European Journal of Human Genetics (2001) 9, 97–104


www.nature.com/ejhg

ARTICLE

Patterns of inter- and intra-group genetic diversity in the Vlax Roma as revealed by Y chromosome and mitochondrial DNA lineages

Luba Kalaydjieva*[,1,2], Francesc Calafell[3], Mark A Jobling[4], Dora Angelicheva[1], Peter de Knijff[5], Zoë H Rosser[4], Matthew E Hurles[4], Peter Underhill[6], Ivailo Tournev[7], Elena Marushiakova[8] and Vesselin Popov[8]

[1]Centre for Human Genetics, Edith Cowan University, Perth, Australia; [2]Western Australian Institute for Medical Research, Perth, Australia; [3]Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain; [4]Department of Genetics, University of Leicester, Leicester, UK; [5]Department of Human Genetics, Leiden University, The Netherlands; [6]Department of Genetics, Stanford University School of Medicine, Stanford, USA; [7]Department of Neurology, Medical University, Sofia, Bulgaria; [8]Institute of Ethnology, Academy of Sciences, Sofia, Bulgaria

Previous genetic studies, supported by linguistic and historical data, suggest that the European Roma, comprising a large number of socially divergent endogamous groups, may be a complex conglomerate of founder populations. The boundaries and characteristics of such founder populations and their relationship to the currently existing social stratification of the Roma have not been investigated. This study is an attempt to address the issues of common *vs* independent origins and the history of population fissioning in three Romani groups that are well defined and strictly endogamous relative to each other. According to linguistic classifications, these groups belong to the Vlax Roma, who account for a large proportion of the European Romani population. The analysis of mtDNA sequence variation has shown that a large proportion of maternal lineages are common to the three groups. The study of a set of Y chromosome markers of different mutability has revealed that over 70% of males belong to a single lineage that appears unique to the Roma and presents with closely related microsatellite haplotypes and MSY1 codes. The study unambiguously points to the common origins of the three Vlax groups and the recent nature of the population fissions, and provides preliminary evidence of limited genetic diversity in this young founder population. *European Journal of Human Genetics* (2001) **9**, 97–104.

Keywords: Roma (gypsies); founder populations; Y chromosome; mtDNA; genetic diversity

Introduction

The Roma (Gypsies) are a population without written history, and the limited information that exists originates from records provided by the surrounding majority populations.[1] While previous linguistic research has given a general indication of the Indian origins of the Roma,[1] recent studies have emphasised the lack of close resemblance between Romani and any living language spoken in the Indian subcontinent and suggested that Romani originated after the exodus from India, as a product of the social interaction between ethnically and linguistically diverse groups forming the proto-Roma population.[2] It is unclear whether this interaction was accompanied by biological amalgamation of the groups or whether the ancestral migrant population adhered to the social organisation and strict endogamy of the *jatis* of India.[1] The Romani group, defined on the basis of self-identity, rules of endogamy, customs and tradition, language

*Correspondence: L Kalaydjieva, Centre for Human Genetics, Edith Cowan University Joondalup Campus, Perth 6027 WA, Australia.
Tel: +61 8 9400 5808; Fax: +61 8 8400 5851;
E-mail: L.Kalaydjieva@cowan.edu.au
Received 22 March 1999; revised 22 September 2000; accepted 10 October 2000

and religion etc., is still the primary unit of the present-day social organisation of the Roma.[1,3] The greatest diversity exists in the Balkans,[1] with over 50 different Romani groups residing in Bulgaria alone.[3]

Population genetic studies of the European Roma (summarised in[4]) have provided evidence of a complex population, with substantial differences between countries, as well as between samples collected in the same country. The random sampling strategies often employed in such studies have ignored social stratification and thus failed to provide an insight into the relationship between the current identity of Romani groups and their biological history.

This study is an attempt to address the issue of genetic relatedness between socially divergent Romani groups and to outline larger founder populations that merit further research. We have investigated three well defined Romani groups from Bulgaria (Table 1), which differ according to most socio-anthropological criteria and are strictly endogamous relative to each other. Two of the groups (Lom and Kalderas) speak Romani dialects of different strata, strongly influenced by Romanian, whereas the third group (Monteni) are not Romani speakers and have adopted an archaic dialect of Romanian instead. All three groups are therefore classified into the large linguistic category of the Vlax Roma[1,3] reflecting external language influences and thus a history of migrations in the Romanian-speaking Balkan provinces north of the Danube. The Vlax Roma have subsequently spread across Europe in several large migration waves, in the late 19th to early 20th century and in the 1990s, and currently account for about one third of the 8 – 10 million European Roma. Given the linguistic and historical data supporting the diverse origins and further splits of the European Roma, a common history of migrations does not provide sufficient evidence that Vlax groups belong to a single founder population. To address the questions of origins and population fissioning, we have studied mtDNA seqeunce variation and a set of Y-chromosome markers of different mutability among the Lom, Kalderas and Monteni in Bulgaria.

