genes

MDPI

Brief Report

Characterization of the Common Genetic Variation in the Spanish Population of Navarre

Alberto Maillo [1,2], Estefania Huergo [1,†], María Apellániz-Ruiz [3,†], Edurne Urrutia-Lafuente [1,3,†], María Miranda [3], Josefa Salgado [3,4,5], Sara Pasalodos-Sanchez [3], Luna Delgado-Mora [3,6], Óscar Teijido [3], Ibai Goicoechea [7], Rosario Carmona [8,9,10,11], Javier Perez-Florido [8,9,10,11], Virginia Aquino [8], Daniel Lopez-Lopez [8,9,11], María Peña-Chilet [8,11], Sergi Beltran [12,13,14], Joaquín Dopazo [8,9,10,11], Iñigo Lasa [15], Juan José Beloqui [3], NAGEN-Scheme ‡, Ángel Alonso [3,*] and David Gomez-Cabrero [1,2,*]

1 Translational Bioinformatics Unit, Navarrabiomed, Hospital Universitario de Navarra (HUN), Universidad Pública de Navarra (UPNA), IdiSNA, 31008 Pamplona, Spain; alberto.ruizdeinfante@kaust.edu.sa (A.M.)
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, 31008 Pamplona, Spain
4 Servicio de Genética Médica, Hospital Universitario de Navarra (HUN), 31008 Pamplona, Spain
5 Dp. Bioquímica y Biología Molecular, Universidad Pública de Navarra (UPNA), 31006 Pamplona, Spain
6 Instituto de Genética Médica y Molecular (INGEMM), Hospital Universitario La Paz, 28046 Madrid, Spain
7 Department of Personalized Medicine, NASERTIC, Government of Navarra, 31011 Pamplona, Spain
8 Computational Medicine Platform, Andalusian Public Foundation Progress and Health-FPS, 41013 Sevilla, Spain
9 Institute of Biomedicine of Seville, IBiS, University Hospital Virgen del Rocio/CSIC/University of Sevilla, 41013 Sevilla, Spain
10 FPS/ELIXIR-ES, Fundación Progreso y Salud (FPS), CDCA, Hospital Virgen del Rocio, 41013 Sevilla, Spain
11 Biomedical Research Networking Center in Rare Diseases (CIBERER), Health Institute Carlos III, 28029 Madrid, Spain
12 Centro Nacional de Analisis Genomico (CNAG-CRG), Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain
13 Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain
14 Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), 08007 Barcelona, Spain
15 Laboratory of Microbial Pathogenesis, Navarrabiomed, 31008 Pamplona, Spain
* Correspondence: aalonsos@navarra.es (A.A.); david.gomez.cabrero@navarra.es (D.G.-C.)
† These authors contributed equally to this work.
‡ Membership of the NAGEN-Scheme is provided in Appendix A.

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 1069 SNVs common in Navarre but rare (MAF $\leq$ 1%) in all other populations. We further corroborated this observation with a second regional cohort of 239 unrelated exomes, which confirmed 676 of the 1069 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

check for updates

Citation: Maillo, A.; Huergo, E.; Apellániz-Ruiz, M.; Urrutia-Lafuente, E.; Miranda, M.; Salgado, J.; Pasalodos-Sanchez, S.; Delgado-Mora, L.; Teijido, Ó.; Goicoechea, I.; et al. Characterization of the Common Genetic Variation in the Spanish Population of Navarre. *Genes* **2024**, *15*, 585. https://doi.org/10.3390/genes15050585

Academic Editor: Hongyan Xu

Received: 1 April 2024
Revised: 23 April 2024
Accepted: 1 May 2024
Published: 4 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. 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, the region of north-eastern Spain populated by 650,000 people, the local Government supported the "NAGEN scheme" to integrate genomic data analysis into the regional public healthcare system. In recent times, NAGEN has generated numerous WES/WGS and associated phenoclinical profiles in seven projects, including *NAGEN1000*, focusing 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 (Figure 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 that are 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 variants using genomic databases, and their clinical and pharmacological effects and pathogenicity were assessed. Additionally, we conducted functional enrichment analyses to provide further insights. The results will significantly contribute to advancing personalized medicine in Navarre.

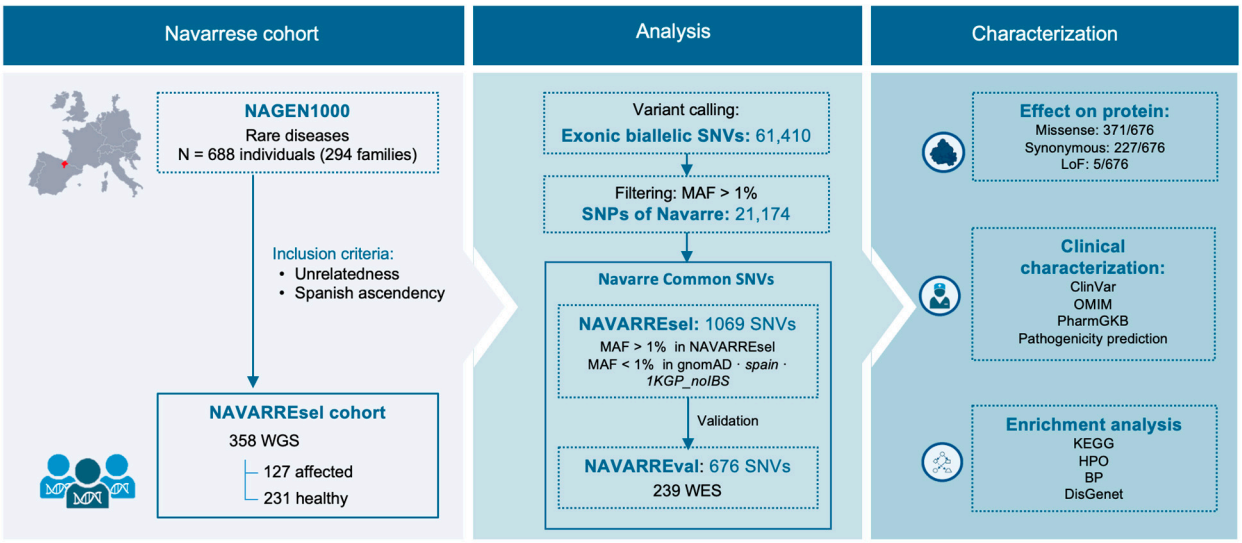


Figure 1. Workflow of this study. Abbreviations: *MGP*, Medical Genome Project; *1KGP*, 1000 Genomes Project; *1KGP_noIBS*, 1000 Genomes Project without Iberian population; *gnomAD*, Genome Aggregation Database; *MAF*, minor allele frequency; *SNV*, single nucleotide variant; *SNP*, single nucleotide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

