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]
Rosario Carmona, rosariom.carmona@juntadeandalucia.es, Spain[7,8,9,10]
Javier Perez-Florido, javier.perez.florido.sspa@juntadeandalucia.es, Spain[7,8,9,10]
Virginia Aquino, virginia.aquino@juntadeandalucia.es, Spain[7]
Daniel Lopez-Lopez, daniel.lopez.lopez@juntadeandalucia.es, Spain[7,8,10]
María Peña-Chilet, maria_pena@iislafe.es, Spain[7,10]
Sergi Beltrán, sergi.beltran@cnag.eu, Spain[11,12,13]
Joaquín Dopazo, joaquin.dopazo@juntadeandalucia.es, Spain[7,8,9,10]
Iñigo Lasa, ilasa@unavarra.es, Spain[14]
Juan José Beloqui, jj.beloqui.lizaso@navarra.es, Spain[3]
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: Computational Medicine Platform, Andalusian Public Foundation Progress and Health-FPS, Sevilla, Spain.
8: Institute of Biomedicine of Seville, IBiS, University Hospital Virgen del Rocio/CSIC/University of Sevilla, Sevilla, Spain.
9: FPS/ELIXIR-ES, Fundación Progreso y Salud (FPS), CDCA, Hospital Virgen del Rocio, Sevilla, Spain.

10: Biomedical Research Networking Center in Rare Diseases (CIBERER), Health Institute Carlos III, 28029 Madrid, Spain.
11: CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.
12: Universitat Pompeu Fabra (UPF), Barcelona, Spain.
13: Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain.
14: Laboratory of Microbial Pathogenesis, Navarrabiomed, Pamplona, Spain.

*corresponding. #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 those associated with disease. To this end, we conducted a comprehensive characterization of 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 using a minor allele frequency (MAF) > 1%. By comparing allele frequency data from 1000 Genome Project Phase 3 (excluding the Iberian cohort of Spain, IBS), Genome Aggregation Database, and a Spanish cohort including IBS individuals as well as data from Medical Genome Project, we identified 1,069 SNVs common in Navarre but rare (MAF ≤ 1%) in all other populations. This observation was further corroborated by 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.

71 ## INTRODUCTION

72 In the last few decades, the use of next-generation sequencing (NGS) in patient healthcare has
73 increased due to technological advances, cost reduction, and enhanced efficiency (1). The
74 advancement of NGS spans a spectrum of applications, encompassing whole exome sequencing
75 (WES), and whole genome sequencing (WGS). These technologies led to the identification of an
76 extensive number of variants, necessitating the implementation of filters to narrow down the list
77 of candidate variants. In this regard, the availability of population-specific catalogues of common
78 variants enables the identification of rare variants (2). Examples of variants catalogues include
79 international initiatives such as the 1000 Genome Project (1KGP) (3) and the Genome Aggregation
80 Database (gnomAD) (4). Countries like the United Kingdom (5), the United States of America (6),
81 and Japan (7) have also developed national databases. Another national initiative is the Medical
82 Genome Project (MGP), which compiles data from unrelated healthy individuals of Spanish origin
83 (8)(9).

84 In Navarre, a 600,000 population region of north-eastern Spain, the local Government supported
85 the "*NAGEN scheme*" to integrate whole genomic data analysis into the regional public healthcare
86 system. NAGEN has, up to date, generated more than 3,000 WGS/WES, and associated
87 phenoclinical profiles in seven consecutive projects on applied genomics, including *NAGEN1000*,
88 focused on rare diseases, and *pharmaNAGEN* on pharmacogenomics in patients with ulcerative
89 colitis and Crohn's disease (10). The NAGEN strategy's success hinges on identifying population-
90 specific common variants, thereby facilitating the establishment of a comprehensive Navarrese
91 population frequency.

92 In the present study (Fig. 1), using two cohorts, WGS and WES, respectively, we aimed to identify
93 and characterize common exonic variants specific to the Navarrese population. The first step was
94 identifying common single nucleotide variants (SNVs) in Navarre that are rare in other populations
95 using WGS data from 358 unrelated Spanish participants in the *NAGEN1000* project. Afterwards,
96 we validated the allele frequency of those variants using 239 unrelated WES samples from the
97 *pharmaNAGEN* project. Finally, we annotated the validated variants using genomic databases,
98 evaluated their clinical and pharmacological effects, and assessed their pathogenicity.
99 Additionally, we conducted functional enrichment analyses to provide further insights. The
100 results will significantly contribute to advancing personalized medicine in Navarre.

101 ## MATERIAL AND METHODS

102 Detailed information on material and methods can be found in the Supplementary Information.

103 **Individual and variant selection**

104 Unrelated individuals from Spain were identified from the Navarrese WGS project, called
105 *NAGEN1000*, resulting in a cohort of 358 participants designated as NAVARREsel. A bioinformatic
106 pipeline was performed, following GATK best practices, to extract biallelic variants in exonic
107 regions (11). WES data from 239 individuals from the Navarrese *pharmaNAGEN* project were
108 selected using the same inclusion criteria. This cohort, labelled as NAVARREval, was used for
109 validation.

110 **Referenced population projects**

111 Population frequencies for three reference cohorts were generated as follows: 1) gnomAD,
112 representing the original frequency from the gnomAD genome project v2.1.1; 2) *1KGP_noIBS*,
113 calculated as the mean of all 1KGP population's frequencies (phase3), with the IBS cohort
114 excluded; and 3) *spain*, comprising frequencies from the MGP project, and the IBS cohort from
115 the 1KGP project.