Subjects and methods

Subjects

Ninety-four unrelated individuals from the three groups were included in the study. The subjects come from the northern part of Bulgaria, with a maximum distance of about 400 km between any two places of residence. Informed consent was obtained from all subjects included in the study.

Male lineages

Y chromosomes were characterised in 48 males (19 Monteni, 19 Lom and 10 Kalderas) using three marker systems of different mutability.

Nine biallelic markers, representing probable unique mutation events in human evolution, included YAP,[5] SRY-8299 (previously SRY_{4064}),[6] sY81,[7] Tat,[8] M9,[9] 92R7,[10,11] SRY-2627,[11,12] SRY-1532[13] and M52. M52 is an A to C transversion within a 534-bp fragment from a single copy region of the non-recombining portion of the Y chromosome, amplified with primers: M52F 5′actgtagcatggtcatctagggtg-3′ and M52R 5′-gacgaagcaaacatttcaagagag-3′. The polymorphic site is at nucleotide position 477 counted from the 5′ end of the M52F primer. The first eight markers were assayed as indicated in the references above. M52 was PCR amplified and typed by DHPLC as described.[9]

Seven Y-specific microsatellites were typed and used to construct haplotypes. DYS19, DYS390, DYS391, DYS392 and DYS393 were analysed as described.[14] A modified protocol was used for the analysis of DYS389[15] and the first two (A+B) and last two (C+D) repeat stretches were combined in the variability analysis. PCR products were fluorescently labelled using a forward primer with a 5′ FITC label and size-separated on the ALF^{TM} DNA sequencer (Pharmacia, Uppsala). Haplotypes followed the order DYS19-DYS389AB-DYS389CD-DYS390-DYS391-DYS392-DYS393. Allele numbers indicate the number of repeats.

The minisatellite MSY1 (DYF155S1) was analysed as described.[16]

Female lineages

Sequence variation in segment I of the hypervariable region was investigated in 82 individuals: 40 Monteni, 19 Lom and 23 Kalderas. PCR amplification and sequencing were performed as previously described.[17] The reactions were analysed on an ABI 373 Stretch DNA Analyzer.

Analysis of Y chromosome data

Nei's unbiased estimator of diversity, including standard error, was calculated as described.[18] Population pairwise F_{ST} calculations were done using Arlequin version 1.1.[19]

Table 1 Socio-anthropological characteristics of the Kalderas, Monteni and Lom Gypsies (modified from[3])

	Kalderas	*Monteni*	*Lom*
Metagroup	Kalderas	Rudari	Yerlides
Traditional trade and ethnonym	Coppersmiths, Tinsmiths	Spoonmakers	Livestock-dealers
Mode of life	Nomadic until late 1950s	Nomadic until 1920 – 1950	Nomadic until 1900 – 1920
Language/dialect	Romani Stratum III (New Vlax)	Archaic Romanian	Romani Stratum II (Old Vlax)
Religion	East Orthodox	East Orthodox	Mostly Baptist, formerly Orthodox
Migrations out of Romania	Late 19th to early 20th century	Late 19th to early 20th century	18th century

Composite minimum-spanning networks of the Y chromosome haplogroup 35 lineage microsatellite/minisatellite haplotypes were constructed as described.[20] MSY1 codes, all of modular structure 3,1,3,4, were treated as four microsatellite loci for this analysis.

Dating of haplogroup 35 was done on microsatellite and minisatellite results separately, using three different methods. The first[21] is based on the mean number of mutational steps from an assumed root of a tree of microsatellite or MSY1 codes. This root (microsatellite haplotype 3; MSY1 code 35e) was chosen from pairwise differences between all haplotypes/MSY1 codes, as the one having the least number of summed mutational steps from all other chromosomes; it was also equivalent to the haplotype constructed from the modal alleles at each locus. The second method[22] is based on the average squared distance, calculated (using the Microsat 1.5d program[23]) between a population of chromosomes and a root haplotype, identified as described above. The third method[24] is based on the accumulated variance since the foundation of the lineage by a single male bearing a single haplotype or code. For all three methods we assume a generation time of 20 years and a mutation rate of 6% per generation (95% CI 1–11%) for MSY1[16] and of 2.1×10^{-1} per generation (95% CI $0.60 - 4.9 \times 10^{-3}$) for microsatellites.[25] For this method, N_e was taken as 4900.[26] Mutation in each MSY1 repeat block was weighted for the number of repeats in that block,[20] assuming an equal chance of mutating for each repeat unit.

Analysis of data on the variation of HVR segment I

MtDNA variation was compared to published European and West Asian,[27–33] Bulgarian[17] and Havik Indian[34] sequences. Genetic distances between these populations and Romani groups were estimated as described[35] and represented as neighbour-joining trees[36] using the PHYLIP 3.57c package.[37]

Admixture was estimated using the triangle method[38] and taking the Havik Indians[34] and the Bulgarians[17] as the parental populations.

