Journal: www.jogg.info
Originally Published: Volume 4, Number 1 (Spring 2008)
Reference Number: 41.004

THE SUBCLADES OF MTDNA HAPLOGROUP J AND PROPOSED MOTIFS FOR ASSIGNING CONTROL-REGION SEQUENCES INTO THESE CLADES

Author(s): Jim Logan

The Subclades of mtDNA Haplogroup J and Proposed Motifs for Assigning Control-Region Sequences into These Clades

Jim Logan

Abstract

This paper presents a study of the phylogeny of mtDNA Haplogroup J using full genome sequence data publicly available through GenBank. It presents a broad history of previous research relative to this haplogroup and the development of motifs for classification of its clades. It then presents a new phylogeny and a set of new motifs for classification where only control region data is available. Finally, it evaluates these motifs relative to full sequence classification and uses them to assess the classic motifs still in use in some projects.

Background

Just over a quarter century ago, it was shown that the mitochondrial DNA (mtDNA) was inherited strictly through the maternal line (Giles, 1980). A few months later, the overall organization of the mitochondrial genome was formally described, and a 16,569-base-pair sequence was established based on the DNA of a single unspecified individual (Anderson, 1981). This sequence, the Cambridge Reference Sequence (CRS)–getting its name in part from the city where the laboratory was located–became the standard reference for reporting mtDNA test results. Any mtDNA sequence could be precisely described by just specifying its differences from CRS. A revised CRS (or rCRS) resulted from reanalysis (Andrews 1999).

Over the past decade the technology of analysis and the nomenclature for describing the analysis of mtDNA has changed dramatically. Thus, to facilitate historical review, it is appropriate to introduce some of that nomenclature. The long string of DNA is connected end to end to form a ring, which in turn is organized into a relatively small region of approximately 1122 base pairs known as the control region, or the D-loop, and a much larger remaining segment known as the coding region. The earliest analyses were effectively spot checks of a few very small segments all over the mtDNA molecule. However, with the recognition of differences in variability throughout that molecule, two sub-regions within the control segment took on special significance for sequencing. These are typically referred to as hypervariable regions 1 and 2 (HVR1 and HVR2). Since, for technical regions, the mtDNA numbering system starts within the control region, HVR2 is numbered from near the beginning while HVR1 is numbered near the end.

Address for correspondence: Jim Logan, jjlnv@comcast.net

Received: December 31, 2007; accepted: March 2, 2008.

For example, in the detailed study of mtDNA variability by Meyer et al (1999), HVR1 is defined as bases 16024 through 16382 (a length of 358 base pairs) and HVR2 as 57 through 371 (a length of 314 base pairs). Many researchers have used similar ranges in their study protocols but there have been differences. For purposes of this project, the hypervariable regions are expanded to cover the entire control region and are defined as 16024 through 16569 for HVR1 and 1 through 576 for HVR2.

The potential for the use of mtDNA in anthropology (and thus genetic genealogy) was demonstrated in a study that concluded that all mitochondrial DNAs stem from one woman who is estimated to have lived about 200,000 years ago, probably in Africa. Thus was born the concept of a "Mitochondrial Eve," supporting the "out of Africa" hypothesis (Cann, 1987). Taking advantage of this maternal-line inheritance, many subsequent studies have sequenced various specific portions of the mtDNA genome, using a variety of evolving techniques, to infer historical relationships among members of selected populations or between populations.

One of the earliest of these studies used the technique of restriction fragment length polymorphism (RFLP) to analyze blood samples from 167 Native American subjects from five widely dispersed populations–three in North America, one in Central America, and one in South America (Toronni, 1992). By applying 14 specific restriction endonucleases they effectively "screened … over 10% of the mtDNA sequence for each individual." They found a total of 67 polymorphic sites and were able to build a phylogenetic tree with four distinct clusters of results (haplogroups) that they arbitrarily called A, B, C, and D.

This study was extended by adding 321 individuals from 17 additional Native American populations (Torroni, 1993a). For 36 of the samples, they also sequenced 341 nucleotides from the displacement loop (D-loop), also

known as the control region, and found that their clustering correlated strongly with the four haplogroups defined by the restriction analysis.

Finally, the Torroni team applied their technology and experience to a study of 411 aboriginal Siberian subjects (Torroni, 1993b). They found similar clusters, but also differences from the Native Americans. Details of their analysis support the theory that the Native American population was genetically derived from early Asian populations. This work also led to the beginning of a formal mtDNA haplogroup system using letters for names and definitions in terms of defined restriction sites based on RFLPs.

In a concurrent study Horai et al. (1993) explored the concept of race using 72 Native American samples for 16 broadly scattered populations throughout North, Central, and South America (Horai, 1993). As distinct from the Torroni studies that used restriction sites scattered over the entire mitochondrial genome, the Horai study relied entirely on the sequencing of a 482-bp segment within the D-loop. They also found four clusters of Native Americans. By comparing the haplotypes with those of world-wide population, including Africans, Europeans, and Asians, they concluded that peopling of the New World was indeed from Asia and from a population of considerable diversity.

The mtDNA Haplogroup J (Hg J) was first distinguished from Eurasian Haplogroups H, I and K through the use of the RFLP technique in an analysis of 175 Caucasians residing primarily in the United States but including 28 French Canadians (Torroni, 1994). Hg J was defined by the RFLP predecessors of nucleotide polymorphisms at rCRS positions 13708 and 16069.

With this background, there was rapid identification of other haplogroups and identification of broad interrelations between them. Haplogroups T, U, V, W, and X were identified in a study using 134 samples from three European populations of Finns, Swedes, and Tuscans (Torroni, 1996). This study found that 99% of mtDNAs fell within the ten haplogroups of H, I, J, K, M, T, U, V, W, and X "suggesting that the identified haplogroups could encompass virtually all European mtDNAs." This study was carried out in the RFLP tradition, but the results were also compared with control region sequences for the Tuscan examples as determined in a separate study (Francalacci 1996). For groups of haplotypes in each haplogroup identified through RFLP analysis, they were able to find identifying concordant nucleotide polymorphisms in the D-loop (control region) that were indicative of that haplogroup. The defining polymorphisms for Hg J were found to be at positions 16069 and 16126 in HVR1 and 295 in HVR2, respectively.

A concurrent, but independent, study that involved sequencing the first hypervariable region (HVR1) of the mtDNA, showed how this technique can be used not only to group haplotypes, but can also use their geographic distribution to infer origin and use their variability to infer age (Richards, 1996). Using 821 widely dispersed test subjects throughout Europe and the Middle East, they identified five primary clusters, 1 through 5, some of which are further divided. The paper also concludes that the frequency of their lineage group 2A (now known as Hg J) varies "widely in Europe, where the range is from 2% (Basques) to 22% (Cornwall)," and that it is probably of Paleolithic Middle East origin.

