MBE Advance Access published May 13, 2008

Research Article:

Mitochondrial DNA phylogeny in Eastern and Western Slavs

B. Malyarchuk[1,*], T. Grzybowski[2], M. Derenko[1], M. Perkova[1], T. Vanecek[3], J. Lazur[4], P. Gomolcak[5], I. Tsybovsky[6]

[1] Institute of Biological Problems of the North, Far-East Branch of the Russian Academy of Sciences, Magadan, Russia

[2] The Nicolaus Copernicus University, Ludwik Rydygier Collegium Medicum, Institute of Forensic Medicine, Department of Molecular and Forensic Genetics, Bydgoszcz, Poland

[3] Department of Pathology, Medical Faculty Hospital, Charles University, Pilsen, Czech Republic

[4] Department of Laboratory Medicine, LABMED, Kosice, Slovakia

[5] Institute of Pathology, Slovak Medical University, Bratislava, Slovakia

[6] Institute of Problems of Criminalistics and Forensic Expertise, Minsk, Belarus

Correspondence:

Boris Malyarchuk, Institute of Biological Problems of the North, Far-East Branch of the Russian Academy of Sciences, Portovaya str., 18, Magadan, 685000, Russia

Tel: +7 4132 63 11 64, Fax: +7 4132 63 44 63, e-mail: malyarchuk@ibpn.ru

Running Title: mtDNA phylogeny in Slavs

Key words: human mitochondrial DNA, complete genome sequencing, population genetics, molecular phylogeography, Eastern Europe, Slavonic populations

© The Author 2008. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org

Abbreviations: CRS, Cambridge reference sequence; HVS1, first hypervariable segment; HVS2, second hypervariable segment; LGM, Last Glacial Maximum; mtDNA, mitochondrial DNA; np(s), nucleotide position(s); RFLP, restriction fragment length polymorphism; YBP, years before present.

Abstract

To resolve the phylogeny of certain mitochondrial DNA (mtDNA) haplogroups in Eastern Europe and estimate their evolutionary age, a total of 73 samples representing mitochondrial haplogroups U4, HV*, and R1 were selected for complete mitochondrial genome sequencing from a collection of about 2000 control-region sequences sampled in Eastern (Russians, Belorussians, Ukrainians) and Western (Poles, Czechs and Slovaks) Slavs. On the basis of whole-genome resolution, we fully characterized a number of haplogroups (HV3, HV4, U4a1, U4a2, U4a3, U4b, U4c, U4d, and R1a) that were previously described only partially. Our findings demonstrate that haplogroups HV3, HV4, and U4a1 could be traced back to the pre-Neolithic times (~ 12,000-19,000 YBP) in Eastern Europe. In addition, an ancient connection between the Caucasus/Europe and India has been revealed by analysis of haplogroup R1 diversity, with a split between the Indian and Caucasus/European R1a lineages occurring about 16,500 years ago. Meanwhile, some mtDNA subgroups detected in Slavs (such as U4a2a, U4a2*, HV3a, R1a1) are definitely younger being dated between 6,400-8,200 YBP. However, robust age estimations appear to be problematic due to the high ratios of non-synonymous to synonymous substitutions found in young mtDNA subclusters.

Introduction

According to archaeological data, the eastern part of Europe was colonized by modern humans about 42,000-45,000 YBP (years before present), probably as early as anywhere else in northern Eurasia (Anikovich et al. 2007). Genetic data suggest that maternal mitochondrial DNA (mtDNA) lineages can be traced farther back into prehistory, through the Last Glacial Maximum (LGM), to the first settlement of Europe by anatomically modern humans almost 50,000 YBP (Richards et al. 2000). The LGM took place between 19,000-22,000 YBP, and during this interval European population was most probably concentrated in refugia in the western Caucasus and southern European peninsulas (Peyron et al. 1998; Tarasov et al. 2000). Thus, a model of postglacial expansion-recolonization from refugia appears to be a major concept to explain the genetic diversity of the present-day Europeans (Torroni et al. 1998; Rootsi et al. 2004; Roostalu et al. 2007).

A founder analysis of mtDNA lineages in populations of Near East and Europe has demonstrated that the majority of extant mtDNA lineages entered Europe in several episodes during the Upper Palaeolithic (Richards et al. 2000). Neolithic founder clusters (9,000 YBP) were mostly members of haplogroups J, T1, and U3. The main contributors to the late Upper Paleolithic expansions (14,500 YBP) were the major subclusters of haplogroups H, K, T*, T2, W, and X. The main components of the middle Upper Paleolithic (26,000 YBP) were HV*, U1, possibly U2, and U4, and the main component of the early Upper Paleolithic (45,000 YBP) was mainly haplogroup U5 (Richards et al. 2000). The age estimates for the major mtDNA clusters in Europe and Near East are consistent with assumption that these clusters are considerably older in the Near East, but there is also the possibility that significant dispersals may have originated in northern Caucasus and Eastern Europe, as has been suggested for the middle Upper Paleolithic component (Soffer 1987). It is known that mtDNA

haplogroup U4 is present in Finno-Ugric and Turkic-speaking populations of Eastern Europe at frequency of about 9%, whereas HV* is infrequent, about 0.5% (Bermisheva et al. 2002). In Russian populations, the picture is somewhat different – haplogroups U4 and HV* are present, on average, at frequencies of 3.5% and 2.5%, respectively (Malyarchuk et al. 2002a; 2004). These values are comparable with those described for other European populations (Richard et al. 2007).

High mutation rates and homoplasy in the mtDNA control region appear to be an obstacle to developing a reliable classification of mitochondrial lineages and in identifying the correct dating of mtDNA clades (Bandelt et al. 2002; Malyarchuk et al. 2002b). Hence, complete mtDNA sequences may represent the best possible solution for phylogeographic analysis (Torroni et al. 2006). However, until now, the reconstruction of the phylogeny of haplogroups U4 and HV* has been only partially performed at the level of complete-genome resolution (Achilli et al. 2004; Palanichamy et al. 2004; Achilli et al. 2005). Therefore, the objective of this study is to provide new information concerning the molecular dissection of haplogroups U4 and HV* present in Eastern European populations, such as Russians, Belorussians, Ukrainians, Poles, Czechs and Slovaks. Moreover, we present here the results of the complete genome characterization of the R1 haplogroup (one rarely seen in Europe) encouraged by its occurrence in the North-Western Russian populations (Malyarchuk et al. 2004; Grzybowski et al. 2007). The vast majority of the complete genomes displayed here belong to novel subhaplogroups, informative in terms of updating the West Eurasian phylogeny and the understanding of some aspects of the genetic history of Europeans.

