Genetic architecture of Multiple Sclerosis patients in the French national OFSEP-HD cohort

Julien Paris[1], Nayane S.B. Silva[1], Igor Faddeenkov[1], Martin Morin[1], Léo Boussamet[1], Stanislas Demuth[1,6], Mitra Barzine[1], Anna Serova-Erard[1,4], François Cornelis[1,4,5], Sonia Bourguiba-Hachemi[1], Sophie Limou[1], Romain Casey[2], Bertrand Fontaine[7], Sandra Vukusic[2], Laureline Berthelot[1], David Laplaud[1], Nicolas Vince*[1], Pierre-Antoine Gourraud[1,3]

[1] Nantes Université, CHU Nantes, INSERM, Centre de recherche en Transplantation et Immunologie Translationnelle, UMR 1064, F-44000, Nantes, France.

[2] Université de Lyon, Université Claude Bernard Lyon 1, F-69000 Lyon, France; Hospices Civils de Lyon, Service de Neurologie, sclérose en plaques, pathologies de la myéline et neuro-inflammation, F-69677 Bron, France ; Observatoire Français de la Sclérose en Plaques, Centre de Recherche en Neurosciences de Lyon, INSERM 1028 et CNRS UMR 5292, F-69003 Lyon, France EUGENE DEVIC EDMUS Foundation against multiple sclerosis, state-approved foundation, F-69677 Bron, France

[3] Nantes Université, CHU Nantes, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, INSERM CIC 1413, F-44000, Nantes.

[4] CHU Clermont-Ferrand, Génétique – Oncogénétique Adulte – Prévention (GENOAP), Clermont-Ferrand.

[5] Université Clermont Auvergne, UFR de Médecine et des professions paramédicales, Clermont-Ferrand.

21 [6] *INSERM CIC 1434, Clinical Investigation Center, University Hospital of Strasbourg, Strasbourg,*
22 *France.*

23 [7] *From the Sorbonne Université - Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS,*
24 *APHP, Hôpital Pitié Salpétrière Univ. Hosp., DMU Neuroscience 6; Inst. of Cardiometabolism*
25 *and Nutrition, Sorbonne-universités-Upmc 06, INSERM, CNRS; Laboratoire des Signaux et*
26 *Systèmes (L2S), CNRS-CentraleSupélec, Université Paris-Saclay; Sorbonne Université (B.S.),*
27 *Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital St. Antoine-HUEP;*
28 *and INSERM (B.F.), SU, AP-HP, Centre de recherche en Myologie-UMR974 and Service of Neuro-*
29 *Myology, Institute of Myology, University hospital Pitié-Salpêtriere.*

30 **Correspondence: Nicolas Vince: nicolas.vince@univ-nantes.fr*

31 *N.V and PA.G contributed equally to this work*

32 Abstract

33 Multiple Sclerosis (MS) is a central nervous system (CNS) autoimmune inflammatory
34 disease targeting the myelin sheath and affecting 2.8 million patients worldwide, mostly in
35 economically advanced countries. The OFSEP-HD (French Multiple Sclerosis Registry - High
36 Definition) multi-centric cohort comprises 2,667 genetic samples of patients with MS including
37 5 years of clinical, biological and imaging follow up. Here we described the genetic background
38 of the cohort using data generated from the Affymetrix Precision Medicine Research Array
39 (PMRA) genotyping chips to collect 888,799 genomic variants, and up to 8.5 million variants
40 after imputation. Our analysis focused on genetic ancestry, admixture analysis and Human
41 Leukocyte Antigen (HLA) including haplotypes inference. Principal Components Analysis (PCA)
42 clustering identified seven ancestral clusters with 2177 patients (85.6 %) from clearly defined

European ancestry. We observed 232 MS patients from North-African genetic ancestry while 120 of those patients (51.7%) did not self-report North-African origins, highlighting once again the limitations of self-assessed population descriptors. To promote data sharing we implemented the generation of a realistic and anonymous synthetic dataset using an adaptation of a known synthetic data generation methodology. This work unveils the genetic landscape and heterogeneous profiles of the OFSEP-HD cohort and proposes an open synthetic genetic dataset for further analyses.

Keywords: Multiple sclerosis, France, OFSEP-HD, genetics, genomics, ancestry analysis, genomic proportions, bioinformatics, population genetics, Human Leukocyte Antigens, HLA haplotypes, HLA alleles.

Introduction

Multiple Sclerosis (MS) is a complex autoimmune disorder characterized by the demyelination of nerve fibers in the central nervous system (CNS) yielding to heterogeneous symptomatology including cognitive dysfunction, muscle-related problems, mental health issues, vertigo, vision and sexual problems as well as bladder and bowel dysfunctions. The number of MS patients worldwide was estimated to 2.8 million with a prevalence of 35.9 per 100,000 inhabitants[1,2]. MS predominantly affects women, accounting for approximately 70% of cases demonstrating the significant gender disparity in susceptibility[1,3]. MS is highly prevalent in most economically advanced countries including France, with a prevalence of 197.6 per 100,000 inhabitants, which is more than five times greater than the worldwide average prevalence[4]. This highlights the substantial disease burden to the public health regarding healthcare quality and financial costs[5].

65 MS pathogenesis is caused by the interplay between both environmental and genetical
66 components. The well-known latitude gradient of MS demonstrates lowest prevalence regions
67 near the equator and high prevalence regions outside of the tropics especially in North-
68 Western countries[1]. A South-West to North-East increasing latitude gradient is also observable
69 in metropolitan France, demonstrating the impact of the gradient even at a lower regional
70 scale[4]. This gradient is thought to be primarily driven by the lack of sun exposure and vitamin
71 D deficiency, more prominent as we get further from the equator[6]. Moreover, Epstein-Barr
72 Virus (EBV) and CytoMegaloVirus (CMV) infections during lifetime, smoking and childhood
73 obesity, contribute significantly to MS onset[6–8]. Nevertheless, this latitude gradient might also
74 partly result in healthcare disparities across the world, with lowest prevalence regions bearing
75 fewer data due to a deficit of neurologist and MS specialist, in turn yielding to a systemic bias
76 towards fewer MS diagnostics in those regions[1].

77 MS is a polygenic disease, resulting from both previously stated environmental factors
78 and the cumulative effect of multiple genetic variations that influence susceptibility to the
79 disease, as opposed to Mendelian diseases, which are mostly caused by a single genetic
80 mutation. The most recent GWAS (Genome Wide Association Study) identified a total of 233
81 MS Single Nucleotide Polymorphism (SNP) associated to MS susceptibility, with 32 located in
82 the Major Histocompatibility Complex (MHC) region explaining most of the phenotypic
83 variance, and 1 located on the X chromosome[9,10]. Recent insights also suggest the role of
84 genetics on disease severity and progression with one significant variant discovered to date[11].

85 The French population genetic architecture originates from two major ancestral DNA.
86 Early Neolithic farmers migrated to South-Eastern France from Anatolia, while steppe
87 pastoralist arrived from the northern plains of Eurasia. Although Neolithic farmers DNA is

predominant in modern French population, the genomic contribution of steppe pastoralists is more pronounced in Northern parts of France[12]. Over centuries, historical cultural and political borders have shaped the genetic architecture of today's French population. Indeed, genetic clusters in France mostly align with neighboring countries, reflecting shared histories and interactions[12].

Moreover, France's strategic location on the western edge of Europe has made it a privileged migration route between the South and North of the continent, further enriching its genetic diversity. Particularly, the history of migration between North Africa and France is well-documented, shaped by deep historical, cultural, and economic ties[13]. This gene flow contributed to human genetic diversity in southern Europe, including France[14].

However, studying genetic ancestry in France is historically sensitive, and faces legal constraints explaining the few genetic studies conducted on French population with genome wide data[12,15]. Additionally, most patients in genetic studies focusing on polygenic diseases like MS are from European genetic ancestry[16,17]. Nevertheless, initiatives like the French Multiple Sclerosis Registry – High Definition Cohort (OFSEP-HD) and its 2,667 samples with genetic data, is an opportunity to study MS genetics with high quality nation-wide data collection[18]. While the OFSEP-HD cohort is instrumental in numerous published MS studies, its genetic architecture has not been studied yet[19,20]. This study provides the first genetic analysis of the OFSEP-HD population, highlighting the role of ancestry and MS variants including MHC variations for future research into genetic factors influencing MS within the French population.

