Am. J. Hum. Genet. 64:232–249, 1999

The Emerging Tree of West Eurasian mtDNAs: A Synthesis of Control-Region Sequences and RFLPs

Vincent Macaulay,[1] Martin Richards,[1] Eileen Hickey,[1] Emilce Vega,[1] Fulvio Cruciani,[2] Valentina Guida,[2] Rosaria Scozzari,[2] Batsheva Bonné-Tamir,[3] Bryan Sykes,[1] and Antonio Torroni[2,4]

[1]Institute of Molecular Medicine, University of Oxford, Oxford; [2]Dipartimento di Genetica e Biologia Molecolare, Università di Roma "La Sapienza," Rome; [3]Department of Human Genetics, Sackler Faculty of Medicine, Ramat-Aviv, Israel; and [4]Istituto di Biochimica, Università di Urbino, Urbino, Italy

Summary

Variation in the human mitochondrial genome (mtDNA) is now routinely described and used to infer the histories of peoples, by means of one of two procedures, namely, the assaying of RFLPs throughout the genome and the sequencing of parts of the control region (CR). Using 95 samples from the Near East and northwest Caucasus, we present an analysis based on both systems, demonstrate their concordance, and, using additional available information, present the most refined phylogeny to date of west Eurasian mtDNA. We describe and apply a nomenclature for mtDNA clusters. Hypervariable nucleotides are identified, and the relative mutation rates of the two systems are evaluated. We point out where ambiguities remain. The identification of signature mutations for each cluster leads us to apply a hierarchical scheme for determining the cluster composition of a sample of Berber speakers, previously analyzed only for CR variation. We show that the main indigenous North African cluster is a sister group to the most ancient cluster of European mtDNAs, from which it diverged ~50,000 years ago.

Introduction

Mapping of the variation in the modern mitochondrial landscape of Europe and the Near East is shedding light on colonization and the dispersal of peoples in this region (Torroni et al. 1994*a*, 1996; Richards et al. 1996) and also on the etiology of a number of diseases (Hofmann et al. 1997; Torroni et al. 1997). Because of the mode of inheritance of mtDNA (maternal transmission and lack of recombination), the mutations that have struck throughout human history trace the maternal genealogy; through this genealogy and the impact that demography has had upon it, we can attempt to infer something about prehistoric processes. Another virtue of mtDNA is that a high rate of base substitutions, compared with that of nuclear DNA, allows the genealogy to be captured in a fair amount of detail. However, this high rate also has its drawbacks, in that the treelike relationship of modern mtDNAs is obscured by recurrent mutations affecting the same nucleotide positions (nps), which are often difficult to resolve.

For some time, most studies of mtDNA variation have been conducted by use of one of two methods that assay largely different portions of the molecule: direct sequencing of the especially fast–evolving control region (CR) and digestion of the entire molecule by means of standard sets of restriction enzymes. Richards et al. (1996) produced a phylogenetic network of sequences from the first hypervariable segment (HVS I) of the CR and identified six major mtDNA clusters among Europeans. By using high-resolution RFLP analysis, Torroni et al. (1994*a*) had previously identified four clusters (H, I, J, and K) among North Americans of European ancestry. Subsequently, Torroni et al. (1996) applied the same methodology to two Scandinavian population samples and identified five additional clusters (T, U, V, W, and X), which, together with the previous four clusters, appeared to encompass virtually all examined European mtDNAs. In the same article (Torroni et al. 1996), a small set of Tuscan mtDNAs previously studied for HVS I variation (Francalacci et al. 1996) was used to demonstrate the correlation of clusters defined with the two systems, by testing the same samples for RFLP cluster–diagnostic mutations. Further work in one system or the other has illuminated the phylogeny (Bertranpetit et al. 1996; Calafell et al. 1996; Wilkinson-Herbots et al. 1996; Torroni et al. 1997, 1998; Richards et al.

Received August 6, 1998; accepted for publication October 19, 1998; electronically published January 6, 1999.

Address for correspondence and reprints: Dr. Vincent Macaulay, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford, OX3 9DS, United Kingdom. E-mail: vincent.macaulay@cellsci.ox.ac.uk

0002-9297/99/6401-0030$02.00

1998). Hofmann et al. (1997) recently reported coding-region and CR sequences from German subjects, who are very representative of the general west Eurasian mtDNA variation, which is relatively homogeneous (Pult et al. 1994; Richards et al. 1996).

In this study, we first performed a phylogenetic analysis of the data of Hofmann et al. (1997), which allowed us to determine some of the deep splits in the west Eurasian phylogeny. We report and analyze combined RFLP haplotypes and HVS I sequences of 95 subjects from two geographic areas (the Near East and Caucasus) that are thought to be close to the origins of important population expansions into Europe, to present a detailed picture of the west Eurasian phylogeny. Finally, we used all available information to determine the cluster composition of a population from North Africa and to draw some conclusions about the origins of this population's mtDNA gene pool.

Subjects and Methods

Subjects

DNA samples from 50 Adygei from the northwest Caucasus and 45 Druze from northern Israel were obtained from maternally unrelated individuals. The subjects of our subsequent case study were a number of Mozabites, Berber speakers from Ghardaia in northern Algeria, a subset of those previously investigated by Côrte-Real et al. (1996).

mtDNA Analysis

The entire mtDNA sequence of each Druze and Adygei sample was amplified in nine overlapping fragments, by PCR using the primer pairs and amplification conditions described by Torroni et al. (1993, 1997). Each of the nine PCR segments was digested with 14 restriction endonucleases (*Alu*I, *Ava*II, *Bam*HI, *Dde*I, *Hae*II, *Hae*III, *Hha*I, *Hin*cII, *Hin*fI, *Hpa*I, *Msp*I, *Mbo*I, *Rsa*I, and *Taq*I). A polymorphism at np 12308 was assayed as described by Torroni et al. (1996). In addition, all subjects were screened for the presence of *Nla*III sites at nps 4216 and 4577, a *Bfa*I site at np 4914, *Acc*I sites at nps 14465 and 15254, a *Bst*OI site at np 13704, and an *Mse*I site at np 14766. We refer to the full set of digests as the "extended 14-enzyme system."

HVS I of the CR was amplified as described elsewhere (Richards et al. 1996) and was sequenced at the University of Florida DNA Sequencing Core Laboratory, by use of ABI Prism Dye Terminator cycle-sequencing protocols developed by Applied Biosystems (Perkin-Elmer). The fluorescently labeled extension products were analyzed on an Applied Biosystems Model 373 Stretch DNA sequencer or on a 377 DNA sequencer (Perkin-Elmer).

Within hypervariable segment II (HVS II), a 241-bp sequence surrounding np 00073 (we followed the numbering scheme used by Anderson et al. [1981]) was amplified by use of primers 5′-GGTCTATCACCCTATTAACCAC-3′ (light chain, nps 00008–00029) and 5′-TCAATTGTTATTATTATGTCCTACAA-3′ (heavy chain, nps 00242–00223), with annealing at 50°C. PCR products were ethanol precipitated to standardize concentrations to 20 ng/μl, and 1 μl of denatured product was dot blotted onto Hybond N membranes (Amersham). Position 00073 was assayed with the probes reported by Stoneking et al. (1991), by use of the Digoxigenin kit (Boehringer Mannheim). Membranes were stripped of probe by washing in sterile water for 3 min at 95°C and were rehybridized with the reciprocal probe, to verify the authenticity of the results.

Phylogenetic Methods

To infer the phylogenetic relationships between haplotypes comprising various combinations of HVS I, 00073, RFLPs, and coding-region polymorphisms, reduced median networks of data sets were constructed (Bandelt et al. 1995) by use of the program NETWORK (Röhl 1997).

Relative-Rate Estimation

To estimate the relative mutation rate of the RFLP and HVS I systems, a Bayesian analysis was developed. Suppose that n_1 and n_2 mutations are observed in two systems evolving along the same genealogy, of total length T, where the mutation rate for system 1 is μ and that for system 2 is $r\mu$; that is, the relative rate is r. The sampling probabilities (or likelihoods) $p(n_1|\mu, T)$ and $p(n_2|\mu, r, T)$ (for which the usual notation for conditional probability has been used) are Poissonian, with parameters μT and $r\mu T$, respectively. Then, the posterior distribution of r, given n_1 and n_2, can be obtained from

$$p(r|n_1, n_2) = \int_0^\infty p(n_1|s)p(n_2|r, s)p(r, s)ds$$

(e.g., see Jeffreys 1983), where $s = \mu T$. Uninformative Jeffreys priors are appropriate to s and r: $p(r, s) \propto (rs)^{-1}$. Hence,

$$p(r|n_1, n_2) = \left[\frac{(n_1 + n_2 - 1)!}{(n_1 - 1)!(n_2 - 1)!}\right]\left[\frac{r^{n_2-1}}{(r + 1)^{n_1+n_2}}\right].$$

The most probable relative rate obtained from this distribution is $(n_2 - 1)/(n_1 + 1)$; credible regions (Berger 1985) that covered the central 95% of the probability distribution were computed.

Cladistic Nomenclature for Human mtDNA

In a previous article (Richards et al. 1998), we proposed a flexible and consistent nomenclature for mtDNA clades, which is reviewed here. The set of all mtDNAs derived by descent from any maternal ancestor could be distinguished, in principle, by a name. In practice, only clades with, for example, interesting geographical patterning or those derived from major early branchings of the phylogeny need to be named. A named clade is called a "cluster." All clusters should be monophyletic or at least should seem so. (Increased resolution may reveal that a mutation defining a cluster is not a single event, in which case the cluster definition must be revised.) Major clusters are denoted by uppercase roman letters, and most clusters approximate the existing RFLP haplogroups (Torroni et al. 1994*a,* 1994*b,* 1997; Chen et al. 1995). It is possible for clusters to be nested; for example, RFLP haplogroups (and, hence, the clusters) C, D, E, and G are contained in M (e.g., see Torroni et al. 1994*b*). Successively nested subclades of major clusters are named by alternating positive integers and lowercase roman letters; for example, J1b1 ⊂ J1b ⊂ J1 ⊂ J, where "⊂" means "is a subcluster of." A cluster that is composed of a set of named subclusters is referred to by concatenating the subcluster names; for example, the smallest cluster that includes H and V is called "HV." To designate the set of mtDNAs—in general, not a clade itself—that coalesce in an (as yet) unresolved multifurcation but that are not members of any of the clusters branching from that node, we append an asterisk (*) to the list of clusters (e.g., "AB*" in fig. 1). Unnamed clades that enclose clusters are indicated by the prefix "pre-" (e.g., "pre-AB" in fig. 1).