2. Material and Methods

2.1. NAGEN1000 and pharmaNAGEN

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

2.2. Whole Genome Sequencing and Data Analysis

High-quality DNA samples from peripheral blood were used in the *NAGEN1000* project to construct short-insert paired-end libraries with an average insert size of 400 bp. DNA fragmentation was performed with Covaris S2, and capillary electrophoresis was performed with a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). Libraries were sequenced on a NovaSeq 6000 (Illumina, San Diego, CA, USA) with a read length of 2 × 150 bp. A 30X coverage per sample was targeted. After quality control assessment using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed on 10 October 2017), the sequenced data were aligned to the hs37d5 version of human genome reference GRCh37/hg19 using GEM3 [11]. Optical and duplicated reads were flagged with Picard MarkDuplicates (https://broadinstitute.github.io/picard/, accessed on 10 October 2017). Following the established Genome Analysis Toolkit (GATK) best practices pipeline (v3.8) [12], indel realignments and recalibration were applied to the previous BAM files. Variant calling on each sample's BAM file was performed using HaplotypeCaller, with default parameters, resulting in a gVCF file. WGS was conducted at Centro Nacional de Análisis Genómico (CNAG, Barcelona, Spain) and stored at Navarra de Servicios y Tecnología (NASERTIC, Navarre, Spain).

2.3. Whole Exome Sequencing and Data Analysis

In the *pharmaNAGEN* project, germline DNA was extracted from saliva or blood samples using a DNA Blood Maxi Kit (Qiagen, Hilden, Germany) and sequenced with a Nextera DNA Exome kit. The raw data were aligned to the GRCh37/hg19 genome, sourced from UCSC (https://genome.ucsc.edu/, accessed on 4 September 2019), utilizing BWA [13]. The resulting BAM files were marked using Picard (https://broadinstitute.github.io/picard/, accessed on 4 September 2019). Utilizing GATK v4.1.0 [12], an updated version, eliminates the need for the indel realignment step. Recalibration and variant calling were executed with BQSR and Haplotype tools. This process yielded the final gVCF file for each sample. WES was conducted at CNAG and stored at NASERTIC.

2.4. Individual Selection

Selection criteria included unrelated individuals with Spanish ancestry. To assess relatedness among individuals, identity by descent (IBD) was calculated using the method of moments (MoM) with the R package SNPRelate [14]. For validation of ethnicity, the CSVS tool [9] was used to determine the degree of alignment of a sample with the genetic variability of the Spanish population. Individuals with a score equal to or higher than 0.9 were categorized as being of Spanish ancestry.

Applying these selection criteria to the *NAGEN1000* project resulted in a cohort of 358 participants denominated NAVARREsel. Within this group, 127 individuals had been diagnosed with various monogenic diseases, and the most prevalent conditions were polycystic kidney disease (14/127), breast cancer (10/127), and hereditary ataxia (8/127).

Regarding the validation dataset, named NAVARREval, the same inclusion criteria were applied to the *pharmaNAGEN* project, resulting in 239 participants with Crohn's disease (153/239) and/or ulcerative colitis (86/239).

2.5. Variant Quality Control and Filtering in NAVARREsel

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

2.6. Variant Quality Control and Filtering in NAVARREval

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

2.7. Variant Annotation

Variants were annotated using ANNOVAR (version available on 24 October 2019, https://annovar.openbioinformatics.org/) [16]. The identification of known variants was performed using the dbSNP database (version GCF_000001405.25, downloaded from https://ftp.ncbi.nih.gov/snp/latest_release/VCF/ accessed on 30 September 2023) [17]. SNVs were also annotated: (1) for clinical significance by referring to ClinVar (v.20230930, https://www.ncbi.nlm.nih.gov/clinvar/, accessed on 2 October 2023) [18], OMIM (downloaded on 2 October 2023, https://www.omim.org/) [19], VarSome [20], and Franklin [21], and (2) for pharmacological relevance using PharmGKB (https://www.pharmgkb.org/, accessed on 2 October 2023) [22]. Additionally, the pathogenicity of the variants was evaluated using CADD (v1.6, https://cadd.gs.washington.edu/, accessed on 2 October 2023) [23], REVEL score (v1.3, https://zenodo.org/records/7072866, accessed on 2 October 2023) [24], spliceAI [25], and Polyphen2 (http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads downloaded on 4 October 2023) [26]. Variants associated with Crohn's disease and/or ulcerative colitis were identified by referencing the Inflammatory Bowel Disease (IBD) database (accessed on 10 October 2023, from https://www.cbrc.kaust.edu.sa/ibd/index.php?p=ibd#) [27].

2.8. Population Projects

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

The population data from 1KGP phase 3 encompassed 2504 genome samples representing 26 populations [28]. European populations within this dataset included 503 samples from Europe (EUR), such as British in England and Scotland (GBR), Finnish in Finland (FIN), Iberian population in Spain (IBS), Toscani in Italy (TSI), and Utah residents with Northern and Western European ancestry (CEU). Additionally, there were 504 samples from East Asia (EAS), 489 from South Asia (SAS), 661 from Africa (AFR), and 347 from America (AMR), covering various populations.

