Charentes (POC)
87	Haute-Vienne	Nouvelle-Aquitaine (NAQ)	Limousin (LIMO)
88	Vosges	Grand Est (GEST)	Lorraine (LOR)
89	Yonne	Bourgogne-Franche-Comté (BFC)	Bourgogne (BOU)
90	Territoire de Belfort	Bourgogne-Franche-Comté (BFC)	Franche-Comté (FC)
91	Essonne	Île-de-France (IDF)	Île-de-France (IDF)
92	Hauts-de-Seine	Île-de-France (IDF)	Île-de-France (IDF)
93	Seine-Saint-Denis	Île-de-France (IDF)	Île-de-France (IDF)
94	Val-de-Marne	Île-de-France (IDF)	Île-de-France (IDF)
95	Val-d'Oise	Île-de-France (IDF)	Île-de-France (IDF)

Table S1: Administrative divisions of Metropolitan France. Note that 20 Corse (Corsica) groups the two departments 2A and 2B; *Corse-du-Sud* and *Haute-Corse*.

		Date format	Geographic		
			In/out of France *	Department	commune
For birth	Parents and grand-parents	YYYY	1 = in France 2 = out of France 3 = I don't know	Name in full Or number	Name in full Or number
		YY Missing Others (unusable)	1+2 1+3 2+3 1+2+3 Missing	✓Post code ✓Department number ✓Commune name ✓Region name ✓Country name ✓Annotation (as like the other...) ✓Missing ✓Others (unusable)	✓Post code ✓Department number ✓Department name ✓Region name ✓Country name ✓Annotation (as like the other...) ✓Missing ✓Others (unusable)
	Volunteers	MM/YYYY	COG or missing		
For Constances inclusion	Volunteers	MM/YYYY	-	-	Name of the city

Table S2: List of date and geographic data formats transmitted by Constances (expected format in bold is given in the first row).

Selection Criteria		Label	Description
Excellent criteria	1	4GP <15km	Data is known for the communes of all 4 grandparents and they are all within a radius of < 15km
	2	4GP 15-30km	Data is known for the communes of all 4 grandparents and they are all within a radius of between 15 and 30km
	3	4GP 30-50km	Data is known for the communes of all 4 grandparents and they are all within a radius of between 30 and 50km
Very good criteria	4	3GP <15km SameDp	Data is known for the communes of 3 grandparents and they are all within a radius of < 15km. The 4 th grandparent has no commune data but comes from one of the same departments as the other three grandparents
	5	3GP 15-30km SameDp	Data is known for the communes of 3 grandparents and they are all within a radius of between 15 and 30km. The 4 th grandparent has no commune data but comes from one of the same departments as the other three grand-parents
	6	3GP 30-40km SameDp	Data is known for the communes of 3 grandparents and they are all within a radius of between 30 and 50km. The 4 th grandparent has no commune data but comes from one of the same departments as the other three grand-parents
	7	2GP <15km SameDp	Data is known for the communes of 2 grandparents and they are all within a radius of < 15km. The other two grandparents have no commune data but come from the same department(s) as the two grand-parents with commune data
	8	2GP 15-30km SameDp	Data is known for the communes of 2 grandparents and they are all within a radius of between 15 and 30km. The other two grand-parents have no commune data but come from the same department(s) as the two grandparents with commune data
	9	2GP 30-50km SameDp	Data is known for the communes of 2 grandparents and they are all within a radius of between 30 and 50km. The other two grandparents have no commune data but come from the same department(s) as the two grandparents with commune data
	10	2P <1945 <30Km GP SameDp, 0 miss	The two parents are both born before 1945, both have commune data, and are born within a radius of < 30km. Department data is non-missing for all 4 grand-parents, and the grandparents must be born in the same department(s) as the two parents.
Good criteria	11	1GP SameDP	Commune data is known for only one grand-parent. The other three are from the same department as the one with commune data.
	12	0GP SameDp	No commune data is present for any grandparents but they are all known to be born in the same department
	13	1-3GP <30km NeigDp	Commune data is present for between 1 and 3 grand-parents are they are located within a radius of <30km. The grandparents with no commune data are not born in the same department(s) as the ones with commune data (or this family would have already been selected with a MODERATE criteria), but are born in a neighbouring department (measured by distance) to the department assigned to this family.

	14	2P <1945 <30Km GP SameDp, 1 miss	Similar to criteria 12 but here we allow 1 grandparent to have missing data for department
	15	2P <1945 <30Km GP SameDp, 2 miss	Similar to criteria 12 but here we allow 2 grandparents to have missing data for department
	16	2P <1945 <30Km GP SameDp, 3 miss	Similar to criteria 12 but here we allow 3 grandparents to have missing data for departments
	17	2P <1945 <30Km, GP SameDp, 4 miss	Similar to criteria 12 but here we allow all 4 grandparents to have missing data for department
	18	REGION	Using department data only, all 4 grandparents are at least from the same region in France (see Table S1). Candidates for this criterion were only sought in the regions where too few candidates were present from the previous 17 criteria.

Table S3: List of the different categories of criteria, based on the geographical data of the volunteers' ancestors, eligible for selection in POPGEN.

Selection criteria	POPGEN eligible candidates	Saliva kits sent	Included individuals	Genotyping data	Selected for WGS	Final WGS set
1	6253	5347	3744	3576	1813	1776
2	1788	1636	1134	1071	598	589
3	1378	1275	904	867	503	498
4	1046	670	463	440	236	231
5	323	230	160	150	79	79
6	203	156	106	97	59	58
7	1431	913	625	591	84	89
8	126	92	71	70	16	15
9	78	54	38	36	10	10
10	1966	1079	732	700	128	139
11	506	113	73	69	14	14
12	2567	698	408	393	106	111
13	938	328	225	219	51	55
14	1004	275	188	181	47	48
15	1350	384	258	241	48	54
16	771	211	144	139	27	32
17	4013	1084	675 (+1*)	644 (+1)*	131	144
18	533	455	298	285	50	56
Partners	0	0	4 (3+1)	2 (1+1)*	0	0
Total	26274	15000	10250	9772	4000	3998

Table S4: Counts of individuals per selection criteria for the list of the 26,274 POPGEN eligible candidates, the 15,000 individuals who were sent saliva collection kits through the post, the 10,250 included individuals to the study, the resulting set of 9,772 individuals with genotyping data, and the 4000 individuals selected for prospective whole-genome sequencing (WGS). The details of the eligibility criteria, sending of saliva kits, genotyping, and WGS selection are provided in subsequent sections. (*) 4 saliva collection kits were returned but it was in fact the partners of the initially chosen individuals (who happened to also be participants of Constances) who supplied their DNA and consent. Of these 4, one partner was in fact an unselected candidate, coming under selection criteria 17, the 3 others were not among the 26,274 candidates. 3 of these additional saliva samples coming from partners are present in the genotyping data of 9,772 individuals, though none were selected for WGS.

		KM	Region	Department	FS Clusters	Model R ²
Model 1 Region & Department	AMOVA		0.094% (p=0.001)	0.071% (p=0.001)		
	IBD regression	-1.071×10^{-4} (p < 2e-16)	9.299×10^{-2} (p < 2e-16)	1.513×10^{-1} (p < 2e-16)		0.0449
Model 2 Region	AMOVA		0.116% (p=0.001)			
	IBD regression	-1.111×10^{-4} (p < 2e-16)	1.356×10^{-1} (p < 2e-16)			0.0395
Model 3 Department	AMOVA			0.161% (p=0.001)		
	IBD regression	-1.480×10^{-4} (p < 2e-16)		2.275×10^{-1} (p < 2e-16)		0.0390
Model 4 FS Clusters	AMOVA				0.173% (p=0.001)	
	IBD regression	-1.609×10^{-4} (p < 2e-16)			1.486×10^{-1} (p < 2e-16)	0.0290

Table S5: AMOVA and IBD-regression analyses comparing different hierarchical grouping levels in POPGEN: the 22 old administrative regions, the 95 departments, or the 87 FS clusters presented in Figure 5. For AMOVA, a 1000-fold permutation test is performed to estimate the significance of the variance explained by different hierarchical levels. For the IBD regression analyses, physical distance (KM) between pairs of individuals is always included in the model, the overall model R² values are also presented.

Department	Male	Female	Total
01	19	25 (28)	44 (47)
02	17	25	42
03	21	21	42
04	9	5	14
05	7	14	21
06	2	4	6
07	15 (16)	30	45 (46)
08	21	21	42
09	16 (17)	26	42 (43)
10	10 (9)	19 (18)	29 (27)
11	21	21 (22)	42 (43)
12	29	26 (27)	55 (56)
13	23	26 (27)	49 (50)
14	29 (30)	26	55
15	16	26	42
16	29	26	55
17	29	26	55
18	21	21	42
19	21	21	42
20	4	7 (6)	11 (10)
21	21	21	42
22	29	26	55
23	15	14	29
24	29	26	55
25	18	24	42
26	15 (14)	21	36 (35)
27	13 (12)	26 (25)	39 (37)
28	21	21	42
29	29	26	55
30	29	26	55
31	29	26	55
32	22	19	41
33	29	26	55
34	19	23	42
35	29	26	55
36	21	21	42
37	29	26	55
38	19	23 (22)	42 (41)
39	19 (20)	23	42 (43)
40	29	22	51
41	21	26	47
42	29	26	55
43	21	21	42
44	29	26	55
45	29 (30)	26	55 (56)
46	19	23	42
47	21	21	42
48	21	21	42
49	29	26	55
50	29	25	54
51	21	21	42
52	24	18	42
53	29	26	55
54	29	26	55

55	20	26	46
56	29	26	55
57	29	26	55
58	12	23	35
59	29	26	55
60	20	19	39
61	19	23	42
62	30	26	56
63	21	21	42
64	29	25	54
65	17	25	42
66	22	9 (8)	31 (30)
67	29	25	54
68	29	25	54
69	29	26	55
70	21	21	42
71	29	26	55
72	29	26	55
73	10	16	26
74	15 (14)	17 (14)	32 (28)
75	22 (24)	26 (27)	48 (51)
76	28	22 (23)	50 (51)
77	13 (12)	5	18 (17)
78	10	9	19
79	29	26	55
80	21 (20)	21	42 (41)
81	29	26	55
82	18	23 (22)	41 (40)
83	6	4	10
84	10	12	22
85	29	26	55
86	29	26	55
87	19	23	42
88	29	26	55
89	25 (26)	20 (21)	45 (47)
90	10	6	16
91	5	4	9
92	6	10	16
93	6 (5)	8 (7)	14 (12)
94	7	3	10
95	7 (6)	3	10 (9)
Total	2000 (1999)	2000 (1999)	4000 (3998)

Table S6: Counts per department and by sex for the 4000 individuals selected for WGS. Bracketed values indicate the final count of the 3998 individuals for whom WGS was successful, accounting for deviations from the initial selection of 4000 individuals and the 118 replacements required to reach the final sample (details provided in the main text).

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)

Silvous 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 S1: Question number 36 of the Constances 2019 annual questionnaire.

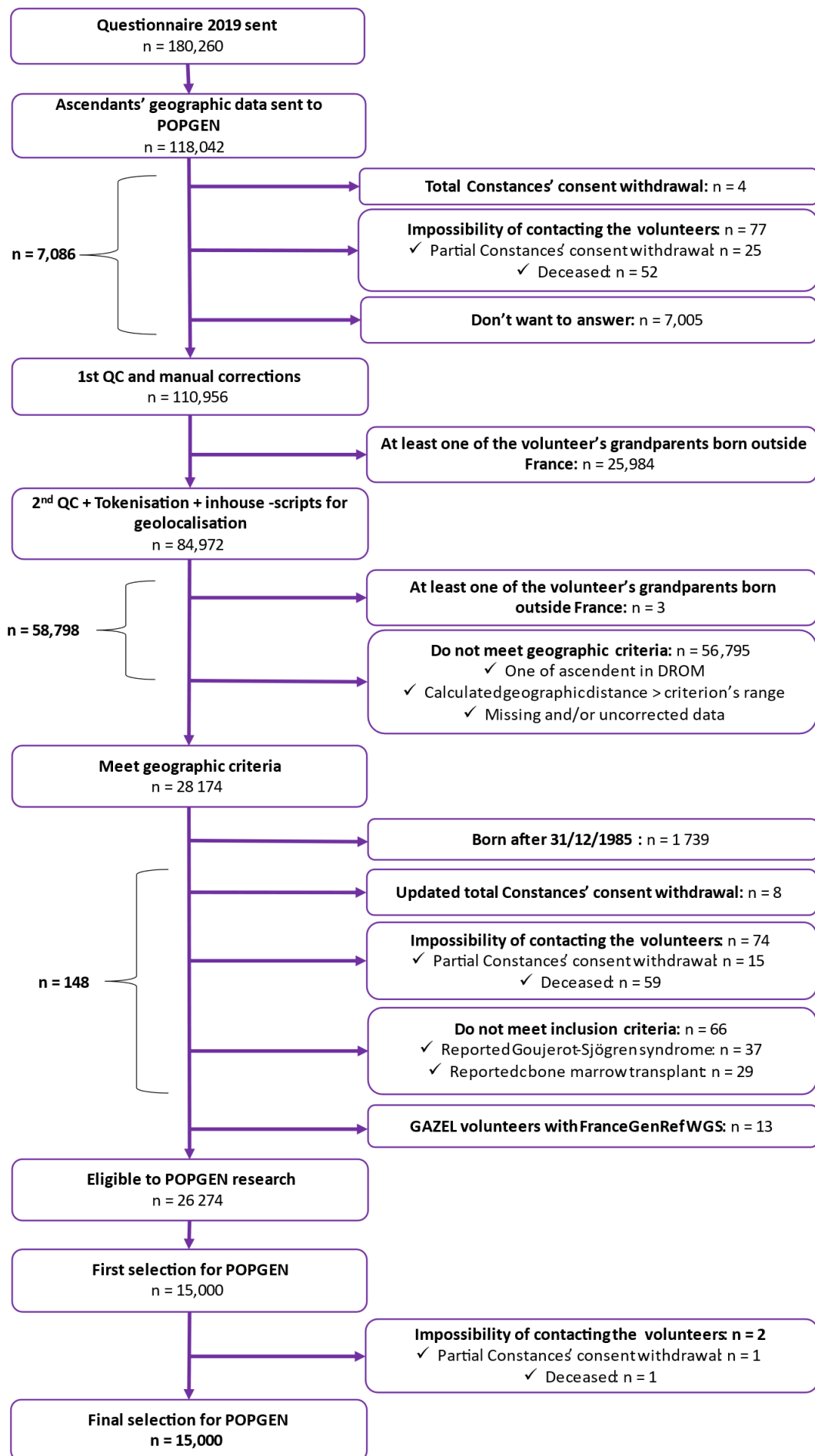


Figure S2: Flowchart of the selection of 15,000 volunteers of Constances for the POPGEN study.