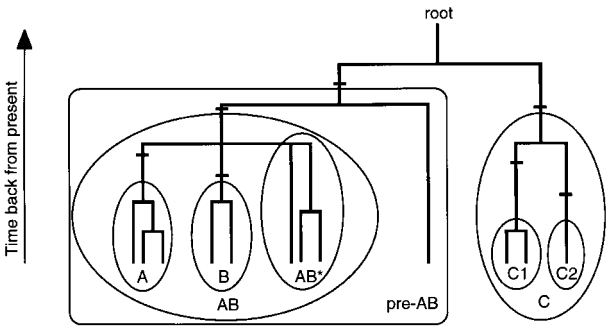


Figure 1 Cartoon mtDNA tree showing the principles of the cladistic notation, for hypothetical clusters A, B, and C.

Results and Discussion

Establishing the mtDNA Phylogeny

Analysis of the coding region and the HVS I and HVS II sequences of Hofmann et al. (1997) established a framework for the west Eurasian phylogeny. The network of this data set (fig. 2) is very treelike, with the slower mutation rate of the majority of sites assayed in the coding region, having allowed homoplasy at sites both in the CR and at hypervariable coding-region positions to be resolved. Relationships between clusters inferred on the basis of RFLP and CR data alone are strengthened by the presence of additional characters: (1) the 16126C shared by clusters T and J is supported by 4216C (i.e., +4216*Nla*III) and 11251G, and (2) the 12308G shared by U and K is supported by 12372A. Most clusters are also better resolved with new characters—for example, J, T, and U4.

The relationship between RFLP haplogroups U and K has been clarified. Previously, these groups branched from an unresolved multifurcation characterized by the 12308G state. Now, K clearly is embedded within U and shares 11467G with at least the subcluster U5 but not with U4. To reinstate U as a valid cluster in our nomenclature, we enlarged it to include K as a subcluster.

The inclusion of the HVS II characters 00150, 00152, and 00195 in the analysis introduced considerable ambiguity in the phylogeny, manifested, for example, by reticulations in a network (not shown). This is in full accordance with their inferred high mutation rates (Torroni et al. 1996; authors' unpublished data). Indeed, a most parsimonious reconstruction (MPR; Swofford and Olsen 1990) of the states of these sites on the network of figure 2 requires at least three, nine, and six mutations, and only two other positions in HVS II—namely, 00146 and 00200—attain a maximum of three hits. (The length polymorphisms in the cytosine tracts 00303–00309 and 00311–00315 are not considered here.) The spectrum of hits on HVS I characters is broadly in line with the published lists (Hasegawa et al. 1993; Wakeley 1993), with 16189 and 16311 leading (with at least five and four hits, respectively) and the usual suspects—for example, 16362—following. Of those coding-region characters studied by Hofmann et al. (1997), 10398 appears to have undergone the most recurrent mutation (three hits), which is in accordance with the behavior of the 10394*Dde*I RFLP (see below).

Supporting evidence emerges for the interpretation of the relationship between African and Eurasian mtDNAs, proposed by Watson et al. (1997). Eurasian mtDNAs are split, by 16223C/T, into two substantial classes (e.g., 16223T is in 7% of Europeans [Richards et al. 1998] and 65% of Mongolians [Kolman et al. 1996]), whereas Africans predominantly have 16223T [91%; Watson et al. 1997]). The 16223T state also characterizes the Neanderthal sequence (Krings et al. 1997). However, whether the thymine-cytosine transition, inferred to have occurred around the time of "out-of-Africa" event, happened just once has never been clear. This np is known

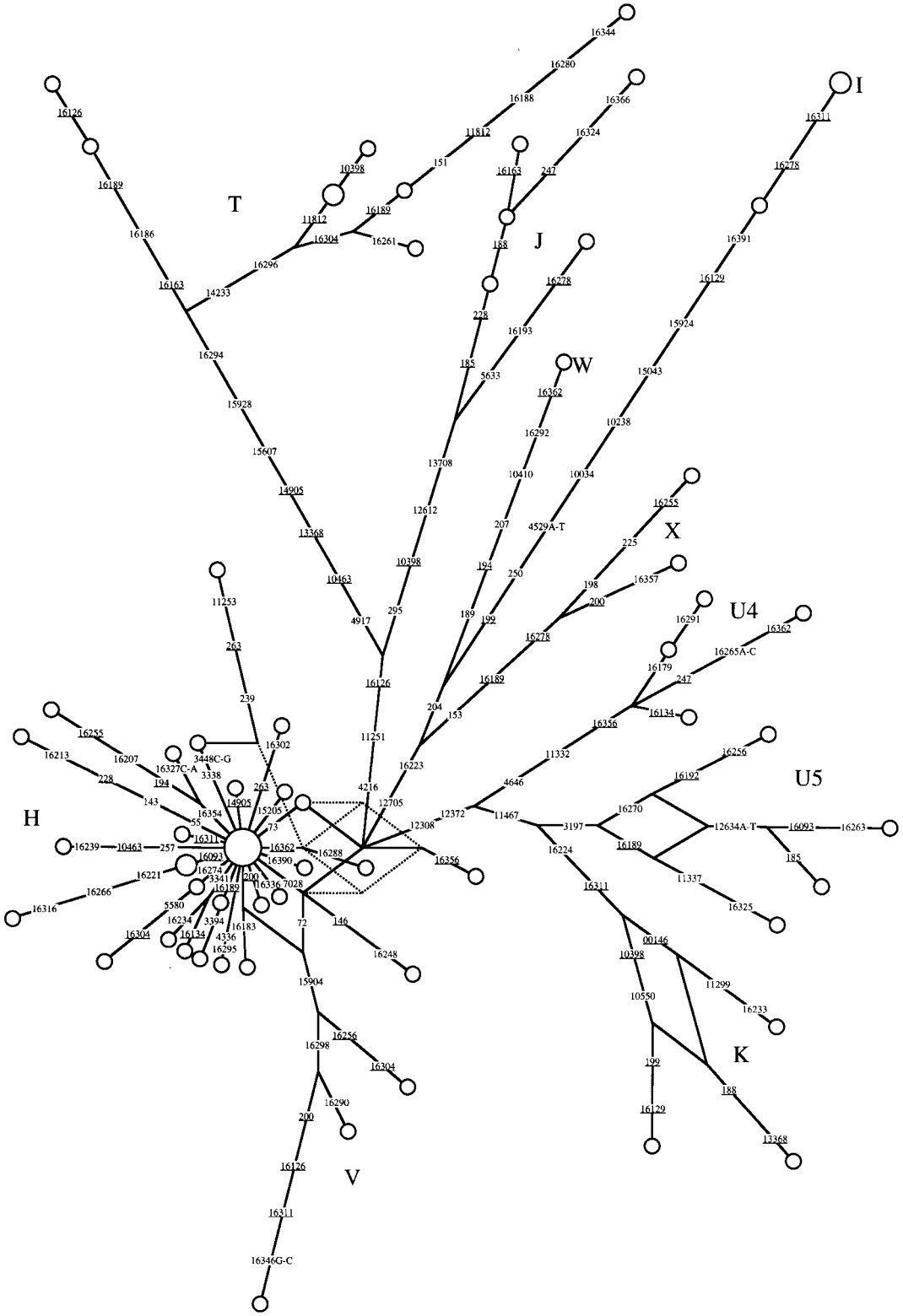


Figure 2 Reduced median network (Bandelt et al. 1995) of the data set of Hofmann et al. (1997). Insertions and deletions were ignored, as was variation at the hypervariable positions 00150, 00152, and 00195 (Torroni et al. 1996) and length variation in the cytosine tracts of HVS II. Circles indicate coding region/HVS I/HVS II haplotypes, with their size proportional to the haplotype frequency in the sample. Mutations are shown on the branches, and the numbering of the CRS was used (Anderson et al. 1981); resolved recurrent mutations are underlined, and unresolved events are shown by reticulation. Mutations are transitions, unless the base change is explicitly indicated. Implausible links are shown as dotted lines. Sample 722 is omitted because a test showed the presence of the 12308*Hin*fI restriction site, suggesting that the published coding-region variants are problematic; it is likely to be a cluster U1 member, as suggested by its HVS I sequence.

to have a moderately fast mutation rate (Hasegawa et al. 1993). However, when a mutation occurs deeply in a phylogeny, its rate may be overestimated because of an incorrect resolution of homoplasy; this is likely to be the case for 16223 in the analysis by Wakeley (1993). Hofmann et al. (1997) identified a thymine-cytosine transition at np 12705 that also splits off the 16223C clusters (J, T, U, H, and V). Thus, we confidently can identify a single common 16223 event, at least for these clusters. The complete African sequence determined by Horai et al. (1995) (chosen to be an early branch of the human phylogeny, designated as cluster L1a in the notation of Watson et al. [1997]) includes 12705T, as do the 16223T sequences in the complete Japanese sequences determined by Ozawa et al. (1991). This is fully consistent with modern Eurasian mtDNA being derived from the 16223 sequence (in HVS I), which, during an Upper Paleolithic expansion, gave rise to, among others, clusters A, I, M, W, X, and, after the 16223T-C/12705T-C events, all the reference-sequence–derived clusters (e.g., B, F, T, J, U, H, and V), in the concomitant rapid branching of the genealogy.