110 ## Materials and methods

111 ### Data collection

112 The OFSEP-HD observational study, totalizing 2,667 genetic samples of MS patients
113 from the historical OFSEP study collected from various centers distributed across France,
114 represents an ongoing cohort study initiated in 2018 with annual high quality follow-up[18,20].
115 All OFSEP-HD patients fulfilled MacDonald2017+ diagnosis criteria at inclusion[21,22]. OFSEP
116 authorization for collecting clinical data and biological samples was granted by the CNIL
117 (Commission Nationale de l'Informatique et des Libertés) in 2015 under CPP (Comité de
118 Protection des Personnes) number 914066. Appropriate informed consent was obtained from
119 all OFSEP-HD patients[22]. The cohort includes different types of demographic variables
120 including patients' and parents' reported origins, sex, birthdate, Expanded-Disability-Status-
121 Scale (EDSS) (non-linear discretized scale measuring the level of disability), number of MRI
122 demyelinating lesions, pregnancies, relapses, treatments and typ e of MS course at inclusion
123 in addition to annual MRI and sample biocollection[22–24]. Data oversight and management are
124 under the responsibility of the EDMUS Foundation[22].

125 Genetic data were acquired through genotyping conducted at the Curie Institute's
126 Genotyping platform in Paris, utilizing the Affymetrix Precision Medicine Research Array
127 (PMRA) chip, representing a total of 888,799 genotyped variants.

128 The OFSEP-HD cohort was compared to a subset of 2,405 reference individuals from
129 the 1000 Genomes Project population, comprising individuals from East-Asian, South-Asian,
130 American, European, and African ancestries[25]. In addition, a North African reference cohort of

131 154 non MS individuals from Arauna et al. was utilized to accurately characterize genetic
132 ancestry in the light of its 486,252 genotyped variants[26,27].

133 **Processing, imputation, and quality controls of genotyping data**

134 Quality control (QC) procedures were applied to OFSEP-HD dataset and North-African
135 reference dataset. No additional quality control or imputation was performed on the 1000
136 Genome reference dataset[28]. A technological QC on OFSEP-HD dataset using Axiom Analysis
137 Suite 2.0 was performed to select high-quality samples and genetic variants. This QC
138 procedure led to the selection of 869,082 variants (97.8 %).

139 A second QC filtering procedure was implemented using Python 3.11 in conjunction
140 with PLINK 1.9/2 and bcftools 1.13 following previously published guidelines[28]. Pre-imputation
141 filters and thresholds applied to OFSEP-HD genetic samples are listed below:

142 1. Individual level and variant level missingness (> 2%)
143 2. Sex-discrepancies (Male: F > 0.8; Female: F ≤ 0.4)
144 3. Exclusion of X/Y/mitochondrial chromosomes
145 4. Minor Allele Frequency (MAF < 1%)
146 5. Hardy-Weinberg Equilibrium (HWE < $1e^{-10}$)
147 6. Relatedness ($\hat{p}$ > 0.2)

148 Details about processing steps of North-African reference dataset are available in Table
149 S1 and a flowchart of the QC procedures in Figure S1[26].

150 Pre-imputation quality controls of OFSEP-HD population resulted in 424,132 remaining
151 SNPs and 2,542 patients. The filters removing most of the variants were the MAF filters and

152 sexual and mitochondrial chromosome variants exclusion. Imputation of OFSEP-HD genetic
153 data was then performed using Topmed Michigan Imputation Server
154 (https://imputation.biodatacataslyst.nhlbi.nih.gov/) to maximize overlap with reference
155 datasets during ancestry analysis. Post-imputation QCs were also performed on OFSEP-HD
156 dataset, excluding variants with insufficient imputation quality ($r^2 < 0.5$) or low frequency
157 (MAF < 10%). Moreover, we filtered again for variants considerably deviating from Hardy-
158 Weinberg Equilibrium (HWE < $1e^{-10}$).

159 **Ancestry analysis**

160 We performed a Principal Component Analysis (PCA) of the 2,542 OFSEP-HD patients
161 with all reference populations using the Eigensoft software[29]. The analysis considered the
162 twenty first principal components obtained from the 9,326 common independent SNPs to
163 OFSEP-HD and reference datasets. A supervised K-Means clustering approach was then
164 implemented to identify ancestry superpopulation clusters on the PCA[30,31]. We described the
165 methodology used to define the best number of clusters k in Figure S2.

166 **Demographic and clinical description: self-reported origins of OFSEP-HD patients**

167 We analyzed demographic and clinical variables of the 2542 remaining patients after
168 quality controls and processing steps providing key insights into the composition of the cohort.
169 We computed descriptive statistics to better understand the self-reported origins distribution
170 of OFSEP-HD patients. The origins of participants in the OFSEP-HD study were self-reported
171 through a Case Report Form (CRF) that included three questions, each asking: "Which
172 geographical region are you originating from?" The participants answered all questions,
173 providing information about both their own and their parents' geographical origins.

174 **Reference populations selection and admixture proportions of European, African**
175 **and East-Asian genome**

176 Admixture analysis was performed using admixture 1.3.0 software. The reference
177 parental populations for admixture analysis were selected from the 1,000 Genome dataset to
178 include the 3 most discretizing ancestries: European, African, and East-Asian, thereby
179 excluding North-African, South-Asian and American admixed ancestries[27,32,33]. More
180 methodological details are available in Supplemental Materials.

181 **Self-reported origins and genetic ancestry: a comparative statistical analysis**

182 Patients' and parents' origins were compared to genetically inferred ancestral
183 backgrounds. Self-reported origins of OFSEP-HD patients and their parents were collected
184 through the previously mentioned CRF. Mapping between ancestry clusters and self-reported
185 origins is reported in Table S2 and the methodology used to select clusters for statistical
186 analyses is discussed in Figures S3 - 4 and Tables S3 – 4. Proportion of discordant pairs and
187 Cohen's kappa score were used as a measure of agreement between patients' self-reported
188 origins and genetic ancestry. Proportion of discordant pairs were calculated as $\frac{\mathrm{b+c}}{\mathrm{a+b+c+d}} * 100$**,**
189 with a, b, c and d being respectively the upper left, upper right, lower left and lower right cells
190 of a 2x2 contingency table. Margin of Errors (95%) ($MoE_{95\%}$) were computed as $\pm 1.96 * \frac{\sigma}{\sqrt{n}}$ with
191 σ the standard deviation and n the sample size[34]. Statistical significance of comparisons were
192 assessed using two-sided statistical association tests from independent data
193 arrangements[35,36]. Additionally, we performed a quantitative statistical analysis comparing
194 European and African genome proportions within genetic clusters regarding patients self-
195 reported origins status. Further details about the methodology are described in Supplemental
196 Materials.

197 **HLA imputation: alleles and haplotypes**

198 Human Leukocytes Antigens (HLA) genotypes were imputed from genotyped SNPs
199 through SNP-HLA Reference Consortium (SHLARC), a widely recognized and extensively used
200 resource for imputing *HLA* alleles from SNP data, employing the HIBAG R package and a large
201 multi-ethnic reference panel[37,38]. Further, we performed individual-level haplotypes
202 imputation using HLA2Haplo tool from Easy-HLA software suite independently for each
203 ancestry cluster using corresponding reference panels (easyhla.univ-nantes.fr)[39]. This tool
204 inferred the phase of imputed HLA alleles for each considered OFSEP-HD individual. Imputed
205 haplotypes include loci *HLA-A, HLA-B, HLA-C, HLA-DRB1,* and *HLA-DQB1*.

206 **Polygenic Risk Scores computation**

207 Polygenic Risk Scores (PRS) provides a summary metric measure of genetic
208 susceptibility risk of an individual to a given genetic-related disease. Here we computed PRS
209 designed specifically for MS, in the OFSEP-HD population[40,41]. We had 210 MS susceptibility
210 variants (including 20 in the MHC region) out of a total of 233 known MS associated
211 susceptibility variations available in our dataset after SNP imputation, filtering (MAF < 1%) and
212 *HLA* alleles imputation (Table S5)[9].

213 We compared a log-additive PRS model to a non-weighted Sum of Risk Alleles Count
214 (SRAC) approach using Pearson's correlation coefficient r^2 computed for the global OFSEP-HD
215 population as well as for specific ancestry clusters[40–42]. Since MS susceptibility variants were
216 mostly found in patients from European genetic ancestry, we compared distributions of SRAC
217 and log-additive models between European and North-African genetic ancestries using Mann-
218 Whitney U tests[9,41,43]. Furthermore, we assessed the phenotypic variance explained by the 20

219 available MHC region genetic variations compared to the 190 non-MHC SNPs at the whole
220 cohort level using Genome-wide Complex Trait Analysis (GCTA) software[44].