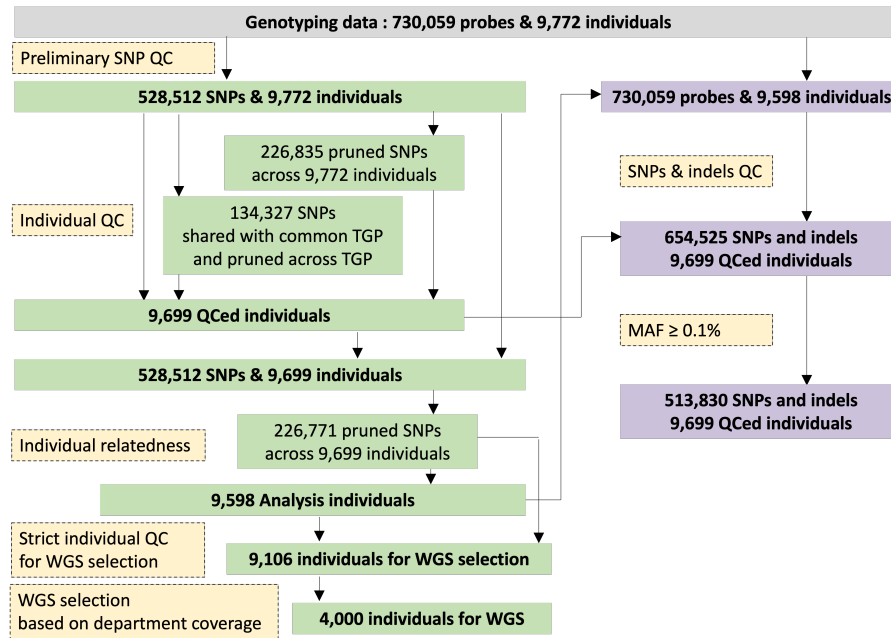


Figure S3: Overview of the quality control (QC) process. First a preliminary SNP QC was performed on which the individual QC was based. Individual QC led to 3 nested sets: a main QCed individual set, an unrelated QCed individual set, and a stringent set of individuals for WGS selection. The final dataset 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.

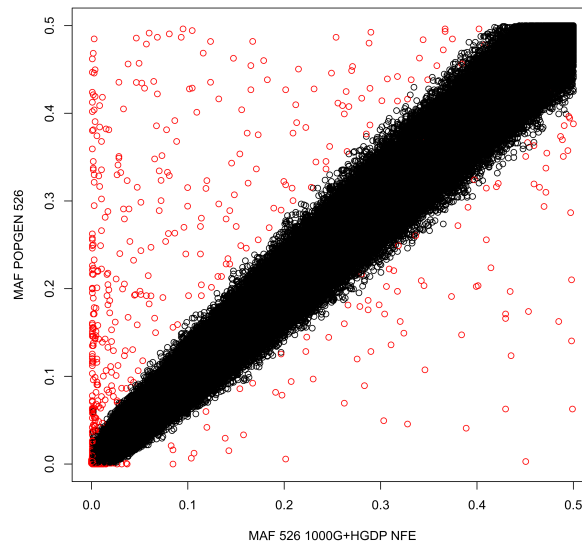


Figure S4: Comparison of Minor Allele Frequencies between POPGEN and NFE_1KG_HGDP [5] during the merge of the two datasets. Points highlighted in red were removed from subsequent analyses (see Methods in the main text).

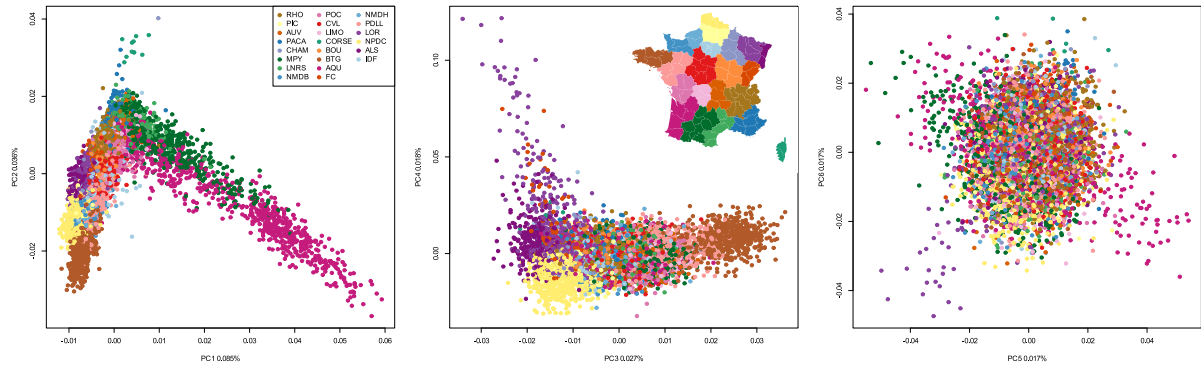


Figure S5: PCA of POPGEN9598. Components 1-6 are plotted with colours corresponding to the ‘old administrative regions’ of France (see Table S1). The plot on the left corresponds to Figure 2d in the main text.

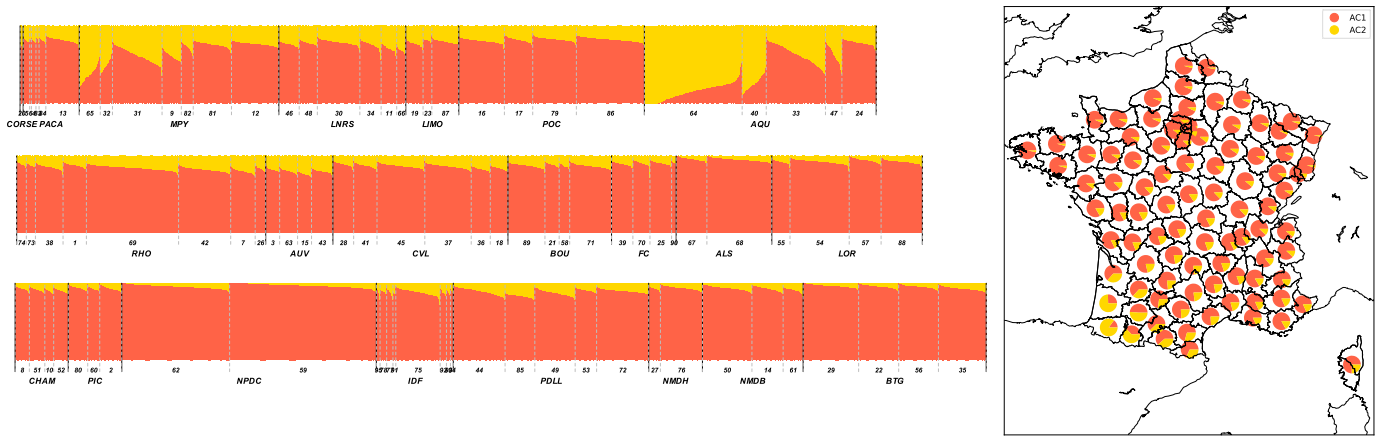
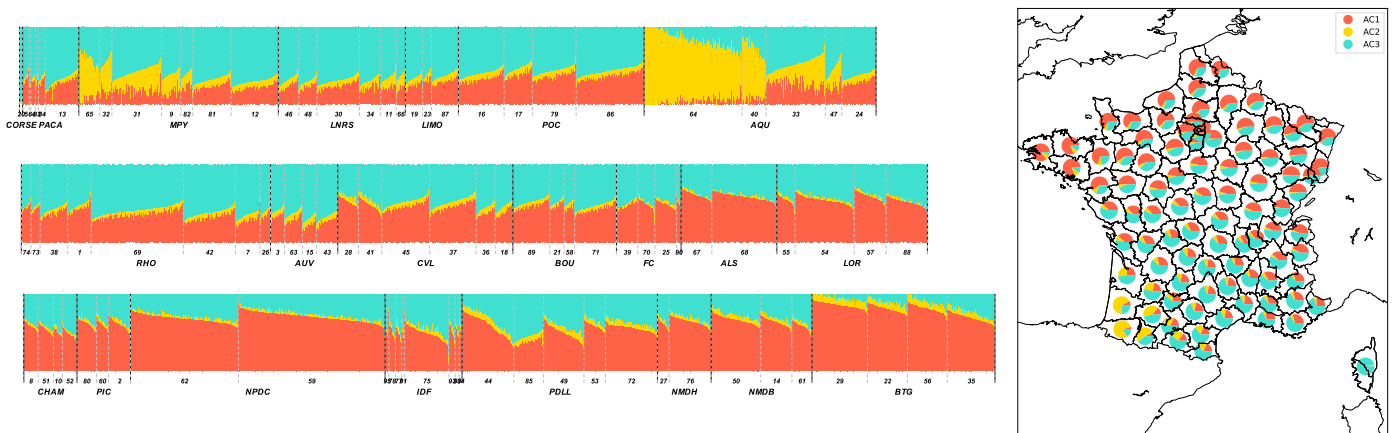
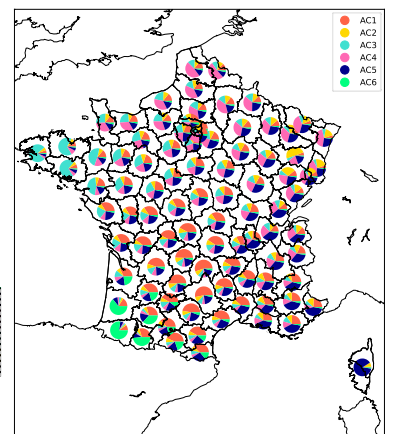
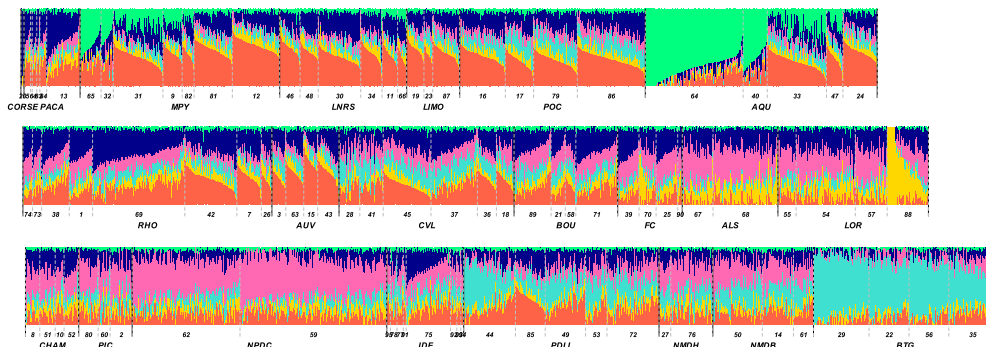
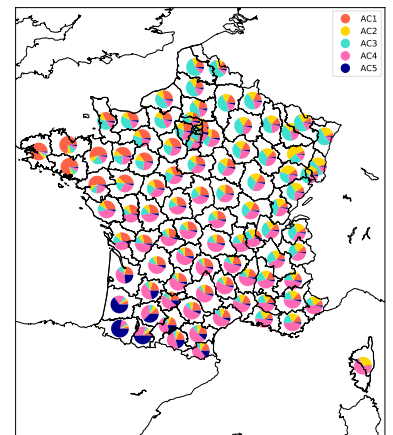
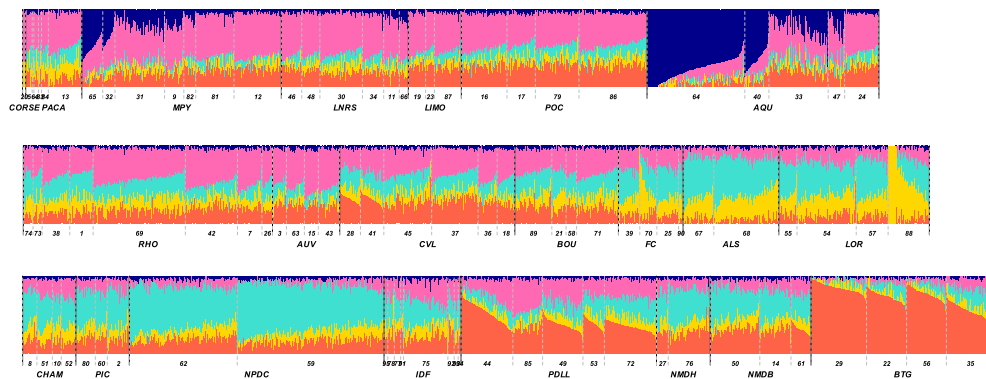
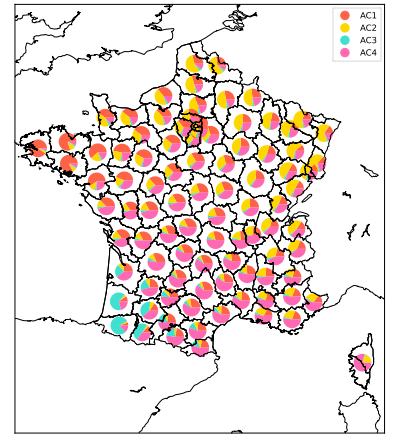
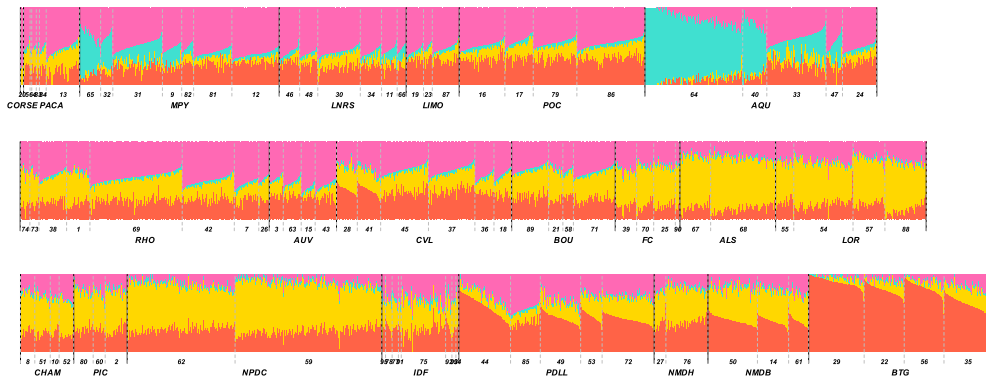