A study of 37 Italian patients with Leber's Hereditary Optic Neuropathy (LHON) disease and 90 matched control subjects found that subjects with the disease were five times more likely to be of Hg J (35.1%) than were the members of the control group (7.1%) (Toronni, 1997). By contrast, there were relatively fewer LHON patients in the Haplogroup U than were in the controls. The associated phylogenetic analysis found four polymorphic sites to be of particular significance for the Hg J. 4216 + 13708 defines the J itself; 15257 in turn defines the J2 subgroup with its absence defining J1; and finally 15812 within 15257 defines (using their notation) J2.2 with its absence defining J2.1. On the other hand, the mutations most commonly associated with LHON (3460, 11778, and 14484) appear to be independent mutational events and are not definitive of any clade. The Hg J apparently provides a genetic background that supports mutations associated with the disease.

Recognizing that a number of distinct mtDNA classification schemes had arisen due to differences in technology of testing and the use of "imperfect phylogenetic analyses and datasets," a team of researchers, centered on Oxford, proposed a new flexible (i.e., expandable and changeable) nomenclatures scheme (Richards, 1998). They adopted the same capital letters for names of the major mtDNA clusters already in use but then suggested a set theoretic approach such that nomenclature could be systematically expanded to accommodate naming discrete subsets as they were recognized and defined. They also developed rules for inserting nomenclature to represent new groupings relative to previously defined sets. Following their recommendations, they applied this scheme to their previous work; for example, the cluster 2 and its two subclusters 2A and 2B were renamed as haplogroup cluster JT and Haplogroups J and T, respectively. They went further and partitioned several of the haplogroups and gave them names and defined HVR1 assignment motifs for them. For example, nested subsets of Hg J included J1, J1a, J1b, J1b1, and J2. The motif for classification of haplogroups based on HVR1 sequence data were given as sites where differences from the CRS occurred. Thus J was defined

as differences at 16069 and 16126, J1 was given as differences at 16069, 16126, and 16261, J1b1 had differences at 16069, 16126, 16145, 16172, 16222, and 16261, etc.

Another study, centered in Oxford, analyzed 95 samples from the Near East and northwest Caucasus to refine the phylogeny of west Eurasia (Macaulay, 1999). The analysis used RFLP analysis, HVR1 sequencing, and spot evaluating position 00073 both to validate previous analysis and to remove some of its remaining ambiguities. For example, the 16126C, which is shared in the definition of both T and J haplogroups, was supported by 4216C, found through restriction analysis. Some ambiguities, however, remained.

The team then turned to researching founder effects, computation of ages of the various clades, and assessing possible geographic regions of the clade origins (Richards, 2000). To provide a better estimate of the Paleolithic and Neolithic contribution to European diversity their research brought together over 4000 samples from various projects, carefully chosen as representative of various regions throughout Europe, the Near East, and the northern-Caucasus. HVR1 sequences were analyzed between nucleotide positions 16090 and 16365 for the older data but typically include the range ~16050 to 16495 for the newer samples. Refined mtDNA motifs were used for classification where appropriate data was found in the sequence data. This analysis was augmented with restriction enzyme analysis as needed, particularly for Haplogroup H. Supplementary data associated with this study included an expanded HVR-based mtDNA classification motif for the entire phylogeny known to date (Macaulay, 2001). This HVR1 based classification scheme, essentially unchanged, is still in occasional use today (e.g., see supplementary data in Behar 2007).

There followed a number of studies that analyzed the coding region and reported on some aspect of Hg J. In the process, some studies showed that the coding region of the mtDNA was a much better source of data for analysis than the control region (Ingman, 2000; Finnila, 2001; Kivisild, 2004). Some were regional studies (Finnila, 2001), whereas others looked at the relationship between haplogroups and the early human expansion (Maca-Meyer, 2001; Richards, 2002 and 2003), linguistics (Forster, 2004), and even longevity of Hg J centenarians (Rose, 2001). There were, of course, also studies that emphasized the technical aspects, such as network analysis (Herrnstadt, 2002, Coble, 2004, 2006; Santos, 2004).

The first known study devoted exclusively to Hg J was a Master of Science thesis by Serk (2004), where she compared populations distributed across Europe, the Near East, and Central and North Africa. Serk started by choosing 712 test samples from the University of Tartu mtDNA bank based on RFLP and HVS1 motifs and then developed phylogenies using both coding region and control region data. In practice, however, the coding region was represented by only 11 loci assessed either by sequencing small segments or through RFLP analysis. Her phylogenies were developed from 306 samples using the coding region results to classify haplotypes into a first level structure for Hg J and then using control region polymorphisms to develop the details.

In a study designed to resolve uncertainties in the relationships between Indian and western Eurasian mtDNA pools through the study of the phylogeny of mtDNA macrohaplogroup N, a "reappraisal of the Western Eurasian mtDNA Phylogeny," was conducted by Palanichamy et. al. (2004). Upon reviewing the works of Finnila (2001), Rose (2001), and Herrnstadt (2002, 2003), it was concluded that "the former J1a . . . is proven to be one subbranch of J2 on the basis of coding-region sequences." This study further recommended that this subbranch be renamed "J2a," retaining the "a," and that the old J1a name be retired from further use. Using complete mtDNA sequences of 75 Indian samples, supplemented by 25 complete sequences taken from the literature, they reconciled "conflicts among published western Eurasian data sets," refined the basal phylogeny, and presented it in four parts covering respectively N, pre-HV and JT, U, and the Indian autochthonous R. This phylogeny uses both coding and control region polymorphisms in their definitions. In defining the structure of Hg J, they define J1 in terms of 462 and 3010 and J2 in terms of 7476 and 15257. There is no J1a on their chart, but there is a J2a that subsumes the previous HVR1-only motif for J1a.

A more detailed synthesis of the mtDNA phylogeny in the form of "A human mitochondrial genome database," called MITOMAP, is maintained at the University of California, Irvine (Ruiz-Pesini, 2007). The core topology of the associated tree was generated by a neighbor-joining program using 1060 human mtDNA sequences – largely coding region only. It thus provides little information of classifying sequence where the only test data available relates to the hypervariable regions as is typical for many of the tests today. It contains a wealth of data but can be overwhelming since it is a composite of individual test results rather than an analysis to identify a basal phylogeny.

The largest standardized human mtDNA database to date has been assembled through the public participation side of the Genographic Project and includes 78,590 genotypes (Behar, 2007). The population is self-selecting and each kit purchased is analyzed for either a 12-marker Y-DNA test or an HVR1 test; participants receive their results by accessing a web page with their kit number and a password. The first status report was

issued just 18 months into the five-year project and that report describes the quality assurance protocols and a series of computational quality checks based on phylogenetic principles. Concurrent with the report, the project has made available a periodically updated database comprising all data donated by participants and a nearest-neighbor haplogroup-prediction tool. The paper recognized that the mtDNA motifs by themselves are inadequate for classification and thus internally to the project they also use a panel of 22 coding region single nucleotide polymorphisms (SNPs). These SNPs apparently provide a broad classification of the genotypes tested and then the more detailed classification is provided by the HVR1 motifs. Unfortunately, the Genographic Project is still using the motifs patterned after Richards (2000) and, at least for Hg J, this is known to have major flaws as pointed out by Palanichamy (2004).