116 **Variant annotation**

117 Variants were annotated using ANNOVAR (12). The dbSNP database was utilized to identify known
118 variants (13). Clinical and pharmacological relevance was assessed using ClinVar (14), Online
119 Mendelian Inheritance in Man (OMIM) (15), and PharmaGKB (16) databases. In silico tools,
120 including CADD (17), REVEL (18), and Polyphen2 (19), were used for pathogenicity prediction, and
121 spliceAI (20) was applied to identify splicing effects. Functional enrichment analysis was
122 conducted using WebGestalt (21).

123 # RESULTS

124 ## Navarrese discovery cohort

125 In the *NAGEN1000* WGS Navarrese project, 688 individuals from 294 families, mainly trios (father,
126 mother, and child), were recruited. Individuals from each family were diagnosed with a rare
127 genetic disorder. The WGS was conducted with a mean coverage of 30X, providing comprehensive
128 genomic data across the entire genome. For our study, we kept a subset of this cohort satisfying
129 two criteria: unrelatedness and Spanish ancestry. This result yielded 358 individuals, referred to
130 as NAVARREsel.

131 Then, biallelic exonic SNVs, located on chromosomes 1 to 22, were extracted. Subsequently, we
132 filtered out variants with read depth (DP) < 10, genotype quality (GQ) < 50, or missing genotype
133 in at least one sample. Additionally, sites significantly deviated from Hardy-Weinberg equilibrium
134 (HWE, p-value < 10^{-5}) were removed (22). Finally, 61,410 SNVs remained; of these, 21,174 were
135 identified as common variants: minor allele frequency (MAF) > 1% (further details in the Material
136 and Methods section). When comparing the number of individuals and the identification of SNVs,
137 we observed that including additional individuals is not expected to reveal new common variants
138 (Fig. S1a), and 21,174 was achieved when considering more than 100 individuals.

139

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

141 We performed a principal component analysis (PCA) on the shared variants between NAVARREsel
142 with 1KGP (26 populations), and MGP to depict its relationship. We observed a clear distinction
143 between Navarre and Asian/African populations, reflecting established genetic differences (Fig.
144 2a). Conversely, an overlap is observed between Navarre and other European populations,

145 emphasizing their genetic affinity. Thus, we performed the PCA with European populations (Fig.
146 2b), and we observed that the Navarrese individuals are close to the Spanish populations (IBS and
147 MGP) and exhibit proximity to the Italian individuals.
148 To further analyse the genetic differentiation between Navarre and other populations (MGP, and
149 1KGP), we calculated the mean pairwise F_{ST} values. Given that the lower the F_{ST} value, the greater
150 the similarity between populations, as expected, the lowest degree of differentiation occurred
151 when comparing the Navarre population with the Spanish ($F_{ST(Navarre\text{-}IBS)}$ = 0.0001 and $F_{ST(Navarre\text{-}MGP)}$
152 = 0.0007), and Italian populations ($F_{ST(Navarre\text{-}TSI)}$ = 0.0014). On the other hand, we observed the
153 highest degree of differentiation with the Asian and African populations ($F_{ST(Navarre\text{-}SAS)}$ = 0.0149,
154 $F_{ST(Navarre\text{-}EAS)}$ = 0.0328, $F_{ST(Navarre\text{-}AFR)}$ = 0.0434).
155 These findings, aligning with biological expectations, underscore the regional and continental
156 genetic affinities, providing insights into historical populations and evolutionary dynamics.
157

158 ## Exclusive common variants in Navarre

159 To detect exclusive common variants in the Navarre population, we examined allele frequency
160 data from three cohorts: *1KGP_noIBS* (1KGP excluding 107 individuals of the IBS cohort), gnomAD,
161 and *spain* (integrated IBS and MGP cohorts). Comparison of the MAF among these cohorts
162 revealed that the majority of variants (17,532 SNVs) were classified as common (MAF > 1%) in all
163 populations. However, 835 variants exhibited higher prevalence, with MAF > 1%, solely in national
164 cohorts of Spain (Navarre and *spain*). Specifically in Navarre, 1,069 SNVs were identified as
165 common (MAF > 1%) in the Navarre population and rare (MAF ≤ 1%) in the rest (Fig. 2c).

166 In order to validate the 1,069 variants common in Navarre, we used the NAVARREval cohort,
167 consisting of 239 unrelated individuals from the current population of Navarre with Spanish
168 origin. These individuals were selected from the WES project *pharmaNAGEN* and had diagnoses
169 of Crohn's (159/239) or ulcerative colitis (86/239) diseases. Before validation, a comprehensive
170 examination of the association of the selected SNVs with these conditions was necessary to
171 eliminate potential bias. This examination involved analysing the overlap between the 1,069 and
172 reported variants in the Inflammatory Bowel Disease database, which catalogues genes and
173 variants highly linked to the mentioned diseases (23). The absence of the 1,069 SNVs in this
174 database ensured a robust and reliable validation process.
175 Among the 1,069 variants initially identified, 998 were successfully detected in NAVARREval with
176 a call rate > 80%, and demonstrated conformity to HWE, bolstering the confidence in the precision
177 of these variants within the validation dataset. Notably, 676/998 of these SNVs (68%; p-value =
178 $2.2e^{-16}$) were consistently classified as common in both NAVARREsel and NAVARREval, confirming
179 their prevalence within the Navarrese population (variants information in Table S1). We observed
180 that the validation cohort was sufficient to validate the majority of identified common variants
181 within the Navarrese population (Fig. S1b). On the contrary, within the non-validated subset
182 (322/998, 32%), 134 SNVs exhibited MAFs in NAVARREsel that did not exceed a 2-fold difference
183 in the validation cohort, indicating close MAF between both datasets. Furthermore, out of 322,
184 220 variants had a MAF greater than 0.005, approaching a level of commonality in NAVARREval
185 as well (Fig. S2). This exploration of MAF patterns ensures a comprehensive understanding of the
186 genetic landscape within the Navarre population and its stability across different datasets.

