

1 **The Genetic History of France**

2 Aude Saint Pierre^{1*}, Joanna Giemza^{2*}, Matilde Karakachoff², Isabel Alves², Philippe
3 Amouyel³⁺, Jean-François Dartigues⁴⁺, Christophe Tzourio⁴⁺, Martial Monteil⁵, Pilar Galan⁶,
4 Serge Hercberg⁶, Richard Redon², Emmanuelle Génin^{1*}, Christian Dina^{2*}

5 *¹Univ Brest, Inserm, EFS, CHU Brest, UMR 1078, GGB, F-29200 Brest, France*

6 *²l'institut du thorax, INSERM, CNRS, Univ Nantes, CHU Nantes, Nantes, France*

7 *³Univ. Lille, Inserm, CHU Lille University Hospital, Institut Pasteur de Lille, LabEx DISTALZ-
8 UMR1167 - RID-AGE - Risk factors and molecular determinants of aging-related, F-59000
9 Lille, France*

10 *⁴Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, CHU
11 Bordeaux, F-33000 Bordeaux, France*

12 *⁵Université de Nantes, UMR 6566 CReAAH, LARA, Nantes, France*

13 *⁶Université Paris 13, Equipe de Recherche en Epidémiologie Nutritionnelle, Centre de
14 Recherche en Epidémiologie et Statistiques, Inserm (U1153), Inra (U1125), Cnam, COMUE
15 Sorbonne Paris Cité, F-93017, Bobigny, France*

16 * These authors contributed equally to this work

17 + On behalf of the 3C study

18 **Running Title:** The Genetic History of France

19 **Key words:** Population stratification, Genetic ancestry, Admixture, Demographic history, Gene
20 flow barriers, Association Study

21 **Corresponding authors:** Aude Saint Pierre, aude.saintpierre@univ-brest.fr

22 Christian Dina, christian.dina@univ-nantes.fr

23

24 **ABSTRACT**

25 The study of the genetic structure of different countries within Europe has provided significant
26 insights into their demographic history and their actual stratification. Although France occupies
27 a particular location at the end of the European peninsula and at the crossroads of migration
28 routes, few population genetic studies have been conducted so far with genome-wide data. In
29 this study, we analyzed SNP-chip genetic data from 2 184 individuals born in France who were
30 enrolled in two independent population cohorts. Using FineStructure, six different genetic
31 clusters of individuals were found that were very consistent between the two cohorts. These
32 clusters match extremely well the geography and overlap with historical and linguistic divisions
33 of France. By modeling the relationship between genetics and geography using EEMS software,
34 we were able to detect gene flow barriers that are similar in the two cohorts and corresponds to
35 major French rivers or mountains. Estimations of effective population sizes using IBDNe
36 program also revealed very similar patterns in both cohorts with a rapid increase of effective
37 population sizes over the last 150 generations similar to what was observed in other European
38 countries. A marked bottleneck is also consistently seen in the two datasets starting in the
39 fourteenth century when the Black Death raged in Europe. In conclusion, by performing the
40 first exhaustive study of the genetic structure of France, we fill a gap in the genetic studies in
41 Europe that would be useful to medical geneticists but also historians and archeologists.

42

43

44 INTRODUCTION

45 *Gallia est omnis divisa in partes tres* [*commentarii de bello gallico*¹] was one of the earliest
46 demographic description of antique France (known as Gaul). These three parts were Aquitania,
47 in South West, with Garonne and the Pyrenees mountains as borders; Belgia in North West,
48 following the Seine as Southern border; and finally what we know as Celtic Gaul, that spanned
49 from the Atlantic Ocean to the Rhine River and Alps. A fourth part of the present-day French
50 territory, already part of Romanized territories at this time, was Gallia Transalpina, a strip of
51 lands from Italy to Iberia, with Alps and Cevennes mountains as northern border.

52 The area that was to be modern France was subject to successive population migrations:
53 Western Hunter-Gatherers (15 kya), Neolithic farmers (7 kya) and later steppe Enolithic Age
54 populations^{2,3}, Celtic expansion, integration in Roman empire, Barbarian Great migrations,
55 whose demographical importance remains to be assessed. France's position in Europe, at the
56 edge of the Eurasian peninsula, has made it not only the final goal of a large number of,
57 potentially massive, migrations but also a place of transit either to the North (British Isles) or
58 the South of Europe (Iberian Peninsula) and North Africa, as well as an important crossroad for
59 trade and exchanges.

60 Before France became a single political entity, its territory was divided into various kingdoms
61 and later provinces, which often displayed fierce independence spirit towards the central power.
62 The pre-Roman Gaul was divided into politically independent territories. After the fall of
63 Roman Empire, the modern French territory was divided into Barbarian Germanic kingdoms
64 (Franks, Wisigoths and Burgunds). After a short period of reunification and extension into the
65 Carolingian Empire (VIIth century), the weakening of the central power led to the reduction of
66 the Occidental France at its western part and the rise of local warlords gaining high
67 independence within the Kingdom itself. The feudality period created provinces that were close

68 to independence, although nominally linked through the oath of allegiance to the King of France
69 (Figure S1).

70 During centuries, in spite of important backlashes such as the Hundred Years War, the French
71 Kings managed to slowly integrate the Eastern lands as well as Brittany, enforcing in parallel
72 the central power until the French Revolution. However, every province kept displaying
73 political, cultural and linguistic differences, which could have left imprints in the genetic
74 structure of modern French populations.

75 Geographically, modern France is a continental country surrounded by natural borders: the
76 Atlantic Ocean on the West side, the Channel Sea up North, mountains (Pyrenees and Alps)
77 closing the South-West and East/South-East borders, as well as the Mediterranean Sea on the
78 South side. The Eastern side has the Rhine as a natural border on less than 500 kilometers while
79 the Northeastern borders shows no notable obstacle and exhibits a continuum with Germany
80 and Belgium. This complex history is expected to have shaped the genetic make-up of the
81 current French population and left some footprints in its genetic structure.