RFLP and CR Data from the Druze and Adygei

Having identified a number of important features at the heart of the phylogeny, we now turn to the new data obtained from the analysis of the 45 Near Eastern Druze and 50 north Caucasian Adygei. The RFLP haplotypes, HVS I sequences, and status at np 00073 in HVS II of these mtDNAs are reported in table 1, and the phylogenetic relationships of the combined haplotypes are shown in figure 3. This phylogeny departs from a reduced median network of the combined RFLP and CR variation (except that the hypervariable character 16517*Hae*III is suppressed; Chen et al. 1995) when additional information, derived both from the above discussion of the data of Hofmann et al. (1997) and from our extensive database of mtDNA HVS I and RFLP types, strongly suggests an alternate topology, as follows.

Cluster U.—The order of mutations in U3 is established from two intermediate HVS I sequences with 00073G—namely, 16343, observed throughout Europe, and 16168–16343, observed in southeast Europe. The order in U5 has been described elsewhere (Richards et al. 1998) and is based on widespread intermediates, the uncovering of the transition and reversion at 16192 depending on the (compatible) mutations 16256C-T and 16399A-G (+16398*Hae*III). Parallel mutations at 16311 have been tentatively proposed for K and R1 (defined below) on the basis of the known high mutation rate of 16311 (e.g., its four appearances in fig. 2), compared with the stability of 12308 (one appearance in fig. 2), and an intermediate observed in a Tuscan (Torroni et al. 1996) with 16311T. On the basis of this interpretation, we would expect R1 samples to have 11467A and 12372G (Hofmann et al. 1997).

Clusters I, W, X, M, and C.—With the exception of X, these clusters occur at very low frequencies in this data set; thus, it was necessary to draw largely on additional information. The global consensus in all these clusters is a T at 16223; therefore, we inferred that the W sequences and a minority of the X sequences have undergone separate 16223T-C mutations, which is not unexpected given the mutation rate at this site (Hasegawa et al. 1993). In addition, two transitions at 16126 can be resolved in X because of the presence of European intermediates (Brown et al. 1998). The likely order of mutations in the predominantly Asian cluster M can be inferred from intermediates in the Mongolian data of Kolman et al. (1996). All the ambiguity that remains is caused by the three RFLPs (1715*Dde*I, 8249*Ava*II/8250*Hae*III, and 10394*Dde*I), leading to the cube in the network, so that the branching order of these clusters is still unclear.

The 16223T-C mutation identified above as causing a deep split in the phylogeny is involved here in reticulation with only the hypervariable 10394*Dde*I site. The clade defined by the 16223C state has been named "R." On the basis of African data (Chen et al. 1995; Watson et al. 1997), we placed the root of this network at the empty node marked with an arrow in figure 3.

Other Clusters.—Six previously undescribed clusters also emerged:

1. U1, a subcluster of U (HVS I motif: 16189–16249; RFLP motif: −4990*Alu*I, +12308*Hinf*I, −13103*Hinf*I/+13104*Mbo*I, +14068*Taq*I), is distributed predominantly in the Near East and Mediterranean Europe.
2. U2, also a subcluster of U (probable HVS I motif: 16051–16129C; probable RFLP motif: +15907*Rsa*I), is present throughout the Near East and Europe.
3. A single haplotype in the Druze that is part of HV* (00073A and 14766*Mse*I; HVS I motif: 16067). A scan of the HVS I database revealed candidate members of this cluster in southern and eastern Europe, as well as in the Near East. The RFLP motif appears to be −8012*Rsa*I (haplotype 44 in Torroni et al. 1997).
4. A cluster observed in the Druze (HVS I motif: 16126–16362). Its phylogenetic position still is somewhat obscure, since it introduces an incompatibility between 00073 and 16126: 16126 might be the same event as that in JT, making it a member of pre-JT that has mutated in parallel to 00073A; alternatively, 16126 might have mutated twice, and the cluster might be part of pre-HV. The 16126–16362 cluster is rare but geographically widespread in southern Europe and the Near East.
5. A cluster (called "R1"; HVS I motif: 16311) that branches from the node marked with an asterisk (*) in

Table 1

RFLP and HVS I Haplotypes and HVS II np 00073 Status of Druze and Adygei Samples

Cluster and Sample[a]	RFLP Haplotype[b]	HVS I Haplotype[c]	00073 Status[d]
H:			
AD07	−**7025a,**−**14766u,**+16517e	0	**A**
AD11	−**7025a,**−**14766u,**+16517e	0	**A**
AD23	−**7025a,**−**14766u,**+16517e	0	**A**
AD25	−**7025a,**−**14766u,**+16517e	0	**A**
AD27	−**7025a,**−**14766u,**+16517e	0	**A**
AD36	−**7025a,**−**14766u,**+16517e	0	**A**
AD31	−**7025a,**−**14766u,**+16517e	0	**A**
AD40	−**7025a,**−**14766u,**+16517e	189 223 356	**A**
AD34	−**7025a,**−**14766u,**−16208k,+16517e	209	**A**
AD26	−951j,+4769a,−5176a,−**7025a,**−**14766u,**+16517e	354	**A**
DR02	−**7025a,**+13100i,−**14766u**	111 288 362	**A**
DR17	−**7025a,**+9493g,−9553e,−**14766u,**+16517e	0	**A**
DR26	−5003c/+5004r,−**7025a,**−**14766u**	093	**A**
DR27	−**7025a,**+*9336k,*−**14766u,**+16517e	192	**A**
DR45	+4793e,−6296c,−**7025a,**−**14766u,**+16517e	093 265	**A**
DR46	−**7025a,**−14258m,−**14766u,**−16310k,16517e	311	**A**
AD09	−951j,+4769a,−**7025a,**−**14766u**	0	**A**
AD14	−**7025a,**+9253e,−**14766u,**−16310k	092 189 293 311	**A**
AD21	−**7025a,**+9253e,−**14766u,**−16310k	092 189 293 311	**A**
AD22	−**7025a,**−16303k,−**14766u**	178 291 304	NA
AD37	−1715c,+4793e,−**7025a,**−**14766u,**+16517e	261 291	**A**
HV*:			
DR07	+5133a,−6262i,−8012k,−**14766u**	067	NA
DR18	+5133a,−6262i,−8012k,−**14766u**	067	**A**
DR20	+5133a,−6262i,−8012k,−**14766u**	067	**A**
DR31	+5133a,−6262i,−8012k,−**14766u**	067	**A**
U*:			
DR30	+**12308g,**−15073c,+16223c/+16226a,+16517e	227 309 318T	**G**
DR44	+**12308g,**−15073c,+16223c/+16226a,+16517e	227 309 318T	**G**
U1:			
AD24	−3337k,−**4990a,**+10032a,+**12308g,**−**13103g/** +**13104j,**+**14068l**	129 **189**[3] **249**	**G**
AD33	−**4990a,**+**12308g,**−**13103g/**+**13104j,**+**14068l**	**189**[2] **249** 327	**G**
AD44	−**4990a,**+**12308g,**−**13103g/**+**13104j,**+**14068l**	093 129 186 **189**[3] **249** 365	**G**
DR32	−**4990a,**+**12308g,**−**13103g/**+**13104j,**+**14068l**	**189**[3] **249** 261	**G**
DR43	−**4990a,**+**12308g,**−**13103g/**+**13104j,**+**14068l**	**189**[3] **249** 261	**G**
DR48	−**4990a,**+**12308g,**−**13103g/**+**13104j,**+**14068l**	**189**[3] **249** 261	**G**
U2:			
AD10	+**12308g,**+**15907k,**−**16049k,**+16517e	**051 129C** 189 214 362	**G**
U3:			
AD02	+**12308g,**+16517e	168 192 **343**	**G**
AD03	+**12308g,**+16517e	168 192 **343**	**G**
AD13	+**12308g,**+16517e	168 192 **343**	**G**
AD28	+**12308g,**+16517e	168 192 **343**	**G**
AD30	+**12308g,**+16517e	168 192 **343**	**G**
AD32	+**12308g,**+16517e	168 192 **343**	**G**
AD42	+**12308g,**+16517e	168 192 **343**	**G**
U4:			
AD01	+**4643k,**+7702k,+**11329a,**+**12308g,**+16517e	**356** 362	**G**
U5:			
AD12	−5584a,−6022a,+7569g,+8249b/ −8250e,+**12308g,**−16310k,+16398e	192 256 **270** 311	**G**
AD15	−5584a,−6022a,+8249b/ −8250e,+**12308g,**−16310k,+16398e	192 256 **270** 311	**G**
AD48	−1715c,+**12308g,**+16398e	192 256 **270** 291	**G**
AD43	+**12308g,**+16398e	189 256 **270** 362	**G**

(*continued*)

Table 1 (continued)

