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

Authors:
Alberto Maillo, alberto.ruizdeinfante@kaust.edu.sa, Spain¹,²
Estefania Huergo, ehuergoi@navarra.es, Spain¹,#
María Apellániz-Ruiz, mv.apellaniz.ruiz@navarra.es, Spain³,#
Edurne Urrutia, eurrutia@navarra.es, Spain¹,³,#
María Miranda, mmirandp@navarra.es, Spain³
Josefa Salgado, jsalgadg@navarra.es, Spain³,⁴,⁵
Sara Pasalodos-Sánchez, spasalos@navarra.es, Spain³
Luna Delgado-Mora, lunadelde@gmail.com, Spain³,⁶
Óscar Teijido, oteijidh@navarra.es, Spain³
Rosario Carmona, rosariom.carmona@juntadeandalucia.es, Spain⁷,⁸,⁹,¹⁰
Javier Perez-Florido, javier.perez.florido.sspa@juntadeandalucia.es, Spain⁷,⁸,⁹,¹⁰
Virginia Aquino, virginia.aquino@juntadeandalucia.es, Spain⁷
Daniel Lopez-Lopez, daniel.lopez.lopez@juntadeandalucia.es, Spain⁷,⁸,¹⁰
María Peña-Chilet, maria_pena@iislafe.es, Spain⁷,¹⁰
Sergi Beltrán, sergi.beltran@cnag.eu, Spain¹¹,¹²,¹³
Joaquín Dopazo, joaquin.dopazo@juntadeandalucia.es, Spain⁷,⁸,⁹,¹⁰
Iñigo Lasa, ilasa@unavarra.es, Spain¹⁴
Juan José Beloqui, jj.beloqui.lizaso@navarra.es, Spain³
Ángel Alonso, aalonson@navarra.es, Spain³,#
David Gomez-Cabrero, david.gomez.cabrero@navarra.es, Spain¹,²,#

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: Dpto. 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: Computing Medicine Platform, Andalusian Public Foundation Progress and Health-FPS, Sevilla, Spain.
⁸: Institute of Biomedicine of Seville, IBiS, University Hospital Virgen del Rocío/CSIC/University of Sevilla, Sevilla, Spain.
⁹: FPS/ELIXIR-ES, Fundación Progreso y Salud (FPS), CDCA, Hospital Virgen del Rocío, Sevilla, Spain.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
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.
INTRODUCTION

In the last few decades, the use of next-generation sequencing (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 sequencing (WES), and whole genome sequencing (WGS). These technologies led to the identification of an extensive number of 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). Examples of variants catalogues include international initiatives such as the 1000 Genome Project (1KGP) (3) and the Genome Aggregation Database (gnomAD) (4). Countries like the United Kingdom (5), the United States of America (6), and Japan (7) have also developed national databases. Another national initiative is the Medical Genome Project (MGP), which compiles data from unrelated healthy individuals of Spanish origin (8)(9).

In Navarre, a 600,000 population region of north-eastern Spain, the local Government supported the “NAGEN scheme” to integrate whole genomic data analysis into the regional public healthcare system. NAGEN has, up to date, generated more than 3,000 WGS/WES, and associated phenoclinical profiles in seven consecutive projects on applied genomics, including NAGEN1000, focused on rare diseases, and pharmaNAGEN on pharmacogenomics in patients with ulcerative colitis and Crohn’s disease (10). The NAGEN strategy’s success hinges on identifying population-specific common variants, thereby facilitating the establishment of a comprehensive Navarrese population frequency.
In the present study (Fig. 1), using two cohorts, WGS and WES, respectively, we aimed to identify and characterize common exonic variants specific to the Navarrese population. The first step was identifying common single nucleotide variants (SNVs) in Navarre that are rare in other populations using WGS data from 358 unrelated Spanish participants in the NAGEN1000 project. Afterwards, we validated the allele frequency of those variants using 239 unrelated WES samples from the pharmaNAGEN project. Finally, we annotated the validated variants using genomic databases, evaluated their clinical and pharmacological effects, and assessed their pathogenicity. Additionally, we conducted functional enrichment analyses to provide further insights. The results will significantly contribute to advancing personalized medicine in Navarre.
MATERIAL AND METHODS

