A Comprehensive Analysis of mtDNA Haplogroup J

Jim Logan

Abstract

In the furtherance of a better understanding of human genetic origins and migration history, the Federal GenBank database was mined for all Haplogroup J full-genome mtDNA sequences plus additional sequences that are complete for the coding region. These data were used to develop a phylogeny for Haplogroup J using a matrix developed to show polymorphisms for each sequence organized within clades of the haplogroup. The diversity within clades was then used to compute estimates of the age of each clade. In the process, polymorphisms were analyzed to show their relationship to various genes as well as their relationship to selected medical conditions as reported in the literature. Finally, the literature was reviewed for relevant phylogeographic data toward the ultimate development of a comprehensive history for human mtDNA Haplogroup J.

Introduction

The analysis of mitochondrial DNA (mtDNA) has made significant contributions to the understanding of human evolution and migration. Using restriction enzyme analysis techniques, i.e., comparing restriction fragment length polymorphisms (RFLP), it has been shown that a natural clustering of test results (Torroni et al., 1992) could be used to group the DNA samples into what were later called haplogroups, and thus infer broad genetic backgrounds (Richards et al., 1996). As techniques were refined and it became feasible to do direct sequencing of significant segments of the mtDNA molecule, this clustering was refined and haplogroup classification motifs were developed (Richards et al., 1998). As the database grew the classification structure was extended and relationships between haplogroups were better defined. See Logan (2008) for a historical review of this process as it specifically relates to mtDNA Haplogroup J. One of the purposes of this paper is to present a refined classification structure for Haplogroup J.

Until very recently, sequencing has typically been limited to the control region (displacement loop) of the mtDNA genome, which contains two hypervariable regions (i.e., regions of significantly higher mutation rates than the coding region) which provided relatively more information for a given length of sequence. It was soon discovered, however, that these hyper variable regions have significantly higher instances of back mutations and homoplasies, (i.e., the occurrence of a given polymorphism in more that one haplogroup or even clades of the same haplogroup), thus leading to ambiguities and uncertainties for haplogroup assignment. Some studies then turned to the use of selected markers from the coding region for the broad classification into haplogroups and then used results of sequencing the hypervariable regions to develop the clade structure within the haplogroup. This has been successful for many purposes, but can lead to errors in specific haplogroups. For example, although polymorphisms at nucleotide positions (nps) 16126 and 16069 are adequate for identifying Haplogroup J, and sequencing the complete control region can provide some substructure for J1, there are no polymorphisms within the hypervariable region 1 (HVR1) to cleanly differentiate the J2 clades from those in J1 (Logan, 2008). Whether the purpose was to study a geographic region, a specific disease, or some other purpose, scientists have now published a sufficient number of full-genome sequences to permit a multi-level phylogeny for Haplogroup J and develop estimates of ages of the various clades.

Methods

All mtDNA sequence data used in the current analysis was extracted from the GenBank database maintained by the National Center for Biotechnology Information (NCBI) of the National Institutes of Health. (See Benson et al., 2007 for a description). The broadest available representation of current worldwide population of haplogroup J was achieved by selecting every full-genome sequence (FGS) plus those sequences that were complete except for the control region. Identification of the sequences used in this study is given in Table 1. The Greasemonkey utility (Logan, 2007) was used to identify the sequences to extract and to list the mutations exhibited by each of these sequences. Note that the term mutation is generally limited to the difference between a

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

Received: August 19, 2008; accepted: September 27, 2008.

Table 1
Studies Cited and the Geographic Locations of the Haplogroup J Sequences Used in the Present Study

Source Citation of Referenced Paper	Location or Ethnicity of Population	GenBank Assession Identifiers	Number of Genotypes
Annunen-Rasila 2006	Finland	AM260571	1
Behar 2008	Iran, Azerbaijan	EF445152, EF556169	2
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
Hartmann 2008	Sardinia, South America, Pakistan	EU597520, EU975552, EU975553	3
Ingman 2000	Germany	AF346983	1
Ingman 2007	Yakut, Mansi (Russia)	EU007859, EU007880	2
Kivisild 2006	European	DQ112793, DQ112795, DQ112800, DQ112826	5
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

single sequence and some reference; when multiple sequences are analyzed and differences are found relative to a reference site, polymorphism is generally the preferred term.

Of the 156 sequences selected, 111 are the same sequences used in the previous study (Logan, 2008), seven represent full sequences added to GenBank since that study, and 38 are sequences omitted from the earlier study because there was missing data within the control region thought to be critical in developing the initial J phylogeny. Verification criteria for membership in Haplogroup J include the following: All 156 sequences contained the A12612G and G13708A polymorphisms

and all except three contained A10398G. In addition all 118 truly full-genome sequences had C295T, C16069T, and T16126C in the control region and all except one sequence had a T489C mutation (Logan, 2008).

Each of these sequences was parsed and a matrix was developed to include a column for each sequence and a row for each polymorphism identified. This matrix is the reference for both a detailed analysis of the polymorphisms (including a survey of medical relationships) and the refinement of the Haplogroup J phylogeny. These data were used to compute the average number of polymorphisms in each branch of the phylogeny and to estimate the age of the clades in the phylogeny.

However, certain limitations of this matrix and its origins should be noted. First, although the ethnic origins of donors were generally from European populations or from those located in the western or southern regions of Asia, they do not represent a formally stratified sample. Although it is assumed that the dataset is adequate for development of a phylogenetic tree and an initial estimation of the age of the clades, any conclusions from the geographic distributions calculated from this data should be used with caution.

