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[1,2], Mitra Barzine[1], Anna Serova-Erard[1,3], François Cornelis[1,3,4], Sonia Bourguiba-Hachemi[1], Sophie Limou[1], Francis Guillemin[5], Sandra Vukusic[6,7,8], Romain Casey[6,7,8], Jonathan Epstein[5], Anne Kerbrat[9], Emmanuelle Leray[10], Eric Thouvenot[11,12], Guillaume Mathey[13], Laure Michel[14], Emmanuelle Le Page[15], Jerome De Seze[2], Christine Lebrun-Frenay[16], Caroline Papeix[17], Jonathan Ciron[18], Pierre Clavelou[19], Eric Berger[20], Aurelie Ruet[21,22], Thibault Moreau[23], Olivier Casez[24,25], Pierre Labauge[26], Abir Wahab[27], Gilles Defer[28], Amélie Dos Santos[29], Thomas David[30], Inès Doghri[31], Elisabeth Maillart[32,33], Laurent Magy[34], Helene Zephir[35], Olivier Heinzlef[36], Bertrand Fontaine[37], Laureline Berthelot[1], David-Axel Laplaud[1], Nicolas Vince*[1], Pierre-Antoine Gourraud[1,38], On behalf of OFSEP Biology group

1. *Nantes Université, CHU Nantes, INSERM, Centre de recherche en Transplantation et Immunologie Translationnelle, UMR 1064, F-44000, Nantes, France.*
2. *CHU de Strasbourg, Department of Neurology and Clinical Investigation Center, CIC 1434, INSERM 1434, F-67000 Strasbourg, France*
3. *CHU Clermont-Ferrand, Génétique – Oncogénétique Adulte – Prévention (GENOAP), Clermont-Ferrand.*
4. *Université Clermont Auvergne, UFR de Médecine et des professions paramédicales, Clermont-Ferrand.*

5. *CHRU, INSERM, Université de Lorraine, CIC Clinical Epidemiology, Nancy, Grand Est, France, Université de Lorraine, INSERM, INSPIIRE, Paris, Île-de-France, France*
6. *Hospices Civils de Lyon, Service de Neurologie, sclérose en plaques, pathologies de la myéline et neuro-inflammation, Bron, France, 69677*
7. *Observatoire Français de la Sclérose en Plaques, Centre de Recherche en Neurosciences de Lyon, INSERM 1028 et CNRS UMR 5292, Lyon, France, 69003*
8. *Université de Lyon, Université Claude Bernard Lyon 1, Lyon, France, 69000*

   Eugène Devic EDMUS Foundation Against Multiple Sclerosis (a government approved foundation) Bron, France, 69677
9. *Neurology Department, CRC-SEP Rennes, Rennes Clinical Investigation Center CIC 1414, Rennes University Hospital Rennes University INSERM, CHU Pontchaillou, University, F-35000 Rennes France.*
10. *Univ Rennes, EHESP, CNRS, INSERM, Arènes—UMR 6051, RSMS (Recherche sur les Services et Management en Santé)—U 1309, Rennes, France*
11. *CHU Nimes, Service de Neurologie, Univ Montpellier, Nimes, France*
12. *Institut de Génomique Fonctionnelle, Univ Montpellier, CNRS, INSERM, Montpellier, France*
13. *Nancy University Hospital, Department of Neurology, Nancy, France. Université de Lorraine, APEMAC, F-54000 Nancy, France.*
14. *Neurology department; Clinical Neuroscience Centre, CIC_P1414 INSERM, Rennes, University Hospital, Rennes University.*
15. *CHU Pontchaillou, CIC1414 INSERM, F-35000 Rennes France.*

16. *Neurology, UR2CA_URRIS, Centre Hospitalier Universitaire Pasteur2, Université Nice Côte d'Azur, Nice, France*
17. Foundation Adolphe de Rothschild Hospital , Department of Neurology, F-75000 Paris, France and université Paris -cité
18. *CHU de Toulouse, CRC-SEP, department of Neurology, F-31059 Toulouse Cedex 9, France; Université Toulouse III, Infinity, INSERM UMR1291 - CNRS UMR5051, F-31024 Toulouse Cedex 3, France*
19. *CHU Clermont-Ferrand, CRC SEP Auvergne, Department of Neurology, and INSERM NeuroDol U1107, 63003 Clermont-Ferrand*
20. *CHU de Besançon, Service de Neurologie 25 030 Besançon, France*
21. *Department of Neurology, University Hospital of Bordeaux, 33076 Bordeaux, France*
22. *Neurocentre Magendie, Bordeaux University, INSERM U1215, 33000 Bordeaux, France*
23. *CHU de Dijon, Department of Neurology, EA4184, F-21000 Dijon, France*
24. *CHU Grenoble Alpes, Department of Neurology, Neurology MS Clinic Grenoble, Grenoble Alpes university hospital, Grenoble, F-38700 La Tronche*
25. *T-RAIG, TIMC-IMAG, Grenoble Alpes University, France*
26. *CHU de Montpellier, MS Unit, F-34295 Montpellier Cedex 5, France; University of Montpellier (MUSE), F-34000 Montpellier, France*
27. *APHP, Hôpital Henri Mondor, Department of neurology, F-94000 Créteil, France*
28. *CHU de Caen, MS expert centre, Department of Neurology, avenue de la Côte-de-Nacre, Normandy University, 14033 Caen, France*