187 ## Characterization of common Navarrese variants

188 The annotation analysis for the 676 SNVs common to both Navarrese cohorts revealed that 227
189 were synonymous, 371 were missense, 73 were non-exonic (including two splicing), and five were
190 loss-of-function (LoF). These LoF variants were identified in five distinct genes, none annotated in
191 the OMIM database (15) with a reported associated phenotype. Additionally, none of these
192 variants were documented in the ClinVar database. Following the American College of Medical
193 Genetics (ACMG) guidelines for variant classification, four of these variants were classified as
194 variants of uncertain significance (VUS), and one was deemed benign (24).

195 Among the common SNVs, 264/676 were reported in ClinVar. Specifically, 1/264 were identified
196 as a risk factor, 181/264 as benign or likely benign, 32/264 as uncertain significance, 48/264 as
197 having conflicting interpretations, and 2/264 as likely pathogenic. These missense likely
198 pathogenic variants were c.1688G>A p.Arg563Gln in gene *SCNN1B* (Variant ID: 561264,
199 $MAF_{NAVARREsel}$=0.01396, $MAF_{NAVARREval}$=0.01673), and c.824G>A p.Arg275Gln in gene *PTGIS* (Variant
200 ID: 208402, $MAF_{NAVARREsel}$ = 0.01396, $MAF_{NAVARREval}$ = 0.02092). Then, c.1688G>A in *SCNN1B* was
201 identified in 10 and 16 heterozygous individuals from NAVARREsel and NAVARREval, respectively.
202 Pathogenic variants in gene *SCNN1B* cause Liddle syndrome 1 (OMIM #177200) or Bronchiectasis
203 with or without elevated sweat chloride 1 (OMIM #211400), both with autosomal dominant
204 inheritance patterns. Besides, one study also reported the same variant associated with arterial
205 hypertension (25). Given its notable prevalence in the Navarrese population, observed in healthy
206 and affected (not related to this phenotype) individuals, this variant might be reconsidered and
207 classified as VUS under the ACMG guidelines.

208 A similar situation occurred with the *PTGIS* variant, associated in ClinVar with childhood-onset
209 Schizophrenia (not reported in OMIM). The evidence supporting this association is limited, with
210 a low score of 1 out of 4, reviewed by a single submitter record. In NAVARREsel, ten heterozygotes
211 and in NAVARREval, eight heterozygotes, and one homozygote were identified, mainly healthy
212 individuals or diagnosed with another disease. Therefore, the PTGIS variant should be reassessed
213 and reclassified potentially as VUS according to ACMG criteria.
214 Analysis of the 676 variants with *in silico* functional predictors revealed eight variants as
215 pathogenic by three different pathogenicity tools (REVEL_score > 0.8 (18), CADD > 20 (17), and
216 Polyphen indicating "probably" or "possibly" (19)). However, a comprehensive examination of
217 clinical databases, including ClinVar, Varsome (26), and Franklin (27), contradicted these
218 predictions based on ACMG criteria. Instead, the majority of these variants were classified as
219 uncertain significance (1/8), likely benign (5/8), and benign (2/8). This discrepancy rejected the
220 notion that these variants were directly linked to causing disease.
221 Additionally, the variant c.387-1G>T in *BPIFB3* gene ($MAF_{NAVARREsel}$=0.01536,
222 $MAF_{NAVARREval}$=0.02092), predicted to impact the canonical splicing acceptor site (spliceAI score =
223 0.99), was reclassified as benign based on the allele frequency and the number of homozygotes
224 as per ACMG criteria, indicating no clinical relevance (20).
225 Moreover, none of the common Navarrese variants demonstrated an impact on drug
226 metabolism/efficacy, according to PharmaGKB (16). They did not exhibit significant enrichment in
227 pathways, biological processes, related diseases, or phenotypic ontologies (see Material and
228 Methods).
229

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

231 In this study, we characterized exclusively common variants (MAF > 1%) in the Navarrese
232 population. This research underscores the significance of exploring population-specific genetic
233 variations in advancing personalized medicine. The primary goal was to improve the identification
234 of potential disease-causing variants during genetic diagnosis when using NGS to reach a genetic
235 diagnosis. We selected WGS data from 127 patients diagnosed with hereditary disorders in
236 NAVARREsel. We analyzed biallelic SNVs from the exome region of chromosomes 1 to 22, resulting
237 in an average of 8,871 variants per patient.

238 To refine this list, we systematically excluded common variants (MAF > 1%) from *1KGP_noIBS*,
239 gnomAD, *spain*, and Navarre. As depicted in Fig. 3, which illustrates the percentage of retained
240 variants after filtering with different population allele frequencies, the Navarrese filtering
241 emerged as the most stringent, resulting in 2.1% of the initial set of variants. In comparison,
242 filtering with gnomAD data retained 2.7%, *spain* frequencies preserved 2.9%, and the least
243 restrictive, 1KGP, maintained 4.9% of the variants. This significant decrease in variants achieved
244 through Navarrese-specific filters exemplifies their effectiveness in prioritizing and streamlining
245 genetic investigations. This highlights their crucial role in advancing personalized medicine
246 tailored explicitly to the Navarre population.