Materials and methods

Out of about 2000 samples that had been screened previously for haplogroup-diagnostic RFLP markers and subjected to control region sequencing (Malyarchuk and Derenko 2001;

Malyarchuk et al. 2002a; 2004; 2006; 2008; Grzybowski et al. 2007), a total of 73 samples representing mitochondrial haplogroups U4 (49 samples), HV* (19 samples) and R1 (5 samples) were selected for complete mitochondrial genome sequencing (see Table S1 in Supplementary Material online). These samples were selected from the populations of Western (Poles, Czechs and Slovaks) and Eastern Slavs (Russians, Belorussians and Ukrainians). In the Belorussian sample (n = 247), only the U4 haplogroup diversity has been studied.

Complete mtDNA sequencing was performed using the methodology described in detail by Torroni et al. (2001). DNA samples representing subhaplogroup U4a2, harboring T>C transition at nucleotide position (np) 310, were additionally subjected to amplification with primer pair F15878/R649 and sequenced with the same primers, using the protocol of Brandstätter et al. (2004). Mutations were scored relative to the revised Cambridge reference sequence (rCRS) (Andrews et al. 1999).

Most-parsimonious trees of the complete mtDNA sequences were reconstructed manually and verified by means of the Network 4.5 program (www.fluxus-engineering.com).

For the tree reconstructions, the data received in this study and those published previously by Finnilä et al. (2001), Palanichamy et al. (2004), Achilli et al. (2004; 2005) and Kivisild et al. (2006) were used. Nucleotide positions showing point indels and transversions located between nps 16180-16193 and 303-315 were excluded from the phylogenetic analysis.

Coalescence time calculations were performed using the *rho* statistic, taking one base substitution between nps 577-16023 equal to 5,140 years (Mishmar et al. 2003). For synonymous substitutions we used the rate of one substitution in 6,764 years (Kivisild et al. 2006). This rate has been used in cases where excess non-synonymous substitutions were detected within mtDNA haplogroups. Standard deviation of the *rho* estimate (s.e.) was calculated as in Saillard et al. (2000).

The M_N/M_S ratio estimating the number of mutational changes inferred from the phylogenetic tree was used to measure non-synonymous (N) to synonymous (S) substitution ratios (Kivisild et al. 2006). In addition, K*a*/(K*a* + K*s*) ratios were calculated using DnaSP 4.20.2 software (Rozas et al. 2003), where K*a* is the number of non-synonymous substitutions per non-synonymous site and K*s* is the number of synonymous substitutions per synonymous site (Nei and Gojobori 1986). Chi-square analysis of haplogroup frequencies in populations was performed by means of the program CHIRXC, which estimates the probability of homogeneity using Monte Carlo simulation (1000 runs) (Zaykin and Pudovkin 1993).

Results

Updating the mitochondrial haplogroup U4 phylogeny

Phylogenetic classification of mtDNAs belonging to haplogroup U4 has been based thus far mainly on the presence of diagnostic control region motifs (Richards et al. 1998; 2000; Tambets et al. 2003). As shown in Figure 1, the complete genome data from Eastern European populations allow us to refine the haplogroup U4 phylogeny. Within U4a, characterized previously by the 8818 transition (Achilli et al. 2005), three subclades can be recognized – U4a1, with the transition motif 152-12937-16134, U4a2 defined by the control region transition at np 310, and U4a3 that has been found only once in our study in Czech population and, therefore, cannot be defined for the time present by minimal nucleotide motif. Haplogroup U4a1 is largely characterized by an Eastern European and West Siberian distribution being found at highest frequencies (7-21%) in such populations as Mari, Chuvash and Kets (Table 1, Table S2). We investigated the structure of haplogroup U4a1 by complete genome sequencing of 12 mtDNAs from populations of Poles, Czechs, Slovaks, Belorussians and Russians. The resulting tree (Figure 1) indicates that there are at least three subclusters within haplogroup U4a1 – U4a1a, with transition at np 961 and insertion of three C's at np

965, and its subclade U4a1a1 defined by additional transitions at nps 8167 and 12618 and insertion of one C at np 5899; U4a2 defined by transitions at nps 745 and 3204; and U4a1c characterized by a rank of mutations at nps 8155, 13158, 14110 and 16234. Interestingly, among Eastern Slavs (Russians and Belorussians) the only specific subclade U4a1a1 has been revealed, whereas in Western Slavs the remaining subclades have been found. The coalescence time estimate for U4a1 complete genomes was 14,650 ± 2,400 YBP. This value is close to those (around 10,000-14,000 YBP) calculated from the HVS1 data for Central European (Germanic-speaking) populations, but not for the Baltic Finno-Ugric and Volga people (around 20,000-22,000 YBP) (Tambets et al. 2003).

Complete genome sequencing of 26 mtDNAs has shown that haplogroup U4a2 comprises at least three subclades – U4a2a, with a back-mutation at np 16356, U4a2b characterized by transition at np 16223 and U4a2c, with diagnostic coding region transitions at np 8567. Its major subcluster U4a2c1 (marked by the variants 10654, 16242A, 16288, and 16362) was observed at elevated frequencies (0.9-3.6%) in Turkic-speaking groups of the Volga basin (Bermisheva et al. 2002), whereas earlier branch of U4a2c was detected in our study in the Belorussian population (SV8 in Table S1; Figure 1). In addition, a number of unclassified U4a2*-haplotypes have been revealed, mainly among Slovaks and Russians (Figure 1).

Based on the combined presence of transitions at nps 7705 and 11339, Achilli et al. (2005) identified haplogroup U4b on the basis of complete mtDNA of an Adygei individual. In our samples from Eastern Europe we completely sequenced four samples with these polymorphisms. Interestingly, one of these samples was lacking the 11339 transition. Therefore we propose to re-define haplogroup U4b by the 7705 transition and nominate lineages inside U4b with the transition at np 11339 as U4b1.