82 The study of the genetic structure of human populations is indeed of major interest in many
83 different fields. It informs on the demographic history of populations and how they have formed
84 and expanded in the past with some consequences on the distribution of traits. Genetic
85 differences between populations can give insights on genetic variants likely to play a major role
86 on different phenotypes, including disease phenotypes⁴. This explains the growing interest of
87 geneticist for human population studies that aim at describing the genetic diversity and are now
88 facilitated by the rich genetic information available over the entire genome. In the last decades,
89 several studies were performed using genome-wide SNP data often collected for genome-wide
90 association studies. These studies have first shown that there exist allele frequency differences
91 at all geographic scales and that these differences increase with geographic distances. Indeed,
92 the first studies have shown differences between individuals of different continental origins⁵⁻⁷

93 and then, as more data were collected and marker density increased, these differences were
94 found within continents and especially within Europe^{8,9}. Several studies have also been
95 performed at the scale of a single country and have shown that differences also exist within
96 country. This was for instance observed in Sweden, where Humphreys et al.¹⁰ reported strong
97 differences between the far northern and the remaining counties, partly explained by remote
98 Finnish or Norwegian ancestry. More recent studies have shown structure in the Netherlands¹¹,
99 Ireland¹², UK¹³ or Iberian peninsula¹⁴. Previous studies of population stratification in France
100 have examined only Western France (mainly *Pays de la Loire* and Brittany) and detected a
101 strong correlation between genetics and geography¹⁵. However, no study so far has investigated
102 the fine-scale population structure of the entire France using unbiased samples from individuals
103 with ancestries all over the country.

104 In this paper, we applied haplotype-based methods that have been shown to provide higher
105 resolution than allele-based approaches¹³ to investigate the pattern of fine-scale population
106 stratification in France. We used two independent cohorts, 3C and SUVIMAX with more than
107 2 000 individuals whose birthplace covered continental France and genotyped at the genome-
108 wide level, to assess the genetic structure of the French population and draw inferences on the
109 demographic history.

110

111 **MATERIAL AND METHODS**

112 **Data on SU.VI.MAX & 3C studies**

113 Genetic data were obtained from two French studies, SU.VI.MAX¹⁶ and the Three-Cities
114 study¹⁷ (3C) with the idea to compare if, by analyzing them independently, concordant results
115 would be obtained. Indeed, one major drawback of genetic inferences obtained on population
116 samples is the fact that they can strongly depends on how individuals were sampled and which
117 genetic markers were used. Here, by using data from two studies that sampled individuals using

118 different criteria and genotyped them on different SNP-chip, we should be able to draw more
119 robust inferences.

120 For every individual, information on places of birth was available, either the exact location (3C
121 study) or the “*département*” (SU.VI.MAX). *Départements* are the smallest administrative
122 subdivisions of France. There are a total of 101 French *départements* and 94 of them are located
123 in continental France. These units were created in year 1 789, during French Revolution, partly
124 based on historical counties.

125 3C Study: The Three-City Study was designed to study the relationship between vascular
126 diseases and dementia in 9 294 persons aged 65 years and over. For more details on the study,
127 see <http://www.three-city-study.com/the-three-city-study.php>. Analyses were performed on
128 individuals who were free of dementia or cognitive impairment by the time their blood sample
129 was taken and who were previously genotyped¹⁸. The geographical locations of individuals
130 were defined according to the latitude and longitude of their place of birth, declared at
131 enrolment. Individuals with missing place of birth or born outside continental France were
132 excluded. A total of 4 659 individuals were included in the present study.

133 SU.VI.MAX: The study was initiated in 1 994 with the aim of collecting information on food
134 consumption and health status of French people. A subset of 2 276 individuals born in any of
135 the 94 continental French *départements* was included in this study. The geographic coordinates
136 of each *départements* were approximated based on the coordinates of the corresponding main
137 city.

138

139 **Quality control**

140 Quality control of the genotypes was performed using the software PLINK version 1.9^{19,20}.

141 3C: raw genotype data were generated in the context of a previous study¹⁸ on Illumina
142 Human610-Quad BeadChip. Following the recommendations from Anderson et al.²¹,

143 individuals were removed if they had a call rate < 99%, heterozygosity level \pm 3 standard
144 deviations (SD) from the mean. Cryptic relatedness was assessed by estimating π_{hat} (the IBD
145 test implemented in PLINK²⁰) in each dataset after doing LD-based pruning. Individuals related
146 to another individual from the sample with an IBD proportion of 0.1875 or above were removed
147 (only one individual was kept from each pair). As a final quality control to exclude outlier
148 individuals from populations, we performed principle component analysis (PCA) using the
149 smartpca software from the EIGENSOFT package version 6.0.1²² and removed outliers across
150 the first 10 eigenvectors. The default procedure was used for outlier removals with up to 5
151 iterative PCA runs and, at each run, removing of individuals with any of the top 10 PCs
152 departing from the mean by more than 6 standard deviations. SNPs in strong linkage
153 disequilibrium (LD) were pruned out with PLINK 1.9 (described in PCA section). Outlier
154 individuals were removed prior to performing further analyses. Applying all these QC filters
155 led to the removal of 226 individuals. To avoid redundant information from individuals born at
156 the same place, we randomly selected only one individual from each place of birth. A total of
157 770 individuals covering the 94 continental French “*départements*” were included. All samples
158 failing sample-level QC were removed prior to performing SNPs QC. Markers were removed
159 if they had a genotype-missing rate > 1%, a minor allele frequency < 1% or departed from
160 Hardy–Weinberg proportion ($P \leq 10^{-7}$). After QC, there were 770 individuals and 490 217
161 autosomal SNPs.

162 SU.VI.MAX: Genotype data of the 2 834 samples were available from previous studies using
163 different SNP chips: 1 978 with Illumina 300k/300kduo and 856 with Illumina 660W.
164 Individuals with an unknown birthplace or a birthplace outside of continental France were
165 removed, 1 416 samples were left. Two individuals were removed because of a call rate < 95%.
166 IBD statistic, calculated in PLINK version 1.9, didn’t identify any related samples with a
167 threshold of 0.1875. SNPs were removed if they had a genotype-missing rate > 2%, a minor

168 allele frequency $< 10\%$ or departed from Hardy–Weinberg proportion ($P \leq 10^{-5}$). After QC,
169 there were 1 414 individuals and 271 886 autosomal SNPs.