29. *CHU La Milétrie, Hôpital Jean Bernard, Department of neurology et Centre d'Investigation Clinique, CIC INSERM 1402, F-86000 Poitiers, France*
30. *CHU de la Martinique, Department of Neurology, F-97200 Fort-de-France, France*
31. *CHU de Tours, Hôpital Bretonneau, CRC SEP and department of neurology, F-37000 Tours, France*
32. *Département de neurologie, Hôpital Pitié-Salpêtrière, APHP, Paris ;*
33. *Centre de Ressources et de Compétences SEP Paris, France.*
34. *CHU de Limoges, Hôpital Dupuytren, Department of Neurology, F-87000 Limoges, France*
35. *CHU Lille, CRCSEP Lille, Univ Lille, U1172, F-59000 Lille, France*
36. *Hôpital de Poissy, Departement of Neurology, F-78300 Poissy, France*
37. *From the Sorbonne Université - Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital Pitié Salpétrière Univ. Hosp., DMU Neuroscience 6; Inst. of Cardiometabolism and Nutrition, Sorbonne-universités-Upmc 06, INSERM, CNRS; Laboratoire des Signaux et Systèmes (L2S), CNRS-CentraleSupélec, Université Paris-Saclay; Sorbonne Université (B.S.), Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital St. Antoine-HUEP; and INSERM (B.F.), SU, AP-HP, Centre de recherche en Myologie-UMR974 and Service of Neuro-Myology, Institute of Myology, University hospital Pitié-Salpêtriere.*
38. *Nantes Université, CHU Nantes, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, INSERM CIC 1413, F-44000, Nantes.*

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

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

Abstract

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

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

121 MS pathogenesis is caused by the interplay between both environmental and genetical
122 components. The well-known latitude gradient of MS demonstrates lowest prevalence regions
123 near the equator and high prevalence regions outside of the tropics especially in North-
124 Western countries[1]. A South-West to North-East increasing latitude gradient is also observable
125 in metropolitan France, demonstrating the impact of the gradient even at a lower regional
126 scale[3]. This gradient is thought to be primarily driven by the lack of sun exposure and vitamin
127 D deficiency, more prominent as we get further from the equator[5]. Moreover, Epstein-Barr
128 Virus (EBV) and CytoMegaloVirus (CMV) infections during lifetime, smoking and childhood
129 obesity, contribute significantly to MS onset[6,7]. Nevertheless, this latitude gradient might also
130 partly result in healthcare disparities across the world, with lowest prevalence regions bearing
131 fewer data due to a deficit of neurologist and MS specialist, in turn yielding to a systemic bias
132 towards fewer MS diagnostics in those regions[1].

133 MS is a polygenic disease, resulting from both previously stated environmental factors
134 and the cumulative effect of multiple genetic variations that influence susceptibility to the
135 disease, as opposed to Mendelian diseases, which are mostly caused by a single genetic
136 mutation. The most recent GWAS (Genome Wide Association Study) identified a total of 233
137 MS Single Nucleotide Polymorphism (SNP) associated to MS susceptibility, with 32 located in
138 the Major Histocompatibility Complex (MHC) region explaining most of the phenotypic
139 variance, and 1 located on the X chromosome[8]. Recent insights also suggest the role of
140 genetics on disease severity and progression with one significant variant discovered to date[9].

141 The French population's genetic architecture stems from two major ancestries: early
142 Neolithic farmers from Anatolia and steppe pastoralists from Eurasia. While Neolithic farmer
143 DNA predominates in modern France, steppe pastoralist contribution is more pronounced in
144 the north[10]. Over centuries, historical cultural and political borders have shaped the genetic
145 architecture of today's French population. Indeed, genetic clusters in France mostly align with
146 neighboring countries, reflecting shared histories and interactions[10]. Moreover, France's
147 strategic location on Europe's western edge has made it a key migration route between the
148 continent's south and north, enriching its genetic diversity. Notably, well-documented
149 migration between North Africa and France, driven by historical, cultural, and economic ties,
150 has contributed to human genetic diversity in southern Europe, including France.

151 However, studying genetic ancestry in France is historically sensitive, and faces legal
152 constraints explaining the few genetic studies conducted on French population with genome
153 wide data[10]. Additionally, most patients in genetic studies focusing on polygenic diseases like
154 MS are from European genetic ancestry[13,14]. Nevertheless, initiatives like the French Multiple
155 Sclerosis Registry – High Definition Cohort (OFSEP-HD) and its 2,667 samples with genetic data,

156 is an opportunity to study MS genetics with high quality nation-wide data collection[15,16]. While
157 the OFSEP-HD cohort is instrumental in numerous published MS studies, its genetic
158 architecture has not been studied yet[17,18]. This study provides the first genetic analysis of the
159 OFSEP-HD population, highlighting the role of ancestry and MS variants including MHC
160 variations for future research into genetic factors influencing MS within the French
161 population.

162

Materials and methods

Data collection

The OFSEP-HD observational study, totalizing 2,667 genetic samples of MS patients from the historical OFSEP study collected from various centers distributed across France, represents an ongoing cohort study initiated in 2018 with annual high quality follow-up[15,18]. All OFSEP-HD patients fulfilled MacDonald2017+ diagnosis criteria at inclusion[19,20]. Appropriate informed consent was obtained from all OFSEP-HD patients[20]. Data were generated by participating neurologists in the framework of the French MS registry (OFSEP)[15,16,18,20,21]. The cohort includes different types of demographic variables including patients' and parents' origins, sex, birthdate, Expanded-Disability-Status-Scale (EDSS) (scale measuring the level of disability), pregnancies, relapses, treatments and type of MS course at inclusion in addition to annual MRI data and sample biocollection[20,22,23].

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

The OFSEP-HD cohort was compared to a subset of 2,405 reference individuals from the 1000 Genomes Project population, comprising individuals from East-Asian, South-Asian, American, European, and African ancestries[24]. In addition, a North African reference cohort of 154 non MS individuals from Arauna et al. was utilized to accurately characterize genetic ancestry in the light of its 486,252 genotyped variants[25,26].

Processing, imputation, and quality controls of genotyping data

184 Quality control (QC) procedures were applied to OFSEP-HD dataset and North-African
185 reference dataset. No additional quality control or imputation was performed on the 1000
186 Genome reference dataset[27]. Technological QC on the OFSEP-HD dataset using Axiom Analysis
187 Suite 2.0 selected 869,082 high-quality variants (97.8%).

188 A second QC filtering procedure was implemented using Python 3.11 with PLINK 1.9/2
189 and bcftools 1.13 following previously published guidelines[27]. Pre-imputation filters and
190 thresholds applied to OFSEP-HD genetic samples are listed below:

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

197 Processing steps of North-African reference dataset are detailed in Table S1 and a
198 flowchart of the QC procedures in Figure S1[25].