247 ## DISCUSSION

248 In this study, we aimed to enhance diagnostic precision in the Navarrese population by exploring
249 common population-specific variants in Navarre. Utilizing WGS data from 358 individuals of
250 Navarre, we identified 61,410 SNVs with 21,174 common variants (MAF > 1%). Genetic analysis
251 shows affinity with European populations and low differentiation with Spanish and Italian
252 populations.

253 We focused on exclusively common variants in Navarre compared with reference populations,
254 which led to the identification of 1,069 SNVs. We validated 676 variants in an independent cohort.
255 Of these, none showed clinical or pharmacological relevance beyond what was observed in the
256 Spanish population (28). This aligns with the expectation that variants common in a population
257 are less likely to be associated with disease ethology.

258 Our finding underscores the significance of considering population-specific factors in genomic
259 diagnostics which provides complementary insights to that derived from pangenome references
260 (29). By identifying and excluding common variants within the Navarrese population, we have
261 successfully narrowed down the list of potential disease-causing variants. This approach
262 contributed to the advancement of precision medicine by refining the identification of clinically
263 relevant variants specific to the Navarrese population. In conclusion, this study highlights
264 population-specific variants in Navarre, which are crucial for advancing personalized medicine.
265 Further research will enhance these insights for broader applications.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology (Basel). 2023 Jul 13;12(7):997.
2. Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, et al. Iranome: A catalog of genomic variations in the Iranian population. Hum Mutat. 2019 Nov 17;40(11):1968–84.
3. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.
4. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat. 2022 Aug 16;43(8):1012–30.
5. Smetana J, Brož P. National Genome Initiatives in Europe and the United Kingdom in the Era of Whole-Genome Sequencing: A Comprehensive Review. Genes (Basel). 2022 Mar 21;13(3):556.
6. Ramirez AH, Sulieman L, Schlueter DJ, Halvorson A, Qian J, Ratsimbazafy F, et al. The All of Us Research Program: Data quality, utility, and diversity. Patterns. 2022 Aug;3(8):100570.
7. Mitsuhashi N, Toyo-oka L, Katayama T, Kawashima M, Kawashima S, Miyazaki K, et al. TogoVar: A comprehensive Japanese genetic variation database. Hum Genome Var. 2022 Dec 12;9(1):44.

289 8. Dopazo J, Amadoz A, Bleda M, Garcia-Alonso L, Alemán A, García-García F, et al. 267
290 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic
291 Variation. Mol Biol Evol. 2016;33(5):1205–18.
292 9. Peña-Chilet M, Roldán G, Perez-Florido J, Ortuño FM, Carmona R, Aquino V, et al. CSVS, a
293 crowdsourcing database of the Spanish population genetic variability. Nucleic Acids Res.
294 2021 Jan 8;49(D1):D1130–7.
295 10. NAGEN | Navarrabiomed [Internet]. [cited 2023 Nov 28]. Available from:
296 https://www.navarrabiomed.es/en/nagen
297 11. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
298 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing
299 data. Genome Res. 2010 Sep;20(9):1297–303.
300 12. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from
301 high-throughput sequencing data. Nucleic Acids Res. 2010 Sep 1;38(16):e164–e164.
302 13. Sherry ST. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan
303 1;29(1):308–11.
304 14. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public
305 archive of relationships among sequence variation and human phenotype. Nucleic Acids
306 Res. 2014 Jan;42(Database issue):D980-5.
307 15. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in
308 Man (OMIM(R)). Nucleic Acids Res. 2009 Jan 1;37(Database):D793–6.
309 16. Thorn CF, Klein TE, Altman RB. PharmGKB: The Pharmacogenomics Knowledge Base. In
310 2013. p. 311–20.

311 17. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the
312 deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan
313 8;47(D1):D886–94.
314 18. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An
315 Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. The
316 American Journal of Human Genetics. 2016 Oct;99(4):877–85.
317 19. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense
318 Mutations Using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan 2;76(1).
319 20. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI,
320 et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019
321 Jan;176(3):535-548.e24.
322 21. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive,
323 powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res.
324 2017 Jul 3;45(W1):W130–7.
325 22. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality
326 control in genetic case-control association studies. Nat Protoc. 2010 Sep 26;5(9):1564–73.
327 23. Khan F, Radovanovic A, Gojobori T, Kaur M. IBDDB: a manually curated and text-mining-
328 enhanced database of genes involved in inflammatory bowel disease. Database. 2021 Apr
329 30;2021.
330 24. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines
331 for the interpretation of sequence variants: a joint consensus recommendation of the

332 American College of Medical Genetics and Genomics and the Association for Molecular
333 Pathology. Genetics in Medicine. 2015 May;17(5):405–24.
334 25. Rayner BL, Owen EP, King JA, Soule SG, Vreede H, Opie LH, et al. A new mutation, R563Q,
335 of the beta subunit of the epithelial sodium channel associated with low-renin, low-
336 aldosterone hypertension. J Hypertens. 2003 May;21(5):921–6.
337 26. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome:
338 the human genomic variant search engine. Bioinformatics. 2019 Jun 1;35(11):1978–80.
339 27. Franklin [Internet]. [cited 2023 Nov 28]. Available from:
340 https://franklin.genoox.com/clinical-db/home
341 28. Nunez-Torres R, Pita G, Peña-Chilet M, López-López D, Zamora J, Roldán G, et al. A
342 Comprehensive Analysis of 21 Actionable Pharmacogenes in the Spanish Population:
343 From Genetic Characterisation to Clinical Impact. Pharmaceutics. 2023 Apr
344 19;15(4):1286.
345 29. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human
346 pangenome reference. Nature. 2023 May 11;617(7960):312–24.
347
348

FUNDING

NAGEN1,000 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

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, JD, IL, JB, AA, DGC).

