J Forensic Sci, Nov. 2002, Vol. 47, No. 6
Paper ID JFS2001368_476
Available online at: www.astm.org

Marc W. Allard,[1]*Ph.D.; Kevin Miller,*[2] *Ph.D.; Mark Wilson,*[3] *M.S.; Keith Monson,*[3] *Ph.D.; and Bruce Budowle,*[4] *Ph.D.*

Characterization of the Caucasian Haplogroups Present in the SWGDAM Forensic mtDNA Dataset for 1771 Human Control Region Sequences

ABSTRACT: Currently, the Scientific Working Group on DNA Analysis Methods (SWGDAM) mtDNA dataset is used to infer the relative rarity of mtDNA profiles (i.e., haplotypes) obtained from evidence samples and for identification of missing persons. The Caucasian haplogroup patterns in this forensic dataset have been characterized using phylogenetic methods. The assessment reveals that the dataset is relevant and representative of U.S. and European Caucasians. The comparisons carried out were both the observation of variable sites within the control region (CR) and the selection of a subset of these sites, which partition the variation within human mtDNA control region sequences into clusters (i.e., haplogroups). The aligned sequence matrix was analyzed to determine both single nucleotide polymorphisms (SNPs) in a phylogenetic context, as well as to check and standardize haplogroup designations with a focus on determining the characters that define these groups. To evaluate the dataset for forensic utility, the haplogroup identifications and frequencies were compared with those reported from other published studies.

KEYWORDS: forensic science, SWGDAM forensic mtDNA dataset, haplogroup designation, SNPs, validation, polymorphism, human variation, control region

Human mtDNA genetic variation has been studied extensively over the last decade, and general patterns have emerged. Approximately 99% of the known European and U.S. Caucasian mitochondrial variation can be categorized within ten defined haplogroups (H, I, J, K, M, T, U, V, W, and X, see Ref 1). Single nucleotide polymorphisms (SNPs) have been defined previously as having numerous partitions of the major haplogroup clusters observed within populations of European and U.S. Caucasians (see Refs 1–3 and references therein). The frequency and distributions of these Caucasian haplogroups have been the focus of considerable study (3). Currently, a large forensic mtDNA dataset consisting of 1771 Caucasians is used to infer the relative rarity of mtDNA profiles (i.e., haplotypes) obtained from evidence samples and for use in identification of missing persons (4). This dataset is maintained by the Scientific Working Group on DNA Analysis Methods (SWGDAM) to support forensic mtDNA analysis. The comparisons carried out herein are both the observation of variable sites within the control region and the selection of a subset of these sites, which best partition the variation into clusters (i.e., haplogroup designation). Additionally, interest has been generated concerning which sites are the most variable within the mitochondrial control region (5,6). These relative rates of change are assessed using a phylogenetic approach by counting the number of times each variable character changes on a tree.

Subjects and Methods

Subjects

The breakdown of the Caucasian portion of the SWGDAM dataset included samples from Austria ($n = 101$, see Ref 7), France ($n = 109$), Spain ($n = 159$), and the U.S. ($n = 1402$), for a total of 369 Europeans and a grand total of 1771 Caucasian individuals. Details of sampling are provided in Budowle et al. (8).

mtDNA Analysis

A data matrix was constructed by building a multiple alignment of all of the control region nucleotide sequences in SWGDAM. This was analyzed independently to determine both single nucleotide polymorphisms (SNPs) in a phylogenetic context, as well as to check and standardize haplogroup designations with a focus on confirming and determining the characters that define these groups. To validate the dataset, the haplogroup identifications and frequencies were compared with those reported from other published studies. The Caucasian dataset is composed of samples whose complete control region has been sequenced ($n = 155$), or their hypervariable Regions I and II (positions 16024 to 16365 HVI, and positions 73 to 340 HVII, respectively) have been sequenced ($n = 1617$). These sequences were aligned according to a series of rules developed for the consistent placement of gaps (Wilson et al., in press. Forensic Sci Int). To accommodate phylogenetic analysis, N's were inserted wherever missing data were pre-

[1] Department of Biological Sciences, George Washington University, Washington DC.
[2] Federal Bureau of Investigation, DNA Analysis Unit, Washington, DC.
[3] Federal Bureau of Investigation Academy, Counterterrorism and Forensic Science Research Unit, Quantico, VA.
[4] Federal Bureau of Investigation, Washington DC.

Received 24 Oct. 2001; and in revised form 2 May 2002; accepted 3 May 2002; published 2 Oct. 2002.

Copyright © 2002 by ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959.

sent. The final alignment was 1155 bp long and incorporated 4292 humans including 1771 Caucasians, the Cambridge Reference Sequence (CRS, Refs 9,10), and one common chimpanzee used for an outgroup (11). Thirty-five insertions were incorporated into the human mtDNA multiple alignment including: 16184.1, 16184.2, 16187.1, 16190.1, 16192.1, 16193.1, 16193.2, 16538.1, 16538.2, 44.1, 54.1, 64.1, 294.1, 309.1, 309.2, 309.3, 315.1, 315.2, 356.1, 455.1, 498.1, 516.1, 516.2, 524.1, 524.2, 524.3, 524.4, 525.1, 526.1, 573.1, 573.2, 573.3, 573.4, 573.5, 573.6.

Data Availability

All sequences are available at *www.fbi.gov* under the library selection and Forensic Sciences Communications (*www.fbi.gov\library\fsc*).

Phylogenetic Methods

Parsimony analysis was conducted using the software packages Winclada and Nona (Refs 12,13; *www.cladistics.com*) for the Caucasian subset of sequences. For each analysis, at least 1000 replicates of the parsimony ratchet were conducted to determine the simplest explanation of the data (i.e., the most parsimonious solution). For details of search strategies for large datasets ($n > 500$) incorporated by the parsimony ratchet program, see Nixon (12). These analyses include equal weighting of data, treating gaps as missing, N's representing any possible base, one tree held per replicate, and trees built using only unambiguous optimizations (amb =–setting in Nona, see Ref 12).

The general strategy was to independently determine an alignment and, after parsimony analysis, to build a phylogenetic tree. Tree statistics were reported including tree length, consistency index, and retention index, as well as the number of times that each character changed on a tree ($L =$ length). The length values will be differentially affected by sampling with sites that are available for more taxa (HVI and HVII), generally showing greater relative lengths. Important defining characters for haplogroups were based on the tree topology. These characters are the variable nucleotide sequence position and state listed in reference to the CRS (9,10). After independently defining the positions and states, the results were compared to those in the mtDNA literature (1–3, 14). A cladistic nomenclature for human mtDNA haplogroups is adopted and promoted, and every effort is made to compare and contrast our results to studies that have used these methods (see Ref 3 or Ref 14). In the spirit of cladistic analysis, we report when we are able to show a single origin for these haplogroups and when clusters appear to be due to multiple and independent origins.

