

THE TRANS-CAUCASUS AND THE EXPANSION OF THE CAUCASOID-SPECIFIC HUMAN MITOCHONDRIAL DNA

Ene Metspalu, Toomas Kivisild, Katrin Kaldma, Jüri Parik, Maere Reidla, Kristiina Tambets, and Richard Villems

Department of Evolutionary Biology
Tartu University and Estonian Biocentre
Tartu, Estonia

ABSTRACT

The topology of the network of western Eurasian mitochondrial DNA (mtDNA) lineage clusters in the context of their expansion and spread in this geographic area is analysed. Special attention is devoted to the inner nodes of the reconstructed median network tree, ancestral to mtDNA lineage clusters H and V, to the Caucasus and Trans-Caucasus area populations and to the problem of timing of the expansion of the Caucasoid-specific mtDNA lineage clusters in western Europe *versus* in the Trans-Caucasus. It appears on several examples that typical for Western Europe mtDNA lineage clusters exhibit significantly earlier expansion in the Trans-Caucasus area. Furthermore, the lineage cluster, radiating from the pre-HV node, is significantly more frequent and divergent in the Trans-Caucasus populations than it is in Europe. Meanwhile, a comparison of the Central Asian and the Trans-Caucasus area populations shows that the former have a significant share of eastern Asian-specific mtDNA lineages, which are almost absent in the latter. Finally, a picture starts to emerge, revealing an ancient Indian—Trans-Caucasian—European continuum of a significant proportion of human maternal lineage clusters, dating back to the period between the Upper and Lower Pleniglacials.

1. INTRODUCTION

When listing specific merits of the investigation of the phylogeny of human mtDNA lineages, properties like maternal inheritance, homoplasmy and lack of recombination are usually stressed. Furthermore, as far as the mitochondrial DNA D-loop is concerned, neutral molecular evolution is usually assumed. From time to time, all the four mentioned special features of mtDNA are questioned, but it seems that the deviations from these rules are in general rare. However, there is one more, equally important but less empha-

sised aspect: unlike many other genetic markers, a network of combinatorial haplogroups has been worked out for mtDNA polymorphisms, allowing to differentiate between the major human races.

In world-wide mtDNA sequence data bases, Europeans are relatively "oversampled". This is, however, a very superficial generalisation; even in western and central Europe, there are wide undersampled areas, including France, Poland, southeastern Europe etc. In fact, many of the European and neighbouring populations are only marginally or not at all investigated. Being this gap only quantitative, it would not be of a significant hindrance. However, very limited data on Slavs, on Balkan area populations and specifically on the populations, living in the Caucasus and the Trans-Caucasus area, make this gap qualitative.

The classification of European-type mtDNA lineages (Torroni *et al.*, 1994; Torroni *et al.*, 1996) led to the defining of a limited number of major (H, J, K, T, U) and minor (I, V, W, X) Caucasoid-specific mtDNA haplogroups (lineage clusters), which describe at least 95% of mtDNA variations in Europe and in the Middle and Near East. The same haplogroups seem to be well represented also in northern Africa and in the Middle and Near East, but they are virtually absent in populations, living in sub-Saharan Africa and in eastern Eurasia, Australia and Polynesia (for review see: Wallace 1995). They are also absent in American Natives, except for haplogroup X, which has found its way to some northern American Amerind populations (Brown *et al.*, 1998).

Perhaps the most intriguing question at present is the origin of these mtDNA clusters and the timing of their spread in Europe. The first massive study of the European mtDNA pool (Richards *et al.*, 1996) postulated overwhelmingly Palaeolithic presence and expansion of the major extant mtDNA clusters in Europe: the observed variability suggested expansion times mostly from Middle to Upper Palaeolithic. Further similar mtDNA studies brought additional details into the picture but did not alter the basic conclusion (e.g. Torroni *et al.*, 1998). This view is at least seemingly in conflict with the demic diffusion model of the Neolithic spread of new genes into Europe, carried by arriving from the Fertile Crescent farmers (summarised in Cavalli-Sforza *et al.*, 1994). The debate is still ongoing and recent interpretation of the diversity of nuclear DNA markers in Europe (Chikhi *et al.*, 1998) suggests a largely Neolithic ancestry of the European genes.