170

171 **Population structure within France**

172 *Chromopainter/FineSTRUCTURE analysis*

173 For investigating fine-scale population structure, we used Chromopainter version 2 and
174 FineSTRUCTURE version 2.0.7²³. Data were phased with SHAPEIT v2.r790²⁴ using the 1000
175 genomes dataset as a reference panel. In the 3C dataset, we removed 932 of these SNPs because
176 of strand issues prior to phasing. Files were then converted to Chromopainter format using the
177 ‘impute2chromopainter2.pl’ script. Chromopainter outputs from the different chromosomes
178 were combined with chromocombine to generate a final coancestry matrix of chunk counts for
179 FineSTRUCTURE. For the FineSTRUCTURE run we sampled values after successive series
180 of 10 000 iterations for 1 million MCMC iterations following 10 million “burn-in” iterations.
181 Starting from the MCMC sample with the highest posterior probability among all samples,
182 FineSTRUCTURE performed 100 000 additional hill-climbing moves to reach its final inferred
183 state (See¹³ for details). The final tree was visualized in R with the help of FineSTRUCTURE
184 and ‘dendextend’ libraries. We checked that the MCMC samples were independent of the
185 algorithm’s initial position by visually comparing the results of two independent runs starting
186 from different random seeds. Good correspondence in the pairwise coancestry matrices of the
187 two runs indicates convergence of the MCMC samples to the posterior distribution. Without
188 loss of generality, we used the first of these two runs in our main analysis.

189 *Ancestry profiles of the French population & Spatial pattern of genetic structure EEMS*

190 We used ADMIXTURE v1.3²⁵ to estimate mixing coefficients of each individual. We
191 performed runs for values of K between 2 and 10, with 5-fold cross-validation using the set of
192 pruned SNPs, as described in the PCA analyses. To identify if cluster differences existed, we

193 performed a one-way analysis of variance (ANOVA) on the admixture components, followed
194 by *post hoc* pairwise comparisons.

195 We estimated an effective migration surface using the software EEMS²⁶. We run EEMS with
196 slightly different grids to investigate how/whether these changes affected the results. Plots were
197 generated in R using the “rEEMSpots” package according to instructions in the manual. For
198 both datasets the full set of SNPs was included. For more information on the specific pipeline,
199 see Supplementary Data.

200 *IBD-estimated population size*

201 We estimated the recent effective population size with IBDNe²⁷. IBDNe was run with the
202 default parameters and a minimum IBD segment length of 4 cM (mincm=4). We used the
203 default settings to filter IBD segments from IBDseq v. r1206 software package²⁸. Breaks and
204 short gaps in IBD segments were removed with the merge-ibd-segments utility program. For
205 IBD detection, we varied the minimum IBD segment length in centiMorgan units by the mincm
206 parameter (mincm argument) from the default value, 2 cM, to 8 cM. IBDNe analysis was
207 applied on the whole SU.VI.MAX and 3C datasets as well as on the major subpopulations from
208 fineSTRUCTURE clustering. Growth rates were calculated with the formula
209 $\frac{\text{end value} - \text{start value}}{\text{start value}}$. We assumed a generation time of 30 years, as assumed in the original
210 paper.

211 *Principal Component Analysis (PCA) and F_{ST}*

212 Both PCA and F_{ST} analyses were carried out on a pruned set of SNPs in each dataset
213 independently and using the smartpca tool in the EIGENSOFT program (v6.1.1)²². The pairwise
214 F_{ST} matrix were estimated using the option ‘fsthiprecision=YES’ in smartpca. We calculated
215 the mean F_{ST} between clusters inferred by FineSTRUCTURE as group labels. In each dataset,
216 SNPs in strong LD were pruned out with PLINK in a two-step procedure. SNPs located in
217 known regions of long range linkage disequilibrium (LD) in European populations were

218 excluded from the analysis²⁹. Then, SNPs in strong LD were pruned out using the ‘indep-
219 pairwise’ command in PLINK. The command was run with a linkage disequilibrium $r^2=0.2$, a
220 window size of 50 SNPs and 5 SNPs to shift the window at each step. This led to a subset of
221 100 973 SNPs and 83 246 SNPs in the 3C and SU.VI.MAX datasets respectively. To evaluate
222 the geographic relevance of PCs, we tested for the significance of association between the
223 latitude and longitude of each *département* and PCs coordinates (‘cor.test’ function in R) using
224 a Spearman’s rank correlation coefficient.

225

226 **Relation with neighbor European populations: 1000G and HGDP**

227 We assembled SNP data matching either the SU.VI.MAX or the 3C genotype data (after quality
228 control) with the European individuals from the 1000G phase 3 reference panel and from the
229 Human Genome Diversity Panel data³⁰ (HGDP, Illumina HuHap 650k), to generate four
230 genome-wide SNP datasets analyzed independently.

231 The 1000G reference panel served as donor populations when estimating ancestry proportions.
232 First, in order to define a set of donor groups from 1000G Europe (EUR), we used the subset
233 of unrelated and outbred individuals generated in the study of Gazal et al.³¹. Four European
234 populations were considered: British heritage (GBR, n=85 and CEU, n=94), Spain (IBS,
235 n=107) and Italy (TSI, n=104). These 390 Europeans individuals were then combined with
236 individuals from both datasets independently resulting in a set of 484 874 common SNPs with
237 3C and a set of 232 148 common SNPs with SU.VI.MAX. The filtered datasets (after pruning)
238 included 1 160 individuals genotyped on 100 851 SNPs in the 3C Study and 1 804 individuals
239 genotyped on 64 653 SNPs in SU.VI.MAX. We inferred European ancestry contributions in
240 France using the novel haplotype-based estimation of ancestry implemented in
241 SOURCEFIND³². SOURCEFIND has been shown to give a greater accuracy than the usual
242 Non-Negative least squares regression for inferring proportion of admixture but because it is

243 recommended to use homogeneous donor groups, we ran FineSTRUCTURE on the four
244 European populations defined above and selected the level of clustering describing the main
245 features of the donor populations. These European donor groups served as reference in
246 SOURCEFIND. We performed analysis of variance (ANOVA) on French admixture
247 component per cluster group to identify whether cluster differences existed.

248 Additional analyses combining the European participants of the HGDP panel were carried out
249 in order to estimate the contribution of Basque population of our South West clusters. A total
250 of 160 European HGDP participants were included from 8 populations: Adygei (n=17), French-
251 Basque (n=24), French (n=29), Italian (n=13), Italian from Tuscany (n=8), Sardinian (n=28),
252 Orcadian (n= 16) and Russian (n=25). Using the same procedure for merging panels, the filtered
253 datasets (after pruning) included 930 individuals and 93 938 SNPs in the 3C Study and 1 574
254 individuals and 57 775 SNPs in SU.VI.MAX.

