doi: 10.1111/j.1469-1809.2005.00222.x

Mitochondrial DNA Diversity in the Polish Roma

B. A. Malyarchuk[1,*,†], T. Grzybowski[2,*,†], M. V. Derenko[1], J. Czarny[2] and D. Miścicka-Śliwka[2]

[1]*Institute of Biological Problems of the North, Russian Academy of Sciences, Portovaya str. 18, 685000 Magadan, Russia*
[2]*The Nicolaus Copernicus University in Toruń, Collegium Medicum, Forensic Medicine Institute, ul. Sklodowskiej-Curie 9, 85-094 Bydgoszcz, Poland*

Summary

Mitochondrial DNA variability in the Polish Roma population has been studied by means of hypervariable segment I and II (HVS I and II) sequencing and restriction fragment-length polymorphism analysis of the mtDNA coding region. The mtDNA haplotypes detected in the Polish Roma fall into the common Eurasian mitochondrial haplogroups (H, U3, K, J1, X, I, W, and M*). The results of complete mtDNA sequencing clearly indicate that the Romani M*-lineage belongs to the Indian-specific haplogroup M5, which is characterized by three transitions in the coding region, at sites 12477, 3921 and 709. Molecular variance analysis inferred from mtDNA data reveals that genetic distances between the Roma groups are considerably larger than those between the surrounding European populations. Also, there are significant differences between the Bulgarian Roma (Balkan and Vlax groups) and West European Roma (Polish, Lithuanian and Spanish groups). Comparative analysis of mtDNA haplotypes in the Roma populations shows that different haplotypes appear to demonstrate impressive founder effects: M5 and H (16261–16304) in all Romani groups; U3, I and J1 in some Romani groups. Interestingly, haplogroup K (with HVS I motif 16224-16234-16311) found in the Polish Roma sample seems to be specific for Ashkenazi Jewish populations.

Keywords: mitochondrial DNA, autosomal microsatellite loci, Polish Roma, molecular phylogeography.

Introduction

The Roma (Gypsies), who are believed to be of Indian origin nowadays represent a large population spread over all of Europe, with their highest concentrations in south-eastern Europe and the Iberian Peninsula (Kalaydjieva *et al.* 2001b). The ancestors of the Roma who inhabited North-West India began to migrate westwards in the 9th and 10th centuries. Linguistic, ethnological and anthropological studies have allowed researchers to reconstruct the Roma routes to Europe, which led them through Persia, Armenia, and the Greek-speaking territory of Byzantium (Ficowski, 1985). By the 13th century the Roma had entered the Balkans and some groups moved slowly through the Slavic-speaking regions until they reached Romania. By the 15th century the Gypsies were already living almost everywhere throughout Europe. The first migration of small groups of Gypsies from Hungary to Poland took place at the beginning of the 15th century. Larger groups started to come to Poland from Germany during the 16th century. Those Gypsies have stayed in the Polish territory ever since, and until quite recently have lived a nomadic life; they call themselves the Polish Roma (*Polska Roma*). Probably from the end of the 18th century some Gypsies travelling along the Carpathians began to settle down in the mountain villages of southern Poland. Some of these groups still live in small villages of the Tatra and Beskidy Mountains and are known as *bergitka Roma* (upland Gypsies). Other large Gypsy tribes living in Poland today are descendants of two major groups - *Kelderari* (boilermakers) and *Lovari* (horse hawkers) - who came to Poland from Transylvania and Wallachia in the middle of the 19th century (Ficowski, 1985). According to

* Corresponding author: Dr. Boris A. Malyarchuk, Genetics Laboratory, Institute of Biological Problems of the North, Portovaya str., 18, 685000 Magadan, Russia. Fax/Phone: 7 41322 34463. E-mail: malyar@ibpn.kolyma.ru
†These two authors contributed equally to this work

the latest Census held in 2002 the overall Polish Roma population numbers up to 12,731.

Genetic studies have shown that the Roma populations share a common genetic history, as evidenced by classical, mtDNA and Y-chromosomal markers (Gresham *et al.* 2001; Kalaydjieva *et al.* 2001a; Chaix *et al.* 2004; Morar *et al.* 2004; Zhivotovsky *et al.* 2004). It has recently been found that Roma individuals from many populations and people from the Indian subcontinent share haplotypes for several disease loci (for instance, the congenital myasthenia 1267delG mutation) suggesting that these mutations were characteristic for the founders of the Roma (Morar *et al.* 2004). Coalescence time estimates show that the entire Roma population was founded ~16–25 generations ago (Morar *et al.* 2004). However, despite an obvious founder effect in the Roma, there are substantial differences between the Roma groups, probably due to admixture between the Roma and surrounding European populations. Thus, the population history of the Roma is a string of bottleneck events with current genetic profiles shaped by differential drift due to endogamous practices in small populations and admixture with the surrounding populations (Gresham *et al.* 2001; Jobling *et al.* 2004). Individual Roma groups can also be considered to be isolates within a larger isolate (Jobling *et al.* 2004). It has been noticed that internal differentiation of a founder population into multiple subisolates, maintained for a long time by endogamy rules, appears to be characteristic for the Roma (Morar *et al.* 2004). This feature of the population structure is very attractive for the purposes of gene mapping and searching for small genomic regions showing the strongest linkage disequilibrium with disease. Therefore, population genetic studies of different Roma groups may be highly important. In this study, we have analyzed the diversity of maternal mtDNA lineages in the Polish Roma population in comparison with different Roma groups and the surrounding European populations.

Materials and Methods

Population Samples and mtDNA Analysis

Population samples of 69 Gypsy individuals belonging to the major subdivision of the Polish Roma (*Polska Roma*) were studied. The samples were collected in the West of the country, in the urban areas of Zielona Góra and Nowa Sól. All individuals were maternally and paternally unrelated and originated from the area considered for this study. Appropriate informed consent was obtained from all participants.

