The Counteracting Effects of Demography on Functional Genomic Variation: The Roma Paradigm

Neus Font-Porterias, Rocio Caro-Consuegra, Marcel Lucas-Sánchez, Marie Lopez, Aaron Giménez, Annabel Carballo-Mesa, Elena Bosch, Francesc Calafell, Lluís Quintana-Murci, David Comas

The Counteracting Effects of Demography on Functional Genomic Variation: The Roma Paradigm

Neus Font-Porterias,[1] Rocio Caro-Consuegra,[1] Marcel Lucas-Sánchez,[1] Marie Lopez,[2] Aaron Giménez,[3] Annabel Carballo-Mesa,[4] Elena Bosch,[1,5] Francesc Calafell,[1] Lluís Quintana-Murci,[2,6] and David Comas *[,1]

[1]Departament de Ciències Experimentals i de la Salut, Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Spain

[2]Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, Paris, France

[3]Facultat de Sociologia, Universitat Autònoma de Barcelona, Barcelona, Spain

[4]Facultat de Geografia i Història, Universitat de Barcelona, Barcelona, Spain

[5]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Reus, Spain

[6]Human Genomics and Evolution, Collège de France, Paris, France

*Corresponding author: E-mail: david.comas@upf.edu.

Associate editor: Kelley Harris

Abstract

Demographic history plays a major role in shaping the distribution of genomic variation. Yet the interaction between different demographic forces and their effects in the genomes is not fully resolved in human populations. Here, we focus on the Roma population, the largest transnational ethnic minority in Europe. They have a South Asian origin and their demographic history is characterized by recent dispersals, multiple founder events, and extensive gene flow from non-Roma groups. Through the analyses of new high-coverage whole exome sequences and genome-wide array data for 89 Iberian Roma individuals together with forward simulations, we show that founder effects have reduced their genetic diversity and proportion of rare variants, gene flow has counteracted the increase in mutational load, runs of homozygosity show ancestry-specific patterns of accumulation of deleterious homozygotes, and selection signals primarily derive from preadmixture adaptation in the Roma population sources. The present study shows how two demographic forces, bottlenecks and admixture, act in opposite directions and have long-term balancing effects on the Roma genomes. Understanding how demography and gene flow shape the genome of an admixed population provides an opportunity to elucidate how genomic variation is modeled in human populations.

Key words: **demography, admixture, adaptation, exomes, Roma, mutational load.**

Introduction

The distribution of human genomic variation is affected by population demographic history, especially regarding low-frequency protein-coding variants. Previous studies show an excess of rare population-specific functional variants, as a result of a recent and explosive human population growth (Coventry et al. 2010; Gravel et al. 2011; Keinan and Clark 2011; Marth et al. 2011; Nelson et al. 2012; Tennessen et al. 2012). Population bottlenecks and founder effects have also had a great impact on modeling the spectrum of functional variation: for example, the French-Canadian founder population contains a large proportion of rare and putatively damaging functional variants (Casals et al. 2013) and the severe bottleneck in the Greenlandic Inuit increased the frequency of the extant deleterious variants (Pedersen et al. 2017), among other examples

As in other human populations, the complex demographic history in the Roma population (also known by the misnomer of *Gypsies*) has influenced their patterns of genetic diversity. The Roma population is a highly heterogeneous and socially persecuted ethnic minority, whose diaspora has been historically poorly documented (Fraser 1992), although several studies have aimed to characterize their demographic history. Previous linguistic, anthropological, and genetic data have shown evidence for a South Asian origin of the Roma 1,500 years ago and a posterior diaspora toward the European continent. Once in Europe, they experienced extensive gene flow with non-Roma populations and suffered multiple founder events (Turner 1927; Boerger 1984; Gresham et al. 2001; Sun et al. 2006; Bouwer et al. 2007; Gusmão et al. 2008; Azmanov et al. 2011; Mendizabal et al. 2011; Mendizabal et al. 2012; Moorjani et al. 2013; Martínez-

Article

Cruz et al. 2016; Melegh et al. 2017; Font-Porterias et al. 2019; Bianco et al. 2020; Dobon et al. 2020; García-Fernández et al. 2020). Thus, the genetic study of the European Roma provides a unique opportunity to evaluate the extent to which recent demographic events impact the patterns of diversity of the human genome.

Previous genetic studies suggest that population history also shapes differences across populations in mutational load (i.e., reduction in population fitness due to the accumulation of deleterious mutations compared with a theoretical optimal fitness), with implications in the genetic architecture of diseases. In small populations, the accumulation of deleterious variants might be the result of random fluctuations in allele frequency (i.e., genetic drift) due to a reduced efficacy of purifying selection (Gravel 2016). Most analyses have focused on differences between African and non-African populations leading to controversial results (Lohmueller et al. 2008; Lohmueller 2014; Simons et al. 2014; Do et al. 2015; Henn et al. 2016; Simons and Sella 2016), but with a general agreement that the demographic history mostly impacts the recessive mutational load, rather than the additive load (Simons et al. 2014). In addition, by studying the temporal trajectories of mutational load, a transient increase in recessive load is observed in a small African hunter-gatherer population, which is balanced by gene flow from an expanding farmer population (Lopez et al. 2018). However, the interaction between increased genetic drift and admixture on mutational load remains poorly resolved in non-African populations.

The effects of recent demographic and social processes have a higher impact on some genomic features, such as runs of homozygosity (ROHs). ROHs are enriched for deleterious homozygous variants, when compared with regions of the genome where ROHs are absent(Szpiech et al. 2013; Kaiser et al. 2015; Ceballos et al. 2018). In populations with reduced genetic diversity, the accumulation of more deleterious than synonymous variants in long ROHs is the result of recent founder events and parental relatedness (Szpiech et al. 2013). Moreover, in admixed populations, this enrichment in deleterious homozygotes inside ROHs depends on the specific ancestry of the segment and the characteristics of the source populations (Szpiech et al. 2019). In the case of the Roma, the extensive admixture between their South Asian and West Eurasian sources, together with multiple bottleneck events (Mendizabal et al. 2012; Moorjani et al. 2013; Font-Porterias et al. 2019), might have left ancestry-specific patterns in these regions.

Likewise, population demography also impacts genetic adaptation through the action of positive selection. In admixed populations, positive selection can be studied in terms of postadmixture and preadmixture selection. In "postadmixture selection," an admixed population receives adaptive alleles through gene flow; subsequently, both the adaptive alleles and the variation linked to them rise in frequency in the admixed population, resulting in a local ancestry deviation (Seldin et al. 2011; Bhatia et al. 2014). However, when local ancestry deviation is weak or absent and the selection signal is present in both the admixed and source populations, the process can be defined as "preadmixture adaptation," since after admixture, genetic drift or weak positive selection maintained the initial signal (Bhatia et al. 2014). A clear example of postadmixture selection is found in the population of Madagascar, where African ancestry is increased around the Duffy blood group gene that confers resistance to malaria (Pierron et al. 2018). In contrast, African Americans carry signals of preadmixture selection, occurring in Africa prior to the slave trade to America, around the *HBB* gene. Although signals of adaptive admixture can be identified with local ancestry deviations, weaker or polygenic selection is more difficult to detect through these approaches (Bhatia et al. 2014).

In the present study, we provide new insights into how demography affects the distribution of functional variation, focusing on the Roma population as a model. Throughout all analyses, the admixture experienced by Roma is assessed since it has been shown that ancestry background highly impacts the genetic variation in human populations (Seldin et al. 2011; Kidd et al. 2012; Pierron et al. 2018; Szpiech et al. 2019). We first evaluate the degree of genetic diversity and frequency distribution of deleterious variants comparing Roma and non-Roma populations. In addition, we estimate both current mutational and its trajectory during the Roma history. We also examine whether ROHs are enriched for deleterious homozygotes in the Roma population and whether this enrichment is ancestry dependent. We finally focus on detecting genomic regions under pre- or postadmixture positive selection.

Results

Reduced Genetic Diversity with an Excess of Common Deleterious Variants

We sequenced 89 new high-coverage whole exome sequences (WES) from Iberian Roma and merged them with previously published European and South Asian WES (Auton et al. 2015; Tombácz et al. 2017) that were used as ancestry sources of the Roma. We also genotyped a single-nucleotide polymorphism (SNP) array in a subset of 62 Iberian Roma. After quality control filters, the WES data set contains 410,225 variants in 527 samples, and the array data set 474,632 variants in 487 samples (see supplementary note 1, Supplementary Material online, supplementary figs. 1–6, Supplementary Material online, and supplementary table 1, Supplementary Material online for further details). We also merged both data sets to increase the number of covered genome-wide variants (878,162 SNPs).

