Reconstructing the population history of European Romani from genome-wide data

Isabel Mendizabal[1,*], Oscar Lao[2,*], Urko M. Marigorta[1], Andreas Wollstein[2,#], Leonor Gusmão[3,4], Vladimir Ferak[5], Mihai Ioana[6,7], Albena Jordanova[8,9], Radka Kaneva[9], Anastasia Kouvatsi[10], Vaidutis Kučinskas[11], Halyna Makukh[12], Andres Mestpalu[13], Mihai G. Netea[14,15], Rosario de Pablo[16], Horolma Pamjav[17], Dragica Radojkovic[18], Sarah J.H. Rolleston[19], Jadranka Sertic[20,21], Milan Macek Jr.[22], David Comas[1,**,$], and Manfred Kayser[2,**,$]

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

[2]Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands

[3]IPATIMUP – Institute of Pathology and Molecular Immunology of the University of Porto, Porto, Portugal

[4]Medical and Human Genetics Laboratory, and Molecular Biology and Genetics Post-graduate Program, Federal University of Pará (UFPA), Belém, Pará, Brazil

[5]Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia

[6]University of Medicine and Pharmacy Craiova, Craiova, Romania

[7]University of Medicine and Pharmacy Carol Davila Bucharest, Bucharest, Romania

[8]VIB Department of Molecular Genetics, University of Antwerp, Antwerp, Belgium

[9]Department of Chemistry and Biochemistry, Molecular Medicine Center, Medical University Sofia, Sofia, Bulgaria

[10]Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece

[11]Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

[12]Institute of Hereditary Pathology of the Ukrainian Academy of Medical Sciences, Lviv, Ukraine

[13]Estonian Genome Center, University of Tartu, Tartu, Estonia

[14]Department of Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

[15]Nijmegen Institute for Infection, Inflammation and Immunity, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands
[16]Servicio de Inmunología, Hospital Universitario Puerta de Hierro, Madrid, Spain
[17]DNA Laboratory, Institute of Forensic Medicine, Network of Forensic Science Institutes, Budapest, Hungary
[18]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
[19]Institute of Medical Genetics, University Hospital of Wales, Cardiff, Wales, United Kingdom
[20]Clinical Institute of Laboratory Diagnosis, Zagreb University Hospital Centre, Zagreb Croatia
[21]Department of Chemistry, Biochemistry and Clinical Biochemistry, School of Medicine, University of Zagreb, Zagreb, Croatia
[22]Department of Biology and Medical Genetics, University Hospital Motol and the 2nd Faculty of Medicine, Charles University, Prague, Czech Republic

* These authors contributed equally to this work
** These authors contributed equally to this work
# Current affiliation: Section of Evolutionary Biology, Department of Biology II, University of Munich LMU, Planegg-Martinsried, Germany
$ Correspondence: david.comas@upf.edu or m.kayser@erasmusmc.nl

Running head:
Genetic history of European Romani

Highlights
- The Romani Diaspora originated in north/north-west India around 1.5 kya
- The European spread of the Romani people was via the Balkans starting ~0.9 kya
- Differential footprints of bottlenecks, endogamy and admixture are found in the Romani genomes across Europe

Summary

The Romani, the largest European minority group with approximately 11 million people [1], constitute a mosaic of languages, religions and lifestyles while sharing a distinct social heritage. Linguistic [2] and genetic studies [3-8] have located the Romani origins in the Indian subcontinent. However, a genome-wide perspective on Romani origins and population substructure, as well as a detailed reconstruction of their demographic history have yet to be provided. Our analyses based on genome-wide data from 13 Romani groups collected across Europe suggest that the Romani Diaspora constitutes a single initial founder population which originated in north/north-western India ~1.5 thousand years ago (kya). Our results further indicate that after a rapid migration with moderate gene flow from the Near/Middle East, the European spread of the Romani people was via the Balkans starting ~0.9 kya. The strong population substructure and high levels of homozygosity we found in the European Romani are in line with genetic isolation as well as differential gene flow in time and space with non-Romani Europeans. Overall, our genome-wide study sheds new light on the origins and demographic history of European Romani.