The gnomAD genomes project v2.1.1 comprises 15,691 genomes and represents diverse populations worldwide, including Africans, Americans, Asians, and Europeans. Within European populations, the majority originated from north-western Europe, Estonia, Finland, and a smaller representation from southern Europe [4].

2.9. Population Frequencies

Based on the population projects described in the previous section, three populations were generated for this study as reference: (1) gnomAD, with the original frequency from the gnomAD genome project; (2) *1KGP_noIBS*, with the mean of all 1KGP population's

frequencies, excluding the IBS cohort of 107 individuals; and (3) *spain*, combining the frequencies of the IBS and MGP cohorts. The integration process for the *spain* population consisted of adding the number of total alternate alleles divided by the sum of the total number of alleles across the two cohorts.

2.10. Principal Components Analysis, Admixture, and F_{ST} Analysis

The original VCF files for each chromosome from the 1KGP phase 3 were downloaded on 27 September 2023, from https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. Subsequently, these files were merged using the Picard MergeVCFs tool. Access to the raw data from MGP was granted upon request, and the multi-sample VCF was retrieved for the EGA repository at https://ega-archive.org/datasets/EGAD00001003101, accessed on 27 September 2023. Then, the VCF files from 1KGP, MGP, and NAVARREsel were combined using the GATK tool's CombineGVCFs function. The resulting combined file was used to perform Principal Component Analysis (PCA) with R libraries SNPRelate and SNPassoc, conduct ADMIXTURE (v1.3.0) [29] from 3 to 7 genetic components (K), and calculate the mean pairwise F_{ST} values between populations using vcftools (v0.1.17) [30].

2.11. Enrichment Analysis

Functional enrichment, including pathway (KEGG), biological process (GO), disease (OMIM), and human phenotype ontology (HPO), was performed using WebGestalt (https://www.webgestalt.org/, accessed on 2 October 2023) [31].

3. Results

3.1. Navarrese Discovery Cohort

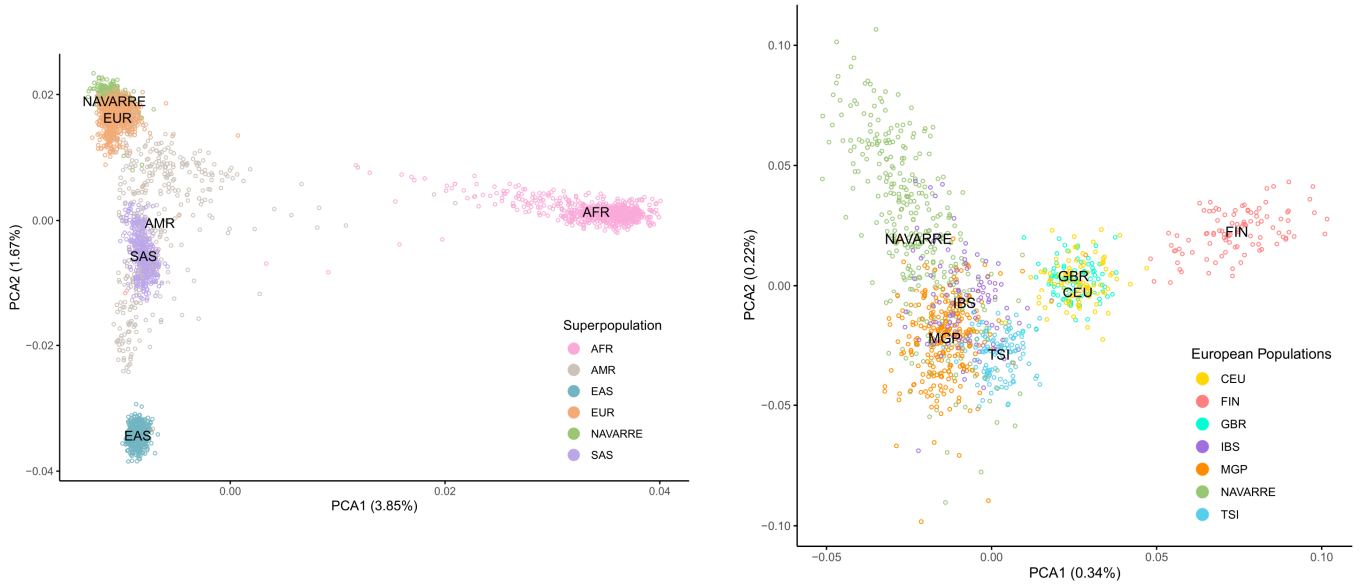
The *NAGEN1000* WGS Navarrese project was composed of 688 individuals from 294 families (mainly trios) with a rare disease. The WGS was conducted with a mean coverage of 30X, providing comprehensive genomic data across the entire genome. For our study, a subset of this cohort was selected satisfying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