Total genomic DNA was extracted from hairs by means of cell lysis in the presence of proteinase K and 1% SDS, followed by phenol/chloroform extractions. RFLP typing was performed by restriction endonuclease analysis of PCR-amplified mtDNA fragments using the same primer pairs and amplification conditions as described elsewhere (Torroni *et al.* 1996; Finnil *et al.* 2000) (Table 1). The samples were typed for a restricted set of RFLPs that were diagnostic of all major Eurasian clusters, on the basis of the hierarchical mtDNA RFLP scheme (Macaulay *et al.* 1999; Richards *et al.* 2000; Yao *et al.* 2002; Malyarchuk *et al.* 2003).

Hypervariable segments I and II (HVS I and II) of the mtDNA noncoding control region were amplified and sequenced as described elsewhere (Malyarchuk *et al.* 2002). The nucleotide sequences from positions 15991 to 16400 (encompassing the HVS I region) and from positions 30 to 407 (encompassing the HVS II) were determined and compared with the revised Cambridge reference sequence (rCRS; Anderson *et al.* 1981; Andrews *et al.* 1999). Complete sequencing of the mtDNA belonging to the Romani-specific M*-lineage was performed as described by Torroni *et al.* (2001). Since DNA samples extracted from hair roots are characterized by low amounts of DNA, for complete mtDNA sequencing we used the DNA extracted from the blood of individual PL173. This Polish individual is characterized by the Romani-specific M*-lineage with HVS I and II sequence 16129-16148-16192-16223-16291-16298-73-263-310 (Malyarchuk *et al.* 2002).

Sequence classification into mtDNA subclusters was based on the nomenclatures of Richards *et al.* (2000) and Palanichamy *et al.* (2004). To classify the mtDNA haplotypes, a phylogeographic approach based on the phylogenetic analysis of the spatial distribution of mitochondrial haplotypes and haplogroups, determined as a monophyletic clade, was performed (Richards *et al.* 1998).

Table 1 RFLP polymorphisms used to identify major Eurasian mtDNA haplogroups

Haplogroups	Characteristic restriction site(s)
West Eurasian:	
HV	−14766 *Mse*I
H	−14766 *Mse*I, − 7025 *Alu*I
pre*V1	−14766 *Mse*I, −15904 *Mse*I, +4577 *Nla*III
pre*V2	−14766 *Mse*I, +15904 *Mse*I, +4577 *Nla*III
V	−14766 *Mse*I, +15904 *Mse*I, −4577 *Nla*III
U	+12308 *Hin*fI
K	+10394 *Dde*I, +12308 *Hin*fI, −9052 *Hae*II
J	+10394 *Dde*I, −13704 *Bst*NI
J1	+10394 *Dde*I, −13704 *Bst*NI, −3007 *Bsh*1236I
T	+13366 *Bam*HI, +15606 *Alu*I
T1	+13366 *Bam*HI, +15606 *Alu*I, −12629*Ava*II
N1	−12498 *Nla*III
I	+8249 *Ava*II, +10032 *Alu*I, +10394 *Dde*I,–12498 *Nla*III
W	+8249 *Ava*II, −8994 *Hae*III
X	−1715 *Dde*I, +14465 *Acc*I
East Eurasian:	
M:	+10394 *Dde*I, +10397 *Alu*I
C	+10394 *Dde*I, +10397 *Alu*I, −13259 *Hinc*II/+13262 *Alu*I
D	+10394 *Dde*I, +10397 *Alu*I, − 5176 *Alu*I
E	+10394 *Dde*I, +10397 *Alu*I, − 7598 *Hha*I
G	+10394 *Dde*I, +10397 *Alu*I, +4830 *Hae*II/+4831 *Hha*I
A	+663 *Hae*III
B	9-bp intergenic deletion between COII and tRNA(Lys)
F	−12406 *Hpa*I/*Hinc*II

Autosomal STRs Analysis

Genotypes for 15 autosomal STR loci (D3S1358, vWA, FGA, TH01, TPOX, CSF1PO, D5S818, D13S317, D7S820, D16S539, D2S1338, D8S1179, D21S11, D18S51 and D19S433) were obtained with the use of AmpFlSTR Profiler and AmpFlSTR SGM Plus PCR amplification Kits (PE Applied Biosystems) according to manufacturers' protocols.

Phylogenetic and Statistical Analysis

The population genetic structure was analyzed using methods implemented in the Arlequin 2.0 software (Schneider *et al.* 2000). The statistical significance of F_{ST}-values was estimated by permutation analysis using 10000 permutations. Intrapopulation diversities (h) were calculated using the formulae (Nei & Tajima, 1981) as implemented in Arlequin 2.0. Multidimensional scaling (MDS) analysis of pairwise interpopulation F_{ST} values was performed with the use of the software package STATISTICA (StatSoft, Inc., Tulsa, OK, USA).

HVS I sequences of the Polish Roma were compared with previously published data on three migrational/linguistic groups of Roma – Balkan, Vlax and West European Roma (Gresham *et al.* 2001). In that study, the Balkan group was represented by populations from early settlements of the Roma in Bulgaria, the Vlax group was represented by the Roma residing in Bulgaria but originating from the Wallachia (present-day Romania) and Moldavia regions, and the West European Roma group comprised subjects from Spain and Lithuania. A 362-bp fragment of the HVS I region, between positions 16023 and 16384, was analyzed.

For population comparison, HVS I data were used from databases published elsewhere (Richards *et al.* 2000; Kasperaviciute & Kucinskas, 2002; Kivisild *et al.* 2003; McEvoy *et al.* 2004; Quintana-Murci *et al.* 2004). At this time, variations between positions 16090 and 16365 of the HVS I region were considered in the AMOVA to allow maximum comparability between all groups of the Roma and the surrounding European populations. Nucleotide positions showing point indels and transversions located between positions 16180-16193 and 303-315 were excluded from all analyses.

Inbreeding coefficient (theta) values based on autosomal STR data within matrilineal groups of the Polish Roma individuals were calculated using Genetic Data Analysis (GDA) software (Lewis & Zaykin, 2001).

