The Iberian Roma genetic variant server; population structure, susceptibility to disease and adaptive traits.

Fabiola Mavillard[1,2]*, Javier Pérez-Florido[3,4,5,6]*, Francisco M Ortuño[3,7], Amador Valladares[1], Miren L Álvarez-Villegas[8], Gema Roldán[3], Rosario Carmona[3], Manuel Soriano[9], Santiago Susarte[9], Pilar Fuentes[9], Daniel López-López[3], Ana María Nuñez-Negrillo[10], Alejandra Carvajal[11], Yolanda Morgado[12], Daniel Arteaga[13], Rosa Ufano[14], Pablo Mir[1, 15, 16], Juan F Gamella[17], Joaquín Dopazo[3,4,5,6#], Carmen Paradas[1,18#], Macarena Cabrera-Serrano[1,18#]

* These authors have equally contributed to this work

1. Instituto de Biomedicina de Sevilla (IBiS), Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla. Sevilla, Spain.

2. Centro Investigación Biomédica en Red Enfermedades Neurodegenerativas (CIBERNED). Instituto de salud Carlos III. Sevilla, Spain.

3. Plataforma Andaluza de Medicina Computacional, Fundación Progreso y Salud (FPS), Hospital Virgen del Rocío, Sevilla, Spain.

4. Grupo de medicina computacional de sistemas, Instituto de Biomedicina de Sevilla (IBiS), Hospital Virgen del Rocío, Sevilla, Spain.

5. Nodo de Genómica Funcional, (INB-ELIXIR-es), Fundación Progreso y Salud (FPS), Hospital Virgen del Rocío, Sevilla 41013, Spain.

6. Bioinformática en Enfermedades raras (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de salud Carlos III. Sevilla, Spain.

7. Departamento de Ingeniería de Computadores, Automática y Robótica, Universidad de Granada, Granada, Spain

8. Servicio de Urgencias de Atención Primaria. Distrito Sevilla. Sevilla, Spain.

9. Centro de Servicios Sociales, Negociado de Servicios Especializados. Ayuntamiento de Sevilla, Sevilla, Spain.

10. Departamento de Enfermería, Facultad de Ciencias de la Salud, Universidad de Granada, Granada, Spain.

11. Departamento de Neurología, Hospital Virgen de las Nieves, Granada, Spain

12. Departamento de Neurología, Hospital Virgen de Valme, Sevilla, Spain

13. Centro de Salud Utrera Sur, Sevilla.

14. Centro de Salud Poligono Sur, Sevilla.

15. Unidad de Trastornos del Movimiento, Servicio de Neurología y Neurofisiología Clínica. Hospital Universitario Virgen del Rocío, Sevilla, Spain.

16. Departamento de Medicina, Facultad de Medicina, Universidad de Sevilla, Seville, Spain

17. Departamento de Antropología Social, Universidad de Granada, Spain.

18. Unidad Enfermedades Neuromusculares, Servicio de Neurología y Neurofisiología Clínica. Hospital Universitario Virgen del Rocío, Sevilla, Spain.

[#] **Correspondance to:**

Macarena Cabrera-Serrano

macabrera@us.es

Department of Neurology. Hospital Universitario Virgen del Rocio. 41013 Sevilla. Spain

+34 955923069

Carmen Paradas

cparadas@us.es

Department of Neurology. Hospital Universitario Virgen del Rocio. 41013 Sevilla. Spain

+34 955923069

Joaquin Dopazo

Grupo de medicina computacional de sistemas, Instituto de Biomedicina de Sevilla (IBiS), Hospital Virgen del Rocío,41013 Sevilla, Spain.

joaquin.dopazo@juntadeandalucia.es

ABSTRACT

The Roma are the most numerous ethnic minority in Europe. The Iberian Roma arrived in the Iberian Peninsula five centuries ago and still today, they keep a strong group identity. Demographic and cultural reasons lie behind a high rate of Mendelian disease often related to founder variants. We have analysed exome data from 119 Iberian Roma individuals collected from 2018 to 2020. A database of variant frequency has been implemented (IRPVS) and made available online. We have analysed the carrier rate of founder private alleles as well as pathogenic variants present in the general population. Significant enrichment in structural variants involving gene clusters related to keratinization and epidermal growth suggest that evolutive mechanisms have developed towards climate and environmental adaptation. IRPVS can be accessed at http://irpvs.clinbioinfosspa.es/

AUTHOR SUMMARY

Reference data is necessary for the correct interpretation of genetic studies. Although most genetic variants are present in all populations, ancestry has an important impact in the genetic background. For that reason databases of genetic variant in populations are developed specifically for different ethnicities, being an important tool for genetic diagnosis. The Roma are the most numerous ethnic minority in Europe. In this study we have collected samples from healthy Roma individuals from Iberian descent and implemented a database of genetic variant to facilitate genetic diagnosis in this population. Analysis of structural variants that are specific to the Iberian Roma not found in other healthy population for which genetic data are available suggest evolution towards environmental adaptation.

INTRODUCTION

In recent years, we have witnessed a thorough transformation of the knowledge and understanding of genetic disease. With massive parallel sequencing becoming available and its cost decreasing rapidly, the number of known disease genes has exponentially grown. So has our understanding of their effects, leading to progressive improvement of methods and protocols for assessing and classifying the variants according to their disease-causing potential. The presence of a large amount of rare variants in populations has become evident[1-3]. Since most pathogenic variants are rare in the general population, and variant frequency being fundamental information during the variant-assessment process, implementation of public repositories of variant frequency, such as Gnomad[4], has become an indispensable tool for genetic diagnosis [5]. Accumulating evidence shows

disease susceptibility is associated with low-frequency variants[6, 7] that are often specific to populations of distinct ancestries [3, 7-11], which makes necessary having population-specific reference data. There are many examples of disease-causing variants that are specific to a population and may explain a large proportion of the cases of a disease in this group[12-14]. There is also the opposite situation in which a benign variant, frequent in a population due to founder effects, is rare in another population. If data from the wrong control population group is consulted, the assumed low frequency could be taken in support of the variant as a potential candidate for causing disease[15]. Using data from an incorrect reference group could not only prevent a correct diagnosis, but lead to misdiagnosis [16]. Commonly used protocols for variant-assessing highlight the necessity of using reference frequency data from the specific population, due to the presence of benign variants common in some populations and very rare in others [17]. The absence of normative data from some minorities causes patients to receive more uncertain results from genetic studies than patients from other populations groups well represented in public repositories [18]. This is particularly important for founder populations and isolated communities with a high number of private variants[19]. Numerous calls, from scientific and non-scientific sources, have been made to include ethnic minorities data in genomic studies for the sake of knowledge as well as equality of opportunities to access tailored medicine [16, 20, 21] (https://www.genome.gov/news/news-release/Genomic-databases-weakened-by-lack-of-non-European-populations),(https://theconversation.com/how-the-genomics-health-revolution-is-failing-ethnic-minorities-86385), (https://blogs.scientificamerican.com/voices/we-need-more-diversity-in-our-genomic-databases/) and (https://www.genome.gov/news/news-release/nih-awards-38-million-dollars-to-improve-utility-of-polygenic-risk-scores-in-diverse-populations). Some population specific databases have been developed[3, 22-28] and large databases of genetic variants are often subdivided by ethnic background or geographical origin[4]. Yet, for some populations there no reference data publicly available to be consulted. Expanded carrier screening programs, designed to include the whole population of a country or region, is less effective to detect pathogenic variants that are private to a minority for which there is not normative data [29]. For the Roma, being the largest ethnic minority in Europe, there is no available information of genetic variant frequency, which renders them in a situation of disadvantage and inequality.

According to estimates of the European Commission, the European Roma, also known as Romani, or by their own ethnonyms, Roma, Sinti, Kale, etc., comprise 10 to 12 million

people (https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/roma-eu/roma-equality-inclusion-and-participation-eu_en). Although there is no official census, the population size has gone through rapid expansion over the last decades[30]. From a genetic point of view, the Roma is a population with high incidence of genetic disease and private features that make it not comparable to the general European population, similarly to other founder populations, like Ashkenazi or Finnish. The Roma descend from a reduced number of ancestors[31]. The original population founders; referred to as proto-Roma, were located to north India by linguistic and later genetic studies. Migrations through Persia, and the Balkans on their way to Europe can be traced through genetic admixture[30, 32]. A high number of founder events together with endogamy, are some of the reasons behind the high rate of Mendelian disease in the Roma. Carrier rates for some diseases have been estimated to be as high as 5 to 20%[33]. Ancient founder mutations are present in all Roma groups, despite being geographically diverse[34]. Younger mutations are restricted to specific groups, like the *CTDP1* IVS6+389C→T causing congenital cataracts facial dysmorphism neuropathy to a few Balkan Roma groups[35] and the *BIN1* p.Arg234Cys variant, causing centronuclear myopathy, to the Iberian Roma, with a carrier rate of 3.5%[36].

The Iberian Roma are one of the most numerous Roma communities in Europe. It is estimated that the Roma first arrived in the Iberian peninsula around 1425 AD and had higher levels of admixture with the host population than other settlements[30]. Still, there is a strong cultural, historical and familial background that preserves a well-defined group identity to the present day.

The knowledge of the genetic background of the Roma through a systematic study of healthy volunteers will improve the diagnostic rate for diseases and thus treatment, prevention of complications, genetic counselling and prevention of further cases in a family. Besides, the identification of disease-causing mutations with high carrier rate among the population, along with an effective information program would enable informed decision-making of the affected individuals.

Prompted by the difficulties in variant assessing and genetic diagnosis in the Roma invidividuals, in this study we have collected blood samples from healthy volunteers of Roma ethnicity for exome sequencing and built a database of genetic variants in the Iberian Roma population that has been made available online. We have performed a comprehensive analysis of variants, autozygosity and evolutionary traces.

RESULTS

One hundred and nineteen DNA samples were included in the study. Of those, 88 were newly recruited for the study and rest were archive samples. The average age was 42 years (range 18 to 80) and 73% were women. Geographical origin was diverse including nine different Autonomous Communities in Spain and Portugal. Province of origin of the individual and both parents was available for 76 samples, with an intergenerational mobility rate of 51%.

A mean coverage of 105.84x±19.30x was obtained for the whole experiment with the majority of target bases (mean of 95.19%±3.18%) covered at ≥30x. These results ensured the quality of the variant calling process.

1. Iberian Roma population variant server (IRPVS) database

IRPVS is an open resource available at http://irpvs.clinbioinfosspa.es. The set of counts of high-quality variants identified in the Iberian Roma cohort have been uploaded into it for search purposes and will be updated with data from new samples as they become available.