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

3.2. Genetic Variation between Navarre, Spanish, and Global Populations

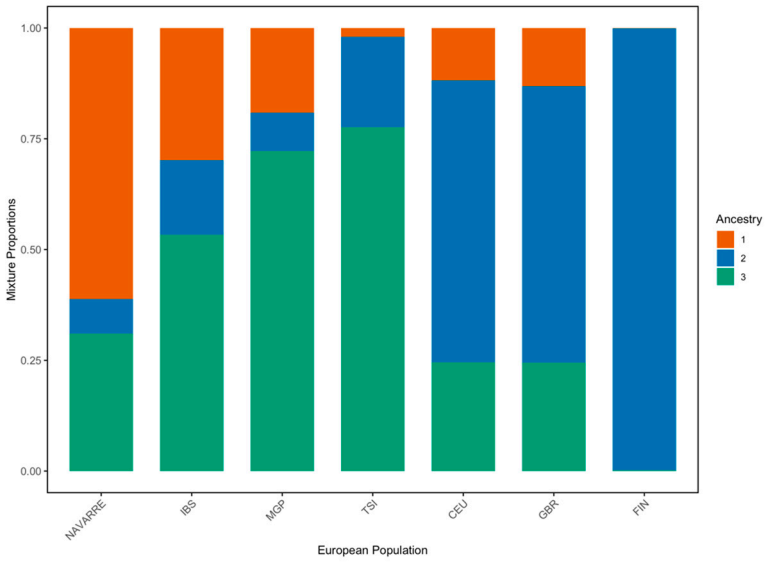
We performed a principal component analysis (PCA) on the shared variants between NAVARREsel, 1KGP, and MGP to depict its relationship. We observed a clear distinction between Navarre and Asian/African populations, reflecting established genetic differences (Figure 2a). Conversely, an overlap is observed between Navarre and European populations, emphasizing their genetic affinity. Thus, focusing on European populations (Figure 2b), we observed that Navarrese individuals are close to the Spanish populations (IBS and MGP) and exhibit proximity to Italian individuals (TSI). Supplementary Figure S2a,b illustrate plots PCA1 against PCA3, and PCA2 against PCA3. This observation is supported when estimating the ancestries of the European populations using ADMIXTURE [29]. The average number of ancestries in each population was calculated with the optimal component $K = 3$, which yielded the lowest cross-validation error (Figure 2c and Figure S3 admixture result at individual level). The Navarrese population showed the highest ancestral proportion on component 1 (61%). In the IBS and MGP populations, component 1 decreased to 30% and 20%, respectively, and was nearly absent (0.1%) in the FIN population. In contrast,

component 2 was predominant in the FIN population (99%), while being the lowest in the Navarrese cohort (7%).



(**a**)

(**b**)



(**c**)

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**) Genetic admixture analysis of 1128 individuals from 7 European populations for the optimal *K* value = 3. Abbreviations: 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.

To further analyse the genetic differentiation, we calculated the mean pairwise F_{ST} values. The lower F_{ST} value indicates greater similarity between populations. This occurred when comparing Navarre with the Spanish ($F_{ST(Navarre\text{-}IBS)}$ = 0.0001 and $F_{ST(Navarre\text{-}MGP)}$ = 0.0007)

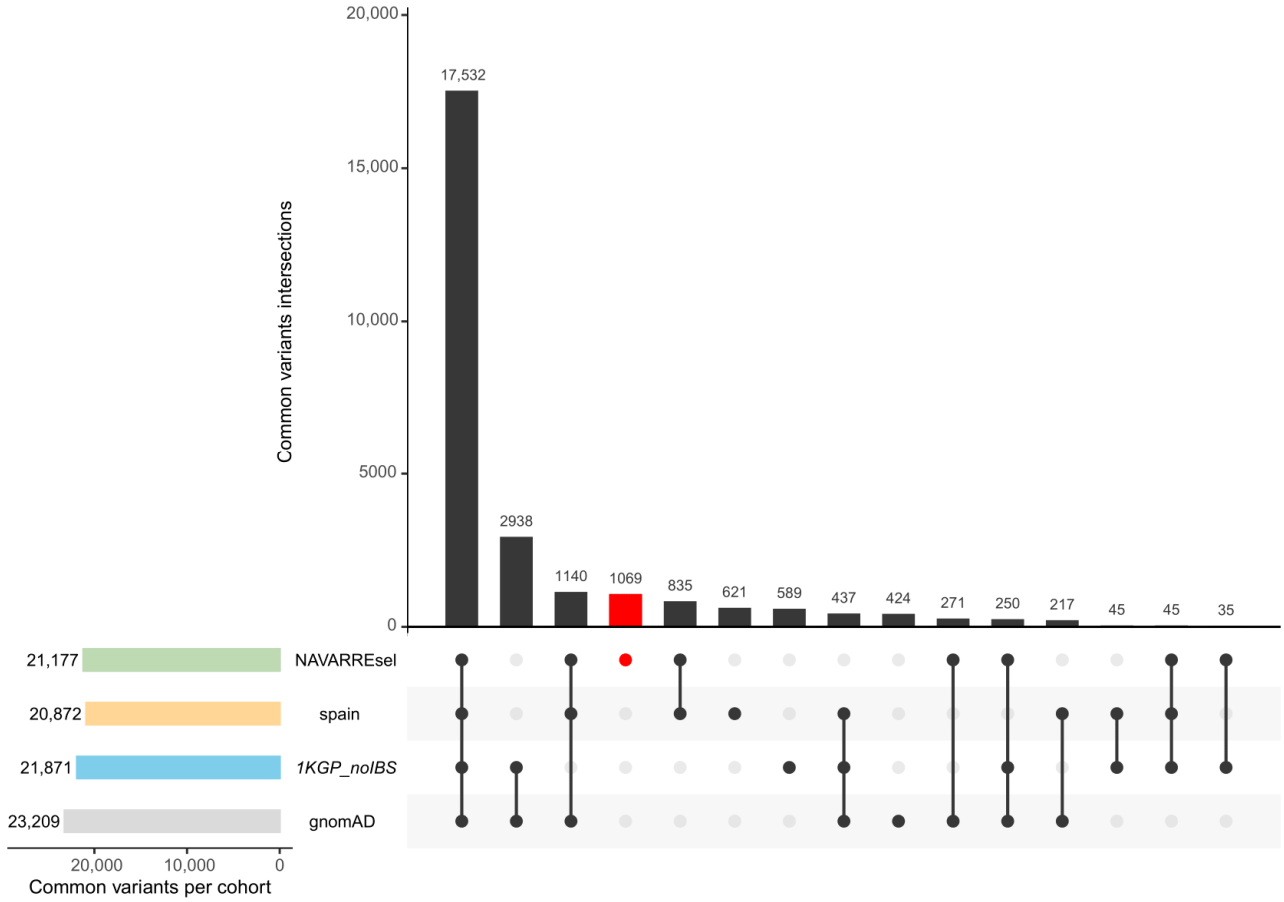
and Italian ($F_{ST(Navarre-TSI)}$ = 0.0014) populations. In contrast, the highest differentiation was observed against East Asian and African populations ($F_{ST(Navarre-EAS)}$ = 0.0328, $F_{ST(Navarre-AFR)}$ = 0.0434, Table S1).