Results and Discussion

Previous studies analyzing the fine-scale genetic substructure of Europeans [9-11] did not include the Romani, even though they are the largest minority group in Europe. Furthermore, the location, dating and magnitude of their suggested Out-of-India Diaspora, as well as their relationships with other populations, remain elusive. To address these issues, we studied the genome-wide diversity of the Romani people by analyzing ~800,000 single nucleotide polymorphisms (SNPs) using the Affymetrix 6.0 platform in 152 individuals from 13 Romani groups from eastern, western and northern parts of Europe (see Figure 1).

European Romani genetic diversity in the worldwide context

First, we explored the genetic relationships of the European Romani with other worldwide populations using previously published genome-wide datasets (4,587 individuals and 51,328 shared SNPs, see Reference datasets section in Supplemental Experimental Procedures). In a first classical multidimensional scaling (MDS or principal coordinates analysis) [12] based on identity-by-state (IBS) distances, worldwide individuals tend to be distributed in the first two dimensions as in [13, 14], with European Romani located with other west Eurasian populations (Figure 2 A and Figure S1 A). We then performed a second MDS focusing on west Eurasians using balanced sample sizes and geographic coverage (Figure 2 B and Figure S1 B). The first dimension separates Indians from non-Romani Europeans, Caucasus and Middle East individuals, and locates in-between the Romani Europeans, Central Asians and Pakistanis. The second places European Romani close to non-Romani Europeans with several Romani individuals included within the latter, which could be indicative of recent admixture.

Next, we constructed a neighbor-joining tree [15] based on F_{ST} distances [16] , using sub-Saharan Africans (Yoruba) as an out-group. All European Romani groups (except the Welsh Romani) appear on the same branch and without any non-Romani European groups (Figure S1 C), which would suggest a shared common origin of the European Romani. Welsh Romani appear to share ancestry with non-Romani Europeans and show evidence of strong genetic drift. However, putative recent admixture with other populations could modify the position of the European Romani with respect to the other populations in the tree. Therefore, we applied the ADMIXTURE clustering method [17] to estimate the membership of each individual to a range of *k* hypothetical

ancestral populations (k=2 to k=15, see Figure 2 C and Figure S1 D and E). At k=2, a longitudinal gradient on the amount of ancestry of each component is observed from India to Europe (|Spearman's rho| = 0.935, p<10-16, after excluding European Romani; Figure S1 F). European Romani show a lower frequency of the main ancestral component in Indians (dark blue) relative to populations from Central Asia and Pakistan (28% vs. 47%, p<10-16, Mann-Whitney test), and higher than Caucasus, Middle East and non-Romani European populations (28% vs. 9%, p<10-16, Mann-Whitney test). This result would suggest that the origin of the European Romani could be located in Central/South Asia (Pakistan and India). Notably, the main ancestry component present in Middle Easterners at k=3 (Figure 2 C, in dark green) shows the lowest average in the European Romani, followed by the Indian populations (3.6% and 6.3%, respectively). This result may indicate a low genetic contribution to the European Romani from the Near/Middle East. At k=5, an ancestral component present mainly in European Romani emerges (Figure 2 C, in red). At k=8 (well-supported k, see Figure S1 G) this ancestry component (red) is almost absent from all non-Romani individuals (on average 1.52%; 95% confidence interval (C.I.) = 0-5.5%). At this k, almost 25% of all European Romani show considerable amounts (above 30%) of the component mainly present in non-Romani Europeans (Figure 2 C, in gray). Further population substructure within the European Romani is observed at k=13. The new component (Figure 2 C, in black) is mainly present in Croatian Romani (average ~76%), less frequent in the remaining Balkan Romani (average 23% across Bulgarian, Serbian, and Greek Romani) and rare in Romani groups from northern and western Europe (e.g. 6.7% in Baltic and Iberian Romani).