Second, there was no uniformity in the DNA collection process used by the various researchers. Many of the early studies extracted blood from the participants and then extracted DNA from various components of that blood (Torroni et al., 1992). In other samples, especially those from studies looking for relationships between DNA and disease, the DNA was extracted from biopsies of muscle tissue or even brain tissue (Zsurka, 2004). Some of the latest studies extracted DNA using buccal swabs or mouth washes as collection processes. Although any of these processes can be expected to include mutations passed down through the germ line, they will vary relative to the presence of somatic mutations and heteroplasmies, especially in testing of older donors.

Third, there is no data about either the age or gender of the donor. As shown in the section on Medical Implications below, both factors are significant in analysis of certain diseases and in longevity studies.

Analysis of the Polymorphisms

The development of a phylogeographic analysis is dependent on both good geographic data and characterization of the DNA haplotypes occurring in each region. However, the polymorphisms observed include both mutations passed down from generation to generation, as well as those that occur within a given organism. Furthermore, a DNA sample extracted from a given tissue involves multiple cells, each of which has multiple mitochondria, which in turn have multiple DNA molecules (Jobling et al., 2004). Since these mitochondria reproduce independently, they also mutate independently producing heteroplasmies, which may or may not be reported in testing or may be of too low level to be detected. Finally, the frequency and type of these mutations are influenced by the type of tissue (blood, muscle, brain, etc.) being used as the DNA source and the age of the organism. See Rand (2001) for a more detailed discussion. Thus, depending on these factors as well as the technology being used for extracting and sequencing the DNA, and reporting standards, there can be significant differences in test results, even from the same individual. Therefore, the statistical data about polymorphisms given in this paper and their use in developing the Haplogroup J phylogeny and age estimates must be considered as initial findings. These findings should to be refined as further data sets are developed with proper stratification for tissue tested, geographic origin, and both gender and age of the participants. A complete list of all observed polymorphisms, their location within various genes, and whether or not they are synonymous, is presented in the supplementary material in conjunction with the data organized for development of the phylogeny.

Analysis of 156 sequences of Haplogroup J identified 411 distinct polymorphisms, of which 106 were observed three or more times. The 243 singletons and 62 doubletons, representing almost three-fourths of the total polymorphisms observed, are apparently rare within J. Although these rare polymorphisms are not significant in defining the basal phylogeny, they are useful for inferring the ages of the clades of that phylogeny, and as additional DNA samples are accumulated and matched with genealogical and archaeological data, they may be significant in pinpointing geographic origins and migrations of specific families.

The 16569 base-pair length of the Cambridge Reference Sequence (CRS) encodes 13 genes important to cell metabolism, two ribosomal RNA genes important to transcription and translation of the mitochondrial genome, and 22 transfer RNA genes important in assembling specific amino acids into the products of the mitochondrial genome (Anderson 1981). It also contains one major non-coding region (i.e., the control region) and several very small non-coding sequences within the coding region. Any small variation in the mtDNA sequence can have biological significance depending on several factors: (1) the type of polymorphism (e.g., a simple nucleotide substitution versus an insertion or deletion), (2) its position within the transcription sequence (e.g., its position within a codon that codes for a given amino acid within the protein to which the gene translates), and (3) the specifics of the change, such the substitution of a T for a C. In order to assess these factors (and others) the set of 411 polymorphisms were subjected to a detailed analysis.

Polymorphisms can be arranged into two major categories–those that involve point substitution (transitions, transversions, and heteroplasmies) and those that affect the length of the sequence (insertions and deletions, or indels). If a mutation of the first category occurs within a gene, that mutation has the potential for making a change in the protein for which the gene encodes and thus affecting the phenotype. However, since there is redundancy in the genetic code, many of these mutations (referred to as synonymous mutations) do not result in an amino acid substitution and, thus no change in the protein for which the gene codes. A mutation of the second category occurring within a gene results in a shift in the reading frame which can cause a complete failure of the production of the prospective protein. On the other hand, with the possible exception of interfering with replication of the DNA itself, polymorphisms occurring in one of the non-coding regions have no known effect. The effects of mutations on the genes that code for ribosomal RNA, have not been researched here.

A catalog of polymorphisms was developed from the results of comparing each sequence in the reference database with the revised Cambridge Reference Sequence (Andrews, 1999). The sequences in the reference database, as identified in Table 1, were extracted from GenBank (Benson et al., 2007) and the polymorphisms were identified through the use of Greasemonkey (Logan, 2007). The mtDB database (Ingman et al., 2006) was then searched for each polymorphism to determine the functional locus within the mtDNA genome, and where appropriate, the codon affected, the position on that codon and any resulting change in the encoded protein. For those polymorphisms not cataloged in mtDB, that database was nevertheless helpful as a general guide in identifying the appropriate codon in the reference rCRS (Genbank sequence AC_000021); the Human Mitochondrial Genetic Code from Table 1 found in Anderson et al. (1981) was then used to determine the change in the amino acid. The catalog includes a complete list of polymorphisms, their locations on the respective gene or RNA, implied change in the resulting amino acid residues on associated protein, and, where appropriate, the point on the phylogeny where it is most significant.

A summary of the type of polymorphism versus its locus class is provided in Table 2. Note, however, that polymorphisms in the control region are probably underreported since some of the sequences in the reference database were not complete in that region. Of the 411 polymorphisms detected, only 20 (4%) were insertions or deletions (indels) and these occurred primarily in the non-coding region with a few occurring within the region that codes for the ribosomal RNA. The fact that none occurred in either the genes or in the transfer RNAs is probably due to the deleterious effects that would result and thus would not be passed along in the germ line.