255

256 **RESULTS**

257 **Chromopainter/FineSTRUCTURE analysis reveals consistent fine-scale genetic** 258 **stratification within France**

259 Results of FineSTRUCTURE analysis reveal fine-scale population patterns within France at a
260 very fine level that are very consistent in the two datasets (Figure 1). FineSTRUCTURE
261 identified respectively 17 and 27 clusters in 3C and SU.VI.M.AX, demonstrating local
262 population structure (Figure S2). Even though the sampling distributions of individuals varies
263 slightly between datasets both analyses show very concordant partitions with a broad
264 correlation between clusters and geographic coordinates. The major axis of genetic
265 differentiation runs from the south to the north of France.

266 In both datasets, the coarsest level of genetic differentiation (i.e. the assignment into two
267 clusters) separates the south-western regions from the rest of France (Figure S3 and S4). Next

268 levels of tree structures slightly differ between the two datasets but converge into a common
269 geographic partitions at $k=6$ clusters in 3C and $k=7$ in SU.VI.MAX (Figure 2). The clusters are
270 geographically stratified and were assigned labels to reflect geographic origin: the South-West
271 SW for the dark-red cluster, the South (SO) for the orange cluster, the Centre (CTR) for the
272 yellow cluster, the North-West (NW) for the pink cluster, the North (NO) for the blue cluster
273 and the South-East (SE) for the cyan cluster. In each dataset, one cluster (labelled “*Others*” and
274 coloured in red) included individuals geographically dispersed over France. Furthermore, one
275 cluster identified in SU.VI.MAX included only one individual and was removed in further
276 analysis so that $k=7$ also resumed to 6 clusters in SU.VI.MAX. At this tree level of 6 clusters,
277 individuals from the NO, NW and CTR clusters are clearly separated in the two datasets. The
278 SW cluster and part of the SO cluster in 3C match geographically the SW cluster identified in
279 SU.VI.MAX while the SE subgroup was not detected in the 3C. This might be explained by
280 differences in the geographic coverage between the two studies especially in the south of
281 France. Indeed, SU.VI.MAX has a better coverage of the south-east whereas 3C lacks data from
282 this region and the reverse is true for the south-west. In the two datasets, two large clusters
283 (CTR and NO) are found that cover most of the central and northern France. Notably, even at
284 the finest level of differentiation (17 and 27 clusters in 3C and SU.VI.MAX respectively), these
285 clusters remain largely intact.

286 The broad-scale genetic structure of France in six clusters strikingly aligns with two major
287 rivers of France, “La Garonne” and “La Loire” (Figure 2). At a finer-scale, the “Adour” river
288 partition the SW to the SO cluster in the 3C dataset. The mean F_{ST} between clusters inferred by
289 FineSTRUCTURE (Table S1 and S2) are small, confirming subtle differentiation. In both
290 datasets, the strongest differentiation is between the SW cluster and all other regions. These F_{ST}
291 values vary from 0.0016 with the SO cluster to 0.004 with the NW cluster in the 3C dataset and
292 from 0.0009 with the CTR cluster to 0.0019 with the NW cluster in SU.VI.MAX. Finally,

293 besides this subtle division, genetic differentiation within France is also due to isolation by
294 distance as shown by the gradient exhibited on the values of the 1st component of the PCA
295 (Figure S5).

296

297 **Different genetic ancestry profiles that could have been shaped by gene flow barriers**

298 Results obtained by using ADMIXTURE corroborate the FineSTRUCTURE analysis with the
299 SW cluster been the most different from the other groups (Figure S6 and S7). At k=2, the SW
300 cluster shows a light blue component that is significantly less frequent in the other groups
301 (ANOVA *post-hoc* tests, $p\text{-value} < 10^{-6}$) (Figure S8). In the 3C dataset, the proportion of light
302 blue tends to decrease gradually from the south-western (SW) part of France to the centre of
303 France (CTR) to finally remain similar in the north of France (NO, NW and *Others*). In
304 SU.VI.MAX, the proportion of light blue component tends to discriminate the north from the
305 south of France (Figure S8). For k=3, a third major component can be defined, the light green
306 ancestry. In the 3C Study this component is predominant in the north of France (NW and NO
307 clusters) and almost absent in the SW while in SU.VI.MAX this component is predominant in
308 the SE and minimal in the extreme west of France (NW and SW). At k=6, both datasets
309 highlight the differentiation of the SW and the NW cluster from the others clusters.

310 We performed EEMS analysis in order to identify gene flow barriers within France; i.e; areas
311 of low migrations. We varied the number of demes from 150 to 300 demes and selected a grid
312 of 250 demes showing good concordance between datasets (Figure S9). In both datasets, we
313 identified a genetic barrier around the south-west region (Figure 3). This barrier mirrors the
314 first division in the FineSTRUCTURE. The plots also reveal a gene flow barrier around
315 Bretagne in the North-West and along the Loire River, which covers the separation of the North

316 cluster. Finally, another barrier is also present on the South-East side that roughly corresponds
317 to the location of the Alp Mountains at the border with Northern Italy.

318

319 **IBD-derived demographic inferences reveal a rapid expansion over the last 150** 320 **generations**

321 Demographic inferences based on IBD patterns in the two datasets were also very concordant.
322 We observed a very rapid increase of the effective population size, N_e , in the last 150
323 generations (Figure S10). The growth rate was 121% (0.8% per generation) in SUVIMAX and
324 100% (0.7% per generation) in 3C. This is in accordance with previous observations³³ which
325 reports a very rapid increase of human population size in Europe in the 150 last generations.
326 However, the increase of N_e was not constant over time and a decrease of N_e is observed in
327 both datasets in-between generations 12-22 (from $\sim 1\ 300$ to $1\ 700$). The growth rates in the
328 period preceding and the period following this decrease were rather different. These growth
329 rates were respectively of 3.1% and 2.4% per generation in 3C and SU.VI.MAX in the last 12
330 generations and of 0.4% and 0.6% per generation in the first period (22 to 150 generations ago).
331 In-between these two periods, a bottleneck could be detected that could reflect the devastating
332 Black Death. This decrease in N_e seems to affect mainly the Northern part of France (Figure
333 S11). However, this result was not robust to change in parameters: the bottleneck effect was no
334 longer seen when longer IBD chunks were used for N_e estimation (Figure S12).