Genetic substructure within the European Romani

To further explore the genetic affinities within European Romani, we ran ADMIXTURE only on the 152 Romani individuals using 277,109 LD pruned SNPs. At k=2 and k=3, Welsh (in gray, see Figure S2 A and B for cross-validation) and Croatian Romani (in dark green) separate from other Romani groups. Further k values tend to distinguish Ukrainian (at k=4), Balkan versus non-Balkan Romani (at k=5) and within the latter, more subtle structure between Central European, North (Baltic) and West (Iberian) Romani populations (at k=6 and k=7 respectively) is observed. The first two dimensions of an MDS on the same dataset separate the Welsh and Croatian Romani from the remaining European Romani groups (see Figure S2 C). The first two

dimensions of an additional MDS after removing individuals with a large percentage of non-Romani ancestry (>20% of gray component in ADMIXTURE at k=5 in Figure 2 C), separate Croatian and Ukrainian Romani, respectively. Notably, Romani individuals from each country tend to cluster together (see Figure S2 D). Supporting this observation, an analysis of molecular variance (AMOVA [18]) using European Romani groups explains 2.71% of the genetic variance ($p<0.0005$). This value is six times larger than that between non-Romani European groups (0.47%; $p<0.0005$), which would suggest a relatively strong genetic isolation of the various European Romani groups tested. Furthermore, in contrast to the association between genetic and geographic distances previously described in non-Romani Europeans [9, 11, 19], we observe here a weak and non-significant correlation between the MDS coordinates and the population geographic coordinates in the European Romani (Pearson correlation $r^2=0.11$ after excluding Welsh Romani from MDS analysis, Mantel test $p = 0.06$ based on 1,000 resamples).

Furthermore, we checked the correlation between pairwise F_{ST} distances [16] and the dates of first records for the presence of the Romani people in each sampled European country. The strongest correlations were observed when considering genetic distances of each Romani population to one of the Balkan Romani populations (i.e. Serbia and Bulgaria) whereas non-Balkan Romani show weak correlations (see Figure 3 and Figure S2 E). In agreement with previous studies [4, 8, 20], this finding would suggest a series of founder colonizations from the Balkan area (Out-of-Balkans) during the Romani European dispersal (see next section for further evidence).

Demographic history of European Romani inferred from approximate Bayesian computation

To test hypotheses about the origin of the European Romani and to estimate the parameters of their demographic history, we performed three approximate Bayesian computation (ABC [21]) analyses. The basic common model considers a proto-Romani population that splits from a given population of the Indian subcontinent (Pakistan and India) and can admix with a hypothetical (unsampled) Central Asian, or Near/Middle Eastern population, as well as with non-Romani Europeans after arriving in Europe (see ABC in Supplemental Experimental Procedures). To avoid any influence in parameter estimation from chip array data [22], we used the correction for Affymetrix data from [23] (see Figure S3 A) and thus restricted our ABC analyses to populations with a

sample size ≥ 5 individuals genotyped on this platform (see Reference datasets in Supplemental Experimental Procedures).

In the first ABC analysis, we attempted to identify the current Romani population that is most genetically similar to the putative founder population of all European Romani groups. For all pairwise comparisons of Romani populations, we computed the Bayes factor between two demographic models, with one as the source and the other as the descendant population, and vice versa in the second model (see Figures S3 B and C). The Bulgarian Romani showed the largest number of comparisons with Bayes factor >1.5 for being the founder population in all comparisons (12 out of the 12 possible pairwise population comparisons, Figure S3 D). This finding delimits the broader geographic area in the Balkans suggested by our previous analyses. This could be due to the fact that in the ABC analysis we are conditioning the effective population size of the parental population as being larger than the descendent one, while controlling for the presence of recent admixture with non-Romani Europeans.