We first assessed the population structure in our data set of Roma samples together with non-Roma (European and South Asian groups, supplementary table 1, Supplementary Material online). Principal component analysis (PCA) results are compatible with Roma being admixed between European and South Asian samples (fig. 1A) and ADMIXTURE analysis (supplementary fig. 7, Supplementary Material online) at $k = 2$ shows the Roma as a mixture of two cluster components found in European and South Asian samples. At $k = 3$ (lowest cross-validation error value), the Roma individuals display membership in a specific component (colored in

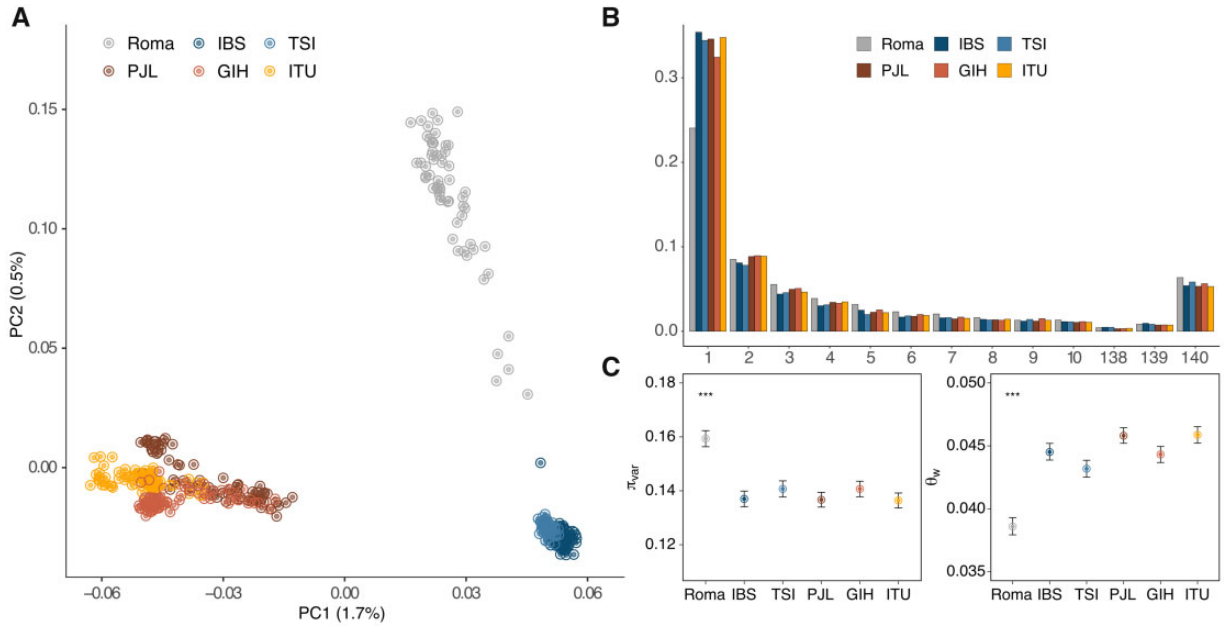

FIG. 1. Population structure and distribution of synonymous variants. (*A*) PCA with the merged data set of genome-wide array and WES variants. (*B*) Unfolded site-frequency spectrum for synonymous WES variants. (*C*) Genetic diversity measures (π_{var} and θ_w) from synonymous WES variants. Other diversity metrics (Tajima's *D* and θ_π) are shown in supplementary figure 8, Supplementary Material online. Significant differences were tested between Roma and non-Roma populations (*** refers to *P* value <0.001 in all comparisons).

orange) and a blue component found in Europe, which reproduces previous results (Mendizabal et al. 2012; Moorjani et al. 2013).

A set of metrics were calculated to describe the genetic diversity in Roma compared with non-Roma. First, the number of segregating sites and private variants in Roma is lower than in non-Roma (supplementary table 2, Supplementary Material online). A depletion of rare variants is further pointed in the unfolded site-frequency spectrum (SFS) and neutrality statistics from synonymous variants (fig. 1*B* and *C*): θ_w is significantly lower in Roma and π_{var} is significantly higher. In addition, θ_π is similar across populations since it assigns more weight to variants at intermediate frequencies, but the Tajima's *D* value of Roma is less negative than that of non-Roma (supplementary Fig. 8, Supplementary Material online).

We further examined the frequency distribution of different functional categories of coding variants. The number of missense derived alleles is similar across Roma and non-Roma populations: 41.14% of the total number of derived alleles are missense in Roma (41.17% in Iberian Population in Spain [IBS], 41.18% in Toscani in Italia [TSI], 41.13% in Punjabi from Lahore, Pakistan [PJL], 41.05% in Gujarati Indian from Houston, Texas [GIH], and 41.02% in Indian Telegu from the UK [ITU]). Missense variants were then grouped in two frequency bins: low-frequency (singletons and doubletons) and common (tripletons or more), stratified in different categories of Genomic Evolutionary Rate Profiling (GERP) (fig. 2), PolyPhen (supplementary fig. 9*A*, Supplementary Material online), and CADD scores (supplementary fig. 9*B*, Supplementary Material online). Roma have significantly more common deleterious variants than non-Roma populations in all functional classifications, especially for the slightly and moderately deleterious categories (*P* value <0.001). Common variants in Roma account for 50% of all deleterious variants, whereas for non-Roma populations these account for 35% or even less. Therefore, the SFS of these variants shows that Roma have significantly fewer singletons and more intermediate and fixed derived variants than non-Roma groups (supplementary fig. 10, Supplementary Material online). In all groups, as expected, the more deleterious the variants, the rarer they are (supplementary fig. 10, Supplementary Material online): the SFS of the most

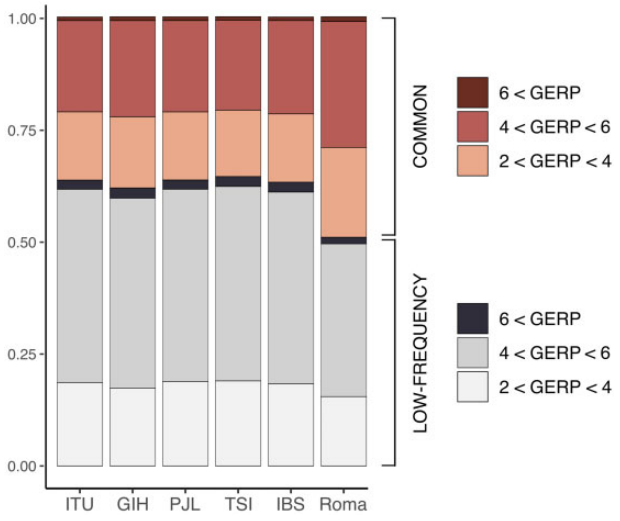

FIG. 2. Proportion of missense variants from each GERP category in each frequency bin (low-frequency, common) for each population. Low-frequency: singletons and doubletons; common: tripletons and above.

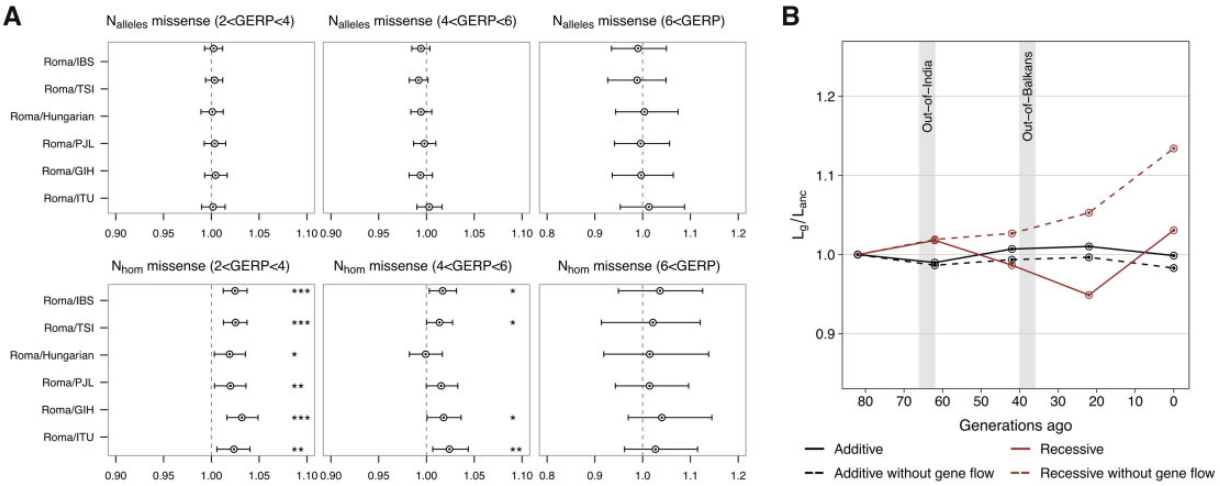

Fig. 3. Deleterious load comparisons and trajectory estimations. (*A*) Mutational load proxy ($N_{alleles}$ and N_{hom}) ratios between Roma and non-Roma for missense variants in each deleterious GERP category. Point estimates and 95% CIs are shown. *P value <0.05; **P value <0.01; ***P value <0.001. Only P values <0.001 are considered significant to account for multiple testing errors. (*B*) Relative mutational load (L_g/L_{ANC}) in the Roma in each sampled generation for each simulated model. L_{ANC}: load in the ancestral population (proto-Roma 20 generations before the "Out-of-India" event). "Out-of-India" and "Out-of-Balkans" represent the two simulated bottlenecks at 63 and 38 generations ago, respectively.

deleterious variants (i.e., GERP >6, probably damaging PolyPhen category, CADD >30) exhibits significantly higher proportions of low-frequency variants (e.g., singletons) than the neutral categories (i.e., $-2 <$ GERP < 2, Benign Polyphen group, $10 <$ CADD) (supplementary fig. 10, Supplementary Material online) within populations, consistent with purifying selection acting on deleterious mutations. This process happens, however, at the same rate for Roma and non-Roma, since the difference between neutral and deleterious categories is not significant among populations, suggesting that Roma experienced higher genetic drift rather than reduced purifying selection. Regarding loss-of-function (LOF) variants, the same trend is observed, although it is not statistically supported due to the low number of high-confidence LOF called in our set of variants and large variation in their distributions (supplementary fig. 11, Supplementary Material online).