Results and Discussion

The analysis of HVS I and II variability in combination with RFLP typing of the coding region haplogroup-specific sites of 69 Polish Roma allowed detection of only 17 different mitochondrial haplotypes (Table 2). Despite the high frequency of identical haplotypes found in different individuals, they do not represent any parent-child pairs. This follows from the results of autosomal microsatellite analysis (see Supplementary

Table 2 mtDNA haplotypes in Polish Roma

HVS I (minus 16000)	HVS II	HG	N	Sample numbers
69 126 145 222 235 261 271	73 295 263 309.1C 315.1C	J1	13	2, 3, 5, 6, 9, 11, 15, 23, 26, 41, 53, 82, 83
343	73 150 263 315.1C	U3	25	8, 21, 24, 25, 27, 29, 32, 34, 39, 42, 44, 46, 49, 52, 69, 70, 71, 72, 79, 89, 95, 96, 97, 98, 102
261 304	64 93 263 315.1C	H	1	86
261 304	93 263 315.1C	H	6	18, 59, 66, 85, 91, 99
304	263 309.1C 315.1C	H	1	10
CRS	263 309.1C 309.2C 315.1C	H	1	45
CRS	263 309.1C 315.1C	H	1	76
145A/G 362	239 263 309.1C 309.2C 315.1C	H	1	75
224 234 311	73 114 263 309.1C 315.1C	K	3	17, 19, 22
145 223	73 189 195 204 207 263 309.1C 315.1C	W	6	1, 14, 16, 43, 51, 54
129 172 223 311 391	73 199 203 204 250 263 315.1C	I	5	28, 47, 55, 56, 94
126 189A 223 278	73 153 195 225 226 263 315.1C	X2e	1	73
183C 189 223 255 278	73 153 195 198 225 263 309.1C 309.2C 315.1C	X2c	1	101
129 223 234 291 298	73 263 309.1C 315.1C	M*	1	57
129 223 291 298	73 263 309.1C 315.1C	M*	1	78
129 223 291 298	73 263 309.1C 309.2C 315.1C	M*	1	93
129 223 291 298	73 146 263 309.1C 315.1C	M*	1	103

Mutations are shown indicating positions relative to the CRS (Anderson *et al.* 1981). The nucleotide positions in HVS I and II sequences correspond to transitions; transversions are further specified. Haplogroup names (HG) are given in capital letters according to the mtDNA classification (Macaulay *et al.* 1999; Richards *et al.* 2000). Heteroplasmic nucleotides are indicated by a slash (/). The presence of insertions is referred to by "." following the nucleotide position.

material Tables 1 and 2). It is noteworthy that extremely high values for the inbreeding coefficient (theta values > 0.1) were found only in some matrilineal groups of individuals (e.g., M*, W and K, see Supplementary Table 2).

A total of eight haplogroups were identified, out of which three - H, J1 and U3 - accounted for 71% of all individuals. The remaining five haplogroups – namely, M*, I, W, X and K – occurred at lower frequencies (<10%). Among these haplogroups, haplogroup M* was found at a frequency of 5.8%, although this haplogroup has been found previously at high frequency in different Romani populations, accounting for 26.5% of the total sample studied (Gresham *et al.* 2001) (Table 3). The ancestral HVS I motif of haplogroup M* in the Roma is 16129-16223-16291-16298. Previously, Gresham *et al.* (2001) identified this cluster of HVS I sequences as belonging to haplogroup M5 described in Indians by Bamshad *et al.* (2001). Meanwhile, the exact position of the Romani-specific M-lineages on the evolutionary tree remains unclear, despite the current progress with mtDNA classification in populations from the Indian subcontinent (Metspalu *et al.* 2004; Palanichamy *et al.* 2004; Rajkumar *et al.* 2005).

Among four Polish Roma individuals characterized by M*-haplotypes we found four HVS I/II sequence types differing by nucleotide substitutions at positions 16234 and 146, as well as by an additional point insertion at position 309 (Table 2). Since the combined HVS I /II sequencing approach did not provide any useful information on the cluster-specific mutations in the HVS II region, we performed a search of diagnostic mutations by means of complete mtDNA sequencing. This study allowed us to reveal a large number of mutations distinguishing the M*-lineage from the rCRS-sequence (Figure 1). Comparison of the Romani M*-lineage with the Indian M5-sequence (Bhovi individual Bho134 from the study by Rajkumar *et al.* 2005) demonstrated that the haplogroup M5 is characterized by three transitions in the coding region, at sites 12477, 3921 and 709. Therefore, the results obtained clearly indicate that the Romani M*-lineage belongs to the Indian-specific haplogroup M5.

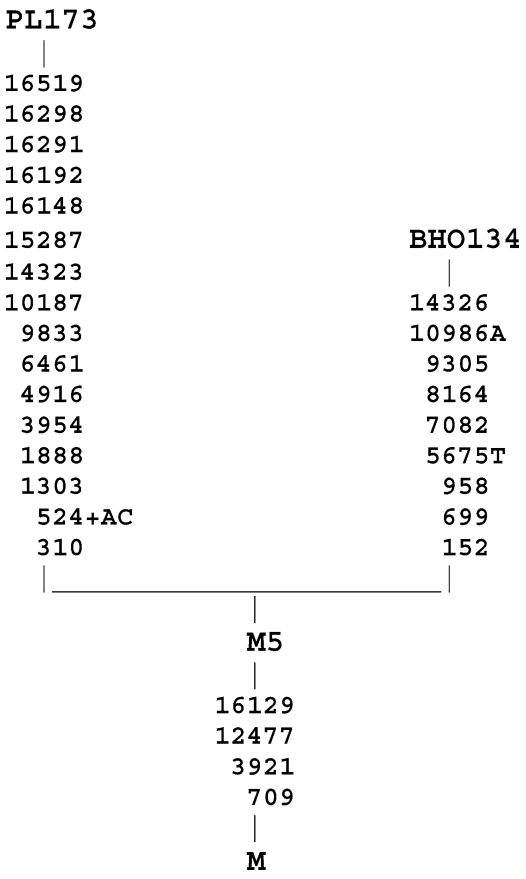

Figure 1 Phylogenetic tree of haplogroup M5 based on complete mitochondrial genome sequences. Numbers along links refer to substitutions scored relative to the rCRS (Andrews *et al.* 1999). Transversions are specified by suffixes. A plus sign (+) denotes an insertion. Nucleotide sequence of Indian individual Bho134 is taken from Rajkumar *et al.* (2005). The haplogroup M differs from the rCRS at sites: 73, 263, 489, 750, 1438, 2706, 4769, 7028, 8701, 8860, 9540, 10398, 10400, 10873, 11719, 12705, 14766, 14783, 15043, 15301, 15326, 16223.