221 **Synthetic data generation**

222 To promote data sharing and reusability of sensitive genetic data, we provided an
223 anonymous – synthetic – GDPR-enforced dataset as an open resource using an adaptation of
224 an already existing "patient-centric" local modelling strategy[45]. The methodology focuses on
225 addressing privacy risks like singling out, linkability, and inference using metrics such as
226 Hidden Rate (HR), Local Cloaking (LC), and Distance To Closest synthetic record (DTC). Signal
227 retention was evaluated using Hellinger distances, correlation differences, and Gower's
228 distance permutation test[46]. Optimal parameters were identified as k = 2 and 20 principal
229 components, with regularization for rare modalities. The privacy-fidelity balance was
230 improved by generating three parallel synthetic datasets and averaging features patient-wise.
231 Post-generation processing ensured biological constraints and recalculated derived features.
232 Methodological details are available in Supplemental Materials and the synthetic dataset on
233 GitHub (https://github.com/jp3142/OFSEP_HD_public_avatar_dataset).

234 ## Results

235 **Ancestry analysis: Principal Components Analysis of OFSEP-HD MS patients**

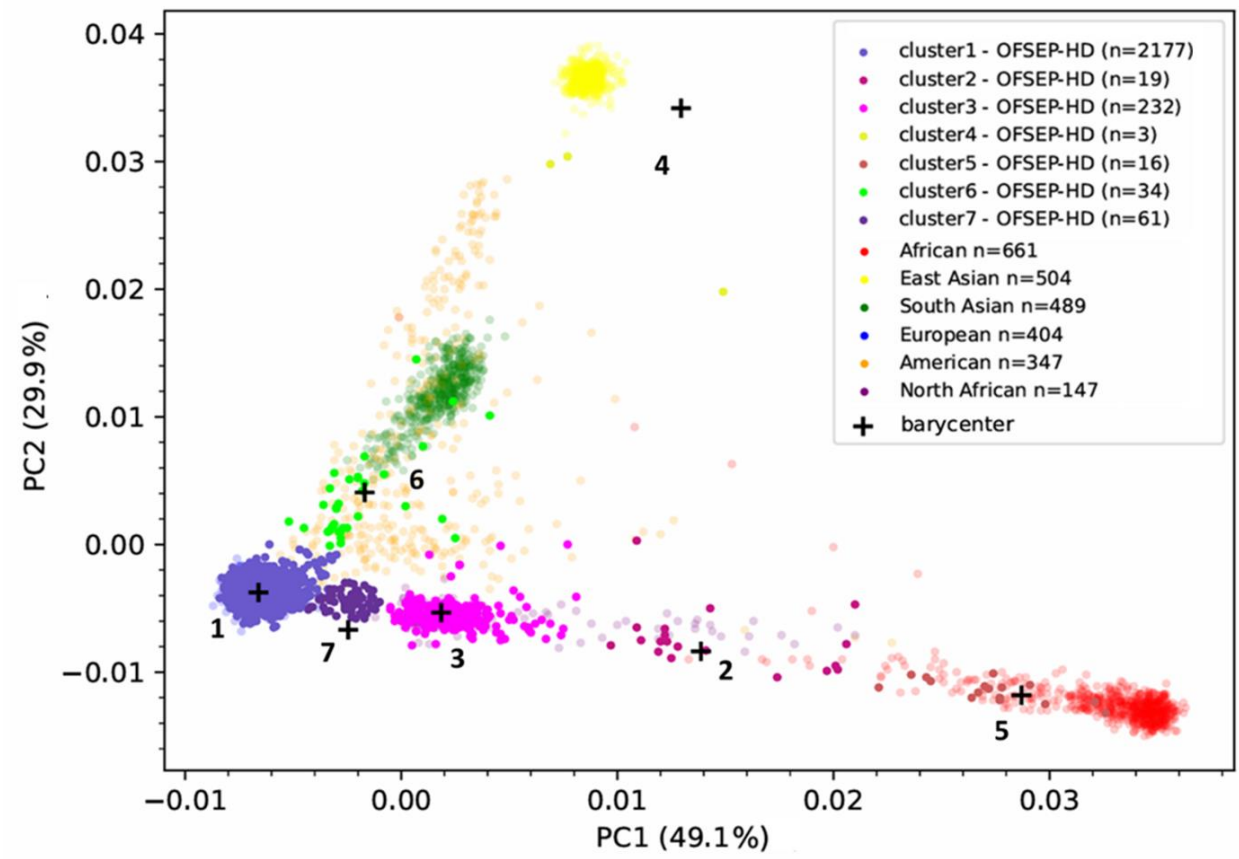
236 Figure 1 shows the Principal Component Analysis (PCA) of OFSEP-HD MS patients
237 compared to the 2 reference populations: 1,000 Genomes and North-African cohort[26]. It
238 highlights the genetic diversity of the MS French population, with a specific focus on the first

239 two principal components explaining respectively 49.1% and 29.9% of the variance (79%). The

240 10 most contributing SNPs for each principal component a

241 re listed in Table S6. The best clustering of the OFSEP-HD population was obtained

242 using 7 distinct clusters.



243

244 ***Figure 1:*** *Principal Component Analysis (PCA) plot of the OFSEP-HD population based on the first two principal components*

245 *(PC1 and PC2), overlaid with reference datasets. Each data point represents an individual. Semi-transparent data points*

246 *represent reference individuals and opaque data points represent OFSEP-HD MS patients. Cluster 1: European. Cluster 2: Afro-*

247 *Caribbean. Cluster 3: North-African. Cluster 4: East-Asian. Cluster 5: African. Cluster 6: South-Asian. Cluster 7: Admixed North-*

248 *African and European.*

249 Findings highlight a major proportion of patients with European genetic ancestry

250 within the OFSEP-HD cohort, totalizing 2,238 MS patients (88 %) assigned to clusters of

251 European ancestry, clusters 1 (n=2,177) and 7 (n=61). Clusters 2 (n=19), 3 (n=232) and 5 (n=16)

account for 267 MS patients (10.5%) with higher African ancestry. Clusters 4 (n=3) and 6 (n=34) totalize 37 MS patients (1.4%) with higher Asian ancestral background.

Self-reported origins of OFSEP-HD patients

Table 1 shows 83 % of patients who reported metropole France origins, while only 71% of fathers and 74% mothers self-reported such origins. Self-reported North-African origins account for 5% of patients and around 10% of their parents. About 6% of patients reported other European origins. The least represented origins are French overseas territories, Asian and Sub-Saharan African origins. Ninety-eight patients did not answer the question or did not know about their origins, and 15 patients reported to be originating from other parts of the world. Details about other demographic and clinical data are showed in Tables S7 – 9.

	Origin count	Origin %	Father's origin count	Father's origin %	Mother's origin count	Mother's origin %
France metropole	2,110	83.01	1,812	71.28	1,889	74.31
French Overseas Territories	14	0.55	19	0.75	16	0.63
Europe (besides France)	156	6.14	238	9.36	231	9.09
North-Africa	127	5.00	272	10.70	233	9.17
Sub-Saharan African	10	0.39	23	0.90	20	0.79
Asia	12	0.47	24	0.94	22	0.87
Others	15	0.59	27	1.06	30	1.18
Unknown	98	3.86	127	5.00	101	3.97

Table 1: *Self-reported origins of 2,542 remaining OFSEP-HD patients after quality controls and processing steps. Count variables correspond to number of patients, and the % variables correspond to proportion in the OFSEP-HD population subset.*

Admixture proportions of European, African and East-Asian genome in OFSEP-HD cohort