In our study, we identified a novel haplogroup U4c, which can be recognized by the 10907 transition. This transition is among of the root markers of the clade (at positions 4811, 6146,

9070, 10907, and 14866) represented in the MitoMap mtDNA tree (Ruiz-Pesini et al. 2007) by several coding region sequences reported previously by Herrnstadt et al. (2002) and Kivisild et al. (2006). For this reason, we propose to name U4c-subclade defined by mutations at positions 4811, 6146, 9070 and 14866 as U4c1. Another new haplogroup is U4d, which is recognizable by two mutations – a transition at np 629 and insertion of one C at np 2405. However, the contemporary geographic distribution of this haplogroup is uncertain due to lack of the control region markers.

Using the calibration method of Mishmar et al. (2003), the coalescence age for haplogroup U4 is 20,460 ± 1,300 YBP, which is consistent with the age of U4 (around 16,000-24,000 YBP) estimated previously by Richards *et al.* (2000) on the basis of HVS1 sequence variation. Coalescence time estimates for haplogroups U4a and U4a2 were 12,100 ± 1,200 and 7,300 ± 1,180 YBP, respectively. Within U4a2, subclades U4a2a (12 genomes) and U4a2* (8 genomes) coalesce at 6,425 ± 1,640 and 7,068 ± 2,134 YBP, respectively. However, one should note that there are considerable differences between U4 clades in the non-synonymous versus synonymous mutations ratio (Table 2, Table S3). This ratio as estimated on the tree (Figure 1) varies from 0.2-0.5 in haplogroups U4d, U4b, U4c, and U4a1 to 1.1 in U4a2. Previous studies have shown that the number of non-synonymous substitutions increases from the average of 0.4 in "older" mtDNA clades to 0.62 in "younger" ones (Kivisild et al. 2006). Hence, a significant excess of non-synonymous mutations in U4a2 can be explained by its young age. Using a mutation rate counting only synonymous substitutions (Kivisild et al. 2006), the age for U4a2 is only 3,100 ± 880 YBP, that is broadly consistent with the beginning of Slavonic prehistory, according to data of archaeology and linguistics (Gimbutas 1971; Sedov 1979).

Haplogroups HV3 and HV4 phylogeny

The majority of West Eurasian mtDNA haplogroups consist of a small number of phylogenetically well-characterized branches of haplogroup R, such as JT, U and HV. Among them, haplogroup HV defined by substitutions at nps 73, 11719 and 14766 in relation to R*-root is represented by haplogroups, H, HV0 (including V) and several less studied branches, namely HV1, HV2, HV3 and still unclassified HV*. It has been suggested that most of the HV-haplogroups presently found in Europe originated in the Near East and Caucasus region (Richards et al. 2000; Tambets et al. 2000), but there are still many questions concerning classification of haplogroups belonging to HV-family.

In the Eastern European populations studied, haplogroup HV1 defined by HVS1 region transition at np 16067 has been revealed at low frequency (0.2%) only in Poles (Malyarchuk et al. 2002a). It is known that in Europe low frequencies of this haplogroup can be seen in populations of Romania, Bulgaria and Hungary (Richards et al. 2000; Egyed et al. 2007). Haplogroup HV2 defined by HVS1 region mutation at np 16217 is also very rare in European populations, being found in single instances in such populations as Slovaks and Bosnians (Malyarchuk et al. 2003; 2008), but in the Indo-Pakistani region its frequency is very high (about 10%) (Metspalu et al. 2004; Quintana-Murci et al. 2004). In a large Russian population, we have not detected any HV1 and HV2 sequences, but have found many haplotypes defined by the HVS1 region mutation at np 16311. These haplotypes are assigned to the haplogroup HV3 as suggested by Metspalu et al. (2004), although such grouping may be artificial due to mutational instability of np 16311 (Bandelt et al. 2002; Malyarchuk et al. 2002b). HV3 haplotypes have been observed with low frequencies (<1%) in populations of north-central and north-western Europe, whereas its contribution to the mtDNA pools of Mediterranean and south Eastern Europeans appears to be slightly higher (1-1.3%) (Richards et al. 2000). The highest frequencies of HV3 (2.3-5.8%) have been detected thus far in some populations of the Near East, Anatolia, Iran and Iraq (Richards et al. 2000; Al-Zahery et al.

2003; Metspalu et al. 2004; Quintana-Murci et al. 2004). In northeastern Europe, the highest frequency of HV3 (2.9%) has been observed in Latvians, although it is represented there only by the nodal HVS1 haplotypes (Pliss et al. 2006). Similarly, individual nodal HVS1 haplotypes belonging to the HV3 have been encountered in several populations of the Volga basin (Bermisheva et al. 2002). Among the Slavs, haplogroup HV3 contributes equally to the Russian, Czech and Slovak mtDNA pools (in average, 1.8%), but essentially less (0.5%) to the gene pool of Poles (these differences are statistically significant only for Russian-Polish comparison, P = 0.015).

Complete mitochondrial genome data from Eastern European populations allows refinement of the haplogroup HV phylogeny. As shown in Figure 2, except for HV3, other haplogroups can be identified – HV4, characterized by the coding region transition at np 7094, and its subcluster HV4a, which can be recognized by the coding region transition at np 709. A novel monophyletic subclade defined by transition at np 8994 emerged within HV3 and is called HV3a here. The mitochondrial genome #As37 has been assigned to HV3a only conditionally (based on transition at np 8994) since only the coding region information is available for this sample and the presence of transition at np 16311 is not verified (Kivisild et al. 2006). In addition, another new clade inside haplogroup HV3, HV3b, is recognizable by two transitions at nps 6755 and 16172. A major subcluster within HV3b is HV3b1 defined by transversion from A to C at np 16113 and three transitions at nps 3507, 8545 and 9266. We have found also that Russian HVS1 sequences with motif 16278-16311 are clustered in subgroup HV3c by transitions at nps 5471 and 14560. One more subcluster, HV3d joins two samples (Russian and Slovak ones) by the 16354 transition in HVS1. The remaining completely sequenced HV3-lineages simply branch off from the HV3-root.

