Characterization of the common genetic variation in the Spanish population of Navarre

Authors:
Alberto Maillo, alberto.ruizdeinfante@kaust.edu.sa, Spain[1,2]
Estefania Huergo, ehuergoi@navarra.es, Spain[1,#]
María Apellániz-Ruiz, mv.apellaniz.ruiz@navarra.es, Spain[3,#]
Edurne Urrutia, eurrutil@navarra.es, Spain[1,3,#]
María Miranda, mmirandp@navarra.es, Spain[3]
Josefa Salgado, jsalgadg@navarra.es, Spain[3,4,5]
Sara Pasalodos-Sánchez, spasalos@navarra.es, Spain[3]
Luna Delgado-Mora, lunadelde@gmail.com, Spain[3,6]
Óscar Teijido, oteijidh@navarra.es, Spain[3]
Ibai Goicoechea, igoicoeo@nasertic.es, Spain[7]
Rosario Carmona, rosariom.carmona@juntadeandalucia.es, Spain[8,9,10,11]
Javier Perez-Florido, javier.perez.florido.sspa@juntadeandalucia.es, Spain[8,9,10,11]
Virginia Aquino, virginia.aquino@juntadeandalucia.es, Spain[8]
Daniel Lopez-Lopez, daniel.lopez.lopez@juntadeandalucia.es, Spain[8,9,11]
María Peña-Chilet, maria_pena@iislafe.es, Spain[8,11]
Sergi Beltrán, sergi.beltran@cnag.eu, Spain[12,13,14]
Joaquín Dopazo, joaquin.dopazo@juntadeandalucia.es, Spain[8,9,10,11]
Iñigo Lasa, ilasa@unavarra.es, Spain[15]
Juan José Beloqui, jj.beloqui.lizaso@navarra.es, Spain[3]
NAGEN-scheme (*see list at the end*)
Ángel Alonso, aalonsos@navarra.es, Spain[3*]
David Gomez-Cabrero, david.gomez.cabrero@navarra.es, Spain[1,2*]

Affiliations:

1: Translational Bioinformatics Unit, Navarrabiomed, Hospital Universitario de Navarra (HUN), Universidad Pública de Navarra (UPNA), IdiSNA, Pamplona, Spain
2: Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.
3: Genomics Medicine Unit, Navarrabiomed, Hospital Universitario de Navarra (HUN), Universidad Pública de Navarra (UPNA), IdiSNA, Pamplona, Spain
4: Servicio de Genética Médica-Hospital Universitario de Navarra (HUN), Pamplona, Spain.
5: Dp. Bioquímica y Biología Molecular-Universidad Pública de Navarra (UPNA), Pamplona, Spain.
6: Instituto de Genética Médica y Molecular (INGEMM), Hospital Universitario La Paz, Madrid, Spain.
7: Department of Personalized Medicine, NASERTIC, Government of Navarra, Pamplona, Spain.
8: Computational Medicine Platform, Andalusian Public Foundation Progress and Health-FPS, Sevilla, Spain.

9: Institute of Biomedicine of Seville, IBiS, University Hospital Virgen del Rocio/CSIC/University of Sevilla, Sevilla, Spain.
10: FPS/ELIXIR-ES, Fundación Progreso y Salud (FPS), CDCA, Hospital Virgen del Rocio, Sevilla, Spain.
11: Biomedical Research Networking Center in Rare Diseases (CIBERER), Health Institute Carlos III, 28029 Madrid, Spain.
12: CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.
13: Universitat Pompeu Fabra (UPF), Barcelona, Spain.
14: Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain.
15: Laboratory of Microbial Pathogenesis, Navarrabiomed, Pamplona, Spain.

*corresponding: aalonsos@navarra.es and david.gomez.cabrero@navarra.es . #contributed equally

ABSTRACT

Large-scale genomic studies have significantly increased our knowledge of genetic variability across populations. Regional genetic profiling is essential for distinguishing common benign variants from disease-causing ones. To this end, we conducted a comprehensive characterization of exonic variants in the population of Navarre (Spain), utilizing whole genome sequencing data from 358 unrelated individuals of Spanish origin. Our analysis revealed 61,410 biallelic single nucleotide variants (SNV) within the Navarrese cohort, with 35% classified as common (MAF > 1%). By comparing allele frequency data from 1000 Genome Project (excluding the Iberian cohort of Spain, IBS), Genome Aggregation Database, and a Spanish cohort (including IBS individuals and data from Medical Genome Project), we identified 1,069 SNVs common in Navarre but rare (MAF ≤ 1%) in all other populations. We further corroborated this observation with a second regional cohort of 239 unrelated exomes, which confirmed 676 of the 1,069 SNVs as common in Navarre. In conclusion, this study highlights the importance of population-specific characterization of genetic variation to improve allele frequency filtering in sequencing data analysis to identify disease-causing variants.

Keywords: personalized medicine, whole genome sequencing WGS, whole exome sequencing WES, single nucleotide variant SNV, population frequencies, genetic variability.

INTRODUCTION

In recent years, the use of NGS in patient healthcare has increased due to technological advances, cost reduction, and enhanced efficiency.[1] The advancement of NGS spans a spectrum of applications, encompassing whole exome/genome sequencing (WES/WGS). These technologies revealed a wealth of genetic variants, necessitating the implementation of filters to narrow down the list of candidate variants. In this regard, the availability of population-specific catalogues of common variants enables the identification of rare variants,[2] such as the international initiatives 1000 Genome Project (1KGP)[3] and Genome Aggregation Database (gnomAD).[4] Moreover, various countries like the UK,[5] USA,[6] and Japan[7] have established their databases. In Spain, for instance, the Medical Genome Project (MGP) compiles data from unrelated healthy individuals.[8,9]

In Navarre, a 650,000 population region of north-eastern Spain, the local Government supported the "*NAGEN scheme*" to integrate genomic data analysis into the regional public healthcare system. Nowadays, NAGEN has generated numerous WES/WGS and associated phenoclinical profiles in seven projects, including *NAGEN1000*, focused on rare diseases, and *pharmaNAGEN* on pharmacogenomics in patients with inflammatory bowel diseases.[10] The NAGEN strategy's success hinges on identifying population-specific common variants to establish a comprehensive Navarrese population frequency catalogue.

In this study (Fig. 1), we aimed to identify and characterize common exonic variants specific to the Navarrese population. Firstly, we identified common single nucleotide variants (SNVs) in Navarre and rare in other populations. Secondly, we validated the allele frequency of these variants in another Navarrese cohort with exome data. Finally, we annotated the resulting

95 variants using genomic databases, and their clinical and pharmacological effects and
96 pathogenicity were assessed. Additionally, we conducted functional enrichment analyses to
97 provide further insights. The results will significantly contribute to advancing personalized
98 medicine in Navarre.

99 ## MATERIAL AND METHODS

100 ***NAGEN1000* and *pharmaNAGEN***

101 In Navarre, the local Government supports the "*NAGEN scheme*", integrating genomic data
102 analysis into regional healthcare across seven projects. In the *NAGEN1000* project, 688
103 participants were recruited to uncover the underlying genetic causes of disease using WGS data.
104 These individuals belonged to 294 families, predominantly trios, affected by rare disorders.
105 Likewise, in the *pharmaNAGEN* project, 274 patients with Crohn's disease or ulcerative colitis
106 were recruited to analyze variants related to drug efficacy and toxicity on WES data.[10]

107

108 **Whole genome sequencing and data analysis**

109 High-quality DNA samples from peripheral blood were used in the *NAGEN1000* project to
110 construct short-insert paired-end libraries with an average insert size of 400 bp. DNA
111 fragmentation was performed with Covaris S2 and capillary electrophoresis with Bioanalyzer 2100
112 (Agilent). Libraries were sequenced on a NovaSeq 6000 (Illumina) with a read length of 2x150 bp.
113 A 30X coverage per sample was targeted. After quality control assessment using FASTQC
114 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), the sequenced data was aligned
115 to hs37d5 version of human genome reference GRCh37/hg19 using GEM3.[11] Optical and
116 duplicated reads were flagged with Picard MarkDuplicates
117 (https://broadinstitute.github.io/picard/). Following the established Genome Analysis Toolkit
118 (GATK) best practices pipeline (v3.8),[12] indel realignments and recalibration were applied to the
119 previous BAM files. Variant calling on each sample's BAM file was performed using
120 HaplotypeCaller, with default parameters, resulting in a gVCF file. WGS was conducted at Centro

