The phylogeography of Y-chromosome haplogroup H1a1a-M82 reveals the likely Indian origin of the European Romani populations


published online 29 November 2012

http://www.plosone.org/article/info:doi/10.1371/journal.pone.0048477
The Phylogeography of Y-Chromosome Haplogroup H1a1a-M82 Reveals the Likely Indian Origin of the European Romani Populations

Niraj Rai1–9, Gyaneshwer Chaubey2–9, Rakesh Tamang1–9, Ajai Kumar Pathak2, Vipin Kumar Singh1, Monika Karmin2,3, Manvendra Singh1, Deepa Selvi Rani1, Sharath Anugula1, Brijesh Kumar Yadav1, Ashish Singh1, Ramkumar Srinivasagaran1, Anita Yadav1, Manju Kashyap1, Sapna Narvariya1, Alla G. Reddy1, George van Driem4, Peter A. Underhill5, Richard Vilems2,3,6, Toomas Kivisild3,7, Lalji Singh1,8,9, Kumarasamy Thangaraj1*

1 CSIR-Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad, India, 2 Evolutionary Biology Group, Estonian Biocentre, Tartu, Estonia, 3 Department of Evolutionary Biology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia, 4 Himalayan Languages Project, Institut für Sprachwissenschaft, Universität Bern, Bern, Switzerland, 5 Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America, 6 Estonian Academy of Sciences, Tallinn, Estonia, 7 Department of Biological Anthropology, University of Cambridge, Cambridge, United Kingdom, 8 Genome Foundation, Hyderabad, India, 9 Banaras Hindu University, Varanasi, India

Abstract

Linguistic and genetic studies on Roma populations inhabiting in Europe have unequivocally traced these populations to the Indian subcontinent. However, the exact parental population group and time of the out-of-India dispersal have remained disputed. In the absence of archaeological records and with only scanty historical documentation of the Roma, comparative linguistic studies were the first to identify their Indian origin. Recently, molecular studies on the basis of disease-causing mutations and haploid DNA markers (i.e. mtDNA and Y-chromosome) supported the linguistic view. The presence of Indian-specific Y-chromosome haplogroup H1a1a-M82 and mtDNA haplogroups M5a1, M18 and M35b among Roma has corroborated that their South Asian origins and later admixture with Near Eastern and European populations. However, previous studies have left unanswered questions about the exact parental population groups in South Asia. Here we present a detailed phylogeographical study of Y-chromosomal haplogroup H1a1a-M82 in a data set of more than 10,000 global samples to discern a more precise ancestral source of European Romani populations. The phylogeographical patterns and diversity estimates indicate an early origin of this haplogroup in the Indian subcontinent and its further expansion to other regions. Tellingly, the short tandem repeat (STR) based network of H1a1a-M82 lineages displayed the closest connection of Romani haplotypes with the traditional scheduled caste and scheduled tribe population groups of northwestern India.


Editor: Yong-Gang Yao, Kunming Institute of Zoology, Chinese Academy of Sciences, China

Received August 8, 2012; Accepted October 2, 2012; Published November 28, 2012

Copyright: © 2012 Rai et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: KT and LS are supported by the Council of Scientific and Industrial Research (CSIR), India. Financial supports to LS and KT from the Department of Biotechnology (DBT), India, and Department of Science and Technology (DST), India, are gratefully acknowledged. LS is also supported by Bhatnagar Fellowship (CSIR) and Bose Fellowship (DST). PAU thanks Carlos D. Bustamante for support. KT and TK acknowledge support from the UK-India Education and Research Initiative (RG-4772). RY, GC and MK thank the European Union European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre, and University of Tartu. RV was supported by Estonian Basic Research grant 3P0270177A08. AKP was supported by the European Social Fund’s Doctoral Studies and Internationalisation Programme DoRa. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.
* E-mail: gyanc@ebc.ee (GC); thangs@ccmb.res.in (KT)
† These authors contributed equally to this work.