Corresponding authors

Correspondence to Ángel Alonso & David Gomez-Cabrero.

407 ## ETHICS DECLARATIONS

408 Competing interests:

409 The authors declare no competing interests.

410 Ethics approval:

411 NAGEN1,000 and PharmaNAGEN were approved by Navarra Ethics Committee for Clinical

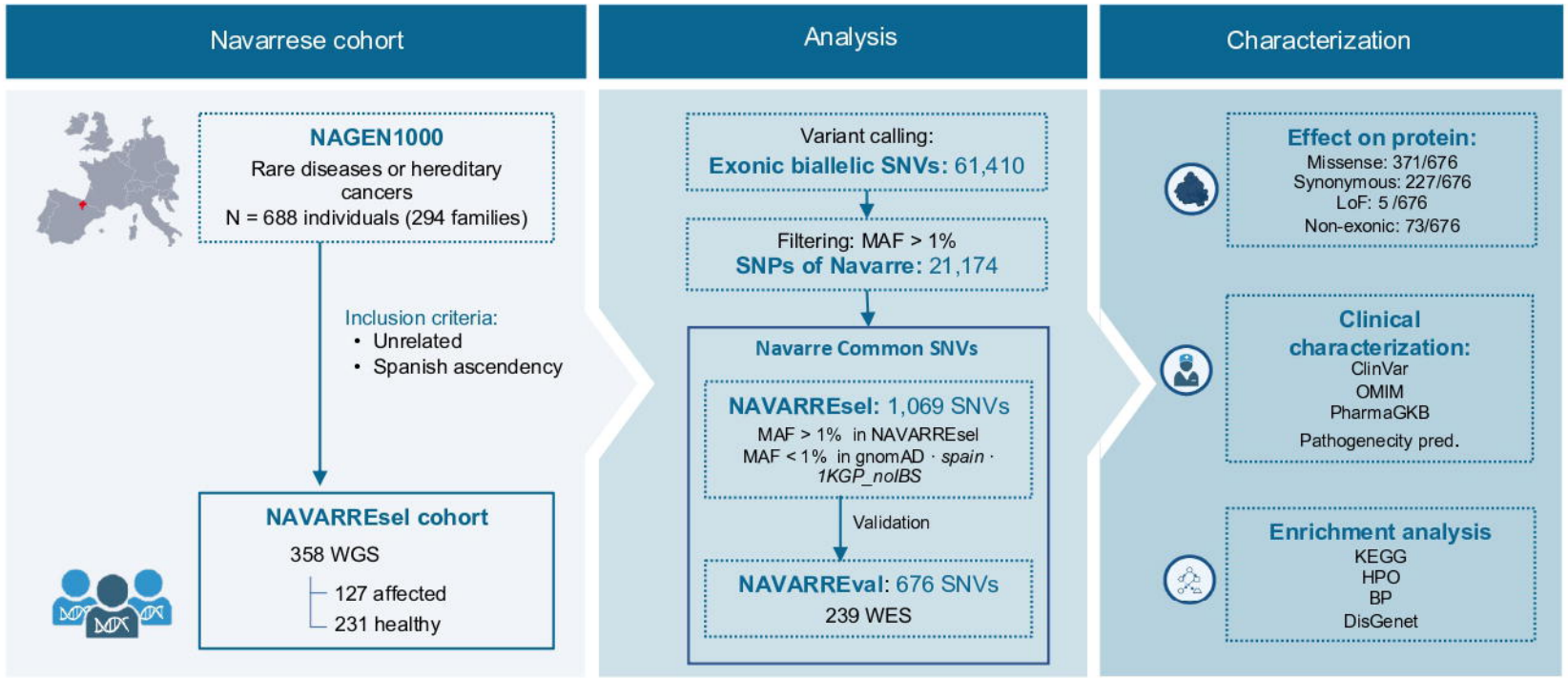
412 Research (CEIC Navarra).