Cluster and Sample[a]	RFLP Haplotype[b]	HVS I Haplotype[c]	00073 Status[d]
K:			
AD06	−1923c,−**9052n/**−**9053f,**+**10394c,**+**12308g,**−**16310k**	093 **224 311**	**G**
DR11	−3337k,−**9052n/**−**9053f,**+10309e,+**10394c,**+**12308g,**+15945c, −**16310k,**+16517e	**224 311**	**G**
DR12	−3337k,−**9052n/**−**9053f,**+10309e,+**10394c,**+**12308g,**+15945c, −**16310k,**+16517e	**224 311** *366*	**G**
DR14	−3337k,−**9052n/**−**9053f,**+10309e,+**10394c,**+**12308g,**+15945c, −**16310k,**+16517e	**224 311** *366*	**G**
DR15	−3337k,−**9052n/**−**9053f,**+10309e,+**10394c,**+**12308g,**+15945c, −**16310k,**+16517e	**224 311**	**G**
DR21	−3337k,−**9052n/**−**9053f,**+10309e,+**10394c,**+**12308g,**+15945c, −**16310k,**+16517e	**224 311** *366*	**G**
DR23	−4360g,−**9052n/**−**9053f,**+**10394c,**−10971g,+**12308g,**+15059m, −**16310k,**+16517e	093 **224 311**	**G**
DR28	−**9052n/**−**9053f,**+**10394c,**−11922j,+**12308g,**+16216o, −**16310k,**+16517e	167 216 **224 311** 368	**G**
J:			
DR06	−1715c,+**4216q,**−8150i,+10394c,−**13704t,**−13916g,−**16065g**	**069 126**	**G**
DR13	+**4216q,**−7474a,+10394c,+11001n/+11002f,+11439j,−12170g/ +12171j,−**13704t,**−15254s,−**16065g,**−16310k	**069 126** 311	**G**
AD18	+**4216q,**+10394c,−**13704t,**−**16065g**	**069 126**	**G**
AD41[e]	+**4216q,**+5260b/−5261e,+10394c,−**13704t,**−**16065g,**+16517e	**069 126** 193 256 335	**G**
DR42	−3063j,+**4216q,**−5779a,−7474a,+10394c,−**13704t,**−15254s,−**16065g**	**069 126 145** 189[2] **231 261**	**G**
T:			
AD05	+**4216q,**+**4914r,**+**13366m/**−**13367b/**+**13367j,**+**15606a,** −**15925i,**+16517e	**126** 256 **294** 296 324	**G**
AD08	+**4216q,**+**4914r,**+**13366m/**−**13367b/**+**13367j,**+13710e,+**15606a,** −**15925i,**+16517e	078 **126** 292 **294** 296	**G**
AD20	+**4216q,**+**4914r,**+**13366m/**−**13367b/**+**13367j,**+13710e,+**15606a,** −**15925i,**+16517e	078 **126 294** 296	**G**
AD17	+**4216q,**+**4914r,**−9751l/−9753g,+**13366m/**−**13367b/**+**13367j,** +**15606a,**−**15925i,**+16517e	**126 294**	**G**
AD29	+**4216q,**+4464k,+**4914r,**+**13366m/**−**13367b/**+**13367j,**+**15606a,** −**15925i,**+16517e	**126 294** 296	**G**
AD46	+**4216q,**+**4914r,**−5261e,+**13366m/**−**13367b/**+**13367j,**+**15606a,** −**15925i,**−16303k, +16517e	**126 294** 296 304	**G**
T1:			
AD38	+**4216q,**+**4914r,**−**12629b,**+**13366m/**−**13367b/**+**13367j,**+**15606a,** −**15925i,**+16517e	**126 163 186 189 294**	**G**
DR16	+**4216q,**+**4914r,**−**12629b,**+**13366m/**−**13367b/**+**13367j,**+**15606a,** −15882b/−15883e,−**15925i,**+16517e	**126 163 186 189 294**	**G**
DR49	−3337k,+**4216q,**+**4914r,**−**12629b,**+**13366m/**−**13367b/**+**13367j,** +**15606a,**−**15925i,**+16517e	**126 163 186 189** 192 234 **294**	**G**
X:			
DR03	−1715c,−6383e,+8391b/−8391e,−13704t,+**14465s,**−15925i,+16517e	126 189[3] **223 278**	**G**
DR05	+255k,+5419l,+**14465s,**+16517e	126 189[3] **223 278**	**G**
DR08	+255k,+5419l,+**14465s,**+16517e	126 189[3] **223 278**	**G**
DR09	+255k,+5419l,+**14465s,**+16517e	126 189[3] **223 278**	**G**
DR19	+255k,+5419l,+**14465s,**+16517e	126 189[3] **223 278**	**G**
DR22	+255k,+5419l,+**14465s,**+16517e	126 189[3] **223 278**	**G**
DR24	+255k,+5419l,+**14465s,**+16517e	126 189[3] **223 278**	**G**
DR29	+255k,+5419l,+**14465s,**+16517e	126 189[3] **223 278**	**G**
DR38	−1715c,+**14465s,**+16517e	189 **278**	**G**
DR39	−1715c,+**14465s,**+16517e	189 **278**	**G**
DR40	−1715c,+**14465s,**+16517e	189 **278**	**G**
DR50	−1715c,+10394c,+**14465s,**+16517e	189 **278**	**G**

(*continued*)

Table 1 (continued)

Cluster and Sample[a]	RFLP Haplotype[b]	HVS I Haplotype[c]	00073 Status[d]
I:			
DR04	−1715c,−**4529n,**+**8249b/**−**8250e,**+**10032a,**+10394c,+**16389m/** −**16390b/**+**16390j,**+16517e	**129 223** 320	**G**
W:			
AD50	+**8249b/**−**8250e,**+8944k,−**8994e,**+16517e	**292**	**G**
M:			
DR10	+**10394c,**+**10397a,**+12345k,−16310k,+16517e	129 189 **223** 249 311 359	**G**
C:			
AD19	+**10394c,**+**10397a,**−**13259o/**+**13262a,**+16517e	129 **223 298 327**	**G**
AD45	−1715c,+**10394c,**+**10397a,**−**13259o/**+**13262a,**+16517e	093 129 **223 298 327**	**G**
AD49	−1715c,+**10394c,**+**10397a,**−**13259o/**+**13262a,**+16517e	093 129 **223 298 327**	**G**
L3a*:			
DR25	−1715c,+11436i,+16517e	**223** 265	**G**
DR47	−1715c,+8249b/−8250e,−11362a,−16049k,+16176j,+16389g/ −16390b,+16517e	051 145 176G **223**	**G**
pre-HV*/pre-JT*:			
DR01	+16517e	114 126 362	A
DR41	−5978a/+5980i	093 126 362	A
R1:			
AD04	+4914r,−5584a/−5586c,−5823a,+15493/94c,−16310k,+16517e	278 311	**G**
AD35	+4914r,−5584a/−5586c,−5823a,+15493/94c,−16310k,+16517e	278 311	**G**
AD39	+4914r,−5584a/−5586c,−5823a,+15493/94c,−16310k,+16517e	278 311	**G**
AD16	−4685a,+4914r,−5584a/−5586c,−5823c,+15493/94c, −16310k,+16517e	278 311	**G**
AD47	+4037j,+4914r,−5584a/−5586c,−5823a,−16310k,+16517e	311	**G**

NOTE.—States diagnostic of haplotype clusters are shown in boldface, and those shown in italics were observed to be heteroplasmic.

[a] Cluster membership was determined as described in the text. AD = Adygei, and DR = Druze.

[b] Sites are numbered from the first nucleotide of the recognition sequence. A plus sign (+) indicates the presence of a restriction site, and a minus sign (−) indicates the absence of a restriction site. The explicit indication of the presence/absence of a site implies the absence/presence in haplotypes not so designated. The restriction enzymes used in the analysis are designated by the following single-letter codes: a = *Alu*I; b = *Ava*II; c = *Dde*I; e = *Hae*III; f = *Hha*I; g = *Hinf*I; h = *Hpa*I; i = *Msp*I; j = *Mbo*I; k = *Rsa*I; l = *Taq*I; m = *Bam*HI; n = *Hae*II; o = *Hinc*II; q = *Nla*III; r = *Bfa*I; s = *Acc*I; t = *Bst*OI; and u = *Mse*I. A slash (/) separating states indicates the simultaneous presence or absence of restriction sites that can be correlated with a single-nucleotide substitution.

[c] nps (−16000) between 16051 and 16368 that are different from the CRS (Anderson et al. 1981); "0" denotes that there is no difference from the CRS. Mutations are transitions (T↔C or A↔G), unless the base change is specified explicitly. When 16189C is present, the tract of four adenines (16180–16183) is prone to heteroplasmic length variation (Bendall and Sykes 1995): when other than four, the number of A's in the majority of mtDNA molecules is given within square brackets, after the 16189C designator.

[d] Nucleotide at position 00073 in HVS II. NA = not available.

[e] This sample has an HVS I motif appropriate to J2 (Richards et al. 1998). However, it does not have −7474*Alu*I or −15254*Acc*I, which, on the basis of information from Howell et al. (1995) and Lamminen et al. (1997), would be expected in this cluster. Hence, the 16193C-T mutation is likely to be recurrent, in which case this sample would be classified as a member of J*.