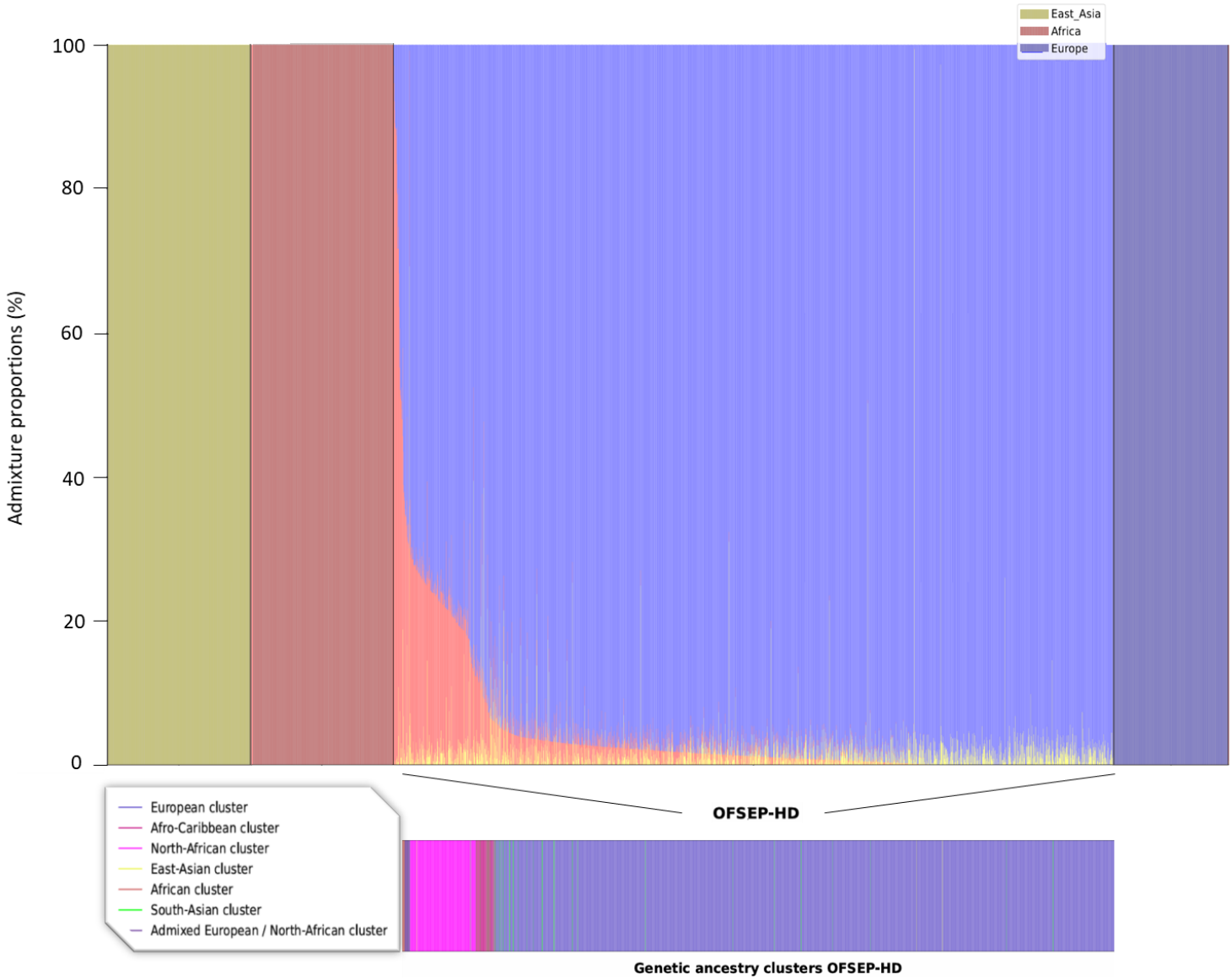


Figure 2: *Admixture plot showing whole-genome admixture proportions of African (red), East Asian (yellow), and European (blue) ancestry for each of the 2,542 OFSEP-HD patients, sorted by African genome proportion. Reference parental populations are shown on both sides of the OFSEP-HD section. Below the plot, genetic clusters identified through PCA are mapped for each individual, providing a comprehensive view of genetic diversity within the cohort.*

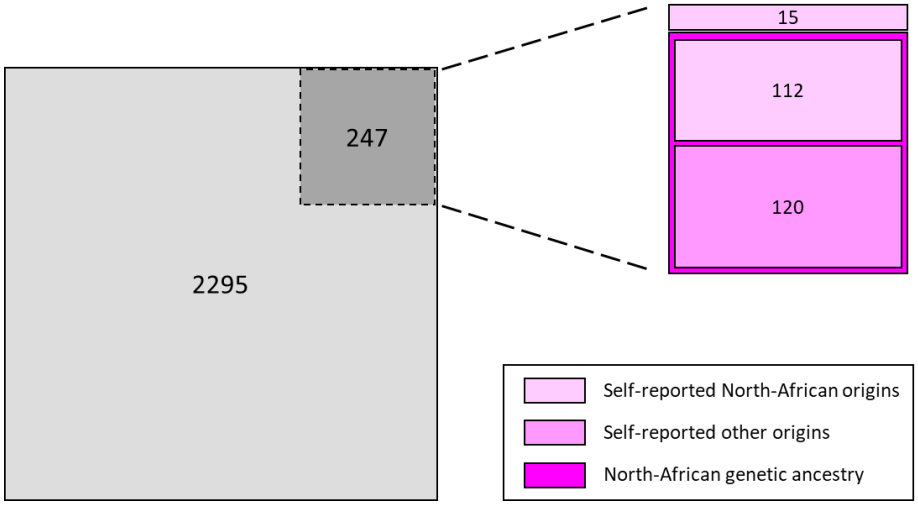
We observed a major European genome proportion in the French MS population with an average above 93.7 % (Figure 2). There is also a substantial number of OFSEP-HD patients (12.3 %) exhibiting more than 10% of African genome (Figure 2 and Table S10). Moreover,

273 most MS patients showing significant African genome ancestry are estimated as part of North-
274 African (cluster 3; n=232) genetic cluster. Distributions of genome proportions among cluster
275 3 is composed in majority by European genetic components with an average of 75.5 %
276 compared to the African genome proportions ranging from 16.7% to 38.9%. Moreover, we
277 found a non-neglectable East-Asian genome proportion for some MS patients, with 101 (< 4%)
278 with at least 5% of East-Asian genome ancestry and 2 patients with more than 85% of East-
279 Asian ancestry (Figure2, Table S10).

280 **An underestimation of North-African self-reported origins among patients from**
281 **North-African genetic ancestry**

282 We focused on qualitative analysis of North-African origins as it represents the second
283 largest genetic group in the OFSEP-HD cohort, after European ancestry. Figure 3 illustrates the
284 distribution of North-African patients based on their self-reported origins and genetic ancestry
285 among the OFSEP-HD population.

286
287
288
289
290
291

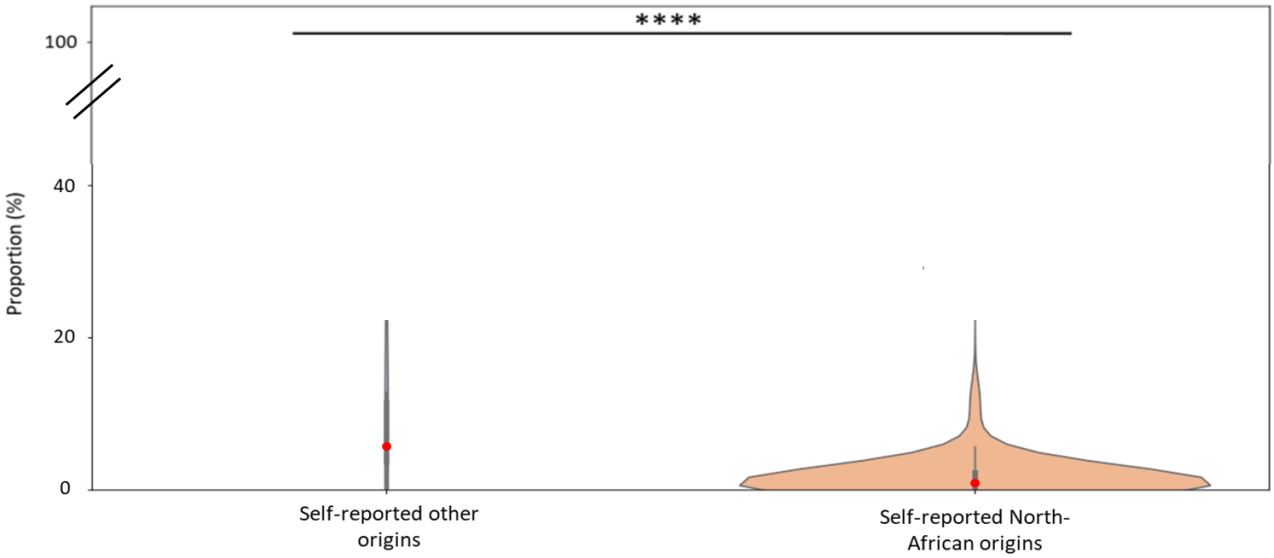


292 ***Figure 3**: Distribution of OFSEP-HD North-African patients by self-reported origins and genetic ancestry. 247 patients self-*
293 *reported North-African origins and 232 of them are from North-African genetic ancestry. The numbers within each section*
294 *correspond to the sample size.*

295 Cohen's Kappa metric (κ = 0.598 ± 0.001) and the low proportion of discordant pairs
296 (5.31 %) illustrate a moderate to substantial agreement between self-reported North-African
297 origins and genetic North-African ancestry (Table S11). Remarkably, among the 127 MS
298 patients who self-reported North African genetic ancestry, 15 patients (11.8%) have a genome
299 that does not reveal marked North-African genetic background (Figure 3). Moreover, 51.7%
300 of patients from North-African genetic ancestry self-identified as from other origins (Figure 3).
301 Contingency tables as well as further analyses about North-African patients are presented in
302 Table S11 and Figure S5. Likewise, we detected a strong underestimation of self-reported Sub-
303 Saharan-African ($p < 0.001$) and Afro-Caribbean origins ($p < 0.05$) compared to their respective
304 PCA-defined ancestry clusters (Figures S6 – 10; Tables S12 – 15).