Figure S6: Unsupervised admixture plots on POPGEN9598 for different numbers of groups ($K=2$ above, and $K=3$ to $K=8$ below). In the three left panels (for each value of K), the individual admixture components of all 9,598 individuals are plotted, arranged first into 22 ‘old regions’ of France and then 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).





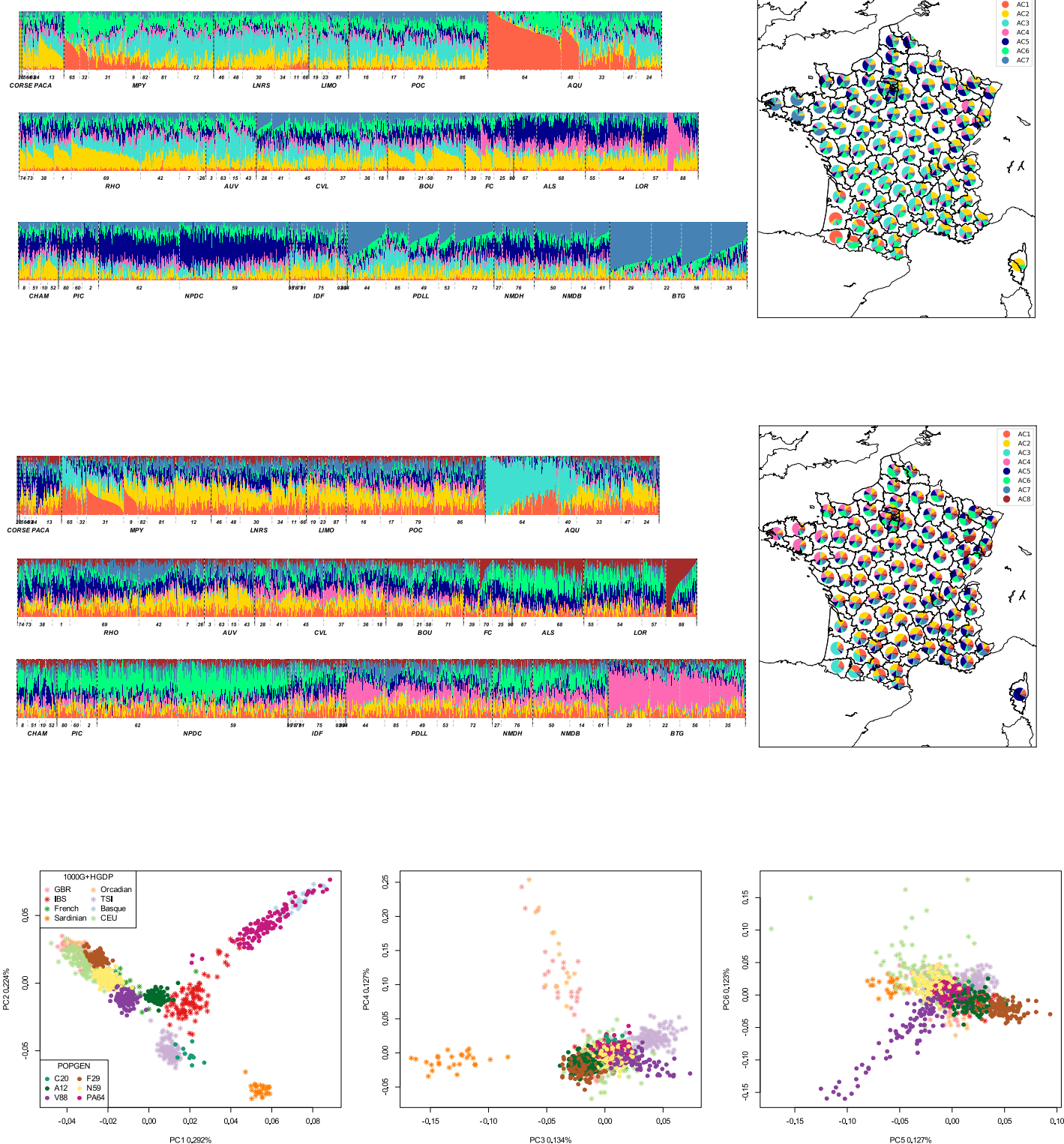


Figure S7: PCA analysis on the combined data of 526 individuals from 6 specific departments of POPGEN with 526 individuals from NFE_1KG_HGDP.

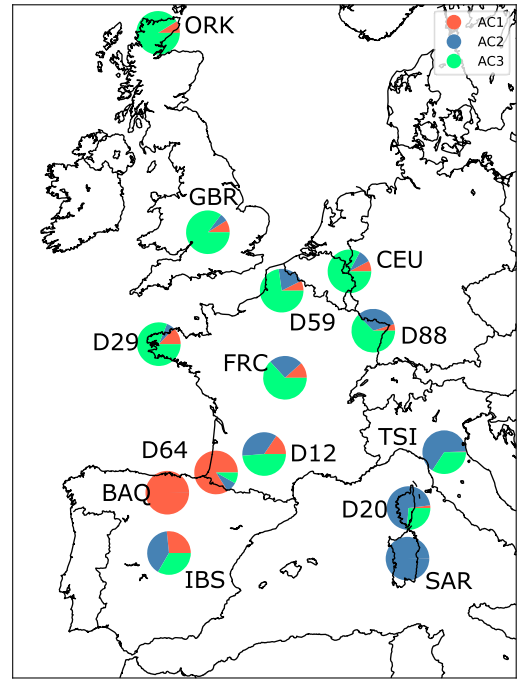
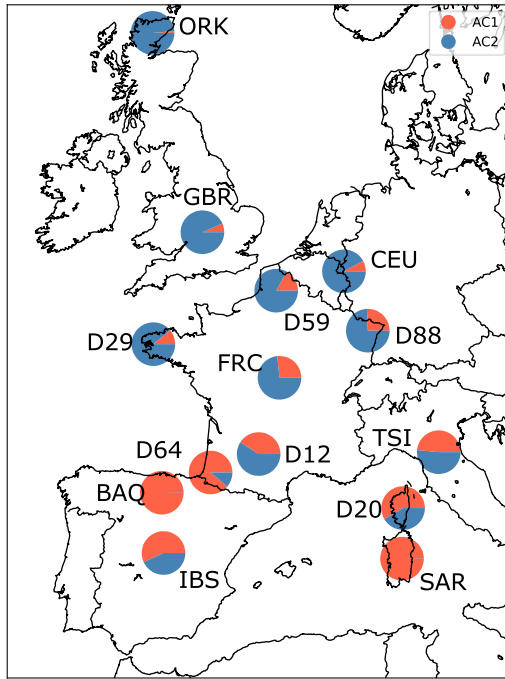
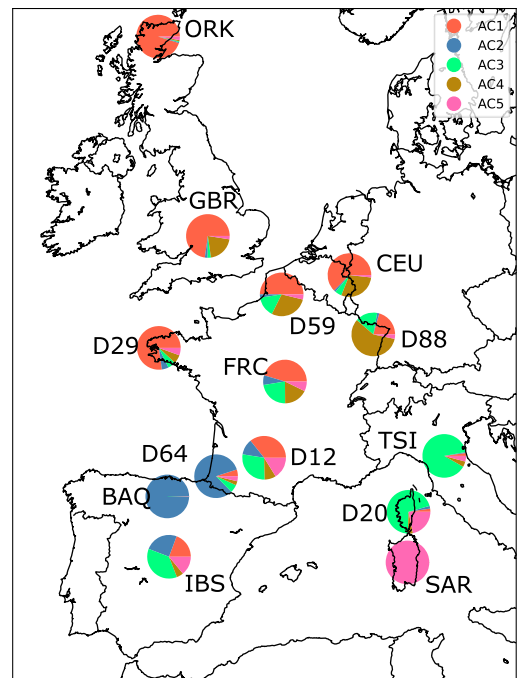
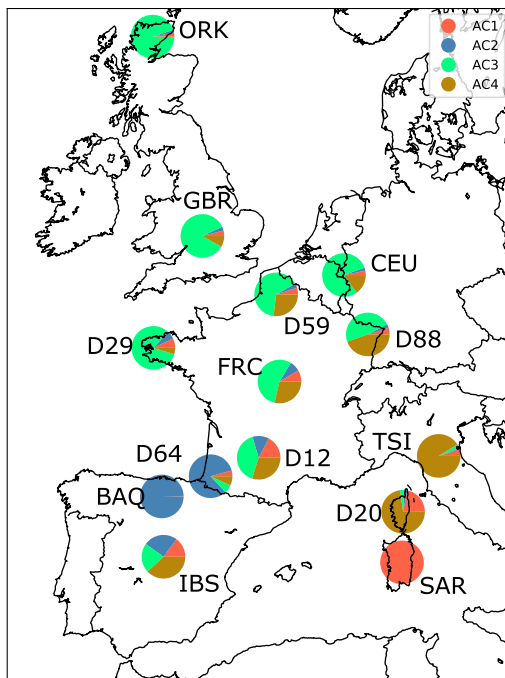
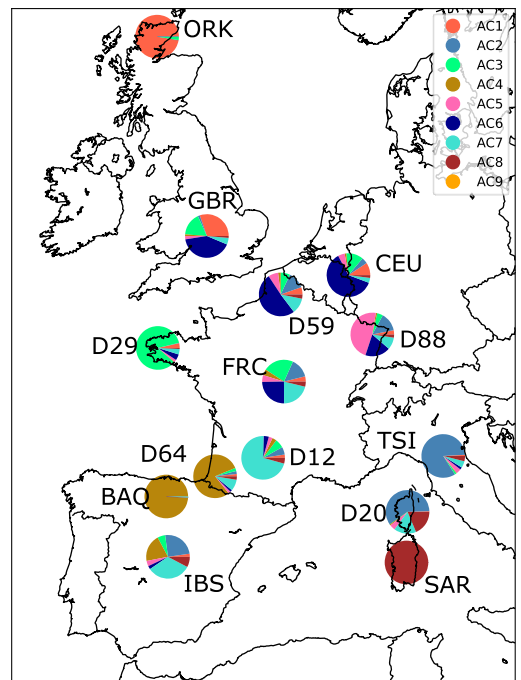
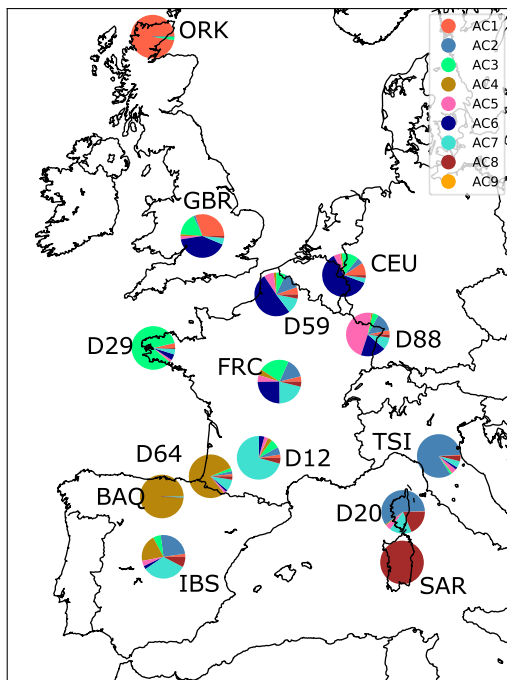
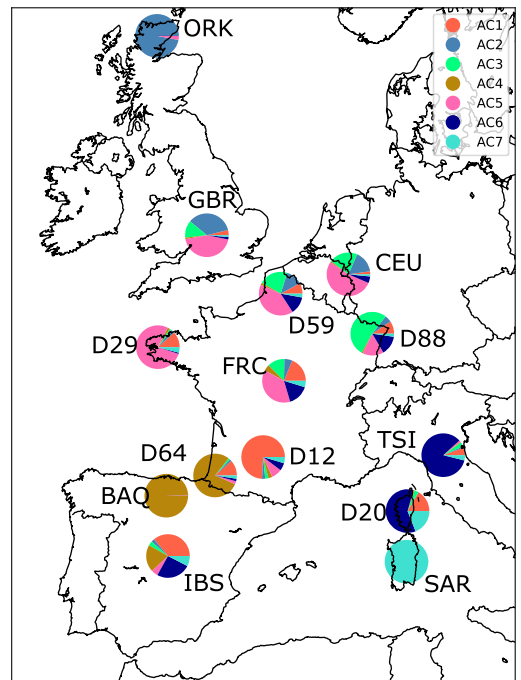
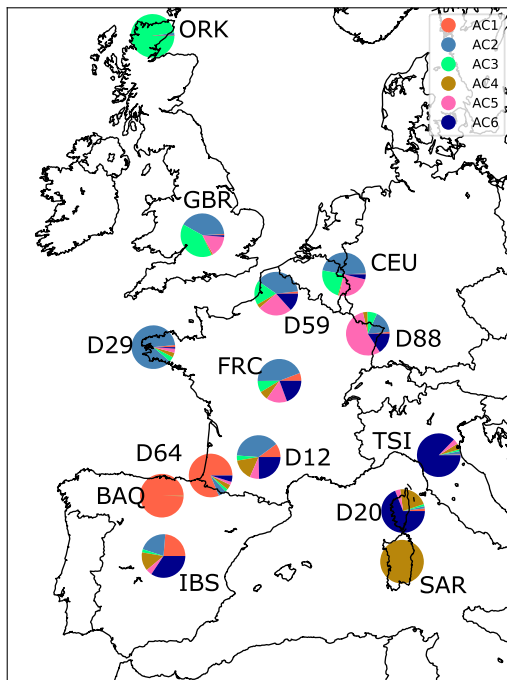
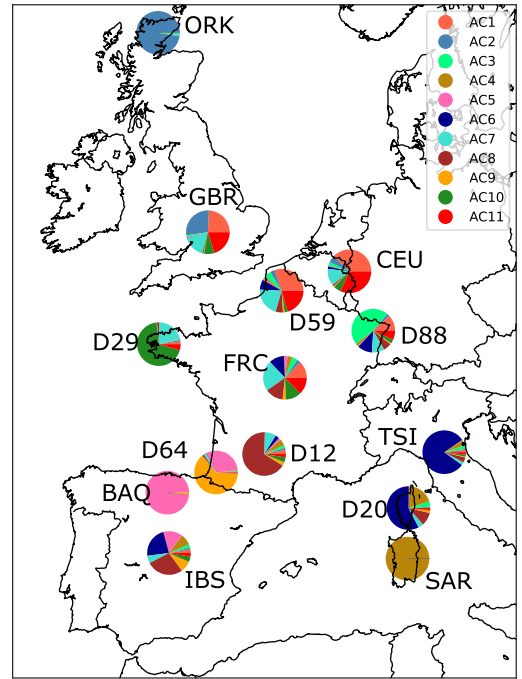
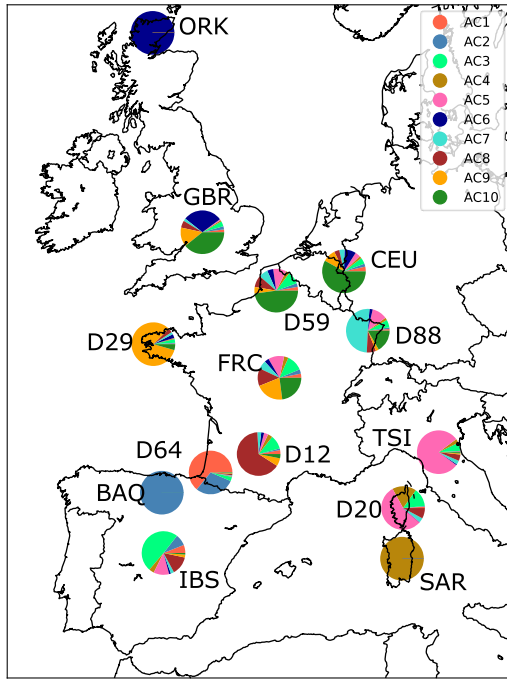