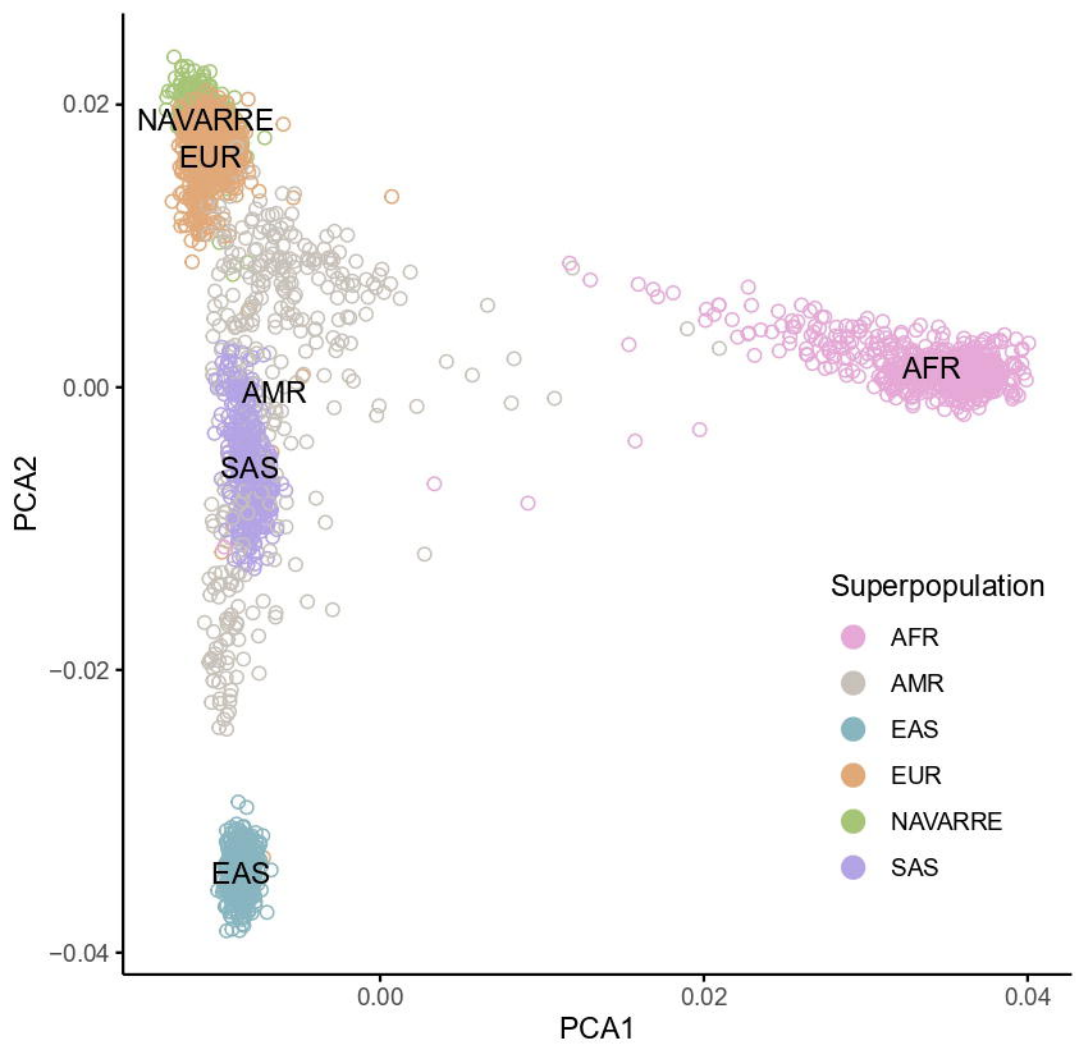
413 ## FIGURES LEGENDS

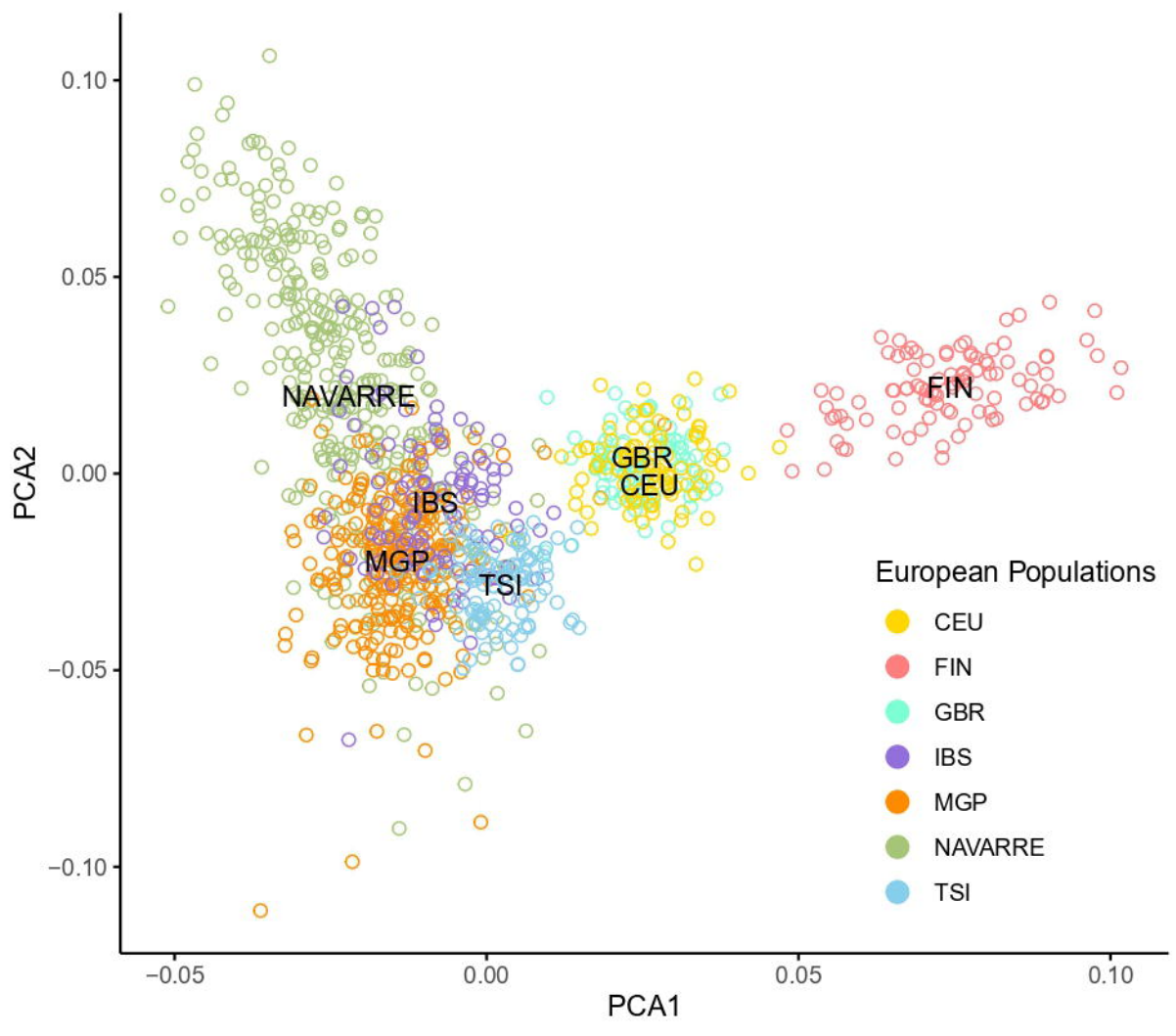
414 **Figure 1**: Workflow of this study. Abbreviations: *MGP*, Medical Genome Project; *1KGP*, 1000
415 Genomes Project; *1KGP_noIBS*, 1000 Genomes Project without Iberian population; *gnomAD,*
416 Genome Aggregation Database; *MAF*, minor allele frequency; *SNV*, single nucleotide variant; *SNP*,
417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