199 Pre-imputation quality controls of OFSEP-HD population resulted in 424,132 remaining
200 SNPs and 2,542 patients. The filters removing most of the variants were the MAF filters and
201 sexual and mitochondrial chromosome variants exclusion. Imputation of OFSEP-HD genetic
202 data was then performed using Topmed Michigan Imputation Server to maximize overlap with
203 reference datasets during ancestry analysis. Post-imputation QCs were also performed on
204 OFSEP-HD dataset, excluding variants with insufficient imputation quality (r^2<0.5) or low

205 frequency (MAF<10%) for genetic ancestry analyses. Moreover, we filtered again for variants
206 considerably deviating from HWE<$1e^{-10}$.

207 **Ancestry analysis**

208 We performed a Principal Component Analysis (PCA) of the 2,542 OFSEP-HD patients
209 with all reference populations using the Eigensoft software. The analysis considered the
210 twenty first principal components obtained from the 9,326 common independent SNPs to
211 OFSEP-HD and reference datasets. A supervised K-Means clustering approach was then
212 implemented to identify ancestry superpopulation clusters on the PCA (Figure S2)[28].

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

214 We analyzed demographic and clinical variables of the 2,542 remaining patients after
215 quality controls and processing steps providing key insights into the composition of the cohort.
216 We computed descriptive statistics to better understand the self-reported origins distribution
217 of OFSEP-HD patients. Participants' origins were reported via a Case Report Form (CRF) with
218 three instances of the question formatted as follows: "Which geographical region are you
219 originating from?". Patients answered all questions, providing their own and their parents'
220 geographical origins.

221 **Reference populations selection and admixture proportions of European, African**
222 **and East-Asian genome**

223 Admixture analysis was performed using admixture 1.3.0 software. The reference
224 parental populations for admixture analysis were selected from the 1,000 Genome dataset to
225 include the three most discretizing ancestries: European, African, and East-Asian, thereby

226 excluding North-African, South-Asian and American admixed ancestries[26,29,30]. More details
227 are available in Supplemental Materials.

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

229 Patients' and parents' origins were compared to genetically inferred ancestral
230 backgrounds. Self-reported origins of OFSEP-HD patients and their parents were collected
231 through the previously mentioned CRF. Mapping between ancestry clusters and self-reported
232 origins is shown in Table S2 and the methodology used to select clusters for statistical analyses
233 is discussed in Figures S3-4 and Tables S3–4. Proportion of discordant pairs and Cohen's kappa
234 score were used as a measure of agreement between patients' self-reported origins and
235 genetic ancestry. Proportion of discordant pairs was calculated as $\frac{\mathrm{b+c}}{\mathrm{a+b+c+d}} * 100$**,** with a, b, c
236 and d being respectively the upper left, upper right, lower left and lower right cells of a 2x2
237 contingency table. Margin of Errors (95%) ($\mathrm{MoE}_{95\%}$) were computed as $\pm 1.96 * \frac{\sigma}{\sqrt{n}}$ with σ the
238 standard deviation and n the sample size[31]. Statistical significance of comparisons were
239 assessed using two-sided statistical association tests from independent data arrangements[32].
240 We performed a quantitative statistical analysis comparing European and African genome
241 proportions within genetic clusters regarding patients self-reported origins status.

242 **HLA imputation: alleles and haplotypes**

243 Human Leukocytes Antigens (HLA) genotypes were imputed from genotyped SNPs
244 through SNP-HLA Reference Consortium (SHLARC), a widely recognized and extensively used
245 resource for imputing *HLA* alleles from SNP data, employing the HIBAG R package and a large
246 multi-ethnic reference panel[33,34]. We also compared the proportion of patients with at least
247 one *HLA-DRB1*15:01* most associated MS susceptibility risk allele between ancestry clusters

248 using Mann-Whitney U test. Further, we performed individual-level haplotypes imputation
249 using HLA2Haplo tool from Easy-HLA software suite independently for each ancestry cluster
250 using corresponding reference panels[35]. This tool inferred the phase of imputed HLA alleles
251 for each considered OFSEP-HD individual. Imputed haplotypes include loci *HLA-A, HLA-B, HLA-*
252 *C, HLA-DRB1,* and *HLA-DQB1*.

253 **Polygenic Risk Scores computation**

254 Polygenic Risk Scores (PRS) provides a summary metric measure of genetic
255 susceptibility risk of an individual to a given genetic-related disease. Here we computed PRS
256 designed specifically for MS, in the OFSEP-HD population[36,37]. We had 210 MS susceptibility
257 variants (including 20 in the MHC region) out of a total of 233 known MS associated
258 susceptibility variations available in our dataset after SNP imputation, filtering (MAF<1%) and
259 *HLA* alleles imputation (Table S5)[8].

260 We compared a log-additive PRS model to a non-weighted Sum of Risk Alleles Count
261 (SRAC) approach using Pearson's correlation coefficient r^2, using scores computed from the
262 global OFSEP-HD population as well as from specific ancestry clusters[36–38]. Since MS
263 susceptibility variants were mostly found in patients from European genetic ancestry, we only
264 compared distributions of both models between European and North-African genetic
265 ancestries using Mann-Whitney U tests[8,37,39]. Furthermore, we assessed the phenotypic
266 variance explained by the 20 available MHC region genetic variations compared to the 190
267 non-MHC SNPs at the whole cohort level using Genome-wide Complex Trait Analysis (GCTA)
268 software[40].

269 **Synthetic data generation**

270 To promote data sharing and reusability of sensitive genetic data, we provided an
271 anonymous – synthetic – GDPR-enforced dataset as an open resource using an adaptation of
272 an already existing "patient-centric" local modelling strategy[41]. The methodology focuses on
273 addressing privacy risks like singling out, linkability, and inference using metrics such as
274 Hidden Rate (HR), Local Cloaking (LC), and Distance To Closest synthetic record (DTC). Signal
275 retention was evaluated using Hellinger distances, correlation differences, and Gower's
276 distance permutation test[42]. Optimal parameters were identified as k=2 and 20 principal
277 components, with regularization for rare modalities ensuring respect of biological constraints.
278 The privacy-fidelity balance was improved by generating three parallel synthetic datasets and
279 averaging features patient-wise. Methodological details are available in Supplemental
280 Materials and the synthetic dataset on GitHub
281 (https://github.com/jp3142/OFSEP_HD_public_avatar_dataset).