The IRPVS interface

The initial screen (Figure 1a) requires the acceptance of the "Terms and conditions for the use of the IRPVS database" (http://irpvs.clinbioinfosspa.es/downloads/IRPVSTermsAndConditions_use.pdf) before any operation is run.

The search option allows querying the IRPVS database. In the left panel (Figure 1b) queries can be done by gene symbol, chromosomal regions, dbSNP entry, HGVS transcript nomenclature (HGVSc) or the HGVS protein nomenclature (HGVSp). The search can be filtered by Sequence Ontology terms for the variation consequences. Also, variants can be highlighted using different types of scores such as SIFT[37], Polyphen[38] , CADD [39] or GERP[40].

The results of the query (Figure 1c) include a list of the positions for which variation has been found in the IRPVS along with complementary data such as: chromosome, position, reference allele and alternative allele, genotype and allelic frequencies in the IRPVS database and impact and conservation indexes (e.g. SIFT, Polyphen, CADD, Gerp) . Allelic

frequency of each variant in other population databases including 1000 genomes project, ESP ("Exome Variant Server · Bio.Tools"), gnomAD v3.1.2 and the Spanish population from the MGP database (CSVS[28]), is also shown. The most deleterious consequence type found and annotation of the canonical transcript among all transcripts for a given variant are also shown. For variants present in the ClinVar[41] or COSMIC[42], the associated phenotype is included. ClinVar and COSMIC are annotated interactively on each query using the Cellbase[43] webservices. Also a visualization of the variant in the genomic context of the selected variant is shown (Figure 1d) in a parallel window based on the Genome Maps browser[44]. Additionally, some extra detailed information can be found on the population frequencies observed for the variant, the phenotype or the effect. Full documentation of IRPVS and a link to source code can be found at http://irpvs.clinbioinfosspa.es (Figure 1e).

2. The Relationship of the Iberian Roma with other populations

The comparison of the genetic variants present in the Iberian Roma (cohort of study) with all other subpopulations present in 1000 Genomes (including general Spanish population-IBS, among others) and MGP (Spanish Medical Genome Project cohort), shows proximity of the Iberian Roma to European and South Asian populations, in good agreement with previous genetic and demographic studies[30, 45]. The principal component analysis (PCA) shows the Iberian Roma cluster being well delimited and in close continuity with the general Spanish cluster and European populations (Figure 2). When compared to specific non-Iberian Roma cluster[46] and taking into account the migrant group, a very close proximity and an overlapping can be observed with samples from Eastern Europe and North/Western Europe respectively (Supplementary Figure 1).

3. Variant analysis

Variability distribution in the Iberian Roma population

Total numbers of variants observed in the cohort of study are shown in Table 1. Of note, 39.5% of the SNVs identified are homozygous (38% when considering both SNVs and Indels) compared to 36.6% of homozygous SNVs previously described for the Spanish MGP population[3]. Table 2 shows the number of variants identified in the Iberian Roma in this study comparing it to Spanish MGP, 1000 genomes (including IBS) and other non-iberian Roma[46] populations. For comparative purposes, we have calculated average numbers per individual in each population with similar results amongst populations in the case of detected variants. When taking into account singleton variants , the Iberian Roma

population shows the lowest average value, being in line with previous studies that have shown less singletons in Iberian Roma when compared to other populations[47].

Private variants

Variants that are private to the Iberian Roma were obtained by subtracting those present in Spanish MGP and 1000 genomes (Table 3). Approximately, 13% of the private variants (SNVs and Indels) are homozygous (9% homozygous SNVs and 17% homozygous indels) compared to 7% of the Spanish private MGP population where only SNVs are reported [3]. We found that 56% of the private variants of the IRPVS (including SNP and Indels) are present in only one individual (singletons), much lower than the 85% found in the Spanish MGP that includes only SNPs[3]. Moreover, 1% of the IRPVS population shares 45% of the private variants in the cohort, whereas for the Spanish MGP 1% of the population shares 5% of the private variants. These differences are higher when we look at all the variants in the population, including private and non-private. To illustrate this notion, Figure 3 compares Iberian Roma and Spanish MGP populations in terms of variants shared by growing fractions of the corresponding population showing that in the Iberian Roma, private variants are shared by more individuals compared to Spanish MGP.

In a similar fashion, Figure 4 illustrates the increment of new variants as more individuals are included in our cohort, showing saturation of the curve is not reached. Non-private variants grow more rapidly while private variants contribute to the growth to a lesser extent. When we look specifically at private variants, we can see how the contribution of singletons is similar to the contribution of polymorphic (present in more than one individual) private variants, showing parallel lines of growth. This is strikingly different from what was described for the Spanish MGP population, where the private variants grow mainly from addition of singletons with a small contribution of polymorphic variants [3].

To check, in terms of Roma population, the uniqueness of private Iberian Roma variants, the intersection of such dataset with non-Iberian Roma population was assessed, arising a significant set of 12,863 SNPs and 3,189 Indels that are unique of Iberian Roma population (84% of the private Iberian Roma variants). The distribution of variants in IRPVS according to the main consequence type categories given by *Cellbase* shows a lower proportion of non-disruptive variants that might change protein effectiveness (MODERATE category) and

a greater proportion of non-coding variants (MODIFIER category) when compared to other populations (MGP+1000G) in the same regions of interest (Figure 5). Interestingly, when considering only private Roma variants (i.e. not present in MGP+1000G population), there is a low proportion of non-disruptive and mostly harmless variants (MODERATE and LOW categories, respectively). On the contrary, there is a high proportion of non-coding variants (MODIFIER category) and sites with high (disruptive) impact in the protein (HIGH category), which includes splice-sites, stop-loss and frameshift variants (see Supplementary Figure 2 for details on each category). These results are also in line with previous studies that have shown more proportion of deleterious variants in Roma compare to non-Roma populations[47].

4. Autozygosity

Runs of homozygosity are more common and on average larger in the Iberian Roma than other populations. Specifically, the Iberian Roma population had an average of 14.3 RoH (>1Mb) per sample whereas this metric decreases to 10.8 RoH for the 1000G and MGP. RoH were also significantly longer ($P < 2.2e^{-16}$, Wilcoxon non-parametric test) for the Iberian Roma, with an average length of 8.3Mb compared to 3.78Mb for 1000G and MGP populations combined (Figure 6a). When we analyse the length of RoH per chromosome we see an even distribution involving all autosomes similarly (Figure 6b).

Another interesting metric is the proportion of the autosomal genome in RoH over the specified target regions, termed as F_{ROH}. The F_{ROH} provides an accurate estimation of the total inbreeding coefficient of each individual[48]. The Iberian Roma population revealed a mean F_{ROH} of 4.78% whereas only 1.6% was reached by the reference population (1000 Genomes and MGP datasets).

5. Deviation from Hardy-Weinberg (HW) equilibrium

A total of 95,765 SNVs from sites with a single alternative allele in the Iberian Roma cohort restricted to common captured regions in Iberian Roma and Spanish MGP populations and exonic regions from Refseq[49] were tested for possible deviations from the HW equilibrium. As a result, 5,818 (6.1%, Supplementary Table 1) variants deviated significantly from the equilibrium ($P < 0.05$), 1,468 (1.5%) variants could not be computed for numerical restrictions (NA) and the rest of variants (88,479, 92.4%) were in HW equilibrium.

Most of the out-of-equilibrium variants observed in the Iberian Roma cohort are also present in MGP+1000G population with the same single alternative allele (5,068 variants,

87%). For these variants, a chi-square comparison test has been carried out to assess the differences in allele frequencies between the Iberian Roma cohort and the Spanish MGP+1000G populations. In this comparison, allelic frequencies distribution was significantly different ($P < 0.05$) for 3,601 variants (71.1%, Supplementary Table 1), while there were no differences in 1,267 variants (25%). The rest of the variants (200, 3.9%) could not be assessed for numerical restrictions (NA). Amongst the small percentage of variants that clearly deviated from equilibrium ($P < 0.05$), some are known pathogenic variants described in the general population (not being Roma private variants). This is the case for the *GJB2* p.Met34Thr variant, which causes autosomal non-syndromic deafness. Its genotypes distribution in the Iberian Roma population is also significantly different when compared with MGP+1000G population ($P < 0.05$). *HOGA1* p.Gly287Val, responsible for autosomal recessive hyperoxaluria and *HINT1* p.His112Asn), cause of autosomal recessive axonal neuropathy are also in disequilibrium in the Iberian Roma, not being present in the MGP+1000G populations.

6. Pathogenic alleles

Forty one individuals in the cohort (34%) were carriers of a known pathogenic Roma founder variant. For some of these variants, the carrier rate was as high 10.9 % for the *GJB2* p.Trp24Ter, being very rare in the general European population. The *GJB2* p.Trp24Ter is a frequent cause of autosomal recessive non syndromic deafness among the Roma[50], with an allele frequency in our cohort of 5.5%, higher than previously described in other Roma groups[51]. Three individuals are heterozygous carriers of two different founder pathogenic variants each. Table 4 shows the most frequent Roma founder variants found in the cohort. Moreover, some pathogenic variants, described in the general population, not private to the Roma, have a remarkably high allele frequency in our cohort. Hundred and sixty variants classified as pathogenic or likely pathogenic by ClinVar (to date September 2022) were present in our study population (Supplementary Table 2). These data should be considered cautiously, since some of the variants classified as pathogenic or likely pathogenic by ClinVar correspond risk factors or susceptibility alleles rather than disease causing variants. Moreover, assertion criteria are not always provided for the listed variants. Of those 160 variants, 152 were more frequent in our cohort than in the general population of gnomAD[4]. A total of 113 individuals (95% of the cohort) carry two variants or more classified as pathogenic or likely pathogenic by ClinVar and 18 individuals (15% of the cohort) were carriers of seven or more pathogenic or likely pathogenic variants each. Three individuals are carriers of the *SPG7* p.Leu78Ter variant, known to cause autosomal

recessive hereditary spastic paraplegia[52], with an allele frequency of 1.5% in our cohort, much higher than the 0.2% and 0.01% allele frequencies in the South Asian and European subpopulations of Gnomad v3.1.2[4, 53] respectively. In fact, this variant has previously been reported in a Spanish Roma family[54].

