FENNO-UGRISTICA 23 / HISTORICA FENNO-UGRICA

The Roots of Peoples and Languages of Northern Eurasia II and III

Szombathely 30.9.-2.10.1998 and Loona 29.6.–1.7.1999

Edited by Ago Kiinnap

Editorial Assistant Piret Klesment

University of Tartu. Division of Uralic Languages /
Societas Historiae Fenno-Ugricae

Tartu 2000

Siiri R o o t s i, Toomas K i v i s i 1 d, Kristiina T a m b e t s, Maarja A d o j a a n, Jüri P a r i k, Maere R e i d 1 a, Ene M e t s p a I u, Sirle L a o s, Helle-Viivi T o I k, Richard V i l l e m s (Department of Evolutionary Biology, Tartu University and Estonian Biocentre, Tartu)

ON THE PHYLOGEOGRAPHIC CONTEXT OF SEX-SPECIFIC GENETIC MARKERS OF FINNO-UGRIC POPULATIONS

Summary

Here we extend our earlier analysis of the sex-determined, uni-parentally inherited genetic systems of Finno-Ugric and other populations. In particular, we specify phylogeography of a unique "Nordic" variant of Y chromosomes (the *Tat C* allele; haplogroup 16 in some of the nomenclatures) by showing that it is virtually absent in all Slavic populations studied by us (Poles, Slovaks, Czechs and Croats), except in Russians. Furthermore, we show that *Tat C* is absent in Hungarians, but well present among Latvians and Lithuanians. We discuss these findings in terms of the demographic history of the Nordic people.

We also present some new data about phylogeography of maternally inherited mitochondrial DNA (mtDNA) variation in order to show that despite the basic uniformity of the Western Eurasian mtDNA gene pool, a more detailed analysis starts to reveal patterns of variations which allow to distinguish between the common founders of the maternal lineages of Caucasoids and between the subsequent radiations of these lineages. Many of these phylogeographically clustered variations are of considerable interest in the reconstruction of ancient demographic movements in Europe and in Western Eurasia in general.

Introduction

Classical human population genetics had but a limited number of tools to study separately maternal and paternal inheritance of humans and none of them were, strictly speaking, satisfactory. Modern genetics made it straightforward by looking directly either to the paternally inherited Y-chromosomal DNA variability or to the maternally inherited mitochondrial DNA. In parallel, modern genetics has opened and is extending at a high speed practically endless resource of *possibly* informative autosomal markers to be used in physical anthropology. Nevertheless, a possibility to follow, reliably and separately, paternal and maternal components in genetic heritage of populations, is perhaps one of the most remarkable achievements of the DNA era. These results are informative (and intriguing) not only for geneticists, but perhaps even more for those who consider themselves primarily ethnologists, demographers, historians etc. A large number of possible questions, starting from investigations of the comparative spread of maternally and paternally inherited genes, selective bottlenecks for women and men, sexual practices and behaviours incl. - all these can be discussed on a more firm basis if we have empirical data on both genetic systems at hand. Meanwhile, one should not superficially

overestimate information gathered using sex-linked genetic markers. Coalescence theory tells us that all our genetic loci may be traced back to the most recent common ancestor (MRCA), except that genetic recombination which shuffles different fragments of gene sequences, make this nice easy concept much more complicated to apply in reality. Nevertheless, it is needed and also sufficient to say that as far as the "African Eve" or the "Y-chromosomal Adam" are concerned, it is quite unlikely that she and he, although both real persons, are direct ancestors (MRCAs) of any other genes in our gene pool, consisting of about 100,000 DNA sequences that are considered as genes and much larger in size and number non-coding areas of our genome that is often called "genomic junk". It is even a bit surprising that the coalescence ages of these two uni-parentally inherited parts of our genome are at present traced back to quite comparable time depths around 150,000 years ago - to approximately 7,000 generations. It is also encouraging that the best available present evidence suggests that MRCAs for both of them ("Eve" and "Adam") very likely lived in sub-Saharan Africa (e.g. R.L. Cann et al. 1987; Hammer et al. 1997).

In our previous paper in this series (Villems et al. 1998) we presented our results about the phylogenetics of mtDNA of Estonians and discussed them in a general context of maternal lineages of European populations. Firstly, we have shown that Estonian maternal lineages are, in a way, "uninteresting" - they are a representative sub-set of the corresponding European pool of mtDNA. We also argued in this paper that maternal lineages of Saami differ from those common in Europeans not because they contain any substantial quantity of "unusual" mtDNA variants (except for some Mongoloid-specific haplogroup M variants at frequencies below 10 per cent), but first of all because their mtDNA pool is severely restricted, possibly due to a very strong random genetic drift, including bottleneck phenomena and/or founder effects. We also compared the topology of mtDNA phylogenetic trees of Estonians and several other Finno-Ugric speaking populations and showed the lack of substantial differences between these populations. Even maternal lineages of Volga-Finnic populations like Maris and Mokshas. were largely overlapping with those found in Estonians and in Europeans in general.

Furthermore, in this paper we discussed certain exclusively paternally inherited (Y-chromosomal) genetic markers of Estonians. Using a set of short tandem repeats (STRs) at the background of *Tat C* haplotype of the Y chromosome (haplogroup 16 in one of the contemporary nomenclatures), we questioned conclusions in a recent paper by Zerjal et al. (1997) and considered them premature. According to Zerjal et al. Y-chromosomal heritage of Finno-Ugric speaking populations strongly suggest a Siberian ancestry of a very considerable fraction of their paternal lineages. In particular, we showed that the divergence of all the fast-evolving short tandem repeats of the Estonian *Tat* C chromosomes we have investigated is much higher than that observed in Siberian populations such as Yakuts and Buryats, suggesting that the probable paternal gene flow was not from east to west, but from west to east.

Since then (our paper was written in mid-1997), general understanding of the phylogenetics of mtDNA (see e.g. Richards et al. 1998: Macaulay et al. 1999; Kivisild et al. 1999a; Metspalu et al. 1999) has been improved and there is also a considerable progress in the understanding of the world-wide diversity of the Y-chromosomal lineages. We have now extended our investigations to several other populations, geographically close to Finno-Ugric populations as well as to those, living far away like the populations of India (Kivisild et al. 1999b). paying particular attention to the Anatolian and Trans-

Caucasus area populations since it is commonly believed that this area had an important role to play in (he early spread of anatomically modern humans (AMH) to Eurasia.

