

POPGEN: building a reference panel of local ancestry to study genetic variability across metropolitan France

Anthony Herzig

Inserm UMR1078

Gaëlle Le Folgoc

Inserm UMR1078

Gaëlle Marenne

Inserm U1078 <https://orcid.org/0000-0002-4363-7170>

Hélène Blanché

Fondation Jean Dausset - CEPH

Aude Saint Pierre

Inserm UMR1078 <https://orcid.org/0000-0002-8569-2936>

Maël Guivarch

Thomas Ludwig

<https://orcid.org/0000-0002-9996-1084>

Isabel Alves

Jean-Marc Sebaoun

Delphine Bacq-Daian

Robert Olasso

Céline Besse

Claire Dandine-Rouland

Indu Sharma

Sacha Schultz

Marc Gros La Faige

Christelle Delmas

Frédérique Nowak

Christian Dina

institut du thorax, CNRS, INSERM, Université de Nantes <https://orcid.org/0000-0002-7722-7348>

Richard Redon

L'institut du thorax <https://orcid.org/0000-0001-7751-2280>

Evelyne Heyer

MNHN

Marie Zins

Jean-François Deleuze

Commissariat à l'énergie atomique et aux énergies alternatives <https://orcid.org/0000-0002-5358-4463>

Emmanuelle Génin

emmanuelle.genin@inserm.fr

Inserm UMR1078 <https://orcid.org/0000-0003-4117-2813>

Article

Keywords:

Posted Date: September 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10328978/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is no duality of interest

Abstract

Understanding the fine-scale genetic structure of human populations is essential for both demographical history and genomic medicine. The POPGEN project was initiated as part of the France Genomic Medicine Initiative to construct a whole-genome reference panel representative of the ancestral genetic diversity that existed across metropolitan France at the beginning of the 20th century. Building on previous studies that reported significant fine-scale population structure within France, POPGEN aimed to improve geographic coverage and generate regional allele frequency data to support medical and population genetics applications.

We recruited 15,000 participants from the largest French population-based cohort, CONSTANCES, based on the birthplace proximity of their four grandparents and sent saliva kits at their homes. The high return rate (68.3%) demonstrates feasibility of large-scale self-sampling in the general population. Genotyping data from the 9,772 samples that could be processed confirm the presence of widespread fine-scale structure across France, and extend this finding to previously underrepresented regions. These results highlight the value of recruitment strategies based on grandparents' origin for capturing subtle genetic structure. The development of sampling strategies that best capture genetic diversity at the finest scales will indeed be essential to characterize population structure across Europe and to enable more accurate analyses in medical genetics, especially when addressing rare variant interpretation and fine-scale ancestry adjustment for association tests and polygenic scores calibration. The thorough description of our recruitment strategy would provide useful insights for other initiatives like the Genome of Europe (GoE) project, which aims to generate whole-genome references encompassing European diversity.

Introduction

With the development of high throughput genomic technologies, it has become possible to explore the genetic diversity of human populations around the world based on full genome sequencing at large scale. These studies have confirmed the existence of strong relationships between genetic and geographic distances found several decades ago by the pioneer work of LL Cavalli-Sforza [1]. They have also shown that differences in allele frequencies exist at all geographic scales: between continents [2], among countries within continents (for example in Europe [3, 4]), and even between regions within countries (for example in Finland [5], Italy [6], UK [7] and very recently in Quebec [8] among others). Such observations have emphasized the need for more population genetics studies at fine geographic scales to characterise population structure at a higher resolution and address the role of genetic variation in diseases [9, 10].

Previous genetic studies on populations of French ancestry have shown that metropolitan France presents a relatively high level of genetic structure compared to other European countries. Early findings based on classical markers such as ABO blood groups [11] or HLA polymorphisms already suggested geographic differentiation, see for example the “Provinces Françaises” study [12]. More recent genome-wide analyses have refined these observations [13–17]. Using both allele frequency and haplotype-based

approaches, these studies have revealed distinct geographic clusters, particularly in Brittany and the Basque Country. Identified clusters were often delineated by major rivers or historical-cultural boundaries and the departments were suggested to be the most relevant administrative unit to capture this stratification. Complementary analyses based on surnames distributions over the 19th and 20th centuries have provided further evidence of persistent spatial structuring in France, with strong parallels between patronymic and genetic distances [18]. Explanations of this high level of genetic stratification can be found in the geography. The geographical positioning of France indeed makes it a pivotal crossroads within Europe in between the North and the South and at the extreme West, leading to diverse populations to settle over the centuries. Its various landscapes with mountains and rivers are natural barriers that have limited population mixture and gene flow. However, previous genetic studies were limited by the type of variations that were studied and/or sample sizes and the geographical coverage. With the development of genomic medicine in the country [19], there was a critical need for a more complete assessment of the genetic structure and diversity of the metropolitan French population to better understand the effect of genetic variants and improve diagnosis, treatment and prevention for both rare and common diseases.

The POPGEN project was launched as part of the French Genomic Medicine Initiative [19, 20] with the primary objective of developing a reference panel of whole genome sequences of individuals originating from France, starting with metropolitan France. The creation of the panel aims to provide information on the frequency of genetic variants at fine geographic scales. Such frequencies are needed to filter out neutral variants from patient genomes and help focus on variants that are absent or very rare in the general population. Given the nature of this objective, we applied a recruitment strategy that would enable individuals to be anchored geographically. To achieve the anchoring, we selected participants for whom all four grandparents were born close to one another. This strategy has previously proved successful in capturing fine-scale population structure [7, 21–23], notably in western France as shown in the FranceGenRef project [17, 24] where 856 individuals sampled from existing biobanks were whole genome (30X short-read) sequenced.

In this work, we present the detailed recruitment design of the POPGEN panel within the CONSTANCES cohort [25], the largest French general population epidemiological study. We explain the different steps of the selection process and the choices made to capture fine-scale genomic variability.

Materials and Methods

Recruitment

POPGEN participants were selected from the CONSTANCES cohort [25] using parental and grandparental birthplace data collected in the 2019 annual questionnaire (Supplementary Figures S1 and S2; Supplementary Note 1). Commune names were geocoded using in-house scripts and public geographic databases. Among 118,042 respondents, 84,972 remained eligible after excluding individuals who withdrew consent, did not provide the required information, or reported grandparents born outside

metropolitan France. Only individuals born before 1985 were considered, leading to 26,274 eligible candidates (Supplementary Figure S3; Supplementary Tables S2 and S3).

From these eligible candidates, 15,000 individuals were selected using department-specific targets based on the 1901 census, with substitutions from neighbouring departments where necessary. Departments were chosen as sampling units because they capture historical and fine-scale population structure in France [14, 16]. The final geographic distribution is shown in Supplementary Figure S1.

Overall, 10,250 individuals returned saliva kits (68%). DNA was extracted at CEPH using magnetic beads, quantified fluorometrically, and quality-controlled in a 10% subset. Genotyping was performed at CNRGH using the Global Screening Array 24 v3.0 Multi-Disease panel containing 730,059 probes. After sample exclusions, genotype data were available for 9,772 individuals (4,582 males and 5,190 females). Each participant was assigned geographic coordinates corresponding to the mean birthplace coordinates of their four grandparents. Maps were generated using the R package maps [26].