335

336 **Different contributions of British and Basque heritage in the six French genetic clusters**

337 To study the relationship between the genetic clusters observed in France and neighbor
338 European populations, we combined our two datasets with the 1000G European dataset. As a
339 first step, we run FineSTRUCTURE on the 1000G European populations excluding Finland

340 and found that they could be divided into 3 donor groups as CEU and GBR clustered together
341 (British heritage) (Figure S13). We estimated European ancestry contributions in France with
342 SOURCEFIND and reported the total levels of ancestry proportions for each individual grouped
343 by cluster (Figure 4). We observed similar patterns of admixture between datasets. The
344 proportion of each admixture component from neighboring European countries was
345 significantly different between the six FineSTRUCTURE clusters in both the 3C and
346 SU.VI.MAX datasets (ANOVA, p -value $<10^{-16}$). As expected, the British heritage was more
347 marked in the north than in the south of France where, instead, the contribution from southern
348 Europe was stronger. The overall contribution from British heritage was substantially higher in
349 the NW than in the NO cluster (76% vs 64% in the 3C and 72% vs 63% in SU.VI.MAX). TSI
350 was contributing to the SE cluster while IBS was mainly contributing to the SW cluster, which
351 again was very coherent with the geographic places of birth of individuals. In both dataset, SW
352 had the highest proportions of IBS component. Part of this IBS component could in fact reflect
353 a Basque origin as shown on the PCA plot obtained when combining 3C, SU.VI.MAX and
354 HGDP European dataset (Figure S14). This trend is even more pronounced in the 3C where
355 few individuals are grouped together with Basque individuals in the first three dimensions. This
356 SW region also corresponds to the “Aquitaine” region described by Julius Caesar in his
357 “Commentari de Bello Gallico”¹ (Figure S1).

358

359

360 **DISCUSSION**

361 In this paper, we have studied the genetic structure of France using data from two independent
362 cohorts of individuals born in different regions of France and whose places of birth could be
363 geolocalized. Modern France has a strategic location at the western-most part of Europe and on
364 migration ways from and towards South and North of Europe. Studying its genetic structure is
365 thus of major interest to gain insight on the peopling of Europe. To date however, no exhaustive

366 study had been conducted on the French genetic make-up and our work was intended to fill this
367 gap.

368 The French genomes were found to map at their expected position in between Nordic (British
369 and CEU), Italian and Spanish genomes from 1000 genomes project. Within France,
370 correlations were detected between genetic data and geographical information on the
371 individual's place of birth. Although the relation seems linear – reflecting isolation by distance
372 process – we also observed that, based on genotype patterns, the population can be divided into
373 subgroups, which match geographical regions and are very consistent across the two datasets.

374 An important division separates Northern from Southern France. It may coincide with the von
375 Wartburg line, which divides France into “*Langue d’Oil*” part (influenced by Germanic
376 speaking) and “*Langue d’Oc*” part (closer to Roman speaking) – Figure S15. This border has
377 changed through centuries and our North-South limit is close to the limit as it was estimated in
378 the IXth century^{34,35}. This border also follows the Loire River, which has long been a political
379 and cultural border between kingdoms/counties in the North and in the South (Figure 3).

380 Regions with strong cultural particularities tend to separate. This is for example the case for
381 Aquitaine in the South-West which duchy has long represented a civilization on its own. The
382 Brittany region is also detected as a separate entity in both datasets. This could be explained
383 both by its position at the end of the continent where it forms a peninsula and, by its history
384 since Brittany has been an independent political entity (Kingdom and, later, duchy of *Bretagne*),
385 with stable borders, for a long time³⁶.

386 The extreme South-West regions show the highest differentiation to neighbor clusters. This is
387 particularly strong in 3C dataset, where we even observe an additional cluster. This cluster is
388 likely due to a higher proportion of possibly Basque individuals in 3C, which overlap with
389 HGDP Basque defined individuals. The F_{ST} between the south-west and the other French

390 clusters were markedly higher than the F_{ST} between remaining French clusters. In 3C these
391 values are comparable to what we observed between the Italian and the British heritage clusters
392 ($F_{ST}=0.0035$). Similar trends are observed in SU.VI.MAX even though the level of
393 differentiation with the SW was weaker.

394 We also observe that the broad-scale genetic structure of France strikingly aligns with two
395 major rivers of France “La Garonne” and “La Loire” (Figure 3). At a finer-scale, the “Adour”
396 river partition the SW to the SO cluster in the 3C dataset.

397 While historical, cultural and political borders seem to have shaped the genetic structure of
398 modern-days France, exhibiting visible clusters, the population is quite homogeneous with low
399 F_{ST} values between-clusters ranging from $2 \cdot 10^{-4}$ up to $3 \cdot 10^{-3}$. We find that each cluster is
400 genetically close to the closest neighbor European country, which is in line with a continuous
401 gene flow at the European level. However, we observe that Brittany is substantially closer to
402 British Isles population than North of France, in spite of both being equally geographically
403 close. Migration of Britons in what was at the time Armorica (and is now Brittany) may explain
404 this closeness. These migrations may have been quite constant during centuries although a two
405 waves model is generally assumed. A first wave would have occurred in the Xth century when
406 soldiers from British Isles were sent to Armorica whereas the second wave consisted of Britons
407 escaping the Anglo-Saxon invasions³⁷. Additional analyses, on larger datasets may be required
408 to discriminate between these various models.

409 Studying the evolution of French population size based on genetic data, we observe a very rapid
410 increase in the last generations. This observation is in line with what has been seen in European
411 populations³³. We also observe, in most cases, a depression during a period spanning from 12
412 to 22 generations ago. This may correspond to a period spanning from 1 300 to 1 700. Indeed,
413 this period was characterized by a deep depression in population size due to a long series of
414 plague events. While the population size in kingdom of France was estimated to be 20 Million

415 in 1 348, it dropped down to 12 Million in 1 400, followed by an uneven trajectory to recover
416 the 20 Million at the end of Louis XIVth reign (1 715)³⁸.