121 Nacional de Análisis Genómico (CNAG, Barcelona, Spain) and stored at Navarra de Servicios y
122 Tecnología (NASERTIC, Navarre, Spain).
123
124 **Whole exome sequencing and data analysis**
125 In the *pharmaNAGEN* project, germline DNA was extracted from saliva or blood samples using
126 DNA Blood Maxi Kit (Qiagen) and sequenced with Nextera DNA Exome kit. The raw data was
127 aligned to GRCh37/hg19 genome, sourced from UCSC (https://genome.ucsc.edu/), utilizing
128 BWA.[13] The resulting BAM files were marked using Picard
129 (https://broadinstitute.github.io/picard/). Utilizing GATK v4.1.0,[12] an updated version, eliminates
130 the need for the indel realignment step. Recalibration and variant calling were executed with
131 BQSR and Haplotype tools. This process yielded the final gVCF file for each sample. WES was
132 conducted at CNAG and stored at NASERTIC.
133
134 **Individual selection**
135 Selection criteria included unrelated individuals and Spanish ancestry. To assess relatedness
136 among individuals, identity-by-descent (IBD) was calculated using the method of moments
137 (MoM) with the R package SNPRelate.[14] For validation of ethnicity, the CSVS tool[9] was used to
138 determine the degree of alignment of a sample with the genetic variability of the Spanish
139 population. Individuals with a score equal to or higher than 0.9 were categorized as being of
140 Spanish ancestry.
141 Applying these selection criteria to the *NAGEN1000* project resulted in a cohort of 358
142 participants denominated NAVARREsel. Within this group, 127 individuals had been diagnosed

143 with various monogenic diseases, and the most prevalent conditions were polycystic kidney
144 disease (14/127), breast cancer (10/127), and hereditary ataxia (8/127).
145 Regarding the validation dataset, named NAVARREval, the same inclusion criteria were applied to
146 the *pharmaNAGEN* project, resulting in 239 participants with Crohn's disease (153/239) and/or
147 ulcerative colitis (86/239).
148

149 **Variant quality control and filtering in NAVARREsel**

150 A 358 multi-sample gVCF was generated with the NAVARREsel cohort, and biallelic SNVs located
151 in exonic regions were selected. The targeted exonic interval was extracted from Nextera
152 (https://support.illumina.com/sequencing/sequencing_kits/nextera-dna-
153 exome/downloads.html, downloaded on September 27, 2023). Variants with a read depth of less
154 than 10, a genotype-quality score below 50, or a call rate of less than 100% were removed.
155 Variants on the X and Y chromosomes were eliminated to avoid sex bias, as well as the
156 mitochondrial chromosome, given its complexity. Hardy-Weinberg equilibrium (HWE) score[15] was
157 calculated by PLINK (-hardy), and SNVs significantly deviated with p-value < 10^{-5} were excluded.
158

159 **Variant quality control and filtering in NAVARREval**

160 Genotype information for specific SNVs from NAVARREval samples was extracted using the
161 SelectVariants tool of GATK v4.1.0.[12] SNVs with a call rate lower than 80% and those exhibiting a
162 significant deviation from HWE (p-value < 10^{-5}) were excluded from validation.
163

164 **Variant annotation**

165 Variants were annotated using ANNOVAR (version available on October 24th, 2019,
166 https://annovar.openbioinformatics.org/).[16] The identification of known variants was performed
167 using the dbSNP database (version GCF_000001405.25, downloaded from
168 https://ftp.ncbi.nih.gov/snp/latest_release/VCF/).[17] SNVs were also annotated: 1) for clinical
169 significance by referring to ClinVar (v.20230930, https://www.ncbi.nlm.nih.gov/clinvar/),[18] OMIM
170 (downloaded on October 2, 2023, https://www.omim.org/),[19] VarSome,[20] and Franklin;[21] and 2)
171 for pharmacological relevance using PharmGKB (https://www.pharmgkb.org/).[22] Additionally, the
172 pathogenicity of the variants was evaluated using CADD (v1.6,
173 https://cadd.gs.washington.edu/),[23] REVEL score (v1.3, https://zenodo.org/records/7072866),[24]
174 spliceAI,[25] and Polyphen2 (http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads
175 downloaded on October 4, 2023).[26] Variants associated with Crohn's disease and/or ulcerative
176 colitis were identified by referencing the Inflammatory Bowel Disease (IBD) database (accessed
177 on October 10, 2023, from https://www.cbrc.kaust.edu.sa/ibd/index.php?p=ibd#).[27]

178

179 **Population projects**

180 The MGP project was a Spanish initiative, primarily featuring WES data from 267 healthy and
181 unrelated participants, mainly from Andalusia and Galicia (Spanish regions).[8]

182 The population data from 1KGP phase 3 encompassed 2,504 genome samples representing 26
183 populations.[28] European populations within this dataset included 503 samples from Europe
184 (EUR), such as British in England and Scotland (GBR), Finnish in Finland (FIN), Iberian population
185 in Spain (IBS), Toscani in Italy (TSI), and Utah residents with Northern and Western European

186 ancestry (CEU). Additionally, there were 504 samples from East Asia (EAS), 489 from South Asia
187 (SAS), 661 from Africa (AFR), and 347 from America (AMR), covering various populations.
188 The gnomAD genomes project v2.1.1 comprises 15,691 genomes and represents diverse
189 populations worldwide, including Africans, Americans, Asians, and Europeans. Within European
190 populations, the majority originated from North-western, Estonian, Finnish, and a smaller
191 representation from Southern Europe.[4]
192

193 **Population frequencies**

194 Based on the population projects described in the previous section, three populations were
195 generated for this study as reference: 1) gnomAD, with the original frequency from the gnomAD
196 genome project; 2) *1KGP_noIBS*, with the mean of all 1KGP population's frequencies, excluding
197 the IBS cohort of 107 individuals; 3) *spain*, combining the frequencies of the IBS and MGP cohorts.
198 The integration process for the *spain* population consisted of adding the number of total alternate
199 alleles divided by the sum of the total number of alleles across the two cohorts.
200

201 **Principal Components Analysis, Admixture, and F_{ST} analysis**

202 The original VCF files for each chromosome from the 1KGP phase3 were downloaded on
203 September 27, 2023, from https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.
204 Subsequently, these files were merged using the Picard MergeVCFs tool. Access to the raw data
205 from MGP was granted upon request, and the multi-sample VCF was retrieved for the EGA
206 repository at https://ega-archive.org/datasets/EGAD00001003101. Then, the VCF files from
207 1KGP, MGP, and NAVARREsel were combined using the GATK tool's CombineGVCFs function. The

208 resulting combined file was used to perform Principal Component Analysis (PCA) with R libraries

209 SNPRelate and SNPassoc, conduct ADMIXTURE (v1.3.0)[29] from 3 to 7 genetic components (K), and

210 calculate the mean pairwise F_{ST} values between populations using vcftools (v0.1.17).[30]

211

212 **Enrichment analysis**

213 Functional enrichment, including pathway (KEGG), biological process (GO), disease (OMIM), and

214 human phenotype ontology (HPO), was performed using WebGestalt

215 (https://www.webgestalt.org/).[31]

216 # RESULTS

217 ## Navarrese discovery cohort

218 The *NAGEN1000* WGS Navarrese project comprised 688 individuals from 294 families (mainly
219 trios) with a rare disease. The WGS was conducted with a mean coverage of 30X, providing
220 comprehensive genomic data across the entire genome. For our study, a subset of this cohort was
221 selected satisfying two criteria: unrelatedness and Spanish ancestry. This result yielded 358
222 individuals, referred to as NAVARREsel.

223 Then, biallelic SNVs on chromosomes 1 to 22, from the exonic region covered by the Nextera
224 Exome Enrichment kit, were extracted. Subsequently, variants with read depth < 10, genotype
225 quality < 50, or missing genotype in at least one sample were filtered out. Additionally, sites
226 significantly deviated from Hardy-Weinberg equilibrium (HWE, p-value < 10^{-5}) were removed.[32]
227 Finally, 61,410 SNVs remained, of which 21,174 were identified as common variants (MAF > 1%).
228 We observed that including additional individuals did not reveal new common variants, and
229 21,174 were achieved when considering over 100 individuals (Fig. S1a).