Quality Control of genetic data

Genetic data quality control was performed using PLINK v1.9 [27], as summarized in Supplementary Figure S4 and detailed in Supplementary Note 2. Preliminary variant QC was used to select high-quality autosomal SNPs for individual-level QC and relatedness analyses. Participants were excluded based on genotype call rate, excess heterozygosity, potential DNA contamination, non-European genetic ancestry, sex discordance, missing birthplace information and relatedness ($\hat{\pi} > 0.125$). This resulted in a final dataset of 9,598 QC-passed unrelated individuals, hereafter referred to as POPGEN9598.

Final variant QC excluded non-autosomal, multiallelic and poorly genotyped variants, variants deviating from Hardy–Weinberg equilibrium, and discordant duplicate probes. The final dataset comprised 654,525 variants, including 646,382 SNPs and 8,143 insertions/deletions. Among these, 513,830 variants had a minor allele frequency of at least 0.001.

Description of genetic diversity in POPGEN data

Projecting POPGEN onto a large European reference panel

The POPGEN9598 individual genotype patterns were projected onto previously constructed principal components (PCs) from a collection of published reference data, which had previously been used to show the position of French populations within the European landscape [17]. Briefly, reference individuals were gathered from the following previous studies: People of the British Isles (PoBi) [21], Population Reference Sample (POPRES) [28], 1000 Genomes Project (TGP) [29], and the Human Genetic Diversity Panel (HGDP) [30]. PCA was performed using 73,646 common SNPs (MAF > 5%) shared between the reference and POPGEN9598 genotype data using the Eigenstrat method [31] implemented

in the R-package smartPCA. For comparison purposes, we also included SNP-matched individual genotype data from the FranceGenRef study [17].

POPGEN fine-scale population structure using genotype data

The fine-scale population structure within POPGEN9598 was described using several methods based on genotype data: PCA (performed using PLINK 1.9)[27], non-supervised admixture analyses using the ADMIXTURE software [32], and analyses of molecular variance (AMOVA) [33] using the R-package ‘popr’ [34] with different group levels (administrative regions, administrative departments, and FineSTRUCTURE (FS) [35] clusters as defined below). These analyses were performed on a pruned subset of 196,144 QCed variants: the set of 513,830 QCed variants with a frequency equal or greater than 0.001 was further filtered with a stringent missing rate of 0.001, long-range LD regions were excluded [36], and pruning was performed with PLINK v1.9 [27] (argument `--indep-pairwise 50 5 0.2`), leading to a set of 196,144 pruned variants.

Diversity in POPGEN compared to a wider European context

To contextualise genetic variation in POPGEN, we compared a subset of 526 participants from six French departments with 526 unrelated non-Finnish European individuals from TGP and HGDP datasets reported by Koenig et al. [37]. From the TGP, we included TSI (n = 103), CEU (n = 121), GBR (n = 88), IBS (n = 104) and from HGDP, Basque (n = 23), Sardinia (n = 27), Orkney (n = 14), France (n = 27), Bergamo (n = 11), Tuscan (n = 8). We refer to these samples as the non-Finnish Europeans of TGP and HGDP (NFE_TGP_HGDP). As there are only a few individuals from the Bergamo and Tuscan groups, they were grouped together with the larger Italian group which we still refer to as TSI (n = 122). For POPGEN, individuals were chosen in 6 departments: 11 from department 20 (Corsica), and five groups of 103 randomly chosen individuals from departments 29 (Finistère), 59 (Nord), 64 (Pyrenées-Atlantiques), 88 (Vosges), and 12 (Aveyron). The down-sampled data from these six departments has a similar structure to NFE_TGP_HGDP as it includes several large groups of around 100 individuals (similar in size to groups such as TSI, IBS, and GBR) and the small group of Corsica (similar to groups such as Sardinia and Orkney).

POPGEN genotypes were lifted over to genomic build 38 using the liftOver tool [38] and merged with NFE_TGP_HGDP genotypes. Only the 471,148 sites that were polymorphic in the NFE_TGP_HGDP samples and that passed the variant-QC in POPGEN were retained. Then, we filtered out 430 variants with high allele frequency differences between the POPGEN and NFE_TGP_HGDP datasets (Supplementary Figure S5), and 3,876 variants with a MAF below 0.1% in either datasets. This gave a final dataset of 1,052 individuals and 466,842 variants. To perform PCA and unsupervised admixture on this combined dataset, the same approach as described above for POPGEN9598 was used. We also estimated Fst statistics [39, 40].

POPGEN fine-scale population structure using haplotype data

Phasing was carried out on POPGEN9598 and the subset of 513,629 QCed SNPs (INDELs were excluded) with a frequency equal or greater than 0.001. A consensus phasing approach was used as in [41]: SHAPEIT2 [42], BEAGLE5 [43], and EAGLE2 [44] were run three times before applying the conshap algorithm [41]. For EAGLE2, variation between runs was achieved by shuffling the order of individuals in the input files. We used genetic recombination maps averaged between males and females from European populations [45] and obtained phasing for 513,617 SNPs.

PBWT-paint [46, 47] was applied separately to each phased chromosome, and results were combined using chromocombine [35]. The resulting matrix was adjusted to the nearest positive semi-definite matrix using nearPD from the R package Matrix [48], centred, and analysed by kernel PCA with kernlab [49].

FineSTRUCTURE (FS) was then applied to the PBWT-paint matrix using 100,000 burn-in and 200,000 MCMC iterations, with all individuals initially assigned to a single group. The highest-scoring partition was retained. To improve interpretability, clusters containing fewer than 50 individuals were merged using Total Variation Distance, which was also used to construct a hierarchical TVD tree. This procedure identified 87 clusters among the 9,598 individuals. Additional FS runs were used to identify outliers for exclusion from the whole-genome sequencing selection process (Supplementary Note 3).

Pairwise identity-by-descent sharing was estimated using HAP-IBD [50]. Only segments longer than 2.5 cM and containing at least 250 variants were retained. The logarithm of the total IBD length shared by each pair was regressed on geographic distance and on indicators of shared membership in historical regions, departments and FS clusters.

Results

Global Population Structure Analyses

The principal components analysis (PCA) in Fig. 1a-c shows the individuals of POPGEN projected onto axes of genetic variation in European populations. For a comparison, we also added the projections from the 856 French individuals from the FranceGenRef study described in Alves et al., 2024 [17]. This simple projection shows that the POPGEN participants find their place between Northern and Southern European populations within the genetic landscape of Europe. Indeed, POPGEN covers a wider area in the PCA space compared to the previous results on the FranceGenRef project. This indicates that WGS data arising from POPGEN should help to provide an even more complete catalogue of genetic variation.

Diversity within France as compared to Europe

The existence of stratification within France is visible on the PCA performed on POPGEN9598 only (Fig. 1d for PC1 and PC2 and Supplementary Figure S6 for additional PCs). Several geographical areas

stand out: the Basque region in the South West on PC1, Brittany in the North West on PC2 and PC3, Corsica in the South East on PC2, and individuals from Alsace, Lorraine and Franche-Comté in the North East on PC4.