Relative long-term effective population size was estimated as described.[39] This and F_{ST} statistics were computed using the Arlequin package version 1.1.[19]

Results

Paternal lineages

A single predominant male lineage On the basis of biallelic markers, the 48 males from the three Romani groups could be subdivided into five haplogroups. Listing the states of the biallelic markers in the order YAP/SRY-8299/sY81/Tat/M9/92R7/SRY-2627/SRY-1532/M52, these are defined as follows: haplogroup 1 – 000011010; haplogroup 2 – 000000010; haplogroup 3 – 000011000; haplogroup 21 – 110000010; haplogroup 35 – 000000011, where '0' is the ancestral and '1' the derived state of each marker with the exception of 92R7 and SRY-2627, where ancestral state is unknown and the T allele is given as the '1' allele.

Seventy-three per cent (35/48) of the males from all three groups fell into a single haplogroup, haplogroup 35 (Figure 1A), uniquely and uniformly defined by the binary M52 transversion. The remaining had diverse Y chromosomes of haplogroups 1, 2, 3 and 21. Pairwise F_{ST} analysis[19] showed no significant differences (5% significance level) between the three groups.

Analysis of MSY1 and of the microsatellites revealed that the haplogroup 35 chromosomes are closely related to each other, clustering tightly together in a minimum spanning network[20] of composite microsatellite/MSY1 haplotypes (Figure 1B), and also in networks constructed separately from either microsatellite or minisatellite data (not shown). With a single exception (Figure 2: individual 7, sub-lineage 35d), the MSY1 codes of the haplogroup 35 chromosomes belong to modular structural class '3,1,3,4' and have the general structure $(3)_{6-7}(1)_{10}(3)_{42-46}(4)_{13/15}$ (Figure 2).[16]

Direct data on haplogroup 35 chromosomes in other populations are not available, so it is difficult to say if this lineage is truly confined to the Roma. Previous studies[16] have shown the 3,1,3,4 MSY1 modular structural class to be very common worldwide, however a comparison of the haplogroup 35 lineage with 147 '3,1,3,4' chromosomes of diverse ethnic and geographic origins fails to reveal a single example with fewer than eight mutational differences from any Romani haplogroup 35 chromosome. As assessed using MSY1, this lineage was found at very low frequency (one out of 18 males) in the Bulgarian non-Romani population (data not shown). When the predominant haplogroup 35 lineage was removed, pairwise F_{ST} analysis[19] showed no difference, at the 5% significance level, between Roma and Bulgarians.

Within haplogroup 35 there were two sets of 11 identical alleles, another set of seven, an additional pair and four singletons. The overall haplotype diversity for the Y chromosome microsatellites was 0.7619 ± 0.0620, with 0.5147 ± 0.1450 found among the Monteni, 0.7810 ± 0.1016 in the Lom and 0.7273 ± 0.1444 in the Kalderas. For MSY1 codes, Nei's unbiased estimator of diversity[35] for the overall sample of Roma males was 0.869 ± 0.020. For the three individual Romani groups, it was estimated as follows: Monteni 0.711 ± 0.114; Lom 0.814 ± 0.059; and Kalderas 0.867 ± 0.107.

Age of the haplogroup 35 chromosomes The substantial presence of haplogroup 35 in all three Romani groups (Figures 1 and 2) indicated that its origin predates the fissions that gave rise to the groups. Using three different methods[21,22,24] and treating microsatellites and the minisatellite MSY1 separately, we have estimated the time to most recent common ancestor for the haplogroup 35 chromosomes (Table 2). On the basis of microsatellite data, the age was estimated at about 400 years, with good agreement between the dates and the 95% confidence intervals obtained with the three methods. The dates derived from the MSY1 data are