Of the 20 indels detected, one deletion and three insertions occurred within regions defining ribosomal RNA. Each of these is associated with a successive repeat

Table 2
Distributions of Types of Polymorphisms Across the Mitochondrial Genome

Polymorphism Type	Within Gene and Synonymous	Within Gene and Non-Synonymous	Ribosomal	tRNA	Control Region	Non-Coding	Total Polymorphisms	Percent of Polymorphisms	Singletons	Doubletons
Transitions	150	86	27	28	69	6	366	89.05	213	57
Transversions	3	5	1	2	6		17	4.14	12	1
Heteroplasmies			2		6		8	1.95	6	1
Deletions			1		4		5	1.22	3	
Insertions			3		11	1	15	3.65	9	3
Totals	153	91	34	30	96	7	411	100.00	243	62

sequence within that RNA and thus impact would be expected to be minimal. For example at positions 2141 through 2149 of the revised Cambridge Reference Sequence (rCRS) there is a pattern of four AG repeats. The insertion shown as 2149.1A and 2149.2G simply extends the length of this repeat sequence to five repeats. All remaining indels occurred in non-coding regions and all but three of these are also associated with repeat sequences. For example, at locations 514 through 523 of the rCRS there is a pattern of five CA repeats, CACACACACA. There are eight instances of C522 and A523 deletes, reducing the length to four, but there are also two instances of a 523.1C and 523.2A, extending the length to six, and one instance of a 523.1C and 523.2C (See Hurst (2007) for further discussion on length heteroplasmies).

Most of the insertions observed were associated with repeats of a single nucleotide type – most commonly a C. For example the 309.1C insertion was observed 48 times in the sample set of 118 full genome sequences. This insertion relates to the well known sequence from 303 through 315 of the rCRS which consists of a sequence of seven C repeats followed by a T and this followed by five C repeats, CCCCCCCTCCCCC. The 309.1C indicates that there was the insertion of a C after position 309–that is, insertion of a C somewhere before the T in the above sequence. Associated with the same sequence there were also insertions 309.2C, 310.1T and 315.1C.

Of the substitutions, the vast majority (89 % of the total) were simple transitions where a purine was substituted for a purine or a pyrimidine was substituted for a pyrimidine. A little over 4% of the substitutions, however, were transversions (mostly singletons) where a purine was substituted for a pyrimidine or visa versa. Less than 2% were heteroplasmies – a polymorphism within a single organism where the state at a given locus in some DNA molecules was different from the corresponding state in other molecules. Six of the heteroplasmies occurred in the non-coding region and two in the regions that encode for a ribosomal RNA. It should be noted that heteroplasmies are typically unbalanced with one variant dominating the other. It is thus likely that other heteroplasmies were present in the test sequences but went undetected. For males, their heteroplasmies cannot be passed on. For females, there is potential for them to be passed to offspring and descendants either subsequently reverting back to "wild" state or stabilizing to a new state. The significance of such heteroplasmies is thus gender dependent, but such data not available from GenBank.

Table 3 shows how each of these polymorphism types were distributed throughout the various segments of the mitochondrial genome. Note that due to several small overlaps in segment definitions, the lengths of the segments add to slightly greater that the 16569 base pair length of the rCRS genome. As an indication of variability of polymorphisms across the genome, the table also shows the polymorphism density defined as the ratio of the number of polymorphisms within a gene or region divided by the length of that sequence. Note that considering the small numbers involved, the density of polymorphisms throughout the genes encoding for proteins is fairly uniform with an average of 2.1% compared to the 8.6% for the control regions. This four-to-one ratio is no doubt low because of the incompleteness of some of the available sequences as described above. The frequency of polymorphism in the genes for ribosomal RNA is somewhat lower at 1.0%. The control region, which accounts for less than 7% of the mtDNA genome, produced over 23% of the polymorphisms.

Medical Implications

A single nucleotide change within a sequence can cause deleterious or advantageous changes in the performance of mitochondrial-coded products (e.g., proteins). Such changes can be inherited through the gene line from mother to child or they may occur somatically within selected tissues of the individual. Several recent studies have shown correlation between the frequency of selective mutations and a variety of diseases and longevity itself. Such correlation, however, does not necessarily imply a cause and effect relationship. There are very complex relationships between the workings of mitochondrial DNA and nuclear DNA that are not well understood (Carelli, 2003). In the concluding remarks of their paper Santoro et al. (2006) stated that

> *Aging and longevity, as complex traits having a significant genetic component, likely depend on many nuclear gene variants interacting with mtDNA variability, both inherited and somatic. We also surmise that what we hypothesize for aging and longevity could have more general relevance and be extended to other complex trains, such as age-related diseases like cardiovascular diseases and diabetes . . .*

and both Alzheimer's Disease and Parkinson's Disease. The description of such nuclear and mitochondrial DNA interactions is beyond the scope of this paper. This section simply describes a few major medical conditions that have been found at elevated (or reduced) frequency within the mtDNA Haplogroup J population. The only polymorphisms considered here are the ones that actually appeared in the reference database; they are summarized in Table 4. The disease associations were those available from MitoMap (Ruiz-Pesini et al., 2007).

In a study of the relationships between mtDNA polymorphisms and aging, De Benedictis et al. (1999), found

Table 3
Statistical Distribution of Polymorphisms for Various Regions of the Mitochondrial Genome