Single Nucleotide Polymorphisms (SNPs) Ranking

The first level for ranking a SNP site was to determine whether or not the characters were found in two or more individuals with nucleotide base changes different from the outgroup. These are known as parsimony informative sites. These nucleotide positions and states (i.e., SNPs) were read off the tree, and all shared derived sites that defined groups for more than ten individuals were listed secondarily. Finally, characters that appeared as most discriminating were chosen after examination of the tree topology. This final SNP set was selected to identify and remove redundant sites that are closely associated with one another (i.e., SNPs that defined the same group).

The shared derived characters defining nodes on the tree were compared to data in the literature on human mtDNA, control region sequence variation. By comparing the defining characters to other sites listed as important for distinguishing clusters, the haplogroups that were present in this dataset were determined, and additional sites that are important for identifying haplogroups were identified. In this way, the contents of the dataset were validated through independent analysis, with haplogroup identification improved as well. Additional character combinations for haplogroup definition were uncovered. In particular, the resulting patterns of mtDNA variation were compared to those of Torroni et al. (1), who have examined both European Caucasians from the U.S. and from Europe. Other studies of Europeans that were compared to SWGDAM included Finnila et al. (2) and Helgason et al. (3). These studies were selected for comparison because of the thorough and recent treatment of their data.

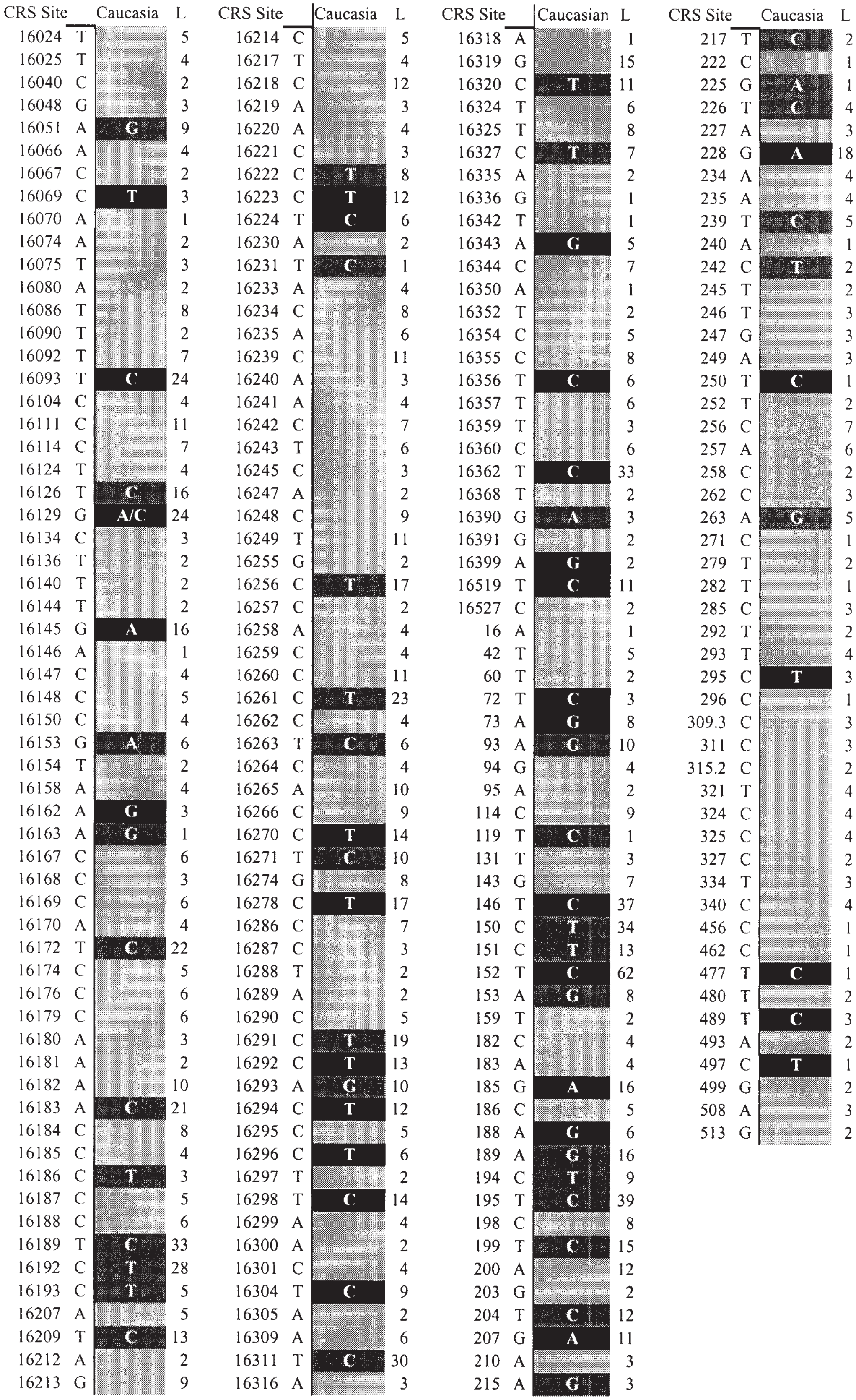
Results and Discussion

Single Nucleotide Polymorphisms (SNPs)

The SNPs were determined from phylogenetic analysis of the Caucasian subdivision of the SWGDAM forensic dataset of human mtDNA, control region sequences. All SNPs observed in two or more individuals were listed ($n = 229$, Fig. 1). There were 28 new sites observed that have not been reported in MITOMAP as human mitochondrial control region sequence polymorphisms (see *www.gen.emory.edu/cgi-bin/MITOMAP/bin/tbl6gen.pl* March 2001 update). Sites sharing a variable character state were parsimony informative and were used for assigning haplogroups. After constructing a phylogeny, characters were optimized on a tree: 212 trees of equal length (1534 steps, $CI = 18$, $RI = 82$) were found. Of the 1771 Caucasians examined, 1278 haplotypes were observed (72%), and, of these, 188 were defined by unique substitutions (i.e., sites observed only in those single individuals).