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

Individual and variant selection

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

Referenced population projects

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

Variant annotation

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

Navarrese discovery cohort

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

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

Genetic variation between Navarre, Spanish, and global populations

We performed a principal component analysis (PCA) on the shared variants between NAVARResel with 1KGP (26 populations), and MGP to depict its relationship. We observed a clear distinction between Navarre and Asian/African populations, reflecting established genetic differences (Fig. 2a). Conversely, an overlap is observed between Navarre and other European populations,
emphasizing their genetic affinity. Thus, we performed the PCA with European populations (Fig. 2b), and we observed that the Navarrese individuals are close to the Spanish populations (IBS and MGP) and exhibit proximity to the Italian individuals.

To further analyse the genetic differentiation between Navarre and other populations (MGP, and 1KGP), we calculated the mean pairwise $F_{ST}$ values. Given that the lower the $F_{ST}$ value, the greater the similarity between populations, as expected, the lowest degree of differentiation occurred when comparing the Navarre population with the Spanish ($F_{ST}(\text{Navarre-IBS}) = 0.0001$ and $F_{ST}(\text{Navarre-MGP}) = 0.0007$), and Italian populations ($F_{ST}(\text{Navarre-TSI}) = 0.0014$). On the other hand, we observed the highest degree of differentiation with the Asian and African populations ($F_{ST}(\text{Navarre-SAS}) = 0.0149$, $F_{ST}(\text{Navarre-EAS}) = 0.0328$, $F_{ST}(\text{Navarre-AFR}) = 0.0434$).

These findings, aligning with biological expectations, underscore the regional and continental genetic affinities, providing insights into historical populations and evolutionary dynamics.

**Exclusive common variants in Navarre**

To detect exclusive common variants in the Navarre population, we examined allele frequency data from three cohorts: 1KGP _noIBS_ (1KGP excluding 107 individuals of the IBS cohort), gnomAD, and _spain_ (integrated IBS and MGP cohorts). Comparison of the MAF among these cohorts revealed that the majority of variants (17,532 SNVs) were classified as common (MAF > 1%) in all populations. However, 835 variants exhibited higher prevalence, with MAF > 1%, solely in national cohorts of Spain (Navarre and _spain_). Specifically in Navarre, 1,069 SNVs were identified as common (MAF > 1%) in the Navarre population and rare (MAF ≤ 1%) in the rest (Fig. 2c).
In order to validate the 1,069 variants common in Navarre, we used the NAVARREval cohort, consisting of 239 unrelated individuals from the current population of Navarre with Spanish origin. These individuals were selected from the WES project pharmNAGEN and had diagnoses of Crohn's (159/239) or ulcerative colitis (86/239) diseases. Before validation, a comprehensive examination of the association of the selected SNVs with these conditions was necessary to eliminate potential bias. This examination involved analysing the overlap between the 1,069 and reported variants in the Inflammatory Bowel Disease database, which catalogues genes and variants highly linked to the mentioned diseases (23). The absence of the 1,069 SNVs in this database ensured a robust and reliable validation process.

Among the 1,069 variants initially identified, 998 were successfully detected in NAVARReval with a call rate > 80%, and demonstrated conformity to HWE, bolstering the confidence in the precision of these variants within the validation dataset. Notably, 676/998 of these SNVs (68%; p-value = 2.2e-16) were consistently classified as common in both NAVARReSel and NAVARREval, confirming their prevalence within the Navarrese population (variants information in Table S1). We observed that the validation cohort was sufficient to validate the majority of identified common variants within the Navarrese population (Fig. S1b). On the contrary, within the non-validated subset (322/998, 32%), 134 SNVs exhibited MAFs in NAVARReSel that did not exceed a 2-fold difference in the validation cohort, indicating close MAF between both datasets. Furthermore, out of 322, 220 variants had a MAF greater than 0.005, approaching a level of commonality in NAVARREval as well (Fig. S2). This exploration of MAF patterns ensures a comprehensive understanding of the genetic landscape within the Navarre population and its stability across different datasets.
Characterization of common Navarrese variants