This internal structure within POPGEN was also apparent running unsupervised ADMIXTURE across all 9,598 individuals as shown for different values of the number of clusters K (Figure S7). The admixture components reflect similar patterns as the PCA plots regarding the Basque region, Brittany, Corsica, and the Vosges. For $K = 6$, we also see components that concentrate in the very North of France, and more widely in the South.

Having observed an important level of diversity within the French populations of POPGEN, we then aimed to contextualize the diversity within the landscape of Europe by focusing on a subset of individuals from POPGEN from six departments (22, 88, 64, 59, 20, and 12) that particularly stand out in the previous results and compared them to the non-Finnish European individuals from the NFE_TGP_HGDP panel. When looking at the F_{st} values, we found that two departments (64 and 20) have high F_{st} values with all other departments and four departments (29, 88, 12, 59) form a group with small F_{st} values between them (Fig. 2a). This overall pattern is reflected in the F_{st} values in the NFE_TGP_HGDP data, where three populations (Orkney, Sardinia, Basque), all known to be European populations with a certain degree of isolation, also have high F_{st} values with all other groups. Notably, the highest F_{st} value in the entire matrix is between Sardinia and Orkney, which are at greatest geographical separation in the dataset. The other groups of NFE_TGP_HGDP have far smaller F_{st} values between each other, often comparable to the F_{st} values between the departments of POPGEN. These results indicate that the diversity within France is similar in scale to the wider diversity in Western Europe. For example, the F_{st} between departments 29 and 88 (0.00134) is in the same range and even slightly greater than between TSI and IBS (0.00110); and the F_{st} between departments 12 and 64 (0.00284) is greater than between GBR and IBS (0.00190). In Fig. 2b, average admixture components (ACs) for an unsupervised analysis with $K = 4$ show good agreement with geography. Interestingly, the French individuals from HGDP (FRC) show different admixture composition than within POPGEN, especially for departments 29, 64, 88 and 20. Similar trends are also seen when considering different values of K (see Supplementary Materials and Figures S9).

Leveraging Haplotype Similarity for Fine Scale Stratification Detection within France

To further investigate fine-scale stratification within France, we move beyond single-marker analyses to genomic segment-based approaches, leveraging haplotype similarity measures. Spectral decomposition of the PBWT-paint similarity matrices (Fig. 3a), along with their UMAP projection (Fig. 3b), reveals a detailed population structure, with clear differentiation observed down to the departmental level. We observe a strong concordance between geographic and genetic proximity, as departments within the same region (and thus sharing the same colour) tend to cluster together in the hierarchical trees (Fig. 4a–b). Careful examination of the trees however reveals some interesting exceptions. First, we find

that some departments such as 71 (Saône-et-Loire in the region Bourgogne-Franche-Comté) and 07 (Ardèche in Auvergne-Rhône-Alpes) appear in the tree outside of their regions. We also observe regions whose departments fall on different parts of the tree such as PDLL and CVL or department 88 (Vosges in the East of France) which separates on its own. This is consistent with the previous analysis on individual genotypes where department 88 was already pointed out.

Similar results were obtained when using the FS clustering algorithm that initially indicated 158 clusters, which were resolved into 87 clusters using TVD analyses. When plotting these clusters on the map of France (Fig. 4c) or when reconstructing the TVD tree for the FS clusters (Fig. 4d), we observe that the South-East detaches first, then a small cluster in the Vosges area, followed by Brittany and a North-South divide. Further subdivisions were then observed in both the North and the South. The geographic coverage of the different clusters is variable, with some compact clusters with relatively few individuals, and others which are much larger and diffuse. There is a clear pattern of geographic proximity of individuals within the same cluster across all 87 groups and across the map of France, but there are also many overlaps as well as a small number of individuals who do not cluster with close geographic neighbours (see Figures S10, S11, S12 for more details on the different clusters). We also see boundaries that follow the main rivers (Loire, Garonne and Adour) as previously observed [14, 17]. Because of the choice we made of merging clusters with less than 50 individuals with their closest neighbour, the 11 individuals from the Island of Corsica do not form their own cluster but they belong to a large cluster including mostly individuals from PACA. What is more, we repeated the entire FS analysis a further two times and found a certain level of instability between runs, especially when we explore the finer levels of the tree to obtain more clusters. Individuals who were in the same cluster in the final outcome of run 1, were in the same cluster only 53.1% and 56.0% of the time in runs 2 and 3 (Supplementary Figure S13).

We also studied the IBD sharing between pairs of individuals and how it correlated with their geographical distance. A total of 7,605,814 IBD segments were estimated across all pairs of individuals on the 22 autosomal chromosomes. Only 6,054,760 (13.1%) of all pairs shared at least one segment and across those sharing pairs, the mean and maximum numbers of shared segments were 1.26 and 65, respectively. The mean and max total length of shared segments were 3.24 cM and 49.36 cM, respectively. The relation between distance and IBD-sharing is not completely homogeneous across the data of POPGEN, as can be seen in Figure S14 where we performed 9,598 separate regression models (indexed $j = 1, \dots, 9598$) with a single variable (KM) and restricting to only pairs within a distance of 300 KM and involving individual j . The slope coefficients of these 9,598 models are then plotted onto the map of France using universal Kriging [51, 52] with or without an inverse normal transformation. The resultant maps display the complex and continuous patterns of population structure, with certain zones standing out that correspond well to earlier analysis such as the unsupervised ADMIXTURE analysis in Figure S7.

Selecting an optimal subset of 4,000 individuals to sequence

To select 4,000 individuals while preserving the genetic diversity of POPGEN, we compared the variance explained by the 87 FS clusters, the 22 historical regions and the 95 departments using AMOVA [33], together with IBD-sharing regression models incorporating geographic distance (Supplementary Table S5). We found that FS clusters explain more variance than regions but perform similarly to departments and that the IBD regression model that includes FS clusters has a smaller R^2 compared to the models including regions and/or departments. We therefore chose to base our selection protocol for the 4,000 individuals on departments even if FS clusters show that the genetic structure does not follow precisely region or departmental boundaries. Department level is indeed the most practical and both AMOVA and IBD-sharing analyses show, using two different approaches, that the department classification captured genetic variability in a manner comparable to the FS clusters. However, FS results were used to identify outlier individuals who might be less appropriate candidates for the WGS panel. A total of 125 individuals were observed to be distal (geographically) from the other individuals of their assigned FS clusters and excluded from the list of possible candidates for WGS. Full details of the exclusion criteria are given in Supplementary Note 3 with an illustration in Figure S15.

We aimed to select 42 individuals per department, balanced by sex (21 men and 21 women), while maximizing geographic coverage and limiting redundancy due to close relatedness or geographic proximity. Additional exclusions were based on relatedness, inbreeding, insufficient DNA quantity and lack of consent for data reuse. After these filters, 9,106 individuals were eligible for selection. Where departmental sample sizes were insufficient, additional individuals were selected from neighbouring departments. The selection procedure also accounted for geographic distance and the reliability of assigned coordinates, as detailed in Supplementary Note 4 and Figure S16.