230

231 ## Genetic variation between Navarre, Spanish, and global populations

232 We performed a principal component analysis (PCA) on the shared variants between NAVARREsel,
233 1KGP, and MGP to depict its relationship. We observed a clear distinction between Navarre and
234 Asian/African populations, reflecting established genetic differences (Fig. 2a). Conversely, an
235 overlap is observed between Navarre and European populations, emphasizing their genetic
236 affinity. Thus, focusing on European populations (Fig. 2b), we observed that Navarrese individuals
237 are close to the Spanish populations (IBS and MGP) and exhibit proximity to Italian individuals

238 (TSI). This observation is supported when estimating the ancestries of the European populations
239 using ADMIXTURE.[29] The average number of ancestries in each population was calculated with
240 the optimal component K=3 (Fig. 2c). The Navarrese population showed the highest ancestral
241 proportion on component 1 at 61%, which started decreasing in the IBS and MGP populations to
242 30% and 20% respectively, and was nearly absent (0.1%) in Finland (FIN). In contrast, component
243 2 was predominant in the Finnish population (99%), while being the lowest in the Navarrese
244 cohort at 7%.
245 To further analyse the genetic differentiation, we calculated the mean pairwise F_{ST} values. The
246 lower F_{ST} value indicates greater similarity between populations. This occurred when comparing
247 Navarre with the Spanish ($F_{ST(Navarre\text{-}IBS)}$ = 0.0001 and $F_{ST(Navarre\text{-}MGP)}$ = 0.0007) and Italian ($F_{ST(Navarre\text{-}}$
248 $_{TSI)}$ = 0.0014) populations. In contrast, the highest differentiation was observed against East-Asian
249 and African populations ($F_{ST(Navarre\text{-}EAS)}$ = 0.0328, $F_{ST(Navarre\text{-}AFR)}$ = 0.0434, Table S1).
250 These findings, aligning with biological expectations, underscore the regional and continental
251 genetic affinities, providing insights into historical populations and evolutionary dynamics.
252

253 ## Exclusive common variants in Navarre

254 To identify exclusive Navarrese common variants, we examined allele frequency among Navarre
255 population and the three referenced populations: *1KGP_noIBS*, gnomAD, and *spain*. A
256 comparison of the MAF revealed that most variants (17,532 SNVs) were classified as common
257 (MAF > 1%) across the four populations. However, 835 variants exhibited higher prevalence solely
258 in Spanish cohorts (Navarre and *spain*). Specifically, 1,069 SNVs were identified as common in
259 Navarre, and rare (MAF ≤ 1%) in the rest (Fig. 3a).

260 To validate these 1,069 variants, we used the NAVARREval cohort, a subset of 239 WES unrelated
261 individuals of Spanish descent from the current Navarrese population and diagnosed with Crohn's
262 disease (159/239) or ulcerative colitis (86/239) (*pharmaNAGEN* project). Before validation, we
263 assessed the association of these SNVs with these conditions by cross-referencing them with
264 reported variants in the Inflammatory Bowel Disease database, which catalogues variants highly
265 linked to the mentioned diseases.[27] The absence of the 1,069 SNVs in this database ensured an
266 unbiased and robust validation process.
267 Among the 1,069 variants initially identified, 998 were detected in NAVARREval with a call rate
268 greater than 80% and demonstrated conformity to HWE. Notably, 676/998 of these SNVs (68%;
269 p-value = $2.2e^{-16}$) were consistently classified as common in NAVARREval, confirming their
270 prevalence within the Navarrese population (variants' information in Table S2). The validation
271 cohort was sufficient to validate the Navarrese common variants, reaching a plateau in Fig. S1b.
272 On the contrary, within the non-validated subset (322/998, 32%), 134 SNVs exhibited MAFs in
273 NAVARREsel that did not exceed a 2-fold difference in NAVARREval, indicating close MAF between
274 both datasets (Fig. S2). This exploration of MAF patterns ensures a comprehensive understanding
275 of the genetic landscape within the Navarre population and its stability across different datasets.
276

277 ## Characterization of common Navarrese variants

278 The annotation of the 676 common Navarrese SNVs revealed 227 synonymous, 371 missense,
279 and five loss-of-function (LoF) variants. These LoF variants were not reported in ClinVar
280 database[18] and were located in five distinct genes without an associated phenotype, according to

281 OMIM.[19] Following the ACMG guidelines for variant classification, four were classified as variants
282 of uncertain significance (VUS) and one as benign.[33]
283 Clinically, 264/676 SNVs were reported in ClinVar: 1/264 as a risk factor, 181/264 as benign/likely-
284 benign, 32/264 as VUS, 48/264 as having conflicting interpretations, and 2/264 as likely-
285 pathogenic. These likely-pathogenic missense variants were in *SCNN1B* [c.1688G>A p.Arg563Gln;
286 $MAF_{NAVARREsel}$=0.013, $MAF_{NAVARREval}$=0.016] and in *PTGIS* [c.824G>A p.Arg275Gln;
287 $MAF_{NAVARREsel}$=0.013, $MAF_{NAVARREval}$=0.021], associated with "low renin hypertension" and
288 "childhood-onset schizophrenia" respectively, according to ClinVar.[18] The evidence supporting
289 these associations is limited, with a score of 1 out of 4 and reviewed by a single submitter record.
290 Therefore, given its notable prevalence in the Navarrese population, observed in healthy and
291 affected (not related to this phenotype) individuals, these variants might be reconsidered and
292 reclassified as VUS under the ACMG guidelines.[34]
293 *In silico* analysis of the 676 variants with functional predictors revealed eight variants as
294 pathogenic by three different pathogenicity tools (REVEL_score > 0.8,[24] CADD > 20,[23] and
295 Polyphen indicating "probably" or "possibly").[26] However, a comprehensive examination of
296 clinical databases, including ClinVar,[18] Varsome,[20] and Franklin,[21] contradicted these predictions
297 based on ACMG criteria. Instead, the majority of these variants were classified as uncertain
298 significance (1/8), likely benign (5/8), and benign (2/8). This suggests that these variants are not
299 disease-causing.
300 Additionally, the variant in *BPIFB3* [c.387-1G>T; $MAF_{NAVARREsel}$=0.01536, $MAF_{NAVARREval}$=0.02092],
301 predicted to impact the canonical splicing acceptor site (spliceAI score = 0.99), was reclassified as

302 benign based on the allele frequency and the number of homozygotes as per ACMG criteria,
303 indicating no clinical relevance.[25]
304 Moreover, common Navarrese variants showed no impact on drug metabolism/efficacy,
305 according to PharmGKB.[22] They did not exhibit significant enrichment in pathways, biological
306 processes, related diseases, or phenotypic ontologies.
307

308 ## Refining disease-causing variant identification in the Navarrese population

309 We identified common variants in the Navarrese population, highlighting population-specific
310 importance in advancing personalized medicine. The aim was to improve the identification of
311 disease-causing variants during genetic diagnosis using NGS. Therefore, we selected 127 WGS
312 Navarrese patients from *NAGEN1000* project diagnosed with rare disorders and extracted exonic
313 SNVs on chromosomes 1 to 22, averaging 8,871 variants per patient.
314 We refined the variant list by excluding common variants from *1KGP_noIBS*, gnomAD, *spain*, and
315 Navarre. The Navarrese filtering emerged as the most stringent, resulting in 2.1% of the initial set,
316 compared to 2.7% with gnomAD, 2.9% with *spain* frequencies, and the least restrictive, 4.9% with
317 *1KGP_noIBS* (Fig. 3). This underscores the effectiveness of Navarrese-specific filter in prioritizing
318 and streamlining genetic investigations.
319

320 ## DISCUSSION

321 In this study, we aimed to enhance diagnostic precision in the Navarrese population by exploring
322 common population-specific variants. Utilizing WGS data from 358 individuals of Navarre, we
323 identified 61,410 SNVs, with 21,174 common. Genetic analysis shows affinity with European
324 populations and low differentiation with Spanish populations.