Figure S8: Unsupervised Admixture plots on the combined data of 526 individuals from 6 specific departments of POPGEN and 526 individuals from NFE_1KG_HGDP for different number of groups (K=2 and K=3 above, and K=4 to K=11 below in subsequent plots). Mean Admixture Components (ACs) for each group are 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).







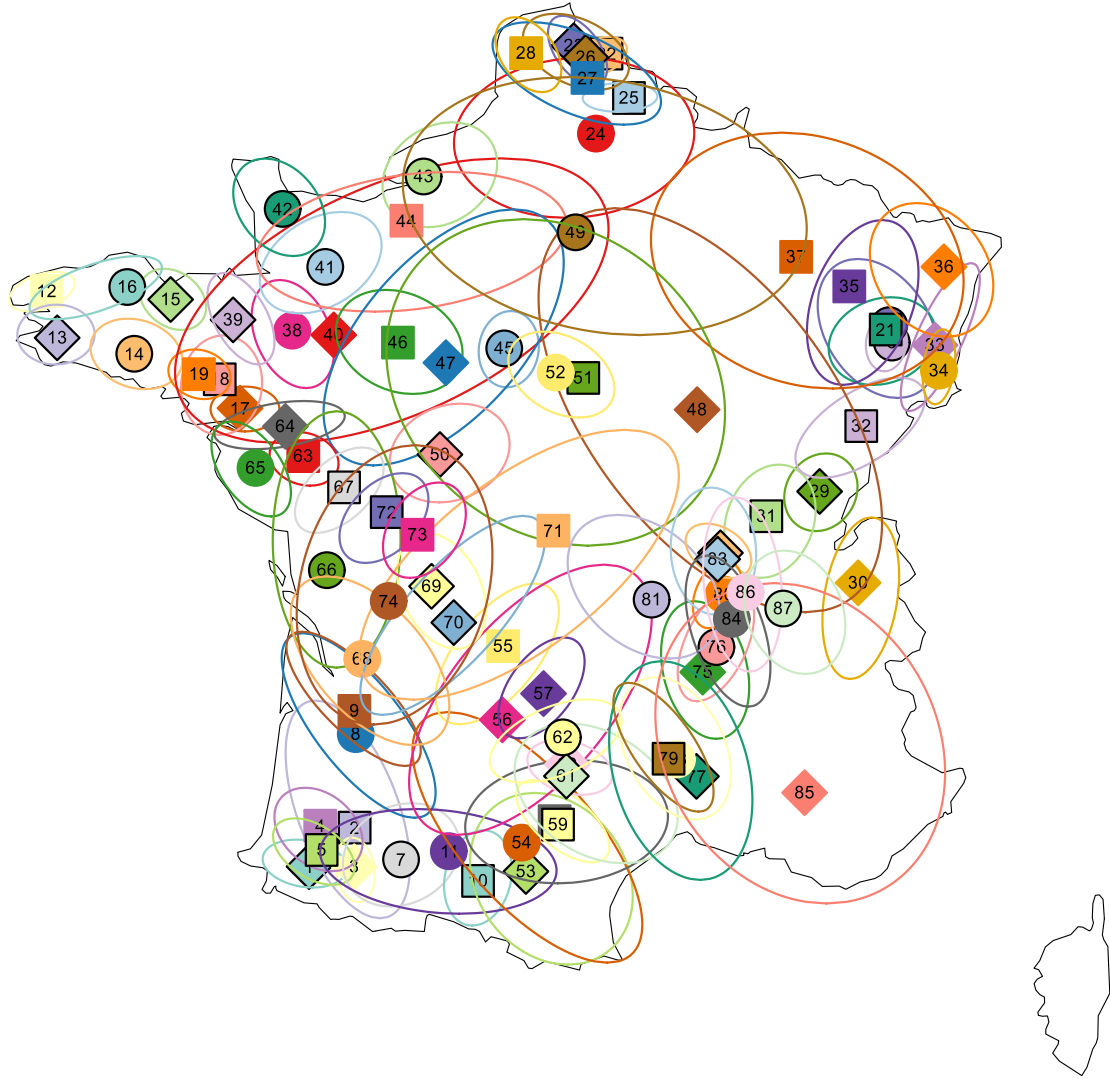


Figure S9: For each of the 87 FineSTRUCTURE clusters presented in Figure 5 of the main text, the centre-of-mass (in longitude and latitude) is plotted onto the map of France along with minimal ellipses containing 90% of the individuals in each cluster.

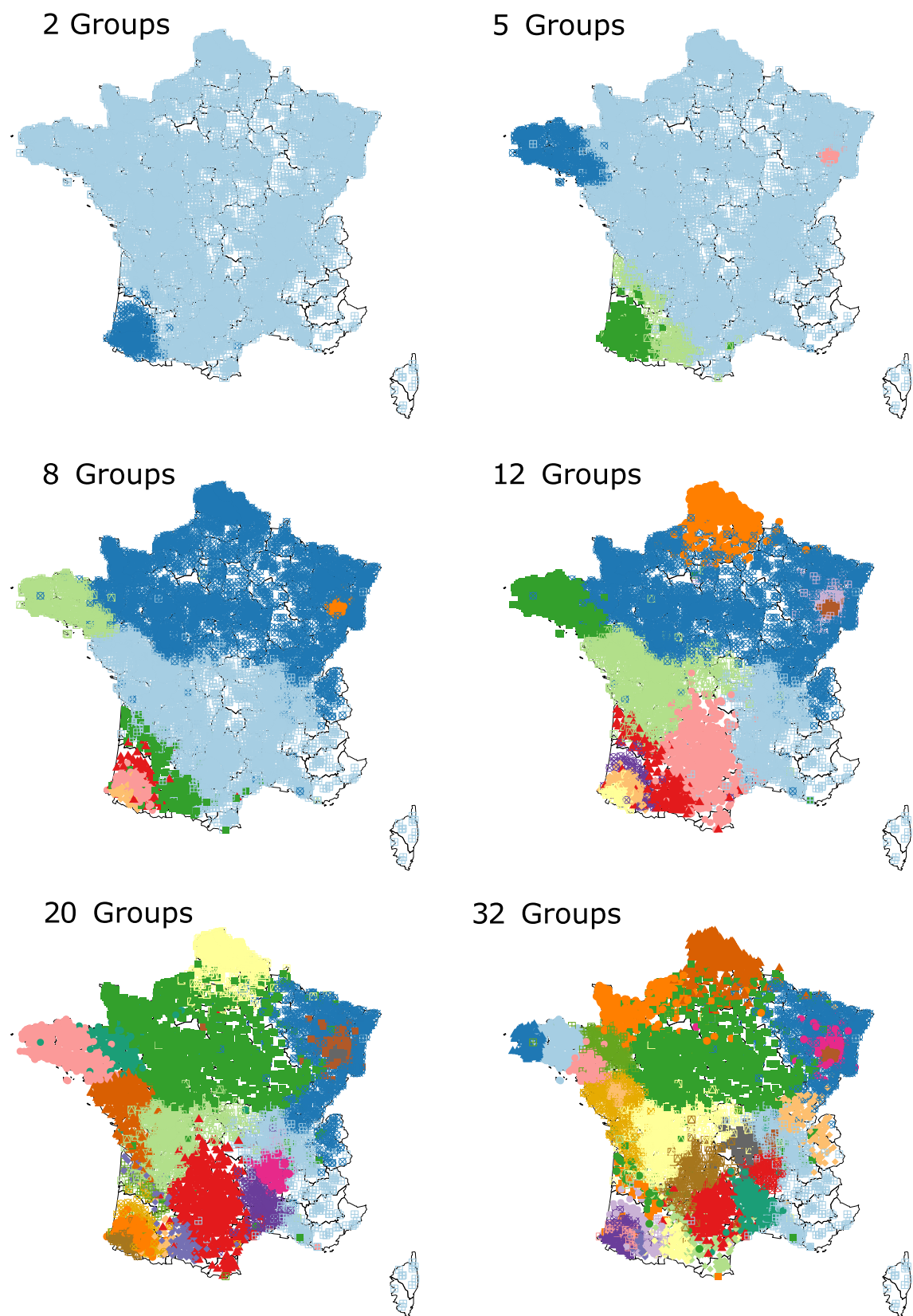


Figure S10: Cuts of the final FineSTRUCTURE TVD tree (see Figure 5d in the main text) at 2, 5, 8, 12, 20, and 32 that correspond to the analysis presented in Figure 5 in the main text.