Among the 4,000 initially selected individuals, 118 failed sequencing and were replaced with geographically close, non-selected individuals matched as closely as possible on department, sex and geographic selection criteria. Two subsequent failures could not be replaced, resulting in a final whole-genome sequencing dataset of 3,998 individuals. Their geographic distribution and sex composition are reported in Supplementary Table S6 and Figure S17. Analyses repeated in this subset confirmed that the population structure observed in POPGEN9598 was preserved (Figures S18–S20).

Discussion

The POPGEN project was initiated within the France Genomic Medicine Initiative [19, 20] to generate regional allele-frequency data across metropolitan France using participants from the CONSTANCES cohort [25]. Because only 4,000 individuals could be sequenced, random sampling was not considered optimal. Instead, participants were selected through a multi-step process based on the birthplaces of their parents and grandparents. Among 15,000 individuals invited to provide saliva samples, 9,772 were genotyped, and genetic diversity analyses were then used to select 4,000 individuals for whole-genome

sequencing. By prioritising individuals whose four grandparents were born in geographically close locations, POPGEN was designed to capture fine-scale population structure and regional genetic variation. This approach provides a useful resource for demographic inference and for characterising the geographic stratification previously reported in France, particularly in northwestern regions [13, 14, 16, 17].

This design nevertheless limits the representativeness of POPGEN. Individuals with recent migration or admixture histories and populations from French overseas territories are not adequately represented. POPGEN may therefore have limited use as a direct control dataset in case–control studies, because comparable grandparental birthplace criteria are unlikely to be available in clinical cohorts, as previously discussed for the People of the British Isles project [53]. More broadly, underrepresentation of some populations may contribute to unequal access to genomic medicine. POPGEN should consequently be regarded as a complementary rather than standalone reference resource. Its data can be integrated into broader initiatives such as the Genome of Europe project, which aims to provide more than 100,000 whole-genome sequences representative of contemporary European populations, including more than 17,000 individuals from France. POPGEN may also support methods that do not require individual-level matching, including local-ancestry approaches such as SURFBAT [54], and may improve phasing and imputation, particularly for rare variants [24].

The geographic sampling strategy was another central design choice. Population genetic studies may sample within predefined geographic units or according to a regular spatial grid, reflecting the contrasting approaches historically advocated by Cavalli-Sforza and Wilson [55]. The former approach informed the Human Genome Diversity Panel, which sampled ethnolinguistically and geographically defined populations [1, 56], whereas the PoBI project mainly recruited within predefined rural regions [21]. In POPGEN, no region was prioritised during birthplace data collection. Sample sizes for saliva-kit invitations were initially defined by department using the 1901 census, while the final sequencing panel aimed for similar numbers per department, broad geographic coverage and a balanced sex ratio.

Although uniform departmental sampling may not maximise genetic diversity, clustering based on FineSTRUCTURE was not sufficiently stable beyond the coarsest level to guide selection reliably. Balanced sample sizes were also advantageous for geographic comparisons, as unequal group sizes often require down-sampling and loss of information [14, 17]. Similar considerations were made for the design of the Genome of the Netherlands project where authors chose to include an approximately equal representation from the original Dutch provinces [57]. French departments also provide relatively comparable geographic units. Their boundaries, established largely in 1790 and partly aligned with natural features, may capture both cultural and physical barriers to gene flow. Our analyses confirmed that departmental groupings reflect relevant components of fine-scale population structure.

POPGEN also differs from many genomic projects in its use of saliva rather than blood. Mailed self-collection kits achieved a participation rate of 68.3%, supporting the feasibility of this strategy for large-scale studies. A pilot study further showed that saliva-derived DNA performs comparably to blood-

derived DNA for whole-genome sequencing [58]. In addition, non-human sequencing reads can provide information on the oral microbiome. Recent studies have demonstrated that oral microbial profiles can be inferred from saliva whole-genome sequencing data and used to investigate associations between human genetic variation and microbial composition [59, 60]. Saliva-based sequencing may therefore provide a dual resource for human genomic and microbiome research.

Consistent with previous studies [13, 14, 16, 17], POPGEN confirms fine-scale genetic structure across France, with differentiation approaching that observed between some European countries. France's position between northern and southern Europe may partly contribute to this pattern. By extending geographic coverage beyond that of earlier studies, POPGEN shows that such structure is widespread and not restricted to a few culturally distinctive regions. Its strong geographic anchorage is therefore particularly valuable for reconstructing demographic history and modelling regional genetic diversity, although its direct applicability to clinical genetics remains limited.

Finally, POPGEN provides insights into sampling strategies for capturing genetic diversity with implications for future projects, including the Genome of Europe initiative. While POPGEN offers a detailed snapshot of French ancestry as shaped by historical populations, it does not capture recent demographic changes. Therefore, future efforts should aim to include more diverse and underrepresented populations, including recent migrants and overseas populations, to ensure equitable access to genomic medicine. In this perspective, POPGEN should be seen as a foundational contribution to a broader international effort, where integration across datasets will be essential to fully represent human genetic diversity.

Data Availability Statement

Access to the data can be requested through CONSTANCES website
<https://www.constances.fr/en/scientific-area/access-to-constances-2/>

Declarations

Ethical Approval

The POPGEN study (C19-39) was sponsored by Inserm. It was granted approval by the Ethics Committee Sud-Ouest et Outre-Mer 2 on 01/08/2019 (initial approval), 22/01/2021, 04/05/2023 and 04/09/2025 (corresponding to substantial amendments n° 1, 2 and 3 respectively). It was also approved by the French data privacy commission Cnil on 11/12/2020 (application N° 920379) and 23/11/2025 (applications No. 920379v1 and No. 925238). The protocol was registered with ClinicalTrials.gov (NCT04183023).

Competing Interests

None to declare

Fundings

This work was supported by The French Ministry of Research and Innovation in the framework of the French initiative for genomic medicine (Plan France Médecine Génomique 2025; PFMG 2025; https://pfmg2025.fr/wp-content/uploads/2025/05/genomic_medicine_france_2025.pdf).

The CONSTANCES cohort benefits from grant ANR-11-INBS-0002 from the French National Research Agency. Data from FranceGenRef project were generated by GENMED labex with funding ANR-10-LABX-0013 from the French National Research Agency.

Author contributions

A.F.H., G.L.F., G.M., A.S.P., R.R., E.H., J-F.D. and E.G. contributed to the conceptualization of the study. Data curation was performed by A.F.H., G.L.F., G.M., H.B., T.L., J-M.S., D.B-D. and S.S. Formal analysis was carried out by A.F.H., G.L.F., G.M., A.S.P., M.G., I.A., C.Di. and E.G. Funding was acquired by F.N. and E.G. Investigation was conducted by A.F.H., G.L.F., G.M., H.B., A.S.P., M.G., J-M.S. and E.G. A.F.H., G.L.F., G.M., H.B., A.S.P., M.G., I.A., J-M.S., D.B-D., R.O., C.Di., J-F.D. and E.G. contributed to the methodology. Project administration was handled by G.L.F., H.B., J-M.S., R.O., C.De., F.N., J-F.D. and E.G. Resources were provided by A.F.H., G.L.F., G.M., H.B., J-M.S., C.B., C.D-R., M.Z., J-F.D. and E.G. Supervision was provided by C.De., R.R., M.Z., J-F.D. and E.G. Validation was performed by A.F.H., G.L.F., G.M., H.B., J-M.S., E.H. and E.G. The original draft was written by A.F.H., G.L.F., G.M., I.S., R.R., J-F.D. and E.G., and all authors contributed to reviewing and editing the manuscript.