Shared derived states were used as diagnostic sites for distinguishing clusters of human mtDNA sequences (i.e., haplogroups). To reduce the number of SNPs, a cutoff was made for sites that defined clusters of $n \leq 10$. To improve and refine the SNP choice, a third rating of sites was made, which included sites that appeared best for distinguishing the major groups. This final selection was an attempt to reduce the redundancy among SNPs by eliminating those that were closely associated with others. For example, Haplogroup I was defined by variable sites 16223T, 199C, 204C, and 250C, although only two of these sites (e.g., 16223T and 250C) is needed to recognize this cluster (Fig. 1). Often there were numerous alternatives to choosing a minimal set of SNPs.

SNPs were rated by the degree of informative value, and this rating was done by color-coding CRS sites (see Fig. 1). The light gray defines SNPs shared by two or more individuals in the forensic dataset. The middle gray and black boxes, with states listed, are SNPs that defined clusters of ten or more individuals, and the black boxes refer to one potential subset of the larger set (middle gray and black) where some redundancy of SNPs sites is omitted. This last set (black) was our initial choice of a minimal set of SNPs. Nonetheless, this is not the absolute minimal set, and other investigators may wish to modify this list and/or substitute among the available positions depending on their purposes. All nucleotides exhibited at these sites and the states observed are listed in Fig. 1. All common states that were observed on the tree topology were listed when more than one was seen (e.g., sometimes the extra states present are due to convergent gains or reversal). For the Caucasian dataset, we observed 229 parsimony informative SNPs, 72 SNPs that defined clades of ten or more individuals, and a reduced set of 32 resulting from the removal of some of the redundancy



CRS Site		Caucasia	L	CRS Site		Caucasia	L	CRS Site		Caucasian	L	CRS Site		Caucasia	L
16024	T		5	16214	C		5	16318	A		1	217	T	C	2
16025	T		4	16217	T		4	16319	G		15	222	C		1
16040	C		2	16218	C		12	16320	C	T	11	225	G	A	1
16048	G		3	16219	A		3	16324	T		6	226	T	C	4
16051	A	G	9	16220	A		4	16325	T		8	227	A		3
16066	A		4	16221	C		3	16327	C	T	7	228	G	A	18
16067	C		2	16222	C	T	8	16335	A		2	234	A		4
16069	C	T	3	16223	C	T	12	16336	G		1	235	A		4
16070	A		1	16224	T	C	6	16342	T		1	239	T	C	5
16074	A		2	16230	A		2	16343	A	G	5	240	A		1
16075	T		3	16231	T	C	1	16344	C		7	242	C	T	2
16080	A		2	16233	A		4	16350	A		1	245	T		2
16086	T		8	16234	C		8	16352	T		2	246	T		3
16090	T		2	16235	A		6	16354	C		5	247	G		3
16092	T		7	16239	C		11	16355	C		8	249	A		3
16093	T	C	24	16240	A		3	16356	T	C	6	250	T	C	1
16104	C		4	16241	A		4	16357	T		6	252	T		2
16111	C		11	16242	C		7	16359	T		3	256	C		7
16114	C		7	16243	T		6	16360	C		6	257	A		6
16124	T		4	16245	C		3	16362	T	C	33	258	C		2
16126	T	C	16	16247	A		2	16368	T		2	262	C		3
16129	G	A/C	24	16248	C		9	16390	G	A	3	263	A	G	5
16134	C		3	16249	T		11	16391	G		2	271	C		1
16136	T		2	16255	G		2	16399	A	G	2	279	T		2
16140	T		2	16256	C	T	17	16519	T	C	11	282	T		1
16144	T		2	16257	C		2	16527	C		2	285	C		3
16145	G	A	16	16258	A		4	16	A		1	292	T		2
16146	A		1	16259	C		4	42	T		5	293	T		4
16147	C		4	16260	C		11	60	T		2	295	C	T	3
16148	C		5	16261	C	T	23	72	T	C	3	296	C		1
16150	C		4	16262	C		4	73	A	G	8	309.3	C		3
16153	G	A	6	16263	T	C	6	93	A	G	10	311	C		3
16154	T		2	16264	C		4	94	G		4	315.2	C		2
16158	A		4	16265	A		10	95	A		2	321	T		4
16162	A	G	3	16266	C		9	114	C		9	324	C		4
16163	A	G	1	16270	C	T	14	119	T	C	1	325	C		4
16167	C		6	16271	T	C	10	131	T		3	327	C		2
16168	C		3	16274	G		8	143	G		7	334	T		3
16169	C		6	16278	C	T	17	146	T	C	37	340	C		4
16170	A		4	16286	C		7	150	C	T	34	456	C		1
16172	T	C	22	16287	C		3	151	C	T	13	462	C		1
16174	C		5	16288	T		2	152	T	C	62	477	T	C	1
16176	C		6	16289	A		2	153	A	G	8	480	T		2
16179	C		6	16290	C		5	159	T		2	489	T	C	3
16180	A		3	16291	C	T	19	182	C		4	493	A		2
16181	A		2	16292	C	T	13	183	A		4	497	C	T	1
16182	A		10	16293	A	G	10	185	G	A	16	499	G		2
16183	A	C	21	16294	C	T	12	186	C		5	508	A		3
16184	C		8	16295	C		5	188	A	G	6	513	G		2
16185	C		4	16296	C	T	6	189	A	G	16				
16186	C	T	3	16297	T		2	194	C	T	9				
16187	C		5	16298	T	C	14	195	T	C	39				
16188	C		6	16299	A		4	198	C		8				
16189	T	C	33	16300	A		2	199	T	C	15				
16192	C	T	28	16301	C		4	200	A		12				
16193	C	T	5	16304	T	C	9	203	G		2				
16207	A		5	16305	A		2	204	T	C	12				
16209	T	C	13	16309	A		6	207	G	A	11				
16212	A		2	16311	T	C	30	210	A		3				
16213	G		9	16316	A		3	215	A	G	3				

FIG. 1—*Single nucleotide polymorphisms (SNPs) determined from phylogenetic analysis of the Caucasian mtDNA control region sequences. The numbers refer to the Cambridge Reference Sequence (CRS) numbering system for sites (9,10). Light bars refer to the presence of a SNP in the dataset (variable site found in two or more individuals). Medium gray and black bars are SNPs that define groups with more than ten individuals. Black bars refer to the most informative SNPs based on phylogenetic analysis and a close examination of the evolution of character data on a tree. This entails removal of some of the redundant characters. Character states are listed both for the CRS and for the more common variable sites (medium gray and black bars). All nucleotides that were observed as defining characters are listed. When more than one character is listed this refers to multiple states at a site or the presence of reversals. Length (L) of characters is determined by counting the numbers of character changes occurring on a most parsimonious tree.*