Here we summarise our own and published by others experimental data on some hitherto under-investigated populations and analyse the results in the context of the diversity and phylogeny of western Eurasian mtDNA lineages. We pay specific attention to the Trans-Caucasus area and show that a number of mtDNA lineage clusters, shared by those present also in Europe, started to expand in this area much earlier.

Some definitions and parallel terms used in this paper.

To avoid misunderstandings, we note that:

- a) *haplogroup* and *lineage cluster* are used as synonyms;
- b) *lineage* is used as an equivalent to *haplotype*;
- c) as far as direct sequence analysis of the mtDNA is concerned (i.e. mostly the sequence of the first hypervariable region (HVR I)), we speak about a haplotype as about *an array of mutations*, accumulated in mtDNA genome, relative to its *ancestral state*;
- d) all nucleotide positions (nps) are according to the Cambridge Reference Sequence (Anderson *et al.*, 1981). In figures, an exact nature of mutation is specified only in case of transversions (e.g. 16,318 AG means an A to G transition at np 16,318);
- e) for HVR 1 sequence, all mutation positions are shown less 16,000 relative to CRS; for RFLP positions, an exact np of loss or gain of a restriction enzyme cutting site is given. The latter does not necessarily coincide with a mutation.

which created a particular RFLP polymorphic site; most of the precise locations of these particular mutations are now known (Richards *et al.*, 1998).

2. THE GLOBAL NETWORK OF mtDNA LINEAGES

The accompanying paper (T. Kivisild *et al.*, this book) brings western Eurasian maternal lineages together with those found in India. Therefore, the present contribution mentions India only occasionally. Nevertheless, the reconstruction of human mtDNA lineages in the two papers is based on an identical ground: on parsimonious median networks (Bandelt *et al.*, 1995). In contemporary variations of such networks their "skeleton" is largely based on informative restriction fragment length polymorphisms, whereas its "flesh" comes from the pattern of mutations, found in mtDNA hypervariable region I. One of such trees was recently published (Richards *et al.*, 1998); here we refer to a slightly different variety (see Fig. 1 in Kivisild *et al.*, this book). We unified our haplogroup nomenclature with that in Richards *et al.*, (1998); minor differences are commented below.

Bearing in mind this phylogenetic reconstruction, we first analyse the topology of the skeleton network from the point of view of the origin of the major Caucasoid-specific mtDNA lineage clusters. There are two such nodes: R* and 6 (see Fig. 1 in Kivisild *et al.*, this book). The first is a branching point for a large variety of not only Caucasoid-specific, but also Indian- and eastern Eurasian-specific mtDNA clusters. These lineage clusters encompass more than 90% of mtDNA haplotypes, found so far in Europe. The significance of this node is discussed in detail in Kivisild *et al.* (this book). Below we reconstruct and discuss its immediate descendant: node HV* (Richards *et al.*, 1998). This node is designated as HV* to indicate that mtDNA haplogroups H and V descend from it. And because the latter two encompass about a half of the extant western European mtDNA lineages, this node must be considered as highly important in the understanding of the demographic history of the Caucasoid populations in general.

Secondly, in this paper we analyse a few mtDNA lineage clusters in greater details, in order to illustrate the approach we suggest for the coalescence time estimation. It brings us close to the recently reactivated debate on the ancestry of the European gene pool (e.g. Richards *et al.*, 1996; Jackes *et al.*, 1997; Chikhi *et al.*, 1998)

3. THE ORIGIN OF HAPLOGROUPS H AND V-

Caucasoids (*pro* Caucasians) is not a term used universally. Here we apply it in a traditional sense, covering not really well defined groups of populations, speaking largely Indo-European, Afro-Asiatic and Finno-Ugric languages. It is quite common to add here Indian populations as well. The mtDNA polymorphisms gathered so far is strongly biased: while the western and northern Europe, except France, is investigated in some depth, less is known about Slavic populations and even less about Caucasians living in the Caucasus area and in the Trans-Caucasus. Without them, however, the picture cannot be complete even as the first approximation.