7. Structural variants and adaptive mechanisms

Structural variants (SV) in the population are thought to have a role in adaptability and response to external pressures, with a great potential to lead to rapid evolution[55]. They are possibly involved in phenotypic diversity and disease susceptibility. We analysed the genomic regions harbouring SV that are specific to the Iberian Roma, and the genes present in these regions. Our analysis shows significant enrichment of deletions, inversions, insertions and duplications, but not translocations (Supplementary Table 3). Gene Ontology terms such as those related to inflammation, immune response and defence, including antigen presentation and processing, and cell-cell adhesion are significantly enriched (FDR adjusted p-value < 0.05). Keratinization and epidermal development were also significantly enriched in SV in our study population (FDR adjusted p-value < 0.05). Of note, 68% of the individuals in the cohort had SV involving KRTAP1-3, KRTAP5-7 or KRTAP5-8, encoding for keratin associated proteins. When we look specifically at inversions, over a 100-fold enrichment is observed in gene clusters related to oxygen and gas transport, and in particular to haemoglobin and haptoglobin complexes formation (Supplementary Table 3).

DISCUSSION

The Roma constitutes an ethnic group with a unique demographic history and cultural tradition that that distinguish them from neighbouring populations. Nowadays, we are witnessing an exponential growth of the diagnostic and therapeutic options for genetic disease, which makes it even more necessary to provide this population group with the resources to take advantage of the recent developments. In this work we have developed a database of population variant frequency in the Iberian Roma to help in the variant assessment processes during genetic diagnosis.

We have relied on the self-definition of the individuals as Roma for their inclusion in the study. Our PCA analysis shows a perfect clustering of the samples together and apart from

other non-Roma populations, in continuity with the Spanish and European clusters, supporting the accuracy of the classification of individual ethnicity based on self-identity (Figure 2). It is interesting to note these clusters do not overlap after five centuries of coexistence, although we know from previous studies that there has been admixture throughout the years[46]. However, there is a clear overlap with other non-Iberian Roma populations such as the North/Western Roma. The high number of RoH identified can be partially explained by the demographic history of the Roma with successive bottlenecks and reduced effective population size, while the presence of significantly long RoHs throughout the genome confirms the persistence of current and recent inbreeding in the population. These data are in keeping with the results of recent anthropological studies that showed consanguineous marriages amongst the Spanish Roma are decreasing in recent years, although still frequent[56].

Previous population studies have shown great numbers of SV, involving 4.8–9.5% of the genome, having around 100 genes completely deleted without apparent phenotypic consequences[57]. SV are thought to be related to adaptive evolution responding to environmental pressures and contributing to human diversity but also to disease susceptibility[58]. Our analysis of SV suggests the Iberian Roma have developed adaptive mechanisms involving immune response, similarly to other populations[55, 59]. More striking is the finding of significant SV enrichment in gene clusters related to keratinization and epidermal growth, suggesting the development of adaptive mechanisms related to climate and environmental stressors. Keratinization is associated with the evolution of hair in mammals. This gene family has evolved under selection among mammals as a response to environmental pressures to hair structure[60]. Interestingly, similar findings have been reported in Indian subpopulations[61]. Enrichment in SV involving other keratin related genes (KRTAP9-2, KRTAP9-3, and KRTAP9-8) have been described to be specific to African populations[55]. While these results show strong statistical significance, we must bear in mind that exome sequencing carries significant bias for the analysis of SV, missing most of the non-coding regions.

We have observed that some variants, including known pathogenic variants, are more common in the Iberian Roma than would be expected according to Hardy-Weinberg equilibrium. This deviation could be explained by small population size as well as genetic drift and selective coupling. Knowledge of the most prevalent pathogenic variants in a population facilitates diagnosis, more so in the field of rare disease, where phenotypes overlap and the process of finding a molecular diagnosis is often complex. For the Roma,

this information is particularly helpful, as it is often one or a few variants that causes the majority of the cases of a disease in the population. However, in this study, we have seen a high carrier rate for some non-Roma pathogenic variants that were not known to be prevalent among the Roma. In fact, some of these variants have a much higher carrier rate among the Iberian Roma than in the general European population. Admixture and gene flow can explain the transmission of a specific allele from a host population (p.e. general European) to Iberian Roma. Genetic drift and selective coupling are possible mechanisms causing an increase of the new allele in the Iberian Roma, however, ethnicity is often not registered in clinical notes, therefore, it is possible that the original descriptions of some of these pathogenic variants, where the ethnic background of patients is not mentioned, correspond to Roma patients. The relatively small sample size analysed may not reflect accurately the actual prevalence of these conditions in the population, however, for many of them, figures shown are not too different from those seen in other Roma subpopulations.

The absence of a cohort of Roma ancestry in population databases of genetic variant frequencies often hinders a precise diagnosis for patients of this ethnicity. The open access database generated in the course of this study will hopefully contribute to improve genetic diagnosis in the Roma by increasing the resources available to this population.

METHODS

Recruitment of participants and sample collection

Inclusion criteria were being over the age of 18, self-reported Roma ethnicity of both parents and four grandparents and understanding and signing informed consent. Exclusion criteria were having a first degree relative already included in the study and having a known hereditary monogenic disease, a neurodegenerative condition of unknown cause or other disorder estimated by the researchers to be likely genetic. Vascular risk factors such as diabetes or hypertension were not considered a cause for exclusion from the study. Minimal demographic information was gathered from each volunteer including place of birth of the individual, place of birth of both parents, family history of disease and age at sample collection. A blood sample was taken from each volunteer between 2017 and 2020, for DNA extraction. Additionally, archive DNA samples from healthy Roma individuals were provided by Biobank Galicia Sur Health Research Institute (PT13/0010/0022) among other collaborations in Spain. For these samples, no demographic or biographic information was available. Samples were de-identified for sequencing and information was pseudo-

anonymized and stored in a secure physical drive with restricted access under the custody of the researchers. IRB approval was obtained from bioethics and scientific committees of *Hospitales Virgen del Rocio-Macarena -Junta de Andalucía, Consejería de salud, igualdad y políticas sociales. (VºBº CEI33160037)*

QC and exome sequencing

All DNA samples went through quality control before sequencing. Samples were run in a 1% agarose gel to ensure integrity. For purity control, absorbance was checked using Qbit, discarding samples with a A260/A280 below 1.7 or above 2. Two different capture kits were used for exome sequencing; MedExome capture was used for the first 20 samples, spanning a ~47Mb target region. For the latter 99 samples, Xgen-exome-research-panel v1.0 was used, spanning a ~39Mb target region. A Nextseq 500 Illumina platform was used for sequencing.

Sequencing Data analysis

Two pipelines for processing the raw sequences (FastQ files) were developed. One for the discovery of SNVs and small indels (<50bp) and another for the discovery of structural variants (>=50bp).

SNV and small indel pipeline is based on GATK best practices[62] for a cohort study. For each sample, *FASTQC[63]* was used to assess quality of raw data and *fastp[64]* was run for quality pre-processing so that clean data is provided to downstream analysis. Then, filtered sequence reads were aligned to the reference human genome build hs37d5 (hg19) by using the BWA alignment tool[64]. The obtained mapped reads (BAM files) are then sorted by *samtools[65]* and duplicate reads are marked to mitigate biases introduced by data generation steps such as PCR amplification by means of *Picard tools* ("Picard Tools - By Broad Institute"). BAM files are later analysed in terms of QC using in-house scripts and the ngsCAT tool[66]. Then, *Base Quality Score Recalibration* (BQSR) is applied in a two-step procedure through GATK *BaseRecalibrator* and *ApplyBQSR* tools[67] with the aim of detecting and correcting for patterns of systematic errors in the base quality scores. After BQSR, per-sample variants (SNVs and small Indels) were identified through GATK's *HaplotypeCaller* tool, by generating a GVCF file per sample. Then, a joint genotyping taking into account all samples in the cohort is run through GATK's *GenomicsDBImport* and *GenotypeGVCF* tools. This way, the different records are merged together in a sophisticated manner, obtaining a set of joint-called SNPs and indels in a single multisample VCF file. With the aim of reducing putative false positives, a variant filtering step is run through the

GATK's Variant Quality Score Recalibration (VQSR) procedure restricted to the union of the two exome captures, producing a set of high-quality variants. This dataset is then annotated using the Cellbase database[43]. Finally, the set of high-quality variants is transformed into a set of counts of variants for the whole cohort which are inserted in the IRPVS database.

Structural variants (>=50bp)[68] pipeline is based on GRIDSS v2.7.3 software[69]. In order to characterize Iberian Roma-specific SVs, a Panel of Normal (PoN) SVs predicted using the same protocol in 138 samples belonging to the Navarra 1000 Genomes Project NAGEN1000 ("Proyecto Genoma Navarra NAGEN 1000 Navarra") (https://www.navarrabiomed.es/en/research/projects/nagen1000) was used. This way, we excluded from our Iberian Roma cohort the SVs found in three or more samples belonging to the PoN, according to GRIDSS recommendations. Finally, using Cellbase database, the list of genes affected by the remaining set of SVs was extracted. For deletions and duplications, we considered genes falling between the boundaries of the SVs. In the case of inversions, insertions and translocations, we considered only genes falling in the SV breakpoints. Finally, we performed a gene ontology[70] functional enrichment analysis with PANTHER[71] over the set of genes affected by SVs in 10 or more samples in our cohort to get enough representation of SVs in the Iberian Roma population

Comparison with other populations

For the analysis of Roma Iberian population in terms of overlapping with other populations for SNVs and small indels, three main reference datasets have been used: the 1000 Genomes Project[72], the Spanish population dataset from the Medical Genome Project[3] and the non-Iberian Roma dataset[46]. For the analysis in terms of autozygosity, the 1000 genomes and MGP datasets have been used. The Spanish population, sequenced in the context of the Medical Genome Project (http://www.clinbioinfosspa.es/content/medical-genome-project) includes 267 healthy, unrelated exome samples of Spanish origin (EGA, accession: EGAS00001000938) in a multisample VCF file[3]. We refer to this population as MGP.

For the 1000 genomes project, a total of five human macro-populations were used in this study, which included 661 African samples (YRI Yoruba in Ibadan from Nigeria, LWK Luhya in Webuye from Kenya, GWD Gambian in Western Divisions in the Gambia, MSL Mende from Sierra Leona, ESN Esan in Nigeria, ASW Americans of African Ancestry in SW USA, ACB African Caribbeans in Barbados), 347 Ad Mixed American samples (MXL Mexican Ancestry

from Los Angeles USA, PUR Puerto Ricans from Puerto Rico, CLM Colombians from Medellin Colombia and PEL Peruvians from Lima Peru), 504 East Asian samples (CHB Han Chinese in Beijing China, JPT Japanese in Tokyo Japan, CHS Southern Han Chinese, CDX Chinese Dai in xishuangbanna China and KHV Kink in Ho Chi Minh City Vietnam), 489 South Asian samples (GIH Gujarati Indian from Houston Texas, PJL Punjabi from Lahore Pakistan, BEB Bengali from Bangladesh, STU Sri Lankan Tamil from the UK and ITU Indian Telugu from the UK) and 503 European samples (CEU residents of UTAH, TSI from Tuscany in Italy, FIN Finnish from Finland, GBR British from England and Scotland and the IBS from Spain). The genome sequences of all 2,504 individuals corresponding to the five super populations were downloaded from the 1000 genomes web page (https://www.internationalgenome.org/), last accessed July 31, 2019 in multisample variant calling format (VCF).