282 ## Results

283 ### Ancestry analysis: Principal Components Analysis of OFSEP-HD MS patients

284 Figure 1 shows the Principal Component Analysis (PCA) of OFSEP-HD MS patients
285 compared to the 2 reference datasets: 1,000 Genomes and North-African cohort[25]. It
286 highlights the genetic diversity of the MS French population, with a specific focus on the first
287 two principal components explaining respectively 49.1% and 29.9% of the variance (79%). The
288 10 most contributing SNPs for each principal component are listed in Table S6. The best
289 clustering of the OFSEP-HD population was obtained using 7 distinct clusters.

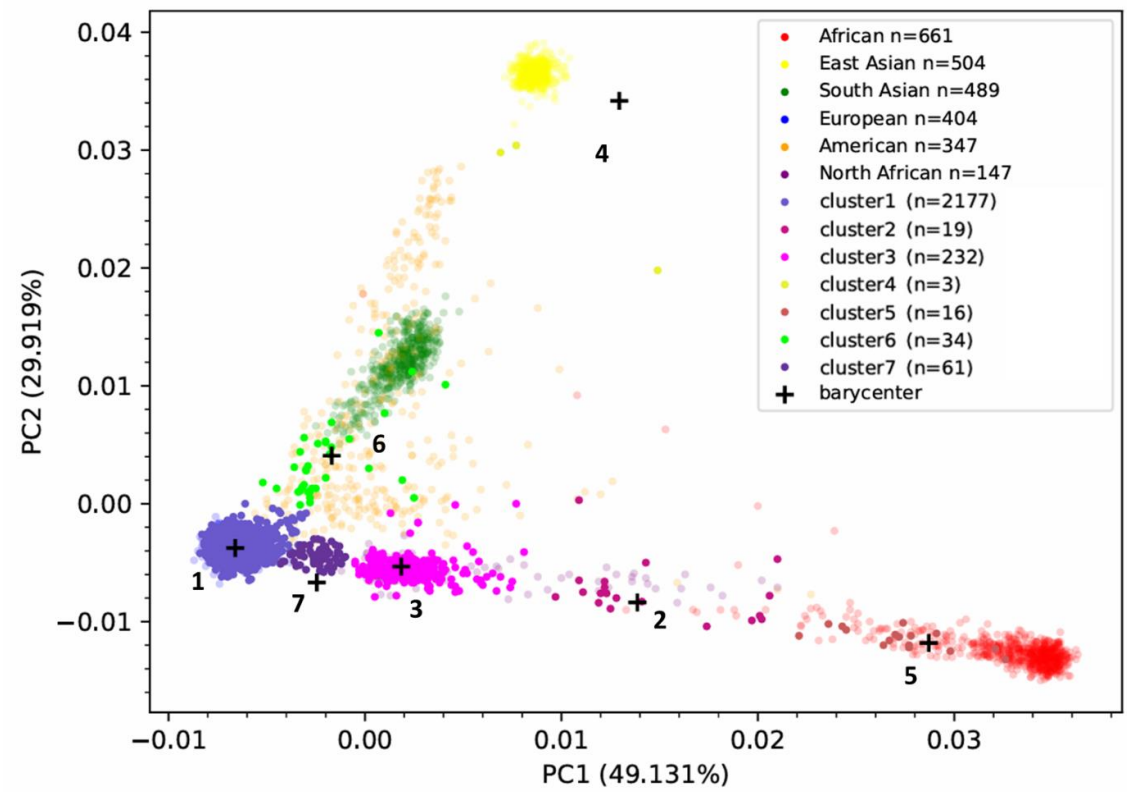

***Figure 1**: Principal Component Analysis (PCA) plot of the OFSEP-HD population based on the first two principal components (PC1 and PC2), overlaid with reference datasets. Each data point represents an individual. Semi-transparent data points represent reference individuals and opaque data points represent OFSEP-HD patients. Cluster 1: European. Cluster 2: Afro-Caribbean. Cluster 3: North-African. Cluster 4: East-Asian. Cluster 5: African. Cluster 6: South-Asian. Cluster 7: Admixed North-African and European.*

Findings highlight a major proportion of patients with European genetic ancestry within the OFSEP-HD cohort, totalizing 2,238 MS patients (88%) assigned to clusters of 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