Introduction

The Roma in England are traditionally known as gypsies because it was thought that they came from Egypt and were therefore ‘gypcians’. German Zigeuner, French tziganes and names in several other European languages derive from a designation for a Manichean sect that practiced sorcery and soothsaying in the last centuries of the Byzantine Empire. On linguistic grounds, Grellmann [1] pointed out that the Roma must have originated in the Indian subcontinent. A detailed linguistic study by Pott [2] established that the various dialects spoken by the Roma derive specifically from North India. Roma populations are distributed widely within Europe including the Balkans and Scandinavia as well as throughout the Near East.

The name by which Roma designate themselves is Rom (singular Room), whereby the double rr in Romani orthography represents a uvular ‘r’ [R] as opposed to an apical ‘r’ [r]. The autonym Roma is held to be cognate with Dom, a collective term for the ancient aboriginal populations of the Indian subcontinent. Many Dom remained outcastes or tribals, whereas some were
assimilated into the lower strata of the caste system by the Indo-European speaking Indians [3,4].

The Roma route of migration has long been the object of linguistic study. Grierson [5] propagated the idea that the 'Gipsy languages' were of 'Dardic origin', but Turner [6] demonstrated that the Romani languages were not Dardic, but belonged to the same central Indo-Aryan subgroup as Hindi. The presence of Burushaski loans in Romani [7], the lack of Arabic loans and the presence of Dardic, Georgian, Ossetian, Armenian and mediaeval Greek loans [9] indicate that the Roma migrated to Europe by a northerly route, beginning around Gilgit in the northernmost Hindu Kush, thence along the southern Caspian littoral, the southern flank of the Caucasus, the southern shoreline of the Black Sea, across the Bosporus, and subsequently spreading across Europe since 13th century. A legacy of the migration is that some Roma refer to themselves as Sinti, an adjectival form derived from Sindh, the name of the Indus river. Morgenstierne [9] argued that Roma refer to themselves as Sinti, an adjectival form derived from European speaking Indians [3,4].

Europe since 13th century. A legacy of the migration is that some Roma migrated to Europe by a northerly route, beginning around Gilgit in the northernmost Hindu Kush, thence along the southern Caspian littoral, the southern flank of the Caucasus, the southern shoreline of the Black Sea, across the Bosporus, and subsequently spreading across Europe since 13th century. A legacy of the migration is that some Roma refer to themselves as Sinti, an adjectival form derived from Sindh, the name of the Indus river. Morgenstierne [9] argued that Roma refer to themselves as Sinti, an adjectival form derived from

Results and Discussion

Hg H1a1a-M82 accounts for more than 12% of all male lineages in South Asia, with a highest incidence of 20% amongst South Indians and a lowest incidence of 0.2% amongst northeast Indian populations (Table S1 and S3). Hg H1a1a-M82 has a decreasing gradient from its peak frequency in southern India toward both the northwestern and the northeastern peripheries and a virtual absence further to the extreme east and west (Table S1, S3 and S4). Interestingly, substantial frequency changes occur despite the short geographical distance between the East Indian and their adjacent northeast Indian populations, suggesting different paternal population histories for these two neighboring regions. Consistent with the frequency distribution of H1a1a-M82 within India, the Y-STR variance and expansion ages basically reflect a similar pattern (Fig. 1 and Table 1). The overall TMRCAs estimate of the H1a1a-M82 lineage in India using p-statistics [46], was ~22KYA. The regional expansion times within India range between 16 and 24 KYA (Table 1). The higher expansion time with the associated mean pair-wise difference (MPD) and haplotype diversity (HD) (Fig. 1 and, Table 1 and 2) suggests that hg H1a1a-M82 originated in the same South Asian pre-LGM gene pool of Y-chromosomes.