The non-Iberian Roma dataset includes 30 samples belonging to four main migrant groups: 10 Balkan, 5 Vlax, 10 Romungro and 5 North/Western Roma coming from four countries: Macedonia, Hungary, Lithuania and Ukraine. FASTQ files where downloaded from EGA database (accession EGAD00001006024) and processed similar to Iberian Roma cohort (see Sequencing Data analysis section) to obtain VCF files.

This way and using our Iberian Roma dataset, a total of 2,920 samples have been studied. When comparing populations and with the aim of reducing bias in number of variants found per individual due to heterogeneity of data (genomes and exomes with different captured regions), the analysis was restricted to the same regions of the 2,920 samples, namely, common captured regions in Iberian Roma and MGP populations and exonic regions from RefSeq (downloaded in October, 2019). As a consequence, studied regions cover ~ 36.5 Mb.

PCA generation

To compare genomic structure of the Iberian Roma population against the aforementioned populations, a Principal Component Analysis (PCA) was performed with PLINK v1.90. The first two dimensions were considered and drawn using R package ggplot2.

Test for selection

In order to study possible deviations from the Hardy–Weinberg equilibrium we used the R package HWChisqStats[73], a function for the fast computation of chi-square statistics (or the corresponding p-values) for a large set of genetic variants (typically SNVs from sites with a single alternative allele). This test allows to evaluate all categories of variants:

autosomal, X chromosomal and pseudo-autosomal regions (PAR1 and PAR2) of the X-chromosome variants, with their own particularities. Additionally, a chi-square test was used to assess the significance of differences in allele frequencies in the Iberian Roma population when compared with the MGP+1000G populations.

Autozygosity study

The autozygosity of the Iberian Roma population was studied by detecting the runs of homozygosity (RoH) of each sample using *bcftools roh[74]*. The standard parameters recommended by the tool were applied to obtain those runs. The generated regions of autozygosity were then compared against homozygosity in 1000 genomes and MGP datasets, both in terms of their averaged length and the proportion of the autosomal genome (Froh). The Wilcoxon non-parametric statistical test was performed to determine whether homozygosity was significantly different among these datasets.

DATA AVAILABILITY

Sequence data has been deposited at the European Genome-phenome Archive (EGA, see https://ega-archive.org/ last accessed December 5, 2022), under accession number EGAS00001006758.

CODE AVAILAVILITY

IRPVS database code is available at https://github.com/babelomics/csvs/tree/gpvs . All the bioinformatics tools used are publicly available and referenced accordingly.

ACKNOWLEDGEMENTS

We thank all the participants and particularly the community of “El Vacie” (Sevilla), for taking part in the study and for their necessary insight. We also thank David Comas (Universitat Pompeu Fabra, Barcelona) for his contribution, and Luba Kalaydjieva, Gianina Ravenscroft and Nigel Laing (University of Western Australia) for their fruitful suggestions and revision of the manuscript.

FUNDING

This study has been funded by Instituto de Salud Carlos III through the projects PI16/00612 and PI20/01200 (MCS) (Co-funded by European Regional Development Fund/European Social Fund "A way to make Europe"/"Investing in your future") and Junta de Andalucia-Consejeria de Salud through the project PIER-0468-2019 (MCS). MCS has been supported by

ISCIII (JR15/00042) and Junta de Andalucia-Consejeria de Salud (B-0005-2017), JD has been supported by grants PID2020-117979RB-I00 from the Spanish Ministry of Science and Innovation and IMP/00019 from the Instituto de Salud Carlos III (ISCIII). RC has been supported by Junta de Andalucía-Consejería de Salud y Familias (RH-0052-2021) co-funded by the European Union, European Social Fund (FSE) 2014-2020.

CONTRIBUTIONS

M.C.S., F.M. and C.P. contributed to the study concept and design, F.M, A.V., M.L.A.V., M.S., S.S., P.F., A.M.N.N., A.C., Y.M., D.A., R.U, J.G.., P.M., C.P., and M.C.S. contributed to the recruitment of participants and acquisition of data, J.P.F., F.M.O., R.C. and D.L.L. performed the data analysis. G.R. developed the IRPVS website and the corresponding database. J.P.F. and M.C.S. drafted the manuscript, J.G. reviewed the manuscript and provided valuable feedback. C.P., J.D. and M.C.S. supervised the study and reviewed the manuscript for important intellectual content. All authors approved the submission of this manuscript.

COMPETING INTERESTS

All authors declare no competing interest.

REFERENCES

1. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012;336(6082):740-3. Epub 2012/05/15. doi: 10.1126/science.1217283. PubMed PMID: 22582263; PubMed Central PMCID: PMCPMC3586590.

2. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337(6090):100-4. Epub 20120517. doi: 10.1126/science.1217876. PubMed PMID: 22604722; PubMed Central PMCID: PMCPMC4319976.

3. Dopazo J, Amadoz A, Bleda M, Garcia-Alonso L, Aleman A, Garcia-Garcia F, et al. 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. Mol Biol Evol. 2016;33(5):1205-18. Epub 20160113. doi: 10.1093/molbev/msw005. PubMed PMID: 26764160; PubMed Central PMCID: PMCPMC4839216.

4. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43. Epub 20200527. doi: 10.1038/s41586-020-2308-7. PubMed PMID: 32461654; PubMed Central PMCID: PMCPMC7334197.

5. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Human mutation. 2021. Epub 20211202. doi: 10.1002/humu.24309. PubMed PMID: 34859531.

6. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. Human molecular genetics. 2010;19(R2):R125-30. Epub 20100830. doi: 10.1093/hmg/ddq365. PubMed PMID: 20805107; PubMed Central PMCID: PMCPMC2953739.

7. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-9. Epub 20120517. doi: 10.1126/science.1219240. PubMed PMID: 22604720; PubMed Central PMCID: PMCPMC3708544.

8. Kurki MI, Gaal EI, Kettunen J, Lappalainen T, Menelaou A, Anttila V, et al. High risk population isolate reveals low frequency variants predisposing to intracranial aneurysms. PLoS genetics. 2014;10(1):e1004134. Epub 20140130. doi: 10.1371/journal.pgen.1004134. PubMed PMID: 24497844; PubMed Central PMCID: PMCPMC3907358.

9. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Clan genomics and the complex architecture of human disease. Cell. 2011;147(1):32-43. doi: 10.1016/j.cell.2011.09.008. PubMed PMID: 21962505; PubMed Central PMCID: PMCPMC3656718.

10. Wiszniewski W, Hunter JV, Hanchard NA, Willer JR, Shaw C, Tian Q, et al. TM4SF20 ancestral deletion and susceptibility to a pediatric disorder of early language delay and cerebral white matter hyperintensities. American journal of human genetics. 2013;93(2):197-210. Epub 20130627. doi: 10.1016/j.ajhg.2013.05.027. PubMed PMID: 23810381; PubMed Central PMCID: PMCPMC3738832.

11. Cohen JC, Boerwinkle E, Mosley TH, Jr., Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. The New England journal of medicine. 2006;354(12):1264-72. doi: 10.1056/NEJMoa054013. PubMed PMID: 16554528.

12. Avela K, Sankila EM, Seitsonen S, Kuuluvainen L, Barton S, Gillies S, et al. A founder mutation in CERKL is a major cause of retinal dystrophy in Finland. Acta Ophthalmol. 2018;96(2):183-91. Epub 20171025. doi: 10.1111/aos.13551. PubMed PMID: 29068140.

13. Jalas C, Anderson SL, Laufer T, Martimucci K, Bulanov A, Xie X, et al. A founder mutation in the MPL gene causes congenital amegakaryocytic thrombocytopenia (CAMT) in the Ashkenazi Jewish population. Blood Cells Mol Dis. 2011;47(1):79-83. Epub 20110413. doi: 10.1016/j.bcmd.2011.03.006. PubMed PMID: 21489838.

14. Zeegers MP, van Poppel F, Vlietinck R, Spruijt L, Ostrer H. Founder mutations among the Dutch. European journal of human genetics : EJHG. 2004;12(7):591-600. doi: 10.1038/sj.ejhg.5201151. PubMed PMID: 15010701.

15. Azmanov DN, Chamova T, Tankard R, Gelev V, Bynevelt M, Florez L, et al. Challenges of diagnostic exome sequencing in an inbred founder population. Molecular genetics & genomic medicine. 2013;1(2):71-6. Epub 20130422. doi: 10.1002/mgg3.7. PubMed PMID: 24498604; PubMed Central PMCID: PMCPMC3865571.

16. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. The New England journal of medicine. 2016;375(7):655-65. doi: 10.1056/NEJMsa1507092. PubMed PMID: 27532831; PubMed Central PMCID: PMCPMC5292722.

17. Nykamp K, Anderson M, Powers M, Garcia J, Herrera B, Ho YY, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genetics in medicine : official journal of the American College of Medical Genetics. 2017;19(10):1105-17. Epub 20170511. doi: 10.1038/gim.2017.37. PubMed PMID: 28492532; PubMed Central PMCID: PMCPMC5632818.

18. Caswell-Jin JL, Gupta T, Hall E, Petrovchich IM, Mills MA, Kingham KE, et al. Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. Genetics in medicine : official journal of the American College of Medical Genetics. 2018;20(2):234-9. Epub 20170727. doi: 10.1038/gim.2017.96. PubMed PMID: 28749474.

19. Tan-Sindhunata MB, Mathijssen IB, Smit M, Baas F, de Vries JI, van der Voorn JP, et al. Identification of a Dutch founder mutation in MUSK causing fetal akinesia deformation sequence. European journal of human genetics : EJHG. 2015;23(9):1151-7. Epub 20141224. doi: 10.1038/ejhg.2014.273. PubMed PMID: 25537362; PubMed Central PMCID: PMCPMC4538208.

20. Jacobs C, Pearce B, Hoosain N, Benjeddou M. Lack of genomic diversity in the SLC47A1 gene within the indigenous Xhosa population. Drug Metab Pers Ther. 2016;31(2):107-14. doi: 10.1515/dmpt-2016-0007. PubMed PMID: 27226103.

21. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. Health Aff (Millwood). 2018;37(5):780-5. doi: 10.1377/hlthaff.2017.1595. PubMed PMID: 29733732.