325 Focusing on exclusively common variants in Navarre compared with referenced populations, we
326 obtained 1,069 SNVs, of which 676 were validated in another Navarrese cohort. Of these, none
327 showed clinical or pharmacological relevance beyond what was observed in the Spanish
328 population.[35] This aligns with the expectation that common population variants are less likely to
329 be associated with disease ethology.

330 Our findings underscore the relevance of considering population-specific factors in genomic
331 diagnostics, which provides complementary insights alongside pangenome references.[36] In
332 conclusion, by identifying and excluding common variants within the Navarrese population, we
333 have successfully refined the identification of potential disease-causing variants, contributing to
334 the advancement of personalized medicine for individuals from Navarre. Further research will
335 enhance these insights for broader applications.

DECLARATION OF INTEREST

Competing interests:

The authors declare no competing interests.

Ethics approval:

NAGEN1000 and *PharmaNAGEN* were approved by the Navarra Ethics Committee for Clinical Research (CEIC Navarra).

FUNDING

NAGEN1000 and *PharmaNAGEN* were supported by Navarra Gov (Dirección General de Industria, Energia y Proyectos Estrategicos S3). GRANTS_NUMBERS: 0011-1411-2017-000032, 0011-1411-2018-000047.

AUTHOR INFORMATION

Authors and Affiliations

Translational Bioinformatics Unit, Navarrabiomed, Hospital Universitario de Navarra (HUN), Universidad Pública de Navarra (UPNA), IdiSNA, Pamplona, Spain
Alberto Maillo, Estefania Huergo, Edurne Urrutia & David Gomez-Cabrero

Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
Alberto Maillo & David Gomez-Cabrero

Genomics Medicine Unit, Navarrabiomed, Hospital Universitario de Navarra (HUN), Universidad Pública de Navarra (UPNA), IdiSNA, Pamplona, Spain
María Apellániz-Ruiz, Edurne Urrutia, María Miranda, Josefa Salgado, Sara Pasalodos-Sánchez, Luna Delgado-Mora, Óscar Teijido, Juan José Beloqui & Ángel Alonso

Servicio de Genética Médica-Hospital Universitario de Navarra (HUN), Pamplona, Spain
Josefa Salgado

Dp. Bioquímica y Biología Molecular-Universidad Pública de Navarra (UPNA), Pamplona, Spain
Josefa Salgado

Instituto de Genética Médica y Molecular (INGEMM), Hospital Universitario La Paz, Madrid, Spain
Luna Delgado-Mora

Department of Personalized Medicine, NASERTIC, Government of Navarra, Pamplona, Spain.
Ibai Goicoechea

Computational Medicine Platform, Andalusian Public Foundation Progress and Health-FPS, Sevilla, Spain.
Rosario Carmona, Javier Perez-Florido, Virginia Aquino, Daniel Lopez-Lopez, María Peña-Chilet & Joaquín Dopazo

Institute of Biomedicine of Seville, IBiS, University Hospital Virgen del Rocio/CSIC/University of Sevilla, Sevilla, Spain.
Rosario Carmona, Javier Perez-Florido, Daniel Lopez-Lopez & Joaquín Dopazo

FPS/ELIXIR-ES, Fundación Progreso y Salud (FPS), CDCA, Hospital Virgen del Rocio, Sevilla, Spain.
Rosario Carmona, Javier Perez-Florido & Joaquín Dopazo

Biomedical Research Networking Center in Rare Diseases (CIBERER), Health Institute Carlos III, 28029 Madrid, Spain

Rosario Carmona, Javier Perez-Florido, Daniel Lopez-Lopez, María Peña-Chilet & Joaquín Dopazo

CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
Sergi Beltrán

Universitat Pompeu Fabra (UPF), Barcelona, Spain
Sergi Beltrán

Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain
Sergi Beltrán

Laboratory of Microbial Pathogenesis, Navarrabiomed, Pamplona, Spain
Iñigo Lasa

Contributions

Conceptualisation (AM, AA, DGC); Clinical and sample collection: (MM, LDM, OT, JS, MAR, SPS); Formal analysis (AM, RC, JPF, VA, DLL, MPC); Data curation (AM); Investigation (AM, EH, MAR, EU, DGC); Funding acquisition (JB, AA); Visualisation (AM, EH, MAR, EU); Writing-original draft (AM, EH, MAR, EU, DGC) Writing-review & editing (AM, MAR, EH, EU, SB, SPS, IG, JD, IL, JB, AA, DGC).

Corresponding authors

Correspondence to Ángel Alonso & David Gomez-Cabrero (david.gomez.cabrero@navarra.es).

NAGEN scheme

Name	Department	NAGEN project
Anda Apiñaniz, Emma	Servicio de Endocrinología y Nutrición, HUN	NAGEN 1,000

Artigas López, Mercedes	Servicio de Genética Médica, HUN	NAGEN 1,000
Bandrés Elizalde, Eva	Servicio de Hematología, HUN	NAGEN 1,000
Basurte Elorz, Mª Teresa	Servicio de Cardiología, CHN	NAGEN 1,000
Brennan, Paul	NENC NHS Genomic Medicine Centre. Newcastle upon Tyne, UK.	NAGEN 1,000
Celaya Lecea, Concepción	Subdirección de Farmacia, SNS-O	PharmaNAGEN
Cuesta Zorita, Manuel Jesús	Servicio de Psiquiatría, Salud mental	NAGEN 1,000
Curi Chercoles, Sergio Miguel	Servicio de Neumología, HUN	NAGEN 1,000
De la Cruz Sánchez, Susana	Servicio de Oncología Médica, HUN	NAGEN 1,000
Erviti López, Juan	Subdirección de Farmacia, SNS-O	PharmaNAGEN
Fanlo Mateo, Patricia	Servicio de Medicina Interna, HUN	NAGEN 1,000
González, Luis Angel	AVANTIA 400+.	NAGEN 1,000
Gonzalo Etayo	Navarra de Servicios y Tecnología NASERTIC. Spain	NAGEN 1,000
Gorricho Mendívil, Javier	Subdirección de Farmacia, SNS-O	PharmaNAGEN

Guerra Lacunza, Ana	Servicio de Aparato Digestivo, HUN	NAGEN 1,000
Gut, Ivo	Centro Nacional de Análisis Genómicos CNAG. Spain	NAGEN 1,000
Ibáñez Bosch, Rosario	Servicio de Endocrinología y Nutrición, HUN	NAGEN 1,000
Jiménez, Jorge	Navarra de Servicios y Tecnología NASERTIC. Spain	NAGEN 1,000
Lasheras, Gorka	Navarra de Servicios y Tecnología NASERTIC. Spain	NAGEN 1,000
Lorea Bueno	Pharmamodelling	PharmaNAGEN
Maite Sarobe Carricas	Servicio de Farmacia Hospitalaria, HUN	PharmaNAGEN
Mendioroz Iriarte, Maite	Servicio de Neurología, HUN	NAGEN 1,000
Molinuevo Ruiz de Zarate, José Ignacio	Servicio de Oftalmología, HUN	NAGEN 1,000
Montes Díaz, Marta	Servicio de Anatomía Patológica, HUN	NAGEN 1,000
Navarro, Adela	Servicio de Cardiología, HUN	PharmaNAGEN
Onintza Sayar	Pharmamodelling	PharmaNAGEN
Pinillos, Iñaki	Navarra de Servicios y Tecnología NASERTIC. Spain	NAGEN 1,000

Purroy Irurzon, Carolina Eugenia	Servicio de Nefrología, HUN	NAGEN 1,000
Sagaseta de Ilurdoz Uranga, Mª Josefa	Servicio de Pediatría, HUN	NAGEN 1,000
Santesteban Muruzabal, Raquel	Servicio de Dermatología, HUN	NAGEN 1,000
Vicuña, Miren	Servicio de Digestivo, HUN	PharmaNAGEN
Viguria, Mª Cruz	Servicio de Hematología, HUN	PharmaNAGEN
Yoldi Petri, Mª Eugenia	Servicio de Pediatría, HUN	NAGEN 1,000
Zubicaray Ugarteche, José Jacinto	Servicio de Otorrinolaringología, HUN	NAGEN 1,000
Zudaire, Maite	Servicio de Hematología, HUN	PharmaNAGEN

404

DATA AND CODE AVAILABILITY

Data is available from the corresponding author upon reasonable request.