Mutational Load Changes through Time with Minor Present-Day Differences

We next explored the present-day mutational load in the Roma compared with non-Roma populations. To approximate the additive and recessive mutational loads, the number of derived alleles per individual ($N_{alleles}$) and the number of derived homozygotes per individual (N_{hom}) were used as proxies, respectively (Lohmueller 2014) (fig. 3A and supplementary figs. 12–14, Supplementary Material online). Roma show the same $N_{alleles}$ compared with non-Roma for all categories (GERP, PolyPhen, and CADD); however, they show a discrete but significant increase in N_{hom} in the slightly deleterious categories ($2 <$ GERP < 4; $20 <$ CADD < 30) (fig. 3A and supplementary figs. 13A and 14A, Supplementary Material online). We applied the same analysis to non-CpG sites to avoid the bias produced by their hypermutability, and we found no differences in mutational load between Roma and non-Roma populations (supplementary figs. 12, 13B, and 14B, Supplementary Material online). In addition, the $R_{X/Y}$ statistics do not show statistical differences between Roma and non-Roma (table 1). We also tested the relationship between mutational load and gene flow, and we found no correlation between the proportion of South Asian ancestry and $N_{alleles}$ and N_{hom} in the Roma samples (supplementary table 3, Supplementary Material online).

To study the temporal trajectory of mutational load through forward simulations, we first estimated the distribution of fitness effects (DFE) of new deleterious mutations, which was then used on the simulations. Based on the estimated demographic parameters (supplementary table 4, Supplementary Material online), the DFE of new mutations was inferred following a gamma distribution with shape and scale estimates (supplementary table 5, Supplementary Material online). The observed and expected SFS from the neutral and selection models (from synonymous and missense variants, respectively) were not significantly different (supplementary fig. 15, Supplementary Material online), showing a good fit of the inferred parameters. The DFE does not differ between Roma and non-Roma groups: all populations show ~25–30% neutral, ~15% weakly deleterious, ~20% moderately deleterious, and ~35–40% strongly deleterious mutations (supplementary fig. 16, Supplementary Material online).

We next performed forward simulations under the previously described Roma demographic model (Mendizabal et al. 2012). This model includes two bottlenecks ("Out-of-India" with 47% of population reduction at 63 generations ago; "Out-of-Balkans" with 70% of reduction at 38 generations ago), with 2.2% gene flow from the Middle East during 13 generations and 5% gene flow from non-Roma Europeans during 38 generations (Mendizabal et al. 2012). The effects of the bottlenecks and of non-Roma to

Table 1. $R_{X/Y}$ ratios between Roma and non-Roma populations for missense variants in each deleterious GERP category normalized by synonymous variants.

$R_{X/Y}$	2 < GERP < 4	4 < GERP < 6	6 < GERP
Roma-IBS	**0.986 (0.954–1.0172)**	**1.032 (0.965–1.107)**	**1.152 (−4.333 to 3.917)**
Roma-TSI	**0.989 (0.957–1.018)**	**1.053 (0.983–1.132)**	**1.225 (−5.176 to 4.417)**
Roma-Hungarian	**0.993 (0.954–1.030)**	**1.033 (0.951–1.114)**	**0.946 (−3.918 to 4.207)**
Roma-PJL	**0.987 (0.945–1.028)**	**1.013 (0.922–1.112)**	**1.058 (−2.953 to 4.287)**
Roma-GIH	**0.973 (0.930–1.014)**	**1.034 (0.938–1.154)**	**1.048 (−2.842 to 4.934)**
Roma-ITU	**0.984 (0.938–1.028)**	**0.970 (0.861–1.074)**	**0.833 (−2.496 to 3.553)**

NOTE.—Point estimates and 95% CIs are shown.

Table 2. Correlations (Spearman's ρ) between the global proportion of South Asian ancestry in the Roma population inferred with RFMix and the number/length of ROHs per-individual.

	All ROHs	0.5 < ROHs ≤ 2.5 (Mb)	2.5 < ROHs (Mb)
Number of ROHs	**0.1051**	**−0.0148**	**0.3587****
Total ROH length	**0.3766****	**0.2518***	**0.3563****

*P value <0.05;
**P value <0.01.

Roma gene flow were investigated with four different sets of forward simulations: full model with only additive mutations (additive model); full model with only recessive mutations (recessive model); model without non-Roma to Roma gene flow and only additive mutations (additive model without gene flow); and model without non-Roma to Roma gene flow and only recessive mutations (recessive model without gene flow).

We show that the mutational load of additive mutations is insensitive to the reduction of effective population size (N_e), with or without gene flow, since both additive models have relative mutational load values (L_g/L_{anc}) ~1 throughout all sampled generations (from 60 generations ago to present) (fig. 3*B*). Conversely, the mutational load of recessive mutations appears to be more sensitive to demographic events since both recessive models have relative mutational load values (L_g/L_{anc}) departing from 1. When recessive mutations are simulated under a model with gene flow from non-Roma to Roma, mutational load increases slightly ($L_g/L_{anc} = 1.018$) after the first bottleneck ("Out-of-India"), but it decreases as soon as gene flow starts acting. When recessive mutations are simulated under a model without gene flow, mutational load starts to increase ($L_g/L_{anc} = 1.019$) after the first bottleneck ("Out-of-India"), rising at a slightly higher rate after the second bottleneck ("Out-of-Balkans"). Interestingly, this accumulation of mutational load in the latter model continues to increase without returning to equilibrium: the simulated population has suffered two recent bottlenecks without recovery. At the present day (0 generations ago), the relative mutational load values for additive models without and with gene flow are stabilized at $L_g/L_{anc} = 0.983$ and $L_g/L_{anc} = 0.999$, respectively; and for recessive models without and with gene flow, they reach $L_g/L_{anc} = 1.134$ and $L_g/L_{anc} = 1.031$, respectively. These simulated values with gene flow are in agreement with the observed load estimations: additive proxy ($N_{alleles}$) is centered ~1 and recessive proxy (N_{hom}) is found within 1 and 1.05 (fig. 3*A*).

Accumulation of Deleterious Mutations in Ancestry-Specific ROHs

As previously suggested, ROHs are highly sensitive to demographic events (Szpiech et al. 2013). Thus, we tested whether ROHs are enriched for deleterious variants in the Roma population. The ratio of deleterious/synonymous N_{hom} is higher inside than outside ROHs, especially in ROHs >2.5 Mb (supplementary fig. 17, Supplementary Material online). In Roma, as in other populations (Szpiech et al. 2013; Pemberton and Szpiech 2018), the rate at which deleterious homozygotes increase inside ROHs is higher than the decrease outside ROHs. The increase in homozygotes in ROHs is higher for deleterious than for synonymous, especially in long ROHs (>2.5 Mb). And, in fact, these long ROHs (>2.5 Mb) are highly and significantly correlated with the Roma inbreeding coefficient (see supplementary note 2, Supplementary Material online, supplementary figs. 18–20, Supplementary Material online, and supplementary table 6, Supplementary Material online for more details).

To test whether this enrichment of deleterious variants in ROHs is ancestry specific, we first examined the relationship between ROHs and ancestry proportions. The proportion of South Asian ancestry per individual is positively and significantly correlated both with the number and length of ROHs (>2.5 Mb) (table 2). Furthermore, the number of SNPs inside ROHs (normalized by base pairs of each ancestry) in South Asian regions is higher than in European regions (25.6% vs. 13.72%) (supplementary fig. 21, Supplementary Material online). Thus, South Asian ancestry in Roma is related with more ROHs. We then focused on the relationship between deleterious alleles and ancestry-specific ROHs (>2.5 Mb). In both European and South Asian regions, the ratio of deleterious/synonymous variation is higher inside than outside ROHs, although the statistical significance is higher in PolyPhen and CADD than in GERP comparisons (supplementary fig. 22, Supplementary Material online). In both South Asian and European segments, the fraction of deleterious and synonymous N_{hom} in ROHs increases linearly with the total ROH

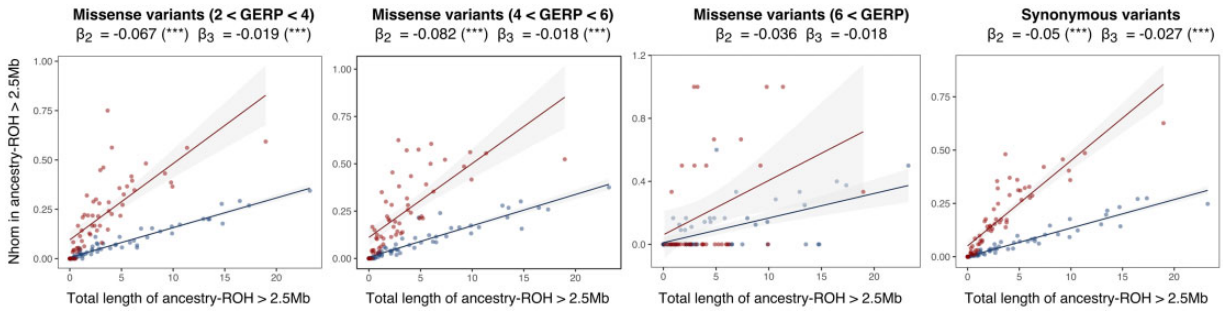

FIG. 4. Fraction of N_{hom} in ancestry-specific ROHs versus the total length of ancestry-specific ROHs per individual. South Asian ROHs in red, and European ROHs in blue. The first three panels show a deleterious GERP category each and the last panel shows synonymous variants. β_2 and β_3 show intercept and slope differences between regressions. *P value <0.05; **P value <0.01; ***P value <0.001.

length per individual. However, an ancestry-specific pattern is observed: only in European segments (though not in South Asian), the rate at which deleterious N_{hom} in ROHs increase is higher than the rate of synonymous increase (test applied following eq. 10 in Szpiech et al. 2013) for CADD and PolyPhen comparisons (supplementary figs. 23 and 24, Supplementary Material online). Moreover, when comparing directly European and South Asian ROHs, two additional patterns appear. First, the overall proportion of deleterious N_{hom} in South Asian ROHs is higher than in European ROHs (significantly different intercept β_2) (fig. 4 and supplementary fig. 25, Supplementary Material online). Second, the rate at which deleterious and synonymous N_{hom} in South Asian ROHs increase is higher than in European ROHs (significantly different slope β_3), except for the most deleterious categories (fig. 4 and supplementary fig. 25, Supplementary Material online).