Acknowledgements

We would like to thank Garrett Hellenthal for his advice on model parameter choices for the FS algorithm. We would like to thank all of the volunteers from the CONSTANCES cohort for their participation. We also thank the CEPH-Biobank and the CNRGH for their technical expertise.

References

1. Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes. Princeton university press; 1994.
2. Hofer T, Ray N, Wegmann D, Excoffier L. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet.* 2009;73: 95–108. doi:10.1111/j.1469-1809.2008.00489.x

3. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. *Nature*. 2008;456: 98–101. doi:10.1038/nature07331
4. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, et al. Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet*. 2008;16: 1413–1429. doi:10.1038/ejhg.2008.210
5. Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, Pedersen NL, et al. The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet*. 2008;83: 787–794. doi:10.1016/j.ajhg.2008.11.005
6. Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, Di Blasio AM, et al. An overview of the genetic structure within the Italian population from genome-wide data. *PLoS One*. 2012;7: e43759. doi:10.1371/journal.pone.0043759
7. 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
8. Morin G-P, Moreau C, Barry A, Girard SL. Fine-scale structure of a whole regional population through genetics and genealogies. *Nat Commun*. 2026 [cited 20 Mar 2026]. doi:10.1038/s41467-026-70175-y
9. Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD. Fine-scale population structure and the era of next-generation sequencing. *Hum Mol Genet*. 2010;19: R221–R226. doi:10.1093/hmg/ddq403
10. Gloss AD, Steiner MC, Novembre J, Bergelson J. The design of mapping populations: Impacts of geographic scale on genetic architecture and mapping efficacy for defense and immunity. *Curr Opin Plant Biol*. 2023;74: 102399. doi:10.1016/j.pbi.2023.102399
11. Vallois HV, Marquer P. La répartition en France des groupes sanguins ABO. *Bulletins et Mémoires de la Société d'Anthropologie de Paris*. 1964;6: 1–200. doi:10.3406/bmsap.1964.1245
12. Ohayon E, Cambon-Thomsen A. *Génétique des populations humaines*. Paris: Édition INSERM; 1986.
13. Karakachoff M, Duforet-Frebourg N, Simonet F, Le Scouarnec S, Pellen N, Lecointe S, et al. Fine-scale human genetic structure in Western France. *Eur J Hum Genet*. 2015;23: 831–836. doi:10.1038/ejhg.2014.175
14. Saint Pierre A, Giemza J, Alves I, Karakachoff M, Gaudin M, Amouyel P, et al. The genetic history of France. *Eur J Hum Genet*. 2020;28: 853–865. doi:10.1038/s41431-020-0584-1
15. Brunel S, Bennett EA, Cardin L, Garraud D, Barrand Emam H, Beylier A, et al. Ancient genomes from present-day France unveil 7,000 years of its demographic history. *Proceedings of the National Academy of Sciences*. 2020;117: 12791–12798. doi:10.1073/pnas.1918034117
16. Biagini SA, Ramos-Luis E, Comas D, Calafell F. The place of metropolitan France in the European genomic landscape. *Hum Genet*. 2020;139: 1091–1105. doi:10.1007/s00439-020-02158-y
17. Alves I, Giemza J, Blum MGB, Bernhardsson C, Chatel S, Karakachoff M, et al. Human genetic structure in Northwest France provides new insights into West European historical demography. *Nat Commun*. 2024;15: 6710. doi:10.1038/s41467-024-51087-1

18. Darlu P, Chareille P. Analyse de la structure génétique et patronymique de la France métropolitaine (XIXe-XXe siècles). *Bulletins et mémoires de la Société d'Anthropologie de Paris BMSAP*. 2022;34. doi:10.4000/bmsap.9972
19. Lévy Y. Genomic medicine 2025: France in the race for precision medicine. *The Lancet*. 2016;388: 2872. doi:10.1016/S0140-6736(16)32467-9
20. Abadie C, Abderrahmane A, Abdous O, Abel C, Ackermann O, Acquaviva C, et al. PFMG2025– integrating genomic medicine into the national healthcare system in France. *The Lancet Regional Health – Europe*. 2025;50. doi:10.1016/j.lanepe.2024.101183
21. Winney B, Boumertit A, Day T, Davison D, Echeta C, Evseeva I, et al. People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *Eur J Hum Genet*. 2012;20: 203–210. doi:10.1038/ejhg.2011.127
22. Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C, Quintela I, Carracedo Á, Donnelly P, et al. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun*. 2019;10: 551. doi:10.1038/s41467-018-08272-w
23. Raveane A, Aneli S, Montinaro F, Athanasiadis G, Barlera S, Birolo G, et al. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Sci Adv*. 2019;5: eaaw3492. doi:10.1126/sciadv.aaw3492
24. Herzig AF, Velo-Suárez L, FrEx Consortium, FranceGenRef Consortium, Dina C, Redon R, et al. How local reference panels improve imputation in French populations. *Sci Rep*. 2024;14: 370. doi:10.1038/s41598-023-49931-3
25. Zins M, Goldberg M. The French CONSTANCES population-based cohort: design, inclusion and follow-up. *European journal of epidemiology*. 2015;30. doi:10.1007/s10654-015-0096-4
26. Becker RA, Wilks AR, Brownrigg R, Minka TP, Deckmyn A. maps: Draw Geographical Maps. 2003. p. 3.4.3. doi:10.32614/CRAN.package.maps
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 2007;81: 559–575.
28. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet*. 2008;83: 347–358. doi:10.1016/j.ajhg.2008.08.005
29. Consortium T 1000 GP. A global reference for human genetic variation. *Nature*. 2015;526: 68. doi:10.1038/nature15393
30. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. *Science*. 2002;298: 2381–2385. doi:10.1126/science.1078311
31. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. *PLOS Genetics*. 2006;2: e190. doi:10.1371/journal.pgen.0020190
32. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

33. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*. 1992;131: 479–491. doi:10.1093/genetics/131.2.479
34. Kamvar ZN, Tabima JF, Grünwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*. 2014;2: e281. doi:10.7717/peerj.281
35. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. Copenhaver GP, editor. *PLoS Genet*. 2012;8: e1002453. doi:10.1371/journal.pgen.1002453
36. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, et al. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet*. 2008;83: 132–135; author reply 135–139. doi:10.1016/j.ajhg.2008.06.005
37. Koenig Z, Yohannes MT, Nkambule LL, Zhao X, Goodrich JK, Kim HA, et al. A harmonized public resource of deeply sequenced diverse human genomes. *Genome Res*. 2024;34: 796–809. doi:10.1101/gr.278378.123
38. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. 2006;34: D590-598. doi:10.1093/nar/gkj144
39. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. *Genetics*. 1992;132: 583–589. doi:10.1093/genetics/132.2.583
40. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting F_{ST} : The impact of rare variants. *Genome Res*. 2013;23: 1514–1521. doi:10.1101/gr.154831.113
41. Al Bkhetan Z, Chana G, Ramamohanarao K, Verspoor K, Goudey B. Evaluation of consensus strategies for haplotype phasing. *Briefings in Bioinformatics*. 2021;22: bbaa280. doi:10.1093/bib/bbaa280
42. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 2013;10: 5–6. doi:10.1038/nmeth.2307
43. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*. 2021;108: 1880–1890. doi:10.1016/j.ajhg.2021.08.005
44. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48: 1443–1448. doi:10.1038/ng.3679
45. Bhérér C, Campbell CL, Auton A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat Commun*. 2017;8: 14994. doi:10.1038/ncomms14994
46. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics*. 2014;30: 1266–1272. doi:10.1093/bioinformatics/btu014
47. Yang Y, Durbin R, Iversen AKN, Lawson DJ. Sparse haplotype-based fine-scale local ancestry inference at scale reveals recent selection on immune responses. *Nat Commun*. 2025;16: 2742. doi:10.1038/s41467-025-57601-3

48. Bates D, Maechler M, Jagan M, libraries TAD (SuiteSparse, dir(system.file("doc" collaborators listed in, "SuiteSparse," et al. Matrix: Sparse and Dense Matrix Classes and Methods. 2025. Available: <https://cran.r-project.org/web/packages/Matrix/index.html>
49. Karatzoglou A, Smola A, Hornik K, Australia (NICTA) NI, Maniscalco MA, Teo CH. kernlab: Kernel-Based Machine Learning Lab. 2024. Available: <https://cran.r-project.org/web/packages/kernlab/index.html>
50. Zhou Y, Browning SR, Browning BL. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *Am J Hum Genet.* 2020;106: 426–437. doi:10.1016/j.ajhg.2020.02.010
51. Pebesma EJ. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences.* 2004;30: 683–691. doi:10.1016/j.cageo.2004.03.012
52. Gräler B, Pebesma E, Heuvelink G. Spatio-Temporal Interpolation using gstat. *The R Journal.* 2016;8: 204. doi:10.32614/RJ-2016-014
53. Tyler-Smith C, Xue Y. A British approach to sampling. *Eur J Hum Genet.* 2012;20: 129–130. doi:10.1038/ejhg.2011.153
54. Herzig AF, Rubinacci S, Marenne G, Perdry H, FrEx Consortium, FranceGenRef Consortium, et al. SURFBAT: a surrogate family-based association test building on large imputation reference panels. *G3 Genes|Genomes|Genetics.* 2024; jkae287. doi:10.1093/g3journal/jkae287
55. Roberts L. A Genetic Survey of Vanishing Peoples. *Science.* 1991;252: 1614–1617. doi:10.1126/science.2047869
56. Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, et al. A human genome diversity cell line panel. *Science.* 2002;296: 261–262. doi:10.1126/science.296.5566.261b
57. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet.* 2014;22: 221–227. doi:10.1038/ejhg.2013.118
58. Herzig AF, Velo-Suárez L, Le Folgoc G, Boland A, Blanché H, Olaso R, et al. Evaluation of saliva as a source of accurate whole-genome and microbiome sequencing data. *Genetic Epidemiology.* 2021;45: 537–548. doi:10.1002/gepi.22386
59. Kamitaki N, Handsaker RE, Hujoel MLA, Mukamel RE, Usher CL, McCarroll SA, et al. Human and bacterial genetic variation shape oral microbiomes and health. *Nature.* 2026;651: 429–439. doi:10.1038/s41586-025-10037-7
60. Velo-Suárez L, Herzig AF, Bocher O, Le Folgoc G, Le Roux L, Delmas C, et al. Sequencing depth overcomes extraction bias: repurposing human WGS data for salivary microbiome profiling. *Genomics;* 2026. doi:10.64898/2026.03.27.714786

Figures

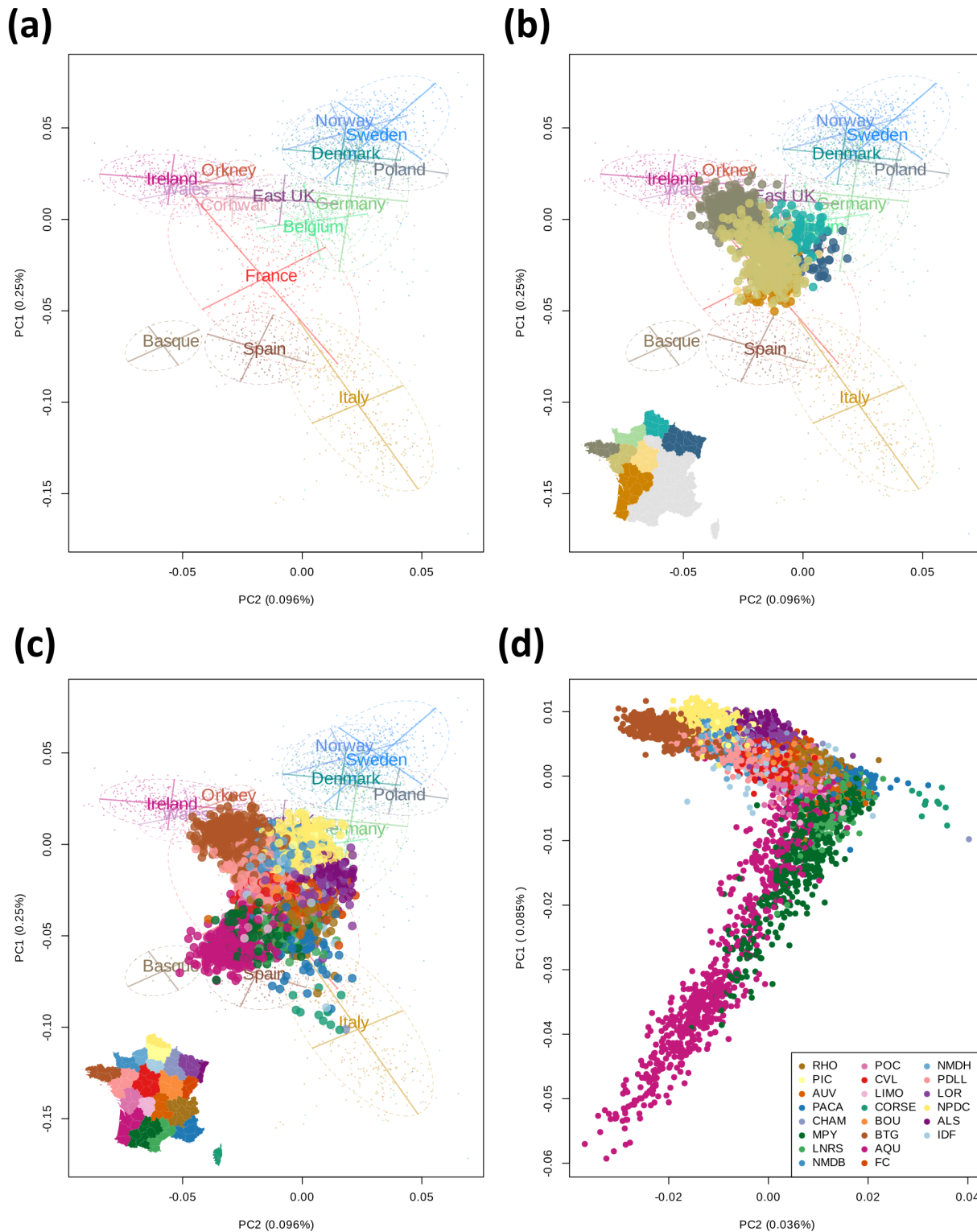


Figure 1

Projection of POPGEN data onto known axes of variation in Western Europe.