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

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

311

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

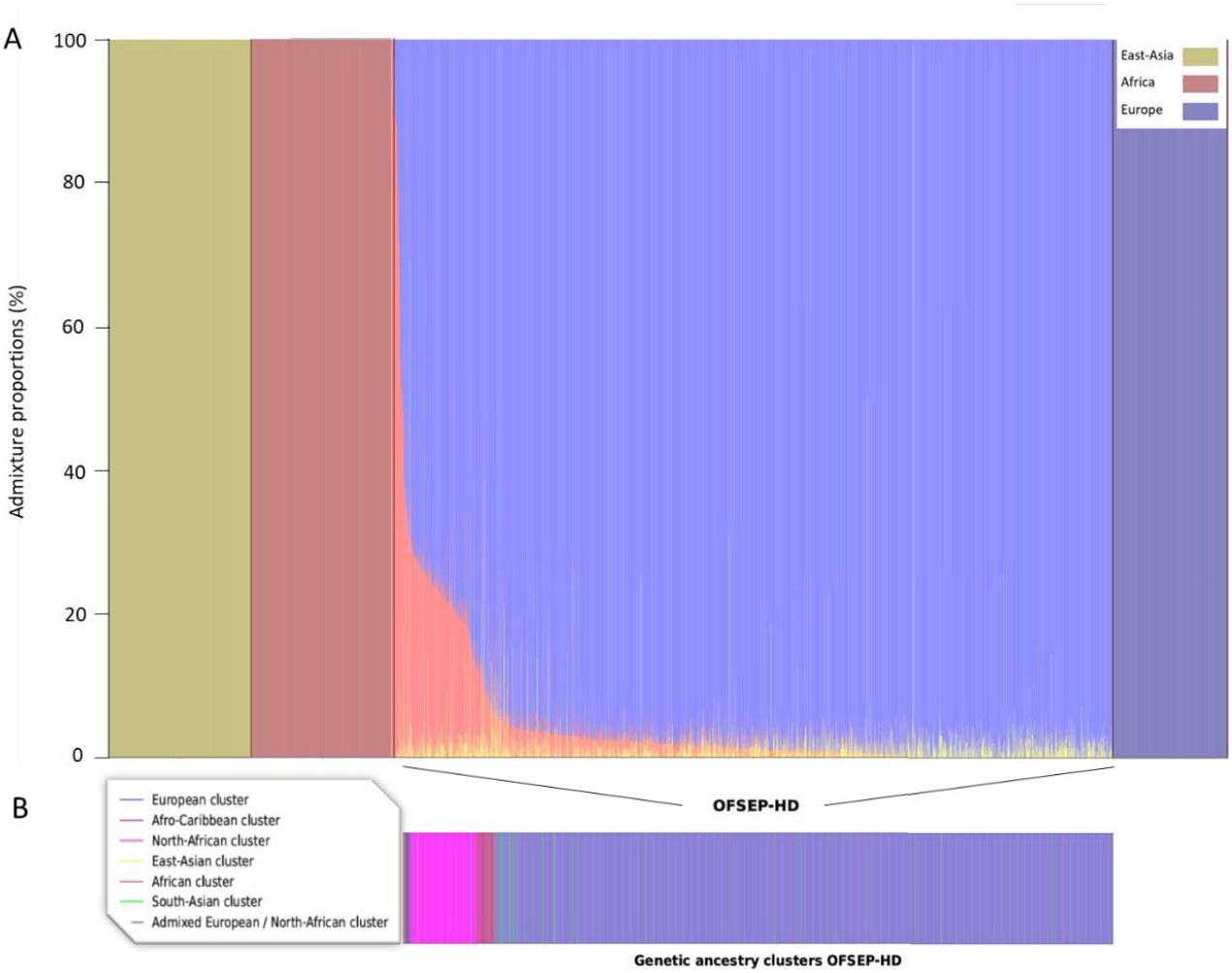

Figure 2: *(A) Admixture plot showing whole-genome admixture proportions of African (red), East Asian (yellow), and European (blue) ancestry for each of the 2,542 OFSEP-HD patients, sorted by African genome proportion. Reference parental populations are shown on both sides of the OFSEP-HD section. (B) Mapping of genetic ancestry clusters previously identified using PCA for each patient.*

We observed a major European genome proportion in the OFSEP-HD 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, most MS patients showing significant African genome ancestry are estimated as part of North-African (cluster 3; n=232) genetic cluster. Distributions of genome proportions among cluster

324 3 is composed in majority by European genetic components (75.5% in average) compared to
325 the African genome proportions ranging from 16.7% to 38.9%. Moreover, we found a non-
326 neglectable East-Asian genome proportion for some MS patients, with 101 (<4%) with at least
327 5% of East-Asian genome ancestry and 2 patients with more than 85% of East-Asian ancestry
328 (Figure 2, Table S10).

329 **Self-reported origins and African-related genetic ancestry**

330 We focused on qualitative analysis of North-African origins as it represents the second
331 largest genetic group in the OFSEP-HD cohort, after European ancestry. Table 2 illustrates the
332 distribution of OFSEP-HD patients based on their North-African self-reported origins and
333 genetic ancestry statuses.

	North-African genetic ancestry	**Other genetic ancestry**	**Total**
North-African self-reported origins	112	15	127
Other self-reported origins	120	2295	2415
Total	232	2310	2542

334 ***Table 2****: Distribution of OFSEP-HD patients regarding North-African self-reported origins and genetic ancestry.*

335 Cohen's Kappa metric (κ = 0.598±0.001) and the low proportion of discordant pairs
336 (5.31%) illustrate a moderate to substantial agreement between self-reported North-African
337 origins and genetic North-African ancestry (Table 2). Remarkably, among 127 patients self-
338 reporting North-African origins, 15 patients (11.8%) have a genome that does not reveal
339 marked North-African genetic background. Moreover, 51.7% of patients from North-African
340 genetic ancestry self-identified as from other origin. Further analyses about North-African
341 patients are presented in Figure S5. Likewise, we detected a strong underestimation of self-

342 reported Sub-Saharan-African ($p<0.001$) and Afro-Caribbean origins ($p<0.05$) compared to
343 their respective PCA-defined ancestry clusters (Figures S6–10; Tables S11–14).

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

345 Among non-North African genetic ancestry patients, no significant differences in
346 African and European genomic components were found between those self-reporting North
347 African origins and those who did not (Table S15), confirming consistency between PCA-
348 defined ancestry clusters and ancestral genomic composition. Thus, we focus on non-North
349 African genetic ancestry patients.

350 We found that self-reported North-African patients from other genetic ancestry have
351 a statistically significant higher proportion (Mann-Whitney U: $p<1e-4$) of African genome
352 compared to patients who did not report such origins (Table S15; Figure S15). Nevertheless,
353 the high variability associated to African genome proportions ($CI_{95\%}$ [3.01%; 28.65%]) and
354 consequently European genome proportions ($CI_{95\%}$ [70.44%; 95.78%]; Table S16) of self-
355 reported North-African patients from other genetic ancestry demonstrates the admixed
356 nature of this subpopulation. Further details on performed analyses and European genome
357 proportions of OFSEP-HD North-African MS patients are discussed in Figures S11–14 and S16-
358 19. Details and analyses of other genetic ancestry clusters and origins are available in Tables
359 S17–28.