Goals of the Current Study

Building upon all these prior studies, the aim of the present study is to gather full genome sequence data that can be classified as belonging specifically to mtDNA Hg J and to analyze that data to develop a new phylogenetic tree for that haplogroup. A second aim is to use that tree as a basis for development of consistent classification motifs for use when the only data available is the HVR1 sequence or when there is both HVR1 and HVR2.

Table 1
Identification of Studies Cited and the Geographic Location of the Haplogroup J mtDNA Sequences that Were Used in the Present Study

Source Citation of Referenced Paper	Location or Ethnicity of Population	GenBank Assession Identifiers	Number of Genotypes
Carelli 2006	Italy	DQ341085-DQ341090	6
Coble 2004	European	AY495195-AY495238	44
Datjen 2006	Germany	DQ358973, DQ358974	2
Derenko 2007	Russia	EF397558, EF397562	2
Fraumene 2006	Sardinia	DQ523640, DQ523653, DQ523659, DQ523671	4
Gasparre 2007	Italy	EF660915, EF660916, EF660926, EF660929, EF660952, EF660962, EF660967, EF660981, EF660984, EF660985	10
Gonder 2007	Tanzania	EF184636	1
Greenspan 2007	Primarily United States	DQ787109, EF452293, EU459669, EU007859, EU007880, EU073970, EU155191	7
Ingman 2000	Germany	AF346983	1
Ingman 2007	Yakut, Mansi (Russia)	EU007859, EU007880	2
Maca-Meyer 2001	Morocco, Maragato	AF381987, AF382001	2
Mishmar 2000	Caucasian	AY195754, AY195774, AY195778	3
Moilanen 2003	Finland	AY339577-AF339593	17
Palanichamy 2004	India	AY714033-AY714035	3
Parsons 2005	Hispanic	DQ282488-DQ282492	5
Pereira 2006	Portugal	EF177420, EF177422, EF177431	3
Zsurka 2004	Unknown	AY665667	1

Table 2
Proposed Phylogenetic Tree for Haplogroup J in Table Form. Numbers represent polymorphic sites relative to the rCRS. Those in parentheses () are useful in analysis, but are not definitive by themselves in that they present some degree of homoplasy--they also occur in one or more other clades of Haplogroup J. Numbers in italics represent HVR1 and HVR2 sites.

Hg	Count	Polymorphic Site Definitions
rCRS		
H		*263, 315.1*, 750, 1438, 4769, 8860, 15326
R		*73*, 2706, 7028, 11719, 14766
JT		4216, 11251, 15452, 16126
J	111	*295*, *489*, 10398, 12612, 13708, *16069*
J1	99	*462*, 3010
J1b	22	8269, *(16222), (16145)*
J1b1	15	5460, 13879
J1b1a	12	*242*, 2158, 8557, *(16172)*
J1b1a1	6	15067
J1b2	6	1733
J1b2a	4	6719, 14927
J1c	72	147798, *(185), (228)*
J1c1	25	*188, (185), (16519)*
J1c1a	3	6293
J1c2	14	*482*, 3394
J1c2a	4	9635, 11623, 13879
J1c2b	4	7184
J1c3	11	13934
J1c4	7	9632, 12083
J1c5	5	5198
J1c6	4	4025
J1d	5	7963, *16193, (152)*, (7789)
J2	12	7476, 15257, *(150), (152), (195)*
J2a	6	11377
J2a1	4	10499, 14133, 16231, (7789), (*16145*), (*16261*)
J2b	6	5633, 10172, *16278, (16193)*
J2b1	5	15812

Sources and Methods

Data for the present analysis comes from a variety of world-wide sources previously deposited in GenBank maintained by the National Center for Biotechnology Information (NCBI) of the National Institutes of Health. (See Benson, 2007 for a description). The Hg J sequences that were extracted from GenBank and their sources are shown in Table 1.

Ian Logan's Greasemonkey utility (Logan, 2007) was used as an aid, both in identifying the sequence to extract, and in listing the 'mutations' in each of these sequences. The first search criteria is based on the work of Richards et al (1996) who identified a very distinct cluster defined by a mutation at 16126 that could be cleanly separated in two parts by mutations at 16069 and 16294. That cluster is now referred to as haplo-

Table 3
Matrix Used to Develop the Phylogeny of Table 2 (Table 3 Continued on Next Page)

Haplo-group	ID	C16069T	T16126C	C46262T	G30100A	G82696A	C16262T	C16261T	G16145A	T13879C	G54606A	C24262T	T21558C	G85577A	T16172C	G12007A	T15067C	C2717T	C17337T	C17337G	A14927G	T16719C	T14798C	G18285A	G22884A	A18889G	T16293C	A72455G	G83959A	T48326C	T33945C	A96635C	C11623T	T13899C	A71843G	C13934T	A96632G	T12083G	A51083G	C40255T	T15252C	A79633G	C16193T	G78794A	C74766T	G15257A	C15507T	T19505C	G11377A	A10493G	A14933C	T16231C	A21532G	A13722G	C56337T	G10172A	C16278T	G15812A
J1b*	DQ282491	T	C	T	A	A	T	T	A																																																	
J1b1a	EF660916	T	C	T	A	A	T	T	A	C	A	T	C	A		A																																											
J1b1a	AY495231	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A																																												
J1b1a	AY495234	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A																																												
J1b1a	AY495235	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A																																												
J1b1a	AY495238	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A								A																																			
J1b1a	AY714035	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A																																												
J1b1a1	AY495232	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A	C																																										
J1b1a1	AY495233	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A	C																																										
J1b1a1	AY495236	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A	C																																										
J1b1a1	AY495237	T	C	T	A	A	T	T	A	C	A	T	C	A	C	A	C																																										
J1b1a1	AY339581	T	C	T	A	A	T	T	A	C	A	T	C	A	C		C																																										
J1b1a1	AY339582.	T	C	T	A	A	T	T	A	C	A	T	C	A	C		C																																										
J1b1b	AY714033	T	C	T	A	A	T	T	A	C	A							T																																								
J1b1b	EF397558	T	C	T	A	A		T	A	C	A							T																																								
J1b1b	EF397562	T	C	T	A	A		T	A	C	A							T																																								
J1b2	EF660985	T		T	A	A	T	T	A										T																									C													
J1b2	AF381987	T	C	T	A		T	T	A										T																																						
J1b2a	DQ282488	T	C	T	A	A	T	T	A										T	G	C																																			
J1b2a	DQ282489	T	C	T	A	A	T	T	A										T	G	C																																		
J1b2a	DQ282490	T	C	T	A	A	T	T	A										T	G	C																																		
J1b2a	DQ282492	T	C	T	A	A	T	T	A										T	G	C																																		
J1c	EF459669	T	C	T	A			T														C																																		
J1c	EF660915.	T	C	T	A																	C		A																															
J1c	AY339586	T	C	T	A																	C	A	A																														
J1c	AY495203	T	C	T	A																	C	A	A																														
J1c	AY495206.	T	C	T	A																	C	A	A																														
J1c	AY495205	T	C	T	A																	C	A	A																														
J1c	EF660926.	T	C	T	A																	C	A	A																														
J1c1	AY339587	T	C	T	A				A													C	A		G																				T									
J1c1	AY339590.	T	C	T	A				A													C	A		G																				T									
J1c1	AY339583	T	C	T	A																	C	A		G																													
J1c1	AY339584.	T	C	T	A																	C	A		G																													
J1c1	AY339585	T	C	T	A																	C	A		G																													
J1c1	EF177422	T	C	T	A																	C	A		G																													
J1c1	AY339588	T	C	T	A																	C	A	A	G																													
J1c1	EU155191	T	C	T	A																	C	A		G																													
J1c1	AY495218.	T	C	T	A																	C	A	A	G																													
J1c1	AY495221	T	C	T	A																	C	A	A	G																													
J1c1	AY495222	T	C	T	A																	C	A	A	G																													
J1c1	AY495223	T	C	T	A																	C	A	A	G																													
J1c1	AY495225	T	C	T	A																	C	A	A	G																													
J1c1	AY495226	T	C	T	A																	C	A	A	G																													
J1c1	AY495228	T	C	T	A																	C	A	A	G																													
J1c1	AY495230	T	C	T	A																	C	A	A	G																													
J1c1	DQ358973	T	C	T	A																	C	A	A	G																													
J1c1	EF660929	T	C	T	A																	C	A	A	G																													
J1c1	EF660952	T	C	T	A																	C	A	A	G																				C									
J1c1	AY495224.	T	C	T	A																	C	A	A	G																													
J1c1	AY495219	T	C	T	A																	C	A	A	G																													
J1c1a	AY495220	T	C	T	A																	C	A	A	G	C																												
J1c1a	DQ787109	T	C	T	A																	C	A	A	G	C																												
J1c1a	AY495227	T	C	T	A																	C	A	A	G	C	G	A																								
J1c1a	AY495229	T	C	T	A																	C	A	A	G	C	G	A																								