TABLE 1—*List of the most important sites and states that identify major Caucasian haplogroups.*

Haplogroup	Polymorphism 1	Polymorphism 2	Polymorphism 3	Polymorphism 4	Polymorphism 5
H*	73A				
T	16126C	16294T			
J	16069T	16126C	295T		
K	16224C	16311C			
U5*	16270T				
I	16223T	199C	204C	250C	
V	16298C	72C			
W	16223T	189G	195C	204C	207A
M	16223T	16298C			
X	16189C	16223T	16278T	195C	

* The most common Haplogroup, H, is usually defined based on the absence of these other variants and often it is associated with 73A (3). Subcluster U5 is listed as the most common form of the U haplogroup. For a more detailed discussion of additional informative polymorphisms see the text.

from closely associated SNPs. Our reduced list identified all clusters in the dataset with 10 or more individuals.

Hyper-Variable Sites in the Forensic mtDNA Control Region Sequences

Some sites were observed to show more reversals and independent gains than other sites. The most rapidly changing sites in the forensic dataset were determined by the number of times they change on the phylogeny, and these SNPs ranged from having 1 to 62 changes (Site 152) on the tree topology. The average number of changes for a variable character on the tree was 6.7 changes (L = length = 6.7). The hyper-variable sites that change the most, 20 or more times on the tree, include 16093, 16129, 16172, 16183, 16189, 16192, 16261, 16311, 16362, 146, 150, 152, and 195. Sites that are slower to change than these most rapid positions though still are changing from 15 to 19 times on the tree include 16126, 16145, 16256, 16278, 16291, 16319, 185, 189, 199, and 228. See Fig. 1 and Table 2 for a full listing of the number of times each site changes on the phylogeny including the alignment position and nucleotide state observed to change.

Most analyses of human CR variation are fully consistent with our results regarding the SNPs that show the greatest variability (15–17). When the threshold of variability is lowered to ten or more changes on the phylogeny, this forensic dataset shares 19 out of 40 sites with those reported by Stoneking (Ref 17, his Table 1). Other investigators have used parsimony analysis to determine the positions with the highest substitution rates by counting the number of substitutions for each character in a most parsimonious tree (15,18). The most variable sites in this forensic dataset largely overlap with these earlier phylogenetic analyses. When a cutoff of ten or more changes is used to define a fast-changing site, then 18 of 29 sites are shared between this dataset and the analyses of Hasegawa et al. (18) and Wakeley (Ref 15, as listed in Ref 16, their Table 2). If the cutoff is reduced to five or more substitutions as the designated fast sites, then all the sites overlap except for sites 16166 (not observed as polymorphic in the forensic dataset), 16163, 16219, and 16230 (the last three sites are below the cutoff).

An alternative strategy for determining hyper-variable sites is to analyze the data in a pedigree, looking for sites that change between relatives. Meyer et al. (16) summarized the data of Howell et al. (5) and Parsons et al. (6), listing ten hyper-variable sites. Using similar cutoffs as above for fast variable sites, the forensic data overlap with seven or eight of these sites (for cutoffs of ≥10 and ≥5 substitutions, respectively). Only SNP Sites 94 and 234 were not observed as fast sites in our phylogenetic analyses. This may be due to: 1) the private mutations encountered, 2) the differences encountered between a phylogenetic versus a pedigree analysis; 3) the specific sampling that was conducted; and 4) the populations incorporated into these studies. This forensic analysis only includes European and U.S. Caucasian populations; nonetheless, the data largely agree with the reported patterns of hyper-variable positions.

Haplogroup Designation

The same ten haplogroups defined by Torroni et al. (1) were observed in the SWGDAM forensic data. We also surveyed for additional haplogroups that may be present due to introgression and the diversity found among U.S. Caucasians. Similar site variation and frequency pattern were observed when compared to other published populations (Table 1 and Fig. 2; see Ref 1), suggesting that the SWGDAM data are representative of the general European Caucasian population. Haplogroup H is the most commonly observed mtDNA type, occurring at a frequency of 45.7% of the SWGDAM data. This haplogroup is also the most prevalent one observed in Western Europe, and was found in approximately 40% of Caucasians (1). Helgason et al. (3) also found this common haplogroup H, although they observed it in only 24% of a Finnish population (see their Fig. 2). Some pair-wise comparisons of haplogroup frequency distributions were not significantly different between populations, using a Chi-squared test ($p > 0.05$ for 5000 replicates) (SWGDAM versus Swedes), though others were significant (SWGDAM versus Tuscany Caucasians). It should be noted that some European populations are also significantly different from each other (Swedes versus Tuscany Caucasians; for other differences see also Table 3 in Ref 3).

The next most common haplogroups represented in the dataset include clusters U (15.6%), T (10.5%), J (10.0%), and K (8.9%) (see Fig. 2). Previous studies of U.S. Caucasians reported that these haplogroups were relatively common in the Caucasian population (J = 9.1% and K = 7.4%, Ref 1, Ref 19). The less frequently observed haplogroups were W, X, V, I, and M. Each one of these clusters was observed in approximately 1.9% of the samples (range of 1.6 to 2.0%). Similar patterns of haplogroup frequency distribution have been reported for Caucasians, though there are some differences between Caucasians throughout Western Europe. For example, Swedes had a larger percentage of T, K, and U haplogroups and fewer individuals with the J haplogroup (1). Alternatively, Finns cluster more often as J and U haplogroups, and fewer H haplogroup individuals were observed (Fig. 2). Greater prevalence of the J and X haplogroups were reported in Tuscany (1,20). These frequency differences may be due to sampling strategies and phylogenetic structure in the data (21). The J haplogroup frequency varies globally within populations from 0% (Norwegians and Saami) to 21.7% (Cornish, see Ref 21).

The largest component of the Caucasian dataset is individuals from populations within the U.S. (*n* = 1402, 79%). The observed haplogroup frequencies were nearly the same whether the dataset was analyzed with only the U.S. Caucasians, only the European Caucasians, or both combined. When U.S. Caucasians were calculated separately, the most common haplogroup observed was H (44.2%). The haplogroups found at intermediate frequencies were U, T, J, and K, at 16.8, 10.8, 10.0, and 9.3%, respectively; and the remaining haplogroups I, V, W, M, and X, were observed ranging from 0.7% (M) to 2.3% (V) (Fig. 3).