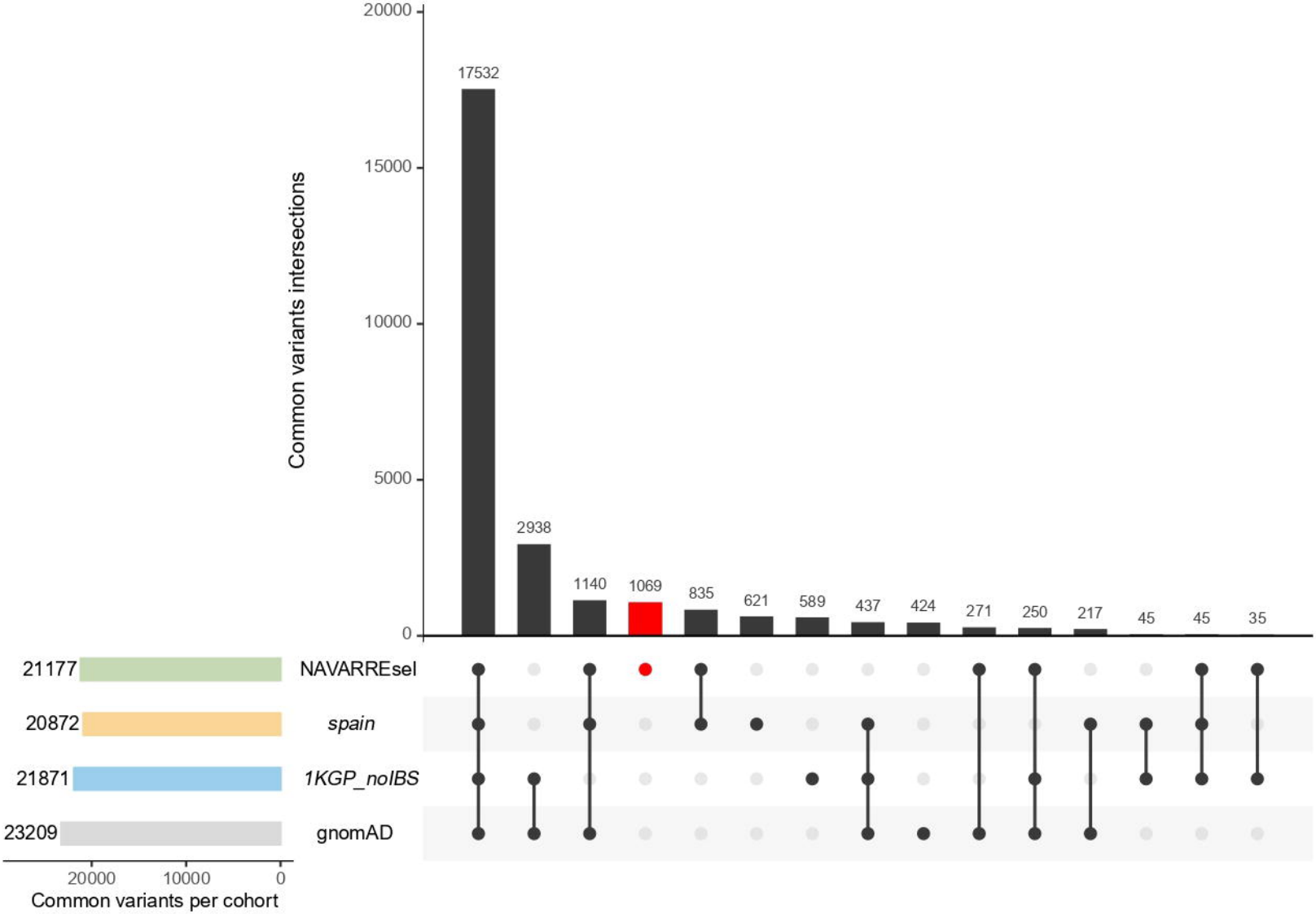
419 **Figure 2: a)** Principal component analysis of overlapped variants between NAVARREsel, MGP, and
420 1KGP (including all populations), and coloured by superpopulations. **b)** Principal component
421 analysis of overlapped variants between NAVARREsel, MGP, and 1KGP (including exclusively
422 European populations). **c)** Upset plot of common variants (MAF > 1%) of each population:
423 NAVARREsel, *spain*, *1KGP_noIBS*, and gnomAD. Abbreviations: *1KGP_noIBS*, 1000 Genomes
424 Project without Iberian population; gnomAD*,* Genome Aggregation Database; *spain*, integration
425 of IBS and MGP populations; *PCA*, principal component analysis; AFR, African populations; AMR,
426 American populations; EAS, east-Asian populations; SAS, south-Asian populations; EUR, European
427 populations; IBS, Iberian populations in Spain; MGP, Medical Genome Project; TSI, Toscani in Italy;
428 CEU, Utah residents with Northern and Western European ancestry; GBR, British in England and
429 Scotland; FIN, Finnish in Finland.

