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 Africanwinoscribed and used to infermenclature f0(are250(u35((usiu35)-24t)*eT*[(2355(tia0t1aaed205(mitohondrial20.32325(Urb0

1998). Hofmann et al. (1997) recently reported coding-region and CR sequences from German subjects, who are very representative of the general west Eurasian

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).

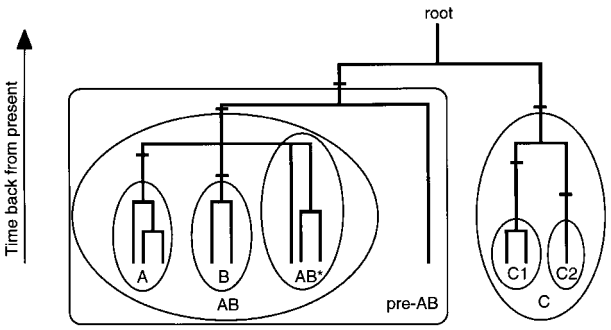


Figure 1 Cartoon mtDNA tree showing the principles of the cladistic notation, for hypothetical clusters A, B, and C.

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

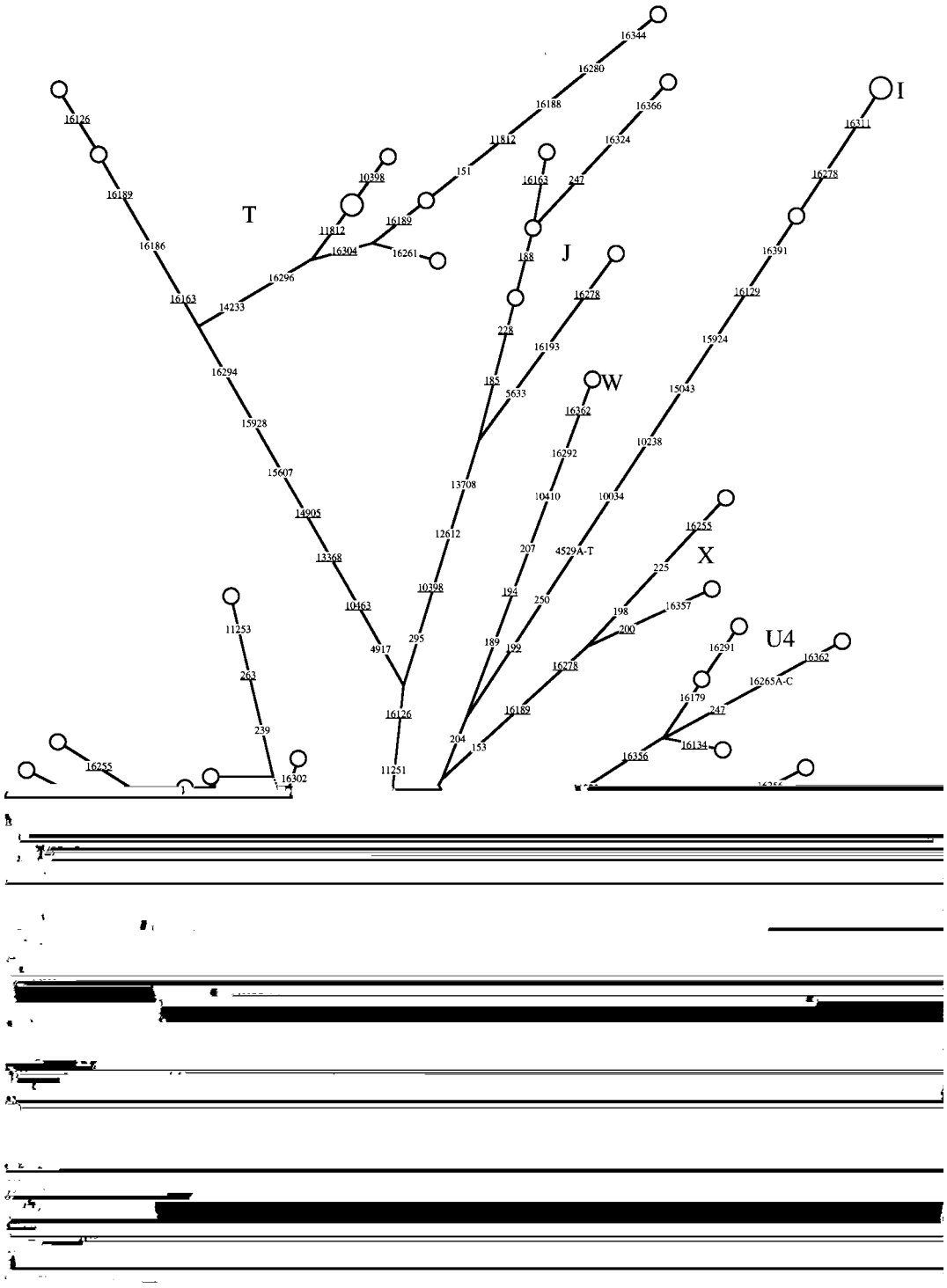


Figure 2 Reduced median network (Bandelt et al. 1995) of the data set of Hofmann et al. (1997). Insertions and deletions were ignored, as was variation at the hypervariable positions 00150, 00152, and 00195 (Torroni et al. 1996) and length variation in the cytosine tracts of HVS II. Circles indicate 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

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

CRodiaor(RFLP)Kaor(Re(in)-380(anncoverislD[(16(RFLP)2T*5380ide55((R4350(np-28o90ide55e2310ide55255e2319(T*5stron09

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	**G16398e**