Locus	Size of Locus	Within Gene and Synonymous	Within Gene and NonSynonymous	Ribosomal	tRNA	Major NonCoding	Other NonCoding	Total Polymor-phisms	Polymorphism Density %	% Synonymous	Synonymous Singletons	Non-Synonymous Singletons	Total Singletons	Percent Singletons	Singleton Density (%)	Singleton % Synonymous	Synonymous Doubletons	Non-Synonymous Doubletons Doubletons	Total Doubletons	Percent Doubletons
ATP6	681	5	10					15	2.20	33	3	6	9	60	0.013	33	2	2	5	33
ATP8	207	5	2					7	3.38	71	4	0	4	57	0.019	100		1	1	14
COI	1542	25	7					32	2.08	78	16	5	21	66	0.014	76	4	2	6	19
COII	684	8	4					12	1.75	67	4	4	8	67	0.012	50				0
COIII	781	11	6					17	2.18	65	7	6	13	76	0.017	54	1		1	6
Cytb	1135	14	15					29	2.56	48	11	7	18	62	0.016	61	1	1	2	7
ND1	957	6	6					12	1.25	50	4	3	7	58	0.007	57	2		2	17
ND2	1042	11	10					21	2.02	52	7	8	15	71	0.014	47	2	1	3	14
ND3	345	7	4					11	3.19	64	5	0	5	45	0.014	100	1		1	9
ND4	1378	20	7					27	1.96	74	12	3	15	56	0.011	80	1	2	3	11
ND4L	297	5	2					7	2.36	71	2	2	4	57	0.013	50	1		1	14
ND5	1812	29	13					42	2.32	69	19	6	25	60	0.014	76	4	3	7	17
ND6	525	7	5					12	2.29	58	5	2	7	58	0.013	71	2	3	5	42
12S rRNA	954			9				9	0.94				6	67	0.006				2	22
16S rRNA	1559			25				25	1.60				16	64	0.010				1	4
tRNA (all)	1509				30			30	1.99				21	70	0.014				5	17
Control Region	1122					96		96	8.56				46	48	0.041				14	15
Other NC	86						7	7	8.14				3	43	0.035				3	43
Totals	16616	153	91	34	30	96	7	411			99	52	243				21	16	62	
Averages										63				59		66				15

Table 4
Polymophisms Observed in the Haplogroup J Reference Database that Have Been Reported as Associated with mtDNA-Related Diseases

Count	Difference from CRS	Locus (gene)	Codon	Codon Base	Amino Acid Change	HG	Medical Relationship (from MitoMap)
153	A10398G	ND3	114	1	Thr-Ala	JT	PD protective factor, longevity
11	C150T	D-Loop				(J2)	Longevity--multiple reports
1	G3460A	ND1	52	1	Ala-Thr		LHON/Confirmed
4	G11778A	ND4	340	2	Arg-His		LHON/Confirmed
2	T14484C	ND6	64	1	Met-Val		Progressive Dystonia--two reports
156	T4216C	ND1	304	1	Thr-His	JT	LHON/Confirmed
155	G13708A	ND5	458	1	Ala-Thr	JT	"LHON/P.M.
22	G15257A	Cytb	171	1	Asp-Asn	J2	Insulin resistance/P.M.
1	G9738A	COIII	178	1	Ala-Thr		LHON/P.M.
3	T10237C	ND3	60	2	Ile-Thr		LHON/P.M.
1	G14831A	Cytb	29	1	Ala-Thr		LHON/Prov
11	G15812A	Cytb	356	1	Val-Met	J2b1	LHON/Prov
17	G5460A	ND2	331	1	Ala-The	J1b1	LHON/Prov
1	T9861C	COIII	219	1	Phe-Leu		LHON/Secondary
2	A11084G	ND4	109	1	Thr-Ala		Alzeimer's, Parkinson's
3	T16189C	D-Loop					Type 2 diabetes, Cardiomyopathy, Endometrial cancer/Prov

that 23% a group of centurions in northern Italy were Haplogroup J, whereas only 2% of a control group of younger persons were Haplogroup J. This contrasted with the results of Haplogroup U that showed centurions were about 2% versus 23.5% for the control group. A subsequent study (Dato et al., 2004) concluded that this effect was population specific since comparable statistics were not found in southern Italy. Ongoing research will likely show that the Haplogroup J population of northern Italy has a higher percentage of J2 than that of southern Italy. As shown in the chart, the polymorphism found to be most significantly related to longevity within J is C150T and as shown below, that polymorphism value is also an indicator for J2. This differentiation likely resulted over many years of separation as one group migrated from the Near East through central Europe and ultimately into northern Italy (with a significant percentage of J2) and the other migrated through the coastal areas of the Mediterranean with a lower percentage of J2. It should be pointed out that the C150T itself (and similarly A10398G) may not have any effect on longevity but rather are markers that are simply statistically correlated.

In a similar study of an Irish population (Ross et al., 2001), Haplogroup J was singled out for special study of longevity. No significant association was found when considering that haplogroup as a whole. However, when they separated the samples into two categories based on restriction fragment analysis, they found that one category had a much higher frequency of centenarians than that the control group whereas the other had a much lower frequency. Then, in a later paper (Ross et al., 2003), and using the same population, they looked specifically at Parkinson's disease. They found of the 12% of the population that was diseased, 2% were in one J group whereas 10%% were in the other J group. They called the first group J1 and the second J2 but unfortunately, their subdivision cannot be correlated with the subclades of J found in the present study since the polymorphic restriction sites have not been identified or to correspond to any polymorphisms found in the reference database.

In a related study of the control region only, Zhang et al. (2003) looked at 207 subjects from Northern, Central, and Southern Italy and found that centenarians and twins had a significantly higher percentage of C150T transitions compared to controls. Based on analysis of multiple tissue types and comparison of twins, as well as longitudinal studies, they concluded the C150T transition can be inherited but it can also occur somatically with age. In considering the possible impact of the C150T transition they noted its proximity to the secondary origin of the replication of the heavy strand of the mtDNA molecule. They found T152C to be fairly common occurrence along with the C150T and also a few T146C in proximity. Further analysis suggested that

> *The somatic event(s) at or near position 150 transition may be part of a general remodeling of the mtDNA replication machinery, probably nuclearly controlled. This remodeling could accelerate mtDNA replication and compensate for the oxidative damage of mtDNA and its functional deterioration occurring in old age.*

The current study found that T150C occurred exclusively in the J2 subclade of Haplogroup J and is thus a strong indicator of that subclade, although not definitive. The reason for this phenomenon has not been determined.

