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

To cite this version:

HAL Id: pasteur-03199596
https://hal-pasteur.archives-ouvertes.fr/pasteur-03199596
Submitted on 15 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial| 4.0 International License
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 Comas1

1Departament de Ciències Experimentals i de la Salut, Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Spain
2Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, Paris, France
3Facultat de Sociologia, Universitat Autònoma de Barcelona, Barcelona, Spain
4Facultat de Geografia i Història, Universitat de Barcelona, Barcelona, Spain
5Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Reus, Spain
6Human Genomics and Evolution, Collège de France, Paris, France

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

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; Tennesen 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-
Cruz et al. 2016; Melegh et al. 2017; Font-Porterias et al. 2019; Bianco et al. 2020; Dobon et al. 2020; Garcia-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; Bhathia 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 (Bhathia 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 (Bhathia 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
(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. 1B and C): $h_w$ is significantly lower in Roma and $p_{var}$ is significantly higher. In addition, $h_w$ 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). Significant differences were tested between Roma and non-Roma populations (*** refers to $P$ value $<0.001$ in all comparisons).

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. 9A, Supplementary Material online), and CADD scores (supplementary fig. 9B, 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 deleterious 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).

**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 ($\pi_{var}$ and $h_w$) from synonymous WES variants. Other diversity metrics (Tajima’s $D$ and $\theta_w$) 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).

**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.
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 (Nalleles) and the number of derived homozygotes per individual (Nhomo) were used as proxies, respectively (Lohmueller 2014) (fig. 3A and supplementary figs. 12–14, Supplementary Material online). Roma show the same Nalleles compared with non-Roma for all categories (GERP, PolyPhen, and CADD); however, they show a discrete but significant increase in Nhomo 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 RX/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 Nalleles and Nhomo 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
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 \( \frac{L_g}{L_{anc}} \sim 1 \) throughout all sampled generations (from 60 generations ago to present) (fig. 3B). Conversely, the mutational load of recessive mutations appears to be more sensitive to demographic events since both recessive models have relative mutational load values \( \frac{L_g}{L_{anc}} \) departing from 1. When recessive mutations are simulated under a model with gene flow from non-Roma to Roma, mutational load increases slightly \( \frac{L_g}{L_{anc}} \sim 1.018 \) 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 \( \frac{L_g}{L_{anc}} \sim 1.019 \) 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 \( \frac{L_g}{L_{anc}} = 0.983 \) and \( \frac{L_g}{L_{anc}} = 0.999 \), respectively; and for recessive models without and with gene flow, they reach \( \frac{L_g}{L_{anc}} = 1.134 \) and \( \frac{L_g}{L_{anc}} = 1.031 \), respectively. These simulated values with gene flow are in agreement with the observed load estimations: additive proxy \( N_{alleles} \) is centered \( \sim 1 \) and recessive proxy \( N_{hom} \) is found within 1 and 1.05 (fig. 3A).

**Table 1.** \( R_{XY} \) ratios between Roma and non-Roma populations for missense variants in each deleterious GERP category normalized by synonymous variants.

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

**Note.**—Point estimates and 95% CIs are shown.

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

<table>
<thead>
<tr>
<th>Number of ROHs</th>
<th>0.5 &lt; ROHs ≤ 2.5 (Mb)</th>
<th>2.5 &lt; ROHs (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All ROHs</td>
<td>0.1051</td>
<td>0.0148</td>
</tr>
<tr>
<td>Total ROH length</td>
<td>0.3766**</td>
<td>0.2518*</td>
</tr>
</tbody>
</table>

*\( P \) value < 0.05; **\( P \) value < 0.01.