Within India, language group-wise expansion times and variance were observed to be highest among Dravidian speakers (25±3.5 KYA, 0.59), followed by Austroasiatic speakers (22.4±3.3 KYA, 0.52), and lowest among Indo-European speakers (19.5±2.2 KYA, 0.47). Just as with the expansion time of hg O2a-M95 among Austroasiatic speakers [35], the H1a1a-M82 expansion time was established to be substantially higher among South Munda (22±3.5 KYA) than in the North Munda (16.3±3.3 KYA) group. The South Munda groups live in close proximity to the Dravidian groups, and it is likely that the already diverse Dravidian H1a1a-M82 might have assimilated into the Munda group after the Austroasiatic incursion into the Indian subcontinent.
The geographical distribution of hg H1a1a-M82 is largely restricted to South Asia, and its significant occurrence among European Roma populations strongly links the Roma to the Indian subcontinent. The high frequency of H1a1a-M82 among all the Roma groups and their reduced genetic diversity relative to South Asian populations can most likely be attributed to their recent migration from India (Table 1 and 2). In the network analysis (Fig. 2), the Roma Y-STR haplotypes cluster predominantly close to the northwestern Indian haplotypes. In addition, the northwestern Indian haplotypes, while diverse, generally radiate from the core of the network while the Roma haplotypes being distributed further away. These patterns point to northwestern India as the source of the Roma H1a1a-M82 chromosomes. The average age estimate of Roma founders considering their distance from Northwest Indian founders is 1405 ± 688 YBP (Table 3), which is largely in agreement with the time frames suggested previously.

In order to ascertain the closest population group among northwestern Indians, we redrew the network of Roma haplotypes exclusively within the northwestern Indian variation (Fig. 3). It is highly revealing that the closest or matching haplotypes with the Roma haplotypes were found in scheduled caste and scheduled tribe populations, while the middle and upper caste haplotypes were more distant to the Roma haplotypes (Fig. 3). Scheduled castes and Scheduled tribes are the endogamous groups in India that are given a special status by the Government of India to uplift their social status (for more details, refer [47]). Historically, the assimilation of so-called tribals into the caste system generally did little to ameliorate the socio-economic barriers or enhance the marriageability of former outcastes to members of the middle or

<table>
<thead>
<tr>
<th>Group</th>
<th>$n$</th>
<th>$h$</th>
<th>HD</th>
<th>MPD</th>
<th>AMD</th>
<th>Variance</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roma Portugal</td>
<td>21</td>
<td>8</td>
<td>0.8286±0.00646</td>
<td>1.5190±0.9501</td>
<td>1.14</td>
<td>0.08</td>
<td>3.69±2.53</td>
</tr>
<tr>
<td>Roma Serbia</td>
<td>41</td>
<td>17</td>
<td>0.8744±0.0363</td>
<td>1.4183±0.8851</td>
<td>1.66</td>
<td>0.1</td>
<td>3.71±2.18</td>
</tr>
<tr>
<td>Roma Croatia</td>
<td>76</td>
<td>13</td>
<td>0.5607±0.0674</td>
<td>1.2758±0.8111</td>
<td>1.36</td>
<td>0.05</td>
<td>1.91±0.72</td>
</tr>
<tr>
<td>Northwest India</td>
<td>64</td>
<td>57</td>
<td>0.9950±0.0042</td>
<td>7.0164±3.3392</td>
<td>5.86</td>
<td>0.38</td>
<td>16.04±2.38</td>
</tr>
<tr>
<td>West India</td>
<td>13</td>
<td>13</td>
<td>1.0000±0.0302</td>
<td>6.6795±3.3706</td>
<td>5.23</td>
<td>0.45</td>
<td>17.65±3.97</td>
</tr>
<tr>
<td>North India</td>
<td>26</td>
<td>24</td>
<td>0.9908±0.0151</td>
<td>8.3108±3.9781</td>
<td>8.04</td>
<td>0.49</td>
<td>19.46±2.95</td>
</tr>
<tr>
<td>East India</td>
<td>14</td>
<td>13</td>
<td>0.9890±0.0314</td>
<td>8.0659±3.9862</td>
<td>9.14</td>
<td>0.48</td>
<td>20.79±3.84</td>
</tr>
<tr>
<td>Northcentral India</td>
<td>35</td>
<td>20</td>
<td>0.9092±0.0389</td>
<td>7.3126±3.5062</td>
<td>7.94</td>
<td>0.46</td>
<td>19.12±3.01</td>
</tr>
<tr>
<td>Southcentral India</td>
<td>43</td>
<td>36</td>
<td>0.9900±0.0078</td>
<td>8.7741±4.1281</td>
<td>8.84</td>
<td>0.59</td>
<td>24.94±3.09</td>
</tr>
<tr>
<td>South India</td>
<td>37</td>
<td>35</td>
<td>0.9970±0.0071</td>
<td>8.8769±4.1865</td>
<td>8.78</td>
<td>0.56</td>
<td>23.63±3.22</td>
</tr>
</tbody>
</table>