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

3.3. Exclusive Common Variants in Navarre

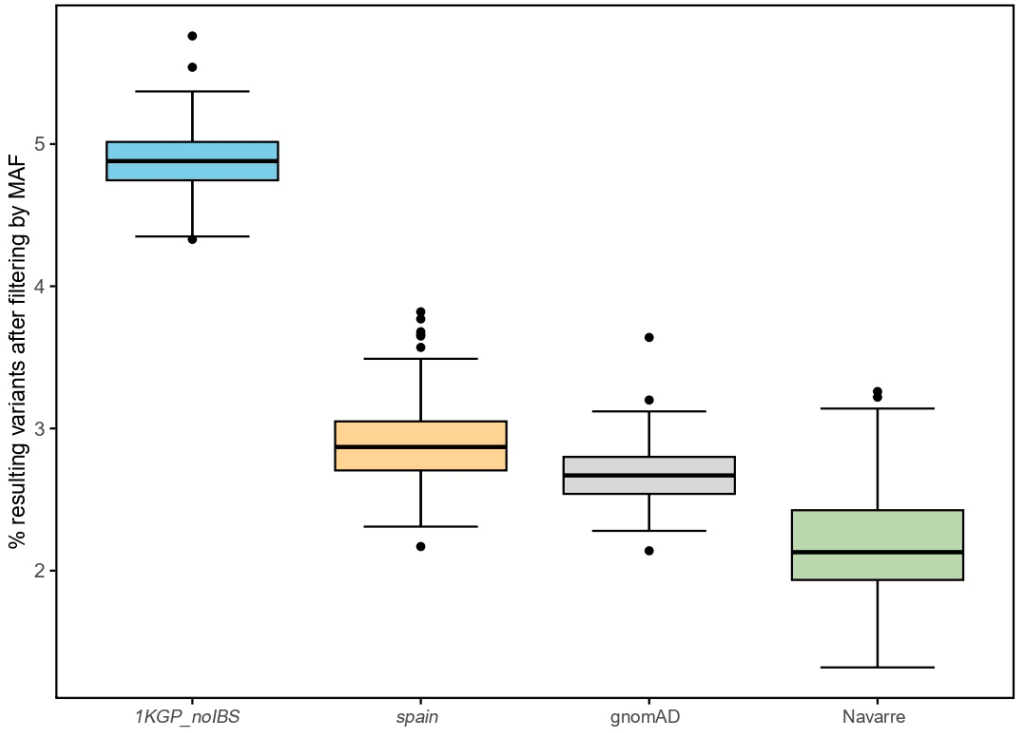
To identify exclusive Navarrese common variants, we examined allele frequency among the Navarre population and the three referenced populations: *1KGP_noIBS*, gnomAD, and *spain*. A comparison of the MAF revealed that most variants (17,532 SNVs) were classified as common (MAF > 1%) across the four populations. However, 835 variants exhibited higher prevalence solely in Spanish cohorts (Navarre and *spain*). Specifically, 1069 SNVs were identified as common in Navarre, and rare (MAF $\leq$ 1%) in the rest (Figure 3a).

To validate these 1069 variants, we used the NAVARREval cohort, a subset of 239 WES samples from the *pharmaNAGEN* project. The validation cohort consists of unrelated individuals of Spanish descent from the current Navarrese population diagnosed with Crohn's disease (159/239) or ulcerative colitis (86/239). Before validation, we assessed the association of these SNVs with these conditions by cross-referencing them with reported variants in the Inflammatory Bowel Disease database, which catalogues variants highly linked to the mentioned diseases [27]. The absence of the 1069 SNVs in this database ensured an unbiased and robust validation process.



(**a**)

Figure 3. *Cont.*



(**b**)

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

Among the 1069 variants initially identified, 998 were detected in NAVARREval with a call rate greater than 80% and demonstrated conformity to HWE. Notably, 676/998 of these SNVs (68%; p-value = 2.2×10^{-16}) were consistently classified as common in NAVARREval, confirming their prevalence within the Navarrese population (variants' information in Table S2). The validation cohort was sufficient to validate the Navarrese common variants, reaching a plateau in Figure 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 NAVARREval, indicating close MAF between both datasets (Figure S4). This exploration of MAF patterns ensures a comprehensive understanding of the genetic landscape within the Navarre population and its stability across different datasets. The MAF spectrum for these sets is depicted in Supplementary Figure S5.

3.4. Characterization of Common Navarrese Variants

The annotation of the 676 common Navarrese SNVs revealed 227 synonymous, 371 missense, and 5 loss-of-function (LoF) variants. These LoF variants were not reported in the ClinVar database [18] and were located in five distinct genes without an associated phenotype, according to OMIM [19]. Following the ACMG guidelines for variant classification, four were classified as variants of uncertain significance (VUS) and one as benign [33].

Clinically, 264/676 SNVs were reported in ClinVar: 1/264 as a risk factor, 181/264 as benign/likely benign, 32/264 as VUS, 48/264 as having conflicting interpretations, and 2/264 as likely pathogenic. These likely pathogenic missense variants were in *SCNN1B* [c.1688G > A p.Arg563Gln; MAFNAVARREsel = 0.013, MAFNAVARREval = 0.016] and in

PTGIS [c.824G > A p.Arg275Gln; MAFNAVARREsel = 0.013, MAFNAVARREval = 0.021], associated with "low renin hypertension" and "childhood-onset schizophrenia", respectively, according to ClinVar [18]. The evidence supporting these associations is limited, with a score of 1 out of 4 as reviewed by a single submitter record. Therefore, given its notable prevalence in the Navarrese population, being observed in healthy and affected (not related to this phenotype) individuals, these variants might be reconsidered and reclassified as VUS under the ACMG guidelines [34].