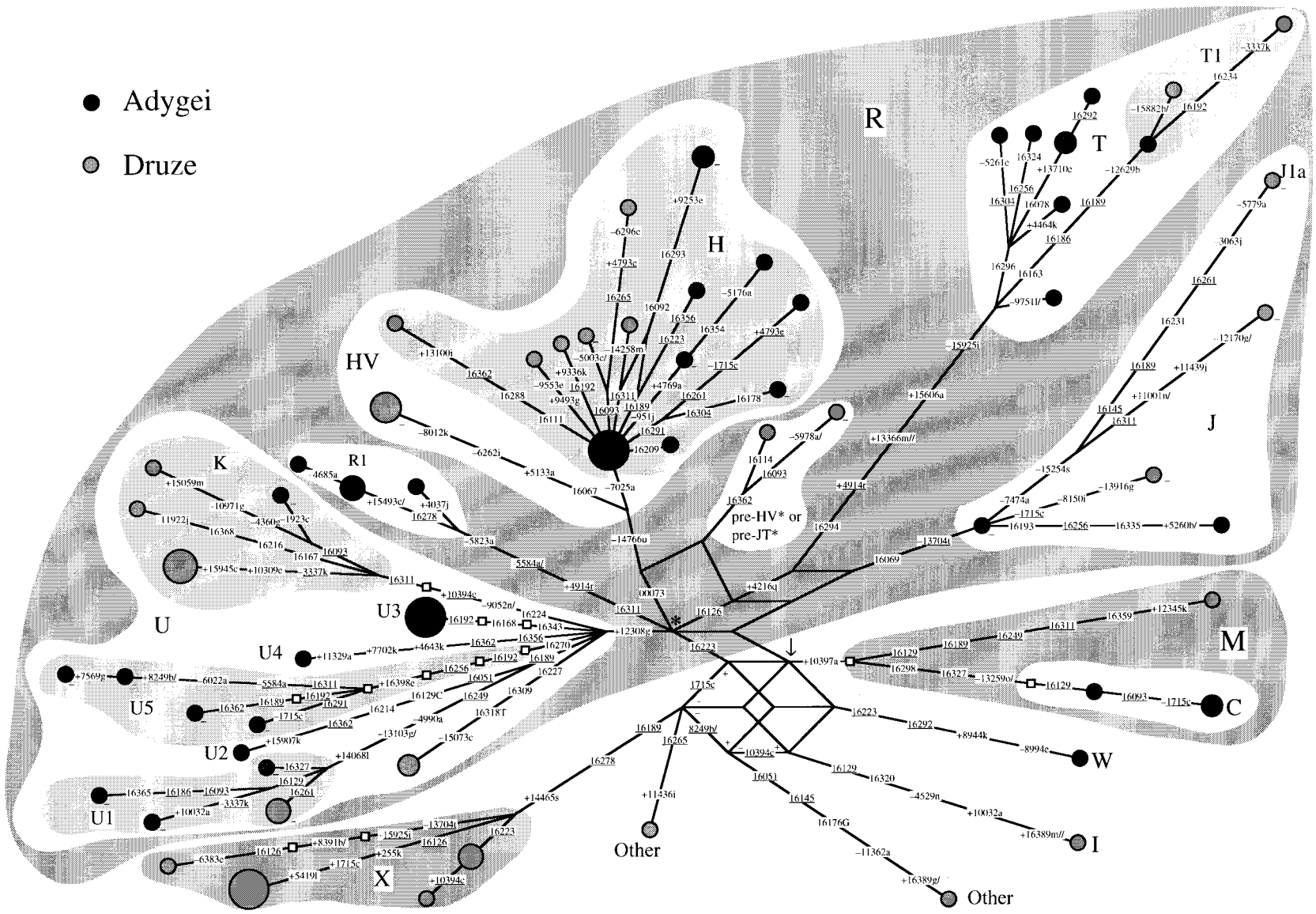
the network and that appears in the Adygei sample. There is little information on the geographic extent of this cluster, since it is impossible to identify on the basis of HVS I alone and since the associated RFLP motif has not been reported in other populations.

6. A singleton Druze haplotype (DR47; denoted as "Other" in fig. 3) that appears to be a member of a small cluster (HVS I motif: 16145–16176G–16223–16390; RFLP motif: +8249*Ava*II/−8250*Hae*III, −11362*Alu*I, +16176*Mbo*I, +16389*Hin*fI/−16390*Ava*II) related to clusters I, W, and X and that has Near Eastern and southern European affinities.

The one remaining unclassified singleton in the Druze (DR25; "Other" in fig. 3) has the ancestral state at np 16223. Networks derived from RFLPs and HVS I separately (not shown) are very much less resolved. The larger number of informative characters in the pooled data set leads to better-differentiated clusters and to better-characterized diversity within the clusters.

Relative-Rate Calibration

To compare studies relying on RFLPs or HVS I only, it is useful to know the relative mutation rates of the two systems. To compare the transition rate of HVS I, between 16090 and 16365 (i.e., the average no. of transitions, per unit time, across this 276-bp stretch of


Adygei
Druze
R
H
HV
K
R1
U
U3
U4
U5
U2
U1
X
T1
T
J1a
J
M
C
W
I
Other
Other
pre-HV* or
pre-JT*

DNA), with the average number of site losses and gains, per unit time, within the extended 14-enzyme system, the network in figure 3 was resolved to a plausible tree, and the number of mutations in each system was counted. (Note that certain HVS I mutations also generate RFLP site gains or losses—e.g., 16227A-G ≡ +16223*Dde*I/+16226*Alu*I—that contribute to both counts. The hypervariable 16517*Hae*III polymorphism was neglected.) There are 95 HVS I changes and 118 RFLP site gains and losses. This implies a most probable relative rate, $\mu_{\text{RFLP}}/\mu_{\text{HVS I}}$, of 1.22, with a 95% credible region of 0.94–1.63. To assess whether the long inner branches of the phylogeny might be leading to a significant loss of mutations at fast sites in HVS I (which are simply not being reconstructed), the analysis was repeated with only mutations within the clusters well represented in this data set, namely, H, pre-HV*/pre-JT*, J, T, K, U1, U5, R1, and X. Then, $\mu_{\text{RFLP}}/\mu_{\text{HVS I}} = 1.14$ (from $n_{\text{HVS I}} = 57$ and $n_{\text{RFLP}} = 67$), with a 95% credible region of 0.82–1.68. Since the credible regions overlap considerably, there is no evidence of the effect on this data. For reference, the relative rate for the original (unextended) 14-enzyme system versus HVS I is 1.14 (from $n_{\text{HVS I}} = 95$ and $n_{\text{RFLP}} = 110$), with a 95% credible region of 0.88–1.53 (cf., a point estimate, by Torroni et al. [1998], of $\mu_{\text{RFLP}}/\mu_{\text{HVS I}} = 1/1.21 \approx 0.83$).

Recurrent Mutation at RFLPs

A number of positions are inferred from the network to have mutated several times. Particularly variable RFLPs include 1715*Dde*I (at least six hits), 10394*Dde*I (at least three hits), and 3337*Rsa*I (three hits). The first two of these markers have been employed as haplogroup diagnostics. However, as discussed below, it is no longer necessary to rely solely on −1715*Dde*I as a marker for cluster X. As for the 10394*Dde*I site—although it probably has occurred once deeply in the phylogeny (e.g., see Chen et al. 1995) and reverted subsequently, distinguishing several major clusters (e.g., J from T, and K from the remainder of U)—it is sufficiently stable within each cluster to be a potentially useful diagnostic in association with other markers. The RFLP 16517*Hae*III is confirmed to be extremely hypervariable, requiring at least 12 hits in an MPR of the network (note, however, that this value could be overestimated, because of potential unresolved recurrent mutations at 16093 and 16189 in H and U1/U2; if we postulate parallelisms at these sites, 16517*Hae*III still has nine hits).

Motifs

To assist in the classification of mtDNAs, we present a table of motifs consisting of mutations—throughout the molecule—that are signatures of the various west Eurasian clusters (table 2). The information was derived from the data sets presented here and also from published data sets, especially when clusters were poorly represented (e.g., C, I, and W). In many cases, either HVS I sequence motifs or diagnostic RFLPs clearly provided enough information to reveal the cluster status of many sequences. However, there are important exceptions. For example, HVS I is not sufficient to dissect H from U: information on 00073, 7025*Alu*I, and/or 12308*Hinf*I is necessary.

For example, note the identification of RFLP markers for HV—namely, −14766*Mse*I, a site loss generated by a T-C transition at position 14766 (Lamminen et al. 1997)—and for T1, namely, −12629*Ava*II. The original 14-enzyme RFLP system led to a characterization of cluster X in terms of the rather variable character 1715*Dde*I, as noted above. However, we observed (among the Druze) individuals whose HVS I motif is manifestly X-like but who are +1715*Dde*I. Fortunately, there exists an RFLP that better defines X—namely, +14465*Acc*I—and, henceforth, we propose to use +14465*Acc*I as an RFLP diagnostic for this cluster.

To set the west Eurasian variation in context, we have summarized, in a schematic genealogy, or coalescent tree (fig. 4), what is known of the global mtDNA phylogeny. The tree is coarse grained in that the taxa are not individuals but groups of individuals comprising clusters.

Figure 3 Network (Bandelt et al. 1995) of the combined Druze and Adygei HVS I, np 00073, and RFLP data sets. Circles represent HVS I/00073/RFLP haplotypes, with their size being proportional to the haplotype frequency in the populations (the smallest indicate singletons, and the largest indicate those with frequency 7); blackened circles indicate Adygei samples, and gray-shaded circles indicate Druze samples. The links in the network indicate mutations: HVS I and RFLP mutations are indicated as described in footnotes b and c intable 1, except that 16,000 was not subtracted from nps in the CR, and single-nucleotide changes that affect more than one enzyme are indicated by the first mutation, followed by one slash (/) if two enzymes are involved or by two slashes for three enzymes; transitions at np 00073 in HVS II are indicated by "00073." Underlined mutations indicate homoplastic events that have been resolved during reduction of the median network. External information (e.g., samples, from other populations, that fill intermediate empty nodes) has been used to refine the reduced median network; these external data points are indicated by unblackened squares. The node with an asterisk (*) has the CRS in HVS I and a guanine at np 00073. The direction of indicated RFLP site losses and gains is from this node outward, toward the tips of the phylogeny; where there is ambiguity about this direction, gains and losses are indicated by a plus sign (+) or a minus sign (−). The phylogenetically uninformative RFLP 16517*Hae*III (Chen et al. 1995) was not used in the construction of the network, but haplotypes that lack this restriction site are indicated by an underscore to the right of the circle. The extent of the clusters is indicated by the shaded backgrounds, and the corresponding cluster names are given in large letters. The inferred root is indicated with an arrow.

Table 2

Motifs Characterizing the Major West Eurasian mtDNA Clusters

Cluster	Subcluster[a]	RFLP Motif[b]	HVS I Motif[c]	00073 Status	Additional HVS II Motif[c]	Additional Coding-Region Motif[c]
H		−7025*Alu*I		A		
		−14766*Mse*I				
V		−4577*Nla*III	16298	A	00072	15904
		−14766*Mse*I				
U		+12308*Hinf*I		G		12372
	K	−9052*Hae*II/−9053*Hha*I	16224			11467
		+10394*Dde*I	16311			
	U1	−4990*Alu*I	16189			
		−13103*Hin*fI/ +13104*Mbo*I	16249			
		+14068*Taq*I				
	U2	+15907*Rsa*I	16051			
			16129C			
	U3		16343			
	U4	+4643*Rsa*I	16356			
		+11329*Alu*I				
	U5		16270			3197
						11467
	U6		16172			
			16219			
J		+4216*Nla*III	16069	G	00295	11251
		+10394*Dde*I	16126			12612
		−13704*Bst*OI				
T		+4216*Nla*III	16126	G		10463
		+4914*Bfa*I	16294			11251
		+13366*Bam*HI				14905
		+15606*Alu*I				
		−15925*Msp*I				
	T1	−12629*Ava*II	16163			
			16186			
			16189			
I		−4529*Hae*II	16129	G	00199	4529T
		+8249*Ava*II/−8250*Hae*III	16223		00204	10238
		+10032*Alu*I	16391		00250	12705
						15043
						15924
W		+8249*Ava*II/−8250*Hae*III	16223	G	00189	12705
		−8994*Hae*III	16292		00204	
					00207	
X		+14465*Acc*I	16223	G		12705
			16278			
M		+10394*Dde*I	16223	G		12705
		+10397*Alu*I				
C		+10394*Dde*I	16223	G		12705
		+10397*Alu*I	16298			
		−13259*Hin*cII/+13262*Alu*I	16327			

NOTE.—For the motifs, we drew on information from this article and elsewhere (Hofmann et al. 1997; Lindholm et al. 1997; Ozawa et al. 1991).