In a second ABC also based on pairwise comparisons, we used the Bulgarian Romani as a proxy to locate the putative source population of the European Romani within the Indian subcontinent (see Figures S3 E and F). The genetically-similar [24] Indo-European speaking groups from north-west India (Meghawal in Rajasthan) and northern India (Kashmiri Pandit in Jammu & Kashmir), were the populations showing the largest number of comparisons with Bayes factor >1.5 (94% each, see Figure 4 A and Table S1). Despite not having samples from that area, the highlighted geographic region in India as the source area for the Romani encloses the Punjab, as suggested previously by anthropological, linguistic [2], and mtDNA [8] evidence. However, given that India is genetically heterogeneous, and endogamy plays an important role in restricting the genetic variation at a regional level and to particular caste/tribes, future dedicated sampling across linguistic and social strata in this Indian sub-region is needed to identify the actual parental population of the European Romani.

Finally, in a third ABC using Meghawal Indians as a proxy for the parental Romani population and Bulgarian and Spanish Romani as proxies for eastern and western European Romani groups (see Figure S3 G), we aimed to estimate the parameters of the Romani demographic history (see Figure 4 B; Figure S3 H and Table S2 for centrality and dispersion statistics). The date of the Out-of-India founder event was estimated at ~1.5 kya. After a strong bottleneck the proto-Romani effective population size became 47% of the parental Indian population. During the migration

towards Europe, the Romani would have undergone modest genetic admixture with the populations encountered, including Middle East, Caucasus and Central Asia (number of migrants per generation estimated to be ~2.2% of the proto-Romani population size during 13 generations, or ~330 years). Around 0.9 kya, the eastern and western European Romani would have diverged. The western European Romani would have undergone an additional bottleneck reducing their population size to 70% of that of eastern European Romani. Finally, both western and eastern European Romani would have admixed with non-Romani European populations (~4% and ~5% of migrants per generation; during ~38 generations or ~940 years). In sum, the increasing genetic distance from the Balkans and the decaying effective population sizes in western Romani point at cumulative drift events within Europe as one of the main forces driving the extensive genetic differentiation observed within the European Romani, regardless of their recent common origin.

Signatures of bottlenecks and endogamy in European Romani inferred from genomic homozygosity

A demographic history of bottlenecks and isolation is expected to leave a footprint in the levels of genomic homozygosity [25]. We investigated runs of homozygosity (ROH; [26]) in Indian, Romani and non-Romani Europeans. The shape of the distribution of the cumulative ROH in the European Romani individuals resembles that expected under a scenario of recent bottlenecks [27] (see Figure S4 A). Furthermore, we found more and longer ROH in the European Romani compared to Indians and non-Romani Europeans (see Figure S4 B and C and Table S3), including very long tracts (>20Mb) absent in non-Romani Europeans, which suggests that consanguineous marriages may be common in all European Romani groups . Interestingly, ROH statistics correlate positively with the blue and red ancestral components (k=2 and k=5 in Figure 2 C), putative Indian and Romani respectively, but negatively with the gray in k=5 (European one, see Table S4). Overall, the extensive ROH patterns in the Romani are in agreement with decreases in the Romani effective population sizes as suggested by the ABC analyses and with endogamous marriage practices. Interestingly, the Welsh Romani also show extensive ROH in their genomes. The finding of typically Indian mtDNA lineages in the Welsh Romani samples (see mtDNA haplotype classification section in Supplemental Experimental Procedures) confirms their maternal Romani origin. Thus, our data suggest that either the Welsh

Romani admixed *in situ* with non-Romani Europeans and afterwards underwent strong isolation, or that they received genetic admixture with an already isolated local population, such as the so called Native Travellers [28]. Future studies are needed to investigate possible admixture between Welsh Romani and Travellers, and any potential sex-bias in the admixture between Welsh Romani and non-Romani Europeans.