Of course, we investigate genes of the extant populations. Why is ancient DNA used so seldom? There are many answers to this question. First of all - technical complications. Even well-preserved ancient remains often give poor results (do not yield DNA in detectable amounts). Secondly, there is a problem of contamination. But the third problem is a major one. Namely, population genetics is genetics of populations: one needs to analyse a large quantity of samples to obtain meaningful and reliable results. A few or even a dozen of samples only seldom allow to get answers to the problems which are important in the reconstruction of the demographic history of human populations. There are of course exceptions, like Neanderthal man (Krings et al. 1998), where a single sequence turned out to be most useful to verify the predicted earlier extent of differences between Neanderthals and AMHs. Many other such general and specific questions exist where the results of the analysis of an ancient DNA might be of a great importance but it is evident that an absolute majority of the problems must be solved using extant populations. And this is not a shortcoming, but to a large extent exactly what is interesting: to understand the present existing diversity of humans. Still, one may hope that in a long run the results obtained in analysing fossil (or simply ancient) DNA start to fill essential gaps in a general picture of the genetic history of humans.

The last point we want to make in this general introduction is the human genome programme. It is by now evident that the prototype genome - all three billion base pairs - will be completely sequenced within a few coming years. Taking the extent of normal polymorphism among humans equal to one out of three hundred (a conservative estimate), it would result in an enormous number of polymorphic genetic markers. Not all of them would be useful for population genetic analysis, but the number of informative sites can easily be in hundreds of thousands. The problem is how to identify them and how to achieve that at least a fraction of them would be used by a sufficient number of authors, so that the results would be directly (phylogenetically) compatible - not only as dines, trends and aggregate statistics. Looking at much less sophisticated targets: at the progress in the understanding of the phylogenetics of mtDNA and of the Y chromosome, some obvious worries might well arise. However, the message here is not for casting doubt in an enormous potential of autosomal studies; it would build up gradually to huge databases. Gradually because the identification of useful polymorphic markers needs effort: the prototype human genomic sequence itself is silent about them.

Our pan-Arctic fathers

There is a steady progress in the understanding of the worldwide distribution of the human paternal lineages - a story told in the variability of the Y chromosome. More than a hundred of single nucleotide polymorphic sites in the Y chromosome are known by now and perhaps ten per cent of them seem to be informative in a global context. These numbers make up approximately a tenth of potentially available SNPs in this chromosome. There seems to be now a clear understanding in how such global phylogenetic trees should be constructed. Specifically, the main effort is now in the reconstruction of a truly cladistic picture, based on historically unique species single point mutations - in contrast to an earlier belief that short tandem repeats (STRs) alone are informative enough for deep phylogenetic reconstructions. A unique event might as well be an insertion (like *Alu)* or

deletion, but the message is clear: first we need a reliable skeleton and only then flesh can be added by using information obtained from the variability (length polymorphisms) of the fast-evolving highly variable STRs. This skeleton is now emerging (Fig. 1) but a considerable further effort is needed, because many Y-chromosomal haplogroups are still not really haplogroups but internal nodes of too complex branching patterns for a detailed analysis. Therefore, combined in-depth analysis of the topology of the Y-chromosomal tree including inferring coalescence times from the variability of STRs is so far limited to nodes that can be considered external.

We admit, however, that we feel somewhat insecure in suggesting absolute lime estimates for Y-chromosomal phylogenetics by taking the evolutionary speed of STR's around $2x10^{-4}$. SNPs or any other unique events like Alu insertions allow to reconstruct trees and sometimes to see what happened earlier, what later, but as such they themselves lack time dimension, usable for detailed calculations. There is an easy possibility of including a relative time scale, using much faster evolving STRs or any other markers, evolving faster than SNPs, Earlier enthusiasm in converting these relative estimates directly into absolute time scale has cooled down to some extent because it became evident that not only the evolutionary speeds of different micro-satellite repeats (even of the same size class) differ considerably but because their modes of molecular evolution are not necessarily simple: one-step increase or decrease. Furthermore, there seems to be a conflict between evolutionary rate calculations based on pedigree analysis and using direct experimental methods. As for mtDNA analysis, the former seems to suggest higher speeds and. consequently, shorter time estimates. Sorting out truth might not be easy but these uncertain aspects do not necessarily interfere with gathering experimental data, constructing phylogenetic trees and inferring relative time depths of various branches of the human Y-chromosomal DNA lineages.

As already mentioned in the Introduction, important in its empirical content was a paper by Zerjal et al. (1997), where it was first shown that the spread of *Tat* C allele is restricted to Buryats, Yakuts and Saamis and Finns. They showed that this Y-chromosomal lineage is very frequent in these populations, covering about a half and more of their paternal inheritance - of their Y chromosomes. Why did Zerjal et al. conclude that this pattern of distribution supports Siberian origin of a substantial fraction of Finno-Ugric paternal lineages? There were two reasons. First, they found that the frequency of *Tat* C among Yakuts and Buryats is somewhat higher than among Finns. Second, according to their data, the diversity of STRs in the background of *Tat C* among the Baltic Finns and these Siberian populations was comparable. As it turned out, this was not really the case: the particular allele is much more divergent in Estonians than in Siberian non-Finno-Ugric populations(Villemset al. 1998). Here we reproduce a network scheme of the diversity of *Tat* C in selected populations, illustrating this point (Fig. 2). Why, then, the diversity of *Tat C* allele is so low in the Finnish population? The answer is probably at hand: already earlier studies revealed a bottleneck in the paternal and maternal DNA lineages in the founding of the Finnish population (Sajantila et al. 1996) and subsequent investigations confirmed a male-specific bottleneck in Finns (Kittles et al. 1999).

The other aspect that was important to understand was a comprehensive phylogeography of the spread of *Tat C*. Zerjal et al. (1997) showed that *Tat* C, although frequent in Buryats, is present only in a small number of sub-populations of Mongols and is not found further south - in China, Korea etc. They also demonstrated that despite its

high frequency in Finns and Saamis. *Tat C* drops abruptly among Indo-European speaking Scandinavians - down to less than 5 per cent in Norwegians and to nil in Western Europe. It was shown earlier (see also Villems et al. 1998) that its frequency in Russians is around 15 per cent. We have now investigated several different Slavic and other populations: Slovaks, Czechs, Poles Croats, Lithuanians, Georgians, Armenians, Ossetes, Turks. The results are informative and allow to draw several firm conclusions. First, none of the western Slavic populations studied possess *Tat C* at frequencies above a few per cent. This result confirms our earlier suggestion that a relatively high incidence of *Tat C* among Russians probably reflects a Finno-Ugric "substratum" in eastern Slavs. Absence or very low incidence of *Tat C* allele in the Caucasus area populations shows that paternal lineages of the populations of the Eastern European Plain had hardly contributed to the Y-chromosomal pool of the extant populations further south. Of course, this statement is true for as long as *Tat C* exists among the populations inhabiting the former area: deeper in time depth connections at the founder node can be seen (discussed below),

More interesting is the situation with Latvians and Lithuanians who linguistically belong to Indo-Europeans and the Baltic branch shared between them is close to the Slavic branch of Indo-European languages. Surprisingly, it turned out that frequencies of the *Tat* C allele in both Latvian (Lahermo et al. 1999) and Lithuanian (our results) Y chromosomes are close to those among Estonians, Karelians and Finns: i.e. significantly higher than among Russians and much higher than among western Slavs: around 29% for Latvians and 33% for Lithuanians. We consider this finding very interesting from the point of view of the ethnogenesis of the extant Baltic and Finno-Ugric; populations. There is no apparent north-south frequency gradient of *Tat C* allele from the Arctic Sea (Saamis) to Lithuanians but a sharp east-west cline both in Scandinavia and on the Baltic area. Finally, *Tat C* Y chromosomes are also very rare among Hungarians (Lahermo et al. 1999 and our data).