22. Chiang CWK, Marcus JH, Sidore C, Biddanda A, Al-Asadi H, Zoledziewska M, et al. Genomic history of the Sardinian population. Nature genetics. 2018;50(10):1426-34. Epub

20180917. doi: 10.1038/s41588-018-0215-8. PubMed PMID: 30224645; PubMed Central PMCID: PMCPMC6168346.

23. Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, et al. Deep whole-genome sequencing of 100 southeast Asian Malays. American journal of human genetics. 2013;92(1):52-66. Epub 20130103. doi: 10.1016/j.ajhg.2012.12.005. PubMed PMID: 23290073; PubMed Central PMCID: PMCPMC3542459.

24. Zlotogora J, Patrinos GP. The Israeli National Genetic database: a 10-year experience. Hum Genomics. 2017;11(1):5. Epub 20170316. doi: 10.1186/s40246-017-0100-z. PubMed PMID: 28302154; PubMed Central PMCID: PMCPMC5356354.

25. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al. The Genome of the Netherlands: design, and project goals. European journal of human genetics : EJHG. 2014;22(2):221-7. Epub 20130529. doi: 10.1038/ejhg.2013.118. PubMed PMID: 23714750; PubMed Central PMCID: PMCPMC3895638.

26. Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, et al. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. PLoS genetics. 2013;9(9):e1003815. Epub 20130926. doi: 10.1371/journal.pgen.1003815. PubMed PMID: 24086152; PubMed Central PMCID: PMCPMC3784517.

27. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nature genetics. 2016;48(9):1071-6. Epub 20160718. doi: 10.1038/ng.3592. PubMed PMID: 27428751; PubMed Central PMCID: PMCPMC5019950.

28. Pena-Chilet M, Roldan G, Perez-Florido J, Ortuno FM, Carmona R, Aquino V, et al. CSVS, a crowdsourcing database of the Spanish population genetic variability. Nucleic acids research. 2021;49(D1):D1130-D7. doi: 10.1093/nar/gkaa794. PubMed PMID: 32990755; PubMed Central PMCID: PMCPMC7778906.

29. Easteal S, Arkell RM, Balboa RF, Bellingham SA, Brown AD, Calma T, et al. Equitable Expanded Carrier Screening Needs Indigenous Clinical and Population Genomic Data. American journal of human genetics. 2020;107(2):175-82. doi: 10.1016/j.ajhg.2020.06.005. PubMed PMID: 32763188; PubMed Central PMCID: PMCPMC7413856.

30. Font-Porterias N, Arauna LR, Poveda A, Bianco E, Rebato E, Prata MJ, et al. European Roma groups show complex West Eurasian admixture footprints and a common South Asian genetic origin. PLoS genetics. 2019;15(9):e1008417. Epub 20190923. doi: 10.1371/journal.pgen.1008417. PubMed PMID: 31545809; PubMed Central PMCID: PMCPMC6779411.

31. Martinez-Cruz B, Mendizabal I, Harmant C, de Pablo R, Ioana M, Angelicheva D, et al. Origins, admixture and founder lineages in European Roma. European journal of human genetics : EJHG. 2015. doi: 10.1038/ejhg.2015.201. PubMed PMID: 26374132.

32. Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, et al. Origins and divergence of the Roma (gypsies). American journal of human genetics. 2001;69(6):1314-31. Epub 20011109. doi: 10.1086/324681. PubMed PMID: 11704928; PubMed Central PMCID: PMCPMC1235543.

33. Kalaydjieva L, Gresham D, Calafell F. Genetic studies of the Roma (Gypsies): a review. BMC medical genetics. 2001;2:5. Epub 20010402. doi: 10.1186/1471-2350-2-5. PubMed PMID: 11299048; PubMed Central PMCID: PMCPMC31389.

34. Mendizabal I, Lao O, Marigorta UM, Kayser M, Comas D. Implications of population history of European Romani on genetic susceptibility to disease. Hum Hered. 2013;76(3-4):194-200. doi: 10.1159/000360762. PubMed PMID: 24861864.

35. Kalaydjieva L, Chamova T. CTDP1-Related Congenital Cataracts, Facial Dysmorphism, and Neuropathy. In: Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJH, et al., editors. GeneReviews((R)). Seattle (WA)1993.

36. Cabrera-Serrano M, Mavillard F, Biancalana V, Rivas E, Morar B, Hernandez-Lain A, et al. A Roma founder BIN1 mutation causes a novel phenotype of centronuclear myopathy with rigid spine. Neurology. 2018;91(4):e339-e48. doi: 10.1212/WNL.0000000000005862. PubMed PMID: 29950440; PubMed Central PMCID: PMCPMC6070382.

37. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003;31(13):3812-4. doi: 10.1093/nar/gkg509. PubMed PMID: 12824425; PubMed Central PMCID: PMCPMC168916.

38. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20. doi: 10.1002/0471142905.hg0720s76. PubMed PMID: 23315928; PubMed Central PMCID: PMCPMC4480630.

39. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761-3. Epub 20141022. doi: 10.1093/bioinformatics/btu703. PubMed PMID: 25338716; PubMed Central PMCID: PMCPMC4341060.

40. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS computational biology. 2010;6(12):e1001025. Epub 20101202. doi: 10.1371/journal.pcbi.1001025. PubMed PMID: 21152010; PubMed Central PMCID: PMCPMC2996323.

41. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research. 2018;46(D1):D1062-D7. doi: 10.1093/nar/gkx1153. PubMed PMID: 29165669; PubMed Central PMCID: PMCPMC5753237.

42. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic acids research. 2019;47(D1):D941-D7. doi: 10.1093/nar/gky1015. PubMed PMID: 30371878; PubMed Central PMCID: PMCPMC6323903.

43. Bleda M, Tarraga J, de Maria A, Salavert F, Garcia-Alonso L, Celma M, et al. CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Nucleic acids research. 2012;40(Web Server issue):W609-14. Epub 20120612. doi: 10.1093/nar/gks575. PubMed PMID: 22693220; PubMed Central PMCID: PMCPMC3394301.

44. Medina I, Salavert F, Sanchez R, de Maria A, Alonso R, Escobar P, et al. Genome Maps, a new generation genome browser. Nucleic acids research. 2013;41(Web Server issue):W41-6. Epub 20130608. doi: 10.1093/nar/gkt530. PubMed PMID: 23748955; PubMed Central PMCID: PMCPMC3692043.

45. Mendizabal I, Lao O, Marigorta UM, Wollstein A, Gusmao L, Ferak V, et al. Reconstructing the population history of European Romani from genome-wide data. Curr Biol. 2012;22(24):2342-9. Epub 20121206. doi: 10.1016/j.cub.2012.10.039. PubMed PMID: 23219723.

46. Bianco E, Laval G, Font-Porterias N, Garcia-Fernandez C, Dobon B, Sabido-Vera R, et al. Recent Common Origin, Reduced Population Size, and Marked Admixture Have Shaped European Roma Genomes. Mol Biol Evol. 2020;37(11):3175-87. doi: 10.1093/molbev/msaa156. PubMed PMID: 32589725.

47. Font-Porterias N, Caro-Consuegra R, Lucas-Sanchez M, Lopez M, Gimenez A, Carballo-Mesa A, et al. The Counteracting Effects of Demography on Functional Genomic Variation: The Roma Paradigm. Mol Biol Evol. 2021;38(7):2804-17. doi: 10.1093/molbev/msab070. PubMed PMID: 33713133; PubMed Central PMCID: PMCPMC8233508.

48. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nat Rev Genet. 2018;19(4):220-34. Epub 20180115. doi: 10.1038/nrg.2017.109. PubMed PMID: 29335644.

49. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2016;44(D1):D733-45. Epub 20151108. doi: 10.1093/nar/gkv1189. PubMed PMID: 26553804; PubMed Central PMCID: PMCPMC4702849.

50. Bouwer S, Angelicheva D, Chandler D, Seeman P, Tournev I, Kalaydjieva L. Carrier rates of the ancestral Indian W24X mutation in GJB2 in the general Gypsy population and individual subisolates. Genetic testing. 2007;11(4):455-8. doi: 10.1089/gte.2007.0048. PubMed PMID: 18294064.

51. Alvarez A, del Castillo I, Villamar M, Aguirre LA, Gonzalez-Neira A, Lopez-Nevot A, et al. High prevalence of the W24X mutation in the gene encoding connexin-26 (GJB2) in Spanish Romani (gypsies) with autosomal recessive non-syndromic hearing loss. American journal of medical genetics Part A. 2005;137A(3):255-8. doi: 10.1002/ajmg.a.30884. PubMed PMID: 16088916.

52. Choquet K, Tetreault M, Yang S, La Piana R, Dicaire MJ, Vanstone MR, et al. SPG7 mutations explain a significant proportion of French Canadian spastic ataxia cases. European journal of human genetics : EJHG. 2016;24(7):1016-21. Epub 20151202. doi: 10.1038/ejhg.2015.240. PubMed PMID: 26626314; PubMed Central PMCID: PMCPMC5070891.

53. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91. doi: 10.1038/nature19057. PubMed PMID: 27535533; PubMed Central PMCID: PMCPMC5018207.

54. Sanchez-Ferrero E, Coto E, Beetz C, Gamez J, Corao AI, Diaz M, et al. SPG7 mutational screening in spastic paraplegia patients supports a dominant effect for some mutations and a pathogenic role for p.A510V. Clinical genetics. 2013;83(3):257-62. Epub 20120521. doi: 10.1111/j.1399-0004.2012.01896.x. PubMed PMID: 22571692.

55. Eaaswarkhanth M, Pavlidis P, Gokcumen O. Geographic distribution and adaptive significance of genomic structural variants: an anthropological genetics perspective. Hum Biol. 2014;86(4):260-75. doi: 10.13110/humanbiology.86.4.0260. PubMed PMID: 25959693.

56. Gamella JF, Nunez-Negrillo AM. The Evolution of Consanguineous Marriages in the Archdiocese of Granada, Spain (1900-1979). Hum Biol. 2019;90(2):97-114. doi: 10.13110/humanbiology.90.2.02. PubMed PMID: 33951885.

57. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172-83. Epub 20150203. doi: 10.1038/nrg3871. PubMed PMID: 25645873.

58. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85-97. doi: 10.1038/nrg1767. PubMed PMID: 16418744.

59. Nguyen DQ, Webber C, Ponting CP. Bias of selection on human copy-number variants. PLoS genetics. 2006;2(2):e20. Epub 20060217. doi: 10.1371/journal.pgen.0020020. PubMed PMID: 16482228; PubMed Central PMCID: PMCPMC1366494.