		AD43 +**12308g** −= ++− − − − + ++	

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)

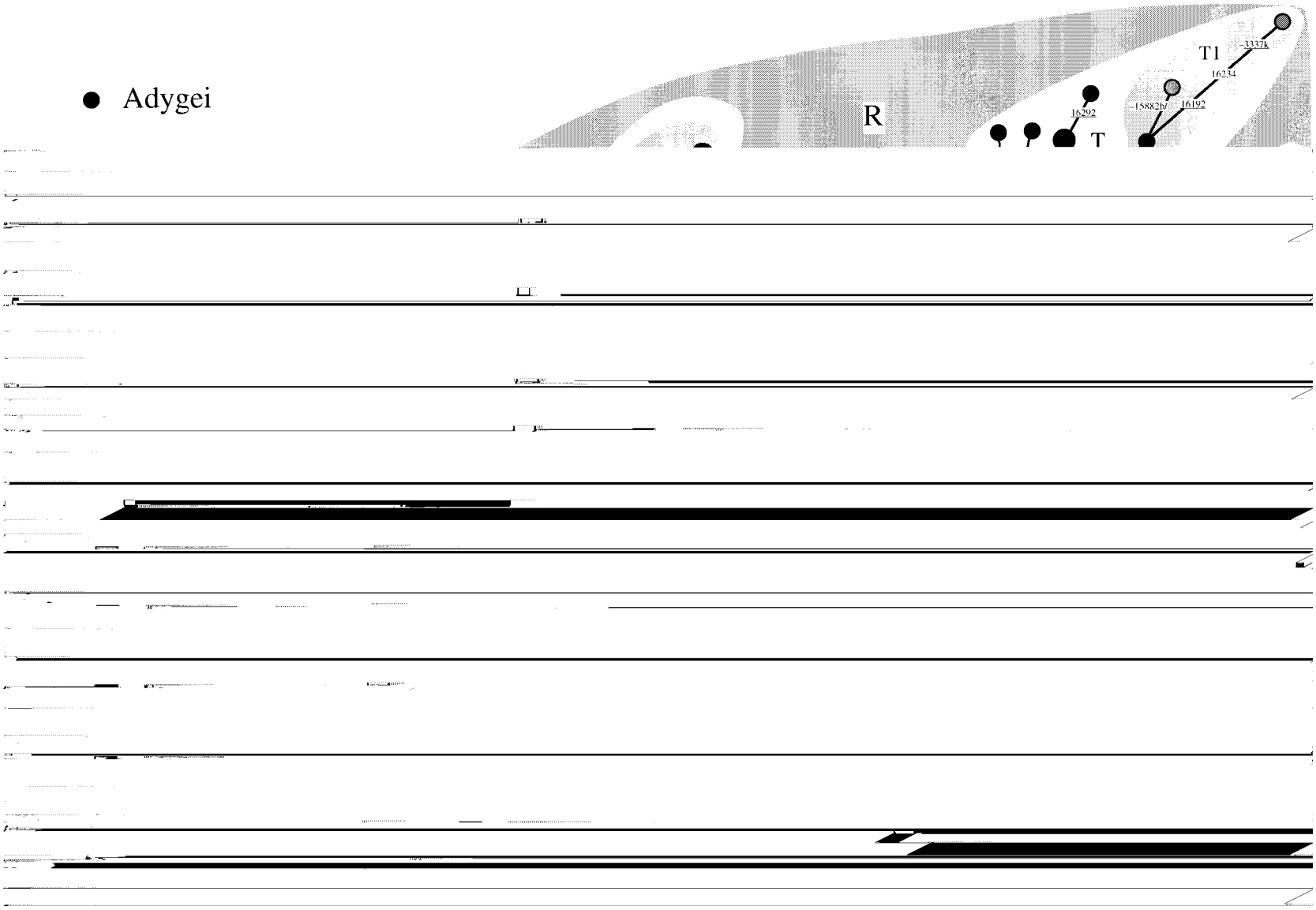
Cluster and Sample[a]	RFLP Haplotype[b]	HVS I Haplotype[c]	00073 Status[d]
I:			
DR04	−1715c,−**4529n,**+**8249b/**−**8250e,**+**10032a**,+10394c,+**16389m/** −**16390b/**+**16390j**,+16517e	**129 223** 320	**G**
W:			
AD50	+**8249b/**−**8250e**,+8944k,−**8994e**,+16517e	**292**	**G**
M:			
DR10	+**10394c,**+**10397a**,+12345k,−16310k,+16517e	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 = *Hin*cII; 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 num-