The coalescence age of the entire HV3 defined by a transition at np 16311 is 12,700 ± 1,960 YBP. We should note, however, that HV3 might be polyphyletic due to the hypervariability

of np 16311. Meanwhile, subclusters HV3a and HV3b are likely monophyletic and their age estimates were 8,200 ± 2,900 and 15,420 ± 4,500 YBP, respectively. However, HV3b is characterized by a high ratio of non-synonymous vs. synonymous substitutions, so its coalescence age can be estimated as 11,837 ± 4,460 YBP, using the rate suggested by Kivisild et al. (2006) (Table 2). The same is probably true for the haplogroup HV4 age estimation.

Refinement of the haplogroup R1 phylogeny

There is very little complete mitochondrial sequence data concerning haplogroup R1, very rarely observed in European populations (Richards et al. 2000). In Eastern Europe, R1 haplotypes were encountered only in northwestern Russians and Poles (Malyarchuk et al. 2004; Grzybowski et al. 2007). To date, only one complete mtDNA sequence belonging to this clade has been published, from the Brahmin population of India (sample C134; Palanichamy et al. 2004). Five additional genomes presented here allow us to refine the R1 phylogeny (Figure 2). The root of R1 can now be defined by fifteen coding region mutations, while three transitions at nps 4026, 5378 and 7424 separate subcluster R1a from haplotypes characterized by transitions at nps 14162, 15497 and 16278. In turn, subcluster R1a consists of two clades – an Indian one represented by sample C134 (Palanichamy et al. 2004) and the Caucasus/European one, R1a1, which is defined by the 13105 and 13368 transitions. It should be noted that haplogroup R1 was described for the first time in Adygei from the Northern Caucasus (Macaulay et al. 1999). Its presence, albeit at low frequencies, was then confirmed in some populations of the Caucasus (e.g., in the Kabardins which are linguistically related to the Adygei people) (Nasidze and Stoneking 2001), the Near East (Richards et al. 2000; Rowold et al. 2007) and the South Caspian region (in Iran and Turkmenistan) (Metspalu et al. 2004; Quintana-Murci et al. 2004).

It has been suggested that haplogroup R1, as well as other haplogroups rarely observed in European populations (R2, N1a) were brought to Europe from the Near East in the Neolithic times (Torroni et al. 2006). Meanwhile, the complete sequencing of R1 mtDNAs suggests a deep split between the more ancient 16278-16311 R1-branch and the R1a subcluster, about 28,300 ± 4,900 YBP (Figure 2). A second split (about 16,450 ± 4,100 YBP) is seen between the Indian haplotype C134 and the R1a1-subcluster. Both R1-types are present in Adygei population of the Northern Caucasus (Macaulay et al. 1999), thus suggesting that R1 evolution occurred in the Caucasus area, from where these lineages have extended in different directions. It is known that the Adygei people (the Adygei, Cherkess and Kabardins) are one of the most ancient indigenous populations of the Caucasus region. Therefore, the presence of R1-haplotypes in populations of northwestern Russians can be explained by the contribution of Northern Caucasus populations to the Russian gene pool.

Discussion

Although the previous studies of mtDNA diversity in European populations suggested that both haplogroup frequency patterns and haplotype composition in Slavs were similar to those characteristic of other Europeans, complete genome sequencing allowed us to describe the specific mtDNA components which are found predominantly in modern-day Slavs, albeit introduced at different times into European gene pools. The results of this study show that some of them (HV3, HV4, R1, U4a1) could possibly be traced back to the pre-Neolithic in Eastern Europe and the Caucasus area. These mtDNA lineages were probably involved in the late-glacial expansions from Eastern European refugia after the LGM. The estimated split between the Indian and Caucasus/European R1a-branches dated as ~16,500 YBP confirms this idea, but more data on the diversity of haplogroup R1in different Eurasian populations are needed.

Meanwhile, some of mtDNA subgroups widespread in Slavonic populations (such as U4a2a, U4a2*, HV3a, R1a1) are definitely younger (dated between 6,400-8,200 YBP), suggesting that their expansions may be related to more recent historical events. Yet the most important component from the viewpoint of ethnic history of Slavs is mitochondrial subcluster U4a2, most probably of a central-eastern European origin. Expansion of this subcluster may be explained by a dispersal of the Corded Ware culture which flourished 5,200-4,300 YBP in Eastern and Central Europe. This culture (also known as the Battle-Axe culture) encompassed most of continental northern Europe from the Volga River in the east to the Rhine River on the west (Gimbutas 1971; Mallory 1989).

However, the results of genetic dating should be interpreted with caution because of the problems with the rate of mutations and possible influence of selection on the evolution of mtDNA (Torroni et al. 2001; Mishmar et al. 2003). Haplogroups U4a2, HV3b, HV4 and R1 demonstrate a substantial fraction of non-synonymous mutations whose proportion to synonymous ones varies from 0.63 to 1.1. For instance, K*a*/(K*a* + K*s*) ratio for 25 U4a2 genomes showing the highest M_N/M_S ratio was 0.31, whereas it was only 0.17 for all 49 U4 genomes studied. Meanwhile, within haplogroup U4 protein-coding genes the pattern of non-synonymous and synonymous changes was different, because K*a*/(K*a* + K*s*) ratios varied from 0.46 and 0.27 in *ND5* and *ATP6* to 0.1 and 0.03 in *ND4* and *CO1*, respectively. It is well-known that the *ATP6* gene has the highest level of non-synonymous substitutions in global mtDNA comparisons (Mishmar et al. 2003; Elson et al. 2004), but *ND5* is usually less variable and even conservative, as in haplogroup J (Moilainen et al. 2003). In this regard, we hope that further extending of the haplogroup U4 complete genome data sets will help to solve the question of whether or not adaptive selection acts on the *ND5* gene in European haplogroup U4.

Due to the relatively high rates of non-synonymous to synonymous substitutions detected in some mtDNA subclusters, the mutation rate counting only synonymous substitutions can be used in such cases for the time estimates (Kivisild et al. 2006). This approach allowed lower estimations of the coalescence times for certain mtDNA clusters (Table 2), pointing even to the possibility that the flowering of some mtDNA clades (such as U4a2 and U4a2a) might be due to the recent expansion (~ 3,000 YBP) of pre-Slavonic tribes in Central and Eastern Europe. Therefore, it seems crucial for the future to develop general principles for using a mtDNA coding-region molecular clock in dating events in human population history.