408 ## REFERENCES

409 1. Satam H, Joshi K, Mangrolia U, et al. Next-Generation Sequencing Technology: Current
410 Trends and Advancements. *Biology (Basel)*. 2023;12(7):997.
411 doi:10.3390/biology12070997
412 2. Fattahi Z, Beheshtian M, Mohseni M, et al. Iranome: A catalog of genomic variations in
413 the Iranian population. *Hum Mutat*. 2019;40(11):1968-1984. doi:10.1002/humu.23880
414 3. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human
415 genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-1073.
416 doi:10.1038/nature09534
417 4. Gudmundsson S, Singer-Berk M, Watts NA, et al. Variant interpretation using population
418 databases: Lessons from gnomAD. *Hum Mutat*. 2022;43(8):1012-1030.
419 doi:10.1002/humu.24309
420 5. Smetana J, Brož P. National Genome Initiatives in Europe and the United Kingdom in the
421 Era of Whole-Genome Sequencing: A Comprehensive Review. *Genes (Basel)*.
422 2022;13(3):556. doi:10.3390/genes13030556
423 6. Ramirez AH, Sulieman L, Schlueter DJ, et al. The All of Us Research Program: Data quality,
424 utility, and diversity. *Patterns*. 2022;3(8):100570. doi:10.1016/j.patter.2022.100570
425 7. Mitsuhashi N, Toyo-oka L, Katayama T, et al. TogoVar: A comprehensive Japanese genetic
426 variation database. *Hum Genome Var*. 2022;9(1):44. doi:10.1038/s41439-022-00222-9
427 8. Dopazo J, Amadoz A, Bleda M, et al. 267 Spanish Exomes Reveal Population-Specific
428 Differences in Disease-Related Genetic Variation. *Mol Biol Evol*. 2016;33(5):1205-1218.
429 doi:10.1093/molbev/msw005

430 9. Peña-Chilet M, Roldán G, Perez-Florido J, et al. CSVS, a crowdsourcing database of the
431 Spanish population genetic variability. *Nucleic Acids Res*. 2021;49(D1):D1130-D1137.
432 doi:10.1093/nar/gkaa794
433 10. NAGEN | Navarrabiomed. Accessed November 28, 2023.
434 https://www.navarrabiomed.es/en/nagen
435 11. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and
436 versatile alignment by filtration. *Nat Methods*. 2012;9(12):1185-1188.
437 doi:10.1038/nmeth.2221
438 12. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce
439 framework for analyzing next-generation DNA sequencing data. *Genome Res*.
440 2010;20(9):1297-1303. doi:10.1101/gr.107524.110
441 13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
442 *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324
443 14. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance
444 computing toolset for relatedness and principal component analysis of SNP data.
445 *Bioinformatics*. 2012;28(24):3326-3328. doi:10.1093/bioinformatics/bts606
446 15. Wigginton JE, Cutler DJ, Abecasis GR. A Note on Exact Tests of Hardy-Weinberg
447 Equilibrium. *The American Journal of Human Genetics*. 2005;76(5):887-893.
448 doi:10.1086/429864
449 16. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from
450 high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164-e164.
451 doi:10.1093/nar/gkq603

452 17. Sherry ST. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*.
453 2001;29(1):308-311. doi:10.1093/nar/29.1.308
454 18. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among
455 sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(Database
456 issue):D980-5. doi:10.1093/nar/gkt1113
457 19. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in
458 Man (OMIM(R)). *Nucleic Acids Res*. 2009;37(Database):D793-D796.
459 doi:10.1093/nar/gkn665
460 20. Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search
461 engine. *Bioinformatics*. 2019;35(11):1978-1980. doi:10.1093/bioinformatics/bty897
462 21. Franklin. Accessed November 28, 2023. https://franklin.genoox.com/clinical-db/home
463 22. Thorn CF, Klein TE, Altman RB. PharmGKB: The Pharmacogenomics Knowledge Base. In: ;
464 2013:311-320. doi:10.1007/978-1-62703-435-7_20
465 23. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the
466 deleteriousness of variants throughout the human genome. *Nucleic Acids Res*.
467 2019;47(D1):D886-D894. doi:10.1093/nar/gky1016
468 24. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting
469 the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*.
470 2016;99(4):877-885. doi:10.1016/j.ajhg.2016.08.016
471 25. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from
472 Primary Sequence with Deep Learning. *Cell*. 2019;176(3):535-548.e24.
473 doi:10.1016/j.cell.2018.12.015

474 26. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense
475 Mutations Using PolyPhen-2. *Curr Protoc Hum Genet*. 2013;76(1).
476 doi:10.1002/0471142905.hg0720s76
477 27. Khan F, Radovanovic A, Gojobori T, Kaur M. IBDDB: a manually curated and text-mining-
478 enhanced database of genes involved in inflammatory bowel disease. *Database*.
479 2021;2021. doi:10.1093/database/baab022
480 28. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation.
481 *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
482 29. Patterson N, Moorjani P, Luo Y, et al. Ancient Admixture in Human History. *Genetics*.
483 2012;192(3):1065-1093. doi:10.1534/genetics.112.145037
484 30. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools.
485 *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330
486 31. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive,
487 powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res*.
488 2017;45(W1):W130-W137. doi:10.1093/nar/gkx356
489 32. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality
490 control in genetic case-control association studies. *Nat Protoc*. 2010;5(9):1564-1573.
491 doi:10.1038/nprot.2010.116
492 33. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of
493 sequence variants: a joint consensus recommendation of the American College of
494 Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in*
495 *Medicine*. 2015;17(5):405-424. doi:10.1038/gim.2015.30

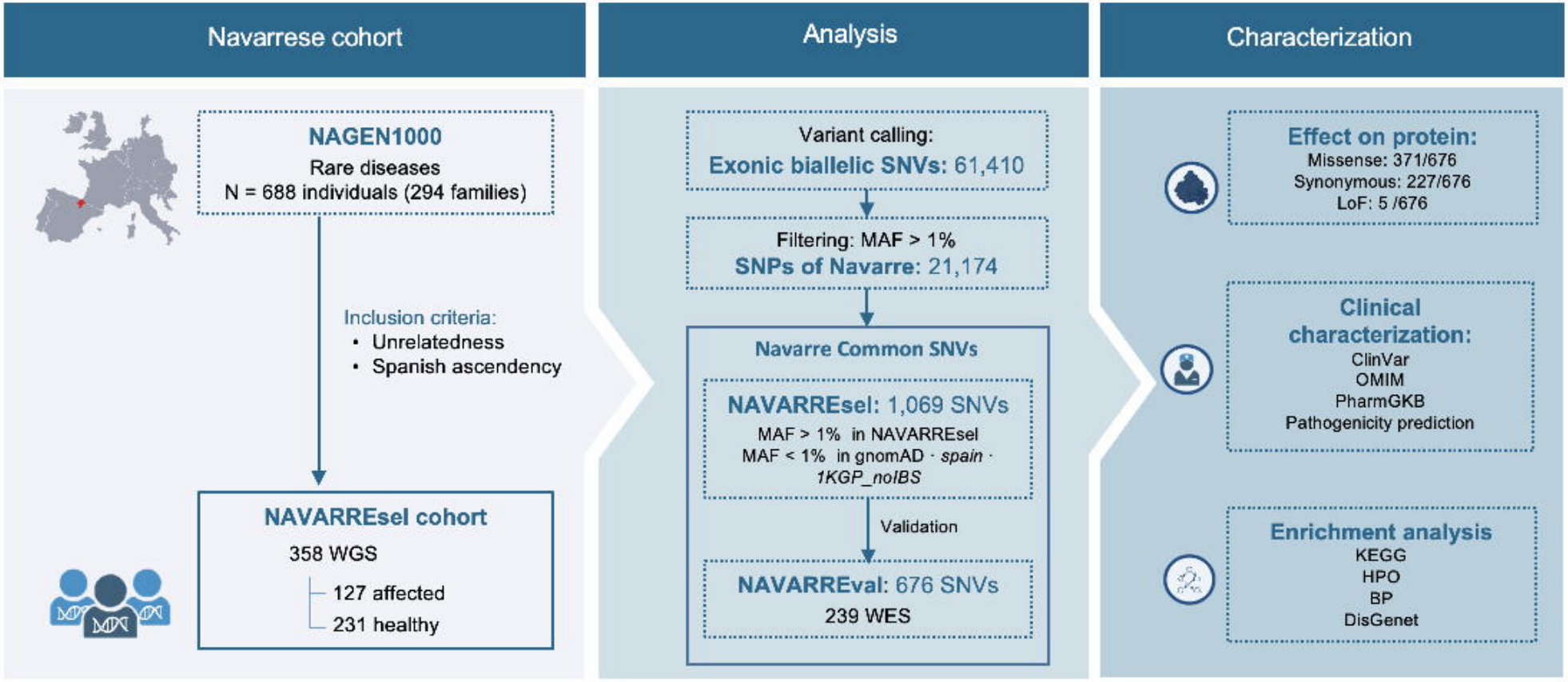
496 34. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP
497 Guidelines. *The American Journal of Human Genetics*. 2017;100(2):267-280.
498 doi:10.1016/j.ajhg.2017.01.004
499 35. Nunez-Torres R, Pita G, Peña-Chilet M, et al. A Comprehensive Analysis of 21 Actionable
500 Pharmacogenes in the Spanish Population: From Genetic Characterisation to Clinical
501 Impact. *Pharmaceutics*. 2023;15(4):1286. doi:10.3390/pharmaceutics15041286
502 36. Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*.
503 2023;617(7960):312-324. doi:10.1038/s41586-023-05896-x
504
505