Meanwhile, several labs (Lahermo et al. 1999 and unpublished so far data) have carried out more dense mapping of *Tat C* allele within Russia including Siberia, as well as in Inuits. These results fully confirmed earlier data by Zerjal et al. (1997) and yielded many additional interesting details. It is now clear that *Tat C* is frequent not only among Baltic Finno-Ugrians but also among Komis and Finno-Ugrians of the Volga basin and in Western Siberia. Furthermore, it is frequent not only in Yakuts but also among populations like Koryaks, Chukchi, Evenks, Evens, Nenetses. Yukaghirs. It has also spread among GreenlandInuits.

Much has been clarified and although several important questions remain unsolved (see below), we can with confidence superimpose the spread of the *Tat C* allele and the map of populations and languages. This variant of the Y chromosome is truly circum-Arctic. Being also by far the dominant variety of Y chromosome on this area, its spread is not restricted to any linguistically defined population: it can be found among Uralic-, Indo-European- and Altaic-speaking populations of the area. The finding is certainly in contrast to the spread of the human maternal lineages where the Siberian Altaic speaking populations share only a small fraction of maternal lineages with European Finno-Ugric populations like Karelians, Estonians etc. and where their mtDNA pool overlaps with those found among Mongoloid populations in general, including Han, Japanese, Mongols etc., among whom the *Tat C* allele is virtually unknown.

Although a more precise comparison of the differential spread of the maternal and paternal lineages in Northern Eurasia is certainly possible and desired, one general lesson

seems to be apparent already. Namely, the spread (and flow) of maternally and paternally inherited genes in humans may differ significantly: a conclusion which cannot be drawn studying "usual" (autosomal) genes alone. During the last five years many authors, specifically L. L. Cavalli-Sforza, have stressed (and presented experimental evidence in favour of) that maternal gene flow seems to cover wider areas than that for paternal. While it might be indeed so in some cases, one should not accept it as a general rule. The present results demonstrate the opposite. And there are other such examples speaking for a wider (more intensive, longer in distance) spread of paternal lineages that we are not going to discuss here.

How can such a cross-linguistic spread of a certain variety of Y chromosomes occur? In case of Russians the possible answer is not complicated, if one accepts the Finno-Ugric "substratum" concept, supported by many independent lines of research and the wealth of historical evidence. However, to explain its circum-Arctic spread is more puzzling. Could it be a general "substratum"? And if yes, what is its time depth - before Last Glacial maximum (LGM) or after? One may keep an eye also on a possibility of a straightforward Darwinian selection as a vehicle: enhanced reproductive success (higher fitness) of the carriers of this variety of Y chromosome in a cold climate compared to the pool of Y chromosomes among that-time Siberian populations. It is known that under selective pressure, advantageous genes (in case of the Y chromosome a single locus) can spread in fact very fast, without carrying much additional (autosomal) genetic information with it, since the latter component can quickly be "diluted out" in a mendelian process. Without any direct evidence available, one cannot go further with these or other speculations. One aspect, at least, is clear - phylogeography of the diversity (i.e. not its mere frequency) of the *Tat C* allele deserves a great deal of attention by anybody who wants to understand genetic history and, applying a more complex and problematic term - ethnogenesis - of the Nordic people. Take, for example, the fact that *Tat C* allele is well presented among Khanties but appears to be very rare in Hungarians. It immediately raises several questions and a need for explanations. One of the suggestions may be that Khanties and Mansis obtained this Y-chromosomal lineage only relatively recently, together with a long list of other Siberian populations. The second possibility is that Hungarians and Siberian Ugric-speaking populations have genetically, in fact, very little in common and that they never had any. The third possibility is that Hungarians have lost the lineage by drift during their migration from Eastern Europe - South-Western Siberia to Pannonia. For that, however, the size of their male population should have dropped really drastically in order to eliminate one major variety of Y chromosomes from their gene pool. Many details can be added to these speculations but since direct comparative genetic investigations are in principle possible, it seems wiser to postpone genetics-based discussions till such investigations are carried out.

Above we concentrated solely on the *Tat C* allele: although dominant, it makes up less than a half of the Estonian (Finno-Ugric) Y-chromosomal pool. All other paternal lineages found in Estonians are those present all over European - Western Eurasian populations. Although frequencies vary, sometimes even significantly (Table 1), they are not in a such a sharp contrast like those for haplogroup 16 - the *Tat C* allele. Nevertheless, there is another significant but "negative" contrast - virtual lack of haplogroup 9 in North-Easlern Europe. Note, however (Table 1), that here the borderline does not go in-between

eastern and western Slavs but we see well pronounced north-south gradient in the spread of this Y-chromosomal lineage. And mere are other examples like that as well. In fact, most of the Y-chromosomal haplogroups display non-random distribution of frequencies inside continental Europe already at this superficial resolution, that can be achieved using a limited set of bi-allelic markers (i.e. usually unique single nucleotide polymorphisms, SNPs). There is no simple pattern behind it but one can expect a number of possibly very informative signs from the point of view of male gene flows in the past- documents of past demographic movements as well as signs of social behaviour, wars etc. However, a detailed analysis of the frequencies of such haplogroups deserve at present to be carried out only for such a subset of them, about which we know that they are not themselves internal nodes (like e.g. I, 2. 26) of a complex network of Y-chromosomal lineages.

Maternal lineages start to reveal informative details

There is no consensus on the horizon about the peopling of Europe yet. However, there seems to be a slight tendency towards accepting that phylogenetically meaningful structuring of human maternal lineages is both possible and superior to mere frequency computations of ill-defined "characters". Yet the old approaches are perfectly alive and reach sometimes even more radical conclusions than the classical version of the Neolithic demic diffusion ever claimed.

We are not going to submerge into these generalities here and will touch more empirical aspects of the present-day maternal lineage studies in our lab and elsewhere. Empirical here means cataloguing phylogeographics of individual clusters of maternal lineages. Formally, an individual mtDNA cluster consists of a founder (can be also an "empty" node) and its descendants. Therefore, a properly defined cluster of maternal lineages (haplogroup) is a monophyletic clade, reflecting common genealogy. This type of analysis was initiated in D. Wallace's laboratory about 10 years ago by identifying a large number of polymorphic sites in slowly evolving (coding) areas of mtDNA. That, in turn, allowed to identify the key elements of the topology of the world-wide evolutionary tree of mtDNA: to classify reliably and phylogenetically the main clusters (e.g. Torroni et al. 1993; Torroni et al. 1994, Torroni et al. 1996). Joint efforts have by now allowed to fuse phylogenetically informative sites in mtDNA hypervariable and coding areas (e.g. Richards et al. 1998; Macaulay et al. 1999, Kivisild et al. 1999b).