360 **Parents' origins: a better proxy of patients' genetic ancestry?**

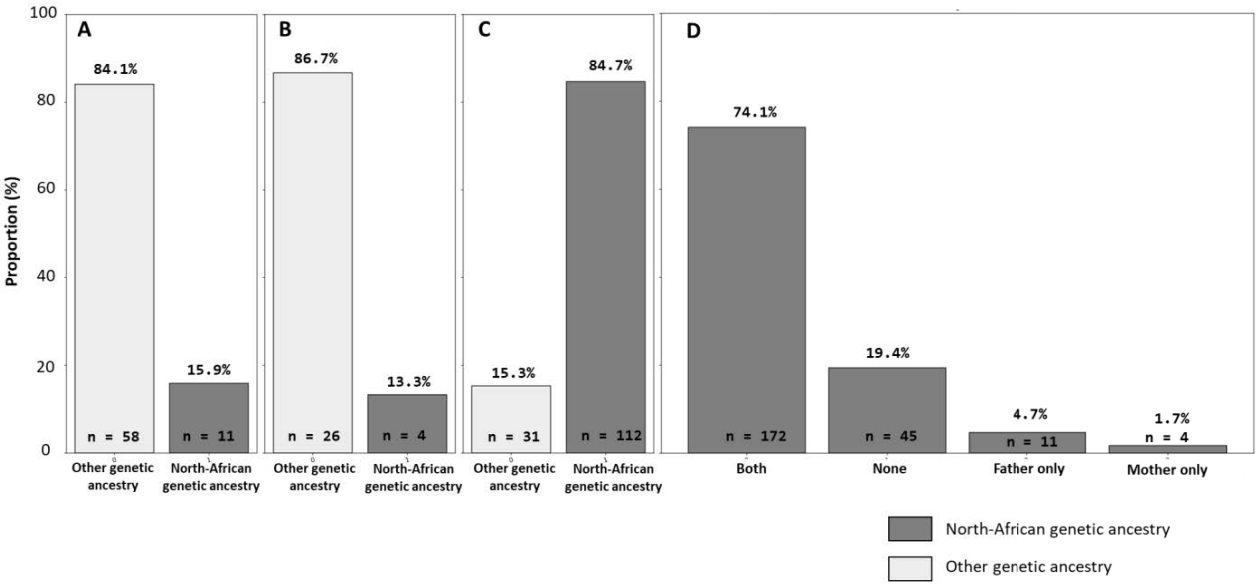

Figure 3*: (A) North-African genetic ancestry status among patients with only the father reported as from North-African origins. (B) North-African genetic ancestry status among patients with only the mother reported as from North-African origins. (C) North-African genetic ancestry status among patients 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 3A and 3B; Table S29). Nevertheless, we observed a substantial agreement with patients' North-African genetic ancestry when both parents have common origins (84.7%, Figure 3C) also highlighted by a higher Cohen's Kappa value ($\kappa=0.64$), and a lower proportion of discordant pairs (3.11%) (Tables S30–33), compared to what was obtained previously with patients' self-reported origins (Table 2). Additionally, 74.1% of MS patients with North-African genetic ancestry reported both parents as from North-African

379 origins (Figure 3D). Analysis of other parents' origins and genetic ancestries are available in
380 Tables S34–37.

381 **HLA alleles imputation**

382 *HLA* alleles imputation was performed on classical *HLA* genes *HLA-A, HLA-B, HLA-C,*
383 *HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1* and *HLA-DRB1* and is particularly accurate for all
384 genes. The average imputation post-probabilities ranging from 86.5% for *HLA-DRB1* to 98.9%
385 for *HLA-DPA1*, indicate trustful results across all considered *HLA* class I and II genes. In
386 addition, we found a significantly higher proportion of patients ($p<1e-5$) with at least one *HLA-*
387 *DRB1*15:01* allele in the European genetic cluster (48.8%) compared to the proportion among
388 the North-African genetic cluster (33.2%)[43]. Additional details about imputation quality
389 distribution statistics are shown in Figure S20 and Table S38.

390 **HLA haplotypes imputation**

391 *HLA* haplotypes imputation from imputed *HLA* alleles *HLA-A, HLA-B, HLA-C, HLA-DRB1*
392 and *HLA-DQB1* achieved very high accuracy with an average post-probability of 82.9% across
393 all haplotypes collectively. Figure S21 and Table S39 show distribution of post-probabilities by
394 HLA locus. We summarize HLA diversity of the 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 %

395 ***Table 3:*** *Description of the 10 most frequent imputed haplotypes including at least one MS associated HLA allele within the*
396 *OFSEP-HD cohort. Bold alleles within haplotypes represent known MS associated HLA alleles*[8,44].

397 All of the ten most frequent HLA haplotypes including at least one HLA MS associated
398 allele present the *HLA-A*02:01* protective allele, but some also contain *HLA-DRB1*15:01* risk
399 allele. The most frequent HLA haplotype is
400 ***A*02:01~**B*07:02~C*07:02**~DRB1*15:01~**DQB1*06:02* (1.91%).

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

402 We found an average log-additive score of 20.29±0.03 for 190 non-MHC SNPs,
403 representing 201 risk alleles per patient. Including the 20 available MHC variations increased
404 the score to 23.38±0.05 for 214 risk alleles in average. The log-additive model and SRAC
405 showed high correlation (r^2=0.91) for non-MHC SNPs, but correlation decreased substantially
406 (r^2=0.54) when MHC variants were included (Figure S22). Moreover, MHC variants increased
407 the log-additive score by 15.23% while only increasing the average number of risk alleles by
408 6.46%, highlighting their larger effect on MS susceptibility. PRS distributions differed
409 significantly between MS patients from North-African and European ancestries for all PRS
410 models (p<1e-6). PRS scores distribution statistic details are available in Table S40 and Figure
411 S23.

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

413 Using our synthetic data generation process, we created a highly anonymized dataset
414 as indicated by key metrics following recommended thresholds: HR=99.49% (>90%); median
415 LC=50 (>5), with DTC=1.8 (>0.2) and CDR=0.99 (>0.3)[41]. Signal retention metrics also show high
416 synthetic dataset fidelity to original data with Hellinger distance distributions difference of
417 0.04 (<0.1) and distributions correlation difference of 1.92% (<10%). Univariate statistical
418 analysis of generated and original datasets indicated four features (rs58394161, rs11809700,
419 rs11578655, rs10801908) out of the 82 generated features where distributions in synthetic
420 data were significantly deviating from original data distributions after Bonferroni correction
421 for multiple testing (p<0.05). Gower's distance permutation test showed no statistically
422 significant overall differences from what would be expected by chance between generated
423 and original datasets (p=0.38)[45]. Figure S24 shows datasets overlap on the two first
424 dimensions of the reduced original Euclidean space.