Adygei
R
16292
T
T1
16234
16192
T

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

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

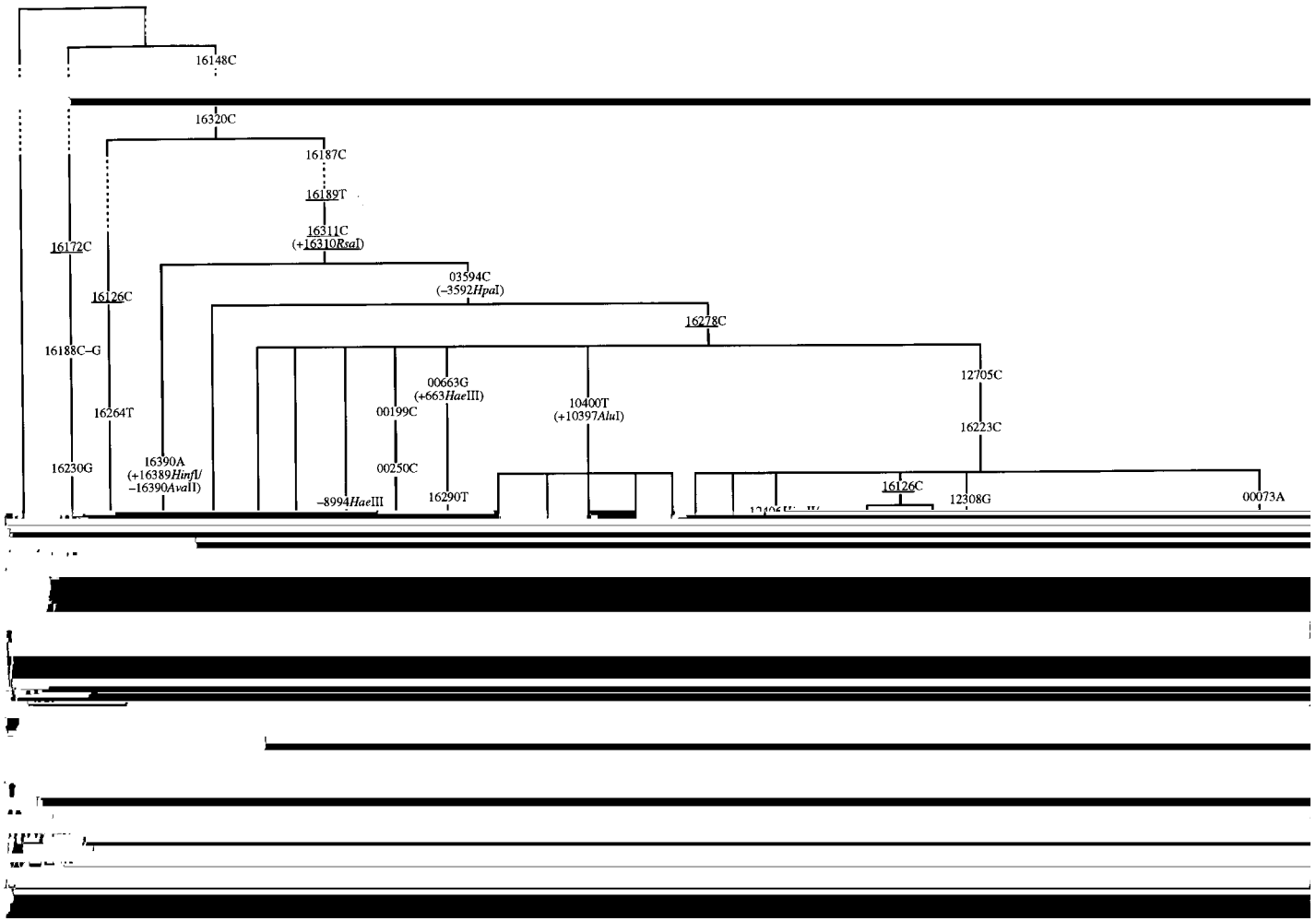


Figure 4 Schematic world mtDNA genealogy, or coalescent tree, rooted with the Neanderthal sequence of Krings et al. (1997). Mutations are indicated by the np involved, followed by either the final base, for transitions, or both the preceding and the resulting bases. Where mutations generate RFLP-site 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), suggesti95(liV0(et)]T7)4.10ag510(1mance-F5

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

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).

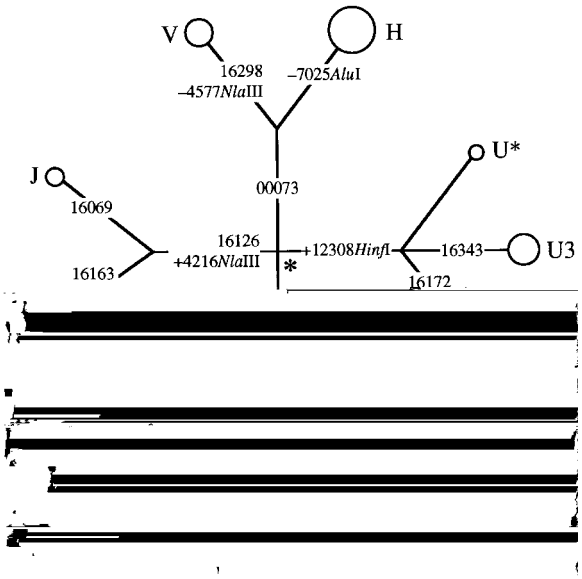


Figure 5 Schematic phylogenetic network of the Berber sample. A selection of diagnostic RFLPs and CR mutations is displayed on the branches. Any diversity within the named clusters is not shown. The sizes of the circles are proportional to the number of samples in the entire Berber data set of Côrte-Real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

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,mf0c55(igrees.)-344(G)1991).22(-6((sequence)chonl)-2.222oJ)-350(Hum
mf0esabrera[(mf0c1.80)-55(igrees.)-344VMDNA JdrialJcatihip A

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