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

Among the common SNVs, 264/676 were reported in ClinVar. Specifically, 1/264 were identified as a risk factor, 181/264 as benign or likely benign, 32/264 as uncertain significance, 48/264 as having conflicting interpretations, and 2/264 as likely pathogenic. These missense likely pathogenic variants were c.1688G>A p.Arg563Gln in gene SCNN1B (Variant ID: 561264, MAF_{NAVAREsel}=0.01396, MAF_{NAVAREval}=0.01673), and c.824G>A p.Arg275Gln in gene PTGIS (Variant ID: 208402, MAF_{NAVAREsel} = 0.01396, MAF_{NAVAREval} = 0.02092). Then, c.1688G>A in SCNN1B was identified in 10 and 16 heterozygous individuals from NAVAREsel and NAVAREval, respectively.

Pathogenic variants in gene SCNN1B cause Liddle syndrome 1 (OMIM #177200) or Bronchiectasis with or without elevated sweat chloride 1 (OMIM #211400), both with autosomal dominant inheritance patterns. Besides, one study also reported the same variant associated with arterial hypertension (25). Given its notable prevalence in the Navarrese population, observed in healthy and affected (not related to this phenotype) individuals, this variant might be reconsidered and classified as VUS under the ACMG guidelines.
A similar situation occurred with the *PTGIS* variant, associated in ClinVar with childhood-onset Schizophrenia (not reported in OMIM). The evidence supporting this association is limited, with a low score of 1 out of 4, reviewed by a single submitter record. In NAVARReSel, ten heterozygotes and in NAVARReVal, eight heterozygotes, and one homozygote were identified, mainly healthy individuals or diagnosed with another disease. Therefore, the PTGIS variant should be reassessed and reclassified potentially as VUS according to ACMG criteria.

Analysis of the 676 variants with *in silico* functional predictors revealed eight variants as pathogenic by three different pathogenicity tools (REVEL score > 0.8 (18), CADD > 20 (17), and Polyphen indicating “probably” or “possibly” (19)). However, a comprehensive examination of clinical databases, including ClinVar, Varsome (26), and Franklin (27), contradicted these predictions based on ACMG criteria. Instead, the majority of these variants were classified as uncertain significance (1/8), likely benign (5/8), and benign (2/8). This discrepancy rejected the notion that these variants were directly linked to causing disease.

Additionally, the variant c.387-1G>T in *BPIFB3* gene (MAFNAVARReSel=0.01536, MAFNAVARReVal=0.02092), predicted to impact the canonical splicing acceptor site (spliceAI score = 0.99), was reclassified as benign based on the allele frequency and the number of homozygotes as per ACMG criteria, indicating no clinical relevance (20).

Moreover, none of the common Navarrese variants demonstrated an impact on drug metabolism/efficacy, according to PharmaGKB (16). They did not exhibit significant enrichment in pathways, biological processes, related diseases, or phenotypic ontologies (see Material and Methods).
Refining disease-causing variant identification in the Navarrese population

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

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

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

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

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

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


27. Franklin [Internet]. [cited 2023 Nov 28]. Available from: https://franklin.genoox.com/clinical-db/home


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-Ruíz, 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.
ETHICS DECLARATIONS

Competing interests:

The authors declare no competing interests.

Ethics approval:

NAGEN1,000 and PharmaNAGEN were approved by Navarra Ethics Committee for Clinical Research (CEIC Navarra).
**FIGURES LEGENDS**

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

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

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

NAGEN1000
Rare diseases or hereditary cancers
N = 688 individuals (294 families)

Inclusion criteria:
- Unrelated
- Spanish ascendency

NAVARREsel cohort
358 WGS
- 127 affected
- 231 healthy

Analysis

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
1KG_no65

Validation

NAVARREval: 676 SNVs
239 WES

Characterization

Effect on protein:
Missense: 371/676
Synonymous: 227/676
Lof: 5/676
Non-exonic: 73/676

Clinical characterization:
ClinVar
OMIM
PharmGKB
Pathogenicity pred.

Enrichment analysis
KEGG
HPO
BP
DisGeNet