Table 3 Continued on Next Page

Table 3 (Continued)
Matrix Used to Develop the Phylogeny of Table 2

Haplo-group	ID	C16069T	T16126C	C4627T	G3010A	G8269A	C16222T	C16261T	G16145A	T13879C	G5460A	C2427T	T2158C	G8557A	T16172C	G16207A	T15067C	C2717T	C17337T	A14927G	T6719C	T14798C	G18585A	G22884A	A18848G	T16293C	A17245G	G8339A	T4832C	T3394C	A9635C	C11623T	T13899C	A7184G	C13934T	A9632G	T12083G	A5198G	C4025T	T1552C	A7963G	C16193T	G7789A	C17476T	G15257A	C1550T	T1995C	G11374A	A11049G	A14433G	T16231C	A21535G	A21322G	C5633T	G10172A	C16178T	G15812A
J1c2	EF660981	T	C	T	A																	C							C	C																											
J1c2	EF177431	T	C	T	A																	C	A						C	C																											
J1c2	AY195754	T	C	T	A																	C	A	A					C	C																											
J1c2	AY495198	T	C	T	A																	C	A	A					C	C																											
J1c2	AY495208	T	C	T	A																	C	A	A					C	C																											
J1c2a	AY339591	T	C	T	A																	C		A					C	C	C	T	C																									
J1c2a	AY339592	T	C	T	A																	C		A					C	C	C	T	C																									
J1c2a	AY339593.	T	C	T	A																	C		A					C	C	C	T	C																									
J1c2a	AY495210	T	C	T	A																	C		A					C	C	C	T	C																									
J1c2a	AY495216	T	C	T	A																	C		A					C	C	C	T	C																									
J1c2b	AY495202	T	C	T	A																	C	A	A					C	C				G																							
J1c2b	EF177420	T	C	T	A																	C	A						C	C				G																							
J1c2b	AY714034	T	C	T	A																	C	A	A					C	C				G																							
J1c2b	EF452293	T	C	T	A																	C	A	A					C	C				G																							
J1c3	AY495211	T	C	T	A																	C		A											T																							
J1c3	AY495213	T	C	T	A																	C		A											T																							
J1c3	AY495217	T	C	T	A																	C		A											T																							
J1c3	AY495214	T	C	T	A																	C		A											T																							
J1c3	AY495215	T	C	T	A																	C		A											T																							
J1c3	AY495195	T	C	T	A																	C	A	A		C									T																							
J1c3	AY495201	T	C	T	A																	C	A	A											T																							
J1c3	AY495207	T	C	T	A																	C	A	A											T																							
J1c3	DQ358974	T	C	T	A																	C	A	A											T																							
J1c3	DQ523659	T	C	T	A																	C	A	A											T																							
J1c3	EF660962	T	C	T	A																	C	A												T																							
J1c4	AY346983	T	C	T	A																	C	A	A												G	G																					
J1c4	AY195774	T	C	T	A																	C	A	A												G	G																					
J1c4	AY495197	T	C	T	A																	C	A	A												G	G																					
J1c4	AY495200	T	C	T	A																	C	A	A												G	G																					
J1c4	AY665667	T	C	T	A																	C		A												G	G																					
J1c4	DQ341085	T	C	T	A																	C	A	A												G	G																					
J1c4	DQ341086	T	C	T	A																	C	A	A												G	G																					
J1c5	EU007859	T	C	T	A																		A	A														G																					
J1c5	AY495212	T	C	T	A																	C		A														G																					
J1c5	AY495199	T	C	T	A																	C	A	A														G																					
J1c5	AY495204	T	C	T	A																	C	A	A														G																					
J1c6	AY339589	T	C	T	A				A													C	A															T																					
J1c6	AY495196	T	C	T	A																	C	A	A														T																					
J1c6	AY495209	T	C	T	A																	C	A	A														T																					
J1c6	EU073970	T	C	T	A																	C	A	A														T																					
J1d	EU007880	T	C	T	A																																		C	G	T	A																	
J1d	AF382001	T	C	T	A																																		C	G	T	A																	
J1d	EF184636	T	C	T	A																																		C	G	T	A																	
J1d	DQ341087	T	C	T	A																																		C	G	T	A																	
J1d	DQ341088	T	C	T	A																																		C	G	T	A																	
J2a	EF660984	T	C																																							T	A		C	A															
J2a	EF660967	T	C																																							T	A	T	C	A	G															
J2a1	AY339579	T	C					T	A						C																								C			A	T	A	T	C	A	G	G	C		G								
J2a1	AY339580	T	C					T	A						C																											A	T	A		C	A	G	G	C	G	G								
J2a1	DQ341089	T	C						A																														C			A	T	A	T	C	A	G	G	C	G								
J2a1	DQ523640	T	C					T	A																														C			A	T	A	T	C	A	G	G	C	G	G								
J2b	DQ523671	T	C																																				C		T		T	A	T									T	A	T	
J2b1	AY195778	T	C																																				C		T		T	A	T									T	A	T	A
J2b1	AY339577	T	C																																				C		T		T	A	T	C								T	A	T	A
J2b1	AY339578	T	C																																						T		T	A		C								T	A	T	A
J2b1	DQ341090	T	C																																				C		T		T	A										T	A	T	A
J2b1	DQ523653	T	C																																				C		T		T	A	T									T	A	T	A