Haplogroup H is one of the most frequent mitochondrial haplogroups in the Roma (Gresham *et al.* 2001). In the Polish Roma it was found at a frequency of 16%. This haplogroup is also the most frequent haplogroup in Europe and is characterized by a considerable branching substructure with several large subclusters (Finnilä *et al.* 2001; Herrnstadt *et al.* 2002; Achilli *et al.* 2004; Loogväli *et al.* 2004). Among the Roma, only one H-haplotype (16261-16304) is widespread in different Romani populations of Europe, but it has the highest frequency (11%) in the Vlax Roma. In Europeans this haplotype is very rare, being found only in several individuals from Belorussian and Bosnian populations (Belyaeva *et al.* 2003; Malyarchuk *et al.* 2003). Interestingly, haplotype 16261-16304 has still not been found in populations from India and Pakistan (Kivisild *et al.* 1999; 2003, Quintana-Murci *et al.* 2004).

Haplogroup U3 is also one of the most frequent haplogroups in the Roma, but its highest frequencies were found in the Spanish, Lithuanian and Polish Roma (Table 3). Diversity of the haplogroup U3 in the Roma is reduced mainly to a single haplotype 16343 that is a root haplotype for haplogroup U3. Note that this haplotype is also present in many populations from Europe and the Middle East (Richards *et al.* 2000). Haplogroup J is characterized by a very high frequency in the Polish Roma (18.8%). This haplogroup has been also found frequently in other Roma populations (Table 3), but it is noteworthy that haplogroup J appears to be very diverse in the Bulgarian Roma - 8 out of 11 HVS I sequence types found in European Roma were observed in individuals belonging to the Balkan and Vlax groups (Gresham *et al.* 2001). In contrast, the Polish Roma are characterized by a marked "founder" effect, because all of their 13 J-individuals have a single HVS I/II sequence type, bearing transitions at positions 16235 and 16271 in the HVS I region and belonging to the subhaplogroup J1. The haplotype with the HVS I motif 16069-16126-16145-16222-16235-16261-16271 is very rare in European Roma populations, being found only in the Spanish Roma (one occurrence). Among Europeans, this haplotype has been revealed only in French (0.5%; Dubut *et al.* 2004) and Czech (2.3%; Vanecek *et al.* 2004) populations. A similar haplotype, lacking only the 16271 transition, has been revealed in a single individual from Bulgaria (Kalaidjii North population) (Gresham *et al.* 2001). It is important that this haplotype has also been found in Baluch and Brahui populations from Southwestern Pakistan at frequencies of 5% and 7.9%, respectively (Quintana-Murci *et al.* 2004). Derived haplotypes, with an additional transition at position 16189, were also described in populations from Syria and Turkey (Richards *et al.* 2000). Therefore, one may suggest that J1-haplotypes characterized by mutation at position 16235 might have been characteristic of the ancestral Romani population.

HVS I sequence type 16145-16223 is another mtDNA haplotype frequent in the Polish Roma (8.7%)

Table 3 Haplogroup distributions (no. of individuals and % values in parentheses) in different populations of the Roma

Haplogroup	Polish Roma (n = 69)	Lithuanian Roma (n = 18)	Spanish Roma (n = 25)	Balkan Roma (n = 71)	Vlax Roma (n = 161)
M	4 (5.8)	4 (22.2)	5 (20.0)	19 (26.8)	45 (28.0)
HV	11 (15.9)	4 (22.2)	3 (12.0)	19 (26.8)	72 (44.7)
T	0	0	0	1 (1.4)	5 (3.1)
J	13 (18.8)	0	3 (12.0)	10 (14.1)	12 (7.5)
U1	0	0	0	1 (1.4)	0
U3	25 (36.2)	10 (55.6)	13 (52.0)	1 (1.4)	4 (2.5)
U5	0	0	1 (4.0)	4 (5.6)	1 (0.6)
K	3 (4.3)	0	0	2 (2.8)	2 (1.2)
N1b	0	0	0	1 (1.4)	4 (2.5)
I	5 (7.2)	0	0	1 (1.4)	4 (2.5)
X	2 (2.9)	0	0	9 (12.7)	12 (7.5)
W	6 (8.7)	0	0	3 (4.2)	0
h (± s.e.)	0.80 ± 0.03	0.63 ± 0.09	0.69 ± 0.08	0.83 ± 0.02	0.71 ± 0.03

but absent in other Romani populations. The results of RFLP typing have shown that this haplotype belongs to haplogroup W, so the absence of the W-specific mutation at position 16292 may be due to a back mutation. An analysis of the database of Richards *et al.* (2000) shows that position 16292 appears to be stable within the haplogroup W members. Nevertheless, recent data has indicated several cases of a back mutation at this position within haplogroup W (Palanichamy *et al.* 2004). Population screening of haplotype 16145-16223 in published data sets from different Eurasian populations has shown a lack of this haplotype. Similar W-haplotypes characterised by a transition at position 16145 were found only in some populations from India (Gujarati) and Pakistan (Sindhi) (according to data of Quintana-Murci *et al.* 2004).

Another relatively frequent haplotype among the Polish Roma (4.3%) is the HVS I sequence 16224-16234-16311 belonging to haplogroup K. This haplotype has not been found up to now in any Romani population (Gresham *et al.* 2001). Population screening has shown that haplotype 16224-16234-16311 is rare among European and Near Eastern populations but is very frequent (24%) in Ashkenazi Jewish populations (Behar *et al.* 2004). This haplotype has not been found in Poles (Malyarchuk *et al.* 2002) but is relatively frequent in the Polish Ashkenazi (7.3%) (Behar *et al.* 2004).