Smoothness (lack of clines, contrasts) in the spread of mtDNA lineages all over Europe is a popular point of view at present (e.g. Simoni et al. 2000), but even a slightly closer inspection shows that this randomness is only the first approximation, reflecting the fact that Western Eurasians and Northern Africans can largely be seen as carrying a continuum of the same phylogenetically closely related mtDNA lineage clusters of the largely Paleolithic origin, when it is compared with mtDNA pools of Eastern Asians or sub-Saharan Africans.

This first approximation is not the level where one needs to or should slop. A dedicated comparative analysis of the phylogenetic trees of populations at high resolution reveals that the picture is far from smooth and allows, both in principle and in reality, to go much further and find informative (and sometimes very intriguing) connections between maternal lineages of extant populations. Even taking still a coarse next approximation - sub-clusters of the second most frequent among Caucasoid populations haplogroup U, one starts to see sharp frequency differences in some clusters, while the others are distributed

more evenly (Table 2). It is obvious that while the frequency of haplogroup U as a whole is close for all Western Eurasian phylogeographic categories, then "opening it up" according to phylogenetically defined sub-entities starts to reveal otherwise hidden differences. As far as sub-clusters of haplogroup U are concerned. Table 2 shows quite clearly that there is no single rule. The sub-clusters U4 and U5 are significantly more frequent in Northern Europeans, possibly specifically in Finno-Ugrians, whereas the reverse gradient is seen for Ul and K. Differences in the spread of U2 may not be statistically representative. U3, in turn, is much less frequent in Europe than in Anatolia (Table 2).

Coalescence theory allows to calculate the age of these individual clusters and to get an idea when they started to expand. Approximate as these calculations may be at present, this is nevertheless a very significant progress compared to mere frequency patterns. Furthermore, the topology of the individual lineage clusters, provided a representative set of populations is taken for analysis, allow to reveal monophyletic branches, specific either for defined (restricted) areas, or, on the contrary, to see those which have spread across large geographic areas. As an example we present here one sub-cluster of the most abundant in Europe haplogroup H, characterized by a motif of transitions in nucleotides 16,293 and 16,311 (Fig. 3) . Its spread in Caucasoid populations is rather specific: absent in Turks and Trans-Caucasians (Armenians, Georgians, Ossetes), it is also absent in most of the Mediterranean populations but frequent in Estonians and present in many Central European populations like Slovaks. Germans etc. Although frequent in Estonians, one cannot consider it specific for Finno-Ugrians, since it seems to be rare or absent in Finns, Karelians and Saamis. It might have been removed by random drift, but considering its time of expansion (see below), it may be that it never reached Saamis and Karelians. Its topology reveals (Fig. 3) that it started to expand from two founders and, above those, one can also see an additional, much more ancient founder, characterised by mutation at np 16,092 (Fig. 3).

We cannot say much about an absolute age of this variety of haplogroup H: being a branch of a dominant Western Eurasian haplogroup, it may well be more than 20,000 years old. However, it appears that the two major founders seen in Figure 3, started to expand simultaneously (that is not surprising) about 5200 ± 1700 years ago - at the end of the Neolithic - beginning of the Bronze Age. Low frequency of this variant of haplogroup H in large areas of Europe and lack in Anatolia-Trans-Caucasus suggests strongly that mtDNA gene flow - movement of females - from the areas where it does occur to areas where it is absent, was probably rather limited during the last 5,000 years or so. The time scale roughly coincides with an expansion of the Linear Pottery. Impressed Ware and early Eastern European pottery-bearing sites. One puzzle concerning this maternal lineage remains here without an answer: it is frequent also among Albanians.

In a case study of a different kind we wish to characterize one "unlabelled" variety of haplogroup U. Table 2 shows that almost all Caucasoid populations contain a small fraction of haplogroup U variants which do not belong to any of the formally defined sub-clusters. This group is heterogeneous and possibly contains also lineages which have, thanks to reverse mutations, lost their "diagnostic" sites. Above them it also contains a very rare, but nevertheless wide-spread variety of haplogroup U, characterized by an additional mutation at np 16,146. We found it first among Estonians, but the investigation of other populations and other researchers" published data show its presence among Karelian, Austrian, German, Swiss (German-speaking area), Slovak and Czech populations

and even in one Basque. Again not in Turks and Trans-Caucasians but as a very thin layer over much of Central Europe. The database for this sub-cluster is too small for meaningful coalescence time calculations, but its phylogeography shows sharing specific maternal lineages between linguistically different populations.

The next example we give here is U4, since its frequency seems to be highest among Finno-Ugrians. Figure 4 shows that its topology is not simple: one can see several founders, including some minor, but phylogeographically intriguing branches. For example, motif 16,356; 16,179 leads to a tiny cluster shared by Georgians, Germans, Italians and Turks but not Finno-Ugrians. On the other hand, there is a much more frequent sub-division of U4 - 16,356; 16,134 - where different Finno-Ugric, Germanic and Slavic populations are dominant (Fig. 4). This branch exhibits signs of the beginning of expansion around 19,000 ± 3,500 BP. The founder of U4 (16,356 alone) is a source of individual expansion of this cluster as well. Taking its topology at face value, an expansion of this unit has possibly started after the LGM, in late the Upper Paleolithic and covers lineages, found in extant populations of Armenia, Georgia, Anatolia, Crete, Croatia, Italy, Iberia, UK as well as in a variety of continental Germanic, Slavic and Finno-Ugric populations, as well as among Volga Tartars (Orekhov, personal communication) and even in Indians. What we see here is a representative pan-Western-Eurasian collection of populations who all share maternal lineages of this particular type. Bearing in mind that the expansion of this cluster pre-dates almost certainly the Neolithic period plus the fact that its frequency in the Middle and Near East and in Trans-Caucasus is lowest and seems to be highest in North-Eastern Europe, it seems unlikely that its presence and expansion in Europe can be linked to the Neolithic demic diffusion.

We brought these case studies just to illustrate possibilities of the approach. Much work is still ahead since many areas are so far covered superficially or, at least, unevenly. One of such areas is Sweden and the other one is Denmark - very limited data available about these populations hinder, unfortunately, comprehensive cataloguing of mtDNA variants possibly present in Northern Europe.