60. Wu DD, Irwin DM, Zhang YP. Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. BMC Evol Biol. 2008;8:241. Epub 20080823. doi: 10.1186/1471-2148-8-241. PubMed PMID: 18721477; PubMed Central PMCID: PMCPMC2528016.

61. Gautam P, Chaurasia A, Bhattacharya A, Grover R, Indian Genome Variation C, Mukerji M, et al. Population diversity and adaptive evolution in keratinization genes: impact of environment in shaping skin phenotypes. Mol Biol Evol. 2015;32(3):555-73. Epub 20141221. doi: 10.1093/molbev/msu342. PubMed PMID: 25534032.

62. Poplin R, Ruano-Rubio V, Depristo MA, Fennell TJ, Carneiro MO, Van Der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. 2017.
63. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90. doi: 10.1093/bioinformatics/bty560. PubMed PMID: 30423086; PubMed Central PMCID: PMCPMC6129281.
64. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 20090518. doi: 10.1093/bioinformatics/btp324. PubMed PMID: 19451168; PubMed Central PMCID: PMCPMC2705234.
65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 20090608. doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PubMed Central PMCID: PMCPMC2723002.
66. Lopez-Domingo FJ, Florido JP, Rueda A, Dopazo J, Santoyo-Lopez J. ngsCAT: a tool to assess the efficiency of targeted enrichment sequencing. Bioinformatics. 2014;30(12):1767-8. Epub 20140226. doi: 10.1093/bioinformatics/btu108. PubMed PMID: 24578402.
67. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303. Epub 20100719. doi: 10.1101/gr.107524.110. PubMed PMID: 20644199; PubMed Central PMCID: PMCPMC2928508.
68. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic acids research. 2014;42(Database issue):D986-92. Epub 20131029. doi: 10.1093/nar/gkt958. PubMed PMID: 24174537; PubMed Central PMCID: PMCPMC3965079.
69. Cameron DL, Schroder J, Penington JS, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome research. 2017;27(12):2050-60. Epub 20171102. doi: 10.1101/gr.222109.117. PubMed PMID: 29097403; PubMed Central PMCID: PMCPMC5741059.
70. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000;25(1):25-9. doi: 10.1038/75556. PubMed PMID: 10802651; PubMed Central PMCID: PMCPMC3037419.
71. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic acids research. 2019;47(D1):D419-D26. doi: 10.1093/nar/gky1038. PubMed PMID: 30407594; PubMed Central PMCID: PMCPMC6323939.
72. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi: 10.1038/nature15393. PubMed PMID: 26432245; PubMed Central PMCID: PMCPMC4750478.
73. Graffelman J, Weir BS. Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome. Heredity (Edinb). 2016;116(6):558-68. Epub 20160413. doi: 10.1038/hdy.2016.20. PubMed PMID: 27071844; PubMed Central PMCID: PMCPMC4868269.
74. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32(11):1749-51. Epub 20160130. doi: 10.1093/bioinformatics/btw044. PubMed PMID: 26826718; PubMed Central PMCID: PMCPMC4892413.

FIGURES

Figure 1: The Iberian Roma Population Variant Server (IRPGS) database.

a Initial IRPVS page. **b** Query panel in the Search option. **c** List of variants found in the Iberian Roma population within the selected region along with complementary information on impact, conservation, other's population frequencies and phenotype. **d** Genomic browser that displays the selected variant in its genomic context. **e** Help icon with links to Full documentation, source code, database version and contact details.

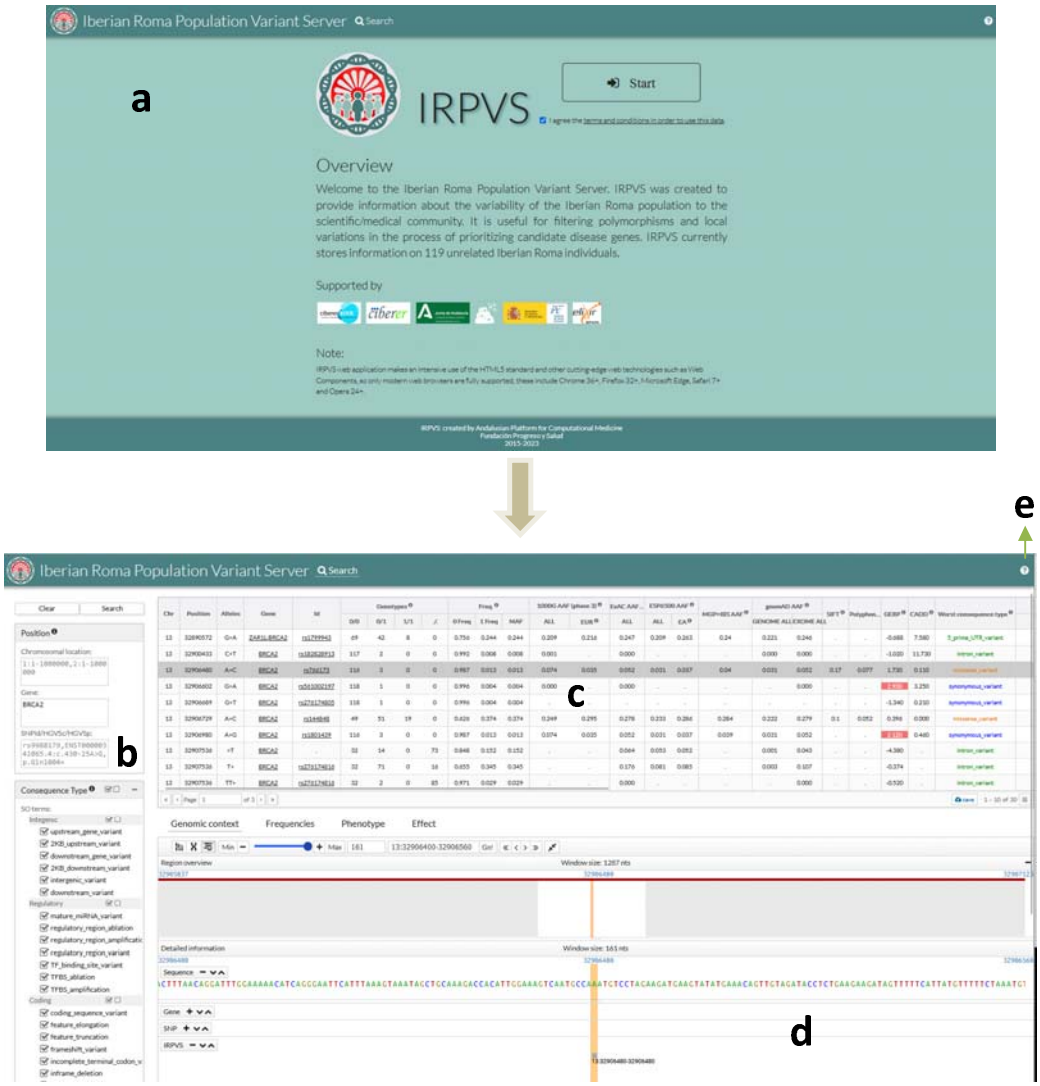


Figure 2. Principal Component analysis (PCA).

PCA analysis performed using the 119 Iberian Roma samples included in this study as well as MGP, Thousand Genomes subpopulations and non-iberian European Roma restricted to common captured regions in Iberian Roma and MGP populations and exonic regions from RefSeq.

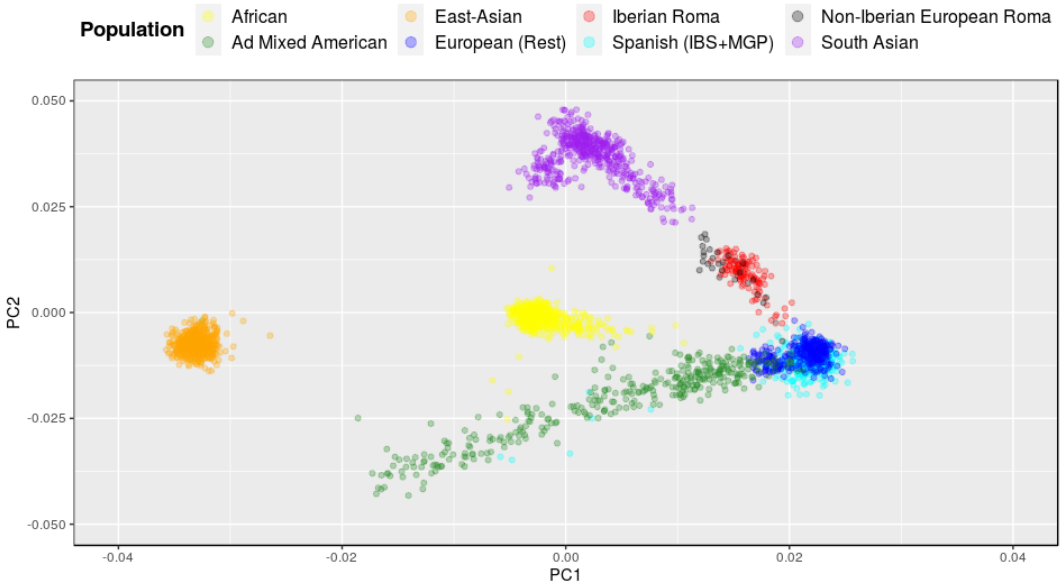


Figure 3. Variants shared by growing fraction of Iberian Roma and MGP populations.

Two comparisons are shown: (i) private Iberian Roma variants obtained by substracting those present in MGP or 1000 genomes (blue line) and private MGP variants obtained by substracting those present in 1000 genomes (red line) and (ii) Iberian Roma and MGP variants (blue and red dashed lines, respectively). In both cases, the analysis was restricted to common captured regions in Iberian Roma and MGP populations and exonic regions from RefSeq.

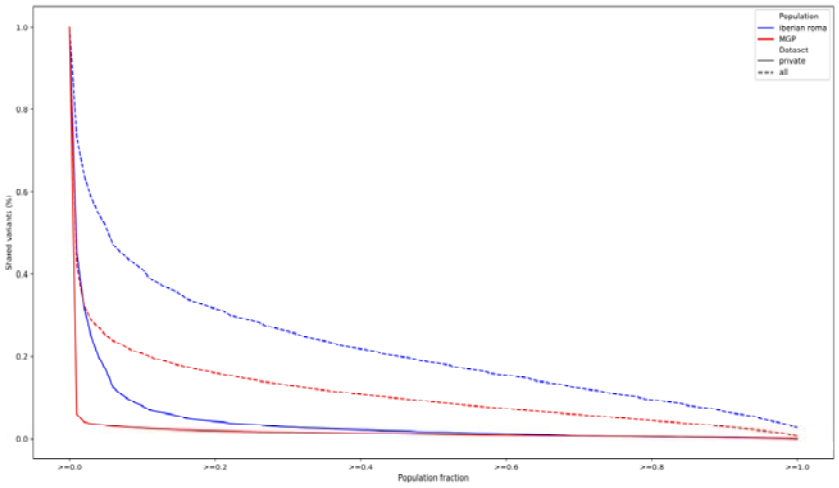


Figure 4. Accumulative number of variants contributed by individuals.