We carried out such an analysis on Ossetes, Armenians and Georgians (summary $n = 536$, where n is a number of samples typed for haplogroup-specific RFLPs and HVR I sequenced). Here, we are particularly interested in selected aspects of these results. First, it turned out that all three populations possess a significant proportion of mtDNA lineages which belong neither to haplogroup H nor V, deriving nevertheless from a node, defined as HV* (Richards *et al.*, 1998). The frequency of such lineages in all three populations is much higher than detected thus far anywhere else in Europe: around 8% (for

comparison, see Table 2 in Kivisild *et al.*, this book). Median network tree of this mtDNA lineage cluster is shown in Fig. 1.

To this reconstruction, we included data from Macaulay *et al.* (1999), where a precise RFLP typing was performed, as well as a few other sequences from available data bases (Miller and Dawson 1998), what we added relying on HVR I sequence alone. We did it because the particular mutations: transversion at np 16,220AC and transition at np 16,067 (see Fig. 1), are very rare or even absent in a world-wide mtDNA sequence databases, making such an identification reliable. A subcluster arising at np 16,311 (Fig. 1) is different: there is a large number of individuals in Europe possessing transition at this np and our analysis of a variety of Europeans shows that an overwhelming majority of such mtDNAs in a super-cluster HV* belong to haplogroup H. We also included our Indian data (Kivisild *et al.*, ms. submitted and Kivisild *et al.*, this book). Note that individuals from the Caucasus area and Anatolia (Armenians, Ossetes, Georgians and Turks) encompass about a half of the sample. However, any comprehensive comparison is at present premature since most of the data available from literature does not include full RFLP typing, like we carried out for Estonians, Armenians, Ossetes, Georgians, Czechs, Gypsies and Indians, which are included into the network, presented in Fig. 1.

Our analysis reveals that node HV*, an ancestral state to the most common for Europe haplogroup H (as well as for haplogroup V, see Fig. 1 in Kivisild *et al.*, this book), is diverse in the Trans-Caucasus area and also in India. Its expansion has probably started independently from a few founders. The node position itself is represented by a few Georgian, Armenian, Kashmir and Czech individuals, while lineages with mutations at np 16,067,16,220AC and 16,311 probably represent independent founders in addition to the node. These three additional founders were likely already present before the expansion phase(s), apparent from Fig. 1, took place. It suggests the occurrence of another, even earlier expansion phase. The presence of several sub-founders makes the coalescence age estimate for the total sample somewhat ambiguous; it is around 40,000 years before present (BP). The three representative sub-founders (the ancestral state and the derivatives after mutations at nps 16,067 and 16,311, all suggest values which are older than the Last Glacial - Maximum LGM): around 25,000-35,000yrs.BP. Note that Europeans (Estonian, Czech, Bulgarian, German, UK, North Irish), although rare in this picture, are distributed over all of the identified subclusters, while Indians from northwestern India (Kashmir, Uttar Pradesh and Lombadis from Punjab) derive from the central node only (Fig. 1).

Taken together, we suggest that the sum of these "pre-HV" lineages can be considered as a separate lineage cluster 6, identified as lacking the *AluI* site at np 7,025, possessing nucleotide A at np 00073 and can be further differentiated the ancestral to it node R* by loss of the *TruI* site at np 14,467 and by possessing A at np 00073. Compared to haplogroup V, haplogroup Ö lineages have *NlaIII* site at np 4,577. The fact that this lineage cluster, ancestral to about a half of all maternal lineages present in western Europe, is frequent and divergent in the Trans-Caucasus area as well as present in north-western India, suggests that it started its expansion (and possibly even originated) from this geographic area. A few western European haplogroup 6 lineages thus far detected, are present in all sub-clusters of this haplogroup and may have been carried to western Europe any time since Europe was inhabited by modern humans.

In this context, it would be intriguing to reveal the presence of haplogroup 6 in yet another possible "fusion" area—in the Central Asia. Recently, about 200 mtDNA HVR I sequences of various populations of this area became available (Comas *et al.*, 1998). Although this study lacked RFLP typing, we reconstructed a tentative median network phylogenetic tree of the published sequences (Fig. 2). Apart of some unavoidable ambi-