Further discussion

At first glance, the pattern of the distribution of paternal and maternal lineages in Europe differs in a sense that the haplogroups of paternal lineages established at present seem to be distributed less evenly. It is certainly true for the *Tat C* allele, since there is not even remotely comparable dominant mtDNA lineage cluster with such a circum-Arctic distribution. Whatever the mechanism, one conclusion seems obvious: if indeed this particular mutation arose first about 10,000-15,000 years ago in (that time!) North-Eastern Europe, then its carriers never penetrated Western and Southern Europe. Or if they did, then they either migrated back or were (males) exterminated. Instead we see a very successful spread of *Tat C* eastwards - up to Kamchatka, to Chukchi and even to Greenland Inuits. If our time estimate is sound then we are discussing events occurring in the late Upper Paleolithic and even much more recently - bearing in mind a very limited STR-linked diversity of *Tat C* in the Altaic languages speaking Siberian populations.

Migration of hunters-gatherers is dictated by two basic forces: availability of food and the space unoccupied by others. Re-colonization of Northern Europe after the Last Glacial Maximum started around 15,000-17,000 years ago and the population density in the "classical" Ice Age *refugiae* areas dropped accordingly (e.g. Dolukhanov, this volume).

As for big game like mammoths of the Eastern European Plain *refugium* area, it is likely that they moved away lo Siberia, with at least some of their specialized hunters in their wake. This is one possible speculation how *Tat C* found its way to North-Eastern Asia - provided that this variant of the Y chromosome existed already at that time. The answer to the question why *Tat C* did not spread westwards may be in a higher density of already existing population in this direction at limes when the spread occurred. As for selective elimination of this lineage in Central - Western Europe then we consider this scenario less likely. At least without "hard evidence" at hand - as. For example, finding high incidence of *Tat C* in the Upper Paleolithic - Mesolithic - Neolithic human remains in Central Europe.

The intellectual strength of the phylogenetic approach consists in a possibility of using molecular evolutionary arguments. Figure 1 demonstrates that at the level of the present resolution, haplogroup 16 (*Tat C*) derives from haplogroup 12 and the latter from haplogroup 26. The latter is an internal node, giving many branches of paternal lineages; some of them defined as individual haplogroups. Table 1 shows that haplogroup 12, although less frequent than 16, is nevertheless well visible in Estonians as well as in Russians. Meanwhile, this precursor haptogroup is even more frequent in Czechs and is also present in other western Slavic populations. Bearing in mind frequencies of haplogroup 16 in Finno-Ugrians and in Russians, one may argue that this slight "contamination" of the western Slavic pool of Y chromosomes with haplogroup 16 is caused by "border line diffusion". However, if the place of origin of this haplogroup had been located within Eastern Siberian Mongoloid populations and this variant of Y chromosome was carried to Europe by Finno-Ugric tribes coming from Siberia, then one would not expect to see its much less frequent precursor - i.e. haplogroup 12 - in western Slavs at such frequencies, perhaps even none at all. It is there, nevertheless. It suggests that haplogroup 12, the precursor variant to 16, which is *ipse facto* phylogenetically (and of course in absolute time scale) older than the latter, may well be considered as Central European and its more detailed phylogeography is worth investigating. Finally, it is intriguing to note that yet another step back in the phylogenetic history - haplogroup 26 - is already well visible not only in Turks and Trans-Caucasians but also in Indians (Fig. 1; Table 1), but being an internal node (see above), it is at present an ill-defined phylogenetic state in the molecular evolution of Y chromosome: it is, so to say, everything that is not yet classified as descendants of this node, like haplogroups 3, 22 etc.

We admit that the synthesis of the Y-chromosomal and mitochondrial DNA data is not there yet. Phylogeography of the two data sets seem to differ - at least as far as Finno-Ugric and the two Baltic populations - Latvians and Lithuanians - are concerned. However, as far as the Baltic Finno-Ugrians are concerned, they may not be so different at all, provided an equally detailed analysis of the mtDNA and Y-chromosomal phylogeny is carried out, analyzed and interpreted in a wider context: together with historical, linguistic and anthropological context in general.

References

Anderson, S., A. t. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin et al. 1981, Sequence and organization of the human mitochondrial genome. -

Nature 290, pp. 457–465; Cann, R. L., M. Stoneking, A. L. Wilson 1987, Mitochondria1 DNA and human evolution. – Nature 325, 31–36; Comas, D., F. Calafell, E. Mateu, A. Perez-Lezuan, E. Bosch, R. Martinez-Arias, J. Clerimon et al. 1998, Trading genes along the silk road: mtDNA sequences and the origin of Central Asian populations. – American Journal of Human Genetics 63, pp. 1824–1838; Hammer, M. F., A. B. Spurdle, T. Karafet, M. R. Bonner, E. T. Wood, A. Novelletto, P. Malaspina, R. J. Mitchell, S. Horai, T. Jenkins et al. 1997, The geographic distribution of human Y chromosome variation. – Genetics 140, pp. 767–782; Kittles, R. A., A. W. Bergen, M. Urbanek, M. Virkkunen, M. Linnoila, D. Goldman, J. C. Long 1999, Autosomal, mitochondrial, and Y-chromosomal DNA variation in Finland: evidence for a male-specific bottleneck. – American Journal of Phys. Anthropology 108, pp. 381–399; Kivisild, T., K. Kaldma, M. Metspalu, J. Parik, S. Papiha, R. Villems 1999a, The place of the Indian mitochondrial DNA variants in the global network of maternal lineages and the peopling of the old world. – S. S. P a p i h a, R. D e k a, R. C h a k r a b o r t h y (eds.), Genome Diversity: Applications to Human Population Genetics, New York, pp. 135–152; Kivisild, T., M. J. Bamshad, K. Kaldma, M. Metspalu, E. Metspalu, M. Reidla, S. Laos, J. Parik, W. S. Watkins, M. E. Dixon, S. S. Papiha, S. S. Mastana, M. R. Mir, V. Ferak, R. Villems 1999b, Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. – Current Biology 9, pp. 1331–1334; Krings, M., A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, S. Paabo 1997, Neanderthal DNA sequences and the origin of modern humans. – Cell 90, pp. 19–30; Lahermo, P., M.-L. Savontaus, P. Sistonen, J. Beres, P. De Knijff, P. Aula, A. Sajantila 1999, Y chromosomal polymorphisms reveal founding lineages in the Finns and the Saami. – European Journal of Human Genetics 7, pp. 447–458; Macaulay, V. A., M. B. Richards, E. Hickey, E. Vega, F. Cruciani, V. Guida, R. Scozzari, B. Bonne-Tamir, B. Sykes, A. Torroni 1999, The emerging tree of the West Eurasian mtDNAs: a synthesis of control region sequences and RFLPs. – American Journal of Human Genetics 64, pp. 232–249; Metspalu, E., T. Kivisild, K. Kaldma, J. Parik, M. Reidla, K. Tambets, R. Villems 1999, The Trans-Caucasus and the expansion of the Caucasoid-specific human mitochondrial DNA. – S. S. P a p i h a, R. D e k a, R. C h a k r a b o r t h y (eds.), Genome Diversity: Applications to Human Population Genetics, New York, pp. 121–133; Richards, M. B., V. A. Macaulay, H.-J. Bandelt, B. C. Sykes 1998, Phylogeography of mitochondrial DNA in western Europe. – Annales Human Genetics 325, pp. 241–261; Sajantila, A., A.-H. Salem, P. Savolainen, K. Bauer, C. Gierig, S. Paabo 1996, Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. – Proc. Natl. Acad. Sci. USA 93, pp. 12035–12039; Simoni, L., F. Calafell, D. Pettener, J. Bertranpetit, G. Barbujani 2000, Geographic patterns of mtDNA diversity in Europe. – American Journal of Human Genetics 66, pp. 262–278; Torroni, A., T. G. Schurr, M. F. Cabell, M. D. Brown, J. V. Neel, M. Larsen, C. M. Vullo, D. C. Wallace 1993, Asian affinities and continental radiation of the four founding Native American mtDNAs. – American Journal of Human Genetics 53, pp. 563–590; Torroni, A., M. T. Lott, M. F. Cabell, Y.-S. Chen, L. Lavergne, D. C. Wallace 1994, mtDNA and the origin of Caucasians: Identification of ancient Caucasian-specific haplogroups, one which is prone to a recurrent somatic duplication in the D-loop region. – American Journal of Human Genetics 55, pp. 760-776; Torroni, A., K. Huopanen, P. Francalacci, M. Petrozzi, L. Morelli, R. Scozzari, D. Obidu, M-L. Savontaus, D. C. Wallace 1996, Classification of