$n$ = number of samples; $h$ = number of haplotypes; HD = haplotype diversity; MPD = mean pairwise difference; AMD = average mutational distance from Roma modal haplotype.
high castes. However their language and means of subsistence were often affected, e.g. assimilation to an Indo-Aryan language and the shift from foraging, hunting and fishing to a more sedentary existence. Not surprisingly, the genetic differences between scheduled tribes and scheduled castes are not found to be substantial [47]. On the basis of our findings, it is therefore most parsimonious to conclude that the genealogically closest patrilineal ancestors of the Roma were among the ancestors of the present scheduled tribes and scheduled caste populations of northwestern India. The genetic data analysed here for the first time provide strong population genetic support for the linguistic

Table 2. The average number of pairwise differences are shown within (PiX along diagonal) and between (PiXY above diagonal) populations; pairwise (δμ)2 genetic distance values are depicted below the diagonal.

<table>
<thead>
<tr>
<th></th>
<th>RP</th>
<th>RS</th>
<th>RC</th>
<th>NWI</th>
<th>WI</th>
<th>NI</th>
<th>NCI</th>
<th>SCI</th>
<th>SI</th>
<th>Afg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roma-Serbia (RS)</td>
<td>0.27117</td>
<td>1.41829</td>
<td>2.42202</td>
<td>5.75191</td>
<td>4.88368</td>
<td>7.59568</td>
<td>8.40592</td>
<td>7.06899</td>
<td>7.64265</td>
<td>7.61437</td>
</tr>
<tr>
<td>Roma-Croatia (RC)</td>
<td>0.7492</td>
<td>1.07497</td>
<td>1.27579</td>
<td>6.03372</td>
<td>4.90891</td>
<td>7.92814</td>
<td>7.98872</td>
<td>7.22744</td>
<td>7.54376</td>
<td>7.98898</td>
</tr>
<tr>
<td>Northwest India (NWI)</td>
<td>1.26057</td>
<td>1.53457</td>
<td>1.88764</td>
<td>7.01637</td>
<td>7.0012</td>
<td>8.0631</td>
<td>8.66853</td>
<td>8.08304</td>
<td>8.45858</td>
<td>8.46917</td>
</tr>
<tr>
<td>West India (WI)</td>
<td>0.63333</td>
<td>0.83479</td>
<td>0.93127</td>
<td>0.15327</td>
<td>6.67949</td>
<td>8.24852</td>
<td>8.47802</td>
<td>7.8044</td>
<td>8.19678</td>
<td>8.35135</td>
</tr>
<tr>
<td>North India (NI)</td>
<td>2.4752</td>
<td>2.73115</td>
<td>3.13486</td>
<td>0.39953</td>
<td>0.75339</td>
<td>8.31077</td>
<td>9.12363</td>
<td>8.14945</td>
<td>8.85599</td>
<td>8.79002</td>
</tr>
<tr>
<td>East India (EI)</td>
<td>3.47622</td>
<td>3.63681</td>
<td>3.31786</td>
<td>1.12738</td>
<td>1.10531</td>
<td>0.93527</td>
<td>8.06593</td>
<td>8.62653</td>
<td>9.26578</td>
<td>8.94402</td>
</tr>
<tr>
<td>Northcentral India (NCI)</td>
<td>2.60866</td>
<td>2.70354</td>
<td>2.93325</td>
<td>0.91855</td>
<td>0.80835</td>
<td>0.33776</td>
<td>0.93726</td>
<td>7.31261</td>
<td>8.6</td>
<td>8.60541</td>
</tr>
<tr>
<td>Southcentral India (SCI)</td>
<td>2.35066</td>
<td>2.54647</td>
<td>2.51882</td>
<td>0.56335</td>
<td>0.46999</td>
<td>0.31357</td>
<td>0.84577</td>
<td>0.55665</td>
<td>8.77409</td>
<td>9.04274</td>
</tr>
<tr>
<td>South India (SI)</td>
<td>2.45712</td>
<td>2.46679</td>
<td>2.91264</td>
<td>0.52255</td>
<td>0.57317</td>
<td>0.1962</td>
<td>0.47261</td>
<td>0.51066</td>
<td>0.21726</td>
<td>8.87688</td>
</tr>
<tr>
<td>Afghani (Afg)</td>
<td>2.92302</td>
<td>3.23801</td>
<td>2.72395</td>
<td>1.34859</td>
<td>1.17308</td>
<td>1.37026</td>
<td>0.78846</td>
<td>1.4056</td>
<td>1.13234</td>
<td>1.17417</td>
</tr>
</tbody>
</table>