Supplementary Material

Supplementary tables S1 (control-region variation of the completely sequenced mtDNAs), S2 (population frequencies of haplogroups U4a1 and U4a2) and S3 (characterization of nucleotide substitutions in completely sequenced mtDNAs) are available at the journal's website. All completely sequenced mitochondrial genomes have been submitted to the GenBank database, under accession numbers EF222232-EF222253, EU545415-EU545465.

Acknowledgments

We are very grateful to Ewa Lewandowska for her excellent technical assistance. This research was supported by Russian Foundation for Basic Researches (grant No. 06-04-48136), the Programme of Russian Academy of Sciences Presidium "Biodiversity and Gene Pools Dynamics" and Collegium Medicum of the Nicolaus Copernicus University (Bydgoszcz, Poland) (grant No. BW 34/06).

Literature Cited

Achilli A, Rengo C, Battaglia V, et al. (13 co-authors). 2005. Saami and Berbers - an unexpected mitochondrial DNA link. Am. J. Hum. Genet. 76: 883-886.

Achilli A, Rengo C, Magri C, et al. (21 co-authors). 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.

Al-Zahery N, Semino O, Benuzzi G, Magri C, Passarino G, Torroni A, Santachiara-Benerecetti AS. 2003. Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. Mol. Phylogenet. Evol. 28: 458-472.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 23: 147.

Anikovich MV, Sinitsyn AA, Hoffecker JF, et al. (15 co-authors). 2007. Early Upper Paleolithic in Eastern Europe and implications for the dispersal of modern humans. Science. 315: 223-226.

Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V. 2002. The fingerprint of phantom mutations in mitochondrial DNA data. Am. J. Hum. Genet. 71: 1150-1160.

Bermisheva M, Tambets K, Villems R, Khusnutdinova E. 2002. Diversity of mitochondrial DNA haplotypes in ethnic populations of the Volga-Ural region of Russia. Mol. Biol. (Mosk.) 36: 990-1001.

Brandstätter A, Peterson CT, Irwin JA, Mpoke S, Koech DK, Parson W, Parsons TJ. 2004. Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. Int. J. Legal Med. 118: 294-306.

Egyed B, Brandstätter A, Irwin JA, Pádár Z, Parsons TJ, Parson W. 2007. Mitochondrial control region sequence variations in the Hungarian population: Analysis of population

samples from Hungary and from Transylvania (Romania). Forensic Sci. Int.: Genetics. 1: 158-162.

Elson JL, Turnbull DM, Howell N. 2004. Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. Am. J. Hum. Genet. 74: 229-238.

Finnilä S, Lehtonen MS, Majamaa K. 2001. Phylogenetic network for European mtDNA. Am. J. Hum. Genet. 68: 1475-1484.

Gimbutas M. 1971. The Slavs. New York, NY: Praeger Publishing.

Grzybowski T, Malyarchuk BA, Derenko MV, Perkova MA, Bednarek J, Woźniak M. 2007. Complex interactions of the Eastern and Western Slavonic populations with other European groups as revealed by mitochondrial DNA analysis. Forensic Sci. Int.: Genetics. 1: 141-147.

Herrnstadt C, Elson JL, Fahy E, et al. (11 co-authors). 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.

Kivisild T, Shen P, Wall DP, et al. (17 co-authors). 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373-387.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonné-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.

Mallory JP. 1989. In search of the Indo-Europeans. London: Thames & Hudson.

Malyarchuk BA, Derenko MV (2001) Mitochondrial DNA variability in Russians and Ukrainians: implication to the origin of the Eastern Slavs. Ann. Hum. Genet. 65: 63-78.

Malyarchuk B, Derenko M, Grzybowski T, Lunkina A., Czarny J., Rychkov S., Morozova I., Denisova G., Miścicka-Śliwka D. 2004. Differentiation of mitochondrial DNA and Y chromosomes in Russian populations. Hum. Biol. 76: 877-900.

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Drobnic K, Miścicka-Śliwka D. 2003. Mitochondrial DNA variability in Bosnians and Slovenians. Ann. Hum. Genet. 67: 412-425.

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Woźniak M, Miścicka-Śliwka D. 2002a. Mitochondrial DNA variability in Poles and Russians. Ann. Hum. Genet. 66: 261-283.

Malyarchuk BA, Perkova MA, Derenko MV, Vanecek T, Lazur J, Gomolcak P. 2008. Mitochondrial DNA variability in Slovaks, with application to the Roma origin. Ann. Hum. Genet. 72: 228-240.

Malyarchuk BA, Rogozin IB, Berikov VB, Derenko MV. 2002b. Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. Hum. Genet. 111: 46-53.

Metspalu M., Kivisild T, Metspalu E, et al. (16 co-authors). 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. 5: 26.

Mishmar D, Ruiz-Pesini E, Golik P, et al. (13 co-authors). 2003. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA 100: 171-176.

Moilanen JS, Majamaa K. 2003. Phylogenetic network and physicochemical properties of nonsynonymous mutations in the protein-coding genes of human mitochondrial DNA. Mol. Biol. Evol. 20: 1195-1210.

Nasidze I, Stoneking M. 2001. Mitochondrial DNA variation and language replacements in the Caucasus. Proc. Biol. Sci. 268: 1197-1206.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3: 418-426.

Palanichamy MG, Sun C, Agrawal S, Bandelt H-J., Kong QP, Khan F, Wang CY, Chaudhuri TK, Palla V, Zhang YP. 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.

Peyron O, Guiot J, Cheddadi R, Tarasov P, Reille M, de Beaulieu J-L, Bottema S, Andrieu V. 1998. Climatic reconstruction in Europe for 18,000 Y.B.P. from pollen data. Quaternary Res. 49:183-196.

Pliss L, Tambets K, Loogväli EL, Pronina N, Lazdins M, Krumina A, Baumanis V, Villems R. 2006. Mitochondrial DNA portrait of Latvians: towards the understanding of the genetic structure of Baltic-speaking populations. Ann. Hum. Genet. 70: 439-458.

Quintana-Murci L, Chaix R, Wells RS, et al. (17 co-authors). 2004. Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. Am. J. Hum. Genet. 74: 827-845.

Richard C, Pennarun E, Kivisild T, et al. (19 co-authors). 2007. An mtDNA perspective of French genetic variation. Ann. Hum. Biol. 34: 68-79.