European mtDNA from an analysis of three European populations. – Genetics 144, pp. 1835–1850; **Villems, R., M. Adojaan, T. Kivisild, J. Parik, G. Pielberg, S. Rootsi, K. Tambets, H.-V. Tolk 1998,** Reconstruction of maternal lineages of Finno-Ugric speaking people and some remarks on their paternal inheritance. – K. J u l k u, K. W i i k (eds.), The Roots of Peoples and languages of Northern Eurasia, Jyvaskyla, pp. 180–200; **Zerjal, T., B. Dashnyam, A. Pandya, A. Kayser, L. Roewer, F. R. Santos et al. 1997,** Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. – American Journal of Human Genetics 60, pp. 1174–1183.

Figures

Fig. 1. Reduced phylogenetic network of the human paternal lineages. One of the existing Y-chromosomal haplogroup nomenclatures according to M. Jobling and C. Tylor-Smith, reduced here to display relationships between haplogroups in Table 1

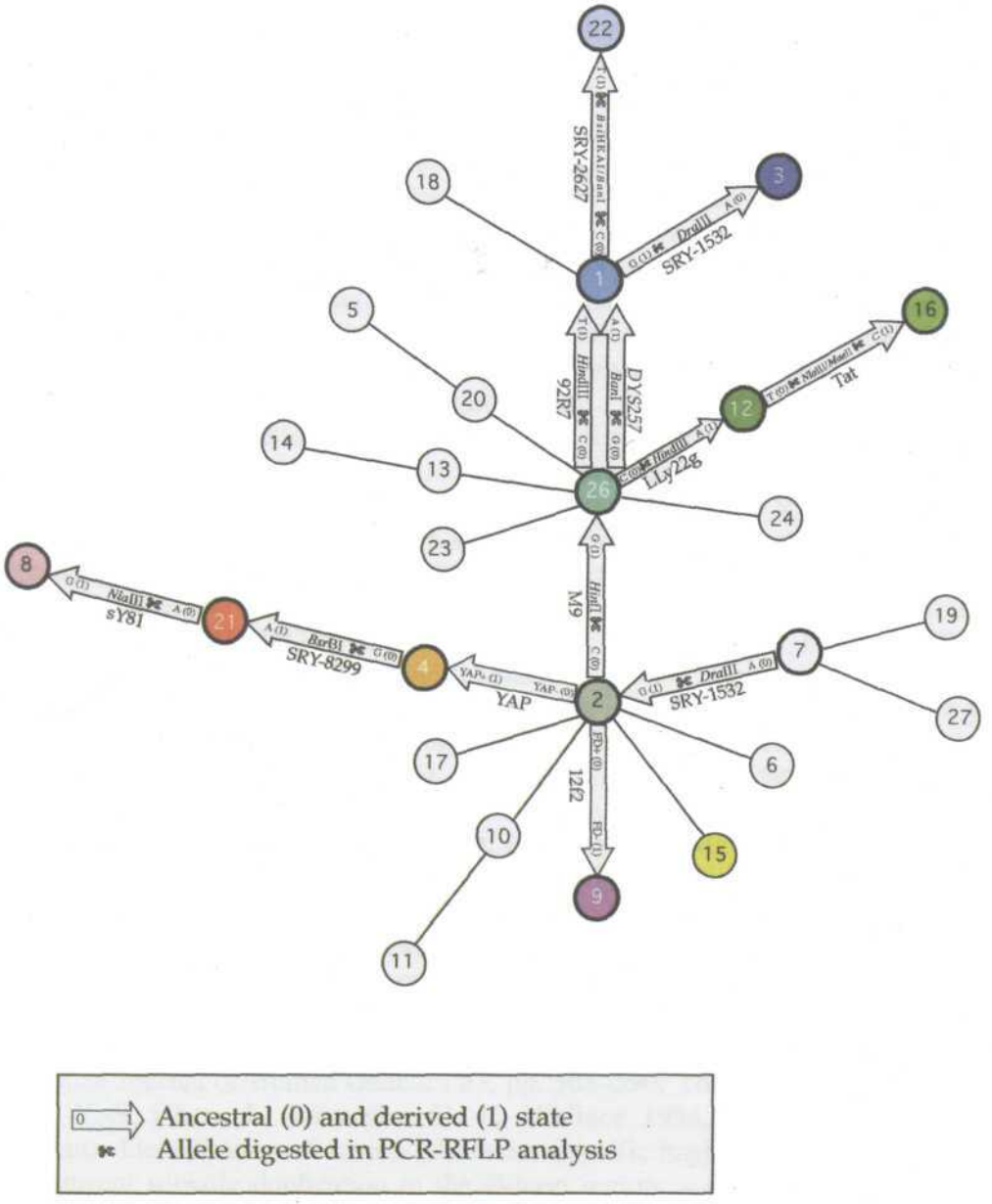


Fig. 2. Microsatellite diversity of the *Tat C* variant of Y chromosomes. Note much higher diversity of the *Tat C* allele in Estonians than in Yakuts and Buryats.