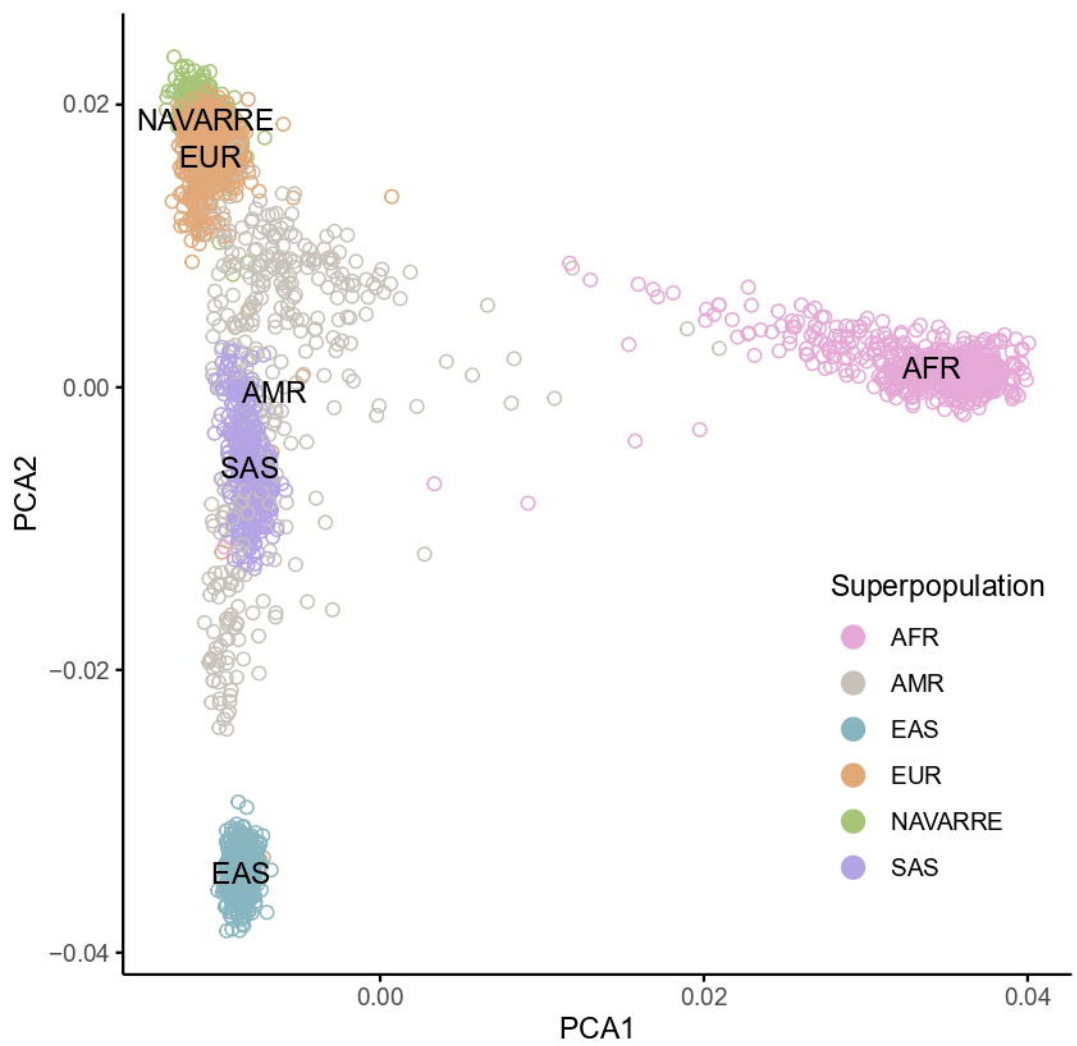
506 ## FIGURES LEGENDS

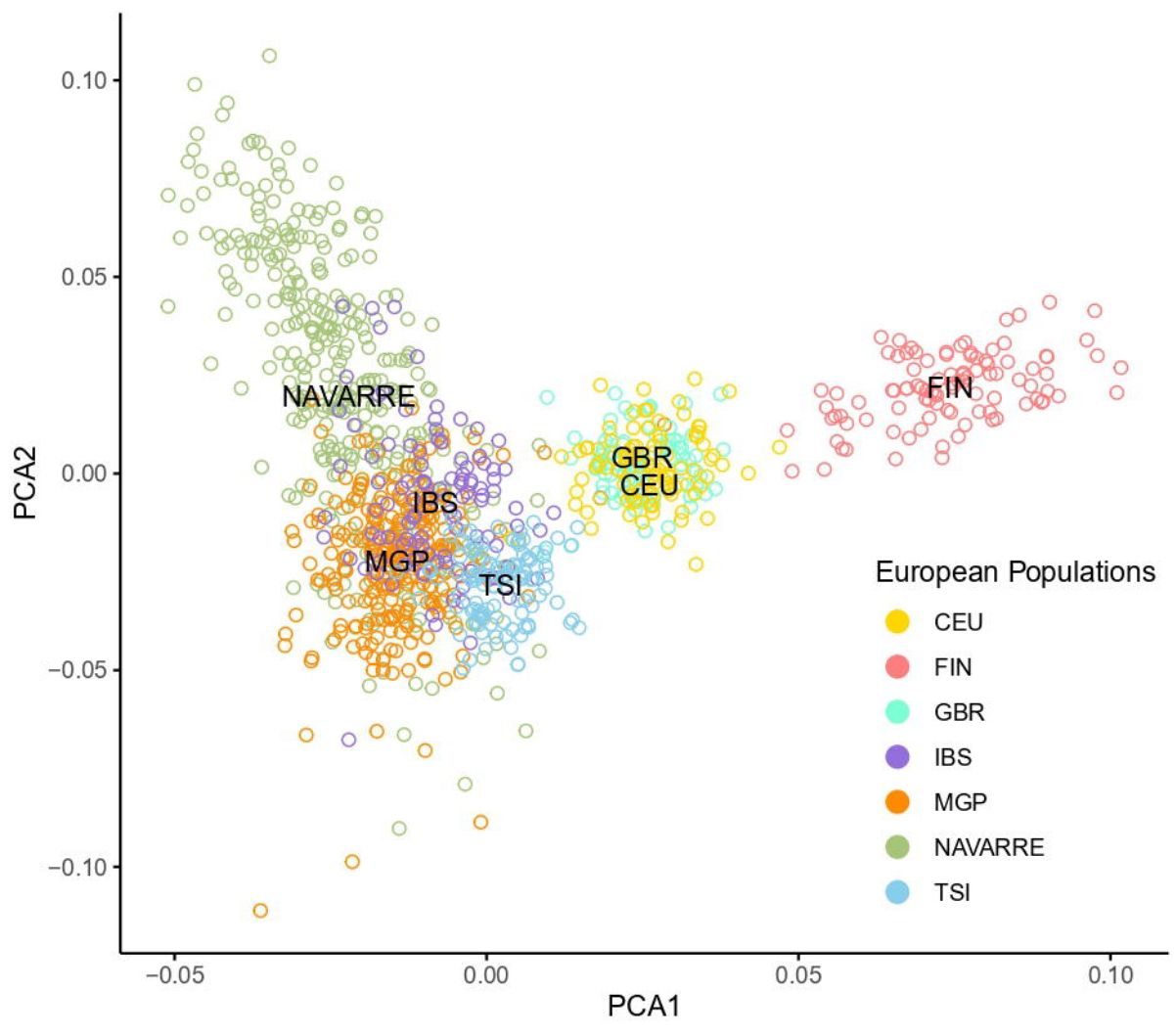
507 **Figure 1**: Workflow of this study. Abbreviations: *MGP*, Medical Genome Project; *1KGP*, 1000
508 Genomes Project; *1KGP_noIBS*, 1000 Genomes Project without Iberian population; *gnomAD,*
509 Genome Aggregation Database; *MAF*, minor allele frequency; *SNV*, single nucleotide variant; *SNP*,
510 single nucleotide polymorphism; *WGS*, whole genome sequencing; *WES*, whole exome
511 sequencing; *LoF*, Loss-of-function; HPO, Human Phenotype Ontology; BP, biological process.

