Article

Discovery section

Recent common origin, reduced population size, and marked admixture have shaped European Roma genomes.

Erica Bianco 1, Guillaume Laval 2, Neus Font-Porterias 1, Carla García-Fernandez 1, Begoña Dobon 1,# Rubén Sabido-Vera 1, Emilija Sukarova Stefanovska 3, Vaidutis Kučinskas 4, Halyna Makukh 5, Horolma Pamjav 6, Lluis Quintana-Murci 2,7 Mihai G. Netea 8,9, Jaume Bertranpetit 1, Francesc Calafell 1, David Comas 1,*

1 Departament de Ciències Experimentals i de la Salut, Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, Spain;
2 Human Evolutionary Genetics Unit, Department of Genomes and Genetics, UMR 2000, CNRS, Institut Pasteur, Paris, France.
3 Research Center for Genetic Engineering and Biotechnology “Georgi D. Efremov”, Macedonian Academy of Science and Arts, Skopje, Macedonia;
4 Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University, Vilnius, Lithuania;
5 Institute of Hereditary Pathology of the Ukrainian Academy of Medical Sciences, Lviv, Ukraine;
6 Department of Reference Sample Analysis, Institute of Forensic Genetics, Hungarian Institute for Forensic Sciences, Budapest, Hungary
7 Chair Human Genomics and Evolution, Collège de France, Paris, France
8 Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, the Netherlands
9 Department for Genomics & Immunoregulation, Life and Medical Sciences 12 Institute (LIMES), University of Bonn, Bonn, Germany
* to whom correspondence should be addressed.
# current address: Department of Anthropology, University of Zurich, Zurich, Switzerland

Corresponding Author
David Comas: david.comas@upf.edu

Abstract

The Roma Diaspora — traditionally known as Gypsies —remains amongst the least explored population migratory events in historical times. It involved the migration of Roma ancestors out-of-India through the plateaus of Western Asia ultimately reaching Europe. The demographic effects of the Diaspora – bottlenecks, endogamy, and gene flow – might have left marked molecular traces in the Roma genomes. Here, we analyze the whole genome sequence of 46 Roma individuals pertaining to four migrant groups in six European countries. Our analyses revealed a strong, early founder effect followed by a drastic reduction of ~44% in effective population size. The Roma common ancestors split from the Punjabi population, from -Northwest India, some generations before the Diaspora started, less than 2,000 years ago. The initial bottleneck and subsequent endogamy are revealed by the occurrence of extensive Runs of Homozygosity and Identity By Descendent segments in all Roma populations. Furthermore, we provide evidence of gene flow from Armenian and Anatolian groups in present-day Roma, although the primary contribution to Roma gene pool comes from non-Roma Europeans, which accounts for more than 50% of their genomes. The linguistic and historical differentiation of Roma in migrant groups is confirmed by the differential proportion, but not a differential source, of European admixture in the Roma groups, which shows a westward cline. In the present study we found that despite the strong admixture Roma had in their diaspora, the signature of the initial bottleneck and the subsequent endogamy is still present in Roma genomes.

Key words: Roma Diaspora; Gypsies; Complete genomes; Demographic history; endogamy; admixture

Introduction

Roma people, also known with the misnomer term of Gypsies, are the largest transnational minority in Europe, accounting for 10-15 million people dispersed across the continent (Liégeois 1994; Fraser 1995); yet, little is known about the details of the Roma Diaspora. According to anthropologic, linguistic (reviewed in Liégeois 1994; Fraser 1995), and genetic studies (Gresham et al. 2001; Kalaydjieva et al. 2001; Morar et al. 2004; Mendizabal et al. 2011; Mendizabal et al. 2012; Moorjani, et al. 2013; Martínez-Cruz et al. 2016; Melegh et al. 2017), Roma left the North Western part of the Indian subcontinent ~15 centuries ago, crossed the Iranian, Armenian, and Anatolian plateaus, and reached Europe in the 9th-10th century CE. In Europe, Roma migrated within the continent in different waves, associated with the Romani dialects spoken by the different groups (Liégeois 1994; Fraser 1995; Hancock 1995; Gresham et al. 2001). There are four main migrant groups: Balkan and Vlax Roma, living in the Balkan Peninsula; Romungro Roma, living in central Europe; and North/Western Roma, living in Northern and Western Europe (Liégeois 1994; Fraser 1995).

Despite the recent origin of Roma people (Fraser 1995; Gresham et al. 2001; Chaix et al. 2004; Morar et al. 2004; Mendizabal et al. 2012; Moorjani, et al. 2013), all studies performed so far revealed a highly complex demographic history, at different levels (Chaix et al. 2004; Mendizabal et al. 2012; Moorjani, et al. 2013; Martínez-Cruz et al. 2016; García-Fernández et al., unpublished data). The first level of complexity is represented by the aforementioned recent origin of Roma people, together with the series of bottlenecks and splits experienced during their Diaspora (Chaix et al. 2004; Mendizabal et al. 2012), and different levels of endogamy (Liégeois 1994; Fraser 1995; Chaix et al. 2004; Kalaydjieva et al. 2005; Mendizabal et al. 2012), which have left traces of low intragroup diversity and high intergroup heterogeneity (Peričić et al. 2005; Malyarchuk et al. 2006; Irwin et al. 2007; Gusmão et al. 2008; Klarić et al. 2009; Zalán et al. 2010; Salihović et al. 2011; Martínez-Cruz et al. 2016). Roma people low effective population size (Ne), due to bottlenecks, founder effects, and endogamy, is thought to have resulted in the occurrence of a number of disease-linked variants that are private to the Roma groups (Kalaydjieva et al. 2001; Morar et al. 2004; Mendizabal et al. 2013). An additional level of complexity is generated by the different admixed ancestry components found in the present Roma groups. Whereas the South Asian component was likely present in proto-Roma people, represented by their Ancestral South Indian component (ASI), the West Eurasian component can be traced back either to the Ancestral North Indian (ANI) component of the proto-Roma (Reich et al. 2009; Moorjani, et al. 2013) or to their recent admixture with other non-Roma European populations (termed ‘Europeans’ from now on), following their arrival to Europe (Chaix et al. 2004; Peričić et al. 2005; Malyarchuk et al. 2006; Gusmão et al. 2008; Gusmão et al. 2010; Mendizabal et al. 2011; Mendizabal et al. 2012; Moorjani, et al. 2013; Martínez-Cruz et al. 2016; Font-Porterias et al. 2019; García-Fernández et al. submitted). Such recent admixture events strongly influenced the present-day Roma West Eurasian component (Moorjani, et al. 2013; Melegh et al. 2017), but its origins have only recently started to be explored (Font-Porterias et al. 2019).

So far, Incomplete genomic data, lack of representative samples, and weak definition of Roma groups have precluded a deep analysis of the Roma genomic history. To explore at high resolution the history of the Roma Diaspora and its consequences on the genomic landscape of current Roma groups, we generated whole genome sequences, at high coverage (~30X), from 40 Roma volunteers from five different countries and belonging to the four main migrant groups. These new data have been compared with genomes of non-Roma surrounding populations with the aims of: (1) analyzing the origins and substructure of Roma migrant groups; (2) defining the admixture patterns and the population origins of the genetic components present in the Roma; (3) exploring the degree of endogamy of Roma groups; and, (4) providing a demographic framework of the Roma Diaspora.

Results

Population structure of Roma

We analyzed 40 newly sequenced Roma complete genomes from unrelated volunteers, belonging to the four main migrant groups (Balkan, Vlax, Romungro, and North/Western) from five different countries, together with six Roma Vlax individuals from Romania (Dobon et al., unpublished data), within the landscape of Europe, the Middle East, and the Indian subcontinent (see Figure 1A for a map of the sampling location and SM Table 1 and SM Table 2 for samples details) (Mallick et al. 2016; Mondal et al. 2016; Serra-Vidal et al. 2019). In total, we analyzed 155 genomes with an average coverage of 12X-35X that contained in total 7,838,351 SNPs, that were reduced to 1,445,921 SNPs, after filtering for missing data and pruning for linkage disequilibrium.

We first assessed the genetic structure of Roma using Principal Component Analysis (PCA, Figure 1B and SM Figure 1) (Patterson et al. 2006; Price et al. 2006), and showed that Roma people form a cluster that falls in a cline between Europe and South Asia, in agreement with their South Asian origin of Roma and their subsequent admixture with Europeans (Mendizabal et al. 2012; Moorjani, et al. 2013; Melegh et al. 2017). Within Roma, individuals from the same Roma group tend to cluster together in the PCA. We found a clear separation within the North/Western Roma, between Spanish and Lithuanian Roma, and between North/Western Roma and the other migrant groups (SM Figure 1).