305 **North-African patients and admixed genome proportions**

306 Among individuals who do not have North African genetic ancestry, we found no
307 significant differences in the proportions of African and European genomic components when
308 comparing those who self-reported North African origins with those who did not report such
309 origins (Table S16), confirming the consistency between PCA-defined ancestry clusters and
310 ancestral genomic composition. Thus, we focus here on patients who are not from North-
311 African genetic ancestry. Figure 4 illustrates African genome proportions of those patients
312 from other genetic ancestries according to their self-reported North-African origins status.



313

314 **Figure 4:** Violin plots showing distribution of African genome proportion of patients not from North-African genetic ancestry
315 regarding self-reported origins status. Boxplots outer borders within each violin plots represent 25th and 75th percentiles and
316 the red dots represent the means. (****p < 1e-4).

317 We found that self-reported North-African patients from other genetic ancestry have
318 a statistically significant higher proportion (Mann-Whitney U test: p < 1e-4) of African genome
319 compared to patients who did not report such origins (Figure 4; Table S16). Nevertheless, the
320 high variability associated to African genome proportions ($CI_{95\%}$ [3.01%; 28.65%]) and
321 consequently European genome proportions ($CI_{95\%}$ [70.44%; 95.78%]; Table S17) of self-
322 reported North-African patients from other genetic ancestry demonstrates the admixed
323 nature of this subpopulation. Further details on performed analyses and European genome
324 proportions of OFSEP-HD North-African MS patients are discussed in Figures S11 - 18. Details
325 and analyses of other genetic ancestry clusters and origins are available in Tables S18 – 29.

Parents' North-African origins: a better proxy of patients' genetic ancestry?

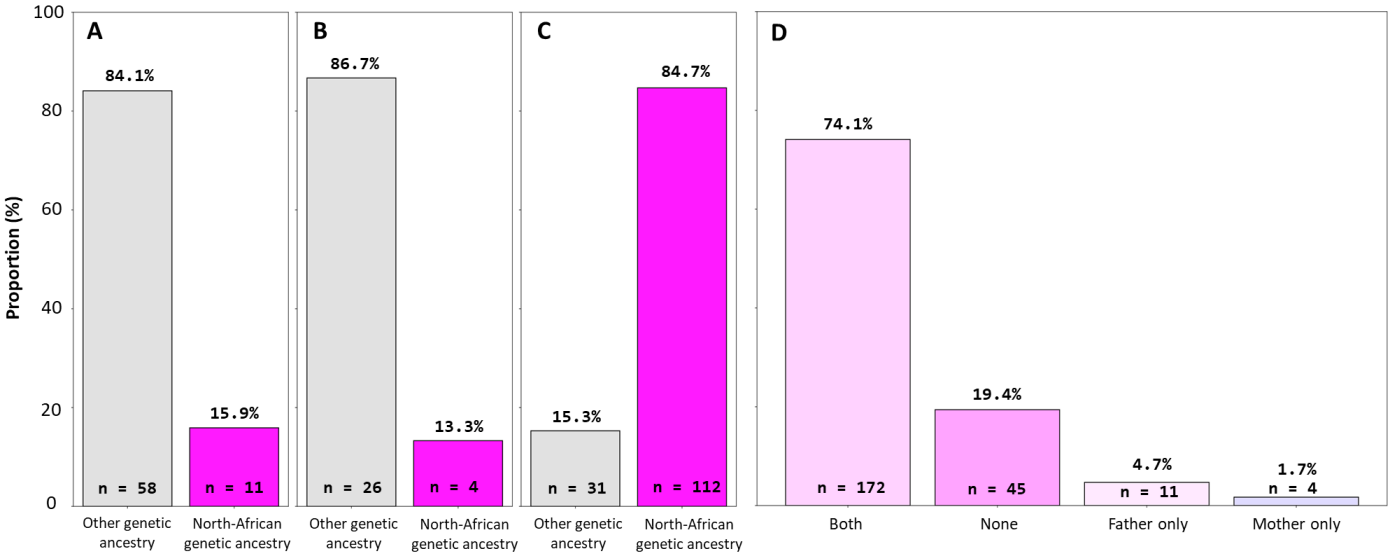


***Figure 5**: (A) North-African genetic ancestry status among patients with only the father reported as from North-African origins. (B) North-African genetic ancestry status among patients with only the mother reported as from North-African origins. (C) North-African genetic ancestry status among patients with both parents reported as from North-African origins. (D) Parents' North-African origins status of patients from North-African genetic ancestry (Both: both parents were reported as from North-African origins, None: no parents were reported as from North-African origins, Father only: the father is the only parent reported as from North-African origins, Mother only: the mother is the only parent reported as from North-African origins).*

We detected a significant association between parents' origins and children estimated genetic ancestry ($p < 1e-4$). However, we found no significant statistical association ($p = 0.77$) between parents' North-African origins and their child's genetic ancestry when only one parent reported North-African origins, demonstrated by high discordance for both fathers (84.1%) and mothers (86.7%) (Figure 5A and 5B; Table S30). Nevertheless, we observed a substantial agreement with patients' genetic ancestry when both parents have common origins (84.7%, Figure 5C) also highlighted by a higher Cohen's Kappa value ($\kappa = 0.64$), and a lower proportion of discordant pairs (3.11%) (Tables S31 – 34), compared to what was obtained previously with patients' self-reported origins (Figure 3). Additionally, 74.1 % of MS

343 patients with North-African genetic ancestry reported both parents as from North-African
344 origins (Figure 5D). Analysis of other parents' origins and genetic ancestries are available in
345 Tables S35 – 38.

346 **HLA alleles imputation**

347 *HLA* alleles imputation was performed on classical *HLA* genes *HLA-A, HLA-B, HLA-C,*
348 *HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1* and *HLA-DRB1* and is particularly good for all
349 genes. The average imputation post-probabilities ranging from 86.5% for *HLA-DRB1* to 98.9%
350 for *HLA-DPA1*, indicate trustful results across all considered *HLA* class I and II genes. In
351 addition, we found 50.3% OFSEP-HD patients presenting at least one *HLA-DRB1*15:01* allele,
352 which is consistent with the literature[47]. Additional details about imputation quality
353 distribution statistics are shown in Figure S19 and Table S39.

354

355 **HLA haplotypes imputation**

356 *HLA* haplotypes imputation from imputed *HLA* alleles *HLA-A, HLA-B, HLA-C, HLA-DRB1*
357 and *HLA-DQB1* achieved very high accuracy with an average post-probability of 82.9 % across
358 all haplotypes collectively. Figure S20 and Table S40 show distribution of post-probabilities by
359 HLA locus. We summarize HLA diversity in the context of OFSEP-HD MS cohort in Table 3.

Rank	Haplotype	Frequency
1	**A*02:01~**B*07:02~C*07:02~**DRB1*15:01~**DQB1*06:02	1.91 %
2	**A*02:01~**B***44:02~**C*05:01**~DRB1*13:01~**DQB1*06:03	0.43 %
3	**A*02:01~B*44:02~**C*05:01~DRB1*04:01**~DQB1*03:01**	0.39 %
4	**A*02:0**1**~B*44:02~**C*05:01**~DRB1*15:01~**DQB1*06:02	0.37 %
5	**A*02:01~**B*15:01~C*03:04~DRB1*04:01**~DQB1*03:02**	0.31 %
6	**A*02:01~**B*08:01~C*07:01**~DRB1*03:01~**DQB1*02:01	0.30 %
7	**A*02:01~**B*18:01~C*07:01~DRB1*11:04**~DQB1*03:01**	0.28 %
8	**A*02:01~**B*18:01~C*05:01**~DRB1*03:01~**DQB1*02:01	0.26 %
9	**A*02:01~**B*40:01~C*03:04**~DRB1*08:01~**DQB1*04:02	0.26 %
10	**A*02:01~**B*40:01~C*03:04**~DRB1*15:01~**DQB1*06:02	0.24 %

360 ***Table 3:*** *Description of the 10 most frequent imputed haplotypes including at least one MS associated HLA allele within the*
361 *OFSEP-HD cohort. Bold alleles within haplotypes represent known MS associated HLA alleles*[9,48].