(a) PCA analysis on the same Western European reference data as in [16], using variants that were also present in the POPGEN9598 dataset. Each individual is plotted as a small point and minimal ellipses with minor and major axes containing 90% of the individuals are added for each group. (b) same as (a) with the addition of projections of the 856 individuals from the FranceGenRef study [16]. Colors refer to

different regions of France as indicated in the legend. (c) same as (a) but now with the 9,598 individuals of POPGEN (following QC) projected. (d) first two principal components re-calculated on just the 9,598 individuals of POPGEN. In all plots, we place PC2 on the x-axis and PC1 on the y-axis so that the projections are more recognizable as reflecting geographic locations in Europe and in France.

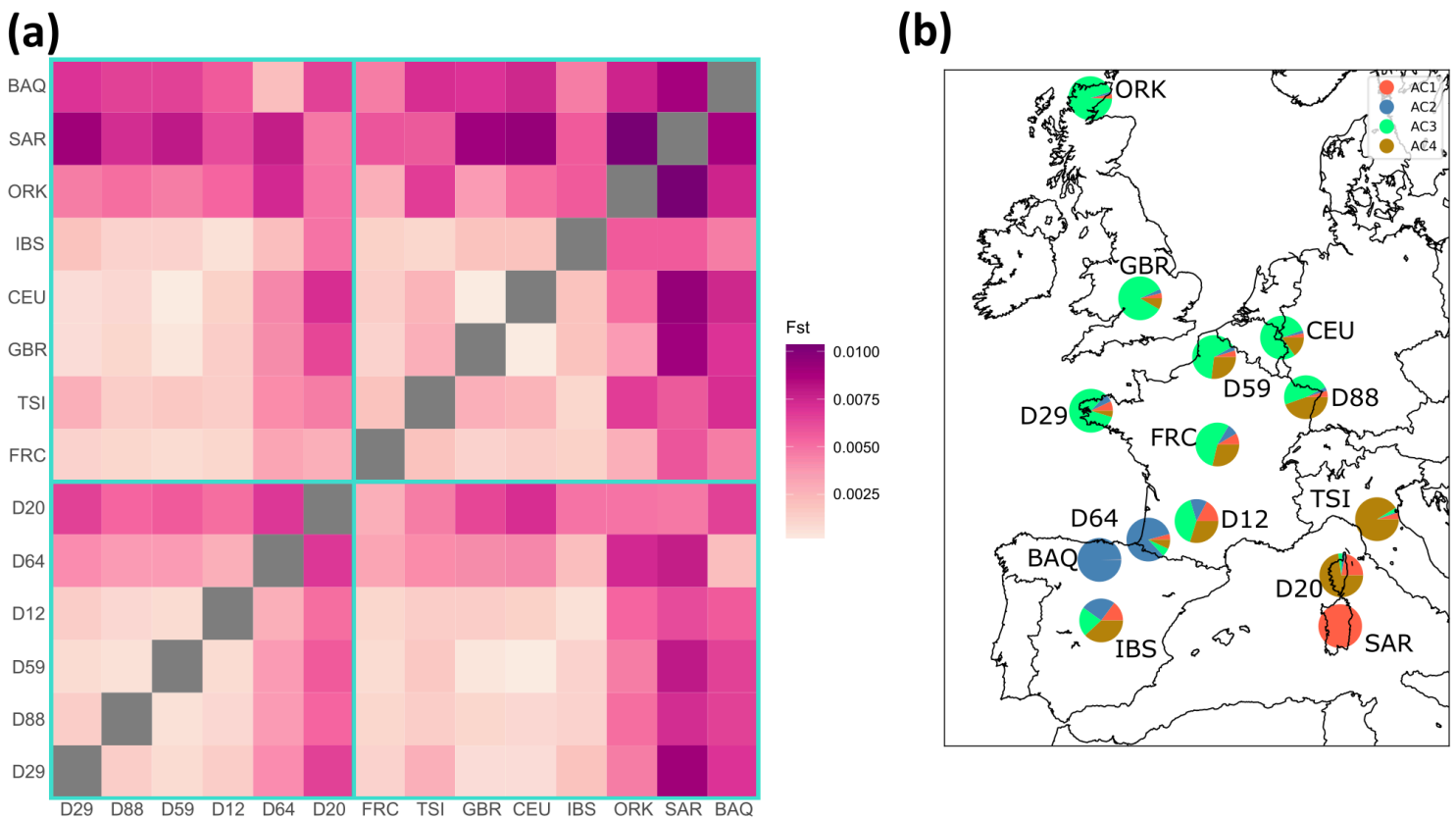


Figure 2

Placing the internal diversity of POPGEN into the wider context of Western Europe.
 (a) F_{st} matrix between groups, including 6 departments from POPGEN and 8 populations from NFE_1KG_HGDP with blue lines separating these two collections. (b) Approximate map of the location of the different groups presented in (a), with average unsupervised admixture components for each group superimposed.