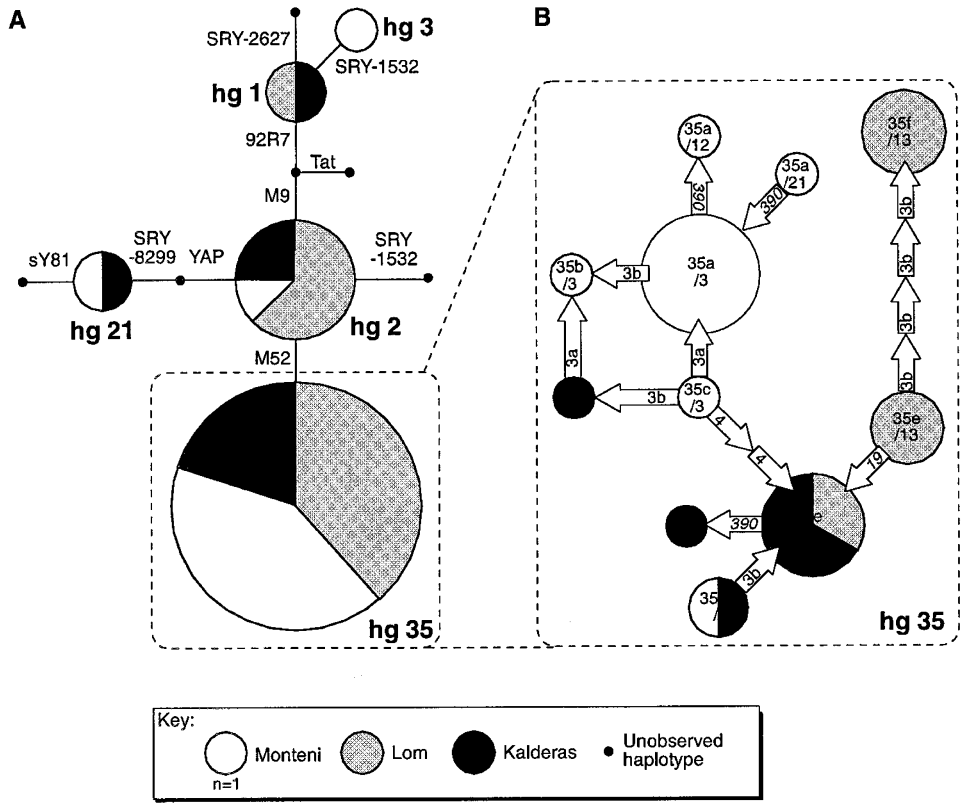

Figure 1 Networks showing haplotype distribution of Gypsy Y chromosome diversity. (**A**) Haplogroup diversity. Diversity assayed using the biallelic polymorphisms is displayed on a maximum parsimony network of haplogroups. A single line between haplogroups represents a single mutation event at the indicated locus. Five of a possible eleven haplogroups are detected. (**B**) Minisatellite and microsatellite diversity within haplogroup 35. Minimum spanning network connecting composite MSY1 and microsatellite haplotypes for haplogroup 35 chromosomes. Sub-lineage 35d, which has a non '3,1,3,4' modular structure, has been omitted. Circles represent observed haplotypes, with areas proportional to the number of individuals. Haplotype names are given as 'MSY1 sub-lineage/ microsatellite haplotype number'. A single arrow between haplotypes indicates a single mutation event, and the direction of the arrow indicates an increase of repeat number. Names of mutating loci are given in italics within the arrows: microsatellite loci are abbreviated by removing the '*DYS*' prefix, and MSY1 repeat blocks as follows: 3a – first block of type 3 repeats; 3b – second block of type 3 repeats; 4 – block of type 4 repeats.

substantially older and there is less good agreement between the results obtained with the different methods, although the confidence limits for two of those overlap with the microsatellite data (Table 2). One reason for the discrepancy may be that the single-step mutation model may be inappropriate for MSY1. In this model, the two Lom-specific sub-lineages 35e/13 and 35f/13, are treated as if they differ by four single step mutations at MSY1 (Figure 2). However, this may in fact represent a single saltatory mutation from 42 to 46 type 3 repeats. Similar considerations apply to the block of type 4 repeats in sub-lineages 35c and 35e. The absence of intermediates makes saltatory mutations likely, further supported by their occurrence in deep-rooting pedigrees.[40] A re-analysis of the age of the haplogroup 35 chromosomes, treating both of these potential saltatory mutations as single steps, gives dates which agree well with the microsatellite data (Table 2), consistent with other studies using both systems.[11,20]

The composite network of haplogroup 35 chromosomes suggested population clustering of the Y chromosome composite haplotypes (Figure 1B). Using Y chromosome microsatellites, an F_{ST} value of 0.0342 ± 0.0333 was obtained (P=0.141, not significantly different from zero), however the faster mutating minisatellite MSY1 gave an F_{ST} value of 0.1752. Given the small current sample size, attempts to date the demographic events that have given rise to the three Vlax groups would be premature.

Maternal lineages

Sequencing analysis of segment I of the mtDNA hypervariable region identified a total of 32 different sequences among 82 individuals (Table 3). More than half of these were shared between the groups. In the Monteni group, 15 sequences were found, of which seven also occurred among the Lom and Kalderas; the proportion of shared sequences was 6/13 in the Lom and 8/15 in the Kalderas group.