Richards M, Macaulay V, Hickey E, et al. (26 co-authors). 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am. J. Hum. Genet. 67: 1251-1276.

Richards MB, Macaulay VA, Bandelt H-J, Sykes BC. 1998. Phylogeography of mitochondrial DNA in western Europe. Ann. Hum. Genet. 62: 241-260.

Roostalu U, Kutuev I, Loogväli EL, Metspalu E, Tambets K, Reidla M, Khusnutdinova EK, Usanga E, Kivisild T, Villems R. 2007. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Mol. Biol. Evol. 24: 436-448.

Rootsi S, Magri C, Kivisild T, et al. (45 co-authors). 2004. Phylogeography of Ychromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. Am. J. Hum. Genet. 75: 128-137.

Rowold DJ, Luis JR, Terreros MC, Herrera RJ. 2007. Mitochondrial DNA gene flow indicates preferred usage of the Levant Corridor over the Horn of Africa passageway. J. Hum. Genet. 52: 436-447.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19: 2496-2497.

Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi P, Wallace DC. 2007. An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. 35 (Database issue): D823-828.

Saillard J, Forster P, Lynnerup N, Bandelt H-J, Norby S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am. J. Hum. Genet. 67: 718-726.

Sedov VV. 1979. The origin and early history of the Slavs. Moscow: Nauka (in Russian).

Soffer O. 1987. Upper Paleolithic connubia, refugia, and the archaeological record from eastern Europe. In: Soffer O, editor. The Pleistocene Old World. Plenum Press, New York, p. 333-348.

Tambets K, Kivisild T, Metspalu E, et al. (13 co-authors). 2000. The topology of the maternal lineages of the Anatolian and Trans-Caucasus populations and the peopling of the Europe: some preliminary considerations. In: Renfrew C, Boyle K, editors. Archaeogenetics: DNA and the population prehistory of Europe: McDonald Institute for Archaeological Research Monograph Series. Cambridge University Press, Cambridge, p. 219-235.

Tambets K, Tolk HV, Kivisild T, Metspalu E, Parik J, Reidla M, Voevoda M, Damba L, Bermisheva M, Khusnutdinova E. 2003. Complex signals for population expansions in Europe and beyond. In: Bellwwood P, Renfrew C, editors. Examining the farming/language dispersal hypothesis. Cambridge: McDonald Institute for Archaeological Research Monograph Series. Cambridge University Press, Cambridge, p. 449-458.

Tarasov PE, Volkova VS, Webb III T, et al. (13 co-authors). 2000. Last glacial maximum biomes reconstructed from pollen and plant macrofossil data from northern Eurasia. J. Biogeogr. 27:609-620.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J. 2006. Harvesting the fruit of the human mtDNA tree. Trends Genet. 22: 339-345.

Torroni A, Bandelt H-J, D'Urbano L, et al. (11 co-authors). 1998. mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am. J. Hum. Genet. 62: 1137-1152.

Torroni A, Rengo C, Guida V, et al. (11 co-authors). 2001. Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am. J. Hum. Genet. 69: 1348-1356.

Zaykin DV, Pudovkin AJ. 1993. Two programs to estimate significance of Chi-square values using pseudo-probability test. J. Heredity. 84: 152.

Table 1.

Haplogroups U4a1 and U4a2 distribution (percentage with number of individuals in parentheses) in different populations of Eastern, Central and Western Europe.

Populations	**U4a1**	**U4a2**	**U4a2a**	**U4a2b**	**U4a2c**
Eastern Europe (n=4008)[a b c]	1.8 (74)	1.4 (57)	0.8 (31)	0.3 (9)	0.3 (13)
Central and Eastern Europe (n=2016)[a]	1.0 (21)	1.9 (38)	1.2 (24)	0.4 (7)	0.1 (3)
Northeastern Europe (n=1013)[b]	0.9 (9)	0.8 (8)	0.6 (6)	0.2 (2)	0
Volga-Ural region (n=979)[c]	4.5 (44)	1.1 (11)	0.1 (1)	0	1.0 (10)
Central and Western Europe (n=3704)[d]	0.5 (19)	0.3 (12)	0.2 (9)	0.1 (2)	0

[a] Russians, Belorussians, Ukrainians, Poles, Slovaks

[b] Finns, Karelians, Estonians, Latvians, Lithuanians

[c] Maris, Komi-Zyrians, Komi-Permians, Mordwin, Udmurts, Chuwashes, Tatars, Bashkirs

[d] British, Germans, Czechs, Austrians, Swiss, Bosnians, Slovenians, Italians, French, Spaniards, Portuguese

The data were taken from the literature cited in Table S2.

Table 2.

Diversity and the rate of non-synonymous/synonymous changes in haplogroups U4, HV3, HV4 and R1

HG	N	Diversity (*rho* ± s.e.)[a]	M_N	M_S	M_N/M_S	Diversity (*rho* ± s.e.)[b]
U4	49	3.98 ± 0.26	26	52	0.5	
U4a	42	2.36 ± 0.24	17	25	0.68	0.78 ± 0.14
U4a1	13	2.85 ± 0.47	5	10	0.5	
U4a2	26	1.42 ± 0.23	12	11	1.09	0.46 ± 0.13
U4a2a	12	1.25 ± 0.32	4	5	0.8	0.42 ± 0.19
U4a2*	8	1.38 ± 0.4	2	4	0.5	
U4b	4	4.0 ± 1.0	3	10	0.3	
U4c	4	5.75 ± 1.2	4	8	0.5	
U4d	6	5.0 ± 0.91	2	9	0.22	
HV3	17	2.47 ± 0.38	7	17	0.41	
HV3a	5	1.6 ± 0.57	1	6	0.17	
HV3b	4	3.0 ± 0.87	3	3	1.0	1.75 ± 0.66
HV4	4	3.75 ± 0.97	5	6	0.83	1.75 ± 0.66
R1	6	5.5 ± 0.96	5	8	0.63	3.83 ± 0.8
R1a	5	3.2 ± 0.8	3	5	0.6	1.6 ± 0.57
R1a1	4	1.25 ± 0.6	1	2	0.5	