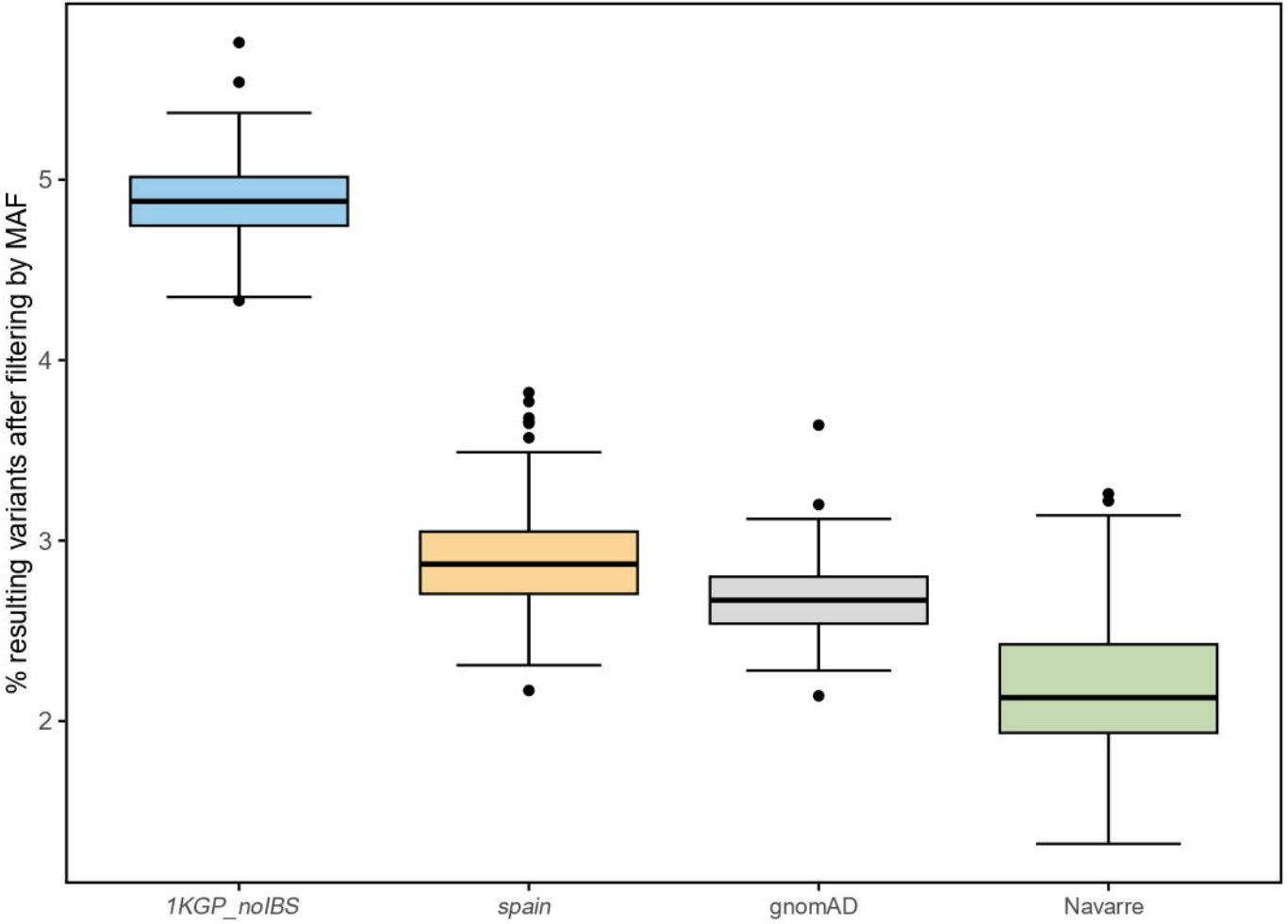
430 **Figure 3:** Resulting percentage of variants per patient (n=127) after removing common variants
431 from Navarre, *spain*, gnomAD, or *1KGP_noIBS* populations. In the box plots, the
432 median, upper, and lower quartiles are represented by the centre line and box
433 bounds, respectively. Whiskers display the largest and smallest values within 1.5
434 times the interquartile range from the quartiles.


Navarrese cohort
Analysis
Characterization
NAGEN1000
Rare diseases or hereditary cancers
N = 688 individuals (294 families)
Inclusion criteria:
• Unrelated
• 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
Non-exonic: 73/676
Clinical characterization:
ClinVar
OMIM
PharmaGKB
Pathogenecity pred.
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


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