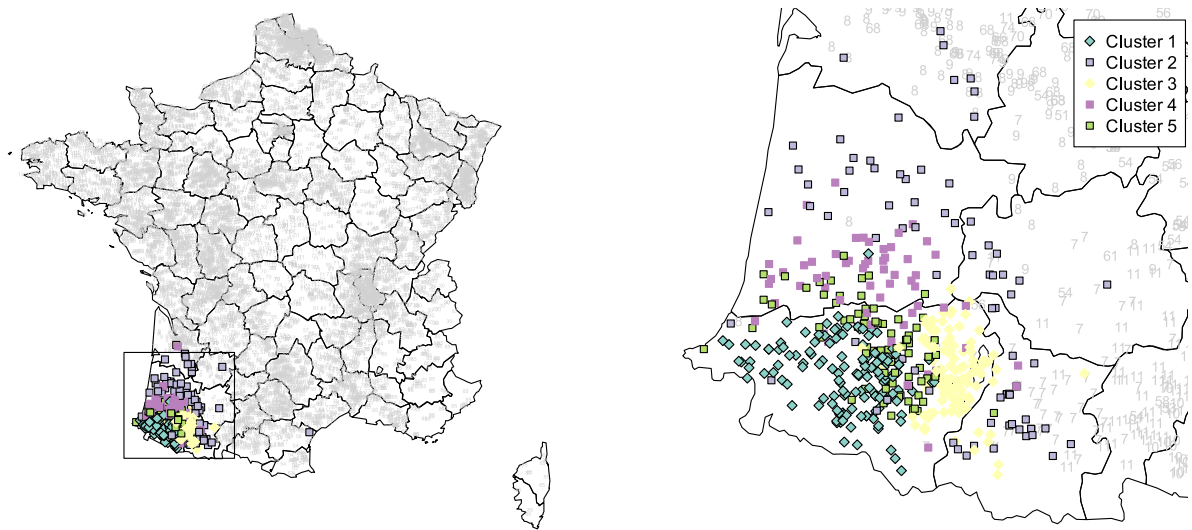
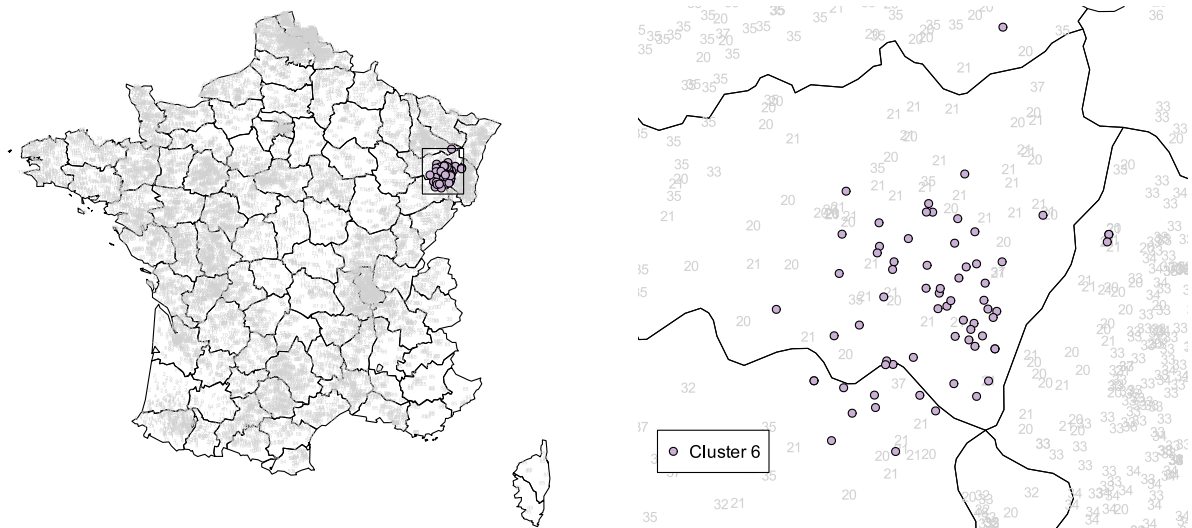
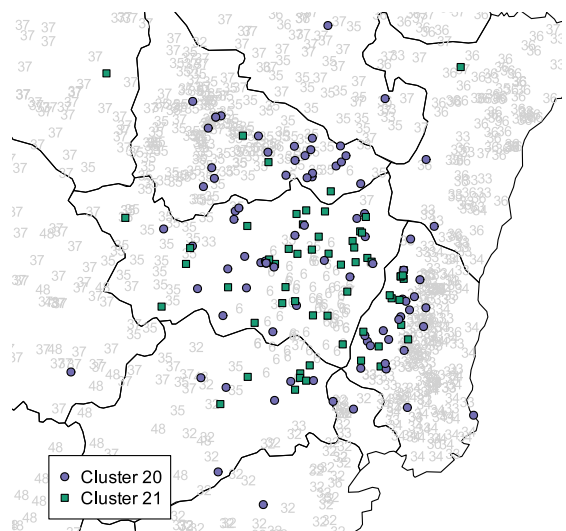
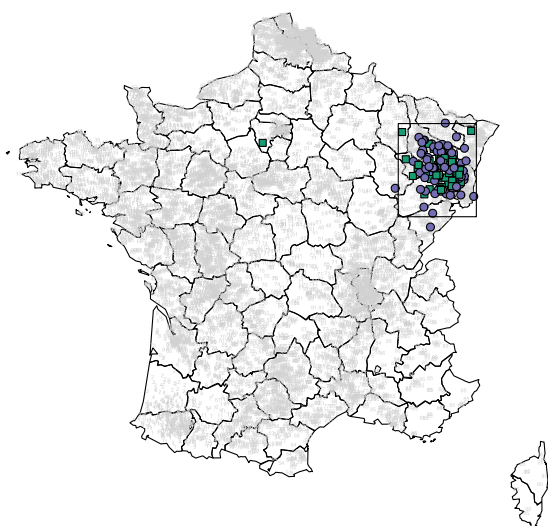
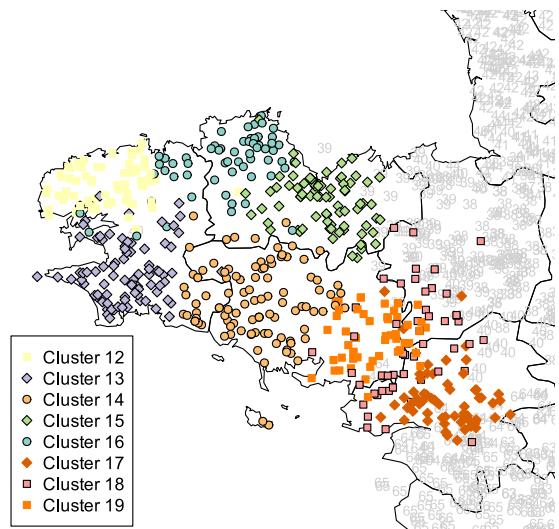
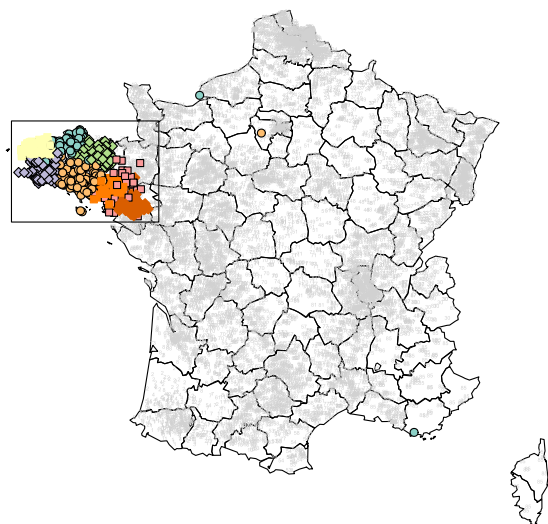
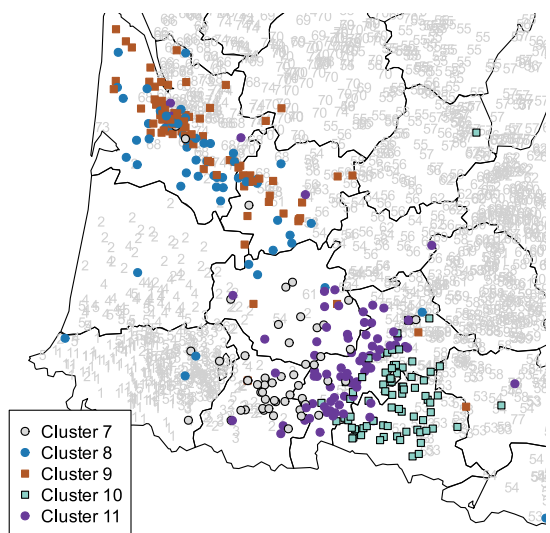
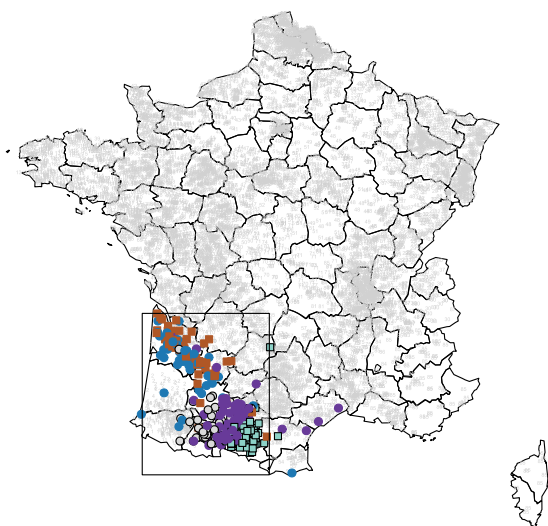
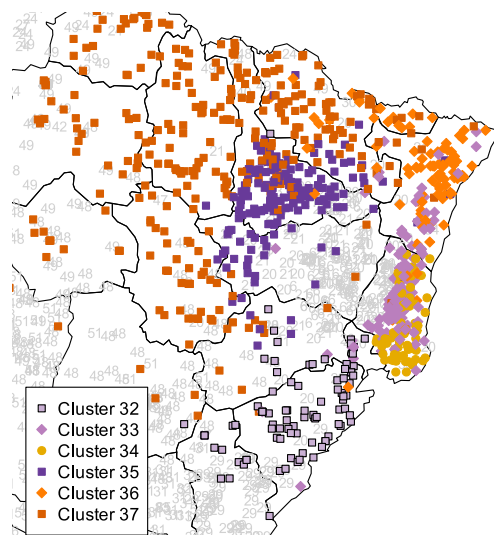
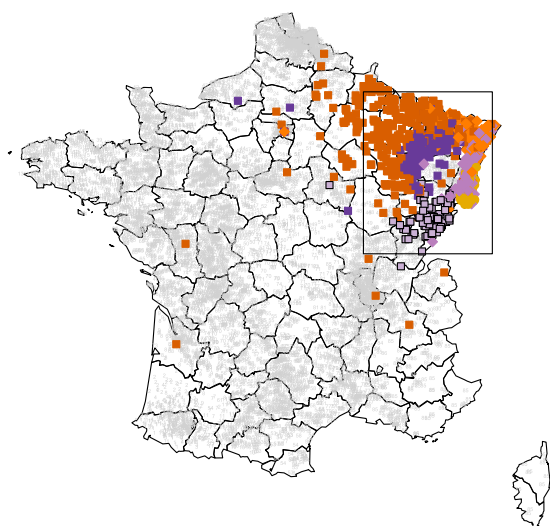
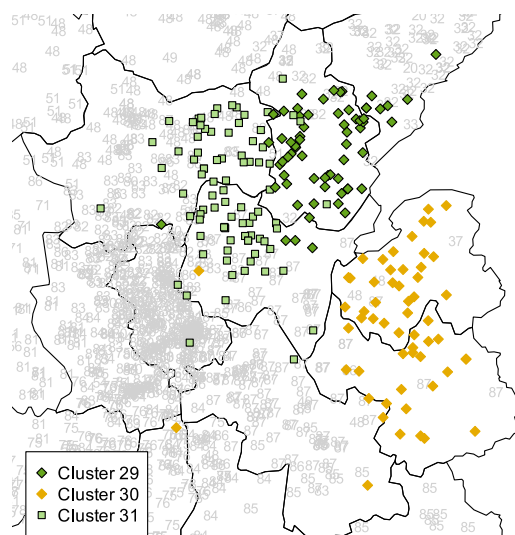
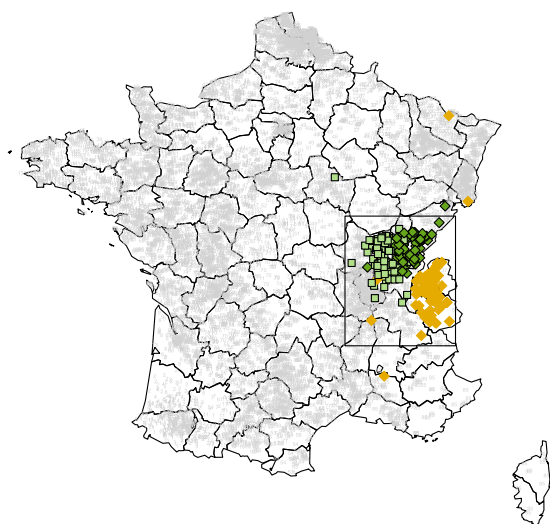
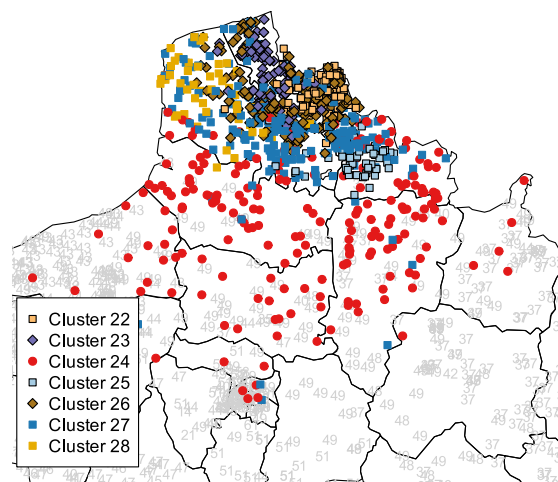
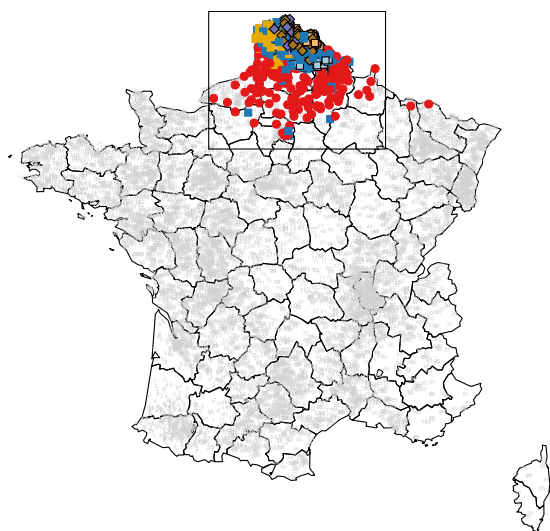
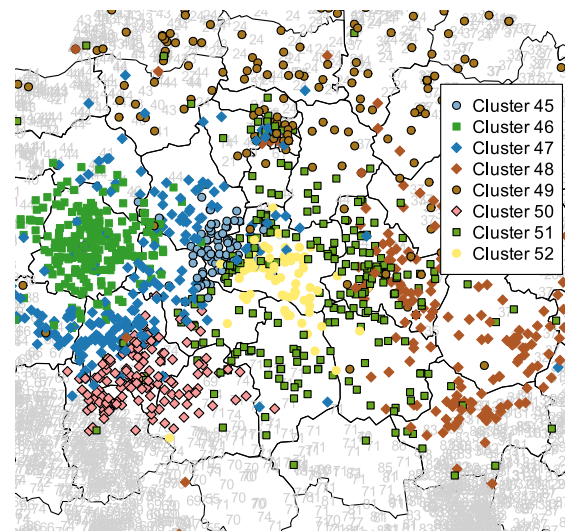
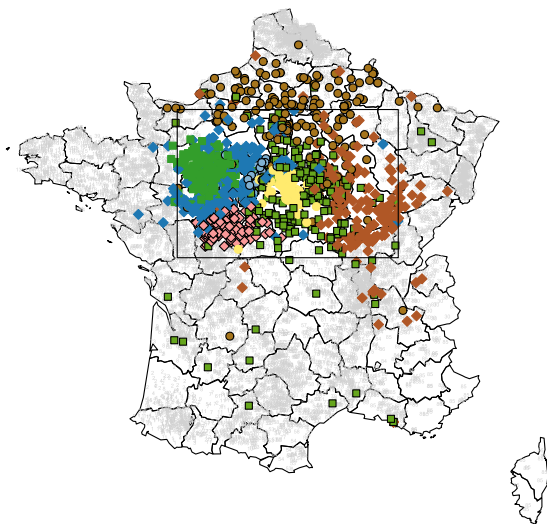
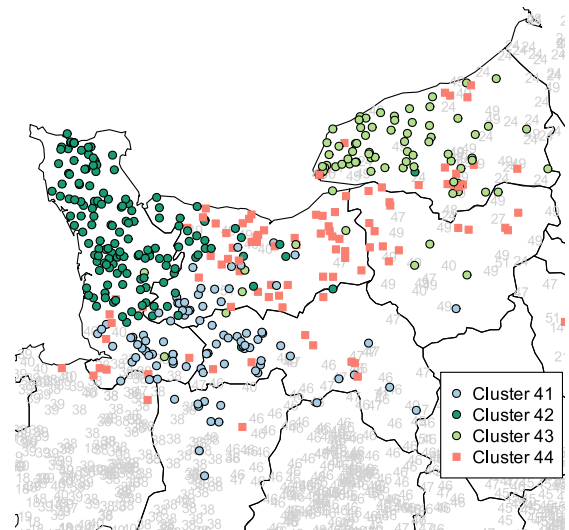
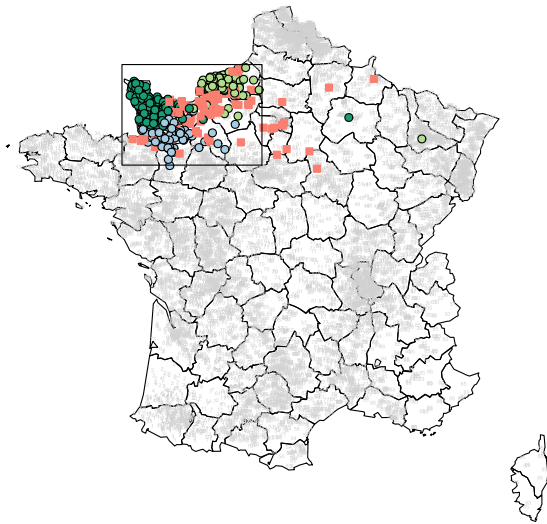
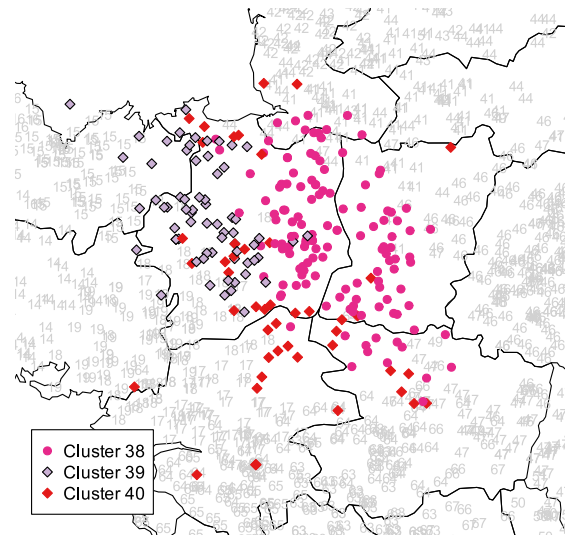
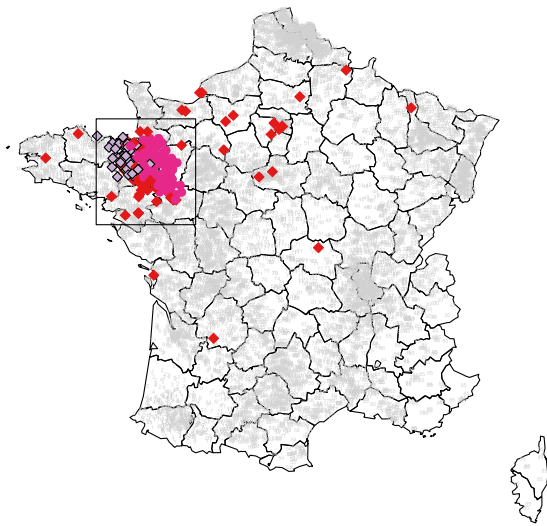