group cluster JT and its two parts are Haplogroups J and T. Although there has been some controversy about the details, those criteria for defining these two haplogroups have been consistently used ever since. Thus, using 16069 as the primary search criterion, 111 full genome sequences previously deposited in GenBank were selected and their differences from rCRS extracted. For those sequences referenced in Ian Logan's script, the Greasemonkey tool was used to analyze the differences between the extracted data and the rCRS; otherwise GEN-SNiP (Argus Biosciences, 2007) was used. The differences were entered into a spreadsheet and parsed to develop a sparse matrix with one column for each polymorphic site identified. Rows and columns were manually rearranged to better show the structure of Hg J. The resulting phylogeny is presented in Table 2, followed by Table 3, containing the matrix used to develop this phylogeny. Before discussing the results in detail, the data sources must be described and evaluated.

One form of validation of this data set is to show that they fit properly in the general mtDNA phylogeny. It is generally agreed that the most recent common ancestor of Hg J and the rCRS is Haplogroup R. Each of the sequences selected would be expected to have each of the mutations back to that point. This path contains 295, 489, 10398, 12612, 13708, and 16069 from J back to JT and 4216, 11251, 15452 and 16126 from JT back to Haplogroup R. Ignoring the 16069, since this was the original search criterion, this logic defines the expected content of 999 cells on the matrix. A check of the matrix finds only three cells that were different than expected – three sequences each presenting a different polymorphic site. This relatively low rate of differences could be explained by back mutation or even errors in the sequencing process. Another source of differences from the rCRS is the artifacts generated from the fact that the rCRS is not a direct ancestor. The polymorphic sites from Haplogroup R to Haplogroup H are 73, 2706, 7028, 11719, and 14766, and similarly from H to rCRS itself the sites are 263, 750, 1438, 4769, 8860, and 15326. Only five differences were found for the 1221 cells so defined.

Finally, a check was made to see if any sequences were missing. Doing a separate search, but using 295 vs 16069 as the criterion, five additional candidates were found, but further analysis showed that one of these belonged to Haplogroup I, another to Haplogroup R1, and three to Haplogroup K1. Thus, the set of 111 sequences extracted from GenBank was judged to be complete.

These 111 sequences cannot be considered strictly representative of Hg J since they were selected simply as all available, rather than being stratified by design. However, since the current goal is the development of the cladistic structure of the haplogroup, as distinct from a description of the geographic distribution or other characteristics, this dataset appears to an adequate sample for the purpose. It certainly contains the greatest number of full sequence Hg J records, representing the widest geographic distribution, of any data set that has been assembled to date. Table 1 provides references to the research papers describing studies that produced the Hg J sequences in the GenBank database, the ethnicity or locality that the research covered, and the accession identifiers for those records that were extracted.

A general analysis of the matrix was conducted by counting the number of entries in each column (i.e., for each polymorphism site found in the data) and analyzing the counts within columns. Of the 333 polymorphic sites found in the matrix, 192 were found to be singletons (i.e., they occurred in only one sequence), 50 were doubletons, and 12 occurred only three times. Since the goal is to develop a basal structure for Hg J, polymorphic sites that occurred less than four times were not included in the analysis. Eliminating the 21 sites that are outside the Hg J phylogenetic structure reduced the working matrix to 58 polymorphic sites. Insertions and deletions were initially considered but their distribution was such that no pattern could be discerned that would be useful in defining a clade or subclades of Hg J.

Results

The Phylogenetic Tree

In satisfaction of the first goal of the current study, Table 2 presents a phylogenetic tree in table form. It is followed by Table 3, the portion of the matrix from which the tree relationships were extracted. In partial satisfaction of the second goal, this tree includes not only identification of the polymorphic sites that identify the branches of the tree, but it also shows sites that might be helpful in predicting a clade or subclades when less than full sequence data is available. Since this tree was ultimately derived using only transition sites, the nucleotide designators have been dropped and only site location relative to the rCRS is shown.

The development of this tree employed a parsimony criterion, where the rows and columns have been rearranged to show the relationships more clearly. Once rearranged, the deepest structure of the tree is fairly obvious. First, 99 of the 111 sequences had both the control region 462 polymorphism site and the coding region 3010, and none of the remaining 12 had either of these mutations. On the other hand the other 12 all had both 7476 and 15257, and neither of these polymorphism sites appeared in the first 99 sequences. Consistent with Herrnstadt (2002), Palanichamy (2004), Carelli (2006), Ruiz-Pesini (2007), and others, these two clades were designated as J1 and J2, respectively. Note that this bifurcation is complete--there were no

Table 4
Motifs for Classifying Haplogroup J Sequences When Only HVR1 and HVR2 Data Are Available

Classic HVR1 Motifs (Richards 2000)						
J1b1	16069T	16126C	16145A	16172C	16222T	16261T
J1b	16069T	16126C	16145A	16222T	16261T	
J1a	16069T	16126C	16145A	16231C	16261T	
J1	16069T	16126C	16261T			
J2	16069T	16126C	16193T			
J*	16069T	16126C				
Proposed Motifs for Use With Both HVR1 and HVR2 (See Text)						
J1c1	16069T	16126C	462T	185A or 228A	188G	
J1c2	16069T	16126C	462T	185A or 228A	482C	
J1c	16069T	16126C	462T	185A or 228A		
J1b1a	16069T	16126C	462T	16145A	16172C	16222T
J1b	16069T	16126C	462T	16145A	16222T	
J1d	16069T	16126C	462T	16193T		
J1	16069T	16126C	462T			
J2b	16069T	16126C	16278T			
J2a1	16069T	16126C	16231T			
J2	16069T	16126C				

sequences left over to be classified as any other clade or as J*. This is in stark contrast with the use of the classic motifs currently still in use for classifying HVR1 only sequence data, as will be discussed further below.