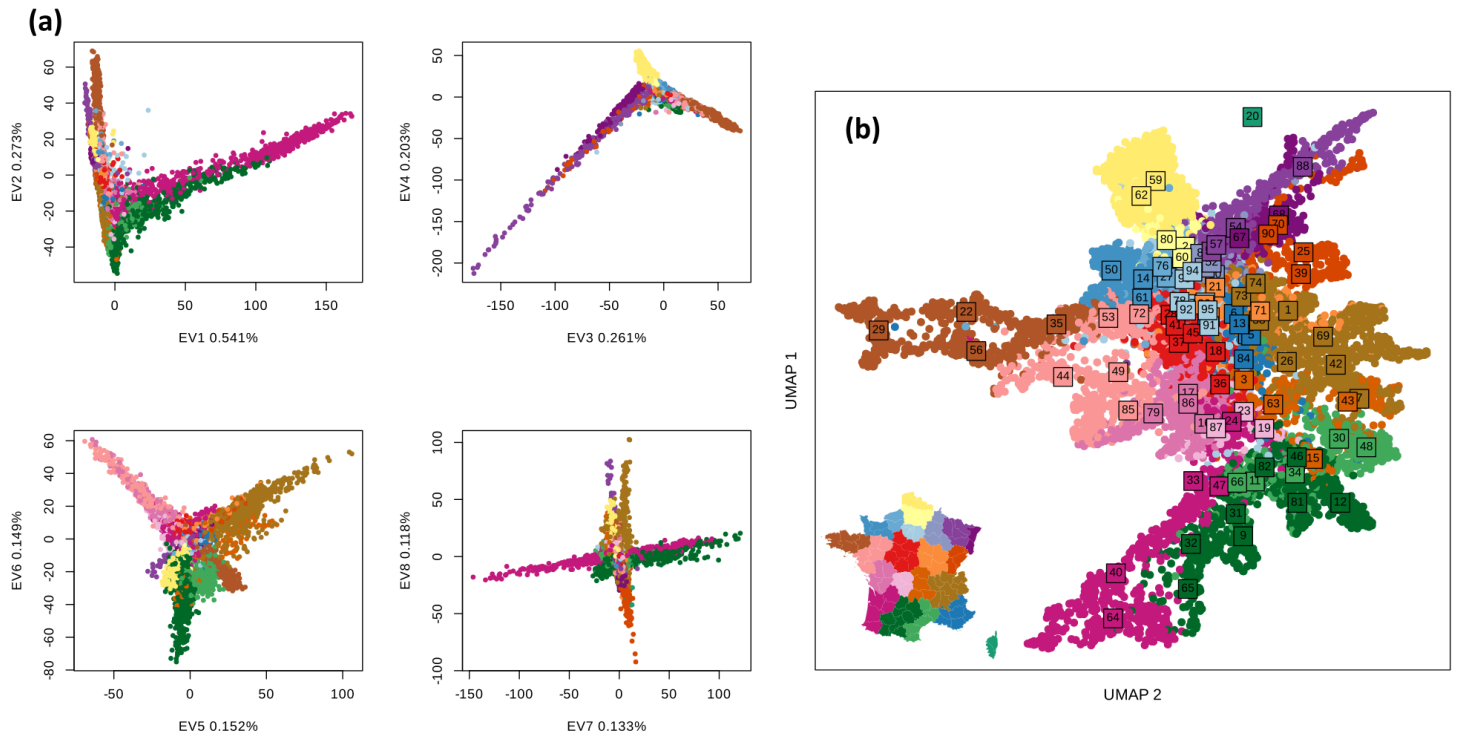


Figure 3

Fine-scale population structure in metropolitan France.

(a) Spectral decomposition of the chunk-counts matrix obtained with PBWT-paint algorithm – results are shown for the first 8 eigen-vectors; (b) UMAP projection of the entire chunk-count matrix with the centres-of-mass of each department overlain. The different colours correspond to the 'old' regions of France as shown in the inset.

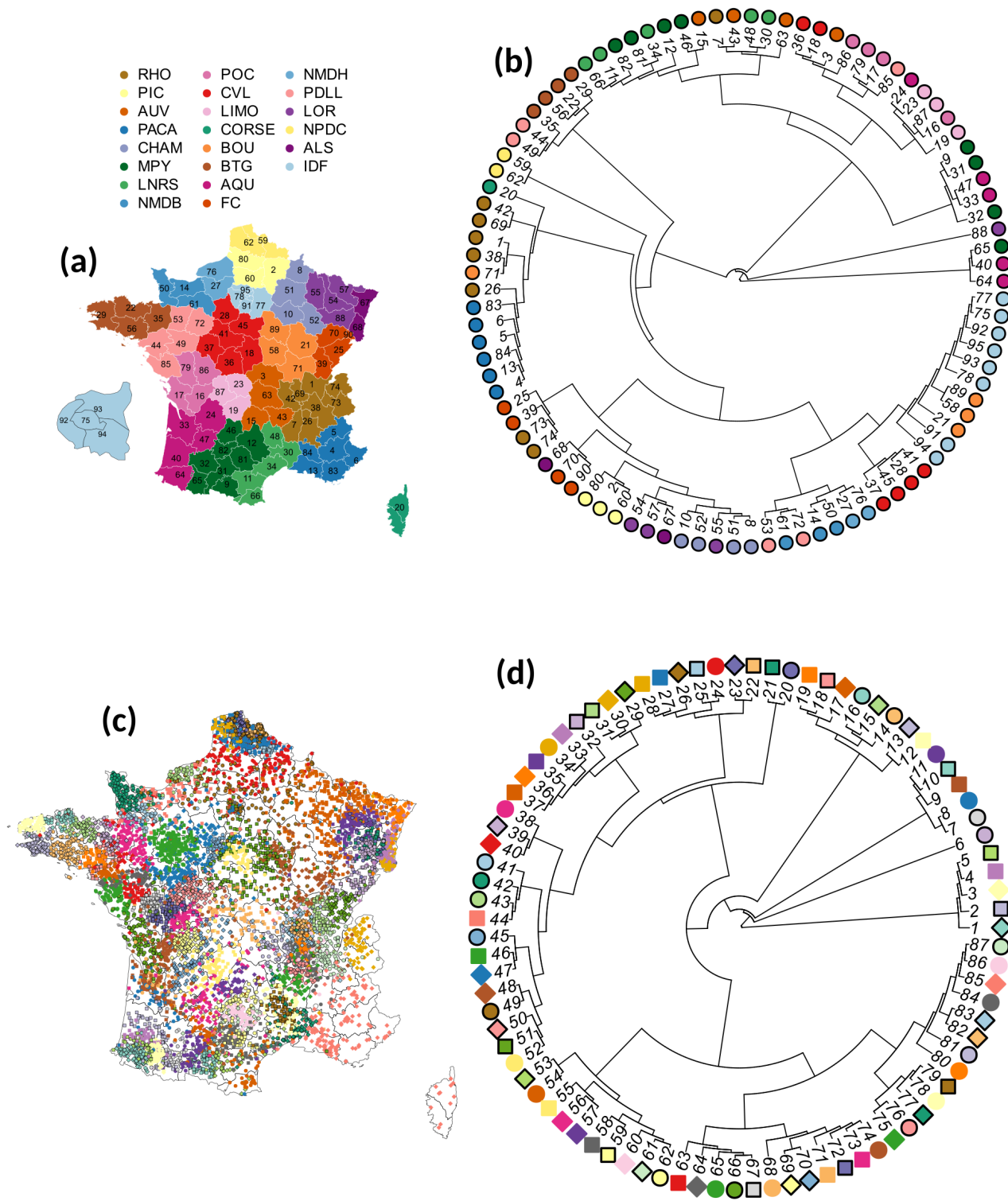


Figure 4

Haplotype-sharing across the 95 administrative departments of France and the 87 FineSTRUCTURE (FS) clusters.

(a) Legend of the different colours and department labels; (b) Neighbour joining tree based on Total-variance distance (TVD-tree) connecting the 95 departments; (c) Map of France with the 9,598

individuals plotted with points that correspond to the 87 FS clusters; (d) TVD-tree connecting the 87 FS clusters.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PopgenresourcesupplementsTextTablecombined.pdf](#)
- [PopgenresourcesupplementsTextTablesSubmittedv2.docx](#)
- [SupplementaryFigS1S20.pdf](#)