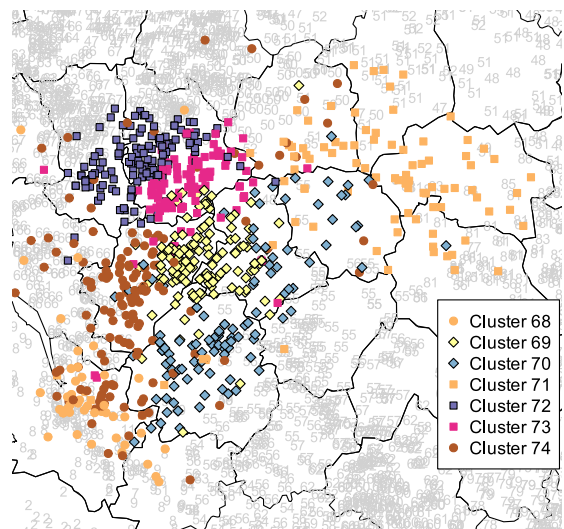
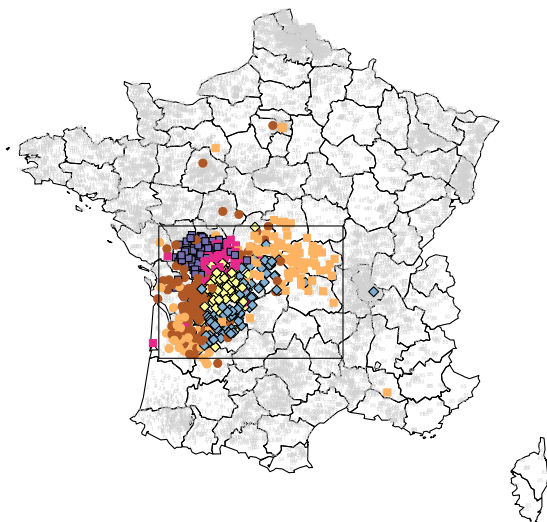
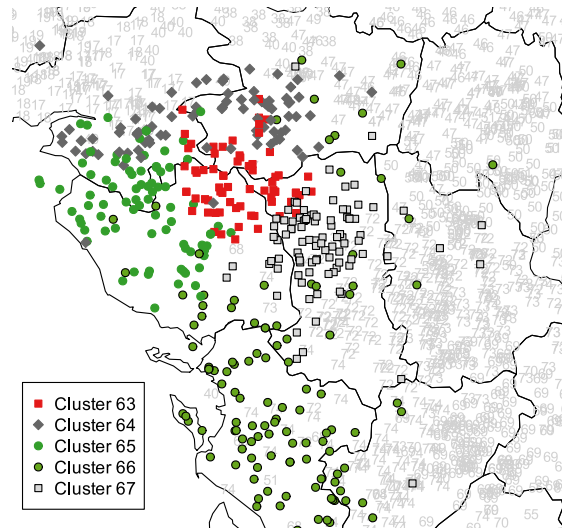
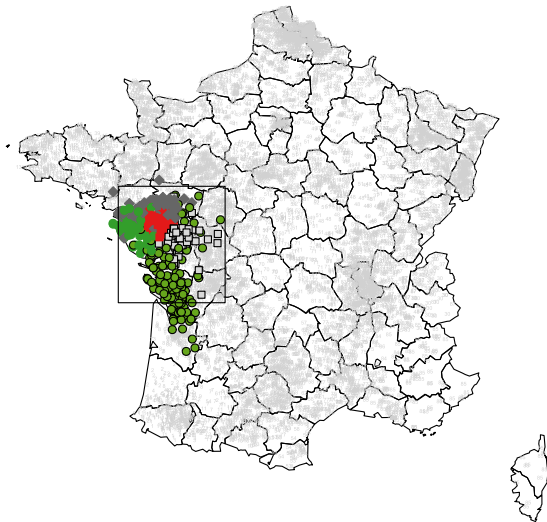
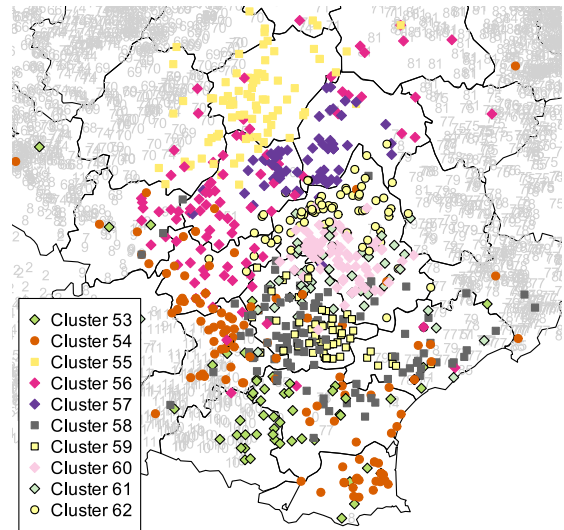
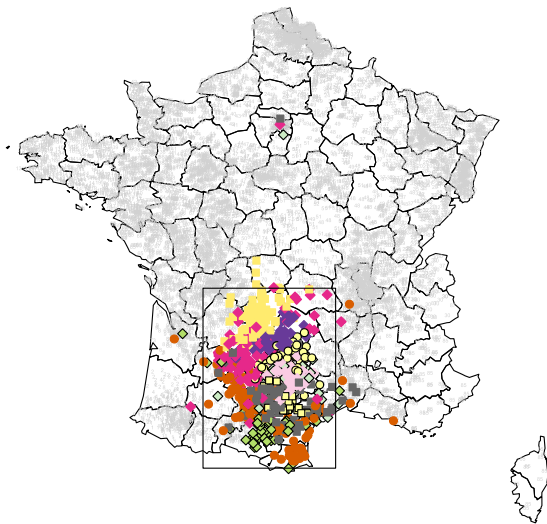
Figure S11: The final FineSTRUCTURE TVD tree (see Figure 5d in the main text) is divided into 16 different sections, and more detailed plots of the results are given in succession. In each plot the map of France with all individuals of POPGEN is presented, individuals within the section of the FineSTRUCTURE 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 5d 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 5c in the main text.