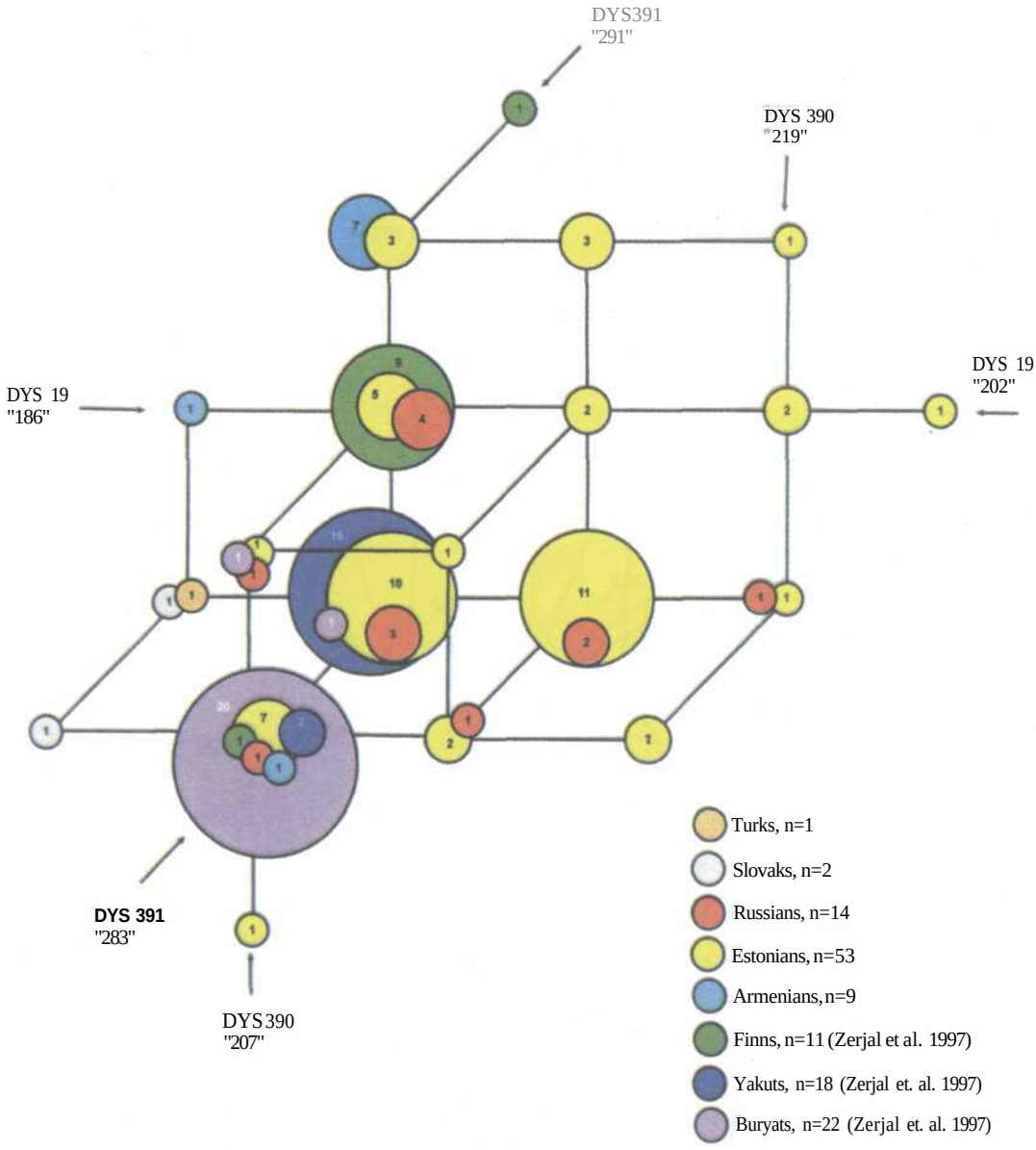


Fig. 3. Phylogeography of one of the human mitochondria1 haplogroup H variants, characterized by a motif of mutations at nucleotides 16,293 and l6,311.
Note the presence of two dominant sub-founders, exhibiting signs of relatively recent expansion events.
Abbreviations for populations as follows: AL – Albanians; CO – Croats; CZ – Czechs; ES – Estonians; FI – Finns; GE – Germans; IT – Italians; PO – Poles; RU – Russians; SA – Sardinians; SL – Slovaks; SP – Spanish; SW – Swiss; UI – Uiguris; UK – British; Data taken from various published articles, databases and our unpublished results. H* defines the nodal position of haplogroup H, identical to Cambridge reference sequence (Anderson et al. 1981). Successive mutations in mtDNA hypervariable region I are shown less 16,000. The exact base substitution is specified only for transversions.