These results point to an ancestry-specific pattern of accumulation of deleterious homozygotes in ROHs. Particularly, they suggest that South Asian ancestry regions in the Roma genomes contain more ROHs and, in turn, these ROHs accumulate more deleterious and synonymous homozygotes than European ROHs.

Selection Signals in Roma Mainly Derive from Preadmixture Adaptation

To explore the consequences of the Roma demographic history on their events of positive selection, we specifically focused on detecting post- and preadmixture adaptation events, using the population branch statistic (PBS), integrated Haplotype Score (iHS), and Cross Population Extended Haplotype Homozygosity (XP-EHH) tests. We first observe that candidates for positive selection in Roma are found in genes with functions primarily related to metabolic and cardiovascular traits, as well as immunity and xenobiotic response (supplementary fig. 26, Supplementary Material online and supplementary table 7, Supplementary Material online). An overrepresentation analysis of each selection test of the top 1% genes reports significant enrichment in xenobiotic detoxification processes (e.g., cellular detoxification of nitrogen compound, glutathione transferase activity, drug metabolism) when comparing Roma against South Asians (Bonferroni-corrected P values below 0.05, supplementary table 8, Supplementary Material online). However, no candidate region observed in Roma shows a local ancestry deviation higher than 4.42 standard deviations (SD) (P value $<10^{-5}$) (supplementary fig. 27, Supplementary Material online), suggesting that these signals derive from either weak adaptive introgression or preadmixture adaptation in the population sources (Bhatia et al. 2014).

Candidates of positive selection with potential metabolic and cardiovascular implications are commonly detected when comparing Roma against Europeans. Among these signals, *DOK5* (chr20: 52,813,832–53,454,024), a gene involved in lipid and insulin metabolism (Cai et al. 2003), shows extreme values in PBS and XP-EHH tests (supplementary table 7, Supplementary Material online, fig. 5*A*, and supplementary fig. 28*A*, Supplementary Material online). Several genome-wide associations with metabolic phenotypes are found within this region such as body mass index or childhood obesity, among others. In addition, several expression quantitative trait loci (eQTLs) are described within this region that change the expression of *DOK5* (or other metabolism-related genes, such as *CYP24A1*) in specific tissues (adipose tissue, adrenal gland, thyroid, among others). The same selection signal is detected when comparing South Asian against European populations (fig. 5*A*), with this gene having been previously identified to be under positive selection in India (Metspalu et al. 2011). These results suggest preadmixture selection in the South Asian source that Roma maintained due to drift or weaker positive selection after admixture. Other signals of positive selection include *PCK1* (gluconeogenesis regulation) (She et al. 2000) and *DAGLB* (linked to cardiovascular traits) (Han et al. 2017) genes (supplementary table 7, Supplementary Material online and supplementary fig. 28*B* and *C*, Supplementary Material online).

Signals related to immunity and xenobiotic response are among the selection candidates identified when comparing Roma against South Asians. The *LAMA3* gene (chr 18: 21,276,048–21,740,878) shows the highest values in PBS and XP-EHH tests and genome-wide associations with immunoglobulins and white blood cell traits (supplementary table 7, Supplementary Material online, fig. 5*B*, and supplementary fig. 28*D*, Supplementary Material online). With respect to gene regulation, the change in the expression of *LAMA3* (and immunity-related genes, such as *HRH4*, *CABLES1*, and *OSBPL1A*) in tissues, such as pancreas, thyroid, and esophagus mucosa, might be the result of multiple eQTLs found within

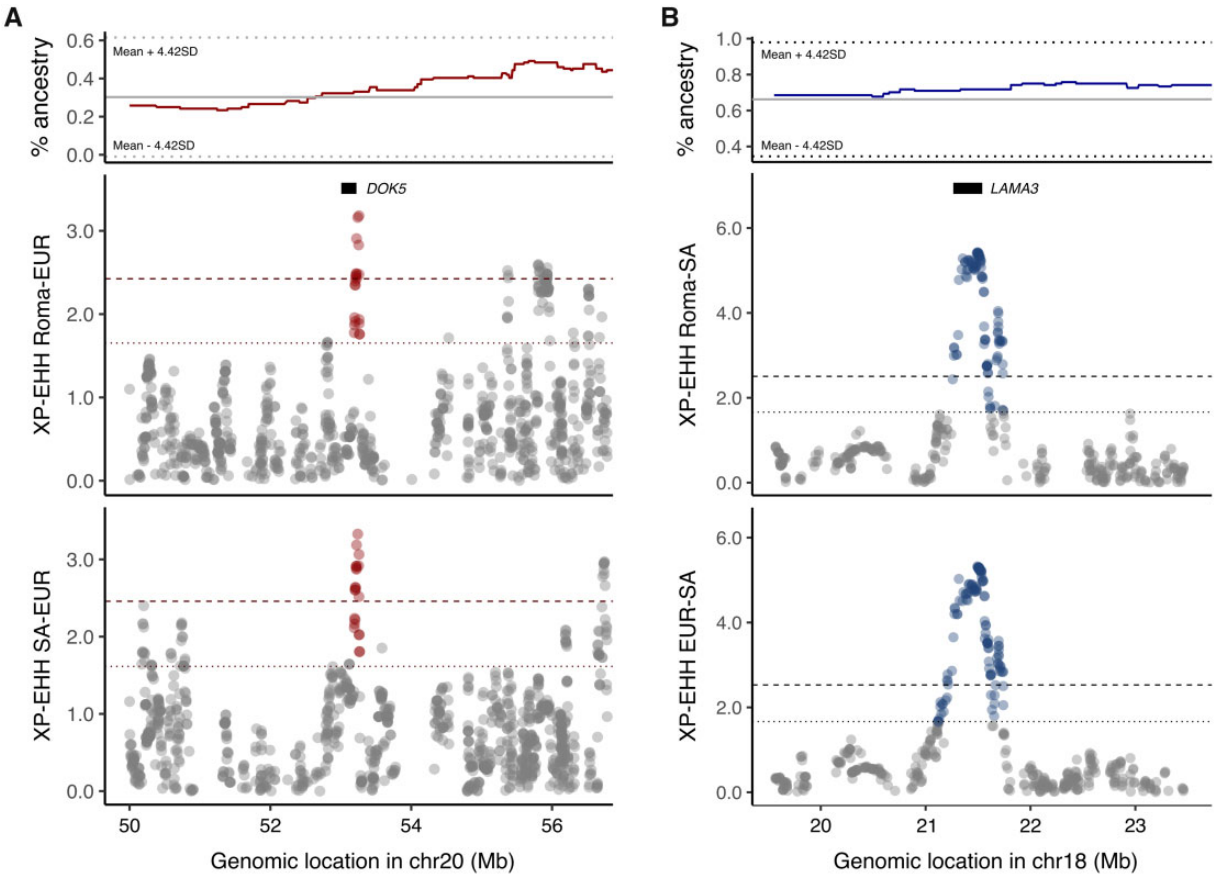

FIG. 5. Selection tests results (XP-EHH) and mean local ancestry in two candidate regions. (*A*) Results for chromosome 20: 50,000,000–56,000,000. Top panel shows South Asian (dark red) ancestry (mean and 4.42 standard deviations in solid and dotted lines, respectively). Genomic location of *DOK5* gene is shown. Middle and bottom panels show XP-EHH analysis comparing Roma against Europe and South Asia against Europe (top 1% and 5% are shown with dashed lines). The region within chr20: 52,813,832–53,454,024 is highlighted in red. (*B*) Results for chromosome 18: 20,000,000–23,000,000. Top panel shows European (blue) ancestry (mean and 4.42 standard deviations in solid and dotted lines, respectively). The genomic location of *LAMA3* gene is shown. Middle and bottom panels show XP-EHH analysis comparing Roma against South Asia and Europe against South Asia (top 1% and 5% are shown with dashed lines). The region within chr18: 21,276,048–21,740,878 is highlighted in blue.

this region. The European population shows the same signal when tested against South Asia (fig. 5*B*), pointing to a pre-admixture selection event in the European source, rather than to postadmixture adaptation in the Roma. Alternatively, the signal in this region could also point to an original selection in South Asia that, in the Roma, has been further selected or drifted to high frequencies compared with present-day South Asians, although selection in Europeans is the most plausible hypothesis. Additional selection candidates show genome-wide associations related to immune system functions, drug response, and toxic substance binding (e.g., *DOCK8* and *SLC6A5* genes) (supplementary table 7, Supplementary Material online and supplementary fig. 28*E* and *F*, Supplementary Material online).

Other candidate regions identified to be under selection are detected in the Roma genomes. Particularly, *MYO5A*, *SLC45A2*, and *APBA2* genes show selection signals specially when comparing Roma against South Asians (supplementary fig. 28*G*–*I*, Supplementary Material online). All three genes are involved in skin pigmentation and have been targeted to be under selection in European populations (Lamason et al. 2005; McEvoy et al. 2006; Voight et al. 2006; Deng and Xu 2018).

Discussion

In the present study, we have shown that the complex demographic history of the Roma has had a multifaceted impact on their genomes. Particularly in this population, two balancing demographic forces are playing a major role. Multiple founder effects driven by political and social persecution against Roma (Fraser 1992) have led to a reduced effective population size and increased genetic drift, whereas extensive admixture throughout their diaspora has resulted in ancestry-specific genetic patterns and decreased their deleterious load.