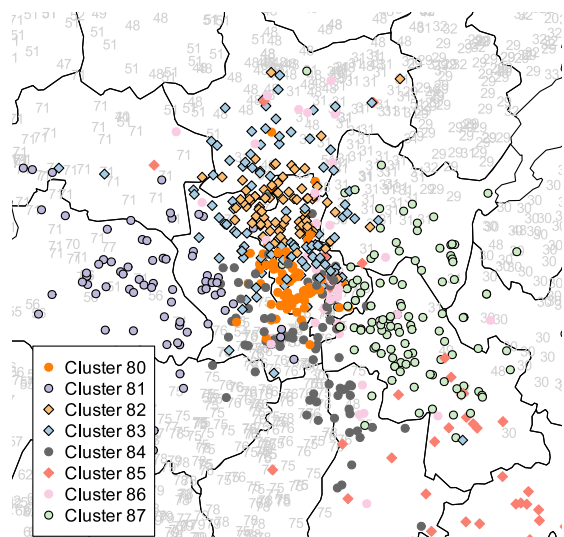
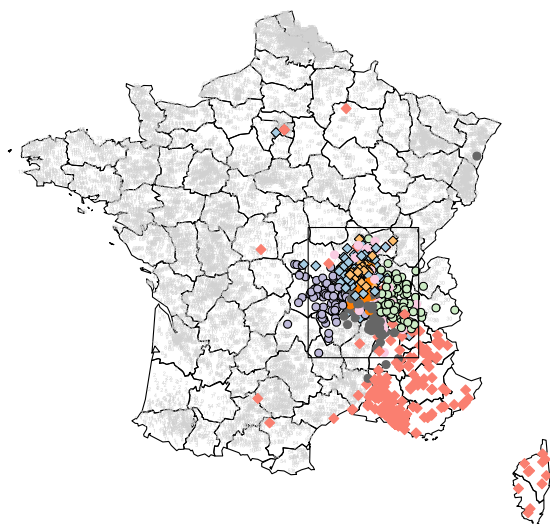
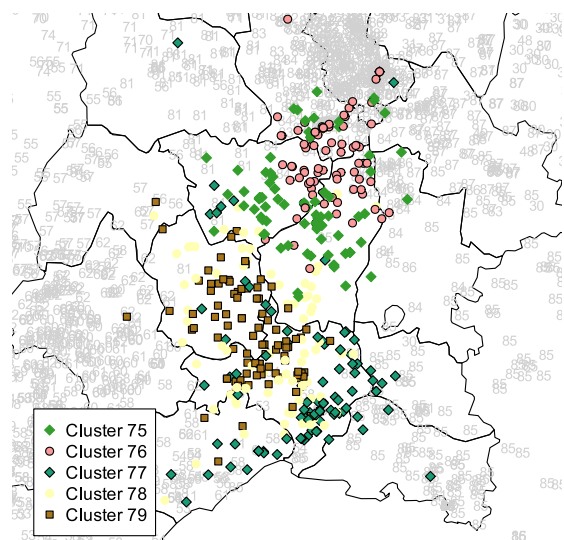
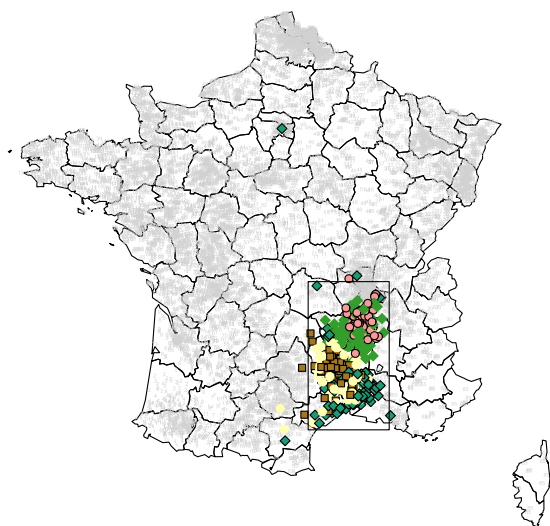












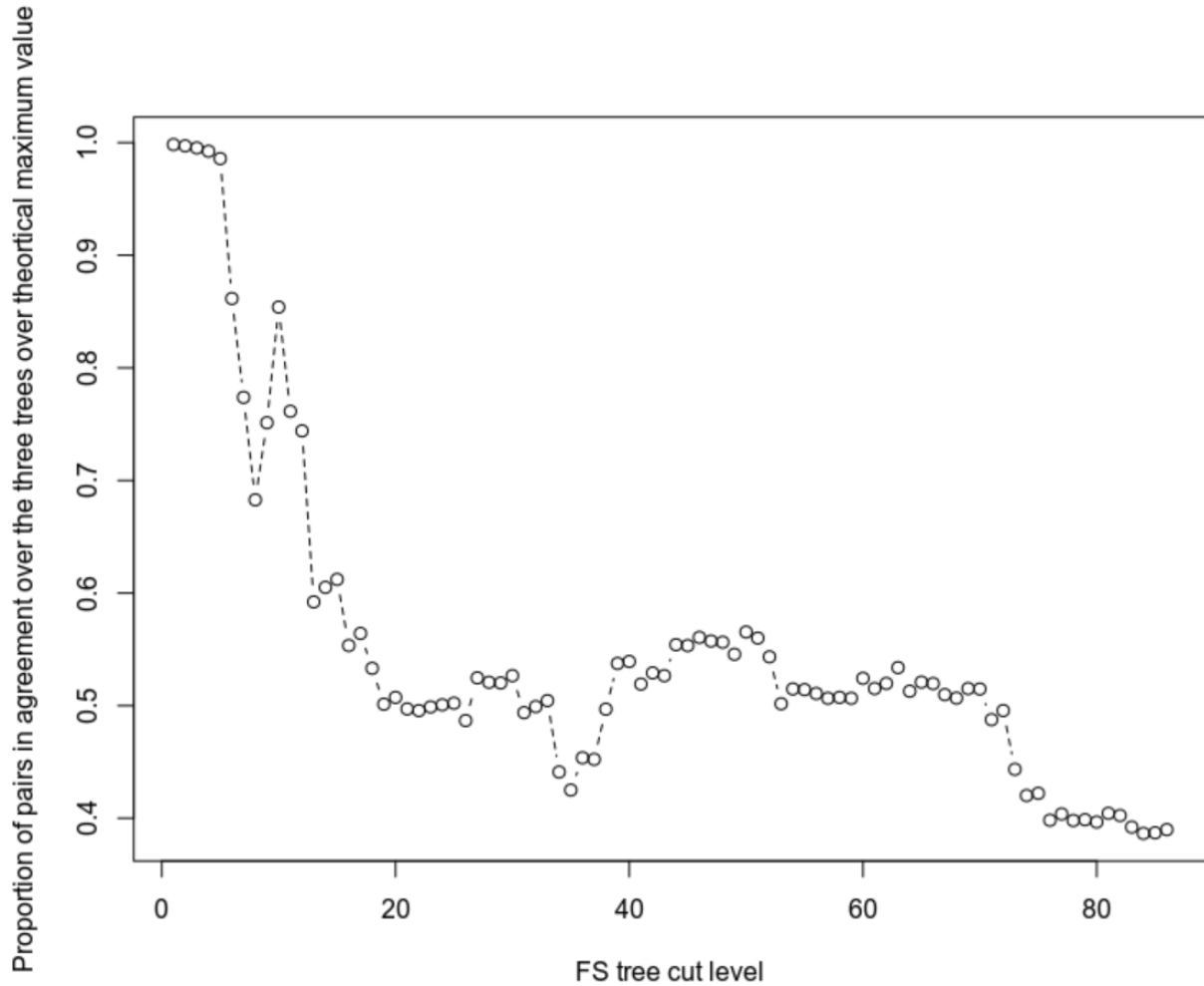


Figure S12: Consistency of FineSTRUCTURE results across three runs. At different cuts of the three trees, we calculated the number of pairs of individuals who are in the same cluster as each other across the three FineSTRUCTURE runs. We then divided this by its maximum possible value which is the minimum 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.

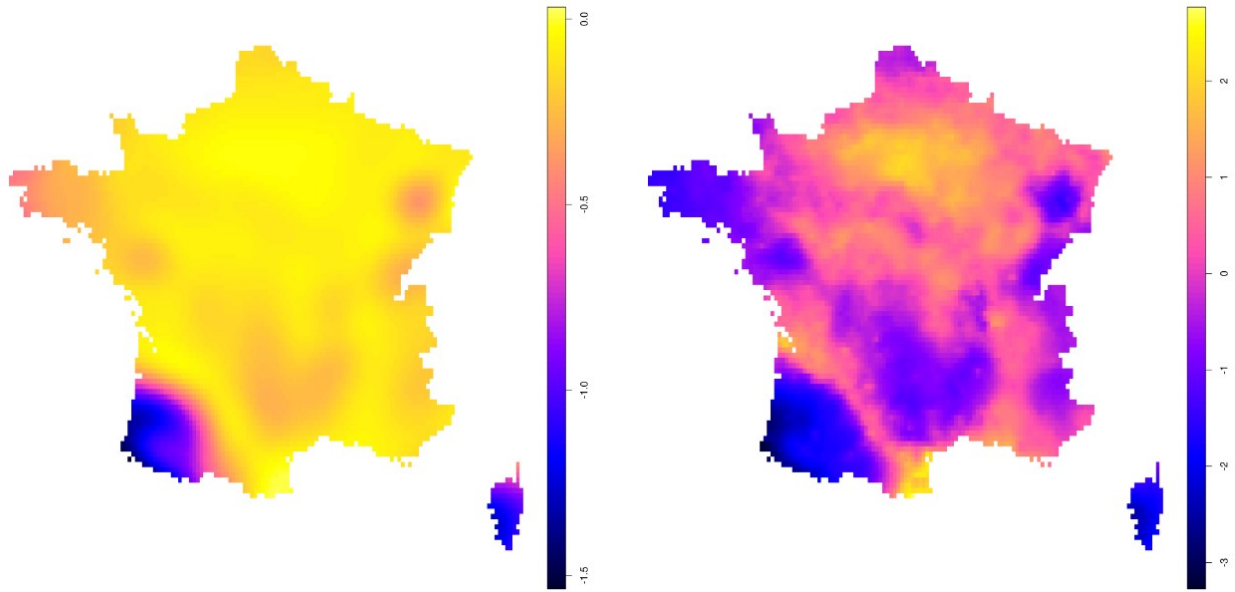


Figure S13: Universal Kriging analysis of the slope coefficient from 9598 regression models that compare physical distance with IBD sharing (both on the log scale) as described in the main text in the analysis pertaining to Table 1. On the left, the raw slope coefficients are presented, with the South 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.

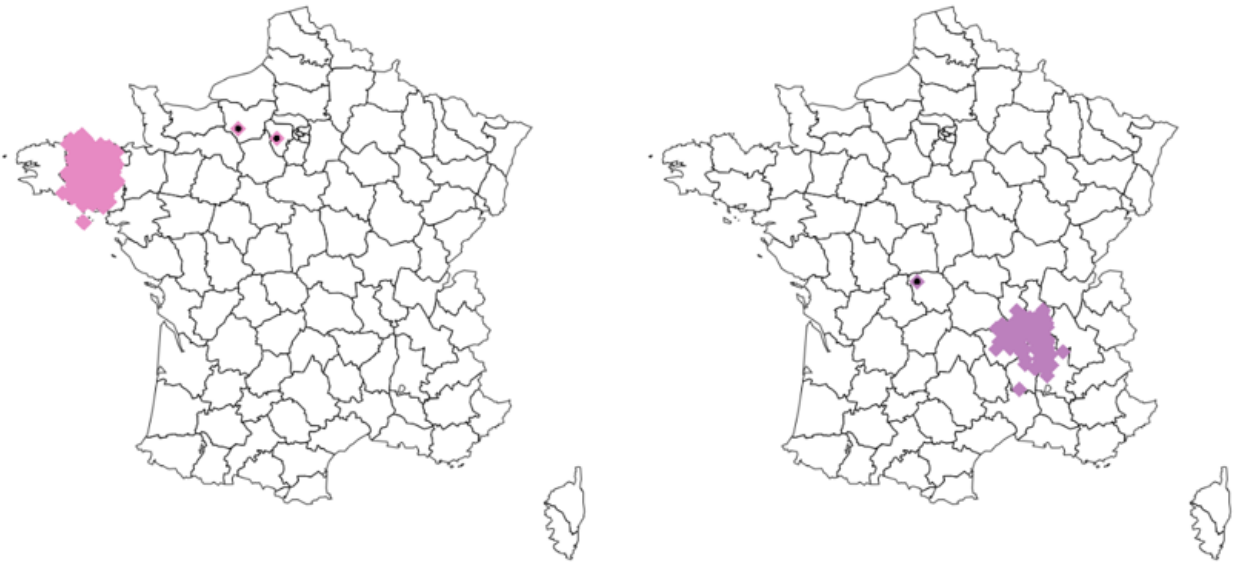


Figure S14: Example of two clusters identified by FineSTRUCTURE with 3 outliers (2 on the left panel, 1 on the right panel identified). The logic behind the exclusion was to avoid individuals whose anchorage in longitude and latitude on the map of France was less well correlated with their genetic data. The definition of the outliers is given in Supplementary Note 3.