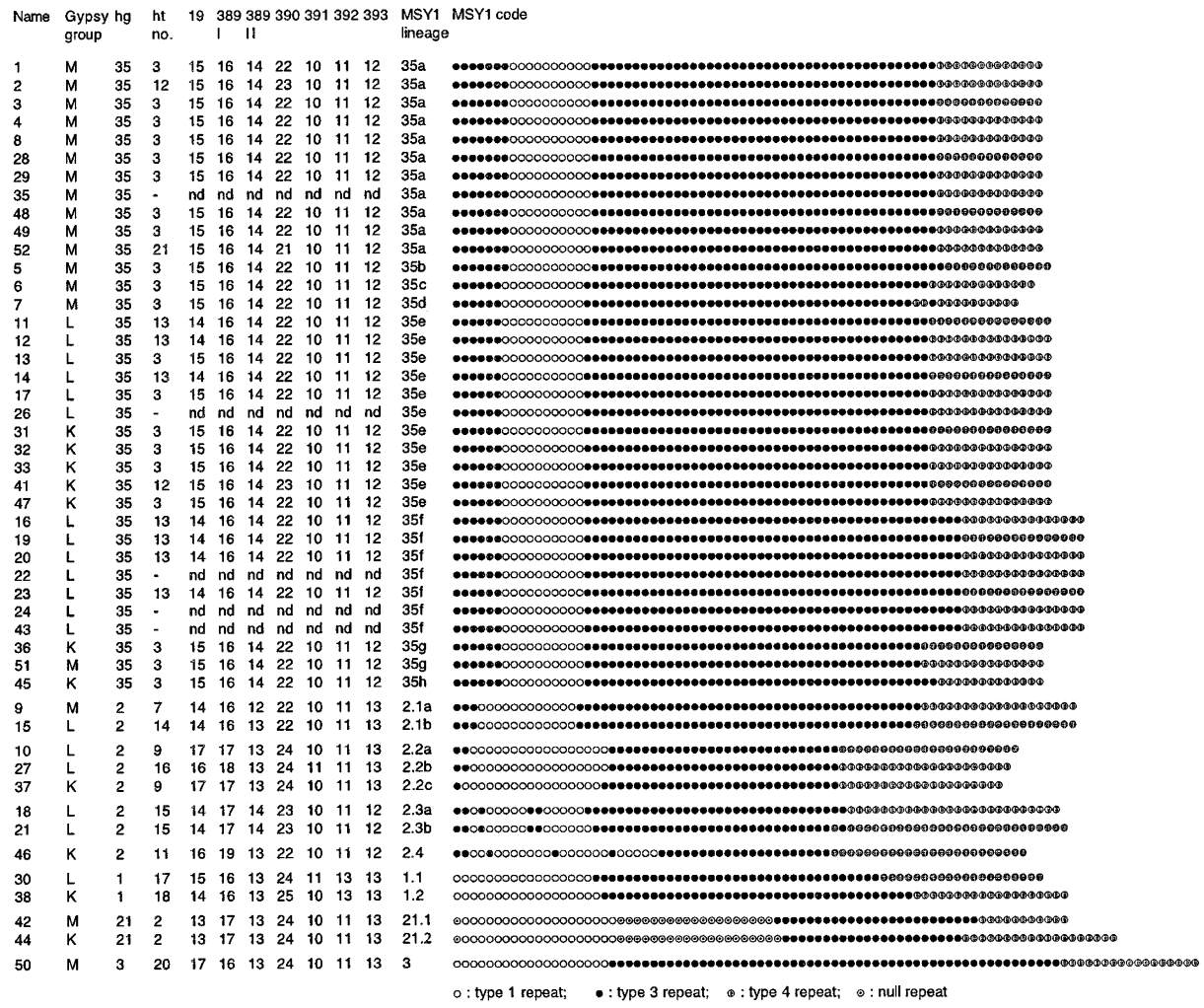

Figure 2 Summary of haplotype data for 48 Gypsy Y chromosomes. Microsatellite loci are abbreviated by removing the '*DYS*' prefix. Microsatellite allele nomenclature is according to Kayser *et al.*[14] M: Monteni; L: Lom; K: Kalderas; hg: haplogroup; ht no.: microsatellite haplotype number; nd: not done.

Table 2 Estimates of time to most recent common ancestor of the Y-chromosomal haplogroup 35

	Microsatellites		*MSY1 single-step*		*MSY1 saltatory*	
Method	*Age*	*95% CI*	*Age*	*Range*	*Age*	*Range*
Mean mutations from root[21]	454	194–1587	706	385–4235	363	198–2176
Accumulated variance[24]	392	150–1874	1340	663–11306	333	169–2531
ASD[22]	457	196–1600	2137	1166–2824	520	284–3120

Ages and 95% confidence intervals (CI) in years are given for microsatellites. For MSY1, the age and expected range have been estimated under a single-step mutation model and also under the assumption that the four-repeat increase between sublineages 35e and 35f, and the two-repeat increase between 35c and 35e, are single mutational events ('saltatory').