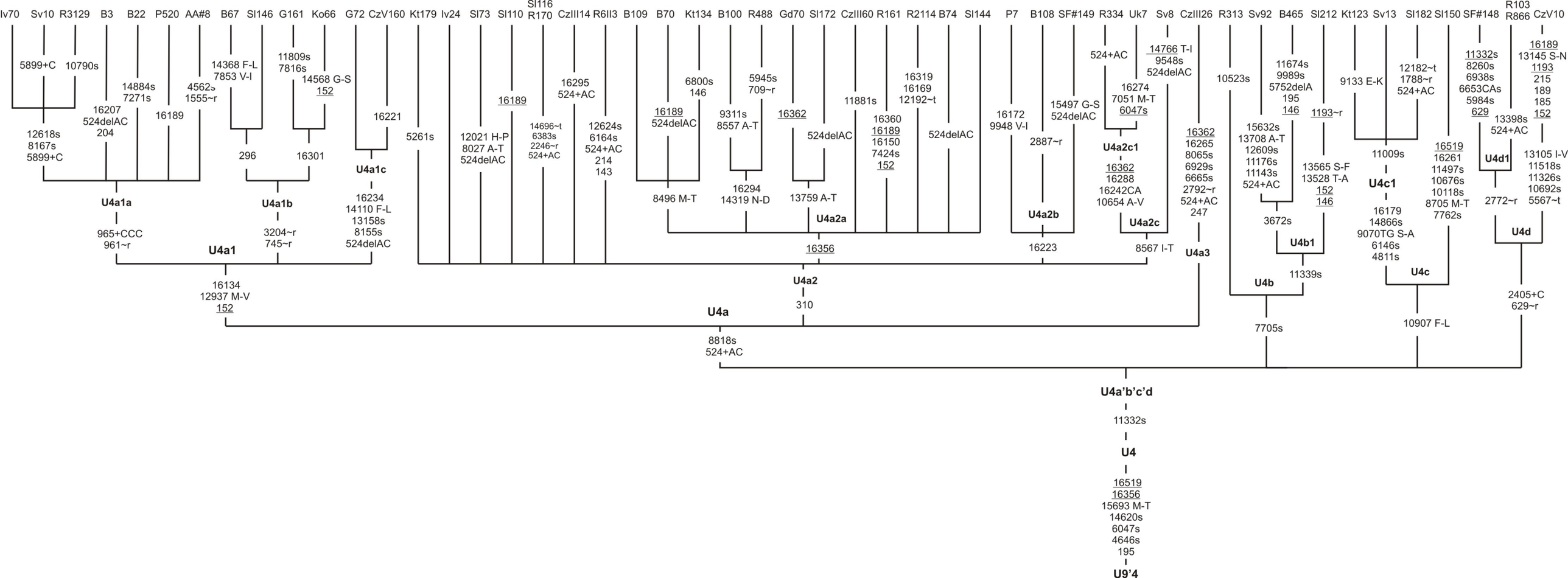
[a] Diversity has been calculated taking into account all coding-region substitutions.

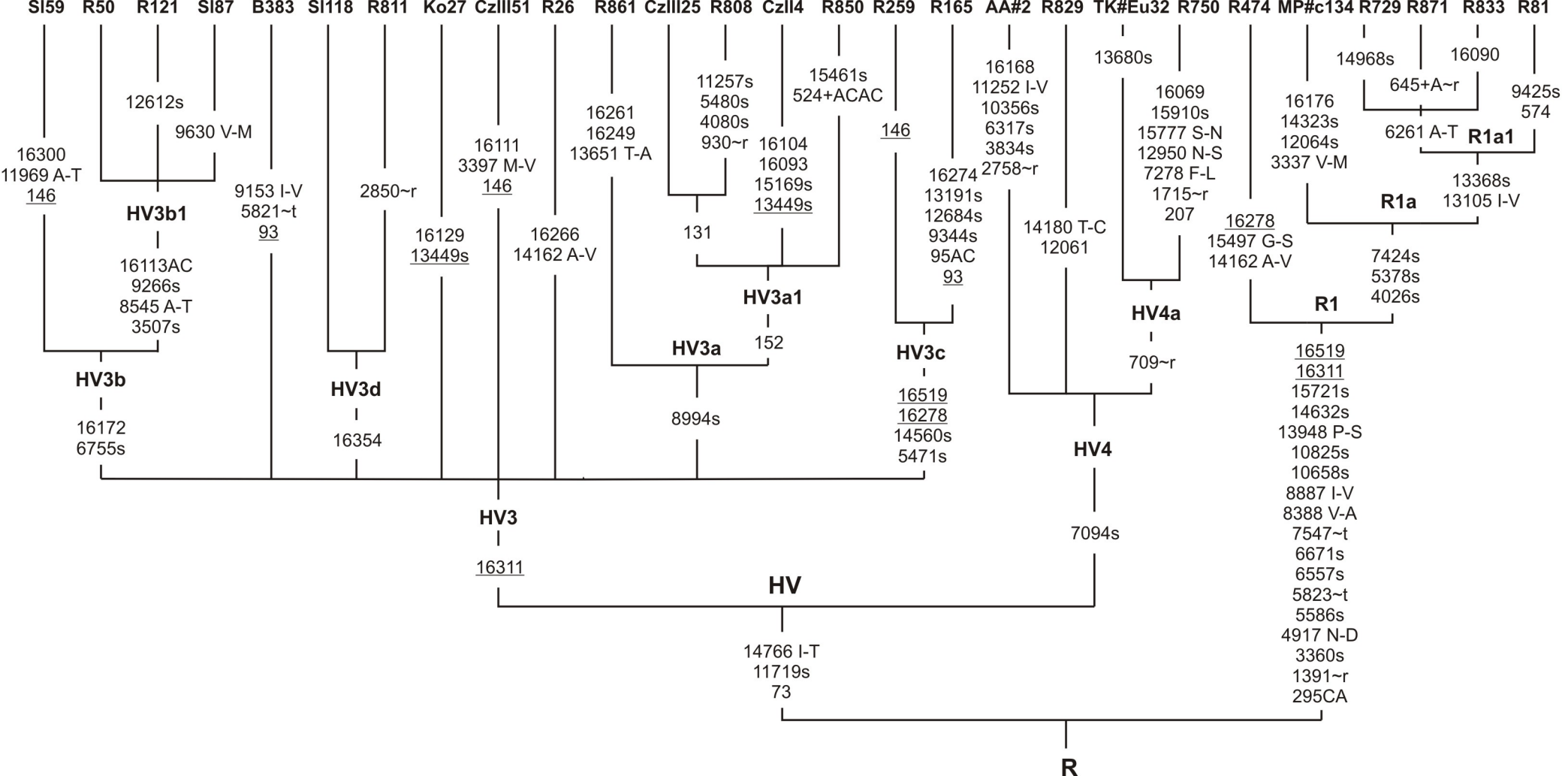
[b] Diversity has been calculated taking into account only synonymous substitutions due to the high (> 0.6) values of the M_N/M_S proportion.