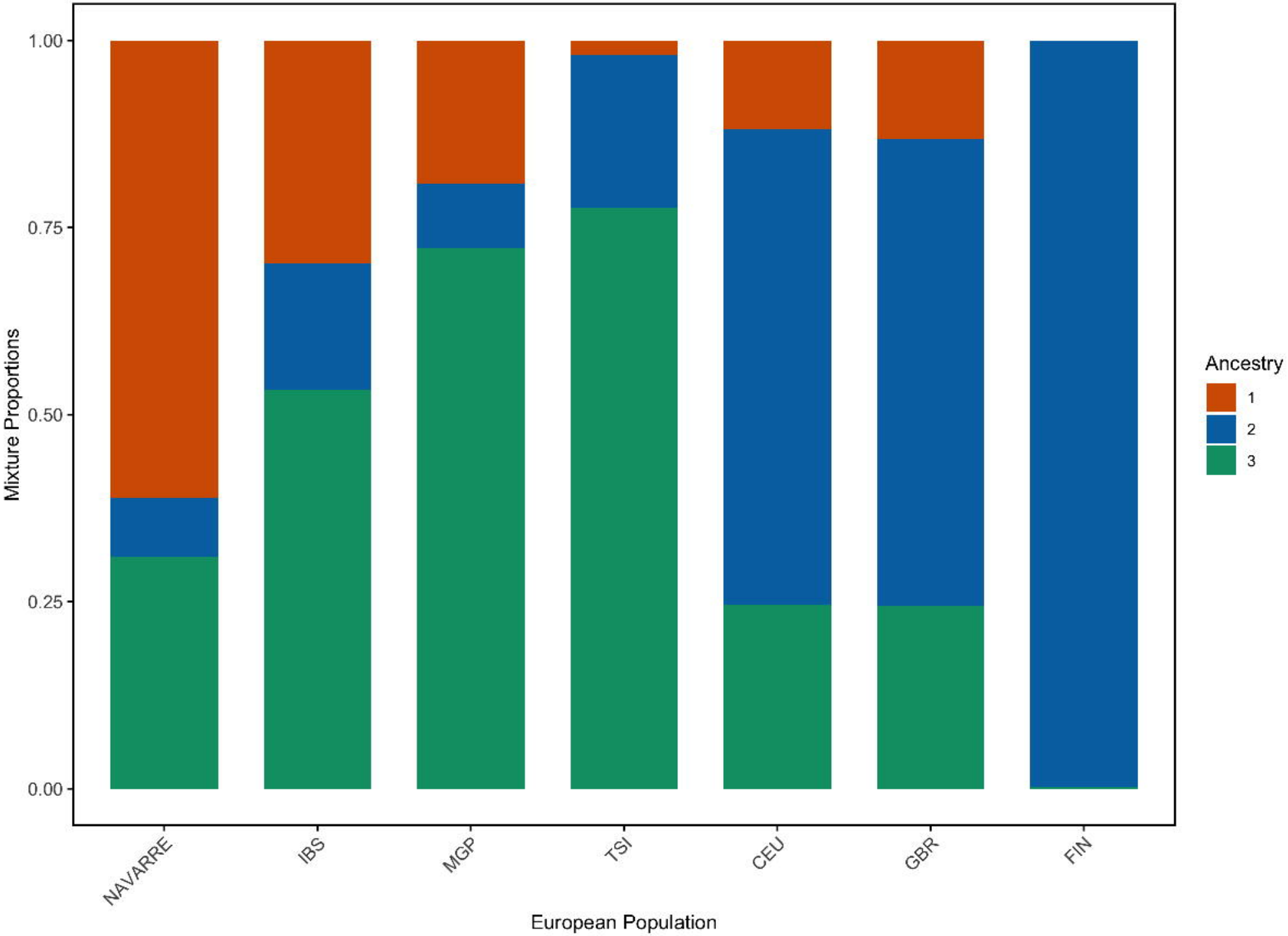
512 **Figure 2: a)** Principal component analysis of overlapped variants between NAVARREsel, MGP, and
513 1KGP (including all populations), and coloured by superpopulations. **b)** Principal component
514 analysis of overlapped variants between NAVARREsel, MGP, and 1KGP (including exclusively
515 European populations). **c)** Genetic admixture analysis of 1,128 individuals from 7 European
516 populations for the optimal *K* value = 3. Abbreviations: PCA, principal component analysis; AFR,
517 African populations; AMR, American populations; EAS, east-Asian populations; SAS, south-Asian
518 populations; EUR, European populations; IBS, Iberian populations in Spain; MGP, Medical
519 Genome Project; TSI, Toscani in Italy; CEU, Utah residents with Northern and Western European
520 ancestry; GBR, British in England and Scotland; FIN, Finnish in Finland.