Accumulative number of variants contributed by individuals. As the number of individuals in the population grows, so does the total number of variants (dark blue line). Variants that are not private to the Iberian Roma account for a major part of this growth (orange line). Private variants show a slower pattern of growth contributing to a lesser extent to the total (light blue line). Private variants are decomposed in singletons, present in a single individual (light blue dashed line) and polymorphic private variants, present in more than one individual (light blue dotted line)

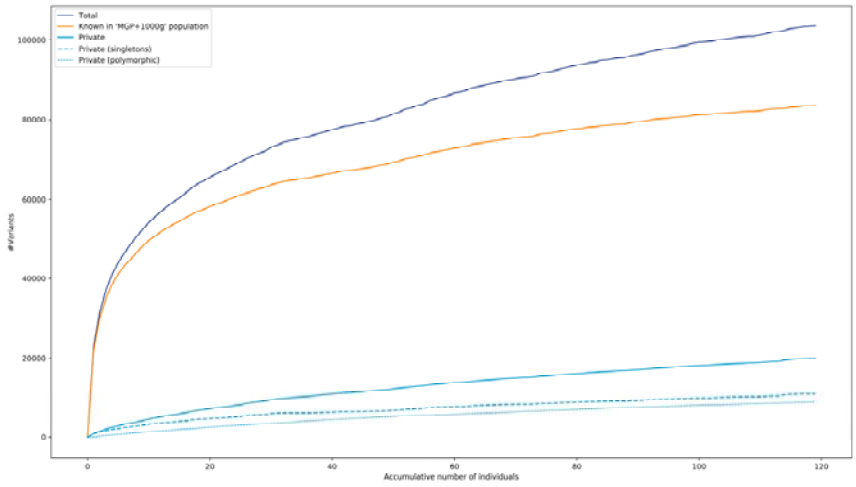


Figure 5. Distribution of variants per consequence type.

Distribution of variants in IRPVS (dark blue bar) and MGP+1000G (orange bar) according to Ensemble's worst consequence type (the worst effect that the variant has on the set of transcripts) obtained through *Cellbase* and restricted to the union of the two exome captures of IRPVS data. Iberian Roma variants not present in MGP+1000G population (light blue bar) is also shown. Consequence type is classified in one of the four main categories (HIGH, MODERATE, LOW and MODIFER) which reflects the severity or impact of the variant consequence.

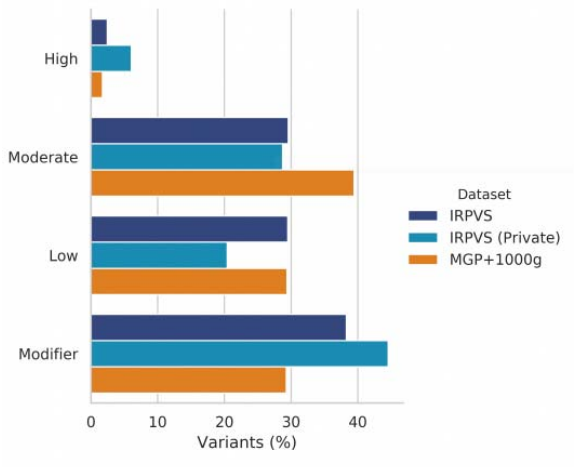
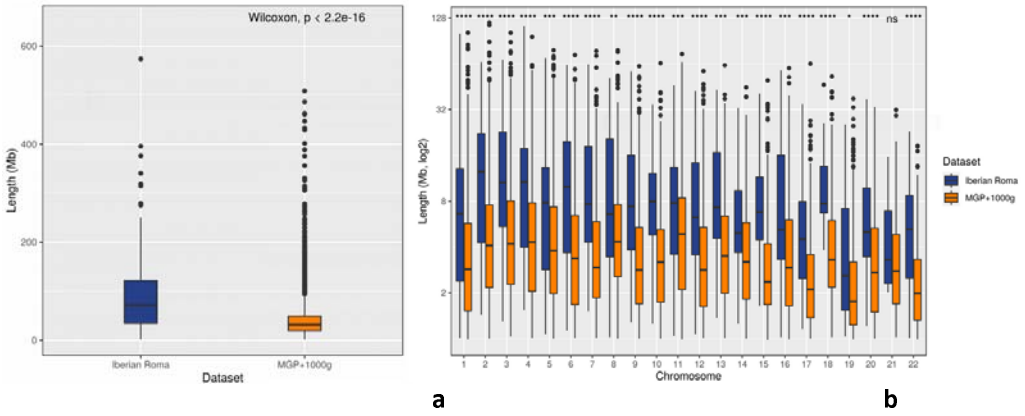


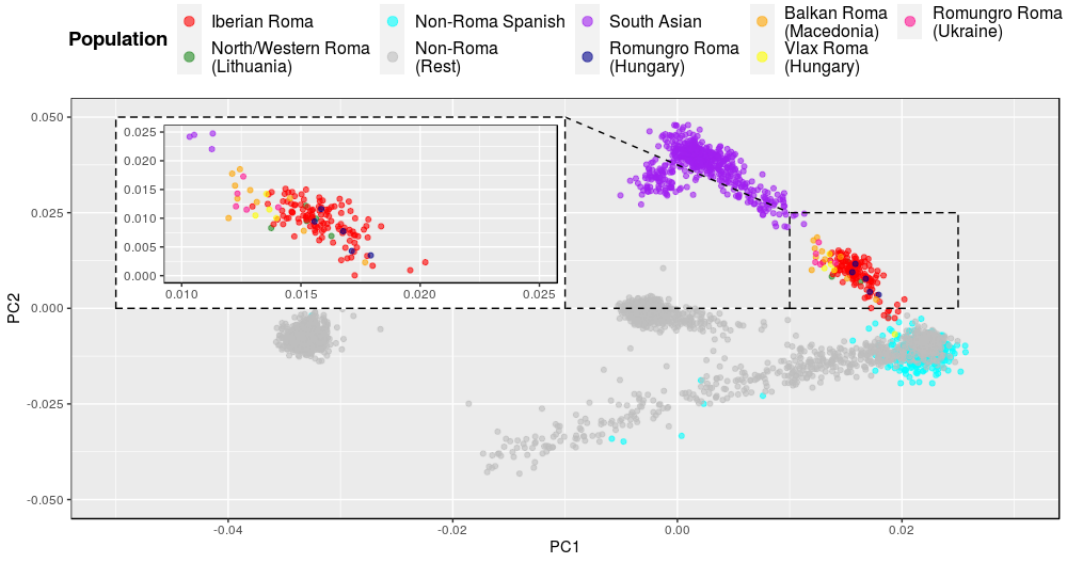
Figure 6 Runs of Homozygosity (RoH).

a Average length of RoH per sample in the IRPVS compared to a reference population (MGP + 1000G). b Length of RoH per chromosome in the IRPVS compared to MGP + 1000G. The following convention for symbols indicating statistical significance were used: (i) ns: p > 0.05, (ii) *: p <= 0.05, (iii) **: p <= 0.01, (iv) *** p <= 0.001 and (v) **** p <= 0.0001



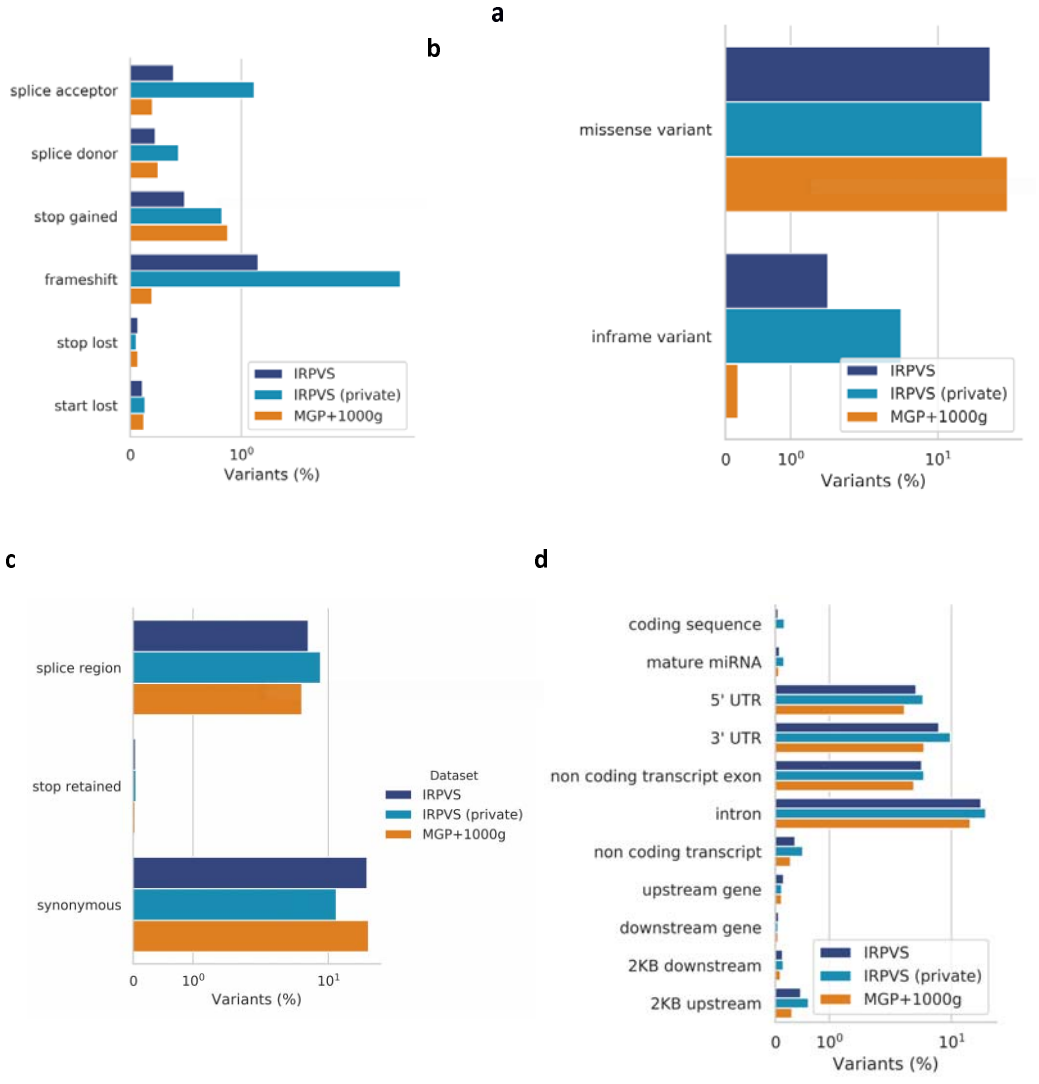
Supplementary Figure 1. Principal Component Analysis (PCA) with a focus on Iberian and non-Iberian Roma population

PCA analysis performed using the 119 Iberian Roma samples included in this study as well as MGP, Thousand Genomes subpopulations and non-iberian European Roma restricted to common captured regions in Iberian Roma and MGP populations and exonic regions from RefSeq. The proximity and overlapping with the different migrant groups from non-iberian European Roma is detailed.