Admixture analysis (Alexander et al. 2009) showed the ancestral component composition of Roma (Figure 1C, SM Figure 2A and SM Figure 2B). At K=4 (the lowest cross validation value, SM Figure 2B and SM Figure 2C), Roma showed two main ancestry components: West Eurasian and South Asian, confirming again the South Asian origin of Roma people and the recent European admixture (Mendizabal et al. 2012; Rai et al. 2012; Moorjani, et al. 2013; Melegh et al. 2017; Font-Porterias et al. 2019). From K=5 onwards (SM Figure 2A), Roma people showed their own genetic component without any migrant group differentiation. Both PCA and Admixture analyses identified an individual with full European ancestry (V5, Hungarian Vlax), who was removed from subsequent analyses (see also SM Figure 2B).

Allele sharing and gene flow in Roma groups

To detect the signatures left by the European gene flow into the Roma gene pool, we analyzed the allele sharing using the outgroup *f3*-statistic in the form of *f3*(YRI; Roma, X), where Yoruba (YRI) from Africa were used as an outgroup, and X stands for the set of samples tested for comparison to Roma (Reich et al. 2009). The outgroup *f3*-statistic values show a cline from Europe to the Indian subcontinent linked to the route of the Roma Diaspora (Figure 2 and SM Figure 3). European populations are the non-Roma group that shared the most alleles with Roma, in agreement with the recent extensive European gene flow in Roma (Font-Porterias et al. 2019)

(regardless the migrant group, SM Figure 3). Eastern Asians (i.e. CHB-Han Chinese), North Africans (i.e. EGY-Egyptians), Southern Indian populations (i.e. ILA-Irula and VLR-Vellalar), together with other populations outside the putative Roma Diaspora route, were the groups that shared less alleles with Roma.

Among European populations, the *f3*-statistics showed no significant differences in the allele sharing by migrant group. We then tested for differences in the source of gene flow using *D*-statistics (Supplementary Note 1) in the form *D* (E1, E2; CHB, Roma), where E1 and E2 are two European populations, Roma is any migrant group, and CHB is the Han Chinese as an outgroup. We found no evidence of specific European populations as the source of gene flow to a specific migrant group, although Southwestern and Southeastern Europeans appeared to have contributed more to the gene flow to Roma (Supplementary Note 1 and SM Table 3). Moreover, we tested whether migrant groups had differential European gene flow using *D*-statistic in the form *D* (E, CHB; MG1, MG2), where MG1 and MG2 are any two Roma migrant groups, and E is any European group. Our results showed a westward cline of increasing European gene flow from Balkan to North/Western Roma (Supplementary Note 1 and SM Table 4).

Among northern Indian populations, the Punjabi (PUN) were the population who shared consistently more alleles with the Roma than any other Northern Indian populations, and more or similar allele sharing than Pakistani populations, with Roma individuals (Figure 2 and SM Figure 3), in agreement with previous observations, that pointed to Punjab as the proto-Roma area of origin (Mendizabal et al. 2012; Font-Porterias et al. 2019). Although other Pakistani populations showed higher or similar allele sharing than Punjab, it has been shown that these populations experienced gene flow from Europe after the Roma Diaspora (Hellenthal et al. 2014), which could increase the allele sharing between Pakistani populations and Roma due to common European gene flow. To test this hypothesis, we compared the allele sharing of Pakistani and Northern Indian populations with Roma and non-Roma Europeans in the form *f3* (X, EUR; YRI), where X is any Pakistani or Northern Indian population, and EUR represents Europeans (SM Figure 4A). Pakistani populations (BAL, BRA, BUR, HAZ, KAL, MAK and SIN) and Uttar Pradesh Brahmins (UBR) from Northern India shared more alleles with Europeans than with Roma (higher *f3* value), or there was no difference in the allele sharing with Roma and with Europeans, suggesting the allele sharing found in outgroup *f3*-statistic between Roma and Pakistan populations may be inflated by the common gene flow with Europeans rather than being just the result of a common origin.

Among Middle Easterner populations, Armenians were those who shared more alleles with Roma people, followed by Turks, with no differences between migrant groups (Figure 2 and SM Figure 3). As for Pakistani populations, the higher allele sharing between Roma and some Middle Easterner populations than with Northern Indian populations is partly explained by the high allele sharing between these populations and Europeans (SM Figure

4B). Furthermore, despite the strong allele sharing with Europeans, the genetic fingerprint of Middle Easterner populations in the Roma genomes confirms the route of the Roma Diaspora from Northern India through this part of the Middle East, before reaching Europe (Bánfai et al. 2019; Font-Porterias et al. 2019).

Within Roma, the *f3*-statistic values show that the amount of shared alleles between migrant groups was systematically higher than between a migrant group and any other tested populations (SM Figure 3), confirming the single and common origin of Roma groups (Gresham et al. 2001; Kalaydjieva et al. 2005; Martínez-Cruz et al. 2016). No significant differences in the allele sharing between migrant groups were observed, being North/Western Roma the groups that shared the least with the other groups due to their larger admixture with Europeans (SM Figure 2B and SM Table 4).

Identity by descent between Roma and other populations

We estimated the Identity by Descent (IBD) segments shared between individuals from different areas and Roma people (Browning and Browning 2013a). For each comparison, we excluded all segments that overlap with segments in IBD between Roma and any individual of another population group, including in the comparison of IBD segments uniquely shared between Roma and the area of interest only.

We found that the IBD load between Roma and Europeans was consistently higher than the IBD load between Roma and any other population (Figure 3A, by migrant group in SM Figure 5), with the exception of the Punjabi (PUN), supporting the strong, very recent gene flow between these two populations, also shown in the *f3*-statistic results (Figure 2). We found no differences that suggest a single or main European source of admixture in the Roma.

Comparing the unique IBD load between migrant groups and Europeans, we found that North/Western Roma shared significantly more IBD with Europeans than the other migrant groups. In agreement with the results of the *D*-statistics (SM Table 4), this result confirms the higher gene flow from Europeans to North/Western Roma (Figure 3B and SM Figure 2BC). We did not detect significant differences between Balkan and Romungro Roma. Although the only Roma individual we found to have full European ancestry was Vlax, Vlax Roma showed the lowest IBD sharing with Europeans (ROMAV in Figure 3B).

Among Northern India/Pakistan populations (Figure 3A and SM Figure 5), we found the Punjabi (PUN) to be a clear outlier, having significant more IBD (overall length) than the rest of populations. The IBD sharing between Roma and Punjab was comparable with the IBD sharing between Roma and Europeans. This result confirms Punjab as the putative region of origin of Roma ancestors within the Indian subcontinent (Mendizabal et al. 2012; Font-Porterias et al. 2019). Moreover, we found no differences among the migrant groups either when comparing the IBD load between Roma and all Indian populations (Figure 3C) or between Roma and Punjabi (SM Figure

6A), a footprint of the single founding event of Roma population (Gresham et al. 2001; Kalaydjieva et al. 2005; Martínez-Cruz et al. 2016) that also excludes gene flow after the Diaspora between Roma and populations from the Indian subcontinent.

Finally, we compared the IBD load between Roma and populations from the Middle East (Figure 3A). The overall IBD load was higher than that of Northern India (except with Punjabi - PUN), in agreement with the occurrence of gene flow between Roma and populations from this area (Bánfai et al. 2019; Font-Porterias et al. 2019). The homogeneity of the IBD load between Middle Easterner populations and Roma migrant group suggest no admixture between these two groups after Roma arrived in Europe (SM Figure 6B).

Estimates of endogamy

We evaluated the extent of Roma endogamy by analyzing the length and number of Runs of Homozygosity (RoH) within Roma individuals and the Identity By Descent (IBD) segments between individuals of the same population and migrant group.

Roma showed a higher RoH number and total length than Europeans or Northern Indian populations, a clear signature of the founder effect and subsequent bottlenecks that Roma suffered in their demographic history (Figure 3D and SM Figure 7). The only two groups that show higher RoH loads than the Roma were the South Indian populations Irula (ILA, tribal population) and Vellalar (VLR, Dravidian non-tribal), who are known to present high levels of endogamy (Figure 3D and SM Figure7) (Juyal et al. 2014; Mondal et al. 2016). Within Roma, Balkan Roma (ROMAB) showed fewer and shorter RoH. In the analysis by length category, we found significant differences between Balkan Roma (ROMAB) and the other migrant groups in long RoH (length > 5Mb), a possible signature of the additional bottlenecks that non-Balkan Roma experienced during their history, after the split from Balkan Roma (SM Figure 7) (Mendizabal et al. 2012).