A clear evidence of this impact is the reduced genetic diversity compared with non-Roma populations (both European and South Asian groups). The depletion of rare alleles and increased high-frequency deleterious variants can be explained by the population decline during the "Out-of-India" bottleneck and subsequent founder events in Europe. This observation is also consistent with what has been

previously reported in other populations with reductions in the effective population size due to recent bottlenecks (20 generations ago in French-Canadians; Casals et al. 2013), or long lasting bottlenecks (20,000 years in the Greenlandic Inuit; Pedersen et al. 2017).

Regarding present-day mutational load, we observe a discrete but significant increase in recessive load (N_{hom}). However, this proxy assumes that all mutations in the genome are recessive, whereas $N_{alleles}$ assumes semidominance (Lohmueller et al. 2008). Some studies suggest that most deleterious variants in the human genome have an additive dominance coefficient, pointing to $N_{alleles}$ as a more reliable proxy to estimate mutational load values in present-day populations (Simons and Sella 2016). $R_{X/Y}$ statistics do not show statistical differences between Roma and non-Roma, in agreement with $N_{alleles}$, further pointing to a similar selection effectiveness between these populations. In addition, we show that the temporal trajectories of mutational load for additive variants are insensitive to population decline and gene flow. This observation is consistent with previous studies showing that mutational load for additive variants does not increase even with changes in N_e, since mutation-selection balance holds and selection remains strong (Simons et al. 2014). Recessive mutations, on the contrary, are more sensitive to changes in N_e and to be drifted to high frequencies: when $N_e s \ll 1$, random genetic drift has more strength than purifying selection (Lohmueller et al. 2008; Simons et al. 2014). However, in some populations, when there is a reduction in N_e followed by gene flow from a larger population, the increase in the recessive load is partially balanced (Lopez et al. 2018), since admixture increases N_e and $N_e s \ll 1$ no longer holds. This trend is also observed in our results: the simulated model without non-Roma to Roma gene flow shows that the recessive load trajectory increases through time after the "Out-of-India" bottleneck, whereas the presence of gene flow attenuates this effect. At the present day, this increase in recessive load, however, only reaches 1.134 relative to the ancestral load in the absence of gene flow and 1.031 with gene flow (the latter corresponding to both the value found in simulations at 0 generations ago and N_{hom} proxy). The small impact on load in this population could be explained by three different factors: extensive gene flow (65% of the Roma genomes have West Eurasian ancestry acquired during the last 700 years) (Font-Porterias et al. 2019), which balances the accumulation of deleterious alleles; a moderate size of the bottleneck (N_e reduction is estimated to be ~47%) (Mendizabal et al. 2012), where genetic drift was increased but not being strong enough; and a short and rapid "Out-of-India" event, where most deleterious mutations did not have enough time to reach fixation.

As we have shown, the study of ROHs can offer new insights into the impact of the demographic history. Particularly, recent inbreeding leading to long ROHs in Roma is responsible for the increase in homozygous deleterious variants, as previously suggested for other populations (Szpiech et al. 2013). In addition, our results point to an ancestry-specific pattern in South Asian ROHs: both deleterious and synonymous homozygous variants accumulate at the same rate in ROHs, as opposite to a higher accumulation of deleterious variants as would be expected (Szpiech et al. 2013; Pemberton and Szpiech 2018). This observation can be explained by an extremely low genetic diversity in the South Asian ancestral source together with the subsequent effects of the "Out-of-India" bottleneck, or due to postadmixture parental relatedness of these ancestry-specific tracks, due to the absence of new gene flow from South Asian sources after the Out-of-India. However, we note that these results could also be driven by a technical artifact since South Asian regions are less abundant than European regions in the Roma population (35% vs. 65% of admixture).

Several cases of positive selection after introgression with archaic hominins have been identified (Kuhlwilm et al. 2016; Enard et al. 2018), whereas for modern humans, postadmixture selection is more difficult to infer (Bhatia et al. 2014; Patin et al. 2017; Secolin et al. 2019). If selection occurred after admixture, one would expect a significant deviation in local ancestry proportions, where a minimum of 4.42 SD should be applied, which corresponds to a P value of $<10^{-5}$ (Bhatia et al. 2014). None of the candidate regions under selection in Roma show a local ancestry deviation higher than 2.5 SD of the mean. The absence of strong local ancestry deviations suggests that postadmixture selection has not had enough time to leave noticeable signals or that selection acting in Roma is weak. Therefore, the observed selection signals most likely represent preadmixture events in Roma source populations. Particularly, the prevalence of metabolic and cardiovascular diseases in the Roma (Vozarova De Courten et al. 2003; Živković et al. 2010) can be the result of an evolutionary mismatch: past positive adaptation in South Asian populations that has become maladaptive in present-day environments and lifestyles (Neel 1962). Selection signals involved in immunity and xenobiotic response, on the other hand, appear to derive from preadmixture adaptation in the European ancestral sources, which Roma could have maintained through drift or even weaker selection due to new pathogen exposure during the changing environment of their diaspora. However, we caution that the approaches based on local ancestry deviation, besides leading to false positives, could lead to false negatives (e.g., due to systematic biases in the local ancestry inference [LAI], genetic drift or small number of generations since admixture) and, as a result it might challenge the detection of regions under weaker or polygenic selection (Seldin et al. 2011; Bhatia et al. 2014; Zhang et al. 2020).

The Roma have been traditionally thought as an isolated and small group. Indeed, they have experienced multiple founder effects that have reduced their genetic diversity, although extensive gene flow has counteracted the increase in mutational load with traceable ancestry-specific patterns in ROHs and with limited evidence of postadmixture selection. Here, we have focused on the Iberian Roma, and due to the heterogeneity found among European Roma (Mendizabal et al. 2012; Font-Porterias et al. 2019), the study of other Roma groups might lead to slightly different results.

The present study is an example of the relevance of accounting for the ancestry components in admixed populations since ancestry-specific patterns can reveal different demographic processes that would otherwise remain hidden. However, we caution that, when working with specific ethnic

groups, we should be aware that ethnicity is not only defined by genetic ancestry since cultural identity is also a major factor. As a concluding remark, we would like to note the potential biomedical implications of the present study. An increased genetic disease prevalence has always been suggested in the Roma, which might be the case for specific disorders where the causal mutation has drifted to high frequencies: for example, galactokinase deficiency (Kalaydjieva et al. 1999), primary congenital glaucoma (Plásilová et al. 1999), and congenital myasthenia (Abicht et al. 1999). However, considering the complexity of our results, a different spectrum of genetic disorders is an interesting hypothesis that needs to be explored, which could point to a different distribution of genetic disease risks: some disease-associated mutations might have accumulated in Roma, whereas some others might be absent in this population.

Materials and Methods

Samples and Sequencing

We sequenced new WES of 89 Iberian Roma samples at 50× from saliva samples, using Agilent SureSelect Human All Exon V6 capture kit. DNA donors and their four grandparents self-identify as Roma from the Iberian Peninsula. Written informed consent was obtained for the participants under the corresponding IRB approvals (CEIC-Parc de Salut Mar 2016/6723/I and 2019/8900/I). Some of the Roma volunteers were collected within the project "El Camí" in collaboration with the Federació d'Associacions Gitanes de Catalunya. In addition, as non-Roma reference populations, we included 1,000 G Exomes (Auton et al. 2015) from IBS, TSI, PJL, ITU, and GIH; and 20 Hungarian WES (Tombácz et al. 2017) (see supplementary note 1, Supplementary Material online for more details). The set of non-Roma populations was chosen based on the population sources of the Roma admixture (Font-Porterias et al. 2019) with available high-coverage exomes. We also genotyped the Iberian Roma samples with Affymetrix Axiom Genome-Wide Human Origins 1 array. Genotype calling was performed with Axiom Analysis Suite 4.0 software with default threshold settings. Genotyping errors were filtered out with PLINK/1.9b (Purcell et al. 2007) using the following quality control filters: SNP missingness of 5%, individual missingness of 10%, SNPs failing Hardy–Weinberg exact test with a P value of 10^{-5}, and minor allele frequency (MAF) threshold of 0.01. After filtering, the genome-wide array data set contains 486,009 SNPs in a subset of 62 of the WES Iberian Roma samples. The Iberian Roma samples were merged with IBS, TSI, PJL, GIH, and ITU from 1,000 G (1000 Genomes Project Consortium 2012), keeping 474,632 genome-wide SNPs and 487 samples (supplementary table 1, Supplementary Material online).

Sequence Preprocessing

The WES preprocessing was performed following the GATK Best Practices recommendations (Van der Auwera et al. 2013). Reads were mapped to the human reference GRCh37 with bwa 0.7.15 (Li and Durbin 2009). Then, duplicates were marked with Picard 2.8.3 (http://broadinstitute.github.io/picard, last accessed March 24, 2021) and indel realignment and base quality score recalibration were performed with GATK 3.7 (McKenna et al. 2010). Variant calling steps were performed with HaplotypeCaller, GenotypeGVCFs, and VariantRecalibrator from GATK 3.7 (McKenna et al. 2010) (see supplementary note 1, Supplementary Material online for more details). We removed indels, nonautosomal chromosomes, and nonbiallelic and fixed sites. Sequencing errors were filtered out with VCFtools 0.1.14 (Danecek et al. 2011) using the following filters: depth of coverage (DP) $<5\times$, genotype quality <20, missingness $>5\%$, and deviations from Hardy–Weinberg equilibrium with P value $<10^{-3}$. Only high-quality individuals were included in the analysis: DP $>40\times$, 85% of the BAM positions covered at $5\times$ minimum, missingness $<5\%$, heterozygosity $<$ mean $+$ 4 SD, and relatedness between pairs of samples lower than second degree (KING; Manichaikul et al. 2010). After sample and variant filtering, our final data set contains 410,225 variants and 527 individuals (supplementary table 1, Supplementary Material online). In those analyses involving per-individual genotypes and allele count analyses, no missing data were allowed (257,452 sites) (see supplementary note 1, Supplementary Material online for more details). The mean genotype concordance between genome-wide array and WES is 99.81 $\pm$ 0.35% for the 6,828 common SNPs in both data sets. We assigned the ancestral state of each variant based on the six primate EPO (Enredo, Pecan, and Ortheus) multialignment Ensembl Compara v59. The genome-wide array data set (474,632 variants) and the WES data set (410,225 variants) were merged, resulting in a data set of 878,162 variants and 487 samples. In supplementary note 3, Supplementary Material online, we assess the quality and potential sequencing biases in the new Roma WES (supplementary figs. 2, 3, and 29, Supplementary Material online and supplementary table 9, Supplementary Material online).