[a] Subclusters have, in addition, the motifs defining the enclosing cluster.

[b] Site losses and gains with respect to the node marked with an asterisk (*) in fig. 3.

[c] Transitions with respect to the CRS, unless the base change is explicit.

Even synthesis of all available data did not provide enough character information to obtain a fully resolved (i.e., bifurcating) tree. The multifurcations are present to indicate the unknown branching order of particular clusters and are not to be interpreted as actual coalescences of clusters occurring in a single generation. Where possible, we have identified the specific nucleotide change responsible for the loss or gain of a restriction site. This was straightforward for site gains but, for site losses, depended on additional sequence information,

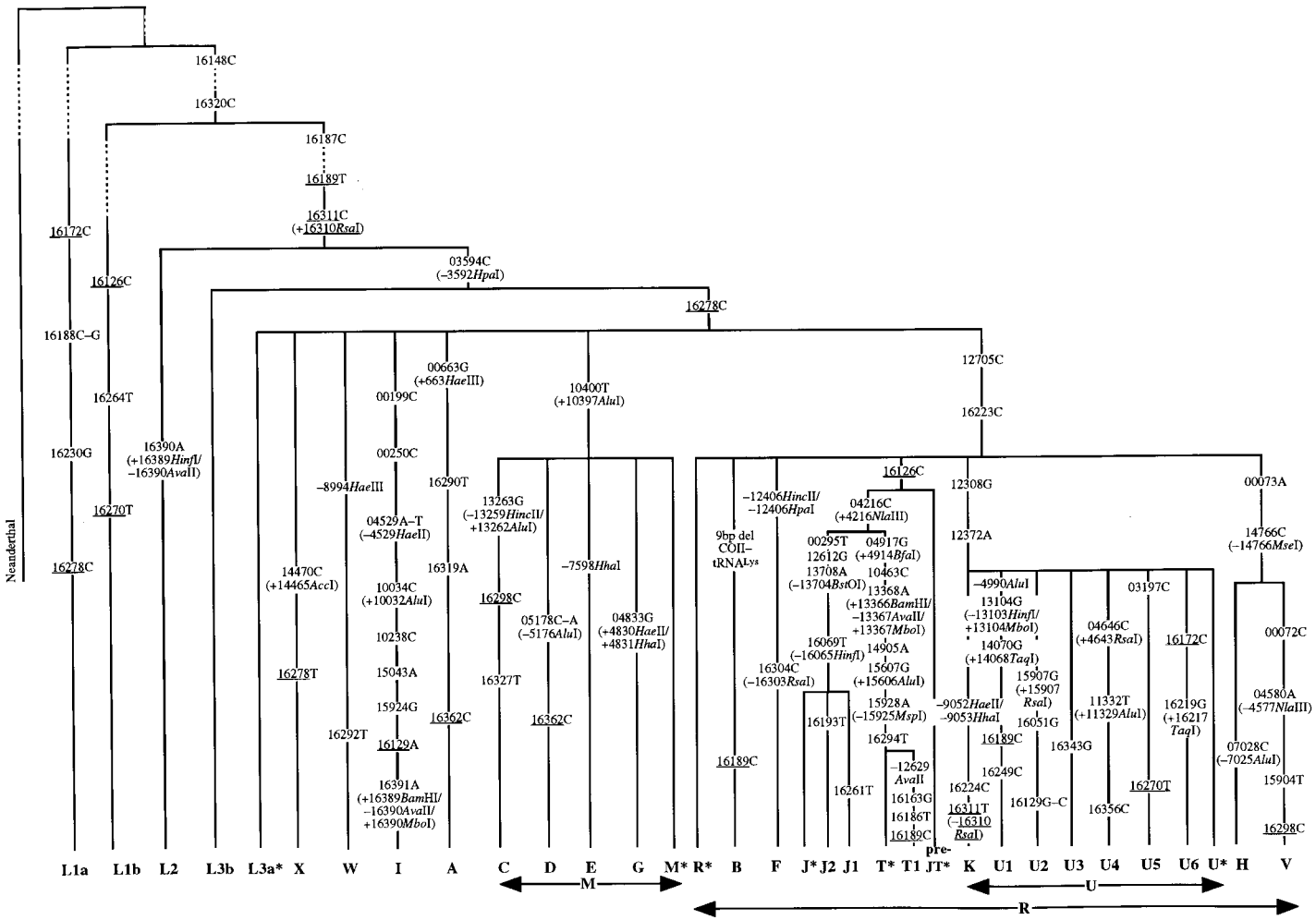


Figure 4 Schematic world mtDNA genealogy, or coalescent tree, rooted with the Neanderthal sequence of Krings et al. (1997). Mutations are indicated by the np involved, followed by either the final base, for transitions, or both the preceding and the resulting bases. Where mutations generate RFLP-site losses or gains in the extended 14-enzyme system, these are indicated by brackets; when the nucleotide responsible was not known, just the RFLP loss or gain is indicated. Recurrent mutations are underlined. The taxa are clusters, the names of which are given at the bottom of the figure. We have used information from published sources (Torroni et al. 1993, 1994*a*, 1994*b*, 1996; Chen et al. 1995; Hofmann et al. 1997; Lamminen et al. 1997; Watson et al. 1997; Howell et al. 1998). Note that the CRS of Anderson et al. (1981) cannot be placed on this diagram: it is known to be a chimera of a number of actual sequences and, as a result, causes a number of incompatibilities with real sequences. For example, although predominantly H-like, it has 14766T (Lamminen et al. 1997).

such as that available in the handful of published complete mtDNA sequences. We have excluded the three RFLPs 1715*Dde*I, 8249*Ava*II/8250*Hae*III, and 10394*Dde*I, since which of these have mutated more than once in the cube at the heart of figure 3is not clear. Other characters were suppressed when there was incomplete information about their status in relevant clusters. For example, some of the established structure of cluster U is not shown, since a site shared by K and U5 (11467G vs. A in U4; Hofmann et al. 1997) has not been assayed in the other U subclusters.

The taxonomy is deliberately west Eurasian heavy, since we have been developing that part here; the classification of Asian and African clusters is still at an early stage. The notation for the African clusters is taken from the report by Watson et al. (1997) and does not comply fully with the scheme described above. For example, L1 is not a clade, since L1a and L1b are separated by the root. The application of our nomenclature to the African clusters requires more information, such as that supplied by the kind of multisystem approach used by us for west Eurasia.

Concordant with a common origin of west and east Eurasians, east Asian clusters A and M are phylogenetically close to west Eurasian clusters I, W, and X, while west Eurasian clusters J, T, U, H, and V are close to east Asian clusters B and F. In contrast, Africans in general branch earlier in the phylogeny, except for members of L3a*. We discuss a striking exception below—namely, cluster U6 of North Africa—but another blurring appears in cluster M, observed in Ethiopians (Passarino et al. 1998). It is tempting to see these patterns as signaling movement from Eurasia back to Africa.

Case Study: Origins of the Berbers

The Berbers, who speak closely related dialects of a distinct branch of the Afro-Asiatic language family, are dispersed in a patchwork throughout a huge area of Cyrenaïca and the Mahgreb, as far west as the Canary Islands, and since antiquity have been widely regarded as aboriginal North Africans (Brett and Fentress 1996). Physically, they resemble other Mediterranean populations, so that an investigation of their origins is of interest not only in its own right but also for the origins of modern west Eurasians (or Caucasians) in general.

Previously, we sequenced HVS I and typed position 00073 of HVS II for 85 Berbers from the villages of the Mzab in northern Algeria, in the context of a study of mitochondrial variation in the Iberian peninsular (Côrte-Real et al. 1996). We identified a novel cluster, referred to as "lineage group 6," comprising one-third of the Berber lineages, which occurred very rarely in other population samples and only in regions known historically to have come under North African influence, such as Iberia (<3% of Romance-speaking Iberians and absent elsewhere in Europe). This cluster therefore appeared to represent the most likely signature of the indigenous North Africans. Most of the sequences in this cluster included the HVS I motif 16172–16189–16219–16278 (with 00073G), but they clustered in a phylogenetic network of the Algerian sequences (fig. 6 in Côrte-Real et al. 1996) with sequence type 16172–16189–16234–16311 (also with 00073G), suggesting that the ancestral sequence was 16172–16189. An alternative ancestral sequence, suggested by the branching pattern of the Portuguese network (fig. 5 in Côrte-Real et al. 1996), which also contained three members of group 6, was 16172–16219, and additional data from the Canary Islands (Pinto et al. 1996) subsequently also suggested that 16172–16219 indeed may be the ancestral sequence type. However, further confusion in the interpretion of the Berber network (which, in fact, should be more highly folded than as shown in fig. 6 in Côrte-Real et al. 1996) is engendered by the conflict of position 16278 with positions 16172 and 16189, so that the African cluster L2 (referred to as "group 3B" in fig. 6 of Côrte-Real et al. 1996) cannot be clearly distinguished from group 6. Therefore (at least in the absence of a suitable weighting scheme), because of the high substitution rate of the CR, it is unclear whether the supposedly indigenous North African cluster evolved from the ancestors of modern west Eurasian or sub-Saharan African lineages. We now have clarified this question by means of RFLP typing.