Additionally, we tested endogamy within migrant groups by evaluating the average cumulative length of the genome that was in IBD between two individuals from the same group (SM Figure 8). We found the cumulative IBD length of Roma population to be comparable to the pattern observed in Punjabi (PUN) and in the endogamic populations of Southern India (ILA and VLR) (Mendizabal et al. 2011; Mondal et al. 2016; Nakatsuka et al. 2017). Among migrant groups, we did not detect any differences in the load of IBD between individuals of the same migrant group. Together with the differences we found in the RoH load of Balkan Roma, the same amount of IBD load in Roma is consistent with the very recent and weaker additional bottlenecks non-Balkan Roma have suffered (Mendizabal et al. 2012). The signature left on the genome of such recent, and weaker, bottlenecks can be detected only on the within-individual diversity (RoH), and not on the between-individuals diversity (IBD)(Severson

et al. 2019). Moreover, the distribution of the total length of IBD fragments was not different from that of Roma people taken together or Southern Indian populations (SM Figure 8).

Defining the Roma demographic history

We finally analyzed the demographic history of Roma by testing different scenarios and using three main approaches: (i) assessing the changes in the effective population size from the IBD pattern; (ii) analyzing the complex admixture history based on allele sharing between populations; and (iii) estimating the best fitting demographic model to Roma people using an Approximate Bayesian Computation approach (which also includes the two first approaches in its calculation).

Changes in the effective population sizes throughout generations

We analyzed the historical effective population size (Ne) of the last 200 generations inferred by the IBD pattern between individuals of the same population using IBDNe (Browning and Browning 2015) (SM Figure 9). By analyzing all Roma groups together, we found that Roma Ne overlapped with the Northern Indian individuals Ne until ~125 generations ago, in agreement with the Roma origins in this region of the subcontinent. Starting ~125 generations ago, Roma Ne suffered a constant reduction and started to differentiate from Northern India Ne. Northern Indian and Southern Indians groups did not exhibit this dramatic reduction in population size. We found the minimum of Roma Ne to be 1,000 (871 – 1,160 95%CI), and it occurred between 1,159 CE – 1,333 CE (between 23 and 29 generations ago, assuming a generation time of 29 years). In recent times, Ne slightly increased and decreased again in the last generations before present. Overall, Roma Ne in the last 125 generations was lower than Northern Indians, and in the last 50 generations was also lower than the Southern Indian Dravidic groups (Juyal et al. 2014; Mondal et al. 2016) (SM Figure 9). At the time of the out of India, which was previously estimated to occur between 35 to 50 generations ago (Liégeois 1994; Fraser 1995; Mendizabal et al. 2012; Moorjani, et al. 2013; Martínez-Cruz et al. 2016), we found Ne to be 1,500 (1,380 – 1,640 95%CI) and 4,600 (3,850 – 6,120 95%CI), respectively.

Admixture history of Roma

The Roma West Eurasian component is the result of the recent gene flow on the Northern Indian background of proto-Roma, who were composed by a Northern Indian (ANI) and a Southern Indian (ASI) ancestral component as any other Indian population (Reich et al. 2009; Moorjani, et al. 2013; Pathak et al. 2018; Yelmen et al. 2019). We tested different complex admixture scenarios to investigate the strength of recent admixture between Roma and Europeans using the admixture graph approach (Patterson et al. 2012). Our analyses found two scenarios that fit our data (Figure 4 and SM Figure 10): both have an Ancestral Northern Indian population (ANI) and an Ancestral Southern Indian population (ASI) that admixed into the ancestors of proto-Roma, confirming the

admixed nature of this population, but the proportion of ANI/ASI differs between the two scenarios. In the first scenario, the ancestors of proto-Roma were the ancestors of the Punjabi population, which presents 51% of ANI (shown in Figure 4 and in agreement with (Mendizabal et al. 2012; Font-Porterias et al. 2019)). In the second scenario, the ancestors were an unknown Northern Indian-like population, with 67% of ANI (SM Figure 10). In both scenarios, two admixture events from a West Eurasian source were necessary for the scenario not to be discarded: a first admixture event between ANI and ASI populations to form the ancestors of the proto-Roma (i.e. the ancestors of Punjabi or the unknown sister population, depending on the model), and a second admixture event between proto-Roma and Europeans. Since both ANI and Europeans are of West Eurasian origin, depending on the ANI ancestry of the ancestral population of Roma, the subsequent admixture with Europeans will have different proportions, 67% in the first scenario (Figure 4) and 53% in the second (SM Figure 10).

A byproduct of the recent admixture with Europeans was the reduction of the ANI and ASI components of Roma ancestors in present-day Roma ancestry (Mendizabal et al. 2012; Moorjani, et al. 2013; Melegh et al. 2017). In our admixture graph analysis (Figure 4 and SM Figure 10), the relative proportion of the ANI component depends on the population of origin of the Roma, but, in both models, the ASI component is ~16%. Although we found that two scenarios fit our data, the first scenario with the ancestral Punjabi (a_pun) as a proxy population for Roma ancestors (Figure 4) seems most likely, since it considers a lower number of 'ghost' populations and it is in agreement with our IBD analyses (Figure 3A) and previous data (Mendizabal et al. 2012; Font-Porterias et al. 2019).

Roma demographic model

We applied an Approximate Bayesian Computation (ABC) approach (Beaumont et al. 2002; Beaumont 2010) to discriminate between 12 demographic scenarios, which included two different branching structures and an increasing number of asymmetrical bidirectional gene flow between Roma migrant groups and Europeans (SM Figure 11, SM Table 5 for model's parameters, Supplementary Note 2). The highest posterior probability was obtained for the model with four migration events and the two-branch structure (2b4m, SM Figure 12, average posterior probability=0.566 using neuralnet, SM Table 6). This model involved a first split between migrant groups in the Balkan Peninsula (Balkan and Vlax common ancestors) and migrant groups outside of the Balkan Peninsula (Romungro and North/Western common ancestors), and two more recent splits, one between Romungro and North/Western Roma and the other between Balkan and Vlax Roma.

Among the parameters that could be accurately estimated (Table 1 and SM Figure 13, see also Supplementary Note 2), our ABC analysis indicated that the split between Roma and their Indian ancestors occurred ~1.6 - 2kya (TbotBAIND, 95%CI 475ya – 3,700ya, assuming a generation time of 29 years, Table 1 and SM Figure 14). The

reduction of the Ne in the bottleneck was ~44% of the ancestor population (95%CI 33% - 59%, mutual of bot1a), reducing the Ne of Roma ancestors to 1,536 (N1, 95% CI 188 – 2,387, Table 1 and SM Figure 14).

Ancestry of described Mendelian variants in Roma

The high level of endogamy present in the Roma genome might have increased the probability to carry deleterious mutations and risk alleles associated with diseases (Kalaydjieva et al. 2001). Among the 48 variants that have been associated with Mendelian diseases, and that have been previously described in Roma (summarized in SM Table 7), eleven were found in the Roma genomes analyzed here (SM Table 8). The local ancestry analysis of these variants, performed with RFMix (Maples et al. 2013), shows that nine variants had European ancestry, pointing to the major European component of the mendelian load in present-day Roma. Only two variants show South Asian ancestry: variant rs121918355, associated with primary congenital glaucoma (Azmanov et al. 2011); and variant rs104894396, associated with deafness(Álvarez et al. 2005), that had mixed ancestry.

Discussion

The analysis of the demographic history of Roma people presents two main levels of complexity. First, Roma population emerged very recently, as they left the Northwestern part of the Indian subcontinent around 1 – 1.5kya, 35 – 50 generations ago (Liégeois 1994; Fraser 1995; Mendizabal et al. 2012; Moorjani, et al. 2013; Martínez-Cruz et al. 2016). Second, despite being considered an isolated population, with extensive bottlenecks, splits, and endogamy, Roma are an admixed population, as in their Diaspora, they experienced gene flow from other non-Roma groups (Mendizabal et al. 2012; Moorjani, et al. 2013; Martínez-Cruz et al. 2016; Bánfai et al. 2019; Font-Porterias et al. 2019).