The latest available study to look at the relationship between longevity and Haplogroup J found no significance in the Ashkenazi Jewish centenarians relative to their control group (Shlush, et al, 2008). Although they referenced the study by Zhang (2003), who pointed out the possible significance of the polymorphism 150C, they missed an opportunity for follow-up testing in their well defined and well understood study population. Unfortunately, 150 is not within the narrow range of the control region they sequenced (16024-16300). Similarly, they would be required to acquire additional test data to permit them to assess the possible broader relationship between longevity and the J2 clade for which 150C is an indicator.

The disease most commonly associated with mtDNA Haplogroup J is Leber's Hereditary Optic Neuropathy (LHON), also known as Leber Optic Atrophy (LOA). This disease occurs about five times more frequently in Haplogroup J than it does in the general population (Torroni et al., 1997). LHON is a maternally inherited disease that presents itself in adolescence or adulthood and can lead to partial or total blindness (Wallace 1988). Although some twenty-five mtDNA variants have been observed to be related, the primary mutations are G3460A, G11778A, and T14484C (Brown et al., 2002). One or another of these mutations is found in ninety percent of the families with LHON, although they rarely occur together (John Hopkins, 2008). Of the 156 sequences in the reference database, G11778A occurred four times (twice in J1c4 and twice in J1d), T14484C occurred twice (once in J1d and once in J2b1), and G3460A occurred once in J1c5. MitoMap (Ruiz-Pesini et al., 2007) also listed two reports of progressive dystonia as associated with LHON and specifically with G11778A. The insulin resistance associated with T4216C may just be due to that position being a point mutation for the super-haplogroup JT.

Within the Haplogroup J population, the polymorphism most commonly associated with either Parkinson's or Alzheimer's disease is G5460A, which, incidentally, is one of the two definitive coding region markers that define subclade J1b1. In addition both Parkinson's and Alzheimer's are highly correlated with deterioration of mitochondrial performance, brought on by increasing frequency of polymorphisms, many, or most of which are in heteroplasmic form.

MitoMap showed a relationship between the T11084C polymorphism and the disease MELAS (mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes). A search of the associated bibliography showed only a weak statistical association and that the most common polymorphism for the disease is at position 3243, which was not observed in the reference database. Finally, T16189C has been reported as being associated with various diseases including type 2 diabetes, cardiomyopathy, and endometrial cancer. No bibliographic references were provided to support these reports.

There is a major study currently underway in Europe which is intended to clarify these relationships and identify others (Franceschi et al., 2007). This "5-year European EU-Integrated Project" is entitled "Genetics of Healthy Aging (GEHA)" and constituted by 25 partner organizations "to identify genes involved in healthy aging and longevity, which allow individuals to survive to advanced old age in good cognitive and physical function and in the absence of major age-related diseases." By agreement of the participating partners, it is scheduled to end April 30, 2009. Results should be forthcoming soon.

A Refined Phylogeny

An initial phylogeny for mtDNA haplogroup J was presented in an earlier paper (Logan, 2008). A slightly refined phylogeny is presented here and includes results of analyzing seven full genome sequences added to GenBank since the earlier analysis plus 34 sequences that are complete for the coding region but not complete in the control region. The inclusion of these last 45 sequences does not cause any changes in the primary structure but does permit identification of some detail at the extremities.

As described in the earlier paper, this phylogeny was developed using a maximum parsimony approach ignoring insertions and deletions (see Analysis of The Polymorphisms above). In addition, the polymorphisms located at sites 16311 and 16519 were excluded from the analysis as being too variable to be useful. However, Hagelberg (2003) has suggested that 16311, and possibly 16519, could be the result of ancient recombination. No recent study has been found to support this hypothesis. Future research may ultimately show utility of these polymorphisms.

The refined phylogeny is present in graphic form in **Figure 1**. The supporting data is available in the supplementary files. Note that this chart includes polymorphisms that are in parentheses or are underlined to indicate special conditions. For example the 185 and 228 shown as markers for J1d are both in parentheses because they appear to be subject to back mutations with neither of them appearing in all samples for the J1c clade, nor either of them defining a proper subclade of J1c. However, of the 74 full-genome sequences that are classified as J1c, all but two include one or both of these markers and there is only one occurrence outside the J1c subclades. Similarly the polymorphisms at 152 and 16193, shown in conjunction with subclades J1c, appear to have originated more than once within the haplogroup. These and similar special markers are included to be used as classification aids for cases that are not full genome sequences, but do have sequences from the control region.

Age of The Clades

One of the first uses of molecular biology to determine the age of the human species was just over 40 years ago. Sarich and Wilson (1967) looked at the variations of serum albumins (a blood protein) in humans and non-human primates and concluded that the split between homo, chimpanzee, and gorilla was approximately 5 to 8 million years ago. For calibration, they used the assumption that hominoids in general separated from the old world monkeys 30 million years ago. Within a decade of that study, techniques were sufficiently developed to analyze the DNA itself. Using restriction fragment techniques to analyze samples from baboon, macaques, guenon, and human samples, Brown (1979) estimated that the average mutation rate of mtDNA was about 2% per site per million years.