To determine the cluster status of these Berber sequences, RFLP typing was performed on at least one member of each sequence type in the data set (43 of 85 samples). The samples were typed for a restricted set of RFLPs that were diagnostic of all west Eurasian and African clusters and of some east Eurasian clusters, on the basis of the information intable 1. The following hierarchical scheme was employed:

1. All samples were tested for 14766*Mse*I, 10394*Dde*I, and 7025*Alu*I, and samples lacking the three sites were assigned to cluster H.
2. All non-H samples harboring –14766*Mse*I and –10394*Dde*I were tested for 4577*Nla*III, and those lacking the *Nla*III site were classified as cluster V. All non-H and non-V samples harboring −14766*Mse*I and −10394*Dde*I were classified as HV*.
3. All non-HV samples were tested for 4216*Nla*III, and those with +4216*Nla*III, −10394*Dde*I were assigned to cluster T, whereas those with +4216*Nla*III, +10394*Dde*I were assigned to cluster J.
4. The remaining samples then were tested for 9052*Hae*II and 12308*Hinf*I. Those with +9052*Hae*II, +12308*Hinf*I,−10394*Dde*I were assigned to cluster U

and those with −9052*Hae*II,+12308*Hinf*I to cluster K, irrespective of the status of 10394*Dde*I.

5. The remainder were tested for 3592*Hpa*I, and those with +3592*Hpa*I were assigned to African clusters L1/L2 (Chen et al. 1995).

6. Those lacking the *Hpa*I site were further classified as follows: +10397*Alu*I to cluster M; −1715*Dde*I, +10032*Alu*I to cluster I; −1715*Dde*I,+14465*Acc*I to cluster X; −8994*Hae*III to cluster W; and +2349*Mbo*I, −8616*Mbo*I, and +10084*Taq*I each defining a distinct African subcluster of L3 (Chen et al. 1995; Watson et al. 1997).

7. Finally, cluster assignments were cross-checked against the CR-sequence motifs.

The results are reported in table 3 and are illustrated by the schematic phylogenetic network in figure 5. The "Berber cluster" is a subclade of cluster U, and we therefore refer to it as "cluster U6." Furthermore, a number of sequences clearly were incorrectly classified as parts of group 3A and group 3B in figure 6 and the appendix in the report by Côrte-Real et al. (1996). Haplotypes 77 and 88 (table 3) were assigned to group 3A (defined by HVS I motif 16129–16223–16311 and, hence, equivalent to cluster I) but are, in fact, in cluster M, having undergone distinct mutations at positions 16129 and 16311. Similarly, haplotypes 25, 32, 44, 94, 100, and 102 (table 3) were assigned to group 3B (on the basis of the HVS I motif 16223–16278, which defines cluster X in Europe) but really belong to African clusters L2 and L3b. RFLP typing therefore has allowed us not only to resolve the phylogeny of the Berber cluster, but also to correct clusters that were conflated into paraphyletic clades because of a lack of phylogenetic resolution in the hypervariable CR. Cluster U6 is a sister cluster to several major and minor clusters in Europe and the Near East, including the most ancient cluster in the region, U5, which is specific to Europe and dates to ~50,000 years ago (Richards et al. 1998). Cluster U also is represented in these data by clusters U3 and potential cluster U*.

An advantage of the resolution of the U6 cluster is that an age estimate now can be made by use of the ρ statistic (Forster et al. 1996), although this is extremely provisional because the genealogy of U6 is far from starlike. For all sequences with motif 16172–16219 from the studies by Di Rienzo and Wilson (1991), Côrte-Real et al. (1996), Pinto et al. (1996), and Watson et al. (1997), we estimated the time to the ancestral sequence by using a mutation rate of 1 in 20,180 years (Forster et al. 1996). The value of ρ for these data is 2.53, which converts to an age of 51,000 years, with a central 95% credible region of 42,500–60,500 years. Since this credible region was derived under the assumption of a starlike genealogy, its width surely is underestimated. Nevertheless, it is suggestive that the age estimate is similar to that of U5, the oldest European-specific cluster, and congruent with archaeological dates for the arrival of anatomically modern humans with Upper Paleolithic (Dabban) industry in Cyrenaïca, which is believed to predate 40,000 years ago (Close and Wendorf 1990). This suggests a model in which U5 and U6 diverged from a common ancestor (the Cambridge reference sequence [CRS]) in the Near East (where traces remain of U6; Di Rienzo and Wilson 1991; authors' unpublished data) ~50,000 years ago and spread along the north and south coasts, respectively, of the Mediterranean, as far as Iberia to the north and Cyrenaïca to the south, ~45,000–50,000 years ago. This model is in accordance with the physical-anthropological view that the aboriginal "Mekta-Afalou" North Africans were closely related to the Cro-Magnon settlers of early Upper Paleolithic Europe (Brett and Fentress 1996). The appearance of the Iberomaurusian industry in the Mahgreb may have been the result of a further pulse of westward expansion at least 22,000 years ago (Close and Wendorf 1990), and population replacement with the arrival of the Epi-Paleolithic Capsian industry, ~9,000 years ago, seems unlikely (Brett and Fentress 1996). A greater number of founder lineages arriving in the first wave of settlement from the Near East would reduce the estimate for the time of settlement, but, at present, there is insufficient evidence for multiple founders, especially given the historical record of interaction (and, therefore, possible gene flow) between North Africa and the Near East. However, because of concordance between the genetic and archaeological dates, the early date proposed here is attractive.

Cluster U6 comprises approximately one-third (24/85) of the Mozabite Berber sample. Two individuals formerly characterized as U6, on the basis of the 16172–16189 motif (Côrte-Real et al. 1996), now can be regarded as a separate part of cluster U, since they lack the 16219 transition; two more individuals, unclassified in the RFLP analysis but clearly non-U, fall into an African subcluster of L3a*, on the basis of the CR-sequence motif 16223–16320 (Côrte-Real et al. 1996), in which not only positions 16172 and 16189 but also position 16278 undergo recurrent mutation with respect to the U6 motif, rendering 16219 the only reasonably stable marker for U6. This highlights the difficulties that may arise when the rapidly evolving HVS I sequences are used to trace the ancestry of deeply diverged lineages, in the absence of additional character information.

In contrast with the Mozabites, U6 comprises 3 (17%) of 18 Moroccan Berbers and 8 (15%) of 54 Canary Islanders (Pinto et al. 1996). Given the low frequency of U6 in Iberia, this observation indicates that a large proportion of modern Canary Islander mtDNAs may

Table 3

Berber Samples Assayed for Cluster-Diagnostic RFLP Polymorphisms

RFLP Haplogroup	Cluster	Sample	HVS I Sequence	00073 Status
H	H	3	0	A
H	H	6	0	A
H	H	13	0	A
H	H	18	0	A
H	H	68	0	A
H	H	84	0	A
H	H	22	311	A
H	H	81	311	A
H	H	21	320 325	G[a]
H	H	16	189 304	A
H	H	45	213 356	A
V	V	12	298	A
V	V	30	298	A
Other[b]	V	29	153 298	A
V	V	4	189 298	A
U	U*/U6	2	172 189 234 311	G
U	U3	20	148 343	G
U	U3	79	148 343	G
U	U3	106	148 343	G
U	U6	8	172 189 219 278	G
U	U6	42	172 189 219 278	G
U	U6	80	172 189 219 278	G
U	U6	107[c]	172 189 219 278	G
U	U6	5	172 189 219 239 278	G
U	U6	26	172 189 219 239 278	G
U	U6	69	172 189 219 239 278	G
U	U6	109	172 189 219 239 278	G
U	U6	89	172 189 219 222 278	G
U	U6	76	145 172 219 235 278	G
U	U6	83	172 189 219 239 278 311	G
J	J	24	069 126	G
J	J	17	069 126 147	G
T	T1	9	126 163 186 189 294	G
M	M	77	185 189 223 249 311	G
M	M	88	129 185 189 223 249 311	G
L1/L2	L2	100	223 278 290 294 309	G
L1/L2	L2	94	189 192 223 278 294 309	G
L1/L2	L2	102	189 223 278 292 294 309	G
African +10084*Taq*I	L3b	25	189 223 278 362	G
African +10084*Taq*I	L3b	32	223 278 318 362	G
African +10084*Taq*I	L3b	44	223 278 362	G
African −8616*Mbo*I	L3b	34	124 223	G
Other	L3a*	7	172 189 223 320	G

NOTE.—The HVS I sequence, between 16069 and 16370, and the 00073 status are from the report by Côrte-Real et al. (1996).

[a] This sample appears to have suffered a reversion to 00073G. It is −7025*Alu*I, and a closely related Basque sequence (Côrte-Real et al. 1996) has 00073A.

[b] Despite being +4577*Nla*III, this sample has a clear V motif in HVS I, and, in addition, it has been confirmed as −14766*Mse*I.

[c] This sample has −10394*Dde*I and −14766*Mse*I and, hence, would be classified as HV*. However, since it matched samples 8, 42, and 80 in the CR, it also was tested for the U marker +12308*Hin*fI, which it had. On this basis, it is inferred to have undergone a recurrent loss of 14766*Mse*I and was placed in U6.

have been derived from the Guanches, the autochthonous Berber-related population of the islands, which disappeared following the Spanish conquest.