Genetic admixture dynamics between Romani and non-Romani Europeans

The demographic model used in ABC assumed a constant migration rate from European non-Romani to Romani populations (see ABC section in Supplemental Experimental Procedures). However, additional information about the timing of such an admixture event can be inferred from the length of ancestral chromosomal segments. Recent genetic migration and admixture from European non-Romani to Romani populations is expected to produce both Romani individuals with long chromosomal segments of non-Romani European ancestry as well as others without any such traces. Over time, cumulative recombination events are expected to shorten and spread these non-Romani European chromosomal tracts across Romani individuals. To identify the segments of Indian and non-Romani European ancestry in the European Romani genome, we used HapMap 3 [29] European (CEU) and Indian (GIH) individuals as proxy parental populations (see Local ancestry analyses in European Romani in Supplemental Experimental Procedures) and applied the HAPMIX [30] algorithm to detect local ancestry in admixed populations. We first performed two analyses to investigate how well HAPMIX distinguishes the ancestry of the two parental populations in the European Romani genome. First, we computed IBS distance matrices between each pair of individuals for each subset of SNPs that HAPMIX ascribes to Indian and European ancestry, and compared them. We observed that the two IBS matrices were significantly less correlated than those calculated from randomly selected SNPs (1,000 random samplings $p< 0.0005$). Second, we observed a high correlation (see Figure S4 D and E) between the averaged ancestry estimates for the Romani individuals by HAPMIX and StepPCO, an independent algorithm for local ancestry estimation [31] ($r=0.935$, p-value < 2.2e-16), as well as when comparing HAPMIX and ADMIXTURE ($r=0.93$, p-value < 2.2e-16). These observations suggest that HAPMIX identifies ancestral chromosomal segments in the Romani genomes.

We then analyzed the length of the genomic segments of non-Romani European origin. Strikingly, several Romani populations from Central Europe

(Slovakia, Hungary and Romania) and from the Balkan area (Bulgaria and Croatia) show low mean values of genetic admixture, but a few individuals present very long segments of non-Romani origin (Figure S4 F and G). This would suggest a recent and ongoing shift in the social rules of the acceptance of Romani and non-Romani couples within Romani groups. Conversely, European Romani from Lithuania, Portugal, and Spain show higher non-Romani European admixture but in shorter chromosomal tracks. This is suggestive of older patterns of genetic admixture and implies higher levels of recent genetic isolation from non-Romani Europeans in these countries. Alternatively, mixed couples may leave the Romani communities and integrate into the non-Romani societies, and thus would not be sampled from Romani groups in these countries.

Conclusions

The present study constitutes the most comprehensive survey available thus far on the genome-wide characterization and demographic history of the European Romani. Our data suggest that European Romani share a common genetic origin, which can be broadly ascribed to north/north-western India around 1.5 kya. After a modest genetic contribution from the populations encountered through their rapid Diaspora from India towards the European continent, our data indicate that the Romani dispersed from the Balkan area around 0.9 kya. We further observe evidence of secondary founding bottlenecks and small population sizes, together with isolation and strong endogamy. Our data further imply that in more recent times, temporally and geographically variable admixture events with non-Romani Europeans have left a footprint in the Romani genomes. Overall, our analyses suggest that despite the relatively short time span, the demographic history of the Romani is rich and complex. Further studies with more dedicated geographical sampling and re-sequencing data would help in defining the Indian parental population of the Romani as well as further details of their migration and subsequent history in Europe.

Experimental Procedures

DNA was isolated from blood and buccal samples collected with informed consent from 206 unrelated volunteers who self-identified as Romani (see Romani samples section in Supplemental Experimental Procedures), and genotyped on Affymetrix 6.0 arrays. After SNP quality-filtering and removal of individuals likely to be related, there were 152 samples genotyped for 807,002 autosomal SNPs for subsequent analyses. For some analyses, we merged our data with data from 4,587 worldwide individuals [9, 13, 14, 29, 32-34], and for others with data from 1,234 west Eurasian individuals, both datasets with 51,328 SNPs. For further details see supplemental information.

Supplemental Information

Supplemental Information includes additional experimental procedures and results and can be found with this article online at doi……..

Data availability

Depending on the research purpose data are available up on request for non-profit scientific research under an inter-institutional data access agreement.