Before another decade was complete, excitement was aroused in the press and anthropology community when Cann et al. (1987) used mtDNA variations to propose that the current human population "stems from one woman who is postulated to have lived about 200,000 years ago, probably in Africa." This woman is commonly referred to as "Mitochondrial Eve." The important concept here is that of a molecular clock. Since that time, there have been numerous studies that apply this concept to estimate the age of various selective populations (e.g., Stoneking et al., 1986; Wills, 1995). Other studies have looked at the variations of mutation rate for selective regions of the mtDNA molecule (e.g., control region vs coding region) and relative to the effect on corresponding coding region (e.g., synonymous versus non synonymous mutations). Ingman et al. (2000) concluded that the control region "has not evolved at a constant rate across all human lineages and is consequently not suitable for dating evolutionary events." Restricting their analysis to the coding region and using a divergence time between humans and chimpanzees of 5 Myr, they proposed a mutation rate of 1.70×10^{-8} substitutions per site. In a follow-on study, Ingman and Gyllenstein (2001) analyzed the mutation rate for individual genes. The average of these mutation rates is 1.26×10^{-8}. It is not clear why the average computed by data in the second paper is different from that in the first. Both studies were based on 53 samples chosen to be representative of 14 major linguistic phyla in an attempt to avoid bias inherent in selecting individuals based on current population size and geographic location.

Subsequently, Mishmar et al. (2003) used the 53 sequences of Ingman and Gullenstein, but added 48 from African, Asian, European, Siberian and North American populations, to conclude that there are significant differences between geographic populations caused by natural selection brought on by differences in climate and diet. Comparing the ratio of non-synonymous to synonymous mutations within the various genes, they found significant differences between tropical, temperate, and arctic-based populations. Based on estimated coalescence dates for various haplogroups, they estimated the mtDNA evolution rate to be 1.26×10^{-8} substitutions per nucleotide per year.

An alternate basis for calibration of substitution rates was demonstrated by Stoneking et al. (1992; 2005) by capitalizing on a founding event to analyze the population of Papua New Guinea. Their analysis showed that this population had a well defined start date that could be estimated and, further, that population had developed into the current population in relative isolation. They arrived at a "rate of human mtDNA evolution" that was in good agreement with the 2-4% per million previously proposed. Atkinson et al. (2008) built on this idea, assumed a founding date of 45,000 years ago, and developed a substitution rate of 1.691×10^{-8} substitutions/site/year for the coding region.

The studies described above estimated mutation rates based on evolutionary models with calibration typically based on assumed date of separation between humans and chimpanzees. Attempts have also been made to compute mutation rates directly from pedigree data. Early divergence estimates were typically obtained using family data developed for disease studies and consisting