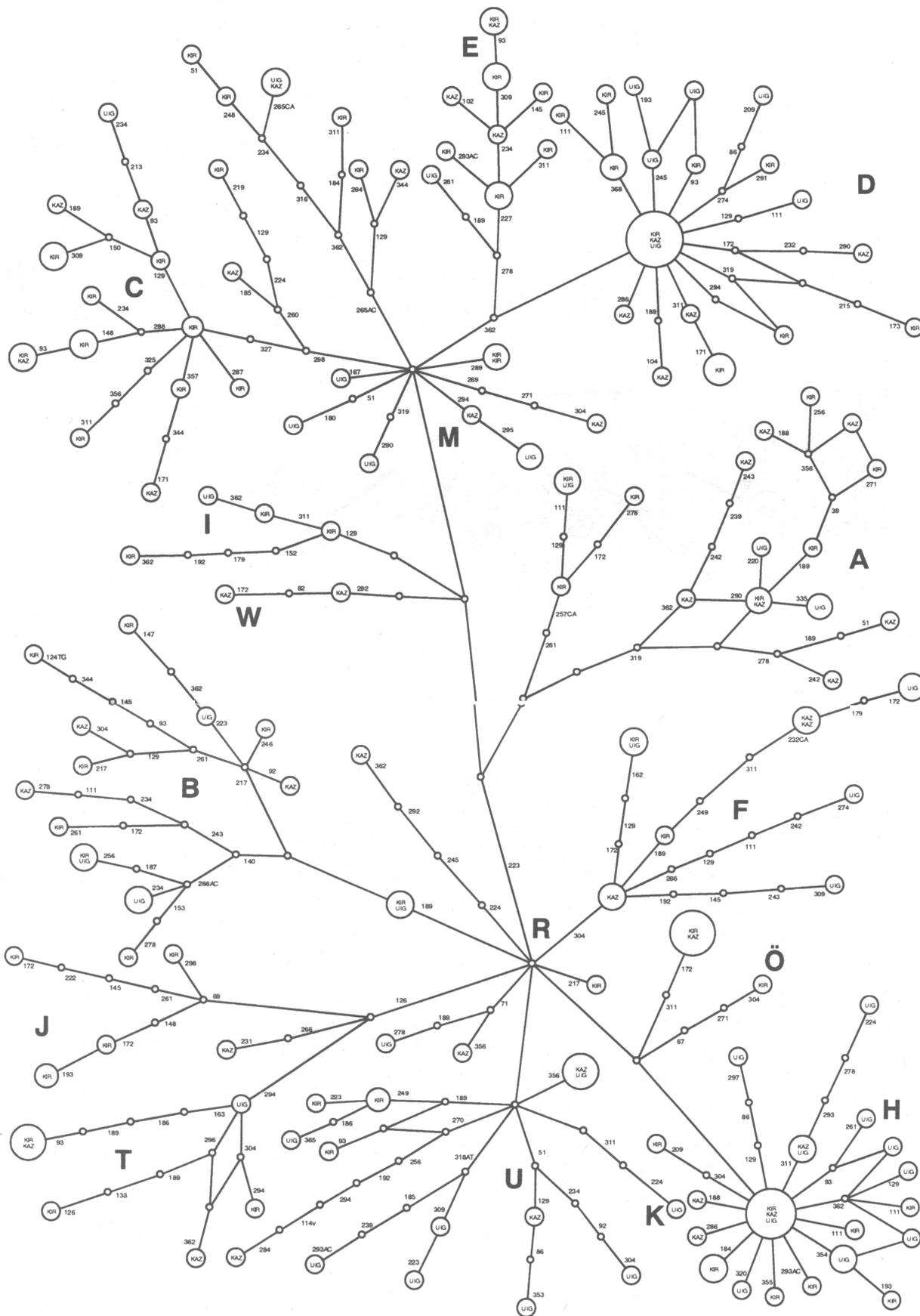


Figure 2. Tentative arranging of the Central Asian mtDNA HVR I sequences into reduced median network. Sequence information from Comas *et al.* (1998).

guities, this reconstruction fully agrees with the conclusion drawn by the authors: the populations studied possess mtDNA lineages, typical both for western and eastern Eurasia. In a set of three populations studied: Altaic language group Kazakhs, Uighurs and Kirgizis, there is one individual having transition at np 16,067 and therefore almost certainly belongs to haplogroup 6. Some of those possessing transition at np 16,311 (motif 16,311 and 16,172, see Fig. 2) may belong to haplogroup 6 as well. The results do show that haplogroup 6 is present also in non Indo-Iranian language group Central Asians, albeit at lower frequency than in the Trans-Caucasian region. On the other hand, the frequency of haplogroup H in Central Asia is also lower: around 15% in these particular populations, compared to about 25% in average for Armenians, Georgians and Ossetes. We note that Ossetes, provided they are direct descendants of historic Alans, probably lived in Central Asia before moving westward about 2,000 BP (Frye and Litvinsky 1996).