Figure 2. Phylogenetic network relating Y-STR haplotypes within haplogroup H1a1a-M82. The network was constructed using a median joining with MP (maximum parsimony) algorithm as implemented in the Network 4.6 program. The size of the circles is proportional to the number of samples. The data used for comparison have been taken from the literature (See Text S1). doi:10.1371/journal.pone.0048477.g002
based identification of the ancestral Roma with the presumed aboriginal Đoma of northwestern India and the Gangetic plain. This suggestion also enables a resolution of the controversy on the Punjab and the Gangetic plain. The Đoma are indigenous to the Punjab as well as to the entire Gangetic plain as far east as the lower course of the Brahmaputra. Yet not all Đoma groups were historically ancestral to the European Roma, and in fact most descendants of the aboriginal Đoma reside in the Indian subcontinent today [4]. It is held that a large number of aboriginal Đoma were recruited in the Punjab to repel the Ghaznavid invasions of the kingdom of Jayapāla between 1001 and 1026. These Đoma were rewarded by nominal promotion to KṢatriya ‘warrior’ or Ra¯jpu¯t caste status, but such nominal promotions in the social context of the Indian subcontinent generally did not lead to a genuine enhancement of eligibility in marriage or social status because of the enduring nature of local memory [4]. With the fall of the Hindu polities in what today is Pakistan, the westward migration of the Đoma was set into motion. Arguably, the humble social position of the Đoma could frequently have led to circumstances, which could have promoted their geographical mobility. Yet the timing of the Ghaznavid invasion presents the probable temporal and spatial historical correlation.

In order to conduct an intra-group comparison of Y-chromosome variation, we estimated the genetic distances (RST values) from 15 loci Y-STR haplotypes and made a multidimensional scaling plot (Fig. S1). The first dimension largely separated all the

<table>
<thead>
<tr>
<th>n (number of samples)</th>
<th>TMRCA (years)</th>
<th>SD (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1110</td>
<td>785</td>
</tr>
<tr>
<td>65</td>
<td>1024</td>
<td>407</td>
</tr>
<tr>
<td>61</td>
<td>2081</td>
<td>873</td>
</tr>
<tr>
<td><strong>1405</strong></td>
<td><strong>688</strong></td>
<td></td>
</tr>
</tbody>
</table>

Different founders were identified based on the Network analysis (Text S1). The age was estimated from the \(\rho\) statistic (the mean number of mutations from the assumed root of each and every founder), using a 25-year generation time and the TD statistic, assuming a mutation rate of \(6.9 \times 10^{-5}\) [11], based on variation at 15 common Y-STR loci.
doi:10.1371/journal.pone.0048477.t003