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. 13.7%) (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
length per individual. However, an ancestry-specific pattern is observed: only in European segments (though not in South Asian), the rate at which deleterious \(N_{\text{hom}}\) 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 \(N_{\text{hom}}\) in South Asian ROHs is higher than in European ROHs (significantly different intercept \(\beta_2\)) (fig. 4 and supplementary fig. 25, Supplementary Material online). Second, the rate at which deleterious and synonymous \(N_{\text{hom}}\) in South Asian ROHs increase is higher than in European ROHs (significantly different slope \(\beta_3\), 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 transerase 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. 5A, and supplementary fig. 28A, 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. 5A), 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. 28B 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. 5B, and supplementary fig. 28D, 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. 4.** Fraction of \(N_{\text{hom}}\) 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. \(\beta_2\) and \(\beta_3\) show intercept and slope differences between regressions. **P** value < 0.05; ***P** value < 0.01; ****P** value < 0.001.
this region. The European population shows the same signal when tested against South Asia (fig. 5B), 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. 28E 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. 28G–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 ($N_{\text{het}}$). However, this proxy assumes that all mutations in the genome are recessive, whereas $N_{\text{alle}}$ assumes semidominance (Lohmueller et al. 2008). Some studies suggest that most deleterious variants in the human genome have an additive dominance coefficient, pointing to $N_{\text{alle}}$ as a more reliable proxy to estimate mutational load values in present-day populations (Simons and Sella 2016). $R_{XY}$ statistics do not show statistical differences between Roma and non-Roma, in agreement with $N_{\text{alle}}$ 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 $N_{s},s \ll 1$, 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 $N_{s},s \ll 1$ 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 simulated model without non-Roma to Roma gene flow. As expected, the Roma has more strength than purifying selection (Lohmueller et al. 2008). Nonetheless, the presence of gene flow from a larger ancestral population results in a faster increase in recessive load, whereas the presence of gene flow has counteracted the increase in recessive load, thereby reducing the overall mutational load.

In summary, our results show that the recessive load 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 simulated model without non-Roma to Roma gene flow. As expected, the Roma has more strength than purifying selection (Lohmueller et al. 2008). Nonetheless, the presence of gene flow from a larger ancestral population results in a faster increase in recessive load, whereas the presence of gene flow has counteracted the increase in recessive load, thereby reducing the overall mutational load.

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 (Páslóva 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 Cami” 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 486,099 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://broadinstitutegithubio/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) $<5$, 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× minimum, missingness $<5\%$, heterozygosity $<\text{mean} + 4 \times 4 \text{SD}$, 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 ± 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), GERD (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 GERD RS groups: neutral ($-2 < \text{GERD} < 2$), slightly deleterious ($2 < \text{GERD} < 4$), moderate ($4 < \text{GERD} < 6$), and extremely deleterious ($\text{GERD} > 6$) (Henn et al. 2016). For PolyPhen-2, three categories were used: benign ($<0.0446$), 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

Population Structure Analysis
PCA and ADMIXTURE were performed using the merged data set of genome-wide array data 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 $\pm$ 6.57% and 21.48 $\pm$ 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 ($\theta_{w}$), Watterson’s $\theta$ ($\theta_{o}$), 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 ($\pi_{vcw}$), 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 variants grouped in GERP, PolyPhen, and CADD scores. In addition, it is specially detected in Roma samples (26.15 $\pm$ 6.57% and 21.48 $\pm$ 6.39%; mean dark-red component with genome-wide array and WES data, respectively).

Mutational Load Proxies
Two summary statistics were used as proxies for mutational load: number of derived alleles per individual ($N_{\text{alleles}}$) and number of derived homozygotes per individual ($N_{\text{homo}}$). $N_{\text{alleles}}$ and $N_{\text{homo}}$ 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 $R_{XY}$ 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 ($N_{\text{alleles}}$, $N_{\text{homo}}$, and $R_{XY}$) 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 $\lambda_{\text{a/h}}/F_{\text{it}/\lambda_{\text{a/h}}}$ (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. $\lambda_{\text{a/h}}/F_{\text{it}/\lambda_{\text{a/h}}}$ infers the mean $E(N_{\text{s}})$; thus, we estimated $E(s) = E(N_{\text{s}})/N_{\text{w}}$, where $N_{\text{w}}$ is the weighted effective population size along time (Lopez et al. 2018) from the $N_{\text{ANC}}$ calculated from $\theta_{S} = 4N_{\text{ANC}}/\mu L_{S}$ (Kim et al. 2017), where $\theta_{S}$ is the population-scaled synonymous mutation rate and $\mu = 1.5 \times 10^{-8}$ (Ségurel et al. 2014). $L_{S}$ derives from $L = L_{\text{NS}} + L_{S}$, where L is the number of bases from which variants were called and assuming a ratio of synonymous to nonsynonymous sites $L_{S}/L_{\text{NS}} = 2.31$ (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 $1.36 \times 10^{-8}$ 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 DEEs 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, $l = s(2hq + (1 - 2h) \times q^2)$, $s$ 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 reference source: European (IBS and TSI populations) and ancestry-specific segments inferred from merged data sets 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): $PBS = (T^{WX} + T^{WY} - T^{XY})/2$; $T^{WY} = -\log(1 - F_{ST}^{WY})$, where $Y$ represents YRI, $X$ either European or South Asian group populations, and $W$ the target population (Roma, European, 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
Maetzu” (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**


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