425 ## Discussion

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

431 Regarding PCA ancestry analysis, we believe the proportion of variance explained by
432 the two first components (79%) is adequate to reveal the key structure of the OFSEP-HD
433 cohort genetic background. Cluster 7, an admixed North-African–European population, was
434 treated as part of the European cluster in statistical analyses due to its closer proximity to the
435 European reference population and lower average African genome proportion compared to
436 the North-African cluster (Tables S3–4, Figures S3–4).

437 The study reveals differences between self-reported origins and genetic ancestry of
438 French North-African MS patients. Only half North-African genetic ancestry patient self-
439 identified as such, highlighting the significance of cultural affiliation over ancestral genetic
440 backgrounds for patients born in France. Approximately 10% of these patients reported
441 "unknown" or "other" origins, suggesting complexities in identity and societal perceptions
442 within multicultural societies (Figure S10)[46].

443 The phrasing of the OFSEP-HD survey question, “From which geographical region are
444 you originating from?”, may lead to diverse interpretations, such as birthplace or upbringing[47].
445 Other biases, like form completion errors or practitioner misinterpretations, should also be
446 considered[48]. Using the example of the MS French population from North-African origins, we

447 highlight the misuse of self-reported origins as population descriptors, which can perpetuate
448 inaccurate genetic ancestry presuppositions. We recommend against using such descriptors
449 alone as genetic ancestry proxies. However, despite the contentious nature of geographical
450 origins and ethnicity as population descriptors, we suggest not fully dismissing them in
451 genetics and genomics research, as they provide insights to better characterize groups
452 experiencing healthcare disparities, including genomic healthcare access[49].

453 Our findings indicate a stronger alignment between reported North-African origins and
454 genetic ancestry when both parents share the same origins. Discrepancies occur with differing
455 parental origins. Caution is advised when using parental origins as a genetic ancestry proxy, as
456 it may not fully represent the child's genetic background, particularly with only one North-
457 African parent. Among other genetic clusters, self-reported North-African origins show a
458 significantly higher ($p<1e-4$) African genomic ancestry proportion (15.83%±12.82%) compared
459 to other origins (2.54%±0.36%). However, the high variability (MoE95%=12.82%) suggests
460 diverse reasons for self-reporting North-African origins.

461 Log-additive models for PRS calculation offer no significant advantage over the simpler
462 SRAC approach for non-MHC autosomal SNPs, due to similar effect sizes and nearly identical
463 impacts on PRS (0R≈1; Figure S22A)[8,44]. A log-additive approach is more suitable for MHC
464 variants with larger effect sizes, unlike SRAC where all variants have equal impact (Figure
465 S22B). Most MS variants, identified in European populations, may not fully capture genetic
466 risks for other ancestries[37,39]. Our findings reveal significant PRS score differences and
467 significantly lower *HLA-DRB1*15:01* allele proportions in patients from North-African genetic
468 ancestry, suggesting reduced MS susceptibility. This must be contextualized within European-
469 focused research, healthcare disparities, and environmental factors like vitamin D,

470 contributing to a South-to-North prevalence gradient[3,50]. Current PRS may perpetuate
471 healthcare disparities, emphasizing the need for more inclusive models[51].

472 Finally, our adapted methodology of the Avatar algorithm significantly improved
473 fidelity and privacy compared to single-generation synthetic avatar data[41]. Averaging parallel
474 synthetic data generation in a patient-wise manner enhanced fidelity and privacy by
475 combining multiple synthetic patients for each original patient. All privacy and signal metrics
476 significantly exceeded standard thresholds, demonstrating the methodology's flexibility in
477 producing high-fidelity synthetic datasets compliant with GDPR[41]. This makes it a viable tool
478 for future research. However, our study only performed univariate comparisons between
479 original and synthetic datasets. Future work should verify the preservation of multivariate
480 dependencies to maintain signal integrity.

481 **Conclusion**

482 In this study, we characterized the genetic landscape of the French OFSEP-HD MS
483 cohort into seven genetic ancestral clusters. We found that self-reported origins significantly
484 underestimate North-African ancestry, highlighting their limitations as proxies for genetic
485 ancestry; parental origins proved more reliable. In addition, French MS patients from North-
486 African genetic ancestry exhibited admixture profiles closer to European than to sub-Saharan
487 African populations. We also imputed HLA alleles/haplotypes and computed MS PRS.
488 However, our findings demonstrate that current MS susceptibility variants are ill-suited for
489 calculating PRS in non-European ancestries. This work establishes the first comprehensive
490 framework for analyzing genetic diversity in the French OFSEP-HD cohort and propose a first
491 open synthetic and GDPR-enforced genetic dataset for this population, including re-
492 identification risk assessment. This advances synthetic data generation methodologies and

paves the way for new methods tailored to genetic data's sensitivity and complexity. Overall, it establishes a foundation for future MS research in France, addressing the national context of limited genetic ancestral studies due to cultural and ethical considerations.

Disclosure of interests

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

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

D-A. Laplaud served on scientific advisory boards for Alexion, BMS, Roche, Sanofi, Novartis, Merck, Janssen and Biogen, received conference travel support and/or speaker honoraria from Alexion, Novartis, Biogen, Roche, Sanofi, BMS and Merck and received research support from Fondation ARSEP, Fondation EDMUS and Agence Nationale de la Recherche.

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

E. Leray reported consulting and lecture fees or travel grants from Alexion, Biogen, Genzyme, MedDay, Merck, Novartis and Roche.

514 E. Thouvenot reported consulting and lecturing fees, travel grants or unconditional
515 research support from the following pharmaceutical companies: Actelion, Biogen, Janssen,
516 Merck Serono, Novartis, Roche, Sanofi.

517 E. Le Page consulting or lectures, and invitations for national and international
518 congresses from Biogen, Merck, Teva, Sanofi-Genzyme, Novartis Alexion, research support
519 from Teva and Biogen academic research grants from PHRC and LFSEP, and travel grant from
520 ARSEP Foundation.