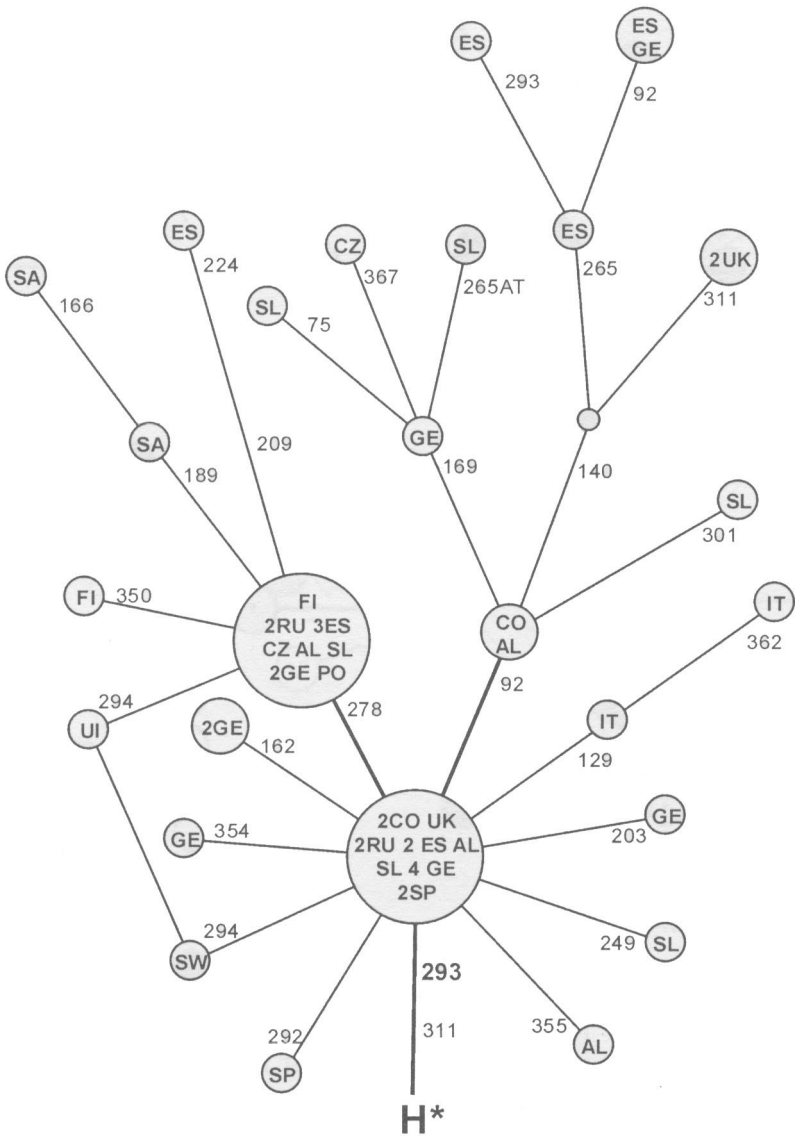
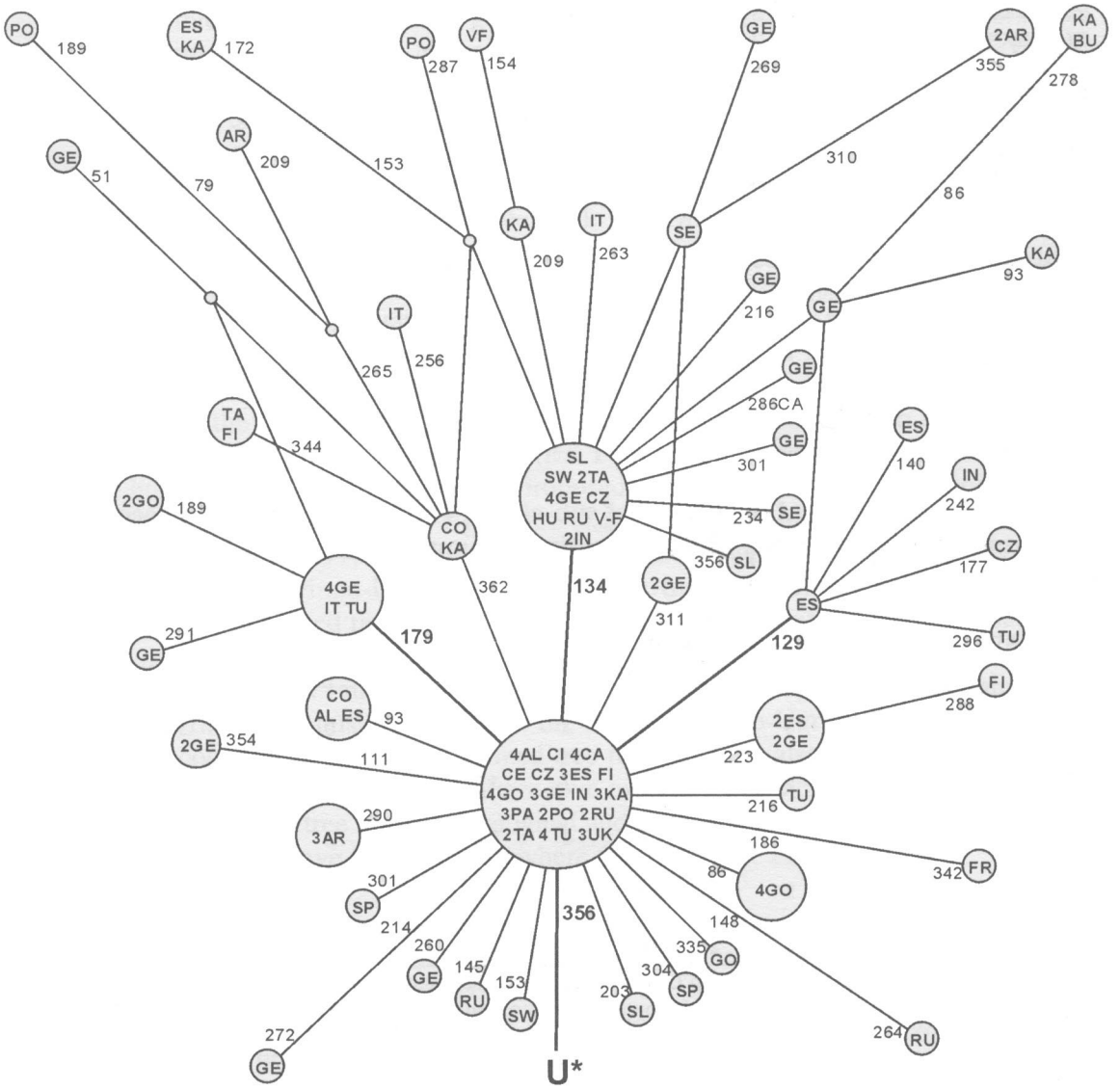


Fig. 4. Phylogeography of the human mitochondria1 haplogroup U4.
Abbreviations for populations as follows: AL – Albanians; AR – Armenians; BU – Bulgarians; CA – Central Asians; CI – Canary Islanders; CO – Croatians; CZ – Czechs; ES – Estonians; FI - Finns; FR – French; GE – Germans; GO – Georgians; HU – Hungarians; IN – Indians; IT – Italians; KA – Karelians; PA – Parsis; PO – Polish; RU – Russians; SE – Swedish; SL – Slovaks; SP – Spanish; SW – Swiss; TA – Tatars; TU – Turks; UK – British; V-F – Volga Finno-Ugrians; Data from various published articles, databases and our unpublished results. Data on Tatars – personal communication from V. Orekhov. U* defines the nodal position of haplogroup U, identical to Cambridge reference sequence for mtDNA hypervariable region I (Anderson et al. 1981). Successive mutations in mtDNA hypervariable region I are shown less than 16,000. The exact base substitution is specified only for transversions.



Tables

Table 1

Distribution of paternally inherited lineage clusters among selected human populations as % from all paternal lineages*

Population	Y-chromosomal haplogroup							
	1	2	3	21	9	12	16	26
ESTONIANS	10	10	27	4	1	5	37	6
HUNGARIANS	24	35	17	6	14	1	1	0
RUSSIANS	7	18	48	6	1	5	14	1
SLOVAKS	17	17	47	10	2	2	3	2
POLES	20	20	52	2	3	1	2	0
CZECHS	19	19	38	7	11	6	0	0
GEORGIANS	19	48	6	2	23	0	0	2
TURKS	24	26	3	5	30	2	1	9
OSSETES	43	11	2	6	34	0	0	4
INDIANS	17	28	31	0	12	0	0	12

* All numbers are rounded to closest integer.

Table 2

Phylogeography of the spread of sub-clusters of haplogroup U, the second most frequent variety of maternal lineages among Caucasoid populations

population	%U	sub-cluster of U (in % of U)								
		U*	U1	U2	U3	U4	U5	U6	U7	K
Finno-Urgians	26	3	1	2	1	20	62	0	0	11
Slavs	23	5	3	10	7	15	41	0	0	19
Mediter.	23	3	8	4	4	10	33	2	1	35
Turks	24	5	15	4	22	4	20	0	6	24
Indians	13	0	2	78	0	5	1	0	13	1

U*- unclassified varieties of haplogroup U mtDNAs; Slavs - Russians, Poles, Slovaks, Czechs; Mediter. – European Mediterranean populations from southern France, Italy, Greece, Iberian peninsula. All numbers are rounded to closest integer. Sizes of sample populations lie between 300 to 1,000.