The remaining haplotypes (Tables 2 and 4) found in the Polish Roma belong to haplogroups H, I, and X. These haplotypes have been observed in different populations of Eurasia (Table 4). Among haplogroup

Table 4 Frequency of the mtDNA HVS-I haplotypes (% values in parentheses) found in the Polish Roma in comparison with other European Roma populations

HVS-I sequence (minus 16000)	HG	Polish Roma (n = 69)	Lithuanian Roma (n = 18)	Spanish Roma (n = 25)	Balkan Roma (n = 71)	Vlax Roma (n = 161)
343	U3	25 (36.2)	10 (55.6)	11 (44.0)	1 (1.4)	4 (2.5)
129 223 291 298	M5	3 (4.3)	4 (22.2)	4 (16.0)	7 (9.9)	14 (8.7)
129 223 234 291 298	M5	1 (1.4)	0	0	0	1 (0.6)
261 304	H	7 (10.1)	2 (11.1)	0	3 (4.2)	18 (11.2)
CRS	H	2 (2.9)	0	0	4 (5.6)	2 (1.2)
129 172 223 311 391	I	5 (7.2)	0	0	1 (1.4)	4 (2.5)
126 189A 223 278	X2e	1 (1.4)	0	0	3 (4.2)	9 (5.6)
304	H	1 (1.4)	0	0	0	1 (0.6)
69 126 145 222 235 261 271	J1	13 (18.8)	0	1 (4.0)	0	0
145 223	W	6 (8.7)	0	0	0	0
224 234 311	K	3 (4.3)	0	0	0	0
145A/G 362	H	1 (1.4)	0	0	0	0
183C 189 223 255 278	X2c	1 (1.4)	0	0	0	0

Table 5 Matrix of F_{ST} values from mitochondrial haplogroup frequencies in the Roma populations

	Polish Roma	Lithuanian Roma	Balkan Roma	Vlax Roma	Spanish Roma
Polish Roma		0.060	0.068	0.093	0.040
Lithuanian Roma	0.041*		0.078	0.067	−0.024*
Balkan Roma	0.097	0.153		0.022	0.064
Vlax Roma	0.149	0.178	0.020		0.075
Spanish Roma	0.017*	−0.027*	0.141	0.192	

F_{ST} values based on pairwise differences between HVS I sequences (under diagonal), F_{ST} values based on haplogroup frequencies (below diagonal), * - non-significant differences ($p > 0.05$).

X lineages found in the Roma one specific subcluster defined by a transversion at position 16189 is interesting, since it appears to be non-typical for European populations. This subcluster was not found in the Lithuanian and Spanish Roma, and is rare in the Polish Roma (1.4%), but is common in the Bulgarian Roma groups being found at frequencies of 5.6 and 8.5% in the Vlax and Balkan Roma, respectively (Gresham *et al.* 2001). Recent population screening of haplogroup X diversity in Eurasia and North Africa has shown that this subcluster (within subgroup X2e) is only found in several individuals from southern Europe, but the Roma-specific branch defined by the 16126 transition is virtually absent in Eurasian populations (Reidla *et al.* 2003). The same is true for other X-haplotypes defined by a mutation at position 16241, which were found in the Roma populations (Gresham *et al.* 2001). These haplotypes have been described in only two Russian individuals from South Russia (Malyarchuk *et al.* 2002). In general, only the presence of haplogroup M5 in different Roma populations clearly points to the Asian origin of this founding Romani lineage. Note that according to Y-chromosome variation data, the paternal lineage of Asian origin (similar to maternal M5) identified in all Romani populations is haplogroup H1, defined by the M82 marker (Gresham *et al.* 2001). Thus, the high frequency of several West Eurasian mtDNA haplotypes that are rare or absent in European populations (such as J1, H (261–304), and W) but present in the Polish Roma may be an indication of the effects of genetic drift acting on this population (Table 4).

In order to study differentiation of the Roma populations, an analysis of molecular variance (AMOVA) was performed separately on the level of mtDNA haplogroups and HVS I sequences. The analysis of between-population differentiation based on the frequencies of the mtDNA haplogroups (as shown in Table 3) revealed that 10.7% of variation was due to differences among the Roma populations. Non-significant pairwise F_{ST}-differences ($p > 0.05$) were found only between the Polish Roma and Romani groups from Lithuania and Spain. The AMOVA results of the HVS I sequencing data show that the between-population F_{ST} value based on the pairwise nucleotide differences is high (6.2%). Non-significant differences were only revealed between the Lithuanian and Spanish Roma populations ($p = 0.8$). In general, the data indicate that there is a significant differentiation between different Roma populations – Polish, Lithuanian and Spanish Roma appear to be distinct from the Balkan and Vlax Roma groups.

To further investigate genetic relationships between the Roma groups and the surrounding European populations, additional published data on the mtDNA HVS I variability in different European populations has been used. The AMOVA results show that 4.4% of the variance is due to differences between populations. Pairwise between-population comparisons (Table 6) reveal that F_{ST}-values vary in an interval of 0-0.005 among European populations, 0-0.095 among the Roma populations, and 0.07–0.115 between the Roma groups and surrounding European populations. F_{ST}-differences between Poles and Polish Roma are estimated as 0.094, and the genetic distance between Polish and Vlax Roma groups is almost the same ($F_{ST} = 0.095$). Highly significant differences are observed between Europeans and Roma when they are treated as two separate groups of populations ($F_{CT} = 5.49\%$, $F_{SC} = 1.9\%$, $p = 0$ in both cases). In general, the data indicate that genetic distances between Roma groups are typically

Table 6 Matrix of F_{ST} values derived from mtDNA HVS I sequences in the Roma groups and the surrounding European populations

	LIT	POL	BUL	ROM	SPA	R_POL	R_VLA	R_BAL	R_LIT
LIT	0								
POL	0.0017*	0							
BUL	−0.0003*	0.0026*	0						
ROM	0.0028*	0.0038	0.0007*	0					
SPA	0.0015*	0.0032	−0.0009*	0.0053	0				
R_POL	0.0876	0.0943	0.0949	0.1009	0.0933	0			
R_VLA	0.0829	0.0777	0.0813	0.0877	0.0796	0.0949	0		
R_BAL	0.0708	0.0742	0.0699	0.0783	0.0673	0.0740	0.0228	0	
R_LIT	0.0958	0.0986	0.1005	0.1151	0.1020	0.0591	0.0696	0.0815	0
R_SPA	0.0786	0.0845	0.0803	0.0963	0.0820	0.0433	0.0780	0.0697	−0.0281*

Populations designated as: LIT – Lithuanians, POL - Poles, BUL – Bulgarians, ROM – Romanians, SPA – Spanish, R_POL – Polish Roma, R_VLA – Vlax Roma, R_BAL – Balkan Roma, R_LIT – Lithuanian Roma, R_SPA – Spanish Roma. * - non-significant differences ($p > 0.05$).