417 However, the decrease we observe in the genetic data does seem to affect mainly the Northern
418 part of France, and for instance is mainly observed in the NO cluster. We see no reason for this
419 trend based on historical records (Figure S16) except perhaps the last plague epidemics in 1 666
420 - 1 670 that was limited to the North of France. Alternatively, a more spread population in the
421 South (which is in general hilly or mountainous) may explain a lower impact of these dramatic
422 episodes. Plague is expected to have had a very strong impact on the population demography
423 in the past as some epidemics led to substantial reduction in the population sizes³⁹. However,
424 we could not detect in our data any footprint of the Justinian plague (541-767 PC) although,
425 according to historical records it had a major impact on the population at that time. This may
426 be due to difficulty to estimate population changes in ancient times, deeper than 50-100
427 generations, especially in presence of more recent bottleneck and given our reduced sample
428 sizes in some of the groups and IBD resolution power. We expect that increasing sample size
429 especially for the FineSTRUCTURE subgroups with small sample sizes will help getting more
430 detailed information farther in the past.

431 Identification of genetic structure is important to guide future studies of association both for
432 common, but more importantly, for rare variants⁴⁰. In the near future, interrogating the
433 demographical history of France from genetic data will bring more precise results thanks to
434 whole genome sequencing that, along with new methods, could allow testing formal models of
435 demographic inference.

436

437 **ACKNOWLEDGMENTS**

438 Part of this work was supported by the French National Research Agency (FROGH: ANR-16-
439 CE12-0033) and the European Union via the Marie Skłodowska-Curie actions (PRESTIGE-
440 2017-4-0018).

441 **Conflict of interest:** none

442

443 **REFERENCES**

- 444 1. Caesar CJ: De bello Gallico, Commentarius Primus: *chapter 1, section 1*.
- 445 2. Lazaridis I: The evolutionary history of human populations in Europe. *Curr Opin Genet*
446 *Dev* 2018; **53**: 21-27.
- 447 3. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K *et al*: Ancient
448 human genomes suggest three ancestral populations for present-day Europeans. *Nature* 2014;
449 **513**: 409-413.
- 450 4. Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population
451 structure on large genetic association studies. *Nature genetics* 2004; **36**: 512-517.
- 452 5. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA *et al*:
453 Genetic structure of human populations. *Science* 2002; **298**: 2381-2385.
- 454 6. Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT *et al*: The
455 distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-
456 chromosome data. *American journal of human genetics* 2000; **66**: 979-988.
- 457 7. A haplotype map of the human genome. *Nature* 2005; **437**: 1299-1320.
- 458 8. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A *et al*: Genes mirror
459 geography within Europe. *Nature* 2008; **456**: 98-101.
- 460 9. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E *et al*: Investigation
461 of the fine structure of European populations with applications to disease association studies.
462 *European journal of human genetics : EJHG* 2008; **16**: 1413-1429.
- 463 10. Humphreys K, Grankvist A, Leu M, Hall P, Liu J, Ripatti S *et al*: The genetic structure
464 of the Swedish population. *PLoS One* 2011; **6**: e22547.
- 465 11. Abdellaoui A, Hottenga JJ, de Knijff P, Nivard MG, Xiao X, Scheet P *et al*: Population
466 structure, migration, and diversifying selection in the Netherlands. *European journal of human*
467 *genetics : EJHG* 2013; **21**: 1277-1285.

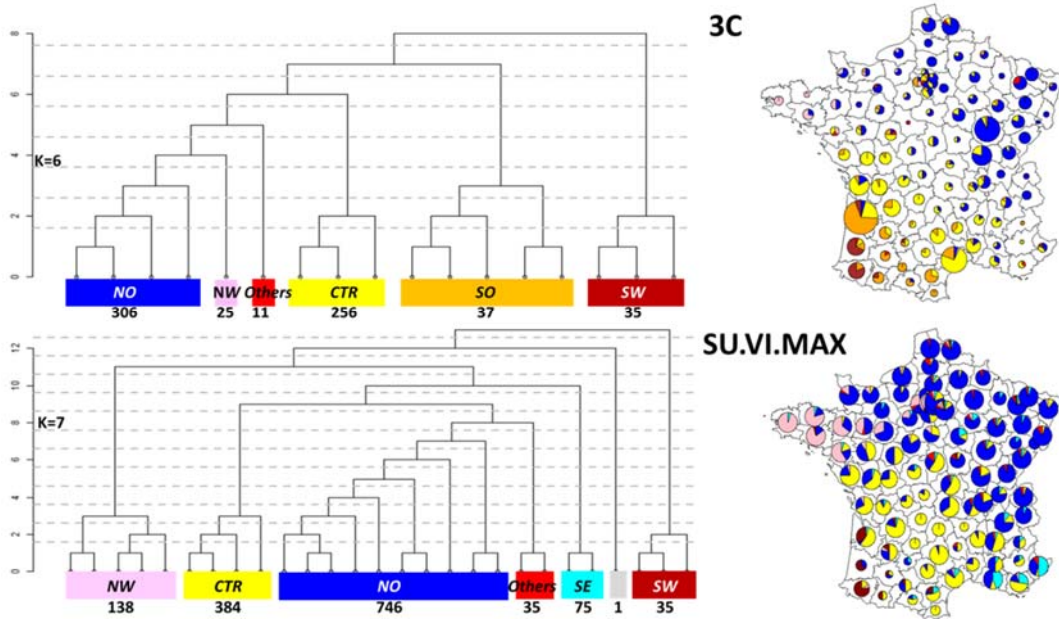
- 468 12. Gilbert E, O'Reilly S, Merrigan M, McGettigan D, Molloy AM, Brody LC *et al*: Author
469 Correction: The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History
470 within Ireland. *Sci Rep* 2018; **8**: 7208.
- 471 13. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T *et al*: The fine-scale
472 genetic structure of the British population. *Nature* 2015; **519**: 309-314.
- 473 14. Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C, Quintela I, Carracedo A, Donnelly P
474 *et al*: Patterns of genetic differentiation and the footprints of historical migrations in the Iberian
475 Peninsula. *Nat Commun* 2019; **10**: 551.
- 476 15. Karakachoff M, Duforet-Frebourg N, Simonet F, Le Scouarnec S, Pellen N, Lecointe S
477 *et al*: Fine-scale human genetic structure in Western France. *European journal of human*
478 *genetics* : *EJHG* 2015; **23**: 831-836.
- 479 16. Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D *et al*: The SU.VI.MAX
480 Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and
481 minerals. *Arch Intern Med* 2004; **164**: 2335-2342.
- 482 17. Vascular factors and risk of dementia: design of the Three-City Study and baseline
483 characteristics of the study population. *Neuroepidemiology* 2003; **22**: 316-325.
- 484 18. Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M *et al*: Genome-wide
485 association study identifies variants at CLU and CR1 associated with Alzheimer's disease.
486 *Nature genetics* 2009; **41**: 1094-1099.
- 487 19. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: Second-generation
488 PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**: 7.
- 489 20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al*: PLINK: a
490 tool set for whole-genome association and population-based linkage analyses. *American*
491 *journal of human genetics* 2007; **81**: 559-575.