Thus, in conclusion we can say that haplogroup 6 is present among Indo-European-, Afro-Asiatic-, Finno-Ugric-, Kartvelian-, North Caucasian- and Altaic-speaking populations. Among Indo-European linguistic super-family, this haplogroup can be found in Europe, in Indo-Iranian-speaking Ossetes as well as among northwestern Indian populations of different localities. The fact that the three Caucasus area populations encompass these mtDNA lineages at nearly equal diversity and frequency (Fig. 1), while revealing several sharp frequency differences in some other mtDNA clusters (Metspalu *et al.*, ms in preparation), is somewhat puzzling. Ossetes are "newcomers" to the area—their arrival to the Caucasus is traditionally linked to the invasion of Hunns in late 4th century, pushing more northern Alans to south, behind Kuban and Terek rivers, from where several later invaders caused their further movement to the High Caucasus. The arrival of Indo-European-speaking Armenians to the political landscape of Trans-Caucasus some 2,600 years ago, although well documented historically, is nevertheless unclear from the point of view of their origin. It can be linked to the massive invasion of Thracians to Anatolia during the Late Iron Age, but possibly also to an invasion from the north—from the Pontus area. And, even if either of the two scenarios is valid, it is still unclear whether the linguistic change (Urartians, in contrast to Armenians, were linguistically not Indo-Europeans) caused any profound changes in the repertoire of maternal lineages of the local people. Finally, Kartvelians (Georgians) are considered to be autochthonous Caucasus-area population. All that lends credence to a possibility that the mtDNA gene pool in this geographic region obtained the founder of this haplogroup long ago, likely early in Upper Palaeolithic. After that, the population perhaps underwent several expansion events, at least one before the LGM. Note that we found haplogroup R individuals (defined as derivatives of the node R*, not belonging to haplogroups D, F, J, T, U and K, see Kivisild *et al.*, this book) also in all three Caucasus populations, albeit at a significantly lower frequencies than in India. Unfortunately, we lack data on Afghani and Iranian populations.

Taking now together all derivatives of mtDNA super-cluster HV* (i.e. H, V and 6), one can see that haplogroup H is the most frequent in Europe, frequent in the Trans-Caucasus and present in northwestern India at a low frequency. Its sister-group, haplogroup V, is present in Europe at a low frequency, virtually absent in the Caucasus and Anatolia, not sampled so far in India and the corresponding lineages are not present in the published so far Central Asian data set. Meanwhile, haplogroup 6 (one can say—the rest of the super-cluster HV*) is uniformly (~8%) present in the Caucasus and found at a low but still well detectable frequency both in India and in Europe and seems to be also present in Central Asia. Further sampling of different populations in Iran, Afghanistan

and Iraq, perhaps also in Turkey, southeastern Europe and in the Ukraine, will possibly add valuable details to the emerging picture of the historic spread of the descendants of HV* in the western Eurasia.

4. CASE STUDIES

Super-cluster HV* is only one of the derivatives of R* (see Fig. 1 in Kivisild *et al.*, this book). All other "classical" European-specific haplogroups that radiate from this node—T, J and U + K, are also present in the populations, living in the Trans-Caucasus (Table 2 in Kivisild *et al.*, this book). Thus, it is interesting to compare, where enough data available, the topology of individual haplogroups of various localities (Europe, the Caucasus area, India, Central Asia etc). Although this analysis does not necessarily reveal the absolute time depth of mtDNA lineage clusters, it does give a reliable comparative estimate of the relative time since their expansion in various geographic localities. Specifically, it allows to share light on the debate about the Palaeolithic *versus* Neolithic ancestry of the European gene pool. Below we analyse several mtDNA lineage clusters, derivatives of R*.