In silico analysis of the 676 variants with functional predictors revealed eight variants as pathogenic by three different pathogenicity tools (REVEL_score > 0.8 [24], CADD > 20 [23], and Polyphen indicating "probably" or "possibly") [26]. However, a comprehensive examination of clinical databases, including ClinVar [18], Varsome [20], and Franklin [21], 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 suggests that these variants are not disease causing.

Additionally, the variant in *BPIFB3* [c.387-1G > T; 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 [25].

Moreover, common Navarrese variants showed no impact on drug metabolism/efficacy, according to PharmGKB [22]. They did not exhibit significant enrichment in pathways, biological processes, related diseases, or phenotypic ontologies.

3.5. Refining Disease-Causing Variant Identification in the Navarrese Population

We identified common variants in the Navarrese population, highlighting population-specific importance in advancing personalized medicine. The aim was to improve the identification of disease-causing variants during genetic diagnosis using NGS. Therefore, we selected 127 WGS Navarrese patients from the *NAGEN1000* project diagnosed with rare disorders and extracted exonic SNVs on chromosomes 1 to 22, averaging 8871 variants per patient.

We refined the variant list by excluding common variants (MAF > 1%) from *1KGP_noIBS*, gnomAD, *spain*, and Navarre. The Navarrese filtering emerged as the most stringent, resulting in 2.1% of the initial set, compared to 2.7% with gnomAD, 2.9% with *spain* frequencies, and the least restrictive, 4.9% with *1KGP_noIBS* (Figure 3b). This underscores the effectiveness of the Navarrese-specific filter in prioritizing and streamlining genetic investigations.

4. Conclusions

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

Focusing on exclusively common variants in residents in Navarre compared with referenced populations, we obtained 1069 SNVs, of which 676 were validated in another Navarrese cohort. Of these, none showed clinical or pharmacological relevance beyond what was observed in the Spanish population [35]. This aligns with the expectation that common population variants are less likely to be associated with disease ethology.

Our findings underscore the relevance of considering population-specific factors in genomic diagnostics, which provides complementary insights alongside pangenome references [36]. However, the study would benefit from being expanded to include a larger cohort of participants (to provide greater statistical power to identify common variants) and increase the number of healthy individuals. In conclusion, by identifying and excluding common variants within the Navarrese population, we have successfully refined the identification of potential disease-causing variants, contributing to the advancement of

personalized medicine for individuals from Navarre. Further research will enhance these insights for broader applications.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes15050585/s1, Figure S1: Accumulative number of new variants contributed by individuals. (a) All common variants in NAVARREsel (21,174 SNVs). (b) The 676 variants are exclusively common in NAVARREsel and validated in NAVARREval. Figure S2: (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). Figure S3: Genetic admixture analysis of 1128 individuals from 7 European populations, at an individual level, for the optimal *K* value = 3. Figure S4: Comparison of MAF between NAVARREsel and NAVARREval encompassing both validated and non-validated variants, with a total of 998 SNVs. Figure S5: Minor allele frequency spectrum for the following distinct sets: 21,174 common variants in NAVARREsel (blue colour); 1069 common variants in NAVARREsel but not in *1KGP_noIBS*, gnomAD and spain (pink); and 676 common variants in both NAVARREsel (green) and NAVARREval (yellow). Common variants are defined with a MAF > 1%. Table S1: F_{st} values between Navarre against MGP and 1KGP populations. Table S2: Variants' information of the 676 exclusively common SNVs of the Navarre population. Details: "CHR": chromosome; "Position": variant position; "REF": reference allele; "ALT": alternate allele; "NAVARREsel_maf": MAF in NAVARREsel cohort; "NAVARREval_maf": MAF in NAVARREval cohort; "1KGP_phase3noIBS": MAF in 1KGP excluding IBS cohort; "Spain": MAF in spain cohort (combining IBS and MGP); "gnomAD": MAF in gnomAD; "Gene": gene where the variant is located; "Type": type of variant; "ExonicInfo": exonic variant type, if applicable; "ClinVar": ClinVar information; "SpliceAI": Splice AI results; "REVEL": REVEL results; "CADD": CADD results; "Polyphen2": Polyphen2 results.

Author Contributions: Conceptualisation: A.M., Á.A. and D.G.-C.; Clinical and sample collection: M.M., L.D.-M., Ó.T., J.S., M.A.-R. and S.P.-S.; Formal analysis: A.M., R.C., J.P.-F., V.A., D.L.-L. and M.P.-C.; Data curation: A.M. and NAGEN-Scheme; Investigation: A.M., E.H., M.A.-R., E.U.-L. and D.G.-C.; Funding acquisition: J.J.B. and Á.A.; Visualisation: A.M., E.H., M.A.-R. and E.U.-L.; Writing—original draft: A.M., E.H., M.A.-R., E.U.-L. and D.G.-C.; Writing—review and editing: A.M., M.A.-R., E.H., E.U.-L., S.B., S.P.-S., I.G., J.D., I.L., J.J.B., Á.A. and D.G.-C. All authors have read and agreed to the published version of the manuscript.

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. M.A.-R has a Postdoctoral Junior Leader—INCOMMING Fellowship from "la Caixa" Foundation (ID: 100010434) and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No. 847648 (fellowship code: LCF/BQ/PI21/11830009).

Institutional Review Board Statement: *NAGEN1000* and *pharmaNAGEN* were approved by the Navarra Ethics Committee for Clinical Research (CEIC Navarra).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no competing interests.

Appendix A

Table A1. NAGEN-Scheme.