521 J. De Seze reported consulting and lecturing fees, travel grants and unconditional
522 research support from Biogen, Genzyme, Novartis, Roche, Sanofi Aventis and Teva Pharma.

523 J. Ciron reported consulting, serving on a scientific advisory board, speaking, or other
524 activities with Biogen, Novartis, Merck, Sanofi, Roche, Alexion and Horizon Therapeutics-
525 Amgen.

526 P. Clavelou reported consulting and lecturing fees, travel grants and unconditional
527 research support from Biogen, Janssen, Medday, Merck, Novartis, Roche, Sanofi-Genzyme and
528 Teva Pharma.

529 E. Berger received honoraria and reported consulting fees from Novartis, Sanofi
530 Aventis, Biogen, Genzyme, Roche and Teva Pharma.

531 A. Ruet received honoraria for meeting speaking from Merck, Alexion, Horizon Th, and
532 Sanofi Genzyme. A Ruet received support for traveling from Biogen, Novartis, and Merck. Her
533 institution received research grants from Biogen, Roche, Sanofi-Genzyme, and BMS.

534 T. Moreau received fees as scientific adviser from Biogen, Medday, Novartis, Genzyme,
535 Sanofi.

Q. Casez received personal fees from Biogen, Roche, Merck, Novartis, Jansen and Sanofi and reported nn financial support for travelling and congress for Roche, Merck, and Novartis.

P. Labauge received fees and grants from Biogen, Sanofi Genzyme, Novartis, Alexion, Merck.

W. Abir received expert testimony from Roche and travel grants from Biogen.

G. Defer received Consulting and lecturing fees for Biogen, Novartis, Genzyme, Merck-Serono, Roche and Teva and funding for travel from Merck Serono, Biogen, Sanofi-Genzyme, Novartis and Teva. Institution granted for research supporting from Merck Serono, Biogen, Genzyme and Novartis.

E. Maillart received consulting and lecturing fees from Alexion, Biogen, Horizon, Janssen, Merck Serono, Novartis, Roche, Sandoz, Sanofi-Genzyme, Teva Pharmaceuticals, and research support from Biogen.

H. Zephir received consulting or lectures, and invitations for national and international congresses from Biogen, Merck, Teva, Sanofi-Genzyme, Novartis and Bayer, as well as research support from Teva and Roche, and academic research grants from Académie de Médecine, LFSEP, FHU Imminent and ARSEP Foundation.

H. Olivier received consulting and lecturing fees from Bayer Schering, Merck, Teva, Genzyme, Novartis, Almirall and BiogenIdec, travel grants from Novartis, Teva, Genzyme, Merck Serono and Biogen Idec and research support from Roche, Merck and Novartis.

Acknowledgments

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" programme, under the reference ANR-10-COHO-002" OFSEP; and the support of the "Eugène Devic EDMUS Foundation against multiple sclerosis"; as well as the contribution of the "ARSEP Foundation" and the contribution of the Biological Resources Centres (BRC) that provided biological sample. In addition, we thank the historical contributors of REFGENSEP, Lena Guillot-Noël and Isabelle Rebeix.

This work was also 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

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

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

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

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

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

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

589 **Writing – Original Draft**: J.P.

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

592 **Web ressources**

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

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

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

596 TopMed Michigan Imputation Server: https://imputation.biodatacataslyst.nhlbi.nih.gov/

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

598 EasyHLA : https://easyhla.univ-nantes.fr/

599 HLA2Haplo (EasyHLA feature): https://easyhla.univ-nantes.fr/prediction/prediction.php

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

601 Genome-wide Complex Trait Analysis (GCTA) 1v.94.1 Linux:

602 https://yanglab.westlake.edu.cn/software/gcta/#Overview

603 Data and code availability

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

605 Declaration of generative AI in scientific writing

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

610 References

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

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

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

617 4. Sawad, A. B., Seoane-Vazquez, E., Rodriguez-Monguio, R. & Turkistani, F. Cost-effectiveness
618 of different strategies for treatment relapsing-remitting multiple sclerosis. *J Comp Eff Res* **6**, 97–108
619 (2017).

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

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

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

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

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

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

630 11. Botigué, L. R. *et al.* Gene flow from North Africa contributes to differential human genetic
631 diversity in southern Europe. *Proceedings of the National Academy of Sciences* **110**, 11791–11796
632 (2013).

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

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

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

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

641 16. OFSEP - Accueil. https://www.ofsep.org/fr/.

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

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

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

648 20. EDMUS, a European database for multiple sclerosis. Confavreux C, Compston DAS, Hommes
649 OR, McDonald WI, Thompson AJ. J Neurol Neurosurg Psychiatry 1992; 55: 671-676

650 21. Brocard, G. *et al.* The biological sample collection of the OFSEP French MS registry: An
651 essential tool dedicated to researchers. *Multiple Sclerosis and Related Disorders* **77**, 104872 (2023).

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

654 23. Lublin, F. D. *et al.* Defining the clinical course of multiple sclerosis. *Neurology* **83**, 278–286
655 (2014).

656 24. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
657 *Nature* **526**, 68–74 (2015).

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

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

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

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

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

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

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

672 32. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* **22**, 276–282
673 (2012).

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

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

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

680 36. Isobe, N. *et al.* Genetic burden in multiple sclerosis families. *Genes Immun* **14**, 434–440
681 (2013).

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

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

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

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

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

691 42. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27**,
692 857–871 (1971).

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

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

698 45. Armstrong, R. A. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* **34**, 502–508
699 (2014).

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

702 47. Gotlieb, R. *et al.* Accuracy in Patient Understanding of Common Medical Phrases. *JAMA*
703 *Network Open* **5**, e2242972 (2022).

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

706 49. Feero, W. G. *et al.* Guidance on use of race, ethnicity, and geographic origin as proxies for
707 genetic ancestry groups in biomedical publications. *The American Journal of Human Genetics* **111**,
708 621–623 (2024).

709 50. Simpson, S. *et al.* Latitude continues to be significantly associated with the prevalence of
710 multiple sclerosis: an updated meta-analysis. *J Neurol Neurosurg Psychiatry* **90**, 1193–1200 (2019).

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

713