4.1. Sub-Haplogroup U3

This sub-cluster is characterised by transition at np 16,343 and is rare in Europe (see e.g. Fig. 6 in Richards *et al.*, 1998). It seems to be absent also in the published (Comas *et al.*, 1998) Central Asian sample (Fig. 2). In contrast to that, we found this haplogroup present in all three Trans-Caucasus populations studied by us. Furthermore, this sub-cluster was also found among Adygei population, living in the northern Caucasus (Macaulay *et al.*, 1999). Because mutations at np 16,343 are relatively rare, and bearing in mind geographic proximity, we included also several putative sub-cluster U3 sequence variants from Turkey (sequences from Calafell *et al.*, 1996; Comas *et al.*, 1996). The topology of the corresponding cluster (Fig. 3) reveals a star-like expansion event. The Adygei sub-sample (Macaulay *et al.*, 1999) is limited to a single haplotype, whereas Armenians, Turks and Georgians are nearly equally divergent. Omitting Adygei sequences from calculations, we obtained $\rho = 1.42$ and time since the beginning of expansion about 30,000yrs. BP ($29,500 \pm 5,000$), i.e. well before the LGM.

4.2. Haplogroup J in the Trans-Caucasus Populations

Haplogroup J has complex topology in Europe with several sub-founders (Richards *et al.*, 1998). This cluster is also well represented in Armenians, Georgians and Ossetes (Fig. 4). The basic topology of the European (see Fig. 7 in Richards *et al.*, 1998) and of the Trans-Caucasus clusters overlap, but there the similarity ends. In Europe, the estimated expansion times for the node J* (which is also clearly the predominant sub-cluster of haplogroup J in Europe) is about 8,000yrs. BP (Richards *et al.*, 1998 and our unpublished results). For the same sub-cluster in the Caucasus—Anatolian area, the time since its expansion is about 27,000yrs. BP. Sub-cluster J2, infrequent in Europe and present there predominantly in the Mediterranean area, is equal in frequency to J* in the Caucasus area.

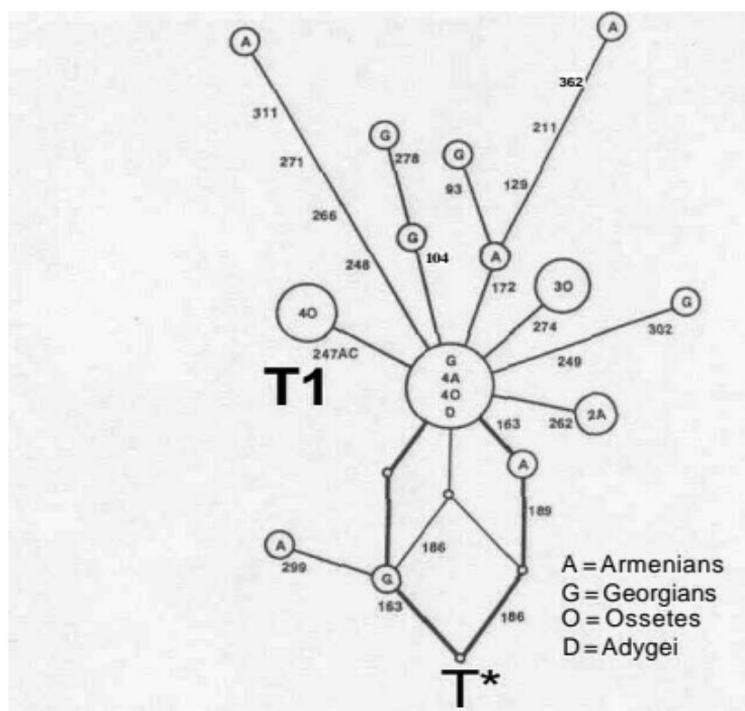


Figure 5. Median network tree for mtDNA haplogroup T sub-cluster T1 for Armenians, Georgians, Ossetes and Adygeis.

expansion time estimates for different mtDNA lineage clusters in the populations of the Caucasus area yield very close values. This seems to make sense: Palaeolithic population expansion waves were likely initiated by changes (climatic or else), which affected equally and simultaneously all present at that time founders. Indeed, it would much complicate any explanation, if all the varieties of mtDNA lineages have had different expansion times in the same area. Third, maternal lineages of the three different Caucasus area populations are in principle close to those found in western Europeans. Frequencies of certain haplogroups vary, those of many sub-clusters even more so, but none of the three populations studied possess at any significant quantities mtDNA lineages, which are either novel or belong to haplogroups, characteristic for eastern Asia or Africa.