Figure 3. Phylogenetic network relating Y-STR haplotypes within haplogroup H1a1a-M82- of Roma with northwest Indian populations. The network was constructed using a median joining with MP (maximum parsimony) algorithm as implemented in the Network 4.6 program. The size of the circles is proportional to the number of samples. SC = Scheduled Caste and ST = Scheduled Tribe.
doi:10.1371/journal.pone.0048477.g003
Indian groups from one another, while the second dimension had a maximum impact in separating east Indian from north-central Indian and northwestern Indian from the Roma groups as well as Croatian Roma from Portuguese and Serbian Roma. We also noted other significant genetic differences between Croatian Roma and Portuguese and Serbian Roma. Portuguese and Serbian Roma share a closer affinity than Croatian Roma (Fig. S1 and Table S5). South and south-central Indian groups occupy a pivotal position in the plot, with other groups scattered around them in accordance with their geographical affinity. The modal haplotypes of different population groups are given in Table S6 and their distance to it in Table S7. The Analysis of Molecular Variance (AMOVA) based on Y-STR data shows that the Roma populations are closest to northwestern Indian populations (Table 4). The average mutational distance from the Roma modal haplotype shows a consistent pattern of northwestern Indian populations as representing the closest Indian groups to European Roma (Table S7).

In conclusion, the analysis of Y-chromosome hg H1a1a-M82 variation in 214 ethnic groups from India shows that northwest Indian populations are the closest to the hg H1a1a-M82 variants observed in the present-day European Roma populations. Although only entire genome analyses can provide the richest narrative of history and migration, the Y-chromosome hg H1a provides an exceptional record of Indian and European Roma-specific paternal heritage, including their exodus from northwestern India and subsequent recent expansion. Our phylogeographical study of hg H1a1a-M82 enables now to better understand the temporal and spatial parameters of this migration. This first genetic evidence of this nature allows us to develop a more detailed picture of the paternal genetic history of European Roma, revealing that the ancestors of present scheduled tribes and scheduled caste populations of northern India, traditionally referred to collectively as the Doma, are the likely ancestral populations of modern European Roma. Our findings corroborate the hypothesized cognacy of the terms Roma and Doma and resolve the controversy about the Gangetic plain and the Punjab in favour of the northwestern portion of the diffuse widespread range of the Doma ancestral population of northern India.

Materials and Methods

Samples were collected with the informed written consent from 3498 unrelated healthy individuals belonging to 57 populations from all the four linguistic groups of India (Table S1 and S3). Y-chromosome marker M82 (defining hg H1a) was genotyped in all the samples. From one to five M82 derived samples of each population were randomly selected for Y-STR genotyping based upon the frequency of H1a1a-M82 in the respective population. Our main aim was to cover highest geographic area (Table S1 and S3). In total, 204, M82 derived samples were genotyped by using the AmpFlSTR® Y-filer™ PCR amplification Kit (Applied Biosystems). The detailed material and method section is given in Text S1.

Supporting Information

Text S1 Detailed materials and methods.

Figure S1 Multidimensional scaling plot of the Rst distances of the population groups based on haplogroup H1a1a-M82 Y-STR data.

Table S1 Details of the Indian and Nepali samples included in the present study and haplogroup H1a1a-M82 frequencies.

Table S2 The 15 loci Y-STR profile of haplogroup H1a1a-M82 belonging to Indian, Afghan and European Roma population, used in the present analysis.

Table S3 Regionwise haplogroup frequency in India and Nepal.

Table S4 The frequency of haplogroup H1a1a-M82 among different world populations.

Table S5 Mean pairwise Fst between different studied groups for haplogroup H1a1a-M82.

Table S6 Modal H1a1a-M82 Y-STR haplotype of different population groups.

Table S7 Average mutational distances from Roma Modal haplotype.

Acknowledgments

We thank all the donors for participating in this study. We are grateful to the anonymous reviewers for their constructive suggestions in improving the quality of our manuscript.

Author Contributions

Conceived and designed the experiments: GC KT. Performed the experiments: NR GC RT AKP VKS M. Karmin MS DSR SA BKY AS RS AY M. Kashyap SN AGR. Analyzed the data: NR GC RT AKP. Contributed reagents/materials/analysis tools: AGR RV LS KT. Wrote the paper: GC KT NR RT AKP. Provided input on manuscript writing: TK PAU RV M. Karmin.
References