An outstanding question of Roma history concerns their origins before their Diaspora towards Europe from the Indian subcontinent. To tackle this question, a complexity factor arise because of the admixed nature of both Roma and the populations of the Indian subcontinent (Mendizabal et al. 2012; Moorjani, et al. 2013; Moorjani, et al. 2013; Pathak et al. 2018; Font-Porterias et al. 2019; Tätte et al. 2019; Yelmen et al. 2019). Indian groups have been described as a mixture of two main ancestral components named ANI and ASI due to their higher frequencies in Northern and Southern India, respectively, although both components are found in most Indian subcontinent populations(Reich et al. 2009; Moorjani, et al. 2013). The ANI component has a Western Eurasian origin and seems to have reached the Indian subcontinent in more recent times compared to the ASI component (Reich et al. 2009). The ANI/ASI components proportion of the proto-Roma in our admixture graph analysis (Figure 4) confirms the Northwestern part of the Indian subcontinent to be the area of origin of Roma (Mendizabal et al. 2012; Melegh et al. 2017; Font-Porterias et al. 2019), and it is supported by our IBD and outgroup *f3*-statistic

results (Figure 2 and SM Figure 3, Figure 3A and SM Figure 5). The populations from this area showed the highest *f3*-statistic and IBD load among the Indian and Pakistani populations. Among them, our results support the Punjabi region as the region of origin of the Roma Diaspora. Indeed, the Punjabi population showed the highest IBD load, presented one of the highest outgroup *f3*-statistic values for Northern Indian populations, and shared the most recent common ancestor with Roma in the best fit scenario in the admixture graph analysis. Moreover, the Punjab as the region of origin of Roma is in agreement with previous linguistic data, because of the similarity between Romaní – Roma people language – and the languages of that area, and genetic studies that pointed to Punjab, despite the population heterogeneity found in the region (Fraser 1995; Gresham et al. 2001; Kalaydjieva et al. 2001; Mendizabal et al. 2011; Mendizabal et al. 2012; Mendizabal et al. 2013; Moorjani, et al. 2013; Martínez-Cruz et al. 2016; Melegh et al. 2017; Alfonso-Sánchez et al. 2018; Font-Porterias et al. 2019).

Our demographic analysis using an ABC framework, using the Punjabi as proxy for Roma ancestors (NIND in SM Figure 12), showed that the ancestors of the present-day Roma split from the Punjabi ~1.6 – 2kya (Table 1 and SM Figure 14, TbotBAIND). Historical records (Liégeois 1994; Fraser 1995) and previous genetic analyses (Mendizabal et al. 2012) pointed to slightly more recent times (1.2 – 1.5 kya), which still fall within our 95% confidence interval, and the overlap increases when correcting for the generation time estimation (25 years in (Mendizabal et al. 2012), 29 years in our analyses). Thus, our analyses suggest that the ancestors of the Roma and the Punjabi were already two differentiated populations within the Indian subcontinent, some generations before the Roma Diaspora started. The split from a common ancestor of two populations is rarely an instantaneous process (i.e. it does not occur in a single generation). Even though it is likely to have been overestimated by mixing Northern Indian individuals from different populations and geographic areas, the range of time it took for Roma ancestors and its Northern Indian ancestors to split it is suggested by the trend of the historical effective population size throughout the generations of the two populations, which started to differentiate before the Roma Diaspora began (~125 generations ago, SM Figure 9).

The complexity of the peopling history of India, where many questions remain to be answered (Reich et al. 2009; Moorjani, et al. 2013; Pathak et al. 2018; Narasimhan et al. 2019; Tätte et al. 2019; Yelmen et al. 2019), further complicates the search for the area of origin of the Roma. Moreover, among the other South Asian populations, the outgroup *f3*-statistic results pointed to other Pakistani populations to share a similar amount of alleles with Roma as Roma do with Punjabi (Figure 2 and SM Figure 3), but the IBD results show a lower amount of IBD sharing between Roma and these populations (SM Figure 6). Since the geographical border between India and Pakistan was redefined in the last century, and the Punjabi region is shared between modern India and Pakistan (Melegh et al. 2017), we explored whether other Pakistani populations contributed to the Roma genomic landscape. We tested whether the high allele sharing between Roma and Pakistan populations (Figure 2) could

be due to a common ancestor or to the recent common gene flow with non–Roma Europeans, which occurred after the Roma Diaspora (Hellenthal et al. 2014). In a further outgroup *f3*-statistics analysis (SM Figure 4), we found non-Punjabi Pakistani populations to share more alleles, or a similar amount, with Europeans than with Roma. This observation could reflect a signature of the recent gene flow from Europeans to Pakistani groups, which led to the signal of allele sharing between Roma and Pakistani populations occurring after Roma left India (Hellenthal et al. 2014).

The present genome analysis of the Roma suggests that, during the Diaspora, the route followed by their ancestors included the Armenian highlands and Anatolia, as shown in the *f3*-statistic and IBD analyses, in agreement with historical, linguistic (reviewed in (Liégeois 1994; Fraser 1995; Hancock 1995)), and genetic data (Bánfai et al. 2019; Font-Porterias et al. 2019). By the other hand, our data clearly discard the North African origin of Roma groups or a route throughout North Africa, in our analysis represented by Egyptians, in contrast to previous historical hypotheses (Fraser 1995) and in agreement with previous genetic data (Gresham et al. 2001; Mendizabal et al. 2012; Moorjani, et al. 2013; Martínez-Cruz et al. 2016; Font-Porterias et al. 2019) (Figure 2 and SM Figure 3). Nonetheless, the major genetic contribution to present-day Roma comes from the recent gene flow from European groups, making Roma genetically more similar to Europeans than to their Indian parental population (Font-Porterias et al. 2019; Dobon et al., unpublished data). Indeed, the recent gene flow between Roma and other populations in Europe accounts, in average, for more than 50% of Roma genomes (Figure 4 and SM Figure 10).

The strength of European gene flow is shown also in the amount of allele sharing and IBD fragments between Roma and the rest of European groups, which were higher than Northern Indian populations (Figure 4 and Figure 3A). Previous studies found the IBD load was higher between Roma and European groups from Eastern Europe (Moorjani, et al. 2013; Melegh et al. 2017) with strong influence from the Balkan peninsula (Font-Porterias et al. 2019). In our analysis, we found Southwestern and Southeastern European groups to have provided more gene flow to Roma compared to other European groups (SM Table 3), but no differences were found in the outgroup *f3*-statistic (SM Figure 3), whose power can be limited due to drift (Patterson et al. 2012). Despite being settled in different areas of the European continent, the signal of the admixture between Roma and Europeans is shared across migrant groups according to *D*-statistics analysis (SM Table 3). Besides the strength of the European gene flow, higher in North/Western Roma and lower in Vlax Roma, no consistent differences were found among Roma migrant groups (Figure 3B and SM Table 4), pointing to a recent common origin and lack of differentiation, possibly due to the short time span since the split of Roma groups. This result contrasts to some differentiation found in the analysis of some uniparental lineages (Gresham et al. 2001; Martínez-Cruz et al. 2016), which might be explained due to the higher drift of uniparental genomes.

The strong admixture Roma experienced with European groups did not erase, however, the signature of the strong bottleneck undergone by the Roma at the beginning of their Diaspora, which represented ~44% of the proto-Roma parental population (Table 1), in agreement with previous data (Mendizabal et al. 2012). Despite the increase in effective population size (Ne) Roma had in the generations after the bottleneck (SM Figure 9), which might be explained by the extensive gene from Europeans (Font-Porterias et al. 2019), the degree of genetic homogeneity in the Roma genomes are comparable to populations of Southern India who are known to follow cultural practices of marriages between close relatives (Mondal et al. 2016; Nakatsuka et al. 2017). At its lower point, the Ne of Roma was even lower than the Ne of Southern Indian populations (SM Figure 9), making Roma more prone to carry deleterious alleles at higher frequencies (SM Table 8) (Kalaydjieva et al. 2001). It is noteworthy that the Mendelian associated variants found in the present dataset are located in genome tracks of European origin, pointing to the extensive very recent admixture with Europeans. In the few generations after the admixture event that introduced the deleterious variants, natural selection might have been less effective in removing these harmful alleles from the population, mainly because of the low effective population size of Roma (Mendizabal et al. 2013). Furthermore, the effects of Roma multiple bottlenecks and consanguinity left a signature on both the Roma high number of long RoH (Figure 3E and SM Figure 7) and on the long IBD chunks between Roma individuals (SM Figure 8) (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). The effects of endogamy are stronger on RoH than on IBD (Severson et al. 2019), suggesting Balkan Roma as the migrant group that suffered less bottlenecks than the other migrant groups, in agreement with the Balkan peninsula being the cradle of Roma people currently living in Europe (Liégeois 1994; Fraser 1995; Gresham et al. 2001; Mendizabal et al. 2012; Martínez-Cruz et al. 2016)

Our analysis of complete genomes shed light on the Roma demographic history and their Diaspora from India to Europe. The small group of Roma ancestors that left Northwestern India were already a different population from their common ancestor with Punjabi when the Diaspora started. At the beginning of the Diaspora, proto-Roma underwent a strong bottleneck that reduced their effective population size to less than half their ancestral effective population size, but, along the route to Europe, Roma ancestors admixed with host populations of Middle Eastern highlands. In a few generations, 50% of the Roma ancestral component was replaced by recent European admixture, increasing their effective population size and thus compensating their previous loss of genetic diversity. The signature of their South Asian origin is still present, despite the strong admixture, as well as the signature of the initial and following bottlenecks, reflected by the high long RoH and IBD load of Roma. It is common to consider Roma as an isolated population, with reduced on no genetic and cultural exchanges with their close neighbors,

but our study showed that, at least in the recent past, Roma people have admixed at a high rate with non-Roma people all along their Diaspora route.