- 492 21. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT: Data
493 quality control in genetic case-control association studies. *Nature protocols* 2010; **5**: 1564-
494 1573.
- 495 22. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal
496 components analysis corrects for stratification in genome-wide association studies. *Nature*
497 *genetics* 2006; **38**: 904-909.
- 498 23. Lawson DJ, Hellenthal G, Myers S, Falush D: Inference of population structure using
499 dense haplotype data. *PLoS Genet* 2012; **8**: e1002453.
- 500 24. Delaneau O, Marchini J, Zagury JF: A linear complexity phasing method for thousands
501 of genomes. *Nature methods* 2011; **9**: 179-181.
- 502 25. Alexander DH, Novembre J, Lange K: Fast model-based estimation of ancestry in
503 unrelated individuals. *Genome research* 2009; **19**: 1655-1664.
- 504 26. Petkova D, Novembre J, Stephens M: Visualizing spatial population structure with
505 estimated effective migration surfaces. *Nature genetics* 2016; **48**: 94-100.
- 506 27. Browning SR, Browning BL: Accurate Non-parametric Estimation of Recent Effective
507 Population Size from Segments of Identity by Descent. *American journal of human genetics*
508 2015; **97**: 404-418.
- 509 28. Browning BL, Browning SR: Detecting identity by descent and estimating genotype
510 error rates in sequence data. *American journal of human genetics* 2013; **93**: 840-851.
- 511 29. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV *et al*: Long-range
512 LD can confound genome scans in admixed populations. *American journal of human genetics*
513 2008; **83**: 132-135; author reply 135-139.
- 514 30. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S *et al*:
515 Worldwide human relationships inferred from genome-wide patterns of variation. *Science*
516 2008; **319**: 1100-1104.

- 517 31. Gazal S, Sahbatou M, Babron MC, Genin E, Leutenegger AL: High level of inbreeding
518 in final phase of 1000 Genomes Project. *Sci Rep* 2015; **5**: 17453.
- 519 32. Chacon-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuna-
520 Alonzo V, Barquera R *et al*: Latin Americans show wide-spread Converso ancestry and imprint
521 of local Native ancestry on physical appearance. *Nat Commun* 2018; **9**: 5388.
- 522 33. Keinan A, Clark AG: Recent explosive human population growth has resulted in an
523 excess of rare genetic variants. *Science* 2012; **336**: 740-743.
- 524 34. Wartburg W. von: *Les origines des peuples romans*. Presses Universitaires de France,
525 1941.
- 526 35. Chaurand JD: *Nouvelle histoire de la langue française*. Seuil, 2012.
- 527 36. Leprohon R: *Vie et mort des Bretons sous Louis XIV*. Le Scouezec, 1984.
- 528 37. Fleuriot L: *Les Origines de la Bretagne. L'émigration*. Payot, 1999.
- 529 38. Dupâquier J: *Histoire de la population française Coffret 4 volumes : Volume 1, Des*
530 *origines à la Renaissance. Volume 2, De la Renaissance à 1789. Volume 3, De 1789 à 1914.*
531 *Volume 4, De 1914 à nos jours*. Presses Universitaires de France, 1995.
- 532 39. Biget JL, Bove B, Cornette J: *Le temps de la Guerre de Cent ans (1328-1453)*. Belin,
533 2009.
- 534 40. Persyn E, Redon R, Bellanger L, Dina C: The impact of a fine-scale population
535 stratification on rare variant association test results. *PLoS One* 2018; **13**: e0207677.
- 536
- 537
- 538

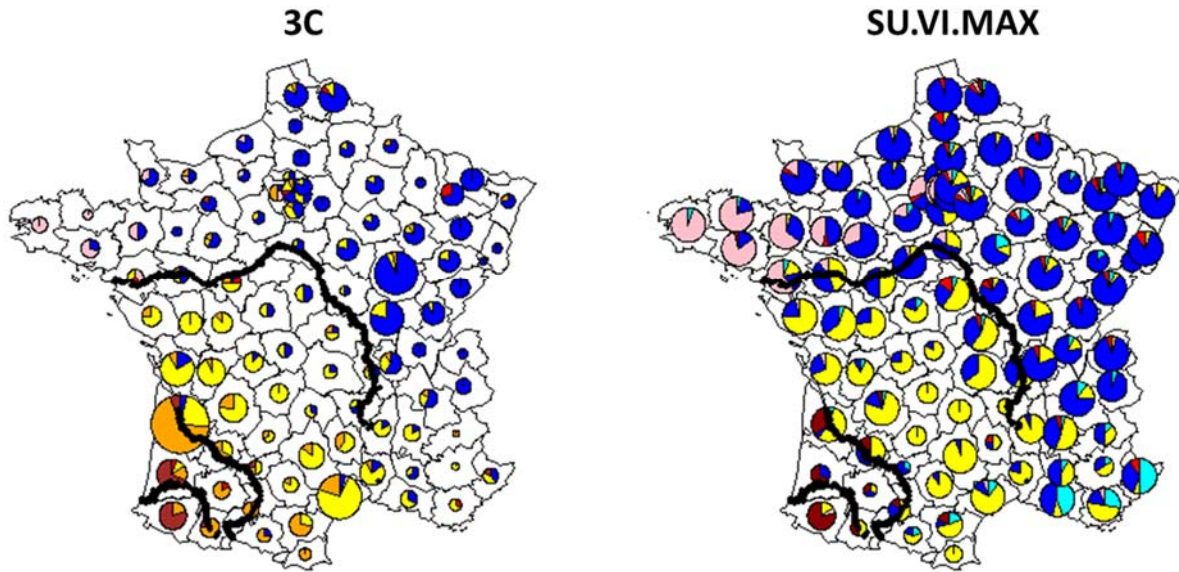
539 **FIGURES**

540
541



543 Figure 1: FineSTRUCTURE clustering of the 3C Study (770 individuals) and SU.VI.MAX (1
544 414 individuals) pooled in 6 and 7 clusters respectively. Left side shows the tree structure and
545 right side shows by *département* pie charts indicating to which of the six clusters the individuals
546 belong to.