Figure legends

Figure 1. Most parsimonious tree of the complete haplogroup U4 sequences, rooted in haplogroup U9'4. Mutations are shown on the branches and are transitions; transversions are further specified. The presence of deletions and insertions is referred to by "del" and "+", respectively, followed by deleted or inserted bases. Underlined nucleotide positions indicate recurrent mutations. Amino acid replacements are specified by single-letter code; s, synonymous replacements; ~t, change in tRNA; ~r, change in rRNA gene. The tree includes mtDNAs designated as follows: R – Russians from different regions of European Russia, Kt – Poles from the Upper Silesia region (southern Poland); B – Poles from the Pomerania-Kujawy region (northern Poland); Gd – Poles from the Pomerania-Gdansk region (northern Poland); P – Poles from the Malopolska region (southern Poland); Ko – Poles from Kaszuby population (northern Poland); Uk – Ukrainian; Iv and Sv – Belorussians, Sl – Slovaks and Cz – Czechs. More detailed information can be found in Table S2. Additional complete sequences were taken from the literature and their sample codes are as follows: AA (Achilli et al. 2005); SF (Finnilä et al. 2001).

Figure 2. Most parsimonious tree of the complete haplogroups HV3, HV4 and R1 sequences, rooted in haplogroup R. The tree includes mtDNAs designated in accordance with population codes explained in the legend of Figure 1. Sequences taken from the literature are designated as: TK (only coding region information is available) (Kivisild et al. 2006); MP (Palanichamy et al. 2004).


Iv70
Sv10
R3129
B3
B22
P520
AA#8
B67
Sl146
G161
Ko66
G72
CzV160
Kt179
Iv24
Sl73
Sl110
Sl116
R170
CzIII14
R6II3
B109
B70
Kt134
B100
R488
Gd70
Sl172
CzIII60
R161
R2114
B74
Sl144
P7
B108
SF#149
R334
Uk7
Sv8
CzIII26
R313
Sv92
B465
Sl212
Kt123
Sv13
Sl182
Sl150
SF#148
R103
R866
CzV10
5899+C
10790s
16207
524delAC
204
14884s
7271s
16189
4562s
1555~r
14368 F-L
7853 V-I
11809s
7816s
14568 G-S
152
16221
16189
14696~t
6383s
2246~r
524+AC
16295
524+AC
12624s
6164s
524+AC
214
143
6800s
146
5945s
709~r
16362
524delAC
11881s
16360
16189
16150
7424s
152
16319
16169
12192~t
524delAC
16172
9948 V-I
2887~r
15497 G-S
524delAC
524+AC
14766 T-I
9548s
524delAC
16274
7051 M-T
6047s
16362
16265
8065s
6929s
6665s
2792~r
524+AC
247
10523s
11674s
9989s
5752delA
195
146
15632s
13708 A-T
12609s
11176s
11143s
524+AC
3672s
1193~r
13565 S-F
13528 T-A
152
146
9133 E-K
11009s
16179
14866s
9070TG S-A
6146s
4811s
12182~t
1788~r
524+AC
16519
16261
11497s
10676s
10118s
8705 M-T
7762s
11332s
8260s
6938s
6653CAs
5984s
629
13398s
524+AC
16189
13145 S-N
1193
215
189
185
152
13105 I-V
11518s
11326s
10692s
5567~t
12618s
8167s
5899+C
5261s
12021 H-P
8027 A-T
524delAC
16189
8496 M-T
9311s
8557 A-T
16294
14319 N-D
16362
13759 A-T
524delAC
16234
14110 F-L
13158s
8155s
524delAC
296
16301
965+CCC
961~r
3204~r
745~r
16134
12937 M-V
152
16356
310
16223
8567 I-T
16274
U4a1a
U4a1b
U4a1c
U4a1
U4a2
U4a2a
U4a2b
U4a2c
U4a2c1
U4a3
U4b
U4b1
U4c
U4c1
U4d
U4d1
U4a
8818s
524+AC
7705s
10907 F-L
2405+C
629~r
2772~r
U4a’b’c’d
11332s
U4
16519
16356
15693 M-T
14620s
6047s
4646s
195
U9’4


SI59
R50
R121
SI87
B383
SI118
R811
Ko27
CzIII51
R26
R861
CzIII25
R808
CzII4
R850
R259
R165
AA#2
R829
TK#Eu32
R750
R474
MP#c134
R729
R871
R833
R81
16300
11969 A-T
146
12612s
9630 V-M
HV3b1
16113AC
9266s
8545 A-T
3507s
HV3b
16172
6755s
9153 I-V
5821~t
93
2850~r
HV3d
16354
16129
13449s
16111
3397 M-V
146
16266
14162 A-V
16261
16249
13651 T-A
HV3a
8994s
11257s
5480s
4080s
930~r
131
16104
16093
15169s
13449s
HV3a1
152
15461s
524+ACAC
146
HV3c
16519
16278
14560s
5471s
16274
13191s
12684s
9344s
95AC
93
16168
11252 I-V
10356s
6317s
3834s
2758~r
14180 T-C
12061
HV4
7094s
13680s
16069
15910s
15777 S-N
12950 N-S
7278 F-L
1715~r
207
HV4a
709~r
16278
15497 G-S
14162 A-V
R1
16519
16311
15721s
14632s
13948 P-S
10825s
10658s
8887 I-V
8388 V-A
7547~t
6671s
6557s
5823~t
5586s
4917 N-D
3360s
1391~r
295CA
16176
14323s
12064s
3337 V-M
R1a
7424s
5378s
4026s
14968s
645+A~r
6261 A-T
16090
R1a1
13368s
13105 I-V
9425s
574
HV3
16311
HV
14766 I-T
11719s
73
R