Within J1, indicators for three subclades were found. Polymorphic site 8269 is present in 21 sequences and none other; 14798 is present in 71 sequences and none other; and 7963 is present in 5 sequences and none other. Following the recommendation of Palanichamy (2004) and others, these three clades were designated as J1b, J1c, and J1d, respectively. These strict criteria account for 97 of the 99 sequences designated as J1, leaving two unassigned. A closer look, however, reveals that 16222 also occurs in all but two of the J1b sequences as assigned (and does not occur elsewhere in the dataset), and that the combination of 16222 and 16261 occurs primarily in J1b sequences and in very few others. Since one of the unassigned J sequences, AF381987, has 16222, 16261, and 16145, all characteristic of J1b, the parsimonious approach suggested its inclusion in the J1b clade, even without 8269. Possible explanations include a back mutation of the 8269 or a processing error. Similarly, the other unassigned J1 sequence, EU007859, can be assigned to J1c on the basis that it has both the 185 and 228 sites, one or both of which occur in all but two of the already assigned J1c sequences and none other. Again, the possible explanation is a back mutation at 14798. Thus, for the purposes of further analysis, all J1 sequences have been assigned to either J1b, J1c, or J1d, with none left over to be called J1*.

The development process of selecting defining characteristics and assigning names to these subclades continued in like manner through the entire matrix. Note that in the tree thus produced, the locations in parentheses are informative but are not definitive by themselves due to homoplasy (that is, these mutations appear in more than one subclade), whereas those without parentheses are definitive. It is important to note that every clade shown has at least one such absolute defining mutation, several of which are from the HVR2 region, but a few are from the coding region only.

In defining Hg J itself, 16069 is a good indicator for classifying a sequence, but HVR1 data alone is grossly

inadequate in classifying sequences to its clades and subclades—there are no HVR1 only criteria for distinguishing J1 from J2 at their root. However, where HVR2 results are available, site 462 is a good indicator for J1 and its absence is a good indicator for J2. Thus while any attempt at identifying the clades of Hg J using only HVR1 data will produce major errors and using both HVR1 and HVR2 will work much better, coding region indicators are required for definitive classification of all subclades of both J1 and J2.

Several homoplasies were observed, mostly within the second hypervariable region. As concluded by others (e.g., Halgason 2000, Behar 2007), sites 16311 and 16519 are too variable across the entire phylogeny to be useful for classification. Sites 16145, 16193, and 16261, three of the six polymorphisms used in the classic HVR1 motifs for predicting the clades of Hg J (Macaulay 2000, and see next section), were also found to be homoplasic. Nevertheless, these three sites were found to be well defined within the phylogenetic struc-

Table 5
Assignment of Full Genome Sequences into Clades of Haplogroup J According to the Proposed HVR1 + HVR2 Classification Motifs. Cells where the assignment was incorrect are shown with pink background. Where the assignment was the same as the FGS assignment, cells are shown with a green background. Assignments that are correct to the lowest available subclade in the model are shown with a blue background, while cells where the assignment was correct, but incomplete, are shown with a yellow background.

	Haplogroup Assignment per Proposed HVR1 + HVR2 Motifs										
FGS Haplo-group	**J1**	**J1b**	**J1b1a**	**J1c**	**J1c1**	**J1c2**	**J1d**	**J2**	**J2a1**	**J2b**	**Sum**
J1b		1									1
J1b1a		1	4	1							6
J1b1a1			6								6
J1b1b	2	1									3
J1b2		2									2
J1b2a		4									4
J1c	1			6							7
J1c1					21						21
J1c1a					4						4
J1c2	1					4					5
J1c2a						5					5
J1c2b						4					4
J1c3				11							11
J1c4				7							7
J1c5				4							4
J1c6				4							4
J1d							5				5
J2a								2			2
J2a1									4		4
J2b										1	1
J2b1										5	5
Sum	4	9	10	33	25	13	5	2	4	6	111

Table 6
Assignment of Full Genome Sequences into Clades of Haplogroup J According to the Proposed HVR1 + HVR2 Classification Motifs. Cells where the assignment was incorrect are shown with pink background. Where the assignment was the same as the FGS assignment, cells are shown with a green background. Assignments that are correct to the lowest available subclade in the model are shown with a blue background, while cells where the assignment was correct, but incomplete, are shown with a yellow background. Assignments only to J* are shown with orange background.

	Haplogroup Assignment Using Classic Motifs						
FGS Haplo-group	**J***	**J1**	**J1a**	**J1b**	**J1b1**	**J2**	**Sum**
J1b				1			1
J1b1a				1	5		6
J1b1a1					6		6
J1b1b		2		1			3
J1b2				2			2
J1b2a				4			4
J1c	6	1					7
J1c1	21						21
J1c1a	4						4
J1c2	5						5
J1c2a	5						5
J1c2b	4						4
J1c3	11						11
J1c4	7						7
J1c5	4						4
J1c6	4						4
J1d						5	5
J2a	2						2
J2a1			4				4
J2b						1	1
J2b1						5	5
Sum	73	3	4	9	11	11	111

ture of Hg J, are informative, and are thus shown in both the phylogeny presented here and in the associated HVR1 + HVR2 classification motifs. Sites 152 (in HVR2) and 7789 (in the coding region) were also found to be homoplasic, but well structured and so were similarly included.

Classification Motifs

In satisfaction of the second goal of the current study, a new set of classification motifs was developed for use when only control data is available. Two sets of motif criteria are presented in Table 4. The first represents the "classic" motifs as originally presented by Richards et al. (2000), based on HVR1 sequence data only. Although significantly flawed, as pointed out above, it is still in use, as in the Genographic project (Behar, 2007). The second is a proposed motif chart that includes classification criteria for use when both HVR1 and HVR2 data are available. These criteria cannot provide the same detail as a full genome sequence and they can produce errors in classification, but it is a significant

Table 7
Results of Reclassifying 568 Records from MitoSearch According to the Proposed HVR1 + HVR2 Motifs

	Haplogroup Assignment per Proposed HVR1 + HVR2 Motifs										
Mito-Search Haplo-group	**J1**	**J1b**	**J1b1a**	**J1c**	**J1c1**	**J1c2**	**J1d**	**J2**	**J2a1**	**J2b**	**Sum**
J*	10	2	2	140	73	67		8		1	303
J				28	6	7		3	1		45
J1	15			45	3	1		5		2	71
J1a									56		56
J1b		8									8
J1b1			46					1			47
J2					1		3	9		25	38
Sum	25	10	48	213	83	75	3	26	57	28	568

improvement over the classic approach. To use these motifs, go to whichever chart matches the data you have (HVR1 only, or HVR1 and HVR2), work your way down from the first entry until you satisfy all the criteria in that entry. At that point stop and read off the clade classification from the first column. Note that even though one of the original goals called for a set of new motifs for use with HVR1 only sequences, none has been presented here. As described above, any such attempt would be seriously flawed. Instead, comments on possible modification of the classic set of motifs are provided below, but the decision was made not to create a new, but seriously flawed, HVR1-only alternative.

The usefulness of any proposed set of prediction motifs is dependent on both the completeness and accuracy of predictions. Using the reference sequences as the source, Table 5 shows a comparison of predictions provided by the new motifs to those produced by analysis of the full sequence. Of the 111 predictions, all sequences were correctly placed in either subclade J1 or J2, and within J1 and J2, only one sequence was placed in a subclade where it did not belong (one FGS J1b1a was assigned to J1b). Full sequence data, of course provided more precision for the lower level subclades.