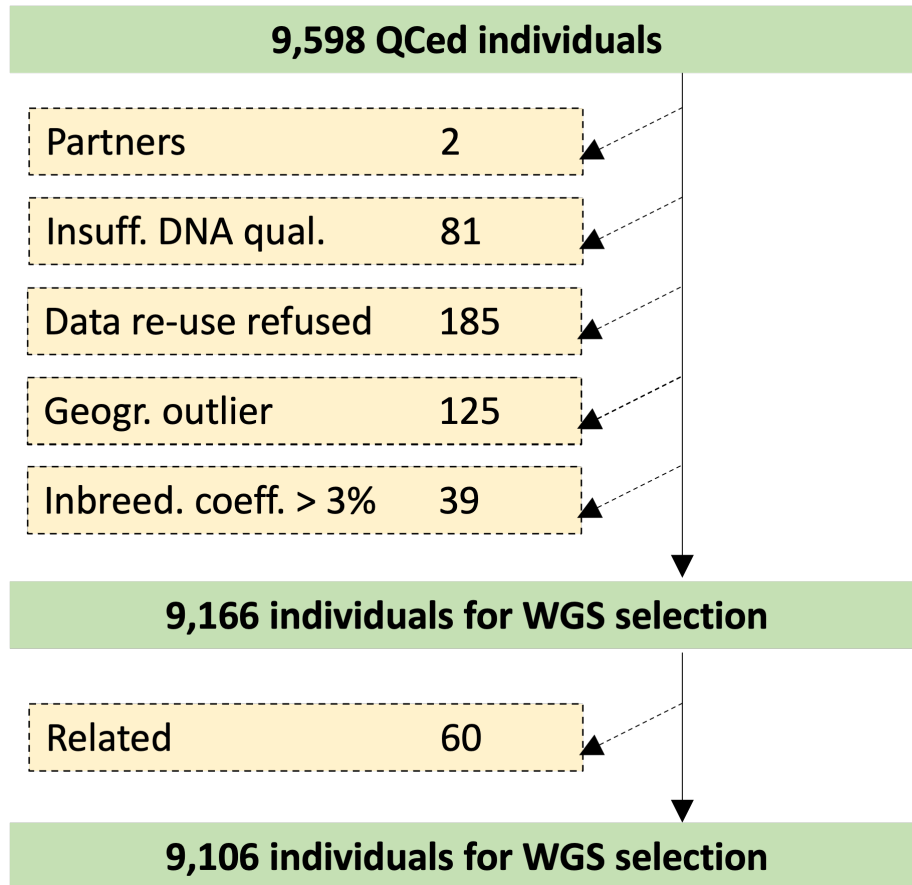


Figure S15: Flow chart detailing the exclusions that lead to a list of 9,106 individuals from which the 4,000 WGS individuals were selected.

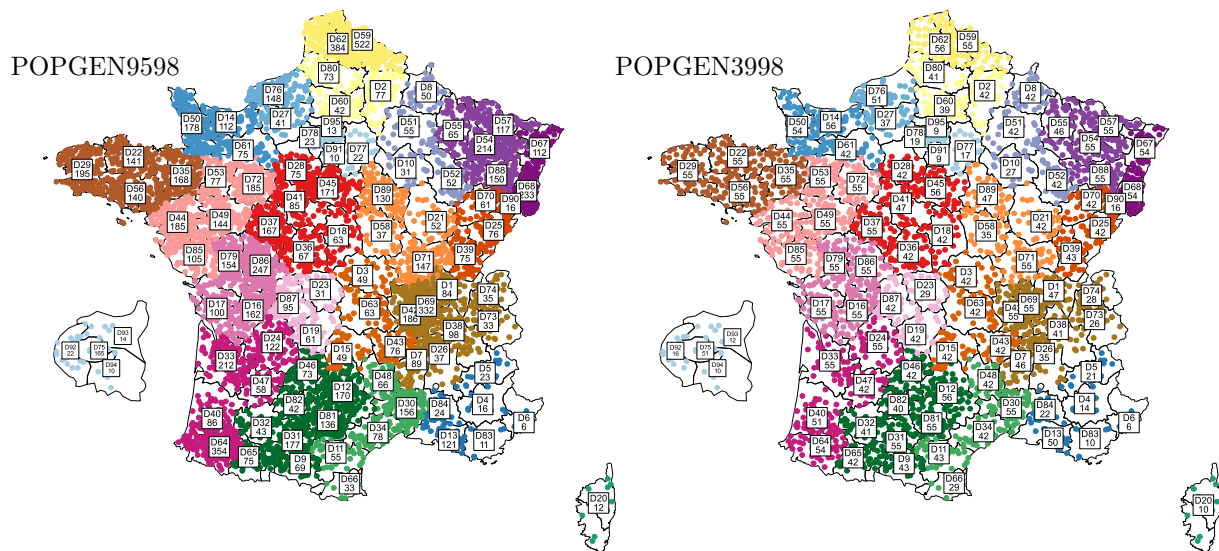


Figure S16: Assigned positions of all individuals in POPGEN9598 (left) and the subset of individuals in POPGEN3998 selected for WGS (right) with colours corresponding to administrative regions in France and department numbers (top) and total number of individuals (bottom) indicated for each department.

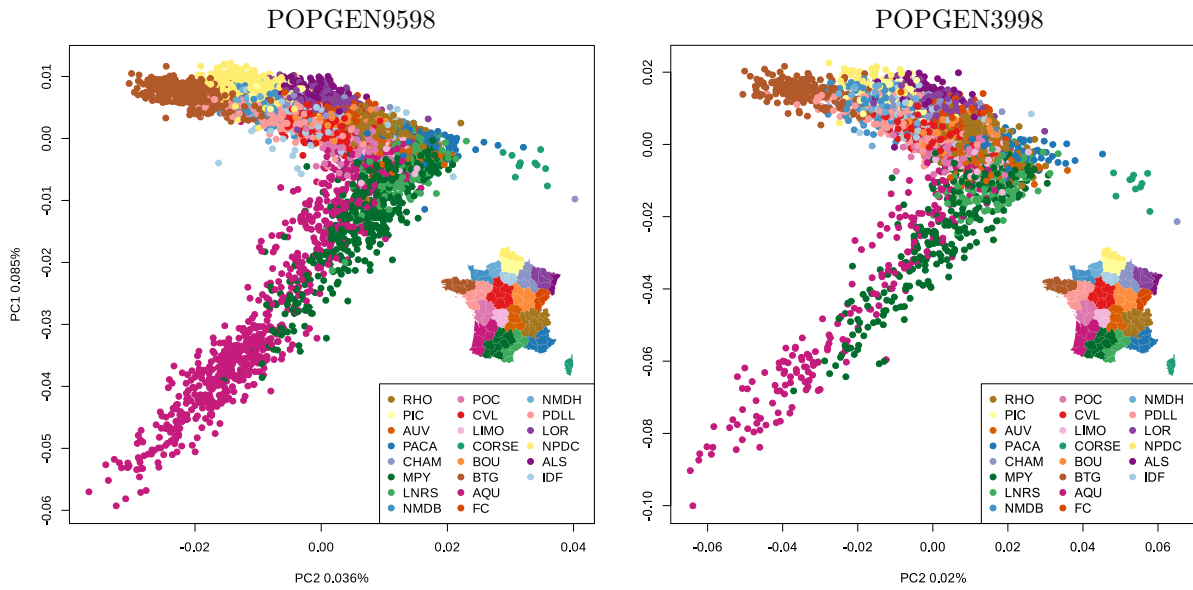


Figure S17: Principal components analysis of POPGEN9598 (left, a direct copy of Figure 2d in main text) compared to the same analysis recalculated on only the POPGEN3998 individuals using the same set of genetic variants (right). Colours correspond to administrative regions in France. The squared-correlation 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).

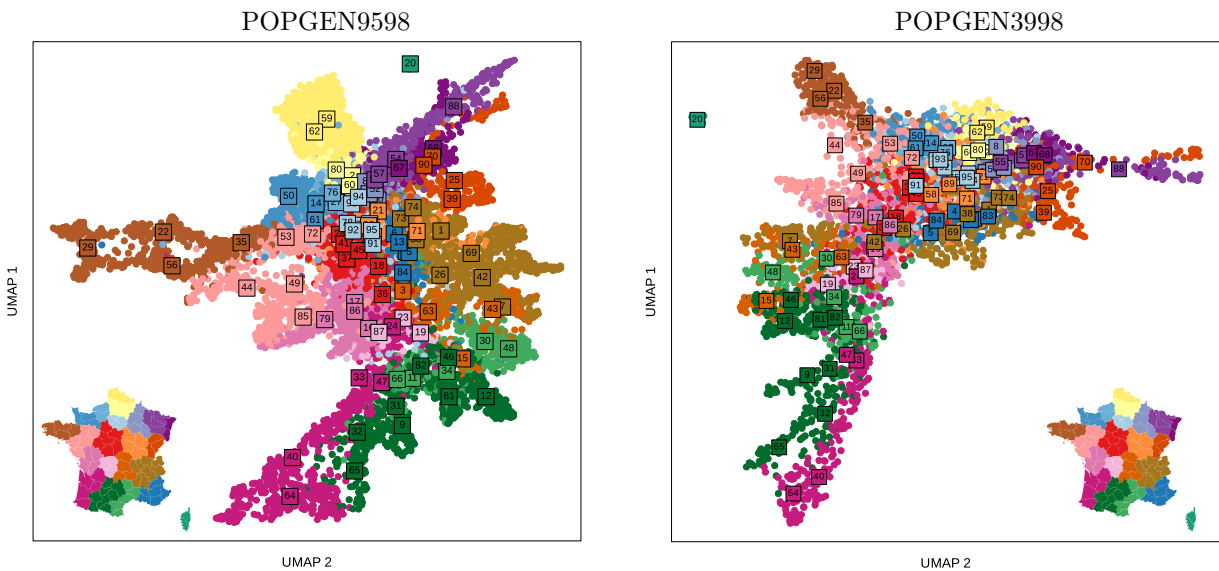


Figure S18: Visual comparison of the UMAP projection of the PBWT-paint matrix of the POPGEN9598 sample (left) and the POPGEN3998 sample (right). Both show separation of individuals at the department level.

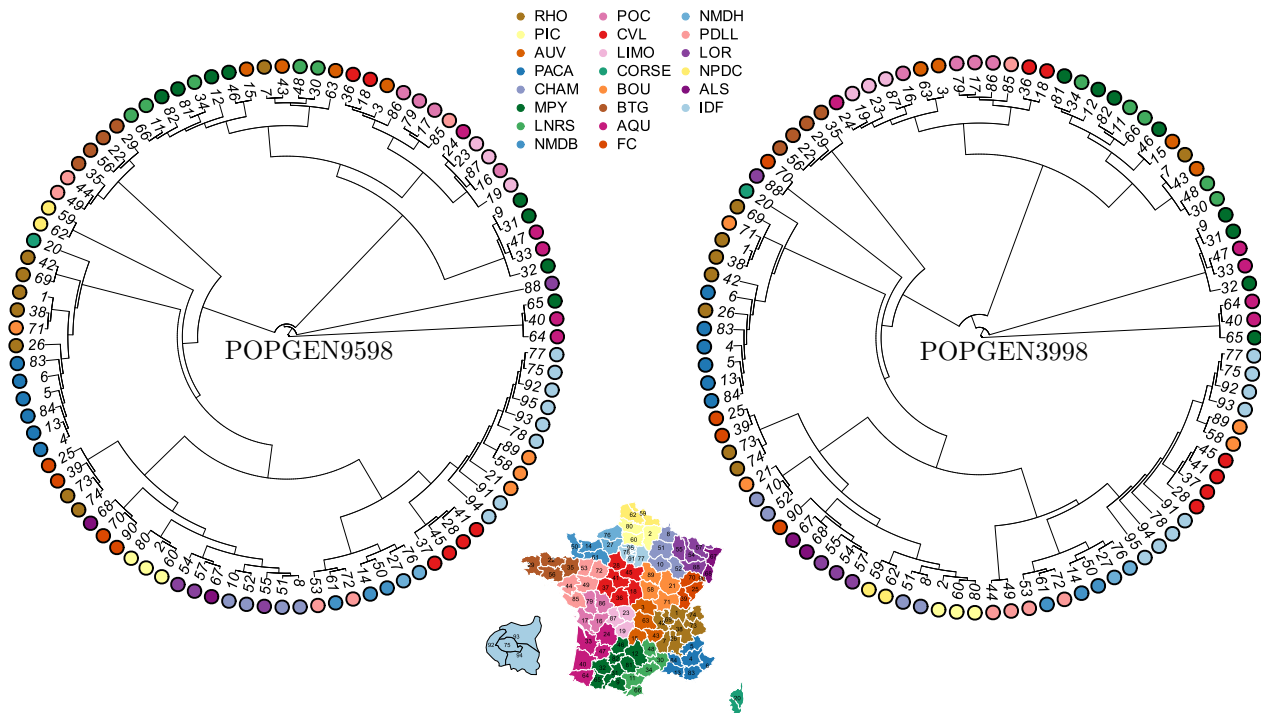


Figure S19: Comparison of the TVD tree of departments calculated using PBWT-paint similarity matrices for the POPGEN9598 sample (left) and the POPGEN3998 sample (right). The normalised Robinson-Foulds distance [6] between the two trees is 0.543 indicating that roughly half of the exact splits are shared 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.

References

1. Loo M van der, Laan J van der, R Core Team, Logan N, Muir C, Gruber J, et al. stringdist: Approximate String Matching, Fuzzy Text Search, and String Distance Functions. 2025. Available: <https://cran.r-project.org/web/packages/stringdist/index.html>
2. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. *Nature*. 2015;519: 309–314. doi:10.1038/nature14230
3. Leutenegger A-L, Prum B, Génin E, Verny C, Lemainque A, Clerget-Darpoux F, et al. Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet*. 2003;73: 516–523. doi:10.1086/378207
4. Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L. FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinforma Oxf Engl*. 2014;30: 1940–1941. doi:10.1093/bioinformatics/btu149
5. Koenig Z, Yohannes MT, Nkambule LL, Zhao X, Goodrich JK, Kim HA, et al. A harmonized public resource of deeply sequenced diverse human genomes. *BioRxiv Prepr Serv Biol*. 2024; 2023.01.23.525248. doi:10.1101/2023.01.23.525248
6. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981;53: 131–147. doi:10.1016/0025-5564(81)90043-2