The remaining Mozabite sequences are typical of either Europe and the Near East (Richards et al. 1998) or sub-Saharan Africa (Watson et al. 1997). The largest cluster of the remaining sequences from the Mzab is cluster H, which is the most frequent cluster in Europe and also is common in the Near East (Torroni et al. 1998), but, unlike those of Europe and the Near East, this cluster is manifestly not starlike in the Mzab, with just three predominant sequence types, suggesting recent

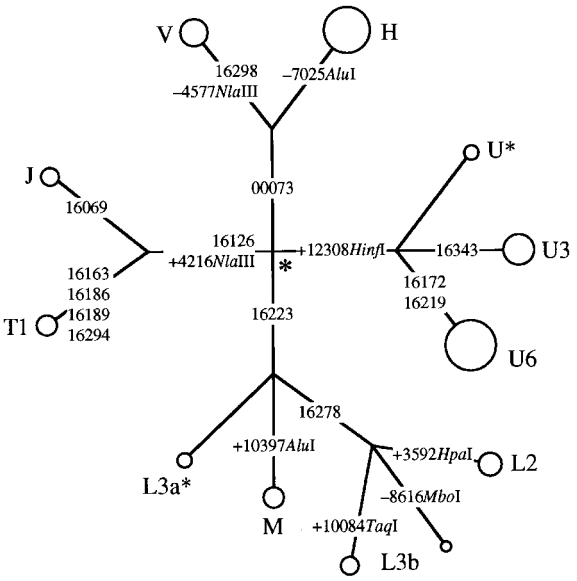


Figure 5 Schematic phylogenetic network of the Berber sample. A selection of diagnostic RFLPs and CR mutations is displayed on the branches. Any diversity within the named clusters is not shown. The sizes of the circles are proportional to the number of samples in the entire Berber data set of Côrte-Real et al. (1996), not just to those that were typed for diagnostic RFLPs in this study.

founder events and/or the action of drift. While two of these sequence types (the CRS and 16311) occur in both Europe and the Near East, one (16213) has been found only in Europe, possibly suggesting a European origin for the H sequences in this population. The cluster derived from the 16213 sequence is only ~10,000 years old in Europe, suggesting that the H sequences may have arrived in North Africa within that time, during the Neolithic period or more recently. A European origin for many of the Berber sequences is also indicated by the presence of cluster V (Torroni et al. 1998). Another common haplotype, 16148–16343, belongs to an additional subcluster of U (U3, see above) and may have been introduced from either Europe or the Near East. Likewise, the 16172–16189–16234–16311 sequence belongs to U, but it may be unrelated to U6 (since 16172 is a hypervariable position) and does not have obvious relatives elsewhere. The cluster T sequence may have been introduced from either Europe or the Near East, but the derived J sequence (16069–16126–16147) perhaps is more likely to have arrived from the Near East, since the 16147 variant so far has been seen previously only in Turkey (Calafell et al. 1996).

Among the 85 Mozabite subjects, there also are five L2 and four L3b sequences—identified on the basis of their HVS I motifs—indicating gene flow from sub-Saharan Africa, and two L3a* sequences, also of sub-Saharan origin (Côrte-Real et al. 1996), indicating a total sub-Saharan component in the Mzab of 14%. This overall picture of high European and little sub-Saharan African input also is reflected in Moroccan and Canary Islander samples (Pinto et al. 1996; J. C. Rando, F. Pinto, A. M. González, M. Hernández, J. M. Larruga, V. M. Cabrera, and H.-J. Bandelt, personal communication).

In summary, one-third of Mozabite Berber mtDNAs have a Near Eastern ancestry, probably having arrived in North Africa ~50,000 years ago, and one-eighth have an origin in sub-Saharan Africa. Europe appears to be the source of many of the remaining sequences, with the rest having arisen either in Europe or in the Near East. Since, in Europe, cluster J appears to have accompanied the Neolithic from the Near East (Richards et al. 1996, 1997, 1998), the J sequences in the Berbers possibly may represent a North African route for the spread of the Neolithic from the Near East. With these exceptions, it is entirely feasible that all the European and Near Eastern sequences present in the northern Berbers arrived from Europe within the last 10,000 years. Iberia and the Mediterranean islands, in particular Bronze Age Sicily and Malta, clearly are implicated in this process (Brett and Fentress 1996). However, since the Nile Valley may have played an important role, particularly in the spread of the Berber language (Brett and Fentress 1996), data from this region will also be necessary before a clear picture can emerge.

Acknowledgments

We thank H.-J. Bandelt for critical advice; A. Cambon-Thomsen, K. K. Kidd, and J. R. Kidd for providing samples; S. Hofmann for discussion; and S. Shanker for her work at the University of Florida DNA Sequencing Core Laboratory. This research received support from The Wellcome Trust (to B.S.) and from the Italian Consiglio Nazionale Della Ricerca (grants 97.04297.CT04 and 98.00524.CT04) and the Ministero Dell'Università e Ricerca Scientifica e Tecnologica. M.R. was supported by a Wellcome Trust bioarchaeology fellowship.

References

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, et al (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457–465

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743–753

Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am J Hum Genet 57:248–256

Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer-Verlag, New York

Bertranpetit J, Calafell F, Comas D, Pérez-Lezaun A, Mateu E (1996) Mitochondrial DNA sequences in Europe: an insight into population history. In: Boyce AJ, Mascie-Taylor CGN (eds) Molecular biology and human diversity. Cambridge University Press, Cambridge, pp 112–129

Brett M, Fentress E (1996) The Berbers. Blackwell, Oxford

Brown MD, Hosseini SH, Torroni A, Bandelt H-J, Allen JC, Schurr TG, Scozzari R, et al (1998) mtDNA haplogroup X: an ancient link between Europe/western Asia and North America? Am J Hm Genet 63:1852–1861

Calafell F, Underhill P, Tolun A, Angelicheva D, Kalaydjieva L (1996) From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. Ann Hum Genet 60:35–49

Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57:133–149

Close AE, Wendoff F (1990) North Africa at 18000 BP. In: Gamble C, Soffer O (eds) The world at 18000 BP. Vol 2: Low latitudes. Unwin Hyman, London, pp 41–57

Côrte-Real HBSM, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, et al (1996) Genetic diversity in the Iberian peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 60:331–350

Di Rienzo A, Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci USA 88:1597–1601

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

Francalacci P, Bertranpetit J, Calafell F, Underhill PA (1996) Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. Am J Phys Anthropol 100:443–460

Hasegawa M, Di Rienzo A, Kocher TD, Wilson AC (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. J Mol Evol 37:347–354

Hofmann S, Jaksch M, Bezold R, Mertens S, Aholt S, Paprotta A, Gerbitz KD (1997) Population genetics and disease susceptibility: characterization of central European haplogroups by mtDNA gene mutations, correlations with D loop variants and association with disease. Hum Mol Genet 6: 1835–1846

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92:532–536

Howell N, Bogolin C, Jamieson R, Marenda DR, Mackey DA (1998) mtDNA mutations that cause optic neuropathy: how do we know? Am J Hum Genet 62:196–202

Howell N, Kubacka I, Halvorson S, Howell B, McCullough DA, Mackey D (1995) Phylogenetic analysis of mitochondrial genomes from Leber hereditary optic neuropathy pedigrees. Genetics 140:285–302

Jeffreys H (1983) The theory of probability. Clarendon Press, Oxford

Kolman C, Sambuughin N, Bermingham E (1996) Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. Genetics 142: 1321–1334

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S (1997) Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30

Lamminen T, Huoponen K, Sistonen P, Juvonen V, Lahermo P, Aula P, Nikoskelainen E, et al (1997) mtDNA haplotype analysis in Finnish families with Leber hereditary optic neuroretinopathy. Eur J Hum Genet 5:271–279

Lindholm E, Cavelier L, Howell WM, Eriksson I, Jalonen P, Adolfsson R, Blackwood DHR, et al (1997) Mitochondrial sequence variants in patients with schizophrenia. Eur J Hum Genet 5:406–412

Ozawa T, Tanaka M, Ino H, Ohno K, Sano T, Wada Y, Yoneda M, et al (1991) Distinct clustering of point mutations in mitochondrial DNA among patients with mitochondrial encephalomyopathies and with Parkinson's disease. Biochem Biophys Res Commun 176:938–946

Passarino G, Semino O, Quintana-Murci L, Excoffier L, Hammer M, Santachiara-Benerecetti AS (1998) Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. Am J Hum Genet 62:420–434

Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM (1996) Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. Ann Hum Genet 60: 321–330

Pult I, Sajantila A, Simanainen J, Georgiev O, Schaffner W, Pääbo S (1994) Mitochondrial DNA sequences from Switzerland reveal striking homogeneity of European populations. Biol Chem Hoppe Seyler 375:837–840

Richards M, Côrte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, et al (1996) Paleolithic and Neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 59:185–203

Richards M, Macaulay V, Sykes B, Pettitt P, Hedges R, Forster P, Bandelt H-J (1997) Reply to Cavalli-Sforza and Minch. Am J Hum Genet 61:251–254

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

Röhl A (1997) Network: a program package for calculating phylogenetic networks. Mathematisches Seminar, University of Hamburg, Hamburg

Stoneking M, Hedgecock D, Higuchi RG, Vigilant L, Erlich HA (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. Am J Hum Genet 48:370–382

Swofford DL, Olsen GJ (1990) Phylogeny reconstruction. In: Hillis DM, Moritz C (eds) Molecular systematics. Sinauer, Sunderland, MA, pp 411–501

Torroni A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, Sellito D, Kengo C, et al (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62: 1137–1152

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli

L, Scozzari R, Obinu D, et al (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144:1835–1850

Torroni A, Lott MT, Cabell MF, Chen YS, Lavergne L, Wallace DC (1994*a*) 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, Miller JA, Moore LG, Zamudio S, Zhuang JG, Droma T, Wallace DC (1994*b*) Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. Am J Phys Anthropol 93:189–199

Torroni A, Petrozzi M, D'Urbano L, Sellitto D, Zeviani M, Carrara F, Carducci C, et al (1997) Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of primary mutations 11778 and 14484. Am J Hum Genet 60: 1107–1121

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, et al (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53:563–590

Wakeley J (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 37:613–623

Watson E, Forster P, Richards M, Bandelt H-J (1997) Mitochondrial footprints of human expansions in Africa. Am J Hum Genet 61:691–704

Wilkinson-Herbots H, Richards M, Forster P, Sykes B (1996) Site 73 in hypervariable region II of the human mitochondrial genome and the origin of European populations. Ann Hum Genet 60:499–508