Table 6 shows comparable results when using the classic motifs. An application of the classic motifs failed to allocate over 65% of the references sequences to a subclade of Hg J (J1 or J2). In addition, nine of the 111 sequences were placed in an inappropriate subclade. The inability to make assignments is primarily due to the fact that there is no indicator for the J1c clade within the HVR1 sequence and the fact that J1c makes up nearly 64% of the reference data set. The inaccuracy stems from the 1998 attempt to develop a phylogeny and associated motifs based solely on HVR1 data (See Richards, 1998 and 2000). Without HVR2 data J1c simply cannot be recognized.

Unfortunately, no other full genome sequence data set is currently available for use in formal validation of either the phylogeny or the prediction motifs. However, the effectiveness of these motifs can be demonstrated by their application to records from MitoSearch that met the criterion of having been sequenced for both HVR1 and HVR2. Table 7 shows the results. Whereas 61% of the 568 Hg J records from MitoSearch were unassigned to a subclades; after applying the new motifs to reclassify the data, all were assigned to J1 or J2 or one of their subclades. Of the 348 previously unassigned (J*), 93% were assigned to J1c or one of its subclades.

It should be pointed out that the reference dataset and that from MitoSearch represent different populations. The reference dataset derives primarily from world-wide academic research projects and is considered to be the most diverse haplogroup dataset available at this time. By contrast, the MitoSearch population is that of genetic genealogy and for economic and cultural reasons is probably heavily biased toward genetic origins in Europe and migration to or through the British Isles. However, the list of full genome sequences from the genetic genealogy community is growing rapidly and it will soon be possible to perform similar analysis using this data. Until it is proven otherwise, it is believed that

the current research database will survive the test of time with only minor refinements.

With respect to Hg J, it is suggested that the research community would be well served if projects and testing companies were to acknowledge the problems described in this report and inform their clients and researchers accordingly. Unfortunately, because of lack of indicators in the HVR1 sequence, no change to the classic motifs can be made to correct for the inability to allocate a large percentage of Hg J test to subclades. On the other hand the large inaccuracy in assignment to what has been referred to as the J1a clade should be changed to designate the 16145-16231-16161 motif to J2a, or even just J2, in consonance with the academic research community as described above.

This is currently an open-ended study. Not only will these results be refined as new data warrants, but also analysis has begun to establish age estimates for each clade. In furtherance of this study and as a service to the community, a discussion group has been established at http://tech.groups.yahoo.com/group/J-mtDNA/. To help control spam, membership is required for posting and access to the archives, but membership is free.

Acknowledgements

Special acknowledgement goes to Ian Logan (no relationship) for his guidance in accessing GenBank and providing the tools that made that access feasible for this study. He also provided constructive criticism of a draft of this paper. Thanks also go to Robert E. Hausman who suggested editorial improvements to the paper. Finally, thanks go to David Pike who coordinated the review for this journal, and to various anonymous reviewers and editors of *JoGG* whose suggestions have significantly improved this paper

Web Resources

http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=nuccore
EntreNucleotide, Portal for GenBank, etc

http://www.mitosearch.org/
Mitosearch mtDNA Database

References

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature*, 290:457-465.

Andrews RM, Hubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genet*, 23:147.

Argus Biosciences, "GEN-SNiP," software tool to analyze mtDNA sequences relative to rCRS or other sequence, http://www.argusbio.com/sooryakiran/gensnip/gensnip.php, accessed December 2007.

Benson DA, Karsch-Mizrachi I, Lipman DJ , Ostell J, Wheeler DL (2007) GenBank. *Nuc Acids Res*, 35:D21-D25 (Database Issue).

Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, Comas D, Mitchell RJ, Quintana-Murci L, Tyler-Smith C, Wells RS, The Genographic Consortium (2007) The Genographic Project Public Participation Mitochondrial DNA Database. *PLoS Genetics*, 3:1083-1095.

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature*, 325:31-36.

Carelli V, Achilli A, Valentino ML, Rengo C, Semino O, Pala M, Olivieri A, Mattiazzi M, Pallotti F, Carrara F, Zeviani M, Leuzzi V, Carducci C, Valle G, Simionati B, Mendieta L, Salomao S, Belfort R, Sadun AA, Torroni A (2006) Haplogroup effects and recombination of mitochondrial DNA: Novel clues from the analysis of Leber hereditary optic neuropathy pedigrees. *Am J Hum Genet*, 78:564-574.

Carter RW (2007) Mitochondrial diversity within modern human populations. *Nucl Acids Res*, 35:3039-3045.

Coble MD, Just RS, O'Callaghan JE, Letmanyl IH, Peterson CT, Irwin JA, Parsons TJ (2004) Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Int J Legal Med*, 118:137-146.

Coble MD, Vallone PM, Just RS, Diegoli TM, Smith BC,Parsons TJ (2006) Effective strategies for forensic analysis in the mitochondrial DNA coding region. *Int J Legal Med*, 120:27-32.

Detjen KA, Tinschert S, Kaufmann SD, Algermissen B, Nurnberg P, Schuelke M (2006) Identical mitochondrial DNA between monozygous twins with discordant neurofibromatosis type 1 phenotype. *Twin Res Hum Genet*, 10:486-495.

Derenko M, Malyarchuk B, Grzybouski T, Denisova G, Dambueva I, Perkova M, Dorzhu C, Luzina F, Lee HK, Vanecek T, Villems R, Zakharov I (2007) Phylogeopgraphic analysis of mitochondrial DNA in northern Asian populations. *Am J Hum Genet*, 81:1025-1041.

Elson JL, Majamaa K, Howell N, Chinnery PF (2007) Associating mitochondrial DNA variation with complex traits. *Am J Hum Genet*, 80:378-381.

Finnila S, Majamaa K (2001) Phylogenetic analysis of mtDNA haplogroup TJ in Finnish population. *J Hum Genet*, 46:64-69.

Finnila S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. *Am J Hum Genet*, 68:1475-1484.

Forster P, Romano V, Cali F, Rohl A, Hurles M (2004) MtDNA Markers in Celtic and Germanic Language Areas in the British Isles" (Chapter 8) in Jones M (ed.), *Traces of Ancestry : Studies in Honour of Colin Renfrew*, McDonald Institute for Archaeological Research, Cambridge.

Francalacci P, Bertranpetit J, Calafell F, Underhill PA (2006) Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am J Phy Anthropol*, 100:443-460.

Fraumene C, Belle EMS, Castri L, Sanna S, Mancosu G, Cosso M, Marras F, Barbujani G, Pirastu M, Angius A (2006) High resolution analysis and phylogenetic network construction using complete mtDNA sequences in Sardinian genetic isolates. *Mol Biol Evol*, 23:2101-2111.