362 The ten most frequent haplotypes carry HLA alleles associated to MS. Indeed, they
363 present the protective allele *HLA-A*02:01*, but also the well-known *HLA-DRB1*15:01* MS risk
364 allele. The most frequent haplotype is ***A*02:01~**B*07:02**~**C*07:02**~DRB1*15:01~**DQB1*06:02*
365 (1.91%) and contain both previously mentioned MS associated HLA alleles.

366

367 **Computing polygenic risk scores of French MS populations**

368 We found an average log-additive score of 20.29 ± 0.03 for 190 non-MHC SNPs,
369 representing 201 risk alleles per patient. Including the 20 available MHC variations increased
370 the score to 23.38 ± 0.05 for 214 risk alleles in average. The log-additive model and SRAC
371 showed high correlation ($r^2 = 0.91$) for non-MHC SNPs, but correlation decreased substantially
372 ($r^2 = 0.54$) when MHC variants were included (Figure S21). Moreover, MHC variants increased
373 the log-additive score by 15.23% while only increasing the average number of risk alleles by
374 6.46%, highlighting their larger effect on MS susceptibility. PRS distributions differed
375 significantly between MS patients from North-African and European ancestries for all PRS
376 models ($p < 1e-6$). PRS scores distribution statistic details are available in Table S41 and Figure
377 S22.

378 **Open-access synthetic data generation: a balance between privacy and fidelity**

379 Using our synthetic data generation process, we created a highly anonymized dataset
380 as indicated by key metrics following recommended thresholds: HR = 99.49% (> 90%); median
381 LC = 50 (> 5), with DTC = 1.8 (> 0.2) and CDR = 0.99 (> 0.3)[45]. Signal retention metrics also show
382 high synthetic dataset fidelity to original data with Hellinger distance distributions difference
383 of 0.04 (< 0.1) and distribution correlation difference of 1.92% (< 10%). Univariate statistical
384 analysis of generated and original datasets indicated four features (rs58394161, rs11809700,
385 rs11578655, rs10801908) out of the 82 generated features where distributions in synthetic
386 data were significantly deviating from original data distributions after Bonferroni correction
387 for multiple testing ($p < 0.05$). Gower's distance permutation test showed no statistically
388 significant differences from what would be expected by chance between generated and

original datasets (p = 0.38)[49,50]. Figure 6 shows datasets overlap on the two first dimensions of the reduced original Euclidean space.

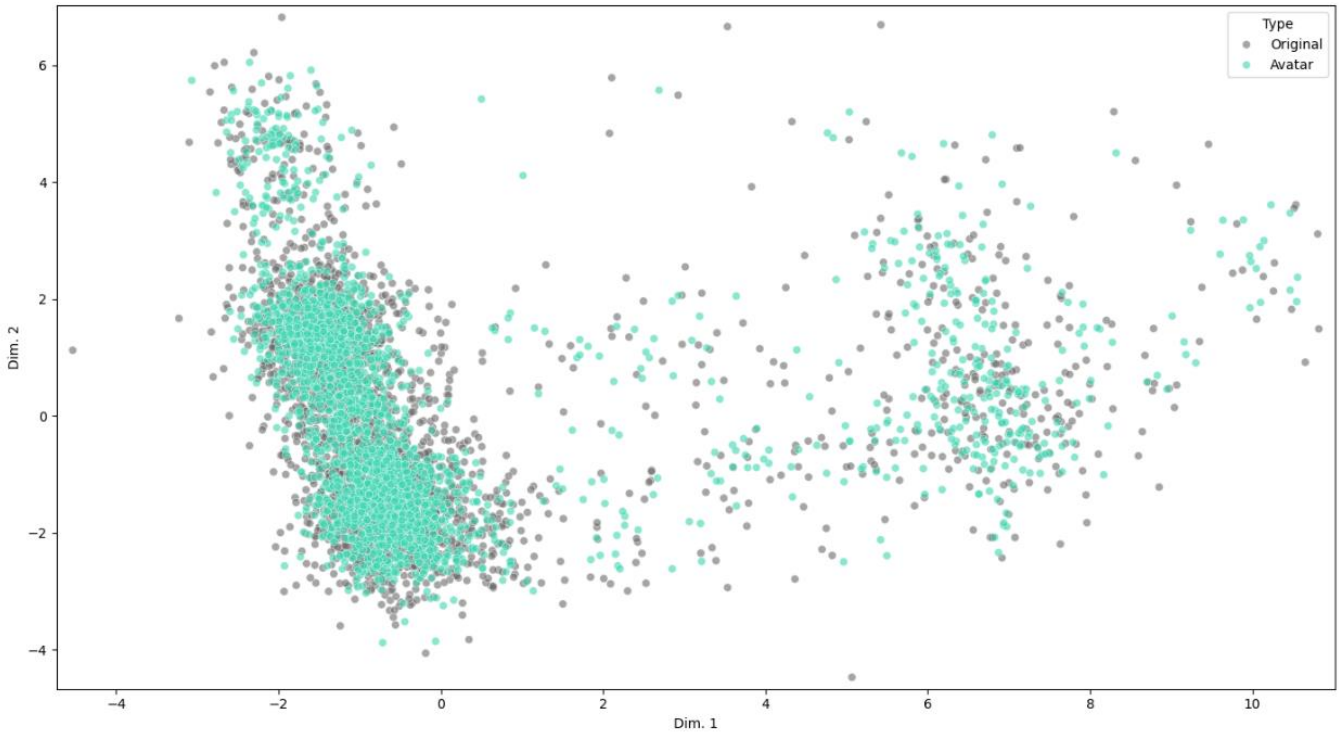


Figure 6: *Comparative projection of original and synthetic dataset structures in original Euclidean space from Octopize's library automatic report generation.*

Discussion

The OFSEP-HD study provides valuable insights into the genetic landscape of the French MS population, revealing significant variations of ancestral origins of MS patients in France. We highlighted a significant proportion (232 patients; 9.1%) of MS French patients from North-African genetic ancestry, illustrating the substantial North-African ancestral background of the French population.

Ancestry analysis and clustering exclusively focused on the two first components of the PCA omitting 21% of the total variance. However, we believe the proportion of variance explained by these components is adequate to reveal the key structure of the OFSEP-HD cohort genetic background.

403 We treated cluster 7 associated with an admixed North-African – European population
404 as part of the European cluster during statistical analyses due to its closer proximity with
405 European reference population and to its lower average African genome proportion compared
406 to North-African cluster (Tables S3 - 4 and Figures S3 - 4).

407 The study reveals differences between self-reported origins and genetic ancestry of
408 French North-African MS patients. Only half North-African genetic ancestry patient self-
409 identified as such, highlighting the significance of cultural affiliation over ancestral
410 backgrounds for patients born in France. This disparity may reflect broader social dynamics,
411 including potential reluctance to acknowledge North-African origins due to historical and
412 ongoing discrimination. Approximately 10% of these patients reported "unknown" or "other"
413 origins, indicating complexities in identity and societal perceptions within multicultural
414 societies (Figure S10)[51].

415 In addition, and maybe most importantly, the way questions on origins are framed may
416 also introduce ambiguities. The phrasing of OFSEP-HD survey question, "From which
417 geographical region are you originating from?", may lead to diverse interpretations, such as
418 reflecting on patients' birthplace, or where they grew up[52]. Other common biases should also
419 be considered, such as errors that may have occurred in form completion, or practitioners
420 misinterpretation and misrecording of information[53]. With this example of MS French
421 population from North-African origins, we highlight the misuse of population descriptors such
422 as self-reported origins that perpetuate inaccurate presuppositions regarding genetic
423 ancestral background. Therefore, we recommend against using such population descriptors
424 alone as proxies for genetic ancestry. Although the use of geographical origins and ethnicity
425 as population descriptors is contentious, we suggest to not fully dismiss them in genetics and

426 genomics research since such descriptors offer additional insights to better characterize
427 groups that have historically experienced health and healthcare disparities, including
428 inequities in genomic healthcare access[54].