Name	Department	Project
Anda Apiñaniz, Emma	Servicio de Endocrinología y Nutrición, HUN	NAGEN1000
Artigas López, Mercedes	Servicio de Genética Médica, HUN	NAGEN1000
Bandrés Elizalde, Eva	Servicio de Hematología, HUN	NAGEN1000
Basurte Elorz, Mª Teresa	Servicio de Cardiología, CHN	NAGEN1000
Brennan, Paul	NENC NHS Genomic Medicine Centre. Newcastle upon Tyne, UK.	NAGEN1000

Table A1. *Cont.*

Name	Department	Project
Celaya Lecea, Concepción	Subdirección de Farmacia, SNS-O	pharmaNAGEN
Cuesta Zorita, Manuel Jesús	Servicio de Psiquiatría, Salud mental	NAGEN1000
Curi Chercoles, Sergio Miguel	Servicio de Neumología, HUN	NAGEN1000
De la Cruz Sánchez, Susana	Servicio de Oncología Médica, HUN	NAGEN1000
Erviti López, Juan	Subdirección de Farmacia, SNS-O	pharmaNAGEN
Fanlo Mateo, Patricia	Servicio de Medicina Interna, HUN	NAGEN1000
González, Luis Angel	AVANTIA 400+.	NAGEN1000
Gonzalo Etayo	Navarra de Servicios y Tecnología NASERTIC. Spain	NAGEN1000
Gorricho Mendívil, Javier	Subdirección de Farmacia, SNS-O	pharmaNAGEN
Guerra Lacunza, Ana	Servicio de Aparato Digestivo, HUN	NAGEN1000
Gut, Ivo	Centro Nacional de Análisis Genómicos CNAG. Spain	NAGEN1000
Ibáñez Bosch, Rosario	Servicio de Endocrinología y Nutrición, HUN	NAGEN1000
Jiménez, Jorge	Navarra de Servicios y Tecnología NASERTIC. Spain	NAGEN1000
Lasheras, Gorka	Navarra de Servicios y Tecnología NASERTIC. Spain	NAGEN1000
Lorea Bueno	Pharmamodelling	pharmaNAGEN
Maite Sarobe Carricas	Servicio de Farmacia Hospitalaria, HUN	pharmaNAGEN
Mendioroz Iriarte, Maite	Servicio de Neurología, HUN	NAGEN1000
Molinuevo Ruiz de Zarate, José Ignacio	Servicio de Oftalmología, HUN	NAGEN1000
Montes Díaz, Marta	Servicio de Anatomía Patológica, HUN	NAGEN1000
Navarro, Adela	Servicio de Cardiología, HUN	pharmaNAGEN
Onintza Sayar	Pharmamodelling	pharmaNAGEN
Pinillos, Iñaki	Navarra de Servicios y Tecnología NASERTIC. Spain	NAGEN1000
Purroy Irurzon, Carolina Eugenia	Servicio de Nefrología, HUN	NAGEN1000
Sagaseta de Ilurdoz Uranga, Mª Josefa	Servicio de Pediatría, HUN	NAGEN1000
Santesteban Muruzabal, Raquel	Servicio de Dermatología, HUN	NAGEN1000
Vicuña, Miren	Servicio de Digestivo, HUN	pharmaNAGEN
Viguria, Mª Cruz	Servicio de Hematología, HUN	pharmaNAGEN
Yoldi Petri, Mª Eugenia	Servicio de Pediatría, HUN	NAGEN1000
Zubicaray Ugarteche, José Jacinto	Servicio de Otorrinolaringología, HUN	NAGEN1000
Zudaire, Maite	Servicio de Hematología, HUN	pharmaNAGEN

References

1. Satam, H.; Joshi, K.; Mangrolia, U.; Waghoo, S.; Zaidi, G.; Rawool, S.; Thakare, R.P.; Banday, S.; Mishra, A.K.; Das, G.; et al. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology* **2023**, *12*, 997. [CrossRef] [PubMed]
2. Fattahi, Z.; Beheshtian, M.; Mohseni, M.; Poustchi, H.; Sellars, E.; Nezhadi, S.H.; Amini, A.; Arzhangi, S.; Jalalvand, K.; Jamali, P.; et al. Iranome: A Catalog of Genomic Variations in the Iranian Population. *Hum. Mutat.* **2019**, *40*, 1968–1984. [CrossRef] [PubMed]
3. 1000 Genomes Project Consortium; Abecasis, G.R.; Altshuler, D.; Auton, A.; Brooks, L.D.; Durbin, R.M.; Gibbs, R.A.; Hurles, M.E.; McVean, G.A. A Map of Human Genome Variation from Population-Scale Sequencing. *Nature* **2010**, *467*, 1061–1073. [CrossRef] [PubMed]
4. Gudmundsson, S.; Singer-Berk, M.; Watts, N.A.; Phu, W.; Goodrich, J.K.; Solomonson, M.; Rehm, H.L.; MacArthur, D.G.; O'Donnell-Luria, A. Variant Interpretation Using Population Databases: Lessons from GnomAD. *Hum. Mutat.* **2022**, *43*, 1012–1030. [CrossRef] [PubMed]
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* **2022**, *13*, 556. [CrossRef]
6. Ramirez, A.H.; Sulieman, L.; Schlueter, D.J.; Halvorson, A.; Qian, J.; Ratsimbazafy, F.; Loperena, R.; Mayo, K.; Basford, M.; Deflaux, N.; et al. The All of Us Research Program: Data Quality, Utility, and Diversity. *Patterns* **2022**, *3*, 100570. [CrossRef] [PubMed]
7. Mitsuhashi, N.; Toyo-oka, L.; Katayama, T.; Kawashima, M.; Kawashima, S.; Miyazaki, K.; Takagi, T. TogoVar: A Comprehensive Japanese Genetic Variation Database. *Hum. Genome Var.* **2022**, *9*, 44. [CrossRef]
8. Dopazo, J.; Amadoz, A.; Bleda, M.; Garcia-Alonso, L.; Alemán, A.; García-García, F.; Rodriguez, J.A.; Daub, J.T.; Muntané, G.; Rueda, A.; et al. 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Mol. Biol. Evol.* **2016**, *33*, 1205–1218. [CrossRef]
9. Peña-Chilet, M.; Roldán, G.; Perez-Florido, J.; Ortuño, F.M.; Carmona, R.; Aquino, V.; Lopez-Lopez, D.; Loucera, C.; Fernandez-Rueda, J.L.; Gallego, A.; et al. CSVS, a Crowdsourcing Database of the Spanish Population Genetic Variability. *Nucleic Acids Res.* **2021**, *49*, D1130–D1137. [CrossRef]
10. NAGEN | Navarrabiomed. Available online: https://www.navarrabiomed.es/en/nagen (accessed on 28 November 2023).