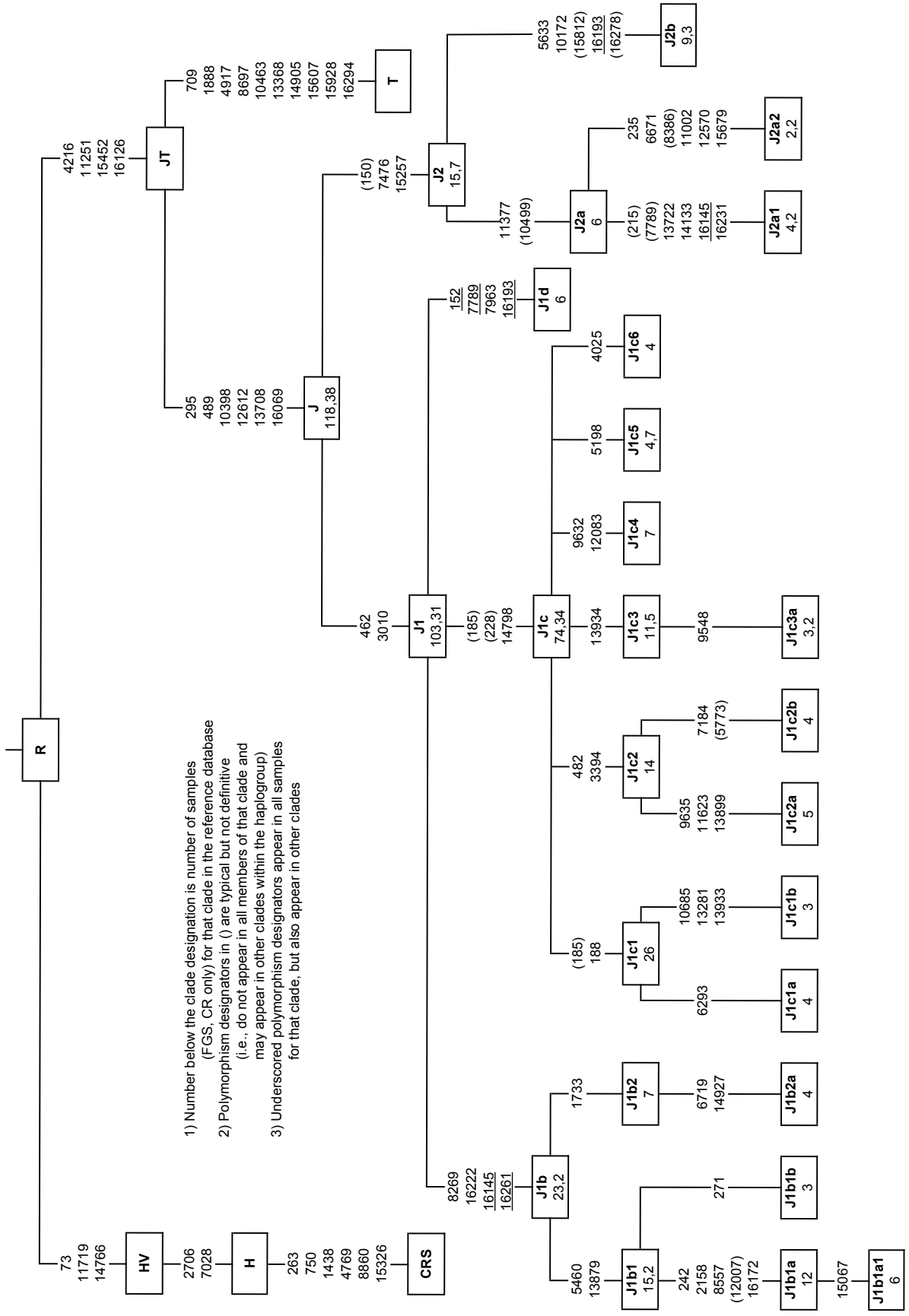

Figure 1. Phylogeny of mtDNA haplogroup J based on 156 sequences from GenBank

Table 5
Estimated ages of the clades of mtDNA Haplogroup J

Subclade of Haplogroup J	Average Length In Subclade	Standard Deviation of Length	Lower Estimate Of Age (kY)	Upper Estimate Of Age (kY)
J	6.071	2.369	14.1	32.1
J1	4.619	1.977	10.1	25.1
J1b	5.200	2.082	11.9	27.7
J1b1	4.412	1.228	12.1	21.5
J1b1a	2.500	1.286	4.6	14.4
J1b1a1	2.143	1.574	2.2	14.2
J1b1b	3.333	0.577	10.5	14.9
J1b2	1.571	0.535	3.9	8.0
J1b2a	0.000	0.000	0.0	0.0
J1c	3.184	1.725	5.6	18.7
J1c1	2.690	1.622	4.1	16.4
J1c1a	3.250	1.708	5.9	18.9
J1c1b	1.500	1.225	1.0	10.4
J1c2	2.929	0.548	9.1	13.2
J1c2a	0.200	0.447	0.0	2.5
J1c2b	3.250	0.500	10.5	14.3
J1c3	2.375	1.088	4.9	13.2
J1c3a	1.800	1.095	2.7	11.0
J1c4	1.143	1.215	0.0	9.0
J1c5	3.182	2.089	4.2	20.1
J1c6	2.500	2.380	0.5	18.6
J1d	3.500	1.643	7.1	19.6
J2	6.500	2.686	14.5	35.0
J2a	7.100	3.071	15.3	38.7
J2a1	2.667	2.066	2.3	18.0
J2a2	3.750	3.403	1.3	27.2
J2b	2.250	1.138	4.2	12.9

of very small sample sizes relative to the rates being estimated. Nevertheless, the general conclusion was that divergence rates for pedigree data were approximately an order of magnitude higher that evolutionary rates (e.g., Howell et al., 2003.) However, as described by Rand (2001), there are many factors that should be considered in stating a final substitution rates. Taking into considerations the gender of the donor, whether the polymorphism appeared to be germ line vs somatic, whether or not the polymorphisms had become fixed, Santos et al. (2008) showed that evolutionary substitution rates and pedigree substitution rates could be reconciled.

This is a good point to note the imprecision of terminology between mutation rates and substitution rates. Mutation rate has to do with the actual change in a DNA molecule with time, whereas *substitution rate*, as used here, has to do with the observable difference with respect to some reference. Many mutations are never observable in testing done for purposes of population genetics. On the other hand, testing of specific tissues may reveal mutations (including heteroplasmies) that have developed somatically as the organism ages.

The problem of calibration and the variability of mutation rates across the mitochondrial genome have been

studied in some detail by Endicott and Ho (2008). Eventually we will be able to account for more of the variability in our analysis. In the meantime, the present work takes a very straightforward but simplified approach for computing the ages of clades of mtDNA Haplogroup J. A substitution rate of 1.7×10^{-8} substitutions/site/year for the coding region was chosen as representative of the literature. Using 15447 for the number of base-pairs in the coding region, this converts to 3808 years per substitution. For each clade the mean *length* of the branches (i.e., the average number of substitutions observed back to the defining polymorphisms) is multiplied by this factor. The result is an estimate of the coalescence time, or Time to the Most Recent Common Ancestor (TMRCA) of the members of that clade. The result of these computations is given in Table 5 and shown on a time-scaled phylogeny in Figure 2. It should be noted, the standard deviation of length, and subsequently the range of ages estimates, is related to the variability of the data; it is not a confidence interval relative to the estimated age.

These ages should be taken as indicating the approximate relative ages of the clades. The astute reader will notice anomalies within these ages. For example, mechanistic computations produced an age for J2 and J2a that are somewhat older than J as the complete clade. This is an artifact of the randomness of mutations and the relative small number of polymorphisms available for defining subclades. It should also be remembered that an overall average substitution rate cannot take into account the fact that mutations do not occur uniformly across the genome, nor can they take into account the fact that clades experienced different migration patterns and thus are subject to different mutation pressures from different climates and diets. Furthermore, since the database used in the analysis was drawn from GenBank, and is thus opportunistic, it most certainly does not represent a random sampling of the current population of the haplogroup.

After describing caveats in their extensive review of status of mutation rates, Bandelt et al. (2006) concluded that the

> *. . . extreme form of weighting that only accepts the coding region but rejects the entire control region is at best provisional and certainly not recommended in the long run. An informed strategy would use rules to decode on a site-by-site basis and contrast synonymous with non-synonymous mutations.*

The technology and data should be available to do such a study in the next few years. For example, data collected in association with the Genographic Project has been used to develop substitution rates for a few selected polymorphisms within the coding region (Rosset et al., 2008).

Origins and Migrations

There is general agreement that there have been three major movements in the peopling of Europe and numerous smaller ones. The first major one, of course, was the initial entry into Europe of anatomically modern humans during the Paleolithic period. Although there is ongoing debate about exact paths through the Near East, there is growing agreement that the ultimate origin of this first set of migrations and initial colonization of Europe was from Africa. The second major set of migrations were from glacial refugia back into the northern regions of Europe where population had been decimated by the Last Glacial Maximum. The third is the inclusion of at least a limited number of migrants associated with the waves of advance of culture, such as agriculture, that occurred during the Neolithic period. The current challenge is to use genetics to develop details about these major movements, to progressively identify and describe the many lesser movements, and to integrate these results with results from other disciplines such as archaeology and linguistics.