Acknowledgements

We thank Jordi Camí and Francesc Valentí for their valuable help in collecting Romani samples from Spain, Natasa Petrovic for collecting Romani samples from Serbia, and Lazarus P. Lazarou for collecting Romani samples from Wales, United Kingdom. We are grateful to Mark Stoneking for his valuable comments on the manuscript. IM was supported by a PhD grant by the Basque Government (Hezkuntza, Unibertsitate eta Ikerketa Saila, Eusko Jaurlaritza, BFI107.4). OL, AW, and MK were supported by the Erasmus MC University Medical Center Rotterdam. UMM was supported by a PhD grant by Universitat Pompeu Fabra. MGN was supported by a Vici grant of the Netherlands Organization of Scientific Research. LG was supported by an Invited Professor grant from CAPES/Brazil. AM was supported by the Estonian Government grant SF0180142s08. This study was supported in parts by the Spanish Government MCINN grant CGL2010-14944/BOS to DC, the Czech Republic Ministry of Health grants CZ.2.16/3.1.00/24022 and 00064203 to MM, the Republic of Serbia Ministry of Education and Science grant ON173008 to DR, by the Belgium University of Antwerp grant IWS BOFUA 2008/23064 to AJ, and by the Portuguese Foundation for Science

and Technology (FCT) project grant PTDC/ANT/70413/2006 to LG. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Education and Science and is partially supported by FCT.

References

1. Council of Europe; Roma and Travellers Division (2010).

2. Fraser, A. ed. (1992). The Gypsies (Oxford: Blackwell Publishers).

3. Ali, M., McKibbin, M., Booth, A., Parry, D.A., Jain, P., Riazuddin, S.A., Hejtmancik, J.F., Khan, S.N., Firasat, S., Shires, M., et al. (2009). Null mutations in LTBP2 cause primary congenital glaucoma. Am J Hum Genet *84*, 664-671.

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

5. Gusmão, A., Gusmão, L., Gomes, V., Alves, C., Calafell, F., Amorim, A., and Prata, M.J. (2008). A perspective on the history of the Iberian gypsies provided by phylogeographic analysis of Y-chromosome lineages. Ann Hum Genet *72*, 215-227.

6. Kalaydjieva, L., Calafell, F., Jobling, M.A., Angelicheva, D., de Knijff, P., Rosser, Z.H., Hurles, M.E., Underhill, P., Tournev, I., Marushiakova, E., et al. (2001). Patterns of inter- and intra-group genetic diversity in the Vlax Roma as revealed by Y chromosome and mitochondrial DNA lineages. Eur J Hum Genet *9*, 97-104.

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

8. Mendizabal, I., Valente, C., Gusmao, 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.

9. Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D., et al. (2008). Correlation between genetic and geographic structure in Europe. Curr Biol *18*, 1241-1248.

10. Nelis, M., Esko, T., Magi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskackova, T., Balascak, I., Peltonen, L., et al. (2009). Genetic structure of Europeans: a view from the North-East. PLoS One *4*, e5472.

11. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98-101.

12. Cox, T.F., and Cox, M.A.A. (2001). Multidimensional Scaling. (Chapman & Hall/CRC).

13. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100-1104.

14. Lopez Herraez, D., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M.R., Gross, A., Scholz, M., and Stoneking, M. (2009). Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One *4*, e7888.

15. Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

16. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 13.

17. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res *19*, 1655-1664.

18. Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics *131*, 479-491.

19. Yang, W.Y., Novembre, J., Eskin, E., and Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. Nat Genet *44*, 725-731.

20. 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.

21. Tavare, S., Balding, D.J., Griffiths, R.C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. Genetics *145*, 505-518.

22. Wollstein, A., Lao, O., Becker, C., Brauer, S., Trent, R.J., Nurnberg, P., Stoneking, M., and Kayser, M. (2010). Demographic history of Oceania inferred from genome-wide data. Curr Biol *20*, 1983-1992.

23. Albrechtsen, A., Nielsen, F.C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol *27*, 2534-2547.

24. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. Nature *461*, 489-494.

25. Pemberton, T., J, Absher, D., Feldman, M., W, Myers, R., M, Rosenberg, N., A, and Li, J., Z (2012). Genomic Patterns of Homozygosity in Worldwide Human Populations. Am J Hum Genet *91*, 275-292.

26. McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa,

A., et al. (2008). Runs of homozygosity in European populations. Am J Hum Genet *83*, 359-372.

27. Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M., Rodriguez-Botigue, L., Ramachandran, S., Hon, L., Brisbin, A., et al. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A *108*, 5154-5162.

28. Matras, Y. ed. (2010). Romani in Britain. The afterlife of a language (Edinburgh: Edinburgh University Press).

29. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52-58.

30. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet *5*, e1000519.

31. Pugach, I., Matveyev, R., Wollstein, A., Kayser, M., and Stoneking, M. (2011). Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biol *12*, R19.

32. Behar, D.M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., et al. (2010). The genome-wide structure of the Jewish people. Nature *466*, 238-242.

33. Metspalu, M., Romero, I.G., Yunusbayev, B., Chaubey, G., Mallick, C.B., Hudjashov, G., Nelis, M., Magi, R., Metspalu, E., Remm, M., et al. (2012). Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet *89*, 731-744.

34. Yunusbayev, B., Metspalu, M., Jarve, M., Kutuev, I., Rootsi, S., Metspalu, E., Behar, D.M., Varendi, K., Sahakyan, H., Khusainova, R., et al. (2011). The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol *29*, 359-365.

Figure legends

Figure 1. Sampling origin of the European Romani samples analyzed in the present study. Geographic origin of the European Romani samples (red dots) analyzed in the present study. Numbers in parentheses indicate sample sizes. Gray shades represent Romani population estimates by country according to the Council of Europe (Roma and Travellers Division, 2010; http://www.coe.int/t/dg3/romatravellers/default_en.asp). Blue numbers indicate the approximate dates for the arrival of the Romani in each country (see Historical data in Supplemental Experimental Procedures).

Figure 2. A) Two-dimensional plot of a multidimensional scaling analysis including European Romani and other worldwide populations and B) European Romani (filled circles) and west Eurasians individuals (empty circles), using a balanced sample sizes and geographic coverage (see Reference Datasets in Supplemental Experimental Procedures). Same plots with population labels are shown in Figures S1 A and B. **C)** ADMIXTURE analysis at k=2, k=3, k=5, k=8 and k=13 ancestral components using the same individuals in panel B. Each vertical bar represents an individual and the proportion of each individual to the k ancestral components is shown in colors. See Figures S1 D and E for more k-s and the names of the populations included in each of the Indian states shown in the figure.

Figure 3. Linear regressions and Spearman's correlations between the oldest historical records of the Romani settlements in each European country and the genetic distances (F_{ST}) between each Romani population and one of three main geographically Romani groups: Balkans (i.e. Bulgaria), West Europe (i.e. Portugal) and North Europe (i.e. Estonia). In the case of Bulgaria the values of each population have been included whereas in other cases only the linear regressions are shown (see also Figure S2 E for all population comparisons and those including Croatia; Welsh Romani were not considered in this analysis).

Figure 4. ABC analyses. A) Contour map (Kriging interpolation) showing north/north-west region of India (including Meghawal and Kashmiri Pandit populations) as the region with the highest probability of representing the homeland of the European Romani. The figure shows the percentage of times the Bayes factor was >1.5 (see also

Table S1, Figure S3 E and F and ABC section in Supplementary Experimental Procedures). The Indian states corresponding to the sampled populations are shown in black. Punjab state (cited in the text but not sampled) is also indicated. Note that the sampling location of Chenchu was originally the same as Vysya [24], but was relocated to avoid the same exact position in the density plot. **B)** Reconstructed demographic history of the European Romani. The width of the branches is proportional to the estimated effective population sizes and the red lines indicate bottleneck events. Arrow width indicates migration rates, in units of number of migrant chromosomes from the donor population per generation. Time of the demographic events was estimated using a generation time of 25 years. See Table S2 and Figure S3 G and H for additional information. See Figure S4, Tables S3 and S4 for inference of additional demographic information not considered in ABC model.