11. Marco-Sola, S.; Sammeth, M.; Guigó, R.; Ribeca, P. The GEM Mapper: Fast, Accurate and Versatile Alignment by Filtration. *Nat. Methods* **2012**, *9*, 1185–1188. [CrossRef]
12. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef] [PubMed]
13. Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
14. Zheng, X.; Levine, D.; Shen, J.; Gogarten, S.M.; Laurie, C.; Weir, B.S. A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* **2012**, *28*, 3326–3328. [CrossRef] [PubMed]
15. Wigginton, J.E.; Cutler, D.J.; Abecasis, G.R. A Note on Exact Tests of Hardy-Weinberg Equilibrium. *Am. J. Hum. Genet.* **2005**, *76*, 887–893. [CrossRef] [PubMed]
16. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef] [PubMed]
17. Sherry, S.T. DbSNP: The NCBI Database of Genetic Variation. *Nucleic Acids Res.* **2001**, *29*, 308–311. [CrossRef] [PubMed]
18. Landrum, M.J.; Lee, J.M.; Riley, G.R.; Jang, W.; Rubinstein, W.S.; Church, D.M.; Maglott, D.R. ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype. *Nucleic Acids Res.* **2014**, *42*, D980–D985. [CrossRef] [PubMed]
19. Amberger, J.; Bocchini, C.A.; Scott, A.F.; Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM(R)). *Nucleic Acids Res.* **2009**, *37*, D793–D796. [CrossRef] [PubMed]
20. Kopanos, C.; Tsiolkas, V.; Kouris, A.; Chapple, C.E.; Albarca Aguilera, M.; Meyer, R.; Massouras, A. VarSome: The Human Genomic Variant Search Engine. *Bioinformatics* **2019**, *35*, 1978–1980. [CrossRef]
21. Franklin. Available online: https://franklin.genoox.com/clinical-db/home (accessed on 28 November 2023).
22. Thorn, C.F.; Klein, T.E.; Altman, R.B. PharmGKB: The Pharmacogenomics Knowledge Base. In *Pharmacogenomics: Methods and Protocols*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 311–320. [CrossRef]
23. Rentzsch, P.; Witten, D.; Cooper, G.M.; Shendure, J.; Kircher, M. CADD: Predicting the Deleteriousness of Variants throughout the Human Genome. *Nucleic Acids Res.* **2019**, *47*, D886–D894. [CrossRef]
24. Ioannidis, N.M.; Rothstein, J.H.; Pejaver, V.; Middha, S.; McDonnell, S.K.; Baheti, S.; Musolf, A.; Li, Q.; Holzinger, E.; Karyadi, D.; et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **2016**, *99*, 877–885. [CrossRef] [PubMed]
25. Jaganathan, K.; Kyriazopoulou Panagiotopoulou, S.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.I.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B.; et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **2019**, *176*, 535–548.e24. [CrossRef] [PubMed]
26. Adzhubei, I.; Jordan, D.M.; Sunyaev, S.R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **2013**, *76*, 7–20. [CrossRef] [PubMed]
27. Khan, F.; Radovanovic, A.; Gojobori, T.; Kaur, M. IBDDB: A Manually Curated and Text-Mining-Enhanced Database of Genes Involved in Inflammatory Bowel Disease. *Database* **2021**, *2021*, baab022. [CrossRef] [PubMed]
28. Auton, A.; Abecasis, G.R.; Altshuler, D.M.; Durbin, R.M.; Abecasis, G.R.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Donnelly, P.; Eichler, E.E.; et al. A Global Reference for Human Genetic Variation. *Nature* **2015**, *526*, 68–74. [CrossRef] [PubMed]
29. Patterson, N.; Moorjani, P.; Luo, Y.; Mallick, S.; Rohland, N.; Zhan, Y.; Genschoreck, T.; Webster, T.; Reich, D. Ancient Admixture in Human History. *Genetics* **2012**, *192*, 1065–1093. [CrossRef] [PubMed]
30. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The Variant Call Format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef] [PubMed]
31. Wang, J.; Vasaikar, S.; Shi, Z.; Greer, M.; Zhang, B. WebGestalt 2017: A More Comprehensive, Powerful, Flexible and Interactive Gene Set Enrichment Analysis Toolkit. *Nucleic Acids Res.* **2017**, *45*, W130–W137. [CrossRef] [PubMed]
32. Anderson, C.A.; Pettersson, F.H.; Clarke, G.M.; Cardon, L.R.; Morris, A.P.; Zondervan, K.T. Data Quality Control in Genetic Case-Control Association Studies. *Nat. Protoc.* **2010**, *5*, 1564–1573. [CrossRef]
33. Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W.W.; Hegde, M.; Lyon, E.; Spector, E.; et al. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **2015**, *17*, 405–424. [CrossRef]
34. Li, Q.; Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am. J. Hum. Genet.* **2017**, *100*, 267–280. [CrossRef] [PubMed]
35. Nunez-Torres, R.; Pita, G.; Peña-Chilet, M.; López-López, D.; Zamora, J.; Roldán, G.; Herráez, B.; Álvarez, N.; Alonso, M.R.; Dopazo, J.; et al. A Comprehensive Analysis of 21 Actionable Pharmacogenes in the Spanish Population: From Genetic Characterisation to Clinical Impact. *Pharmaceutics* **2023**, *15*, 1286. [CrossRef] [PubMed]
36. Liao, W.-W.; Asri, M.; Ebler, J.; Doerr, D.; Haukness, M.; Hickey, G.; Lu, S.; Lucas, J.K.; Monlong, J.; Abel, H.J.; et al. A Draft Human Pangenome Reference. *Nature* **2023**, *617*, 312–324. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.