Variant Annotation

The Variant Effect Predictor (VEP) tool from Ensembl was used to functionally annotate the derived variants in WES data set (McLaren et al. 2016). To avoid exploiting a single type of information, different deleterious prediction scores were taken into account: PolyPhen-2 (Adzhubei et al. 2010), GERP (Davydov et al. 2010), and CADD (Rentzsch et al. 2019). Some variants are annotated as both synonymous and missense since they are in a region with two overlapping genes; in these cases, both annotations were kept. Pooling all damaging variants together can mask the results (e.g., impossibility to find specific patterns concerning specific variant groups). Thus, we classified missense variants into four GERP RS groups: neutral ($-2 <$ GERP < 2), slightly deleterious ($2 <$ GERP < 4), moderate ($4 <$ GERP < 6), and extremely deleterious (GERP > 6) (Henn et al. 2016). For PolyPhen-2, three categories were used: benign (<0.446), possibly damaging (0.446–0.908), and probably damaging (>0.908) (McLaren et al. 2016). For CADD, since values are Phred scaled, variants were split in score changes of 10 into four categories: <10, 10–20, 20–30, and >30. Finally, we also annotated high-confidence LOF variants using the LOF Transcript Effect

Estimator (LOFTEE) VEP plugin (available at https://github.com/konradjk/loftee, last accessed March 24, 2021).

Population Structure Analysis

PCA and ADMIXTURE were performed using the merged data set of genome-wide array and WES variants. Linkage disequilibrium pruning was applied with PLINK/1.9b (Purcell et al. 2007) (window size of 200 SNPs, 25 SNPs shift at each step, and r^2 threshold of 0.5) and MAF >1%, keeping 405,814 variants. PCA was performed with the SmartPCA program implemented in the EIGENSOFT 4.2 package (Patterson et al. 2006), and ADMIXTURE (Alexander et al. 2009) was run 10 independent times with different random seeds for ancestral components $k = 2$–5. Pong (Behr et al. 2016) was used to identify modal ADMIXTURE results. In addition, ADMIXTURE was run, independently, for $k = 2$ for the genome-wide data set (202,724 variants) and the WES with MAF > 1% data set (42,381 variants). As shown in supplementary figure 30, Supplementary Material online, ADMIXTURE analysis with genome-wide array data, when compared with WES data, estimates a higher proportion of the minor component in all populations (Maróti et al. 2018), although it is specially detected in Roma samples (26.15 ± 6.57% and 21.48 ± 6.39%: mean dark-red component with genome-wide array and WES data, respectively).

Genetic Diversity Metrics

To assess the neutral genetic diversity, we used synonymous sites in the WES to estimate pairwise nucleotide diversity (θ_π), Watterson's θ (θ_w), and Tajima's D from the SFS applying the previously defined formulas (Kousathanas et al. 2011). We also computed the pairwise nucleotide diversity only for variant sites (π_{var}), as previously described (Pedersen et al. 2017). We performed 1,000 bootstrap resamples with replacement of the variants divided into 1,000 blocks (Simons and Sella 2016) to obtain 95% confidence intervals (CIs) and P values to compare these diversity metrics among populations. All genetic diversity metrics were calculated using R base software (R Core Team 2019).

Frequency Distribution of Coding Variants

We calculated the SFS for each population stratifying the WES variants in different categories: synonymous, missense, GERP groups in missense variants, PolyPhen groups in missense variants and CADD groups in missense variants. We also grouped the variants into low-frequency (singletons and doubletons) and common (tripletons or more) classes. The same number of individuals was considered for this analysis (70 individuals per population). To obtain 95% CI and P values for Roma and non-Roma comparisons, we performed 1,000 bootstrap resamples with replacement of the variants divided into 1,000 blocks (Simons and Sella 2016). We tested for statistically significant differences between Roma and non-Roma in the proportion of deleterious variants for common and low-frequency categories and for each "number of derived alleles" group in the SFS. We also tested whether the difference between Roma and non-Roma is higher as the variants are more deleterious (for GERP, PolyPhen, and CADD categories). For GERP, we thus tested if, for $1 \leq i \leq 2n$:

$$[2,4]_{i(\text{non}-\text{Roma})} - [-2,2]_{i(\text{non}-\text{Roma})} \neq [2,4]_{i(\text{Roma})} - [-2,2]_{i(\text{Roma})}$$

$$[4,6]_{i(\text{non}-\text{Roma})} - [2,4]_{i(\text{non}-\text{Roma})} \neq [4,6]_{i(\text{Roma})} - [2,4]_{i(\text{Roma})}$$

$$[>6]_{i(\text{non}-\text{Roma})} - [4,6]_{i(\text{non}-\text{Roma})} \neq [>6]_{i(\text{Roma})} - [4,6]_{i(\text{Roma})}$$

Mutational Load Proxies

Two summary statistics were used as proxies for mutational load: number of derived alleles per individual ($N_{alleles}$) and number of derived homozygotes per individual (N_{hom}). $N_{alleles}$ and N_{hom} were calculated by stratifying variants in different categories: synonymous, missense, and missense variants grouped in GERP, PolyPhen, and CADD scores. In addition, we calculated the $R_{X/Y}$ ratio (Do et al. 2015) between each Roma and non-Roma population in each GERP score category normalized by the synonymous sites. We performed 1,000 bootstrap resamples with replacement of the variants divided into 1,000 blocks (Simons and Sella 2016) to obtain 95% CI and P values to compare these proxies ($N_{alleles}$, N_{hom}, and $R_{X/Y}$) among populations. We also tested whether the present-day Roma mutational load is correlated with the South Asian ancestry, estimated with RFMIX v1.5.4 (Maples et al. 2013) (see below).

DFE of New Deleterious Mutations

The DFEs of Roma and non-Roma populations were inferred using ∂a∂i/Fit∂a∂i (Gutenkunst et al. 2009; Kim et al. 2017). We first fitted a three-epoch demographic model using the unfolded SFS for synonymous mutations (as proxies for neutral variation), accounting for ancestral misidentification. Then, conditional on the demographic parameter estimates, the DFE of missense mutations was inferred, assuming a gamma distribution. For both the demographic and DFE parameters, 95% CIs were estimated with 100 bootstraps by site. ∂a∂i/Fit∂a∂i infers the mean $E(N_e s)$; thus, we estimated $E(s) = E(N_e s)/N_w$, where N_w is the weighted effective population size along time (Lopez et al. 2018) from the N_{ANC} calculated from $\theta_S = 4N_{ANC}\mu L_S$ (Kim et al. 2017), where θ_S is the population-scaled synonymous mutation rate and $\mu = 1.5 \times 10^{-8}$ (Ségurel et al. 2014). L_S derives from $L = L_{NS} + L_S$, where L is the number of bases from which variants were called and assuming a ratio of synonymous to nonsynonymous sites $L_S / L_{NS} = 2.31$ (Huber et al. 2017) (supplementary table 10, Supplementary Material online).

Temporal Trajectories of Mutational Load

We performed a set of forward simulations using SLiM 3 (Haller and Messer 2019) using a previously published demographic model that includes Iberian Roma (Mendizabal et al. 2012). The mutation rate was set to 1.36×10^{-8} per base position per generation, recombination rate to 10^{-8} per base per generation, and a burn-in phase of $8N$ generations was applied. The simulated genome structure includes: 20 unlinked chromosomes with 1,000 genes separated by neutral noncoding regions (50,000 base long); genes divided into 8 exon–intron pairs (100- and 5,000-base long, respectively); introns are assumed to be neutral; and exons are based in three-base pair codons, with only the first two positions

under selection (accepting deleterious mutations) (Lopez et al. 2018). Deleterious mutations are subject to a DFE with a gamma distribution with mean $E(s)$ −0.025 and shape 0.18, corresponding to the fitted Roma DFE as Roma and non-Roma DFEs were not statistically different. To reduce the computational time of the simulations, we performed a rescaling of the following parameters: population size and generation time were decreased by ten, whereas mutation and recombination rate, selection coefficient, and migration rate were multiplied by ten to keep population-genetic parameters constant. We periodically sampled nonfixed mutations in proto-Roma or Roma populations and calculated, at each sampled generation, the mutational load as $L = 1 - \exp(\sum_i l_i)$, where i represents each mutation, $l = s \times (2hq + (1 - 2h) \times q^2)$, s the selection coefficient, h the dominance coefficient, and q the frequency of the mutation.