Many of the major clusters were identified based on the presence of a few distinguishing sites, a list of which is presented in Table 1. The analysis indicates that these sites were necessary for defining individuals within these haplogroups. Differences that were found between our analysis of the forensic dataset and the published literature included additional characters with which one can further subdivide haplogroups and/or the addition of characters that may be useful in the primary identification of haplogroups. Some characters published as defining haplogroups were not observed in this forensic dataset. This most likely arose due to the rarity of these particular sites, also known as private mutations (see discussion of 143A for haplogroup W below).

Haplogroup H

Haplogroup H is usually defined based on the presence of site 73A and the lack of other defining polymorphisms from other haplogroups (1), though some repetitive mutations have been reported (e.g., 73G). We also observed numerous reversals of 73A and 73G with eight changes of this site on our tree (Fig. 1). Other investigators (i.e., two show some reversal of site 73) have used a similar approach as ours in their identification of the H haplogroup. Over 800 individuals, approximately 46% (Fig. 2), were observed who fit into this characterization. Finer resolution of H subgroups may be obtained by the presence of the following additional sites (alone or in combination): 16051G, 16093C, 16129A (H4), 16162G (H8), 16189C (H3), 16209C, 16263C, 16278T, 16291T, 16293G, 16304C (H1), 16311C, 16356C (H3), 16362C, 42C, 93G, 146C, 150T, 152C, 195C, 239C, 456T, 477C, and 513A. The numerous small sub-groups within the H haplogroup often are defined by single or paired sequence variants. Whenever this level of variation was observed and matched a published report, these individuals were listed as belonging to the H haplogroup.

Using control region SNPs, the H haplogroup currently is poorly defined and may be a mixture of sub-groups that do not necessarily share a common ancestor. The primary similarity among individuals is based on the absence of defining features. Some of these characters are ancestral states for the control region when using common chimpanzee as outgroup. A single origin for the H haplogroup could not be demonstrated with this large forensic dataset. There were no SNPs that defined this as a single large grouping with shared ancestry (i.e., monophyletic cluster); rather, numerous origins were supported by these data. Other investigators have postulated that this group is a cohesive assemblage possibly because they often include additional data beyond the control region (e.g., inclusion of RFLP data, Ref 1; Ref 2; and Ref 14). Some of the observed variants found in the SWGDAM dataset are most similar to the H1 and H2 subgroups previously defined by Finnila et al. (Ref 2, as well as their unlabeled haplotypes found in their Fig. 2). Other observed clusters are similar to

TABLE 2—*Single nucleotide polymorphisms (SNPs) determined from phylogenetic analysis of the Caucasian mtDNA control region sequences. The numbers refer to the Cambridge Reference Sequence (CRS) numbering system for sites (9,10). Length (L) of characters is determined by counting the numbers of character changes occurring on a most parsimonious tree. Characters are ordered from the fastest changing sites on the tree to the slowest. The length values will be affected by sampling, with sites that are available for more taxa (HVI and HVII) generally showing greater relative lengths as there are more taxa where multiple change could occur.*

L	CRS	L	CRS	L	CRS	L	CRS
62	152	8	153	4	16299	2	16288
39	195	8	198	4	16301	2	16289
37	146	7	16092	4	94	2	16297
34	150	7	16114	4	182	2	16300
33	16189	7	16242	4	183	2	16305
33	16362	7	16286	4	226	2	16335
30	16311	7	16327	4	234	2	16352
28	16192	7	16344	4	235	2	16368
24	16093	7	143	4	293	2	16391
24	16129	7	256	4	321	2	16399
23	16261	6	16153	4	324	2	16527
22	16172	6	16167	4	325	2	60
21	16183	6	16169	4	340	2	95
19	16291	6	16176	3	16048	2	159
18	228	6	16179	3	16069	2	203
17	16256	6	16188	3	16075	2	217
17	16278	6	16224	3	16134	2	242
16	16126	6	16235	3	16162	2	245
16	16145	6	16243	3	16168	2	252
16	185	6	16263	3	16180	2	258
16	189	6	16296	3	16186	2	279
15	16319	6	16309	3	16219	2	292
15	199	6	16324	3	16221	2	315.2
14	16270	6	16356	3	16240	2	327
14	16298	6	16357	3	16245	2	480
13	16209	6	16360	3	16287	2	493
13	16292	6	188	3	16316	2	499
13	151	6	257	3	16359	2	513
12	16218	5	16024	3	16390	1	16070
12	16223	5	16148	3	72	1	16146
12	16294	5	16174	3	131	1	16163
12	200	5	16187	3	210	1	16231
12	204	5	16193	3	215	1	16318
11	16111	5	16207	3	227	1	16336
11	16239	5	16214	3	246	1	16342
11	16249	5	16290	3	247	1	16350
11	16260	5	16295	3	249	1	16
11	16320	5	16343	3	262	1	119
11	16519	5	16354	3	285	1	222
11	207	5	42	3	295	1	225
10	16182	5	186	3	309.3	1	240
10	16265	5	239	3	311	1	250
10	16271	5	263	3	334	1	271
10	16293	4	16025	3	489	1	282
10	93	4	16066	3	508	1	296
9	16051	4	16104	2	16040	1	456
9	16213	4	16124	2	16067	1	462
9	16248	4	16147	2	16074	1	477
9	16266	4	16150	2	16080	1	497
9	16304	4	16158	2	16090		
9	114	4	16170	2	16136		
9	194	4	16185	2	16140		
8	16086	4	16217	2	16144		
8	16184	4	16220	2	16154		
8	16222	4	16233	2	16181		
8	16234	4	16241	2	16212		
8	16274	4	16258	2	16230		
8	16325	4	16259	2	16247		
8	16355	4	16262	2	16255		
8	73	4	16264	2	16257		