Materials and methods

Samples

DNA samples were collected from 40 volunteers, self-defined as Roma that belong to four main migrant groups: 10 Balkan, 5 Vlax, 10 Romungro, and 15 North/Western Roma, and coming from five countries (SM Table 1). All volunteers declared that their eight grandparents were self-defined as Roma. We tested for relatedness using Vcftools (Danecek et al. 2011) --relatedness and retained all the individuals who were less related than third degree cousins. Geographical positions of these samples are shown in Figure 1A (Kahle and Wickham 2013) and detailed in SM Table 1. All samples were collected with informed consent from the participants under the approval of the IRB of the CEIC-Parc Salut Mar 2016/6723/I. Whole genome sequencing data were obtained using Illumina HiSeqX sequencing platform, at average coverage of ~30X. We merged our data with other samples of Roma (6 additional Romanian Vlax individuals) and non-Roma (109 individuals) origin (see SM Table 2). After quality control (FastQC, Brabraham Bioinformatics), all the individuals were mapped against the human reference genome GRCh38 using BWA mem (version 0.7.15)(Li and Durbin 2009). After filtering according to the GATK Best Practices guide (DePristo et al. 2011), variant calling was performed using GATK (McKenna et al. 2010). We performed an extra filter on coverage using a home-made python script to deal with differences in the average coverage of the samples (individual filter for coverage from half to twice the average coverage) to retain only those variants that had coverage between half and twice the individual average coverage.

Population structure

We used vcftools (Danecek et al. 2011) to extract biallelic SNPs. LD pruning (window size of 5 SNPs, step size of 5 SNPs, and variance inflation factor of 2) and missing data filtering were performed using plink 1.9 (Chang et al. 2015). For the analysis including the data of the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015), our data were lifted over to GRCh37 human reference genome using picardtools v1.139 LiftoverVcf (Broad Institute 2017). To explore the relationship between populations, we used the EIGENSOFT smartpca (Patterson et al. 2006; Price et al. 2006) to run principal component analysis (PCA). To explore Roma ancestry components, we run the unsupervised clustering algorithm of ADMXITURE (Alexander et al. 2009). All plots were performed using R (R Core Team 2013).

Endogamy and Ne estimation

We estimated Runs of Homozygosity (RoH) as in Serra-Vidal et al. (2019) using plink 1.9 (Chang et al. 2015), with the parameters --homozyg --homozyg-snp 50 --homozyg-kb 100 --homozyg-density 40 --homozyg-gap 2000. We calculated the cumulative RoH length and number of RoH fragments per individual, and compared the distribution within and between populations using R (R Core Team 2013).

We identified fragments identical by descent (IBD) between individuals, after removing all variants that have more than 20% missing data (--geno 0.2) and without pruning for LD. To call IBD blocks, we used IBDSeq r1206 (Browning and Browning 2013b), with default values on the data lifted over to GRCh37, and using the genetic map provided by the authors on GRCh37 human genome reference. The distribution of the individual pairwise IBD total length was estimated and plotted with R (R Core Team 2013). We extracted IBD segments uniquely shared between two specific groups, excluding all those fragments that intersected between Roma and any other continental area (Europe – EUR, India – IND, Pakistan – PAK and Middle Eastern – ME) using bedtools options (Quinlan and Hall 2010).

We estimated the changes in the effective population size (Ne) in the last 200 generations, with 95% CI, using the IBDseq r1206 (Browning and Browning 2013b) and IBDNe v. 19Sep19.268 (Browning and Browning 2015). We set the minimum IBD segment length to 2 cM, and the maximum number of estimated generations to 200. To increase the number of samples of Indian groups in historical Ne calculation, we merged all Northern Indian (BEN, PUN, UBR, RAJ) and all Southern Indian (ILA, VLR) samples despite being from different, unrelated populations.

Roma admixture history

We modeled the admixture population history of Roma using the admixture graph algorithm implemented in qpGraph, Admixtools package (Patterson et al. 2012). We ran the admixture graph model analysis on our dataset including YRI and CHB populations from the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015). For all scenarios tested, we first generated a skeleton graph that included 10 randomly selected YRI as outgroup, EUR (our European individuals), 10 randomly selected CHB, and populations from North (Punjabi - PUN) and South (Irula - ILA) India. We then added Roma and refined the graph in order to find a model that best fits our data. To fit the data, the model must be without outliers, so that all the expected F statistics are not significantly different from the observed F statistics (Z score < |3|).

Outgroup *f3*-statistic and *D*-statistic

To test the shared drift between Roma and the other populations from a common outgroup, we applied the 3 population test (*f3*) using the same dataset as in the admixture graph. We used *qp3pop* package in Admixtools

(Patterson et al. 2012) in the form of outgroup *f3*-statistic: *f3* (YRI; Roma, X), where YRI are 10 randomly selected Yoruba individuals from the 1000 Genome Project (The 1000 Genomes Project Consortium et al. 2015), X is any other population in the dataset and Roma are the Roma as a single population or the four Roma migrant groups.

D-statistics were calculated using the *qpDstat* package in Admixtools (Patterson et al. 2012) in the form *D* (E1,E2; CHR, R) or in the form *D* (E,CHB;R1, R2), where E stands for European group and R for Roma migrant group, and CHB are 10 randomly selected Chinese from Beijing individuals from the 1000 Genome Project (The 1000 Genomes Project Consortium et al. 2015).

Ancestry identification of disease associated mutations

We performed a literature research for SNPs related to diseases in Roma people using the following terms: “Roma”, “Gypsy”, “mendelian”, “disease”, “genetic”, and “metabolic syndrome”. The genomic position of the mutations was determined using the UCSC Genome Browser. We estimated the frequency of each mutation in Roma and non-Roma using vcftools (Danecek et al. 2011). To estimate the local ancestry around the 11 mutations associated with Mendelian diseases we found in our Roma dataset, we performed a local ancestry analysis. First, we phased the variants using Beagle version 09Feb16.2b7 (Browning and Browning 2013a), with default parameters. Then, we assigned local ancestry using RFMix v1.5.4 (Maples et al. 2013), 5 EM iterations and forward-backward threshold set to 0.8, on the whole chromosome and the information of the region around the mutation was extracted. The ancestry of the haplotypes was inferred setting Europeans and South Asians (SM Table 2) as source individuals and Roma as target individuals.

Roma demographic inference

We explored Roma demographic history using the Approximate Bayesian Computation (ABC) (Beaumont et al. 2002; Beaumont 2010). We tested 12 scenarios that varied in the branching structure within Roma (a series of subsequent bottlenecks or a first split followed by two further splits between migrant groups) (SM Figure 11). We run 100,000 simulations per model using fastSimCoal V2.6 (Excoffier and Foll 2011). Summary statistics (SM Table 9) were calculated using plink (Chang et al. 2015), Admixtools (Patterson et al. 2012) and R(R Core Team 2013) on both simulated and observed data. Accuracy and parameter estimation were calculated on R using the approach implemented in the *abc* library from (Csilléry et al. 2012). See Supplementary Note 2 for extended materials and methods on the ABC approach.

Bibliography

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals.

Genome Res. 19:1655–1664.

Alfonso-Sánchez MA, Espinosa I, Gómez-Pérez L, Poveda A, Rebato E, Peña JA. 2018. Tau haplotypes support the Asian ancestry of the Roma population settled in the Basque Country. Heredity (Edinb). 120:91–99.

Álvarez A, Del Castillo I, Villamar M, Aguirre LA, González-Neira A, López-Nevot A, Moreno-Pelayo MA, Moreno F. 2005. High prevalence of the W24X mutation in the gene encoding connexin-26 (GJB2) in Spanish Romani (gypsies) with autosomal recessive non-syndromic hearing loss. Am. J. Med. Genet. 137 A:255–258.

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. 19:326–333.

Bánfai Z, Melegh BI, Sümegi K, Hadzsiev K, Miseta A, Kásler M, Melegh B. 2019. Revealing the genetic impact of the Ottoman occupation on ethnic groups of East-Central Europe and on the Roma population of the area. Front. Genet. 10:1–11.