Supplementary Figure 2. Distribution of variants per consequence types and impact.

Distribution of variants in IRPVS (dark blue bar) and MGP+1000G (orange bar) according to Ensemble's worst consequence type (the worst effect that the variant has on the set of transcripts) obtained through *Cellbase* and restricted to the union of the two exome captures of IRPVS data. Iberian Roma variants not present in MGP+1000G population (light blue bar) is also shown. Consequence type details are shown for each of the main categories: **a** HIGH, **b** MODERATE, **c** LOW and **d** MODIFIER. For visualization purposes, X-axes are in log scale and only terms with at least 0.03% of variants are shown. Details on consequence types (Sequence Ontology, SO, terms) are available at https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html



TABLES

Table 1. Variants observed in the Iberian Roma population

Total number of variants and average per individual in the Iberian Roma population restricted to the union of the two exome captures. A variant is labelled as singleton if it is present only in a single individual in the Iberian Roma cohort

	TOTAL VARIANTS	AVERAGE VARIANTS PER INDIVIDUAL	AVERAGE VARIANTS PER INDIVIDUAL (HOMOZYGOUS)
Positions with SNVs and Indels	179,597	40,740.83	15,564.47
Monoallelic positions	174,799	40,547.46	15,564.47
Multiallelic positions	4,798	193.37	0
SNVs	163,497	36,647.5	14,490.66
Indels	25,280	4,286.7	1,073.82
Singletons	47,063	395.49	14.87

Table 2. Total and average numbers of variants in different populations.

Total number of variants and average per individual in the Iberian Roma compared to other populations, restricted to common captured regions in Iberian Roma and MGP populations and exonic regions from RefSeq. To be fair in the comparison of singleton variants and their average value amongst cohorts, the whole set of 2,920 individuals have been taken into account, that is, a variant is labelled as a singleton if it is present in a single individual but not in any other of the remaining 2,919 individuals.

	POPULATION				
	MGP (N=267)	MGP+IBS (N=374)	MGP+1000g (N=2,771)	IB Roma (N=119)	Non-IB Roma (N=30)
SNVs	165,140	21,850	1,037,545	97,098	71,233
Indels	-	2,035*	6,665*	5,671	3,238
Singletons	37,962	55,734	520,779	9,402	4,540
Average SNVs per individual	18,077.93	19,256.82	23,102.69	21,698.74	22,624.53
Average Indels per individual	-	531.17*	552.06*	835.00	820.76
Average singletons per individual	142.18	149.02	187.94	79.01	151.33

* For MGP population, indels were not reported. Therefore these samples were not taken into account for indel metrics (total and average per individual)

Table 3. Private variants.

Variants observed in the Iberian Roma not present in the general Spanish population (MGP) and 1000 genomes. A variant is labelled as singleton if it is present only in a single individual in the private variant dataset.

	PRIVATE VARIANTS	AVERAGE PRIVATE PER INDIVIDUAL	AVERAGE PRIVATE VARIANTS PER INDIVIDUAL (HOMOZYGOUS)
Positions with SNVs and Indels	18,434	733.22	94.26
Monoallelic positions	17,975	721.08	94.26
Multiallelic positions	456	12.14	0
SNVs	15,095	388.07	35.13
Indels	4,054	357.29	59.13
Singletons	10,744	90.29	2.52

Table 4. Most frequent Roma founder pathogenic variants identified in the cohort. For each variant, the following information is provided, **Gene**: HGNC symbol of variant carrier gene; **Change**: Change at DNA or RNA level according to HGVS nomenclature; **Genotypes**: R/R refers to homozygous reference genotype counts in the Iberian Roma Cohort, R/A refers to heterozygous genotype counts in the Iberian Roma Cohort, A/A refers to homozygous alternative genotype counts in the Iberian Roma Cohort; **Carrier rate**: percentage of individuals in the Iberian Roma Cohort who carry the variant; **Frequency**: *Freq R* refers to the reference allele frequency in the Iberian Roma Cohort, *Freq A* refers to the alternative allele frequency in the Iberian Roma Cohort and *MAF* refers to the Minor Allele Frequency in the Iberian Roma Cohort; **1000G**: *All* refers to the alternative allele frequency in the whole population of 1000 genomes Project Database (Phase 3) and *Eur* refers to the alternative allele frequency in the European Population of 1000 genomes Project Database (Phase 3); **Exac**: *All* refers to the alternative allele frequency in the whole population of Exome Aggregation Consortium (ExAC); **Associated Disease**: variant-related disease; **Reference**: bibliographic reference describing the association between the variant and the disease; **rs**: Reference single nucleotide polymorphism ID in dbSNP (https://www.ncbi.nlm.nih.gov/snp/); **Variant**: variant in format *chromosome : position within the chromosome* (human genome build Grch37; hg19) *: Reference Allele : Alternative Allele*; **Clinical Significance**: clinical significance according to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)

Gene	Change	Genotypes			Carrier rate (%)	Frequency			1000G		Exac	Associated disease	Reference	rs	Variant	Clinical Significance
		R/R	R/A	A/A		Freq R	Freq A	MAF	All	Eur	All					
GJB2	p.Trp24Ter	106	13	0	10.924	0.945	0.055	0.055	0.0004	0	0.001	Deafness	Schrauwen, 2019	rs104894396	13:20763650:C:T	Pathogenic
SLC12A3	c.1180+1G>T	110	8	1	7.563	0.958	0.042	0.042	-	-	0	Gitelman syndrome	Coto, 2004	rs749098014	16:56912074:G:T	Pathogenic
EIF2AK4	p.Pro1115Leu	112	7	0	5.882	0.971	0.029	0.029	-	-	0	Pulmonary venoocclusive disease 2	Navas-Tejedor, 2017	rs774906916	15:40295502:C:T	Pathogenic
BIN1	p.Arg234Cys	116	3	0	2.521	0.987	0.013	0.013	-	-	0	Centronuclear myopathy 2	Cabrera-Serrano, 2018	rs777176261	2:127821221:G:A	Likely pathogenic
SPG7	p.Leu78ter	116	3	0	2.521	0.987	0.013	0.013	-	-	0	Hereditary spastic paraplegia	Sanchez-Ferrero, 2013	rs121918358	16:89576947:T:A	Pathogenic
GALK1	p. Pro28Thr	117	2	0	1.681	0.992	0.008	0.008	-	-	0	Galactokinase deficiency with cataracts	Kalaydjieva, 1999	rs104894572	17:73761136:G:T	Pathogenic
BCKDHA	p.Trp391Gly	117	2	0	1.681	0.992	0.008	0.008	-	-	0	Mapple syrup urine disease	Quental, 2008	rs398123489	19:41916544:C:-	Pathogenic
CD8A	p.Gly111Ser	117	2	0	1.681	0.992	0.008	0.008	-	-	0	$CD8^+$ T-cell immunodeficiency	Mancebo, 2008	rs121918660	2:87017523:C:T	Pathogenic
ITGA8	c.2982+2T>C	117	2	0	1.681	0.992	0.008	0.008	-	-	-	Bilateral renal agenesis	Humbert, 2014	rs587777279	10:15573047:A:G	Pathogenic/Likely pathogenic
FANCA	p.Gln99ter	117	2	0	1.681	0.992	0.008	0.008	-	-	-	Fanconi anemia	Callén, 2005	rs1057516430	16:89877468:G:A	Pathogenic
SGCG	p.Cys283Tyr	118	1	0	0.840	0.996	0.004	0.004	-	-	-	Muscular dystrophy, limb-girdle, autosomal recessive 5	Piccolo, 1996	rs104894422	13:23898652:G:A	Pathogenic/Likely pathogenic
CHRNE	p.Glu443LysfsTer64	118	1	0	0.840	0.996	0.004	0.004	-	-	0	Myasthenic syndrome, congenital, 4C,	Kalaydjieva, 2001	rs763258280	17:4802186:C:-	Pathogenic
LTBP2	p.Arg299Ter	118	1	0	0.840	0.996	0.004	0.004	-	-	0	Weill-Marchesani syndrome 3/ Glaucoma	Morlini, 2018	rs121918355	14:75022332:G:A	Pathogenic
MANBA	c.2158-2A>G	118	1	0	0.840	0.996	0.004	0.004	-	-	0	Mannosidosis, beta	Brozkova, 2020	rs772852668	4:103556204:T:C	Pathogenic
SHOX	p.Ala170Pro	118	1	0	0.840	0.996	0.004	0.004	-	-	-	Léri-Weill dyschondrosteosis/Langer mesomelic dysplasia	Barca-Tierno, 2011	rs397514461	X:601577:G:C	Pathogenic
ACADS	p.Glu104del	118	1	0	0.840	0.996	0.004	0.004	-	-	0	Short chain acyl-CoA dehydrogenase deficiency	Lisyová, 2018	rs387906308	12:121174884:GGA:-	Pathogenic/Likely pathogenic
SH3TC2	p.Arg1109ter	118	1	0	0.840	0.996	0.004	0.004	-	-	0	Charcot Mary Tooth 4C	Gooding, 2005	rs80338934	5:148389835:G:A	Pathogenic
GLB1	p.Arg59Hys	118	1	0	0.840	0.996	0.004	0.004	-	-	0	GM1 gangliosidosis	Santamaria, 2007	rs72555392	3:33114105:C:T	Pathogenic/Likely pathogenic

Supplementary Table 1. List of genomic variants in the study population which significantly ($p < 0.05$) deviate from Hardy-Weinberg (HW) equilibrium. Table provided as excel file: Supplementary table 1.xlsx

Supplementary Table 2. Pathogenic alleles by Clinvar present in the study population. Table provided as excel file: Supplementary table 2.xlsx

Supplementary Table 3. Gene Ontology terms more significantly involved by structural variants. Table provided as excel file: Supplementary table 3.xlsx