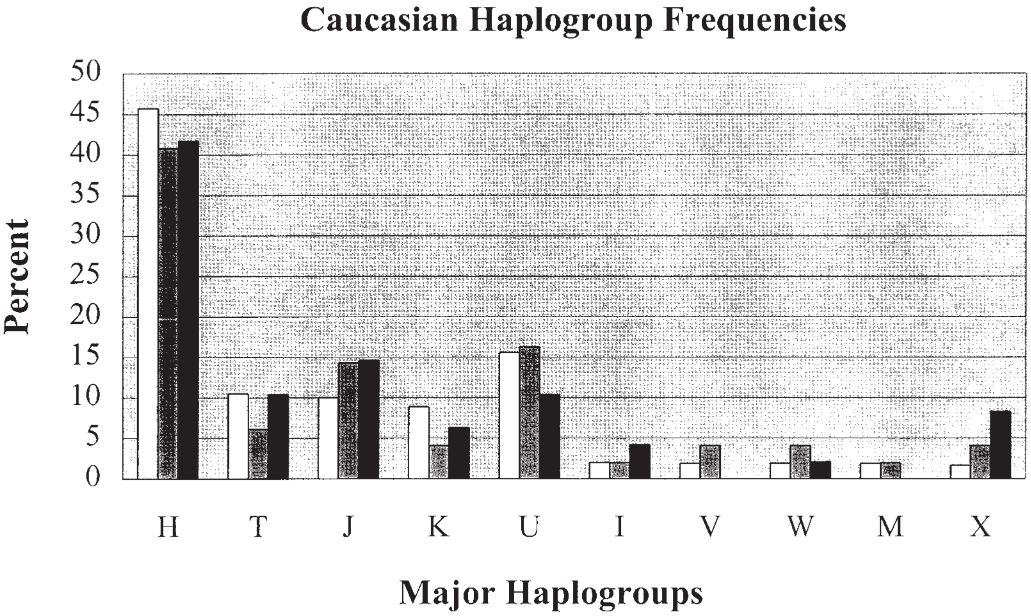


FIG. 2—*Major haplogroup frequencies observed in the SWGDAM Caucasian dataset as well as published Caucasian populations as defined by HVI and HVII. Letters refer to the major haplogroups as defined in the text for three populations: SWGDAM forensic dataset combined (white, n = 1771), Finland (gray, n = 49) and Tuscany (black, n = 48). The data from Finland and Tuscany are described by Torroni et al. (their Table 4, Ref 1) and Francalacci et al. (20).*

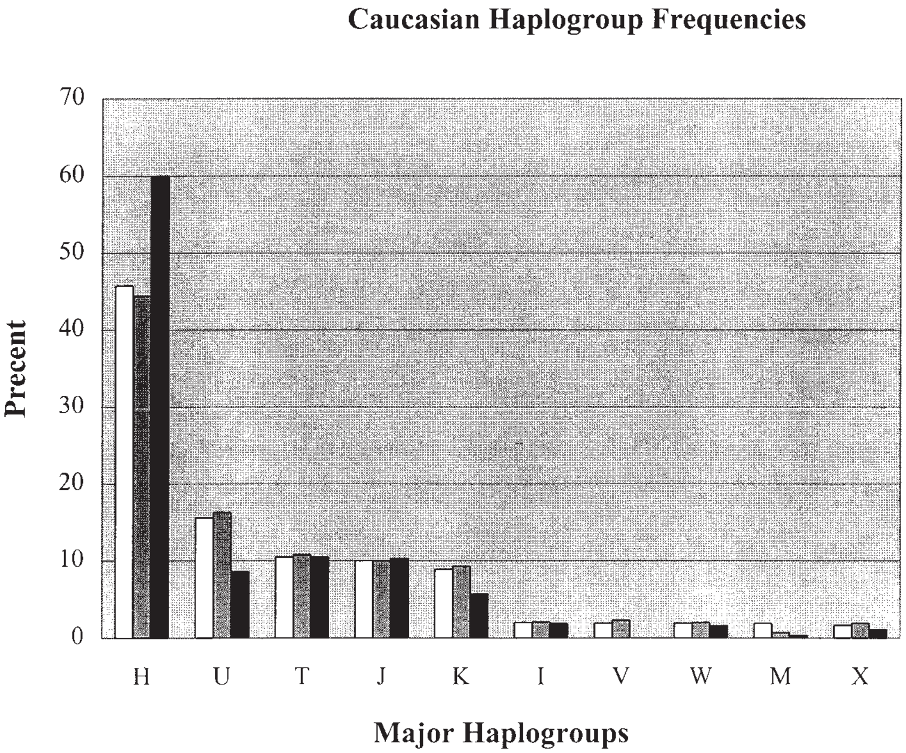


FIG. 3—*Major haplogroup frequencies observed among subdivisions of the SWGDAM Caucasian dataset. Letters refer to the major haplogroups as defined in the text for three subdivisions: SWGDAM forensic dataset combined (white, n = 1771), U.S. subset of SWGDAM (gray, n = 1402) and Europe subset of SWGDAM (black, n = 369).*

those defined by Helgason et al. (3) as H1, H3, H4, and H8 subgroups (see Ref 3's Fig. 2). Some of these SWGDAM sequences are associated closely with the V Haplogroup (see description below). A close association between H and V haplogroups is well documented, indicating that these two groups may share a common ancestry (Ref 2; Ref 3). The forensic dataset is consistent with these findings. Of all of the haplogroups, H is most in need of additional clarification and subdivision.

Haplogroup U

Torroni et al. (1) defined Haplogroup U5 by the presence of sites 16192T and 16270T. Reversals of 16192T and 16192C were seen in our analysis (and also reported independently in Ref 2, $L = 28$ for one of our most variable sites). Using 16270T as the primary designator for this haplogroup is recommended. There were 134 individuals (7.6%) observed carrying this state. This haplogroup can be further subdivided based on the presence of 16256T and 150T. This subdivision has been described previously as U5a and U5b haplogroups, respectively, by Finnila et al. (2). Haplogroup U5a is further sub-divided by 16399G and 16291T. The U5b haplogroup is subdivided by the presence of 16189C or 16311C. Other investigators (Ref 2; Ref 3; and Ref 14) suggest additional division of the U group (U1, U2, U3, U4, U6, U7, and U8) that does not include the 16270T variation. Our analysis identified one cluster that was similar to Finnila et al.'s (2) U2 haplogroup, containing states 16051G, 16129C, 16183C, 16189C, 16362C, 152C, 217C, 340T, and 508G ($n = 15$, 0.85%). The first two sites also are recommended by Helgason et al. (3) for the recognition of this subcluster. This subgroup would not have been considered a U haplogroup by Torroni et al. (1) as they originally recognized 16192T and 16270T as the defining variant for this cluster. Torroni et al.'s (1) earlier description matches Finnila et al.'s (2) U5a and U5b clusters but not the other U haplogroups. Also, we found that less than 1% of the individuals in the dataset ($n = 17$) are defined by states 16343G and 150T. This pattern is consistent with the U3 haplogroup defined by Macaulay et al. (14). Two other sub-clusters are consistent with the U4 Haplogroup of Finnila et al. (2) based on the presence of states 499A ($n = 59$), 16356C, and 195C ($n = 34$). We observed these sites as separate though related sub-clusters, and thus we have designated these as U4a and U4b. We observed six Caucasians that fit the U1 criteria of 16249C and 16189C, and five Caucasians that fit the U6 designation of 16219G and 16172C (3), although both of these frequencies fell below our cutoff for listing these sites as informative at an intermediate level. Similar to U1 and U6, we observed only six Caucasians who fit the U8 criteria of Finnila et al. (2), 16342C, and 282C; and we did not observe any U7 Caucasians.