Beaumont M a. 2010. Approximate Bayesian Computation in Evolution and Ecology. Annu. Rev. Ecol. Evol. Syst. 41:379–406.

Beaumont M a., Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Biagini SA, Solé-Morata N, Matisoo-Smith E, Zalloua P, Comas D, Calafell F. 2019. People from Ibiza: an unexpected isolate in the Western Mediterranean. Eur. J. Hum. Genet. [Internet]:941–951. Available from: http://dx.doi.org/10.1038/s41431-019-0361-1

Brabraham Bioinformatics. FastQC. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Broad Institute. 2017. Picard Tools. Available from: http://broadinstitute.github.io/picard

Browning BL, Browning SR. 2013a. Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. 93:840–851.

Browning BL, Browning SR. 2013b. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194:459–471.

Browning SR, Browning BL. 2015. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum. Genet. 97:404–418.

Chaix R, Austerlitz F, Morar B, Kalaydjieva L, Heyer E. 2004. Vlax Roma history: what do coalescent-based

methods tell us? Eur. J. Hum. Genet. 12:285–292.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4:1–16.

Csilléry K, François O, Blum MGB. 2012. Abc: An R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3:475–479.

Danecek P, Auton A, Abecasis GR, Albers C a, Banks E, DePristo M a, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

DePristo MA, Banks E, Poplin RE, Garimella K V., Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using. Nat. Genet. 43:491–498.

Excoffier L, Foll M. 2011. fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27:1332–1334.

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.Chaix R, editor. PLoS Genet. 15:e1008417.

Fraser A. 1995. The Gypsies. 2nd Editio. John Wiley & Sons Ltd

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. 69:1314–1331.

Gusmão A, Gusmão 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. 72:215–227.

Gusmão A, Valente C, Gomes V, Alves C, Amorim A, Prata MJ, Gusmão L. 2010. A genetic historical sketch of european gypsies: The perspective from autosomal markers. Am. J. Phys. Anthropol. 141:507–514.

Hancock IF. 1995. A Handbook of Vlax Romani. Slavica

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A Genetic Atlas of Human Admixture History. Science 343:747–751.

Irwin J, Egyed B, Saunier J, Szamosi G, O'Callaghan J, Padar Z, Parsons TJ. 2007. Hungarian mtDNA population databases from Budapest and the Baranya county Roma. Int. J. Legal Med. 121:377–383.

Juyal G, Mondal M, Luisi P, Laayouni H, Sood A, Midha V, Heutink P, Bertranpetit J, Thelma BK, Casals F. 2014. Population and genomic lessons from genetic analysis of two Indian populations. Hum. Genet. 133:1273–

1287.

Kahle D, Wickham H. 2013. ggmap: Spatial visualization with ggplot2. R J. 5:144–161.

Kalaydjieva L, Gresham D, Calafell F. 2001. Genetic studies of the Roma (Gypsies): a review. BMC Med. Genet. 2:5.

Kalaydjieva L, Morar B, Chaix R, Tang H. 2005. A newly discovered founder population: The Roma/Gypsies. BioEssays 27:1084–1094.

Klarić IM, Salihović MP, Lauc LB, Zhivotovsky LA, Rootsi S, Janićijevic B. 2009. Dissecting the molecular architecture and origin of bayash romani patrilineages: Genetic influences from south-asia and the balkans. Am. J. Phys. Anthropol. 138:333–342.

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

Liégeois J-P. 1994. Roma, Gypsies, Travellers. Revised Ed. (Strasbourg: Councile of Europe, editor.). Croton-on-Hudson, N.Y.: Manhattan Publishing Company, 468 Albany Post Road, P.O. Box 850, Croton-on-Hudson, NY 10520.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538:201–206.

Malyarchuk BA, Grzybowski T, MV D, Czarny J, Miscicka-Sliwka D, Malyarchuk BA, Grzybowski T, Derenko M V., Czarny J, Miścicka-Śliwka D. 2006. Mitochondrial DNA diversity in the Polish Roma. Ann. Hum. Genet. 70:195–206.

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. 93:278–288.

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. 24:937–943.

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. 20:1297–1303.

Melegh BI, Banfai Z, Hadzsiev K, Miseta A, Melegh B. 2017. Refining the South Asian Origin of the Romani

people. BMC Genet. 18:82.

Mendizabal I, Lao O, Marigorta UM, Kayser M, Comas D. 2013. Implications of Population History of European Romani on Genetic Susceptibility to Disease. Hum. Hered. 76:194–200.

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. 22:2342–2349.

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:e15988.

Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, Comas D, Laayouni H, Li Q, Majumder PP, et al. 2016. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nat. Genet. 48:1066–1070.

Moorjani P, Patterson N, Loh PR, Lipson M, Kisfali P, Melegh BI, Bonin M, Kádaši Ľ, Rieß O, Berger B, et al. 2013. Reconstructing Roma History from Genome-Wide Data. PLoS One 8:e58633.

Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D, Singh L. 2013. Genetic evidence for recent population mixture in India. Am. J. Hum. Genet. 93:422–438.

Morar B, Gresham D, Angelicheva D, Tournev I, Gooding R, Guergueltcheva V, Schmidt C, Abicht A, Lochmuller H, Tordai A, et al. 2004. Mutation history of the roma/gypsies. Am. J. Hum. Genet. 75:596–609.

Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, Bhavani GS, Girisha KM, Mustak MS, Srinivasan S, et al. 2017. The promise of discovering population-specific disease-associated genes in South Asia. Nat. Genet. 49:1403–1407.

Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M, et al. 2019. The formation of human populations in South and Central Asia. Science 365:eaat7487.

Pathak AK, Kadian A, Kushniarevich A, Montinaro F, Mondal M, Ongaro L, Singh M, Kumar P, Rai N, Parik J, et al. 2018. The Genetic Ancestry of Modern Indus Valley Populations from Northwest India. Am. J. Hum. Genet. 103:918–929.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient Admixture in Human History. Genetics 192:1065–1093.

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

Peričić M, Klarić IM, Lauc LB, Janićijević B, Dordević D, Efremovska L, Rudan P. 2005. Population genetics of 8 Y chromosome STR loci in Macedonians and Macedonian Romani (Gypsy). Forensic Sci. Int. 154:257–261.

Price AL, Patterson N, Plenge RM, Weinblatt ME, Shadick N a, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. http://www.r-project.org/ [Internet]. Available from: http://www.r-project.org/

Rai N, Chaubey G, Tamang R, Pathak AK, Singh VK, Karmin M, Singh M, Rani DS, Anugula S, Yadav BK, et al. 2012. The Phylogeography of Y-Chromosome Haplogroup H1a1a-M82 Reveals the Likely Indian Origin of the European Romani Populations. PLoS One 7:e48477.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461:489–494.

Salihović MP, Barešić A, Klarić IM, Cukrov S, Lauc LB, Janićijević B. 2011. The role of the Vlax Roma in shaping the European Romani maternal genetic history. Am. J. Phys. Anthropol. 146:262–270.

Serra-Vidal G, Lucas-Sanchez M, Fadhlaoui-zid K, Bekada A, Zalloua P, Comas D. 2019. Heterogeneity in Palaeolithic Population Continuity and Neolithic Expansion in North Afria. Curr. Biol. 29:1–7.

Severson AL, Carmi S, Rosenberg NA. 2019. The effect of consanguinity on between-individual identity-by-descent sharing. Genetics 212:305–316.

Tätte K, Pagani L, Pathak AK, Kõks S, Ho Duy B, Ho XD, Sultana GNN, Sharif MI, Asaduzzaman M, Behar DM, et al. 2019. The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. Sci. Rep. 9:1–9.

The 1000 Genomes Project Consortium, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, et al. 2015. A global reference for human genetic variation. Nature 526:68–74.

Yelmen B, Mondal M, Marnetto D, Pathak AK, Montinaro F, Gallego Romero I, Kivisild T, Metspalu M, Pagani L. 2019. Ancestry-Specific Analyses Reveal Differential Demographic Histories and Opposite Selective Pressures in Modern South Asian Populations. Mol. Biol. Evol. 36:1628–1642.

Zalán A, Béres J, Pamjav H. 2010. Paternal genetic history of the Vlax Roma. Forensic Sci. Int. Genet. 5:109–113.