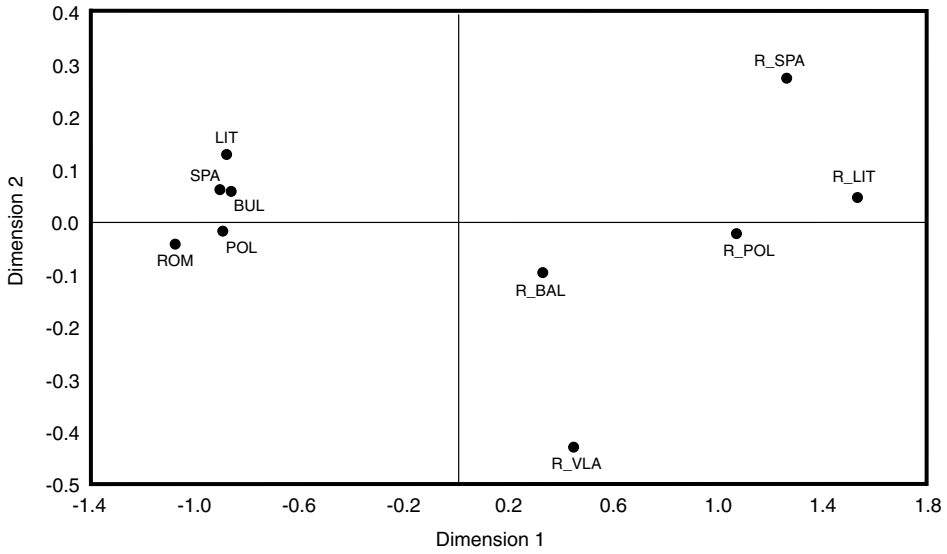

Figure 2 Multidimensional scaling plot of F_{ST} distances between the Roma groups and the surrounding European populations based on mtDNA HVS I variation data (stress value 0.001). Populations designated as in Table 6.

larger than these between the surrounding European populations.

The MDS analysis performed on the basis of pairwise F_{ST} values between European and Romani populations reveals that there is a clear subdivision between the Roma and their European neighbours (Figure 2 and Table 6). Meanwhile, this analysis shows that the Roma populations form two clusters, suggesting the strongest division between the Balkan and Vlax Roma on one hand, and the Polish, Lithuanian and Spanish Roma on the other. The data agree with a previous study of mtDNA differentiation between Roma populations, which showed that the Spanish and Lithuanian Roma are clustered together (Gresham *et al.* 2001). By observing the pattern of distribution of European-specific Y-chromosomal lineages among different Roma populations, it has recently been shown that (i) the degree of admixture between the Roma groups and corresponding "host" populations in Europe is highly variable between different Romani populations, and that (ii) the admixed lineages reflect the lineage distributions within surrounding European populations (Gresham *et al.* 2001; Jobling *et al.* 2004). The analysis of pairwise F_{ST} distances based on mtDNA data (Table 6)

also reveals that admixture events played a significant role in the differentiation of the Roma populations. We have assumed that the pattern of the distribution of pairwise F_{ST} distances between the Roma populations should reflect the pattern observed between the respective pairs of European "host" populations, and quantified this suggestion by Spearman rank correlation analysis. For all pairwise distances compared, the correlation was found to be insignificant ($R = 0.273$, $p = 0.45$). However, the correlation between distributions of pairwise F_{ST} distances was strong ($R = 0.886$, $p = 0.019$) when four pairs of genetic distances (SPA-LIT and R_SPA-R_LIT, SPA-POL and R_SPA-R_POL, SPA-BUL and R_SPA-R_BAL, BUL-LIT and R_BAL-R_LIT) were removed from the analysis. It seems that their negative effect is mostly due to the fact that the Spanish Roma population gives too short distances with the Polish and Lithuanian Roma that are incomparable with genetic distances found between the respective European populations. Thus, the results obtained suggest that, in general, genetic distances between the Roma populations reflect those observed between the "host" European populations. This may indicate that admixture is an important source of genetic differentiation observed between Romani groups; however, the levels of admixture appear to be uneven between different populations.

Conclusions

The previous analysis of the relevance of different criteria (cultural, historical, linguistic, geographic) to the genetic structure of maternal DNA lineages in the Roma performed by Gresham *et al.* (2001) revealed a complex pattern. However, combined analysis of both maternal and paternal lineages allows for the suggestion that classification based on the history of migrations can result in the most highly significant intergroup differences (Gresham *et al.* 2001). It has been indicated that the current genetic structure of the European Roma resulted mainly from early splits and divergent migration routes within Europe (Gresham *et al.* 2001). Genetic drift and different levels and sources of admixture are thought to be two general processes explaining the pattern of observed differentiation of the Roma populations. Our data also indicate that the effects of genetic drift are likely to account for the differences in the distribution of mtDNA lineages in different Romani populations. However, it is difficult to explain the uneven frequency of haplogroup U3 in the Romani populations by the effect of genetic drift alone. Rather, the high frequencies of U3-haplotypes observed in the Polish, Lithuanian and Spanish Roma allow us to suggest that members of a single Roma group migrated independently to the north and southwest of Europe. This scenario is also supported by Y-chromosome data indicating that the Lithuanian and Spanish Roma are characterized by high frequencies (25 and 33%, respectively) of a specific J2f-lineage, defined by the M67 marker (Gresham *et al.* 2001). This lineage has not been shown to be of European populations, but is present in populations from Pakistan, central Asia and the Middle East (Underhill *et al.* 2000; Kivisild *et al.* 2003).