429 Additionally, we showed a stronger alignment between parents reported North-
430 African origins and patients' genetic ancestry when both parents had the same origins.
431 Discrepancies arose when patients reported different parents' origins. Using shared parental
432 origins as a proxy for genetic ancestry should be done cautiously, as it may not fully represent
433 the child's genetic background, especially if only one parent is of North-African origin. Among
434 patients from other genetic clusters, self-reporting North-African origins showed a
435 significantly higher ($p < 1e-4$) proportion of African genomic ancestry (15.83% ± 12.82%)
436 compared to patients reporting other origins (2.54% ± 0.36%). However, the high variability in
437 African genomic ancestry proportion ($MoE_{95\%}$ = 12.82 %) suggests diverse reasons for self-
438 reporting North-African origins.

439 The use of log-additive models to compute PRS does not show considerable
440 advantages compared to a simpler SRAC approach when considering solely non-MHC
441 autosomal SNPs, illustrated by the high correlation between both methods (Figure S21). This
442 is due to the very similar effect sizes associated to non-MHC MS SNPs, yielding to nearly
443 identical impacts of these SNPs on PRS calculation (0R ≈ 1)[9,48]. However, log-additive models
444 seem more suitable when considering MHC variants since it accounts for larger effect sizes
445 associated MHC region variations, as opposed to SRAC method in which all variants have the
446 same impact on score calculation (Figure S21). Although computed for the entire OFSEP-HD
447 cohort, most MS variants were found in European populations and may not fully capture
448 genetic risk for patients from other genetic ancestral backgrounds[41,43]. Indeed, we showed a

449 significant difference in score distributions between North-African and European genetic
450 ancestries, which is more probably due to the lack of MS variants found specifically in North-
451 African MS populations than to a real lower susceptibility of North-African patients to MS
452 compared to European patients. These results indicate that relying on current PRS in clinical
453 use might perpetuate healthcare disparities, highlighting the need for more inclusive models
454 that reflect diverse genetic backgrounds[55].

455 Finally, our customized adapted methodology of the Avatar algorithm showed
456 considerable improvement regarding fidelity and privacy compared to single generation of
457 synthetic avatar data[45]. Indeed, averaging parallel synthetic data generation in a patient-wise
458 fashion conducted in considerably higher fidelity as well as increased data privacy due to the
459 combination of multiple avatar patients to create the synthetic patient for each original
460 patient. All privacy and signal metrics values for our synthetic data generation pipeline meet
461 all standard thresholds criteria[45]. This approach underscores the flexibility of the synthetic
462 data generation methodology in producing high-fidelity synthetic datasets that meet stringent
463 privacy standards (GDPR), making it a viable tool for future research applications. However, in
464 this study we performed only univariate comparisons between original and synthetic datasets.
465 It is equally important to verify that multivariate dependencies are preserved—for example,
466 by confirming that all significant associations between features, as well as between patients
467 and variants, are maintained.

468 **Conclusion**

469 In this study, we characterized the genetic landscape of the French OFSEP-HD MS
470 cohort into seven genetic ancestral clusters. We found that self-reported origins significantly
471 underestimate North-African ancestry, highlighting their limitations as proxies for genetic

472 ancestry; parental origins proved more reliable. In addition, French MS patients from North-
473 African genetic ancestry exhibited admixture profiles closer to European than to sub-Saharan
474 African populations. We also imputed HLA alleles/haplotypes and computed MS PRS,
475 revealing a greater contribution of MHC region variants to PRS values. However, our findings
476 demonstrate that current MS susceptibility variants are ill-suited for calculating PRS in non-
477 European ancestries. This work establishes the first comprehensive framework for analyzing
478 genetic diversity in the French OFSEP-HD cohort and propose a first open synthetic and GDPR-
479 enforced and privacy-aware genetic dataset for this population. This paves the way for
480 improvement of current synthetic data generation methodologies and the development of
481 new methods specifically designed for the sensitivity and complexity of genetic data. Overall,
482 this study lays the foundation for future MS research in France in the French national context
483 of limited genetic ancestral studies due to cultural and ethical considerations.

484 Disclosure of interests

485 J.P, N.S.B, M.M, I.F, S.D, M.B, S.L, L.B, A.S-E, S.B-H, F.C, R.C, L.B, N.V, declare that they
486 have no competing interest.

487 Prof. PA. Gourraud is the founder of Methodomics (2008) and the co-founder of Big
488 data Santé (2018). He consults for major pharmaceutical companies, all of which are handled
489 through academic pipelines (AstraZeneca, Biogen, Boston Scientific, Cook, Edimark, Ellipses,
490 Elsevier, Methodomics, Merck, Mérieux, Sanofi-Genzyme, Octopize). PA Gourraud is a
491 volunteer board member at AXA non-for-profit mutual insurance company (2021). He has no
492 prescription activity with either drugs or devices.

Dr. D. Laplaud has participated in advisory boards for Alexion, Merck, Novartis, and Roche in the last 3 years.

Pr. S Vukusiç has received lecturing fees, travel grants, and research support from Biogen, Janssen, Merck, Novartis, Roche, Sandoz, Sanofi-Genzyme, and Teva.

Acknowledgments

We thank OFSEP for providing source data. Indeed, data collection has been supported by a grant provided by the French State and handled by the "Agence Nationale de la Recherche," within the framework of the "France 2030" program, under the reference ANR-10-COHO-002, Observatoire Français de la Sclérose en Plaques (OFSEP) - Eugène Devic EDMUS Foundation against multiple sclerosis. I also wish to thank the Marie-Curie institute (Paris, France) for genotyping the samples. In addition, we thank the historical contributors of REFGENSEP, Lena Guillot-Noël and Isabelle Rebeix.

This work has been supported by the PRIMUS project (Projection in Multiple Sclerosis), part of a government grant managed by the French National Research Agency (Agence Nationale de la Recherche, ANR) as its 3rd PIA, integrated into the “France 2030” plan under reference [ANR-21-RHUS-0014]. This work was supported by Nantes Métropole, Région des Pays de la Loire and European Union (FEDER) via the Programme d’investissements d’Avenir (NExT, SHLARC Project, Nantes Université).

Moreover, insights and encouragement from following collaborators were instrumental in the completion of this work. I am grateful to CR2TI for providing the necessary facilities and technical support, and Dr. Vincent Mauduit for his assistance with data analysis.

We thank the Curie genomic platforms for technical support and GWAS genotyping. We are also grateful to the Bioinformatics Core Facility BiRD, member of Biogenouest and Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013) for the use of their resources and their technical support.

Authors' contribution

Data Curation: M.M. performed the initial quality control using the Axiom Analysis Tool from Thermofisher, following standardized protocols. J.P. curated the final dataset from the original data.

Formal Analysis: N.S.B.S. conducted the HLA imputation using the HIBAG R tool and data reference from SNP-HLA Reference Consortium. L.B imputed HLA haplotypes and phase using HLA2Haplo interface.

Investigation J.P performed all other analyses and experiments related to the study.

Data Anonymization and Synthetic Data Generation: J.P. generated anonymized synthetic data using the Avatarization method from Octopize in the purpose of data sharing.

Conceptualization: J.P. with input from P.A.G. and N.V, as well as M.M and I.F.

Supervision: P.A.G., N.V.

Writing – Original Draft: J.P.

Writing – Review & Editing: P.A.G. and N.V. as well as I.F, L.B, and S.D contributed to discussions and validated the final manuscript.

Web ressources

Bcftools 1.13: https://samtools.github.io/bcftools/

Plink 1.9 and 2: https://www.cog-genomics.org/plink/

Admixture 1.3.0: https://dalexander.github.io/admixture/

TopMed: https://imputation.biodatacataslyst.nhlbi.nih.gov/

Eigensoft - smartpca: https://github.com/chrchang/eigensoft

HLA2Haplo: https://easyhla.univ-nantes.fr/

Avatar: https://github.com/octopize/avatar-python

Genome-wide Complex Trait Analysis (GCTA) 1v.94.1 Linux:
https://yanglab.westlake.edu.cn/software/gcta/#Overview

Data and code availability

Avatar dataset and pipeline scripts are available on github (https://github.com/jp3142).

Declaration of generative AI in scientific writing

During the preparation of this work the authors used OpenAI's GPT-4 and 3.5, as well as Mistral's model "LeChat" in order to assist with language refining and rephrasing of certain sections. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

550 ## References

551 1. Walton, C. *et al.* Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of
552 MS, third edition. *Mult Scler* **26**, 1816–1821 (2020).

553 2. Nelson, L. M. *et al.* A new way to estimate neurologic disease prevalence in the United
554 States. *Neurology* **92**, 469–480 (2019).

555 3. McGinley, M. P., Goldschmidt, C. H. & Rae-Grant, A. D. Diagnosis and Treatment of Multiple
556 Sclerosis: A Review. *JAMA* **325**, 765–779 (2021).