547



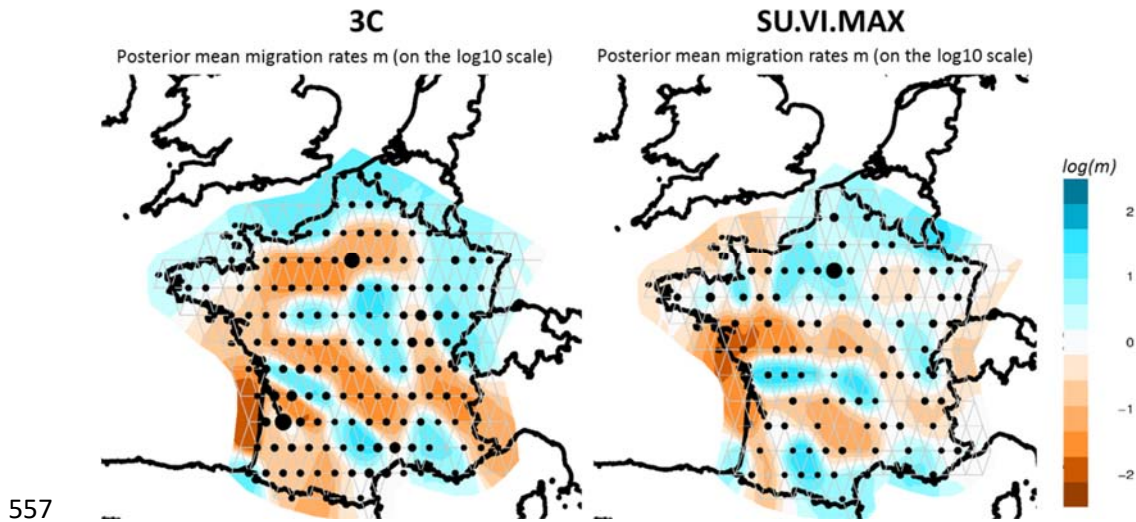
548

549 Figure 2: Pie charts indicating the proportion of individuals from the different “départements”
550 assigned to each cluster. Results are reported for the partition in 6 clusters obtained by running
551 FineSTRUCTURE in the 3C dataset (left) and in SU.VI.MAX (right) independently.
552 Geographic coordinates of three rivers of France are drawn in black: Loire, Garonne and Adour
553 from north to south.

554

555

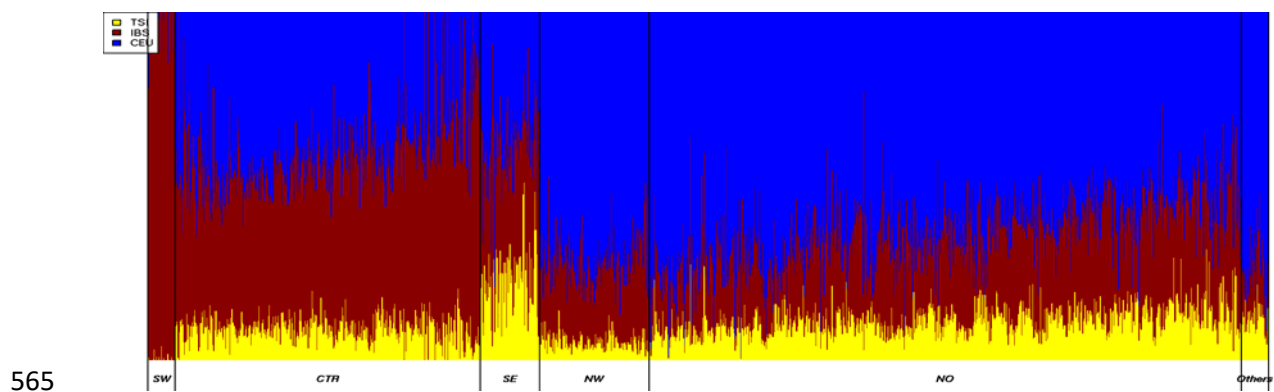
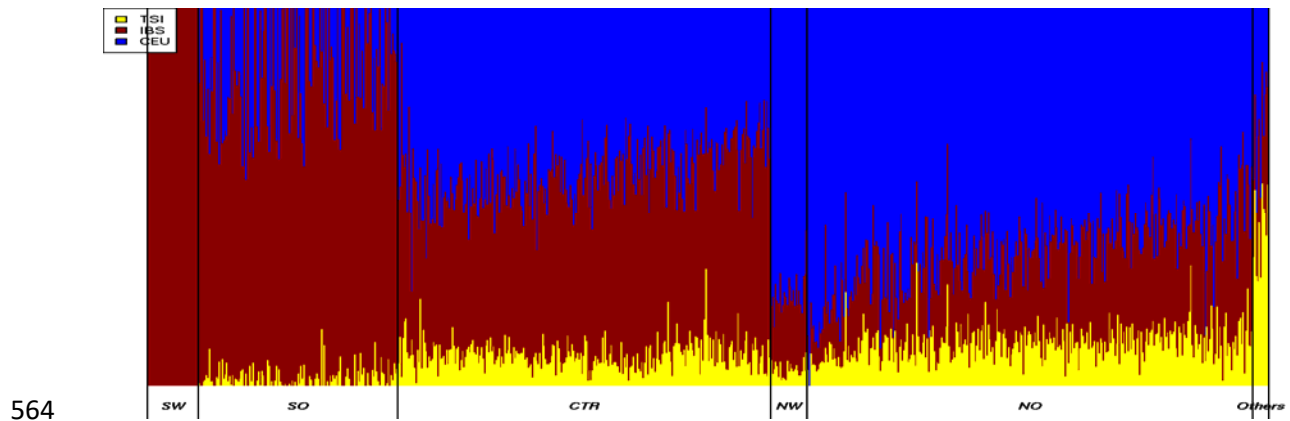
556



558 Figure 3: Estimated effective migration surfaces of France obtained from EEMS on the 3C (left)
559 and SU.VI.MAX (right) datasets. The colour scale reveals low (blue) to high (orange) genetic
560 barriers between populations localized on a grid of 250 demes. Each dot is proportional to the
561 number of populations included.

562

563



566 Figure 4: Ancestry profiles from the three neighbouring European populations inferred by
567 SOURCEFIND in the French 3C (top) and SU.VI.MAX individuals (bottom) datasets. In each
568 cluster, individuals are ordered according to the latitude of their reported birth place.