The phylogenetic analysis did not uncover a single SNP that defines all of the U sub-clusters; thus, we found no evidence for shared ancestry among these variants. Others have proposed that the U haplogroup is a monophyletic cluster based on the combination of both RFLP and sequence analysis, similar to the situation observed with the H haplogroup. We have combined all of these U sub-clusters in defining the U haplogroup frequency ($n = 134 + 15 + 17 + 59 + 34 + 5 + 6 + 6 = 276$, 15.6%, Fig. 2).

Haplogroup T

Haplogroup T was defined by Torroni et al. (Ref 1, their Table 7) by the presence of 16294T and often including 16296T. In these samples, 16294T was found in 186 individuals (10.5%). Another site that is associated closely with 16294T is 16126C (also seen by Torroni et al. (1), and thus 16126C should be included when designating the T haplogroup. Haplogroup T is further subdivided by the presence of 16304C and 16296T. However, both the 16304 and 16296 variants show independent reversals throughout this haplogroup. This additional variation is similar to the T1 and T2 clusters described by Finnila et al. (2). Other sites refining the haplogroup into subclasses in the SWGDAM data include 16153A, 16163G, 16186T, 16189C, 150T, 151T, 152C, and 195C.

Haplogroup J

Haplogroup J was observed in 178 individuals (10%) in the dataset based on the presence of three variable positions (16069T, 16126C, and 295T). This haplogroup may be subdivided into finer subgroups such as the J1 and J2 types of Finnila et al. (2), although their current designations show these subgroups as closely associated and intermixed. Positions and states in the SWGDAM data subdividing the J haplogroup included 16145A, 16172C, 16193T, 16222T, 16231C, 16261T, 16278T, 150T, 152C, 185A, 188G, 195C, 215G, 228A, 242T, and 462T.

Haplogroup K

Haplogroup K was one of the medium-sized groups observed in this forensic Caucasian dataset ($n = 157$, 8.9%). Torroni et al. (1) identified this haplogroup by the character and states 16224C and 16311C, and both of these states were observed in our analysis. Additional states that subdivided this haplogroup included 16093C (K2 observed with multiple independent gains of this state), 16320T (K1 of Ref 3, $n = 10$), 114T, 146C, 195C, and 497T. Additionally, the 152C variant was found in many members of this haplogroup; however, the data suggested numerous independent reversals at this site. Haplogroup sub-cluster K2a (3) was further defined by 16291T ($n = 1$), and K2b was defined by 16319A ($n = 6$).

Haplogroup I

Haplogroup I is a small (2%) cluster that was defined by numerous sites. Torroni et al. (1) defined this group based on nine variable positions (16129A, 16223T, 16311C, 152C, 189G, 199C, 203A, 204C, and 250C). Of these variants, 189G should be excluded as a marker. In our analysis, variant 189G is not associated with Haplogroup I. This observation was also supported by the data of Finnila et al. (2). Position 16129A exhibits the same state as chimpanzee and thus is the ancestral state at this site in our analysis. Four of the above states (16223T, 199C, 204C, and 250C) were designators for Haplogroup I. Additional variants that subdivided this group in the SWGDAM data included 16311C, 152C, and 207A. Variants 16172C and 203A (listed in Ref 2) also were observed; however, these sites fell below our threshold ($n = 10$) for listing these as defining positions ($n = 8$, Fig. 1, light rather than medium gray bars).

Haplogroup V

This haplogroup was defined by Helgason et al. (3) by the presence of 16298C and 72C. In addition, Finnila et al. (2) listed variation at positions 16183 and 485. There were 34 individuals (1.9%) observed with states 16298C and a subset of these with 72C ($n = 17$). The additional variable positions described by Finnila et al. (2) were not observed.

Haplogroup W

Haplogroup W was observed in 34 individuals in the forensic dataset. This relatively small group (1.9%) was defined by a large number of variable positions. Torroni et al. (1) listed seven defining changes relative to the CRS (16223T, 73G, 143A, 189G, 195C, 204C, and 207A). All of these states were observed; however, no justification for the use of two of these to define the W haplogroup could be supported with our data. The character 73G is not specific to this group and should be omitted. According to our analysis, site 143A is independently gained and also is a rare substitution. Additional character states that sub-divide haplogroup W include 16292T, 119C, and 194T.

Haplogroup X

Haplogroup X is another small group with only 29 individuals in the SWGDAM dataset (1.6%). This haplogroup has been defined primarily by states 16223T and 16278T (1). In addition to these characters states, 16189C and 195C are recommended. Finer subdivision of the group is gained with the associated sites 225A, 226C, 153G, and 16183C.

Haplogroup M

Haplogroup M is a small cluster (1.9%) that is often associated with the Z, D, and C haplogroups (3). In the Caucasian forensic dataset, this cluster shares features with Finnila et al.'s (2) Z cluster including the presence of 16298C, 16223T, and 152C. The additional sites 16260, 16224, 16185, 16129, 151, and 489 described in the Finnish data were not encountered in the forensic dataset. The 489C variant is observed in all of these samples, but this character is not specific for the M haplogroup for these analyses. Helgason et al. (3) list 16223T and 489C for the M haplogroup. Helgason et al. (3) list additional states to further resolve the sub-clusters of M (i.e., their C, D, and Z clusters). There were eleven individuals (0.6%) observed that fit these designations, and these have been listed as part of the larger M haplogroup. However, based on the presence of 16223T, 16298C, and 16327T, these would be interpreted as belonging to the C sub-cluster of haplogroup M. Additionally, 22 individuals are only distinguished from CRS by 16223T. According to Helgason et al. (3), these could belong to a large cluster that separates R haplogroups from everything to the right of their Fig. 2 (i.e., a cluster that includes A, X, W, I, L, and M). These Caucasians were closely related to the C sub-cluster on our tree, and thus we have listed these in our M haplogroup (Fig. 2, $11 + 22 = 33$, 1.9%).