557 4. Pierret, C., Mainguy, M. & Leray, E. Prevalence of multiple sclerosis in France in 2021: Data
558 from the French health insurance database. *Revue Neurologique* **180**, 429–437 (2024).

559 5. Sawad, A. B., Seoane-Vazquez, E., Rodriguez-Monguio, R. & Turkistani, F. Cost-effectiveness
560 of different strategies for treatment relapsing-remitting multiple sclerosis. *J Comp Eff Res* **6**, 97–108
561 (2017).

562 6. Simpson, S. *et al.* Sun Exposure across the Life Course Significantly Modulates Early Multiple
563 Sclerosis Clinical Course. *Front Neurol* **9**, 16 (2018).

564 7. Ascherio, A. Environmental factors in multiple sclerosis. *Expert Rev Neurother* **13**, 3–9 (2013).

565 8. Lanz, T. V. *et al.* Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM.
566 *Nature* **603**, 321–327 (2022).

567 9. INTERNATIONAL MULTIPLE SCLEROSIS GENETICS CONSORTIUM. Multiple sclerosis genomic
568 map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).

569 10. Hollenbach, J. A. & Oksenberg, J. R. The immunogenetics of multiple sclerosis: A
570 comprehensive review. *J Autoimmun* **64**, 13–25 (2015).

571 11. Harroud, A. *et al.* Locus for severity implicates CNS resilience in progression of multiple
572 sclerosis. *Nature* 1–9 (2023) doi:10.1038/s41586-023-06250-x.

573 12. Saint Pierre, A. *et al.* The genetic history of France. *Eur J Hum Genet* **28**, 853–865 (2020).

574 13. Kassar, H. *et al.* Emigration flows from North Africa to Europe. *European Journal of Public*
575 *Health* **24**, 2–5 (2014).

576 14. Botigué, L. R. *et al.* Gene flow from North Africa contributes to differential human genetic
577 diversity in southern Europe. *Proceedings of the National Academy of Sciences* **110**, 11791–11796
578 (2013).

579 15. Stoeklé, H. *et al.* French-style genetics v. 2.0: The “e-CohortE” project. *Clin Genet* **96**, 330–
580 340 (2019).

581 16. Pfennig, A., Petersen, L. N., Kachambwa, P. & Lachance, J. Evolutionary Genetics and
582 Admixture in African Populations. *Genome Biology and Evolution* **15**, evad054 (2023).

583 17. Fatumo, S. *et al.* Diversity in Genomic Studies: A Roadmap to Address the Imbalance. *Nat*
584 *Med* **28**, 243–250 (2022).

585 18. Vukusic, S. *et al.* Observatoire Français de la Sclérose en Plaques (OFSEP): A unique
586 multimodal nationwide MS registry in France. *Mult Scler* **26**, 118–122 (2020).

587 19. Gauthier, A. *et al.* Comparison of SimoaTM and EllaTM to assess serum neurofilament-light
588 chain in multiple sclerosis. *Ann Clin Transl Neurol* **8**, 1141–1150 (2021).

589 20. Francis, G. *et al.* Prognostic factors of disability progression in multiple sclerosis in real life:
590 the OFSEP-high definition (OFSEP-HD) prospective cohort in France. *Open access*.

591 21. Thompson, A. J. *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria.
592 *Lancet Neurol* **17**, 162–173 (2018).

593 22. EDMUS, a European database for multiple sclerosis - PubMed. https://pubmed-ncbi-nlm-nih-
594 gov.proxy.insermbiblio.inist.fr/1527537/.

595 23. Kurtzke, J. F. Rating neurologic impairment in multiple sclerosis: an expanded disability status
596 scale (EDSS). *Neurology* **33**, 1444–1452 (1983).

597 24. Lublin, F. D. *et al.* Defining the clinical course of multiple sclerosis. *Neurology* **83**, 278–286
598 (2014).

599 25. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
600 *Nature* **526**, 68–74 (2015).

601 26. Arauna, L. R. *et al.* Recent Historical Migrations Have Shaped the Gene Pool of Arabs and
602 Berbers in North Africa. *Molecular Biology and Evolution* **34**, 318–329 (2017).

603 27. Henn, B. M. *et al.* Genomic Ancestry of North Africans Supports Back-to-Africa Migrations.
604 *PLOS Genetics* **8**, e1002397 (2012).

605 28. Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control
606 and statistical analysis. *Int J Methods Psychiatr Res* **27**, e1608 (2018).

607 29. Jordan, I. K., Sharma, S., Nagar, S. D., Valderrama-Aguirre, A. & Mariño-Ramírez, L. Correction
608 to: Genetic Ancestry Inference for Pharmacogenomics. *Methods Mol Biol* **2547**, C1 (2022).

609 30. Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B. & Heming, J. K-means clustering
610 algorithms: A comprehensive review, variants analysis, and advances in the era of big data.
611 *Information Sciences* **622**, 178–210 (2023).

612 31. Yuan, C. & Yang, H. Research on K-Value Selection Method of K-Means Clustering Algorithm.
613 *J* **2**, 226–235 (2019).

614 32. Guarino-Vignon, P., Marchi, N., Bendezu-Sarmiento, J., Heyer, E. & Bon, C. Genetic continuity
615 of Indo-Iranian speakers since the Iron Age in southern Central Asia. *Sci Rep* **12**, 733 (2022).

616 33. Mendes, M., Alvim, I., Borda, V. & Tarazona-Santos, E. The history behind the mosaic of the
617 Americas. *Curr Opin Genet Dev* **62**, 72–77 (2020).

618 34. Hazra, A. Using the confidence interval confidently. *J Thorac Dis* **9**, 4125–4130 (2017).

619 35. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* **22**, 276–282
620 (2012).

621 36. A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960.
622 https://journals.sagepub.com/doi/10.1177/001316446002000104.

623 37. Vince, N. *et al.* SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for
624 promoting MHC-centric analyses in genomics. *Genet Epidemiol* **44**, 733–740 (2020).

625 38. Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics*
626 *J* **14**, 192–200 (2014).

627 39. Geffard, E. *et al.* Easy-HLA: a validated web application suite to reveal the full details of HLA
628 typing. *Bioinformatics* **36**, 2157–2164 (2020).

629 40. Isobe, N. *et al.* Genetic burden in multiple sclerosis families. *Genes Immun* **14**, 434–440
630 (2013).

631 41. Gourraud, P.-A. *et al.* Aggregation of multiple sclerosis genetic risk variants in multiple and
632 single case families. *Ann Neurol.* **69**, 65–74 (2011).

633 42. Akoglu, H. User's guide to correlation coefficients. *Turk J Emerg Med* **18**, 91–93 (2018).

634 43. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human
635 populations. *Nat Commun* **10**, 3328 (2019).

636 44. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex
637 trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).

638 45. Guillaudeux, M. *et al.* Patient-centric synthetic data generation, no reason to risk re-
639 identification in biomedical data analysis. *npj Digit. Med.* **6**, 37 (2023).

640 46. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27**,
641 857–871 (1971).

642 47. DeLuca, G. C. *et al.* An extremes of outcome strategy provides evidence that multiple
643 sclerosis severity is determined by alleles at the HLA-DRB1 locus. *Proceedings of the National*
644 *Academy of Sciences* **104**, 20896–20901 (2007).

645 48. Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat*
646 *Genet* **47**, 1107–1113 (2015).

647 49. Bonferroni, C. E. *Teoria statistica delle classi e calcolo delle probabilità*. (Seeber, 1936).

648 50. Armstrong, R. A. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* **34**, 502–508
649 (2014).

650 51. Pereira, J. L. *et al.* Genetic Ancestry and Self-Reported "Skin Color/Race" in the Urban
651 Admixed Population of São Paulo City, Brazil. *Genes* **15**, 917 (2024).

652 52. Gotlieb, R. *et al.* Accuracy in Patient Understanding of Common Medical Phrases. *JAMA*
653 *Network Open* **5**, e2242972 (2022).

654 53. Goldberg, S. I., Niemierko, A. & Turchin, A. Analysis of Data Errors in Clinical Research
655 Databases. *AMIA Annu Symp Proc* **2008**, 242–246 (2008).

656 54. Feero, W. G. *et al.* Guidance on use of race, ethnicity, and geographic origin as proxies for
657 genetic ancestry groups in biomedical publications. *The American Journal of Human Genetics* **111**,
658 621–623 (2024).

659 55. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health
660 disparities. *Nat Genet* **51**, 584–591 (2019).

661