Twenty of the 32 sequences identified in this study have not been described previously in other populations.[17,27–34] Two sequences are Eurasian, six have been reported in Europeans and four are a single mutational step away from published European sequences. All 'imported' sequences were shared between the groups. mtDNA variation in the Roma was compared to published European (including Bulgarian) and West Asian sequences and to the Havik

Table 3 mtDNA HVRI sequences found in three Gypsy groups

1111111111111111111111111111111111111			
6666666666666666666666666666666666666			
0000011111111112222222222222333333333	L	M	K
5689922477888912233466799999011224455	O	O	A
3963669526679283403116801468418043445	M	N	L
CCTTGTGGTCCCTCCCTAAACCCCCCCTTTACTACCC			
.....T.....C.....	2	8	1
.....C.....A..T.....T.....	1	2	1
.....A.....T.GG.....C.....T..	1	3	1
.....A.....T.....T..C.....	1	3	0
.....T.....T.....	2	0	3
.....A.C.....T.....C.....	1	0	1
.T...C.....	0	3	4
.....A.....T.GG.....C.....T.T	0	3	1
.....C.....T.....C....	0	1	1
.T...C.A.....	3	0	0
.....T.	2	0	0
..C...AA.G.....T.....	2	0	0
-.....C.....	1	0	0
.....C...T.....C.....	1	0	0
.....T.....T..C.....	1	0	0
...C.....T.....	1	0	0
.....T.....C.....	0	8	0
.....A.....T.....	0	3	0
...C.....T.....	0	1	0
.....G...	0	1	0
...CT.....C..T...G..T.....	0	1	0
.....C.....TT.....	0	1	0
.....TC.....	0	1	0
.....	0	1	0
.....A.....T.....T.....	0	0	3
.....T.....	0	0	2
.T.C.C.....	0	0	1
.....C.....	0	0	1
.....C.....	0	0	1
.....A.....T.....T..T.....	0	0	1
.....T.....T.....T.....	0	0	1

Only variable positions are given. Dots denote identity with the reference sequence.[41] Columns under the 'LOM', 'MON' and 'KAL' headings indicate number of sequences found in the Lom, Monteni and Kalderas groups, respectively.

Indians. In the neighbour-joining genetic tree[36] (Figure 3), Romani sequences clustered together with Havik Indians and away from Europeans and Middle Easterners. The only robust node in the tree was the one separating Roma and Indians from the other populations.

Using the triangle method,[38] and the Havik Indians[34] and the Bulgarians[17] as the parental populations, admixture of Bulgarian sequences for the overall sample was calculated as m=0.48 (95% confidence interval 0.33–0.95). In the individual groups, admixture was estimated at 0.63 (0.34–0.95) for the Monteni; 0.45 (0.20–0.95) for the Lom; and 0.35 (0.13–0.91) for the Kalderas group.

Estimates of mtDNA sequence diversity in the three Romani groups are shown in Table 4. In the Lom and Kalderas groups, diversity was similar to other European

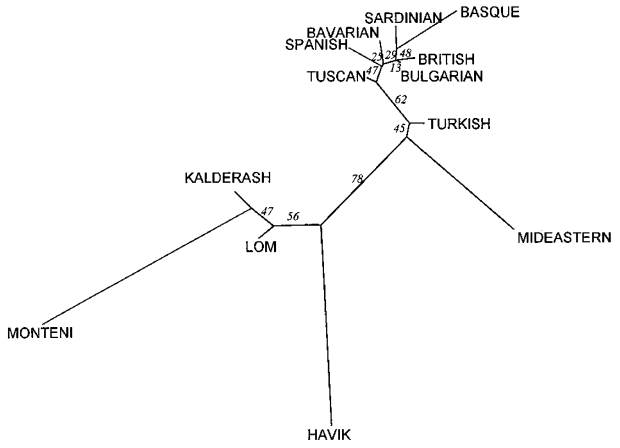

Figure 3 Neighbour-joining tree based on mtDNA sequence of European, West Asian, Roma and Indian populations. Figures are per cent bootstrap support at 1000 iterations. The populations included in the comparison and the references are indicated in the text.

populations. The Monteni presented two sequences at high frequencies, each occurring in eight out of 40 individuals. Each sequence is a single mutation step away from a Kalderas sequence. One is confined to the Monteni, whereas the other is shared with the other two groups. The fact that two mtDNA sequences account for 40% of the Monteni sample is reflected in a lower haplotype diversity (corrected 'heterozygosity') of 0.912 ± 0.024. Based on the number of different sequences and corrected 'heterozygosity', the long-term effective population size of the Monteni was estimated to have been 2–2.5 times lower than that of the Lom and Kalderas (Table 4). However, the estimates derived from polymorphism at the nucleotide level (number of segregating sites and mean pairwise differences) failed to show differences between the three groups. This apparent discrepancy could be the product of an expansion, after a bottleneck, of sequences that were in the periphery of the previously existing tree. Alternatively, it could be an artifact resulting from the small sample size.

The fraction of the genetic variance among the three Romani groups (that is, the F_{ST} value) was estimated at 0.0279.