Acknowledgements and funding

Acknowledgments

We would like to thank all the DNA donors and volunteers who made this study possible. We thank Mònica Vallés for technical support.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness (grant number CGL2016-75389-P (MINEICO/FEDER, UE) and “Unidad de Excelencia María de Maeztu” (funded by AEI – CEX2018-000792-M) to DC and FC; and Agència de Gestió d’Ajuts Universitaris i de la Recerca (Generalitat de Catalunya, grant 2017SGR00702). NF-P was supported by a FPU17/03501 fellowship.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics approval and consent to participate

All samples were collected with informed consent from the participants under the approval of the IRB of the CEIC-Parc Salut Mar 2016/6723/I.

Data availability

All data generated during the current study are available upon request at EGA repository, under the accession number EGAS00001004287.

Competing interests

The authors declare no competing interests.

Author contributions

DC and EB conceived study. EB and CG-F did the preprocessing and quality control of the data. RSV performed the local ancestry and diseases analysis. EB analyzed the data. BD, ESS, VK, HM, HP, MGN, JP provided samples. GL designed and validated the ABC analysis. EB wrote the first draft of the manuscript. NF-P, DC, FC,

CG-F and EB interpreted the results and provided comparative discussions. All authors contributed to the writing and editing of the final manuscript. All the authors approved the final version of the manuscript.

Tables

Table 1. Estimations of parameters with good accuracy under 2b4m model.

		Neuralnet*			Neuralnet_LogTransform*		
Parameter	**Description**	**Mean**	**2.5%CI**	**97.5% CI**	**Mean**	**2.5%CI**	**97.5% CI**
NBEA	Ne common ancestor of all Eurasians	3,187	3,109	3,251	2,427	2,212	2,601
NIND	Ne Northern Indian population (Punjabi)	3,497	317	7,207	2,027	1,590	2,506
N1=NIND/bot1a	Ne Roma founders at the out of India	1,536	188	2,387	977	1,091	853
bot1a	Mutual of the bottleneck Roma had in the out of India	2.28	1.69	3.02	2.07	1.46	2.94
TbotBAIND[a]	Time of split from common ancestors of Roma and Northern Indian population	2,126	475	3,760	1,632	1,103	2,099
TsplitEUIN[a]	Time of split from the common ancestors of Ancestral Northern Indian and Europeans	31,901	23,821	40,087	11,656	8,478	14,675
TsplitEUIS[a]	Time of split from the common ancestors of Ancestral Southern Indians and Europeans	31,509	28,583	37,273	38,813	35,717	44,392
	[a] : years, generation time = 29 years						

Mean and 95% confidence intervals of the posterior distributions were estimated using neural network logistic estimation algorithm, with and without log transformation.

Figures

Figure 1. Map of the samples, Principal Component Analysis (PCA) and ADMIXTURE. (A) Map of the samples used in this study. Sampling locations are approximated. EUR, European non-Roma; INDN, North India; INDS, South India; ME, Middle East; PAK, Pakistan; ROMAB, Balkan Roma; ROMAN, North/Western Roma; ROMAR, Romungro Roma; and, ROMAV, Vlax Roma. (B) PCA of Roma samples together with the rest of the dataset and the 1000 Genomes Project samples from AFR, EAS, EUR and SAS. (C) Admixture analysis of Roma samples together with the rest of the dataset and the 1000 Genomes Project samples from AFR, EAS, EUR and SAS, showing Roma with their two main ancestral components, SAS and EUR. For population codes see SM Table 2.

Figure 2. Outgroup *f3*-statistic statistics. *f3*-statistics was calculated for Roma in the form *f3* (ROMA,X; YRI), where X is any population on the Y axis and YRI is the outgroup (Yoruba). For population codes see SM Table 2.

Figure 3. Identity By Descent (IBD) and Run of Homozygosity (RoH). (A) Average pairwise cumulative length of uniquely shared IBD segments between Roma individuals and the individuals from that specific population, excluding the segments that intersected with populations of other areas (EUR, in green; IND, in turquoise; ME, in brown; and PAK, in pink, SM Table 2). (B) Average pairwise cumulative length of segments in IBD uniquely shared between Roma and Europe, by migrant group. (C) Average pairwise cumulative length of segments in IBD uniquely shared between Roma and Indian populations, by migrant group. In figure B and C, the letters on top indicate whether the distributions are not significantly different in a rank pairwise comparison: same letter means no significant difference. (D) Number and total length of Runs of Homozygosity (RoH) tracts larger than 1Mb, per population: ME* = ARM, EGY, IRN, IRQ, IRJ, BED, DRU, PAL, SAM, JOR, TUR; PAK* = BAL, BRA, BRU, MAK, HAZ, KAL, PAT, SIN. * Individuals from different populations were grouped together, which may decrease the average load and length of IBD fragments.

Figure 4. Admixture graph model for Roma complex admixture scenario. The estimated was Zscore=-1.444 between observed and expected F statistics. Populations in capital letters = sampled populations; small case = unsampled populations. Straight arrows = drift; dashed arrow = admixture, with corresponding admixture proportions between the two populations.