Finally, it seems to us that the deconstruction of mtDNA lineage clusters into sub-founder-based units largely solves the problem of the initial controversy between the experimental data behind the "Neolithic demic diffusion" and "predominantly Palaeolithic mtDNA gene pool" in Europe. Not to add that irrespective of the area where any of those expansions took place, the expanding maternal lineages were phylogenetically close. It appears that neither India nor Central Asia—two localities with equally high incidence of haplogroup M, have contributed significantly into the western Eurasian mtDNA gene pool during many tens of thousands of years. Hence, there is little doubt that the Upper Palaeolithic mtDNA lineages in western Eurasia were qualitatively the same as they are at present. Therefore, their time depth is identical, irrespective of loca-

lion, It does not, however, mean that there was no substantial gene flow from Anatolia to western Europe during early Neolithic. Much earlier expansion phases, observable in the same mtDNA lineage clusters and sub-clusters in the populations, living at present in the Trans-Caucasus area and in Turkey, compared to those in western Europe, support this suggestion.

6. REFERENCES

- Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics* 141:743-753.
- Brown MD, Hosseini SH, Torroni A, Bandelt H-S, Alien JC, Schurr TO, Scozzari R, Cruciani F, Wallace DC (1998) MtDNA haplogroup X: An ancient link between Europe/West Asia and North America? *Am J Hum Genet* 63:1852-1861.
- Calafell F, Underhill P, Tolun A, Anglicheva D, Kalaydjieva L (1996) From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann Hum Genet* 65:35-49.
- Cavalli-Sforza LL, Menozzi P, Piazza A (1994) *The history and geography of human genes*. Princeton, NJ: Princeton University Press.
- Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G (1998) Clines of nuclear DNA markers suggest a largely Neolithic ancestry of the European gene pool. *Proc Natl Acad Sci USA* 95:9053-9058.
- Comas D, Calafell F, Mateu E, Perez-Leuzan A, Bertranpetit J (1996) Geographic variation in human mitochondrial DNA control region sequence: The population history of Turkey and its relationship to the European populations. *Mol Biol Evol* 13:1067-1077.
- Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bosch E, Martinez-Arias R, Clarimon J, Facchini F, Fiori G, Luiselli D, Pettener D, Bertranpalit J (1998) Trading genes along the Silk Road: Mitochondrial DNA sequences and the origin of Central Asian Populations. *Am J Hum Genet* 63:1827-1838.
- Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of Native American mtDNA variation: A reappraisal. *Am J Hum Genet* 59:935-945.
- Frye RN, Litvinsky BA (1996) In: *History of Humanity, Vol III*. pp 461-464 Eds. J Hermann, E. Zurcher. UNESCO and Routledge, Clays Ltd, Surrey, UK.
- Jackes M, Lubell D, Meiklejohn C (1997) On physical anthropological aspects of the Mesolithic-Neolithic transition in the Iberian Peninsula. *Current Anthropology* 38:839-846.
- Larsson NG, Clayton DA (1995) Molecular genetic aspects of human mitochondrial disorders. *Annu Rev Genet* 29:151-178.
- 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.
- Miller KWP, Dawson JL (1997&1998) Mitochondrial DNA Concordance, <http://shelob.bioanth.cam.ac.uk/mtDNA>.
- Richards M, Corte-Real H, Forster H, Macaulay V, Wilkinson-Hebbots H, Demaine A, Papiha S, Hedges R, Bandelt H-J, Sykes B (1996) Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 59:185-203.
- Richards MB, Macaulay VA, Bandelt H-J, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. *Ann Hum Genet* 62:241-260.
- Torroni A, Lott MT, Cabell MF, Chen Y-S, Lavergne L, Wallace DC (1994) mtDNA and the origin of Caucasians: Identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. *Am J Hum Genet* 55:760-776.
- Torroni A, Houpanen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus M-L, Wallace DC (1996) Classification of European mtDNA from an analysis of three European populations. *Genetics* 144:1835-1850.
- Torroni A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, Scillito D, Rengo C, Forster P, Savontaus M-L, Bonne-Tamir B, Scozzari R (1998) mtDNA analysis reveals a major Late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet* 62:1137-1152.
- Wallace DC (1995) Mitochondrial DNA variation in human evolution, degenerative disease, and aging. *Am J Hum Genet* 57:201-223.