Local Ancestry Inference

The merged data set of genome-wide array and WES variants with MAF > 1% (405,814 variants) was phased with SHAPEIT (O'Connell et al. 2014), using the population-averaged genetic map from the HapMap phase II (International HapMap Consortium 2003) and the 1,000 G data set as a reference panel (1000 Genomes Project Consortium 2012). Local ancestry was inferred with RFMix v1.5.4 (Maples et al. 2013), using two reference sources: European (IBS and TSI populations) and South Asian (PJL, GIH, and ITU populations) and one expectation–maximization iteration. The unassigned local ancestry regions comprised around 3.66% of the data. The global South Asian proportion inferred with RFMix is highly correlated ($\rho = 0.9573$, P value $< 2.2 \times 10^{-16}$) with the proportion of the cluster component assigned as dark red (mostly prevalent in South Asia) in ADMIXTURE $k = 2$ (supplementary fig. 31, Supplementary Material online). The mean global proportions of LAI were 68.42% European and 31.58% South Asian (±7.01% SD), whereas for ADMIXTURE $k = 2$; they are 75.14% and 24.86%, respectively (±6.55% SD). The higher proportion of European ancestry inferred with ADMIXTURE compared with RFMix is due to the fact that, in the Roma population, allele-frequency methods overestimate the European component (Font-Porterias et al. 2019).

Identification of ROH Segments

ROHs were identified from the merged data set of genome-wide array and WES variants using PLINK/1.9b (Purcell et al. 2007). ROHs with minimum 50 SNPs, 500 kb, and a maximum gap of 100 kb (between a pair of SNPs) were considered (Kirin et al. 2010). ROHs were partitioned in two length categories (0.5–2.5 and >2.5 Mb), representing "short–medium" and "long" ROHs. Short and medium ROHs are pooled together, whereas long ROHs are classified into its own category because the distribution of lengths is shifted to longer ROHs in the Roma: they show a larger number of longer ROHs than non-Roma, as previously shown (Font-Porterias et al. 2019). Derived missense variants within ROHs were stratified in GERP, PolyPhen, and CADD deleterious categories. ROHs and ancestry-specific segments inferred from merged data set with genome-wide and WES variants were matched to the derived variants from the WES data.

Identification of Genomic Regions under Selection

Candidates for positive selection in Roma were identified from three different selection methods: PBS, iHS, and XP-EHH. F_{ST} values per variant were calculated for Roma, European, South Asian, and YRI (outgroup) populations using VCFtools 0.1.14 (Danecek et al. 2011). PBS was then obtained as previously described (Yi et al. 2010): $PBS = (T^{WX} + T^{WY} - T^{XY})/2$; $T^{WY} = -\log(1 - Fst^{WY})$, where Y represents YRI, X either European or South Asian group populations, and W the target population (Roma, Europe, or South Asia). We performed four different PBS tests per variant: 1) Roma against European; 2) Roma against South Asian; 3) European against South Asian; and 4) South Asian against European, using in all tests YRI as an outgroup. iHS (Voight et al. 2006) was calculated for Roma, European, and South Asian populations with selscan v1.2.0.a (Szpiech and Hernandez 2014). Unstandardized iHS and normalization across frequency bins were computed with default parameters. XP-EHH (Sabeti et al. 2007) was calculated for Roma against Europeans, Roma against South Asians, and European against South Asians with selscan v1.2.0.a (Szpiech and Hernandez 2014). Unstandardized XP-EHH and genome-wide normalization were computed with default parameters. For each statistic (PBS, iHS, and XP-EHH), variants with scores above the top 1% were filtered out when there were less than two other variants in the top 1% within 200 kb (Mathieson et al. 2015; Ilardo et al. 2018). We then selected the top ten signals in each analysis and annotated the variants inside the signal (within the selection score decay) using VEP from Ensembl (McLaren et al. 2016). Genome-wide associations from the GWAS catalog and eQTLs within the top regions were identified (GTEx Consortium 2017; Buniello et al. 2019). For PBS and XP-EHH statistics, we performed a gene annotation enrichment analysis with genes within the top 1% selection signals using DAVID 6.8 (Huang et al. 2009). Gene Ontology (Ashburner et al. 2000) and KEGG (Kanehisa and Goto 2000) pathways were used as functional databases, and the genes present in our data set were used as the background gene list.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank all the DNA donors that made this study possible and the invaluable collaboration of the Federació d'Associacions Gitanes de Catalunya (FAGiC). We also thank Mònica Vallés (Universitat Pompeu Fabra) for technical support. This study was supported by the Spanish Ministry of Science, Innovation and Universities (MCIU) and the Agencia Estatal de Investigación (AEI) (grant numbers CGL2016-75389-P, PID2019-106485GB-I00/AEI/10.13039/501100011033) and "Unidad de Excelencia María de

Maeztu" (AEI, CEX2018-000792-M). N.F.-P. was supported by a FPU17/03501 fellowship.

Data Availability

Iberian Roma whole exome sequences and genome-wide array data are deposited at EGA accession number: EGAS00001004599. Scripts are available in https://figshare.com/articles/software/Whole-exomes_Roma/13848506.

References

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 491(7422):56.

Abicht A, Stucka R, Karcagi V, Herczegfalvi A, Horváth R, Mortier W, Schara U, Ramaekers V, Jost W, Brunner J, et al. 1999. A common mutation (epsilon1267delG) in congenital myasthenic patients of Gypsy ethnic origin. *Neurology*. 53(7):1564–1569.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods*. 7(4):248–249.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 19(9):1655–1664.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet*. 25(1):25–29.

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR; 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*. 526(7571):68–74.

Azmanov DN, Dimitrova S, Florez L, Cherninkova S, Draganov D, Morar B, Saat R, Juan M, Arostegui JI, Ganguly S, et al. 2011. *LTBP2* and *CYP1B1* mutations and associated ocular phenotypes in the Roma/Gypsy founder population. *Eur J Hum Genet*. 19(3):326–333.

Behr AA, Liu KZ, Liu-fang G, Nakka P, Ramachandran S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*. 32(18):2817–2823.

Bhatia G, Tandon A, Patterson N, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ, et al. 2014. Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *Am J Hum Genet*. 95(4):437–444.

Bianco E, Laval G, Font-Porterias N, García-Fernandez C, Dobon B, Sabido-Vera R, Stefanovska ES, Kučinskas V, Halyna Makukh HP, Quintana-Murci L, et al. 2020. Recent common origin, reduced population size, and marked admixture have shaped European Roma genomes. *Mol Biol Evol*. 37(11):3175–3187.

Boerger BH. 1984. Proto-Romanes phonology [dissertation].

Bouwer S, Angelicheva D, Chandler D, Seeman P, Tournev I, Kalaydjieva L. 2007. Carrier rates of the ancestral Indian W24X mutation in GJB2 in the general Gypsy population and individual subisolates. *Genet Test*. 11(4):455–458.

Buniello A, MacArthur J, Cerezo M, Harris L, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 47(D1):D1005–D1012.

Cai D, Dhe-Paganon S, Melendez PA, Lee J, Shoelson SE. 2003. Two new substrates in insulin signaling, IRS5/DOK4 and IRS6/DOK5. *J Biol Chem*. 278(28):25323–25330.

Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, Grenier JC, Gbeha E, Hamdan FF, Girard S, et al. 2013. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet*. 9(9):e1003815.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. 2018. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet*. 19(4):220–234.

Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR, et al. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun*. 1(8):131–136.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al.; 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics*. 27(15):2156–2158.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*. 6(12):e1001025.

Deng L, Xu S. 2018. Adaptation of human skin color in various populations. *Hereditas*. 155:1.

Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet*. 47(2):126–131.

Dobon B, ter Horst R, Laayouni H, Mondal M, Bianco E, Comas D, Ioana M, Bosch E, Bertranpetit J, Netea MG. 2020. The shaping of immunological responses through natural selection after the Roma Diaspora. *Sci Rep*. 10(1):1–12.

Enard D, Petrov DA, Enard D, Petrov DA. 2018. Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. *Cell*. 175(2):360–371.e13.

Font-Porterias N, Arauna LR, Poveda A, Bianco E, Rebato E, Prata MJ, Calafell F, Comas D. 2019. European Roma groups show complex West Eurasian admixture footprints and a common South Asian genetic origin. *PLoS Genet*. 15(9):e1008417.

Fraser A. 1992. The gypsies. Oxford: Wiley-Blackwell.

García-Fernández C, Font-Porterias N, Kučinskas V, Sukarova-Stefanovska E, Pamjav H, Makukh H, Dobon B, Bertranpetit J, Netea MG, Calafell F, et al. 2020. Sex-biased patterns shaped the genetic history of Roma. *Sci Rep*. 10(1):14464.

Gravel S. 2016. When is selection effective? *Genetics*. 203(1):451–462.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD, Altshuler DL, et al.; 1000 Genomes Project. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA*. 108(29):11983–11988.

Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, Angelicheva D, Calafell F, Oefner PJ, Shen P, et al. 2001. Origins and divergence of the Roma (gypsies). *Am J Hum Genet*. 69(6):1314–1331.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature*. 550:204–213.

Gusmão A, Gusmao L, Gomes V, Alves C, Calafell F, Amorim A, Prata MJ. 2008. A perspective on the history of the iberian gypsies provided by phylogeographic analysis of Y-chromosome lineages. *Ann Hum Genet*. 72(Pt 2):215–227.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 5(10):e1000695.

Haller BC, Messer PW. 2019. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Mol Biol Evol*. 36(3):632–637.

Han X, Hu Z, Chen J, Huang J, Huang C, Liu F, Gu C, Yang X, Hixson JE, Lu X, et al. 2017. Associations between genetic variants of NADPH oxidase-related genes and blood pressure responses to dietary sodium intervention: the GenSalt study. *Am J Hypertens*. 30(4):427–434.

Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, et al. 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci USA*. 113(4):E440–E449.