TO BE REMOVED AFTER CREATING THE FINAL PDF TO UPLOAD THE WORD DOCUMENT WITHOUT THEM

Figure 1

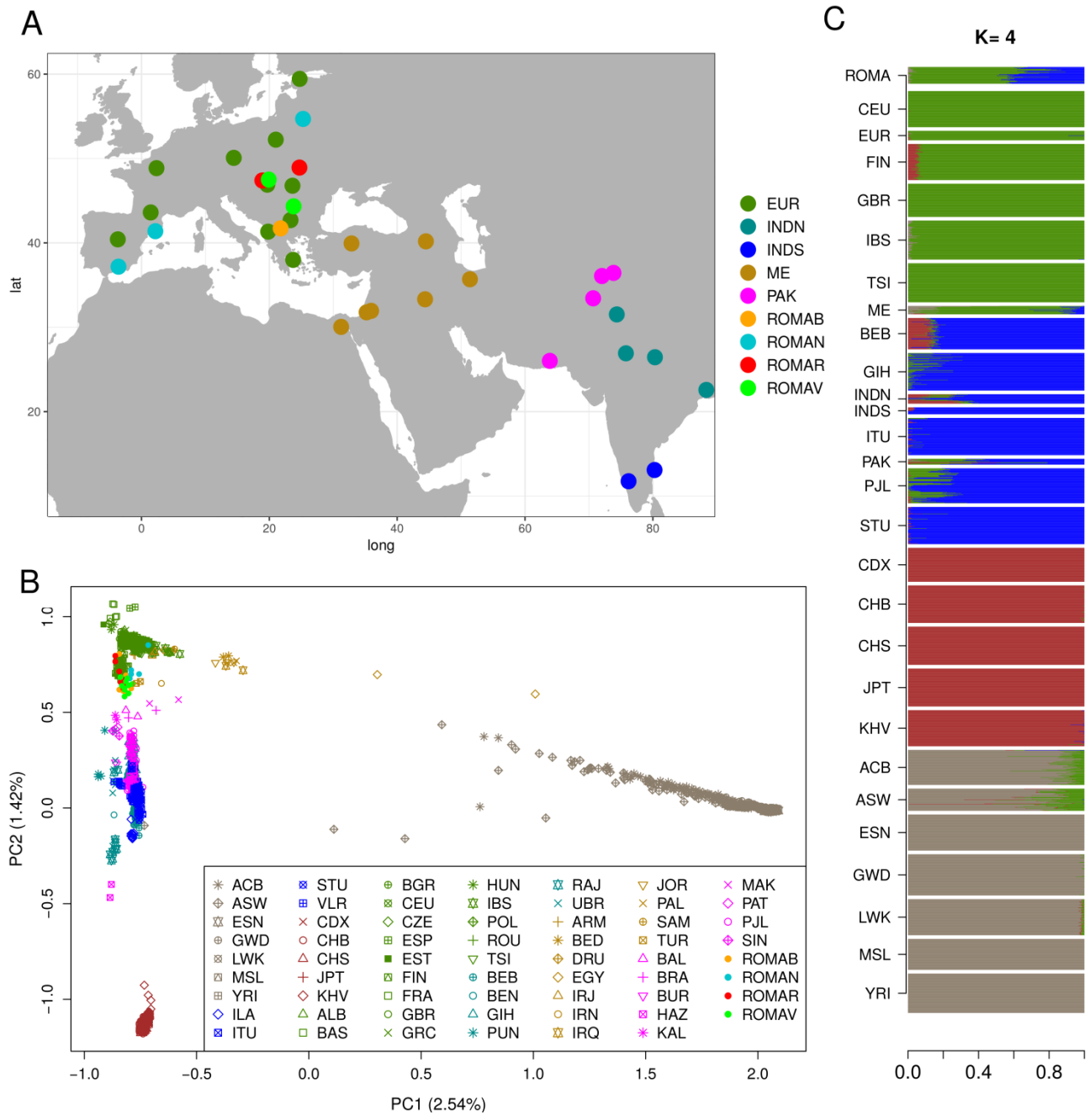


Figure 2

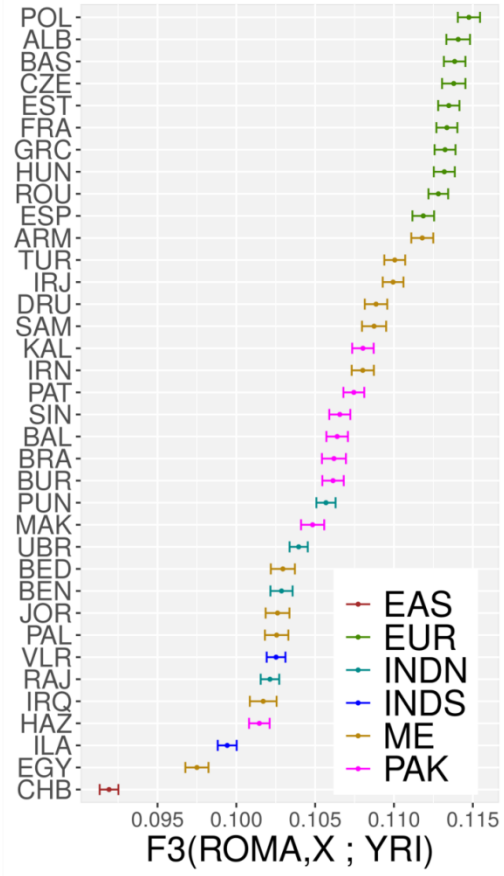

POL
ALB
BAS
CZE
EST
FRA
GRC
HUN
ROU
ESP
ARM
TUR
IRJ
DRU
SAM
KAL
IRN
PAT
SIN
BAL
BRA
BUR
PUN
MAK
UBR
BED
BEN
JOR
PAL
VLR
RAJ
IRQ
HAZ
ILA
EGY
CHB
EAS
EUR
INDN
INDS
ME
PAK
0.095
0.100
0.105
0.110
0.115
F3(ROMA,X ; YRI)

Figure 3

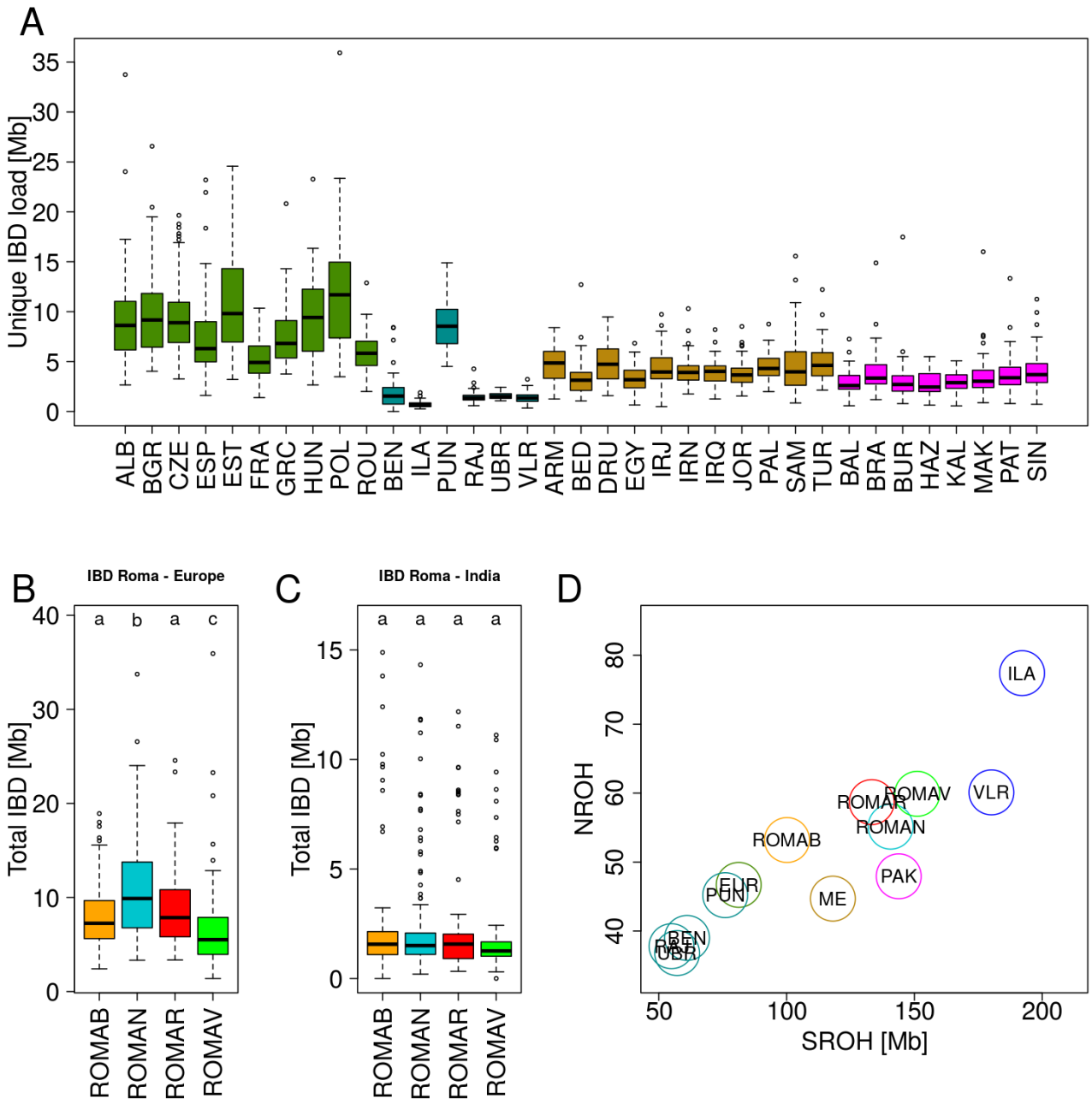
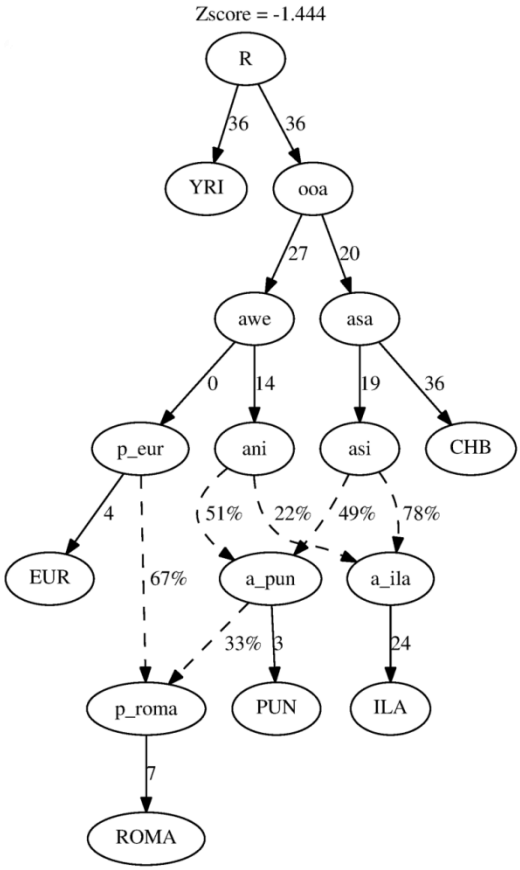

A
Unique IBD load [Mb]
0
5
10
15
20
25
30
35
ALB BGR CZE ESP EST FRA GRC HUN POL ROU BEN ILA PUN RAJ UBR VLR ARM BED DRU EGY IRJ IRN IRQ JOR PAL SAM TUR BAL BRA BUR HAZ KAL MAK PAT SIN
B
IBD Roma - Europe
Total IBD [Mb]
0
10
20
30
40
a b a c
ROMAB ROMAN ROMAR ROMAV
C
IBD Roma - India
Total IBD [Mb]
0
5
10
15
a a a a
ROMAB ROMAN ROMAR ROMAV
D
NROH
40
50
60
70
80
SROH [Mb]
50
100
150
200
ILA
VLR
ROMAV
ROMAR
ROMAN
ROMAB
PAK
ME
FIN
PUN
BEN

Figure 4


Zscore = -1.444
R
36
36
YRI
ooa
27
20
awe
asa
0
14
19
36
p_eur
ani
asi
CHB
4
51%
22%
49%
78%
EUR
67%
a_pun
a_ila
33%
3
24
p_roma
PUN
ILA
7
ROMA