Gasparre G, Porcelli AM, Bonora E, Pennisi LF, Toller M, Iommarini L, Ghelli A, Moretti M, Betts CM, Martinelli GN, Ceroni AR, Curcio F, Carelli V, Rugolo M, Tallini G, Romeo G (2007) Disruptive mitochondrial DNA mutations in complex I subunits are markers of oncocytic phenotype in thyroid tumors, *Proc Nat Acad Sci* (USA), 104(21):9001-9008.

Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA, *Proc Nat Acad Sci* (USA), 77:6715-6719.

Greenspan G (2007) Direct submission of Family Tree DNA full sequence mtDNA test results to GenBank.

Gonder MK, Mortesen HM, Reed FA, de Sousa A, Tiskhoff SA (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol*, 24:757-768.

Helgason A, Sigurdarbottir S, Gulcher JR, Ward R, Stefansson K (2000) MtDNA and the origin of the Icelanders: Deciphering signals of recent population history. *Am J Hum Genet*, 66:999-1016.

Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet*, 70:1152-1171. See also Elson (2007) for an update of the phylogeny.

Hori S, Kondo R, Nakagawa-Hattori Y, Hayashi S, Sonoda S, Tajima K (1993) Peopling of the Americas, founded by four major lineages of mitochondrial DNA. *Mol Biol Evol*, 10:23-47.

Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. *Nature*, 408:708-713.

Ingman M, Gyllensten U (2001) Analysis of the complete human mtDNA genome: Methodology and inferences for human evolution. *J Heredity*, 92:454-461.

Ingman H Gyllensten U (2007) Rate variation between mitochondrial domains and adaptive evolution in humans. *Hum Mol Genet*, 16:2281-2287.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga W, Villems R (2004) Ethiopian mitochondrial DNA heritage: Tracking gene flow across and around the Gate of Tears. *Am J Hum Genet*, 75:752-770.

Logan I (2007a) *Mitochondrial DNA (mtDNA)*, website at http://www.ianlogan.co.uk/mtDNA.htm.

Logan I (2007b) A suggested genome for 'Mitochondrial Eve.' *J Genet Geneal*, 3:72-77.

Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM (2001) Major genetic mitochondrial lineages delineate early human expansions. *BMC Genetics*, 2:13.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, Torroni A (1999) The emerging tree of West Eurasian mtDNAs: A synthesis of control-region sequences and RFLPs. *Am J Hum Genet*, 64:232-249.

Macaulay V (2001) "Supplementary data from Richards et al. (2000)," available at: http://www.stats.gla.ac.uk/~vincent/founder2000/index.html.

Meyer S, Weiss G, von Haeseler A (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics*, 152:1103-1110.

Mishmar D, Ruiz-Pesini E, Golick P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley MK, Chen E, Brown MD, Sukernik RI, Oickers A, Wallace DM (2003) Natural selection shaped regional mtDNA variations in humans. *Proc Nat Acad Sci* (USA), 100:171-176.

Moilanen JS, Finnila A, Majamaa K (2003) Lineage-specific selection in human mtDNA: Lack of polymorphisms in a segment of MTDN5 gene in Haplogroup J. *Mol Biol Evol*, 20:2132-2142.

Palanichamy MG, Sun C, Agrawal S, Bandelt HJ, Kong QP, Khan F, Wang CY, Chaudhuri TK, Palla V, Zhang YP (2004) Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: Implications for the peopling of South Asia. *Am J Hum Genet*, 75:966-975.

Parsons TJ (2005) Singular nucleotide polymorphisms over the entire mtDNA genome that increase the forensic discrimination of common HV1/HV2 types in 'Hispanics.' Unpublished.

Pereira L, Goncalves J, Franco-Duarte R, Silva J, Rocha T, Arnold C, Richards M, Macaulay V (2006) No evidence for a mtDNA role in sperm motility: data from complete sequencing of asthenozoospermic males. *Mol Biol Evol*, 24:868-874.

Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt HJ, Sykes B (1996) Paleolithic and Neolithic lineages in the European

mitochondrial gene pool. *Am J Hum Genet*, 59:185-203. See also the critique by L L. Cavalli-Sforza and E. Minch (1997) in 61:247-251 and the authors' reply in 61:251-254.

Richards MB, Macaulay VA, Bandelt HJ, Sykes BC (1998) Phylogeography of mitochondrial DNA in Western Europe. *Ann Hum Genet*, 62:241-260.

Richards M, Macaulay V, Torroni A, Bandelt HJ (2002) In search of geographic patterns in European mitochondrial DNA. *Am J Hum Genet*, 71:1168-1174.

Richards M (2003) The Neolithic invasion of Europe. *Annu Rev Anthropol*, 32:135-162.

Rose G, Passarino G, Carrieri G, Altomare K, Greco V, Bertolini S, Boafe M, Franceschi C, DeBenedictis G (2001) Paradoxes in longevity: Sequence analysis of mtDNA haplogroup J in centenarians. *Eur J Hum Genet*, 9:701-707.

Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi P, Wallace DC (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucl. Acids Res*, 35:D823-D828 (Database Issue).

Santos C, Montiel R, Angles N, Lima M, Francalacci P, Malgosa A, Abade A, Pilar M (2004) Determination of human Caucasian mitochondrial DNA haplogroups by means of a hierarchical approach. *Hum Biol*, 76:431-453.

Serk P (2004) *Human Mitochondrial DNA Haplogroup J in Europe and the Near East – A M.Sc. Thesis*, Tartu, Estonia, University of Tartu.

Torroni A, Schurr TG, Yang CC, Szathmary EJE, Williams RC, Schanfield MS, Troup GA, Knowler WC, Lawrence DN, Weiss KM, Wallace DC (1992) Native American mitochondrial DNA analysis indicates that the Amerind and Nadene populations were founded by two independent migrations. *Genetics*, 130:153-162.

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, Wallace DC (1993a) Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J of Hum Genet*, 53:563-590.

Torroni A, Sukernik RI, Schurr TG, Starikovskaya YB, Cabell MF, Crawford MH, Comuzzle AG, Wallace DC (1993b) MtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Am J Hum Gent*, 53:591-608.

Torroni A, Lott MT, Cabell MF, Chen YS, Lavergne L, Wallace DC (1994) MtDNA and the origin of Caucasians: Identification of ancient Caucasian-specific haplogroups, one of which is prone to recurrent somatic duplication in the D-loop region. *Am J Hum Genet*, 55:760-776.

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, Wallace DC (1996) Classification of European mtDNA's from an analysis of three European populations. *Genetics*, 144:1835-1850.

Torroni A, Petrozzi P, D'Urbano L, Sellitto D, Zeviani M, Carrara F, Carducci C, Leuzzi V, Carelli V, Barboni P, DeNegri A, Scovvari R (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 the primary mutations 11779 and 14484. *Am J Hum Genet*, 60:1107-1121.

Zsurka G, Schroder R, Hornblum C, Rudolph J, Wiesner RJ, Elger CE, Krunz WS (2004) Tissue dependent co-segregation of the novel pathogenic G12276A mitochondrial tRNALeu(CUN) mutation with the A185G D-loop polymorphism. *J Med Genet*, 41:e124.