Discussion

We have analysed three endogamous groups of Vlax Roma using mtDNA sequences together with a set of paternally inherited markers on the Y chromosome, some of which represent unique mutation events, while others are highly mutable and provide information on more recent population history. The results demonstrate clearly that the three Romani groups share identical or closely related haplotypes in both maternally and paternally inherited systems. Over 60% of the mtDNA sequences are common to all groups. Over 70% of males present with closely related MSY1/

Table 4 Diversity of mtDNA sequences and estimates of relative long-term effective population size

	Kalderas	*Lom*	*Monteni*
Sample size	23	19	40
k	15	13	16
D	0.956 ± 0.027	0.965 ± 0.024	0.912 ± 0.024
Θ_H	20.33	25.77	9.02
Θ_k	21.99	16.89	9.17
Θ_S	5.96	6.29	5.84
Θ_π	4.44	5.11	4.76

k, number of difference sequences; D, haplotype diversity (~corrected 'heterozygosity'). Several estimates of Θ=2Nu are shown based on: H, heterozygosity; k, number of different sequences; S, number of segregating sites, and π, mean pairwise differences.

microsatellite haplotypes that belong to a single Y chromosome lineage defined by markers representing unique mutation events. The findings point unambiguously to the common origins of the three socially divergent groups and indicate that the biological boundaries of founder populations do not mirror the socio-anthropological classification of Romani groups.

The predominant Y chromosome lineage, as well as most mtDNA sequences, appear specific to the Roma. Tracing the geographical and ethnic origins of these lineages in the vast diversity of the Indian subcontinent and along the migration routes of the Roma will become possible with the accumulation of data on global genetic diversity. The current lack of knowledge of the population(s) of origin and of the multiple possible sources of admixture, means that the calculated 50% of non-Asian mtDNA should be regarded only as a crude estimate and a general indicator that female admixture is likely to have been greater than male.

The predominant male lineage, haplogroup 35, is defined by markers representing unique mutation events. To investigate its more recent history, we have examined the highly mutable microsatellites and the minisatellite MSY1. The age of the lineage was estimated at about 400 – 500 years, suggesting that the Vlax Roma are a very young founder population and the splits between the currently existing groups have occurred after their arrival in Europe. While indicating the existence of a larger Vlax founder population comprising socially divergent individual groups, this study has also provided preliminary evidence of internal stratification with clustering of lineages in the different Romani groups. The recent nature of the population fissioning and the small sample size of this study make it difficult to draw conclusions on the divergence between the individual Romani groups. In the case of male lineages, structuring was suggested only by the data obtained for MSY1 which, as a result of its high mutability, can be expected to be faster in accumulating variation in recently subdivided populations. In view of its potential importance for future genetic research, the issue of stratification merits further examination on larger sample sizes.

An interesting result of the study is the apparently limited genetic diversity, particularly with regard to male lineages. A high degree of sharing was found for both haplogroup 35, defined by unique mutation events, and for compound Y chromosome haplotypes defined by the highly mutable microsatellites and the minisatellite MSY1. For MSY1 codes, Nei's unbiased estimator of diversity[35] was 0.865 ± 0.020. A lower figure has been found only for the Surui (0.492 ± 0.081), while Finns, Basques and Cook Islanders all show values in excess of 0.96.[16] In the group of Monteni, Nei's unbiased estimator of diversity[35] for MSY1 codes was 0.711 ± 0.114 and, based on mtDNA diversity, the long-term effective female population size of this group was also estimated to have been smaller than those of the Lom and Kalderas. These initial findings merit further investigations, specifically focusing on the Monteni and other related, Romanian-speaking, Vlax groups in Bulgaria as well as in neighbouring countries.

This study raises the possibility that socially divergent Vlax Romani groups may represent a single young founder population. Further studies of larger samples, especially across countries, will show whether the common origins/recent splits scenario applies to the general Vlax Roma population of Europe. The use of genetic markers of different mutability should provide an insight into the old as well as recent history of this population and its present characteristics in terms of structure and diversity.

Acknowledgements

The study was supported by Edith Cowan University, The Wellcome Trust and the Australian Research Council. We thank D Dye, A Savov, O Kamenov and D Chandler for technical assistance. MA Jobling is a Wellcome Trust Senior Fellow (Grant 057559), ZH Rosser was supported by a BBSRC Studentship, ME Hurles by an MRC Studentship and P Underhill by NIH grant GMS28248.

References

1 Fraser A: *The Gypsies*. Oxford, Blackwell Publishers, 1993.
2 Hancock I. The emergence of Romani as a *koiné* outside of India. Personal communication (preprint).
3 Marushiakova E, Popov V: *Gypsies (Roma) in Bulgaria*. Studien zur Tsiganologie and Folkloristik, Band 18. Frankfurt am Main, Peter Lang, 1997.
4 Mastana SS, Papiha SS: Origin of Romany Gypsies – genetic evidence. *Z Morph Anthrop* 1992; **79**: 43 – 51.
5 Hammer MF, Horai S: Y chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet* 1995; **56**: 951 – 962.
6 Whitfield LS, Sulston JE, Goodfellow PN: Sequence variation of the human Y chromosome. *Nature* 1995; **378**: 379 – 380.
7 Seielstad MT, Hebert JM, Linn AA *et al*: Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Hum Molec Genet* 1994; **3**: 2159 – 2161.
8 Zerjal T, Dashnyam B, Pandya A *et al*: Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. *Am J Hum Genet* 1997; **60**: 1174 – 1183.
9 Underhill PA, Jin L, Lin AA *et al*: Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res* 1997; **7**: 996 – 1005.