Taking into consideration the pattern of the geographic distribution of mtDNA and Y-chromosome haplotypes, it can be seen that mitochondrial haplogroup M5 and Y-chromosomal haplogroup H1 (defined by M82 marker) represent the genetic composition of the ancestral Roma population. Meanwhile, some DNA haplogroups are more restricted geographically, while some haplotypes correspond to the founding lineages of individual populations (subisolates) within the Roma groups. Thus, further genetic studies will be very useful to examine the population history of the Roma, as well as to reveal individual genetic subisolates suitable for the fine mapping of genes involved in complex disorders.

Acknowledgements

We would like to thank Mr. Przemyslaw W. Radzik for his help in sample collection. We are also very grateful to Ewa Lewandowska and Maria Perkova for their excellent technical assistance. This work was supported by the grants from the Polish State Committee for Scientific Research (3P04C 04823, 2002-2005) (to T.G.) and the Program of Basic Research of Russian Academy of Sciences "Dynamics of plant, animal and human gene pools" (to B.M.).

References

Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., Cruciani, F., Zeviani, M., Briem, E., Carelli,

V., Moral, P., Dugoujon, J. M., Roostalu, U., Loogväli, E.-L., Kivisild, T., Bandelt, H.-J., Richards, M., Villems, R., Santachiara-Benerecetti, A. S., Semino, O. & Torroni, A. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am J Hum Genet* **75**, 910–918.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. M., Staden, R. & Young, I. G. (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147.

Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B., Naidu, J. M., Prasad, B. V., Reddy, P. G., Rasanayagam, A., Papiha, S. S., Villems, R., Redd, A. J., Hammer, M. F., Nguyen, S. V., Carroll, M. L., Batzer, M. A. & Jorde, L. B. (2001) Genetic evidence on the origins of Indian caste populations. *Genome Res* **11**, 994–1004.

Behar, D. M., Hammer, M. F., Garrigan, D., Villems, R., Bonne-Tamir, B., Richards, M., Gurwitz, D., Rosengarten, D., Kaplan, M., Pergola, S. D., Quintana-Murci, L. & Skorecki, K. (2004) MtDNA evidence for a genetic bottleneck in the early history of the Ashkenazi Jewish population. *Eur J Hum Genet* **12**, 355–364.

Belyaeva, O., Bermisheva, M., Khrunin, A., Slominsky, P., Bebyakova, N., Khusnutdinova, E., Mikulich, A. & Limborska, S. (2003) Mitochondrial DNA variations in Russian and Belorussian populations. *Hum Biol* **75**, 647–660.

Chaix, R., Austerlitz, F., Morar, B., Kalaydjieva, L. & Heyer, E. (2004) Vlax Roma history: what do coalescent-based methods tell us? *Eur J Hum Genet* **12**, 285–292.

Dubut, V., Chollet, L., Murail, P., Cartault, F., Beraud-Colomb, E., Serre, M. & Mogentale-Profizi, N. (2004) MtDNA polymorphisms in five French groups: importance of regional sampling. *Eur J Hum Genet* **12**, 293–300.

Ficowski, J., (1986) *Cyganie na polskich drogach [Gypsies on the Polish roads – in Polish]*. Wroclaw: Wydawnictwo Literackie Kraków.

Finnilä, S., Hassinen, I. E., Ala-Kokko, L. & Majamaa, K. (2000) Phylogenetic network of the mtDNA haplogroup U in northern Finland based on sequence analysis of the complete coding region by conformation-sensitive gel electrophoresis. *Am J Hum Genet* **66**, 1017–1026.

Finnilä, S., Lehtonen, M. S. & Majamaa, K. (2001) Phylogenetic network for European mtDNA. *Am J Hum Genet* **68**, 1475–1484.

Gresham, D., Morar, B., Underhill, P. A., Passarino, G., Lin, A. A., Wise, C., Angelicheva, D., Calafell, F., Oefner, P. J., Shen, P., Tournev, I., de Pablo, R., Kicinskas, V., Perez-Lezaun, A., Marushiakova, E., Popov, V. & Kalaydjieva, L. (2001) Origins and divergence of the Roma (Gypsies). *Am J Hum Genet* **69**, 1314–1331.

Herrnstadt, C., Elson, J. L., Fahy, E., Preston, G., Turnbull, D. M., Anderson, C., Ghosh, S. S., Olefsky, J. M., Beal, M. F., Davis, R. E. & Howell, N. (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet* **70**, 1152–1171.

Jobling, M. A., Hurles, M. E. & Tyler-Smith, C. (2004) *Human evolutionary genetics. Origins, peoples & disease*. Garland Science: Garland Publishing.

Kalaydjieva, L., Calafell, F., Jobling, M. A., Angelicheva, D., de Knijff, P., Rosser, Z. H., Hurles, M. E., Underhill, P., Tournev, I., Marushiakova, E. & Popov, V. (2001*a*). 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.

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

Kasperaviciute, D. & Kucinskas, V. (2002) Variability of the human mitochondrial DNA control region sequences in the Lithuanian population. *J Appl Genet* **43**, 255–260.

Kivisild, T., Bamshad, M. J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W. S., Dixon, M. E., Papiha, S. S., Mastana, S. S., Mir, M. R., Ferak, V. & Villems, R. (1999) Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr Biol* **9**, 1331–1334.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H.-V., Stepanov, V., Gölge, M., Usanga, E., Papiha, S. S., Cinnioglu, C., King, R., Cavalli-Sforza, L., Underhill, P. A. & Villems, R. (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet* **72**, 313–332.

Lewis, P. O. & Zaykin, D. (2001) Genetic data analysis: computer program for the analysis of allelic data, version 1.0 (d16c). Free program distributed by the authors over the internet from http://lewis.eeb.uconn.edu/lewishome/software.html.

Loogväli, E.-L., Roostalu, U., Malyarchuk, B. A., Derenko, M. V., Kivisild, T., Metspalu, E., Tambets, K., Reidla, M., Tolk, H.-V., Parik, J., Pennarun, E., Laos, S., Lunkina, A., Golubenko, M., Barac, L., Pericic, M., Balanovsky, O. P., Gusar, V., Khusnutdinova, E. K., Stepanov, V., Puzyrev, V., Rudan, P., Balanovska, E. V., Grechanina, E., Richard, C., Moisan, J. P., Chaventre, A., Anagnou, N. P., Pappa,

K. I., Michalodimitrakis, E. N., Claustres, M., Golge, M., Mikerezi, I., Usanga, E. & Villems, R. (2004) Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Mol Biol Evol* **21**, 2012–2021.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-Tamir, B., Sykes, B. & Torroni, A. (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* **64**, 232–249.