Huang D, Sherman B, Lempicki R. 2009. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat Protoc*. 4(1):44–57.

Huber CD, Kim BY, Marsden CD, Lohmueller KE. 2017. Determining the factors driving selective effects of new nonsynonymous mutations. *Proc Natl Acad Sci USA*. 114(17):4465–4470.

Ilardo MA, Moltke I, Korneliussen TS, Cheng J, Stern AJ, Racimo F, de Barros Damgaard P, Sikora M, Seguin-Orlando A, Rasmussen S, et al. 2018. Physiological and genetic adaptations to diving in sea nomads. *Cell*. 173(3):569–580.e15.

International HapMap Consortium. 2003. The international HapMap project. *Nature*. 426(6968):789.

Kaiser VB, Svinti V, Prendergast JG, Chau YY, Campbell A, Patarcic I, Barroso I, Joshi PK, Hastie ND, Miljkovic A, et al.; UK10K. 2015. Homozygous loss-of-function variants in European cosmopolitan and isolate populations. *Hum Mol Genet*. 24(19):5464–5474.

Kalaydjieva L, Perez-Lezaun A, Angelicheva D, Onengut S, Dye D, Bosshard NU, Jordanova A, Savov A, Yanakiev P, Kremensky I, et al. 1999. A founder mutation in the *GK1* gene is responsible for galactokinase deficiency in Roma (Gypsies). *Am J Hum Genet*. 65(5):1299–1307.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 28(1):27–30.

Keinan A, Clark AG. 2011. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*. 336(6042):540–543.

Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, Degenhardt JD, Brisbin A, Sheth V, Chen R, et al. 2012. Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *Am J Hum Genet*. 91(4):660–671.

Kim B, Huber C, Lohmueller K. 2017. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*. 206(1):345–361.

Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. 2010. Genomic runs of homozygosity record population history and consanguinity. *PLoS One*. 5(11):e13996.

Kousathanas A, Oliver F, Halligan DL, Keightley PD. 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol Biol Evol*. 28(3):1183–1191.

Kuhlwilm M, Gronau I, Hubisz M, Filippo C, de Prado-Martinez J, Kircher M, Fu Q, Burbano H, Lalueza-Fox C, Rasilla M, de la, et al. 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*. 530(7591):429–433.

Lamason RL, Mohideen MPK, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, et al. 2005. *SLC24A5*, a putative cation exchanger affects pigmentation in zebrafish and humans. *Science*. 1782–1787.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25(14):1754–1760.

Lohmueller KE. 2014. The distribution of deleterious genetic variation in human populations. *Curr Opin Genet Dev*. 29:139–146.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature*. 451(7181):994–997.

Lopez M, Kousathanas A, Quach H, Harmant C, Mouguiama-Daouda P, Hombert JM, Froment A, Perry GH, Barreiro LB, Verdu P, et al. 2018. The demographic history and mutational load of African hunter-gatherers and farmers. *Nat Ecol Evol*. 2(4):721–730.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 26(22):2867–2873.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet*. 93(2):278–288.

Maróti Z, Boldogkői Z, Tombácz D, Snyder M, Kalmár T. 2018. Evaluation of whole exome sequencing as an alternative to BeadChip and whole genome sequencing in human population genetic analysis. *BMC Genomics*. 19(1):1–13.

Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X, et al.; 1000 Genomes Project. 2011. The functional spectrum of low-frequency coding variation. *Genome Biol*. 12(9):R84.

Martínez-Cruz B, Mendizabal I, Harmant C, de Pablo R, Ioana M, Angelicheva D, Kouvatsi A, Makukh H, Netea MG, Pamjav H, et al. 2016. Origins, admixture and founder lineages in European Roma. *Eur J Hum Genet*. 24(6):937–943.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. 528(7583):499–503.

McEvoy B, Beleza S, Shriver MD. 2006. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. *Hum Mol Genet*. 15(2):176–181.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297–1303.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol*. 17(1):1–14.

Melegh BI, Banfai Z, Hadzsiev K, Miseta A, Melegh B. 2017. Refining the South Asian Origin of the Romani people. *BMC Genet*. 18(1):82–13.

Mendizabal I, Lao O, Marigorta UM, Wollstein A, Gusmão L, Ferak V, Ioana M, Jordanova A, Kaneva R, Kouvatsi A, et al. 2012. Reconstructing the population history of European Romani from genome-wide data. *Curr Biol*. 22(24):2342–2349.

Mendizabal I, Valente C, Gusmão A, Alves C, Gomes V, Goios A, Parson W, Calafell F, Alvarez L, Amorim A, et al. 2011. Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. *PLoS One*. 6(1):e15988.

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Metspalu E, Mallick CB, Hudjashov G, Nelis M, Ma R, Remm M, et al. 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet*. 89(6):731–744.

Moorjani P, Patterson N, Loh P-R, Lipson M, Kisfali P, Melegh BI, Bonin M, Kádaši L, Rieß O, Berger B, et al. 2013. Reconstructing Roma history from genome-wide data. *PLoS One*. 8(3):e58633.

Neel JV. 1962. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet*. 14(4):353–362.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St P, Verzilli C, Shen J, Tang Z, Bacanu S, Warren L, et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*. 337(6090):100–104.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, et al. 2014. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*. 10(4):e1004234.

Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, et al. 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*. 356(6337):543–546.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2(12):e190.

Pedersen CET, Lohmueller KE, Grarup N, Bjerregaard P, Hansen T, Siegismund HR, Moltke I, Albrechtsen A. 2017. The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: insights from the Greenlandic Inuit. *Genetics*. 205(2):787–801.

Pemberton TJ, Szpiech ZA. 2018. Relationship between deleterious variation, genomic autozygosity, and disease risk: insights from the 1000 Genomes Project. *Am J Hum Genet*. 102(4):658–675.

Pierron D, Heiske M, Raza H, Pereda-Ioth V, Sanchez J, Alva O, Arachiche A, Boland A, Olaso R, Deleuze J, et al. 2018. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. *Nat Commun*. 9(1):932–939.

Plásilová M, Stoilov I, Sarfarazi M, Kádasi L, Feráková E, Ferák V. 1999. Identification of a single ancestral CYP1B1 mutation in Slovak Gypsies (Roms) affected with primary congenital glaucoma. *J Med Genet*. 36(4):290–294.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81(3):559–575.

R Core Team 2019. R: A Language and Environment for Statistical Computing.

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 47(D1):D886–D894.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, Mccarroll SA, Gaudet R, et al.; International HapMap Consortium. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*. 449(7164):913–919.

Secolin R, Mas-Sandoval A, Arauna LR, Torres FR, de Araujo TK, Santos ML, Rocha CS, Carvalho BS, Cendes F, Lopes-Cendes I, et al. 2019. Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci Rep*. 9(1):1–12.

Ségurel L, Wyman MJ, Przeworski M. 2014. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet*. 15:47–70.

Seldin MF, Pasaniuc B, Price AL. 2011. New approaches to disease mapping in admixed populations. *Nat Rev Genet*. 12(8):523–528.

She P, Shiota M, Shelton KD, Chalkley R, Postic C, Magnuson MA. 2000. Phosphoenolpyruvate carboxykinase is necessary for the integration of hepatic energy metabolism. *Mol Cell Biol*. 20(17):6508–6517.

Simons YB, Sella G. 2016. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Curr Opin Genet Dev*. 41:150–158.

Simons YB, Turchin MC, Pritchard JK, Sella G. 2014. The deleterious mutation load is insensitive to recent population history. *Nat Genet*. 46(3):220–224.

Sun C, Kong Q-P, Palanichamy M, gounder Agrawal S, Bandelt H-J, Yao Y-G, Khan F, Zhu C-L, Chaudhuri TK, Zhang Y-P. 2006. The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol Biol Evol*. 23(3):683–690.

Szpiech ZA, Hernandez RD. 2014. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*. 31(10):2824–2827.

Szpiech ZA, Mak AC, White MJ, Hu D, Eng C, Burchard EG, Hernandez RD. 2019. Ancestry-dependent enrichment of deleterious homozygotes in runs of homozygosity. *Am J Hum Genet*. 105(4):747–762.

Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zöllner S, Rosenberg NA, Li JZ. 2013. Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet*. 93(1):90–102.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 336(6090):64–69.

Tombácz D, Maróti Z, Kalmár T, Csabai Z, Balázs Z, Takahashi S, Palkovits M, Snyder M, Boldogkői Z. 2017. High-coverage whole-exome sequencing identifies candidate genes for suicide in victims with major depressive disorder. *Sci Rep*. 7(1):1–11.

Turner RL. 1927. The Position of Romani in Indo-Aryan. Gypsy Lore Society. London: B. Quaritch.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From fastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 43:11.10.1–11.10.33.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol*. 4(3):e72.

Vozarova De Courten B, De Courten M, Hanson RL, Zahorakova A, Egyenes HP, Tataranni PA, Bennett PH, Vozar J. 2003. Higher prevalence of type 2 diabetes, metabolic syndrome and cardiovascular diseases in gypsies than in non-gypsies in Slovakia. *Diabetes Res Clin Pract*. 62(2):95–103.

Yi X, Liang Y, Huerta-Sanchez EJX, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS. 2010. Sequencing of fifty human exomes reveals adaptation to high altitude. *Science*. 329(5987):75–78.

Zhang X, Kim B, Lohmueller KE, Huerta-Sánchez E. 2020. The impact of recessive deleterious variation on signals of adaptive introgression in human populations. *Genetics*. 215(3):799–812.

Živković TB, Marjanović M, Prgomelja S, Soldatović I, Koprivica B, Acković D, Živković R. 2010. Screening for diabetes among roma people living in Serbia. *Croat Med J*. 51(2):144–150.