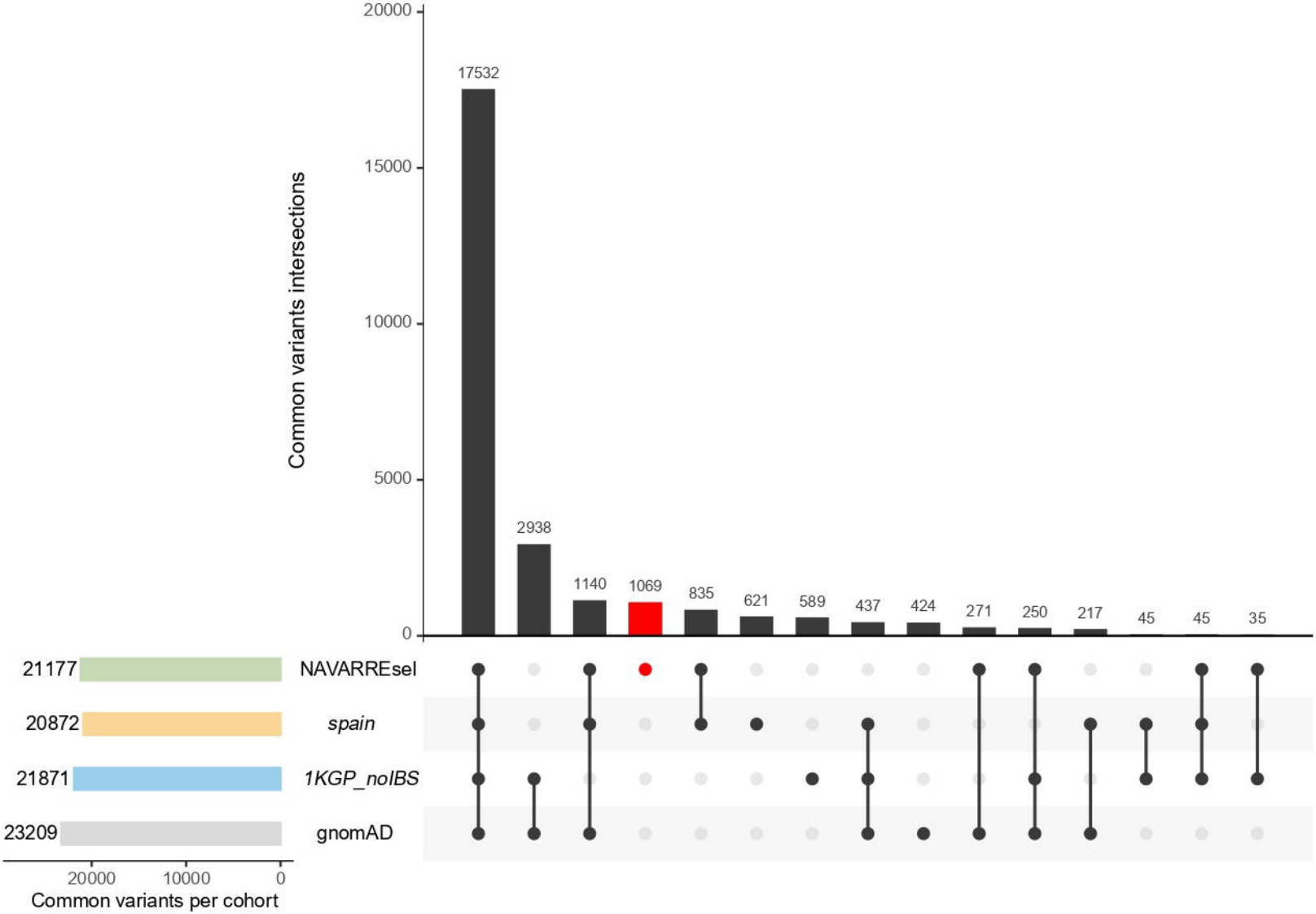
521 **Figure 3: a)** Upset plot of common variants (MAF > 1%) of each population: NAVARREsel, *spain*,
522 *1KGP_noIBS*, and gnomAD **b)** Resulting percentage of variants per patient (n=127) after removing
523 common variants from Navarre, *spain*, gnomAD, or *1KGP_noIBS* populations. The box plots
524 represent the median, upper, and lower quartiles by the centre line and box bounds, respectively.
525 Whiskers display the largest and smallest values within 1.5 times the interquartile range from the
526 quartiles. Abbreviations: *1KGP_noIBS*, 1000 Genomes Project without Iberian population;
527 gnomAD*,* Genome Aggregation Database; *spain*, integration of IBS and MGP populations.

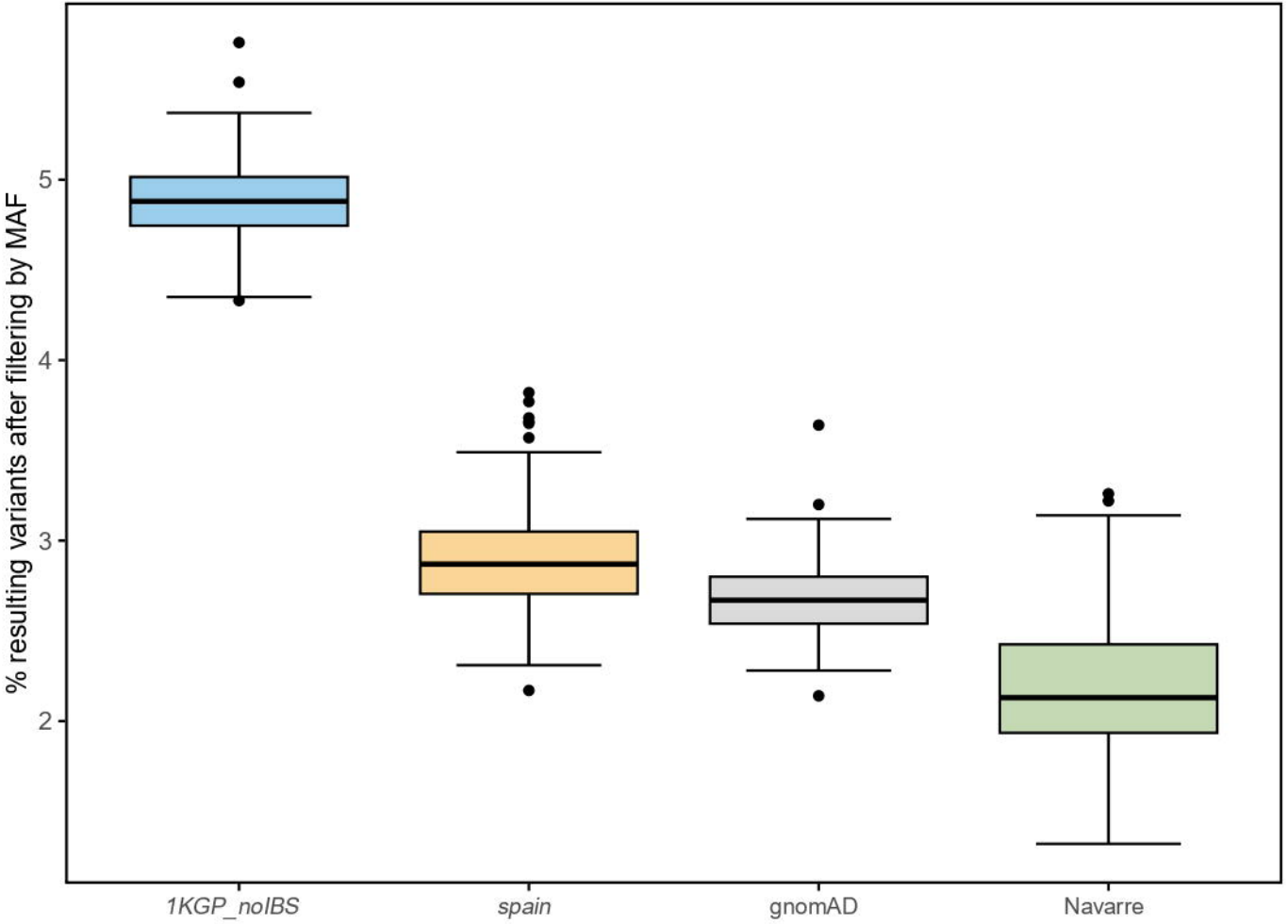

Navarrese cohort
Analysis
Characterization
NAGEN1000
Rare diseases
N = 688 individuals (294 families)
Inclusion criteria:
• Unrelatedness
• Spanish ascendency
NAVARREsel cohort
358 WGS
127 affected
231 healthy
Variant calling:
Exonic biallelic SNVs: 61,410
Filtering: MAF > 1%
SNPs of Navarre: 21,174
Navarre Common SNVs
NAVARREsel: 1,069 SNVs
MAF > 1% in NAVARREsel
MAF < 1% in gnomAD · spain · 1KGP_noIBS
Validation
NAVARREval: 676 SNVs
239 WES
Effect on protein:
Missense: 371/676
Synonymous: 227/676
LoF: 5 /676
Clinical characterization:
ClinVar
OMIM
PharmGKB
Pathogenicity prediction
Enrichment analysis
KEGG
HPO
BP
DisGenet


NAVARRE
EUR
AMR
SAS
AFR
EAS
Superpopulation
AFR
AMR
EAS
EUR
NAVARRE
SAS
PCA2
PCA1
0.02
0.00
−0.02
−0.04
0.00
0.02
0.04


0.10
0.05
0.00
−0.05
−0.10
PCA2
−0.05
0.00
0.05
0.10
PCA1
NAVARRE
IBS
MGP
TSI
GBR
CEU
FIN
European Populations
CEU
FIN
GBR
IBS
MGP
NAVARRE
TSI


1.00
0.75
0.50
0.25
0.00
Mixture Proportions
NAVARRE
IBS
MGP
TSI
CEU
GBR
FIN
European Population
Ancestry
1
2
3


Common variants intersections
20000
15000
10000
5000
0
17532
2938
1140
1069
835
621
589
437
424
271
250
217
45
45
35
21177
20872
21871
23209
NAVARREsel
spain
1KGP_noIBS
gnomAD
20000
10000
0
Common variants per cohort


% resulting variants after filtering by MAF
5
4
3
2
1KGP_noIBS
spain
gnomAD
Navarre