10 Mathias N, Bayes M, Tyler Smith C: Highly informative compound haplotypes for the human Y chromosome. *Hum Mol Genet* 1994; **3**: 115 – 123.
11 Hurles ME, Veitia R, Arroyo E *et al*: Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am J Hum Genet* 1999; **65**: 1437 – 1448.
12 Veitia R, Ion A, Barbaux S *et al*: Mutations and sequence variants in the testis-determining region of the Y chromosome in individuals with a 46,XY female phenotype. *Hum Genet* 1997; **99**: 648 – 652.
13 Kwok C, Tyler-Smith C, Medonca BB *et al*: Mutation analysis of 2kb 5 to SRY in XY females and XX intersex subjects. *J Med Genet* 1996; **33**: 465 – 468.
14 Kayser M, Caglà A, Corach D *et al*: Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med* 1997; **110**: 125 – 133.
15 Rolf B, Mayer E, Brinkmann B, de Knijff P: Polymorphism at the tetranucleotide repeat locus DYS389 in 10 populations reveals strong geographic clustering. *Eur J Hum Genet* 1998; **6**: 583 – 588.
16 Jobling MA, Bouzekri N, Taylor PG: Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). *Hum Mol Genet* 1998; **7**: 643 – 653.
17 Calafell F, Underhill P, Tolun A, Angelicheva D, Kalaydjieva L. From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann Hum Genet* 1996; **60**: 35 – 49.
18 Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 1978; **76**: 379 – 390.
19 Schneider S, Kueffer J-M, Excoffier L: Arlequin ver 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Genevea, Switzerland, 1997.
20 Hurles ME, Irven C, Nicholson J *et al*: European Y-chromosomal lineages in Polynesians: A contrast to the population structure revealed by mtDNA. *Am J Hum Genet* 1998; **63**: 1792 – 1806.
21 Bertranpetit J, Calafell F: Genetic and geographical variability in cystic fibrosis: evolutionary considerations. In: Chadwick D, Cardew G (eds): *Variation in the human genome*. Chichester, Wiley, 1996, pp. 97 – 118.
22 Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB. Origins of Old Testament priests. *Nature* 1998; **394**: 6689 – 6690.
23 Minch E: Microsat vs 1.5d. Department of Genetics, Stanford University, Stanford, 1997.
24 Goldstein DB, Zhivotovsky LA, Nayar K, Ruiz-Linares A, Cavalli-Sforza LL, Feldman MW: Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. *Mol Biol Evol* 1996; **13**: 1213 – 1218.
25 Heyer E, Puymirat J, Dieltjies P, Bakker E, de Knijff P: Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet* 1997; **6**: 799 – 803.
26 Hammer MF: A recent common ancestry for human Y chromosomes. *Nature* 1995; **378**: 376 – 378.
27 Di Rienzo A, Wilson AC: Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc Natl Acad Sci USA* 1991; **88**: 597 – 601.
28 Piercy R, Sullivan KM, Benson N, Gill P: The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int J Legal Med* 1993; **106**: 85 – 90.
29 Côrte-Real HB, Macaulay VA, Richards MB *et al*: Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* 1996; **60**: 331 – 350.
30 Pinto F, Gonzalez AM, Hernandez M, Larruga JM, Cabrera VM: Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. *Ann Hum Genet* 1996; **60**: 321 – 330.
31 Richards M, Corte-Real H, Forster P *et al*: Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 1996; **59**: 185 – 203.
32 Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bertranpetit J: Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Mol Biol Evol* 1996; **13**: 1067 – 1077.
33 Francalacci P, Bertranpetit J, Calafell F, Underhill PA. Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am J Phys Anthropol* 1996; **100**: 443 – 460.
34 Mountain JL, Hebert SS, Bhattacharyya P *et al*: Demographic history of India and mitochondrial DNA sequence diversity. *Am J Hum Genet* 1995; **56**: 979 – 992.
35 Nei M: *Molecular Evolutionary Genetics*. New York, Columbia University Press, 1987.
36 Saitou N, Nei M: The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4**: 406 – 425.
37 Felsenstein J: PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* 1989; **5**: 164 – 166.
38 Cavalli-Sforza LL, Menozzi A, Piazza A: *The History and Geography of Human Genes*. Princeton, Princeton University Press, 1994, pp 54 – 59.
39 Zouros E: Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics* 1979; **92**: 623 – 646.
40 Jobling MA, Heyer E, Dieltjes P, de Knijff P: Y-chromosome-specific microsatellite mutation rates re-examined using a minisatellite, MSY1. *Hum Mol Genet* 1999; **8**: 2117 – 2120.
41 Anderson S, Bankier T, Barrel BG *et al*: Sequence and organization of the human mitochondrial genome. *Nature* 1981; **290**: 457 – 465.