Conclusions

The analyses of the Caucasian SWGDAM forensic dataset ($n = 1771$) demonstrate that these CR sequences are representative of European and U.S. Caucasian populations. The ten major haplogroups described in Caucasian populations were observed in this dataset at similar frequencies as have been reported in the literature (H, I, J, K, M, T, U, V, W, and X). The most common haplogroup, H, was observed for approximately 46% of the Caucasians. Haplogroups observed at intermediate levels included clusters U (15.6%), T (10.5%), J (10.0%), and K (8.9%). The haplogroups observed less frequently include W, X, V, I, and M, and these were observed in approximately 2% or less of the sample.

An alignment of 1155 bp was generated through the incorporation of 35 insertions. Our analyses also found 229 SNPs that were shared by two or more Caucasians, and these were further subdivided by their ability to sort out the genetic variation in a phylogenetic context. Of these 229 sites, 28 have not been reported as variable SNPs in MITOMAP. Many of the 229 identified SNPs correspond to important variable sites that define haplogroups. Variability of sites was measured by the numbers of times a site changed on the phylogenetic tree, and these SNPs ranged from having 1 to 62 changes (Site 152) on the tree topology. The average number of changes for a variable character on the tree was 6.7. The most variable sites observed included 16093, 16129, 16172, 16183, 16189, 16192, 16261, 16311, 16362, 146, 150, 152 and 195 (with 20 or more changes each). These rapidly changing sites are consistent with other published analyses supporting the phylogenetic methodology on which these identifications were based. Sites were chosen as well for their ability to discriminate among these Caucasians, and a ranking of sites is provided. There are 72 SNPs listed that could distinguish clusters containing ten or more Caucasians and a reduced set of 32 SNPs for the same clusters, and all of the sites that varied were ranked based on the number of times they changed on a phylogenetic tree. The general pattern of CR genetic variation is consistent with the literature. These data define the best SNPs for designating haplogroups and for use in subdividing the more common clusters. Our detailed analysis of the large forensic dataset revealed the most important mtDNA CR SNPs useful for discriminating among the haplogroups of Caucasians. Our list of observed SNPs and the ranking of these sites (Fig. 1) provide directions for future expansion of alternative systems to sequencing and for the characterization of human mtDNA. These forensic SNPs could be used to design SNP-based assays for forensic identification purposes.

Our SNP list generated from Caucasians from Europe and the U.S. should be able to discriminate well for this target population. The reduced list identifies all of the members of the major European Caucasian haplogroups (Fig. 1, black bars). A future goal of ours is to provide additional SNP lists for the other subdivisions of SWGDAM. Other large datasets within this database that are slated for future examination include Asians, African Americans, Native Americans, and Hispanics.

Acknowledgments

We thank Dr. D. Foran and two anonymous reviewers for their careful review of the manuscript. This research was partly funded by the Visiting Scientists Program at the FBI Academy, Forensic Science Research Unit, through a grant to Marc W. Allard.

This is publication 01-27 of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only, and inclusion does not imply endorsement by the Federal Bureau of Investigation.

References

1. Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, et al. Classification of European mtDNAs from an analysis of three European populations. Genetics 1996;114:1835–50.
2. Finnila S, Lehtonen MS, Majamaa K. Phylogenetic Network for European mtDNA. Am J Hum Genet 2001;68:1475–84.
3. Helgason A, Hickey E, Goodacre S, Bosnes V, Stefansson K, Ward R, et al. mtDNA and the Islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am J Hum Genet 2001;68:723–37.
4. Miller K, Budowle B. A compendium of human mitochondrial DNA control region: development of an international standard forensic database. Croatian Medical J 2001;42:315–27.
5. Howell N, Kubacka I, Mackey DA. How rapidly does the human genome evolve? Amer J Hum Genet 1996;59:501–9.
6. Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, et al. A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet 1997;15:363–7.

7. Parsons W, Parsons TJ, Scheithauer R, Holland MM. Population data for 101 Austrian Caucasian mitochondrial DNA D-loop sequences: amplification of mtDNA sequence analysis for a forensic case. Int J Legal Med 1998;111:124–32.
8. Budowle B, Wilson MW, DiZinno JA, Staffer Z, Fasano MA, Holland MM, et al. Mitochondrial DNA regions HVI and HVII population data. Forensic Sci Intl 1999;103:23–35.
9. Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature 1981;290:457–65.
10. Andrews R, Kubacka I, Chinnery P, Lightowlers R, Turnbull D, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genet 1999;23:147.
11. Foran D, Hixson JE, BrownWM. Comparisons of ape and human: sequences that regulate mitochondrial DNA transcription and D-loop DNA synthesis. Nuc Acids Res 1988;17:5841–61.
12. Nixon K. The parsimony ratchet, a new method for rapid parsimony analysis Cladistics 1999;15:407–14.
13. Goloboff P. Nona: A tree search program. Program and documentation that is available from ftp.unt.edu.ar/pub/parsimony and www.cladistics.org., 1994.
14. Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, et al. The emerging tree of West Eurasian mtDNAs: a synthesis of control region sequences and RFLPs. Am J Hum Genet 1999;64:232–49.
15. Wakeley J. Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. J Mol Evol 1993;37:613–23.
16. Meyer S, Weiss G, von Haeseler A. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of Human mtDNA. Genetics 1999;152:1103–10.
17. Stoneking M. Hypervariable sites in the mtDNA control region are mutational hot spots. Am J Hum Genet 2000;67:1029–32.
18. Hasegawa M, Rienzo AD, Kocher T, Wilson AC. Toward a more accurate time scale for the human mitochondrial DNA tree. J Mol Evol 1993;37:347–54.
19. Torroni A, Lott M, Cabell MF, Chen Y-S, Lavergne L, Wallace DC. MtDNA and the origin of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am J Hum Genet 1994;55:760–76.
20. Francalacci P, Bertranpetit J, Calafell F, Underhill P. Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. Am J Phys Anthropol 1996;100: 443–60.
21. Torroni A, Richards M, Macaulay V, Forster P, Villems R, Norby S, et al. mtDNA Haplogroups and frequency patterns in Europe. Am J Hum Genet 2000;66:1173–7.

Additional information and reprint requests:
Mark Wilson, M.S.
Federal Bureau of Investigation Academy
Counterterrorism and Forensic Science Research Unit
Quantico, VA 22135
Telephone: (703) 632-4542
E-mail: mwilson@fbiacademy.edu