Malyarchuk, B. A., Grzybowski, T., Derenko, M. V., Czarny, J., Drobnic, K. & Miscicka-Sliwka, D. (2003) Mitochondrial DNA variability in Bosnians and Slovenians. *Ann Hum Genet* **67**, 412–425.

Malyarchuk, B. A., Grzybowski, T., Derenko, M. V., Czarny, J., Wozniak, M. & Miscicka-Sliwka, D. (2002) Mitochondrial DNA variability in Poles and Russians. *Ann Hum Genet* **66**, 261–283.

McEvoy, B., Richards, M., Forster, P. & Bradley, D. G. (2004) The *longue duree* of genetic ancestry: Multiple genetic marker systems and Celtic origins on the Atlantic facade of Europe. *Am J Hum Genet* **75**, 693–702.

Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K., Serk, P., Karmin, M., Behar, D. M., Gilbert, M. T., Endicott, P., Mastana, S., Papiha, S. S., Skorecki, K., Torroni, A. & Villems, R. (2004) Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet* **31**, 26.

Morar, B., Gresham, D., Angelicheva, D., Tournev, I., Gooding, R., Guergueltcheva, V., Schmidt, C., Abicht, A., Lochmuller, H., Tordai, A., Kalmar, L., Nagy, M., Karcagi, V., Jeanpierre, M., Herczegfalvi, A., Beeson, D., Venkataraman, V., Carter, K. W., Reeve, J., de Pablo, R., Kucinskas, V. & Kalaydjieva, L. (2004) Mutation history of the Roma/Gypsies. *Am J Hum Genet* **75**, 596–609.

Nei, M. & Tajima, F. (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics* **97**, 145–163.

Palanichamy, M. G., Sun, C., Agrawal, S., Bandelt, H.-J., Kong, Q.-P., Khan, F., Wang, C.-E., Chaudhuri, T. K., Palla, V. & Zhang, Y.-P. (2004) Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: Implications for the peopling of South Asia. *Am J Hum Genet* **75**, 966–978.

Rajkumar, R., Banerjee, J., Gunturi, H. B., Trivedi, R. & Kashyap, V. K. (2005) Phylogeny and antiquity of M macrohaplogroup inferred from complete mtDNA sequence of Indian specific lineages. *BMC Evol Biol* **5**, 26.

Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., Rengo, C., Al-Zahery, N., Semino, O., Santachiara-Benerecetti, A. S., Coppa, A., Ayub, Q., Mohyuddin, A., Tyler-Smith, C., Mehdi, S. Q., Torroni, A. & McElreavey, K. (2004) Where West meets East: The complex mtDNA landscape of the southwest and central Asian corridor. *Am J Hum Genet* **74**, 827–845.

Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K. & Santachiara-Benerecetti, A. S. (1999) Genetic evidence for an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet* **23**, 437–441.

Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk, H.-V., Parik, J., Loogvali, E.-L., Derenko, M., Malyarchuk, B., Bermisheva, M., Zhadanov, S., Pennarun, E., Gubina, M., Golubenko, M., Damba, L., Fedorova, S., Gusar, V., Mikerezi, I., Moisan, J.-P., Khusnutdinova, E., Osipova, L., Stepanov, V., Voevoda, M., Achilli, A., Rengo, C., Rickards, O., de Stefano, G. F., Papiha, S., Beckman, L., Janicijevich, B., Rudan, P., Anagnou, N., Koziel, S., Usanga, E., Geberhiwot, T., Herrnstadt, C., Howell, N., Torroni, A. & Villems, R. (2003) Origin and diffusion of mtDNA haplogroup X. *Am J Hum Genet* **73**, 1178–1190.

Richards, M. B., Macaulay, V. A., Bandelt, H.-J. & Sykes, B. C. (1998) Phylogeography of mitochondrial DNA in western Europe. *Ann Hum Genet* **62**, 241–260.

Richards, M. B., Macaulay, V. A., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellito, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Yu., Golge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., DiRienzo, A., Novelletto, A., Oppenheim, A., Norby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Scozzari, R., Torroni, A. & Bandelt, H.-J. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* **67**, 1251–1276.

Schneider, S., Roessli, D. & Excoffier, L. (2000) *Arlequin ver.2.0: A software for population genetics data analysis*. Switzerland: Genetics and Biometry laboratory, University of Geneva.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M.-L. & Wallace, D. C. (1996) Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**, 1835–1850.

Torroni, A., Rengo, C., Guida, V., Cruciani, F., Sellitto, D., Coppa, A., Calderon, F. L., Simionati, B., Valle, G., Richards, M., Macaulay, V. & Scozzari, R. (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* **69**, 1348–1356.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L. & Oefner, P. J. (2000) Y-chromosome sequence variation and the history of human populations. *Nat Genet* **26**, 358–361.

Vanecek, T., Vorel, F. & Sip, M. (2004) Mitochondrial DNA D-loop hypervariable regions: Czech population data. *Int J Legal Med* **118**, 14–18.

Yao, Y.-G., Kong, Q.-P., Bandelt, H.-J., Kivisild, T. & Zhang, Y.-P. (2002) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* **70**, 635–651.

Zhivotovsky, L. A., Underhill, P., Cinnioglu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G., Spedini, G., Chambers, G. K., Herrera, R. J., Yong, K. K., Gresham, D., Tournev, I., Feldman, M. W. & Kalaydjieva, L. (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet* **74**, 50–61.

Received: 1 February 2005
Accepted: 27 June 2005

Supplementary Material

The following material is available for this article online:

Table S1. Autosomal STR genotypes of Roma individuals belonging to particular matrilineal groups (mtDNA halpogroups).

Table S2. Inbreeding coefficient values based on autosomal STR data within matrilineal groups of the Polish Roma individuals.