One approach to develop such details is the use of genetics and founder analysis to identify populations, date them through using substitution rates for calibration, and analyze the associated geographic data (Stoneking et al., 1992). Phylogeographic analysis, that is the geographic profile of clusters of haplotypes, can provide the basis for inferring geographic origins of selected populations, and probably migration paths. Such inferences take on additional importance in anthropology and population genetics when they are supported by studies from archaeology, climatology, ecology, and linguistics.

One of the earliest uses of the founder analysis approach was the work of Torroni et al. (1992), which concluded that the Amerind and Nadene populations Native Americans were primarily from two independent migrations that probably occurred several thousand years apart. However, using the modern technique of Bayesian skyline plot analysis (Drummond et al., 2005), Mulligan et al. (2008) have developed a three-stage model for the peopling of the Americas; this was one long migration sequence that included three identifiable stages: (1) divergence of Amerind ancestor from the Asian gene pool, (2) a prolonged period of isolation, and (3) rapid expansion into the Americas with a large population increase.

Comas et al. (1997) demonstrated the potential of mtDNA founder analysis when they analyzed data from nine distinct European and West Asian populations and performed analyses to identify statistical similarities between them. Each population came from published samples from a different research team that focused on a specific geographic area, including a Basque, British,

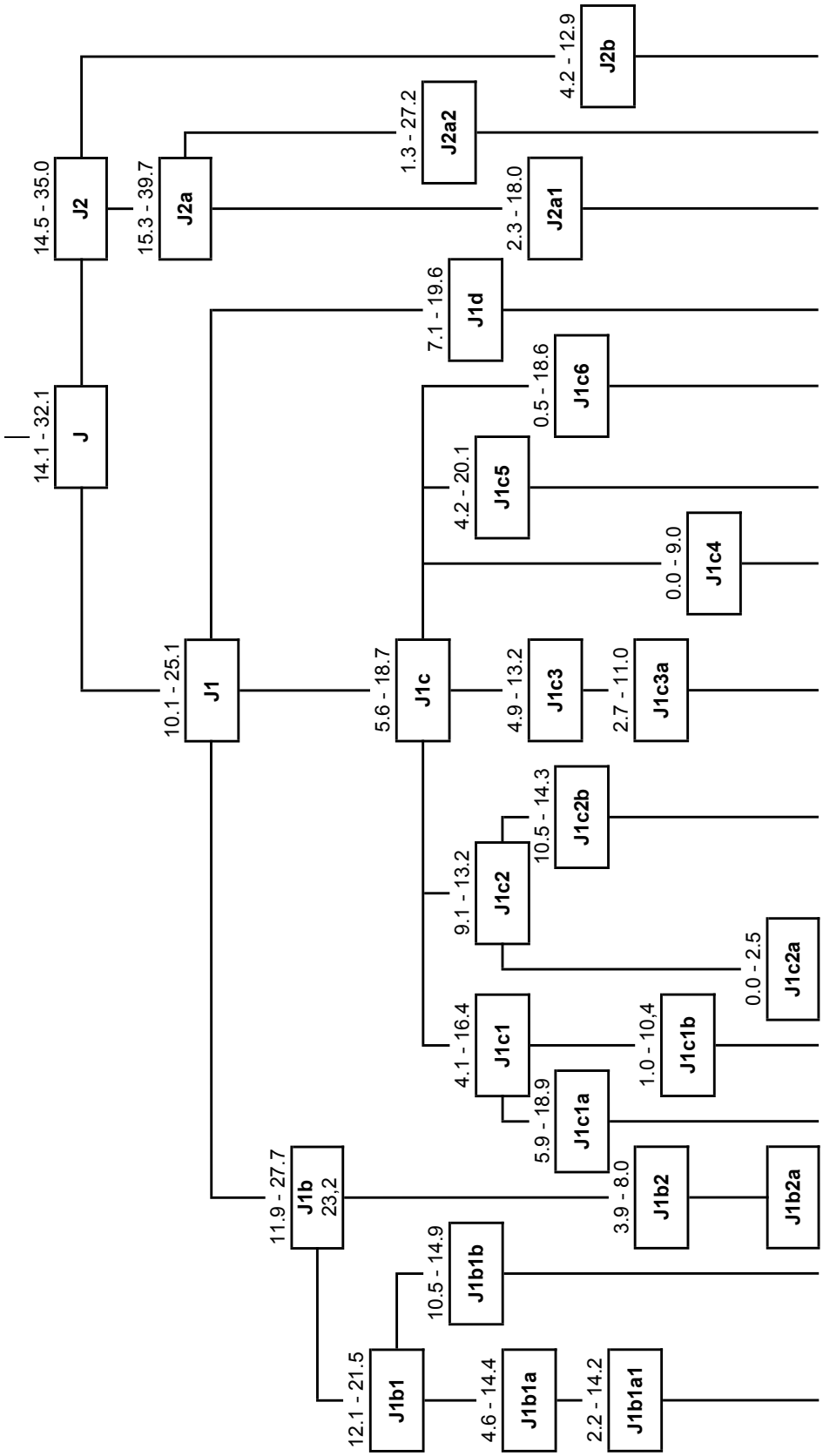

Figure 2. Estimated ages of the clades of mtDNA Haplogroup J

Sardinian, Swiss, Tuscan, Bulgarian, two different Turkish, and a Middle Eastern region. Although differences appeared to be quite low when compared to other world populations (e.g., Africa, Asia, Polynesian, and Native America), a neighbor-joining network developed from the differences turned out to be robust. That network showed "an approximately east-west gradient, with the Middle East at one extreme, followed by the two Turkish samples, most European populations, and, at the other end, the Basque sample."

A large-scale phylogeographic study of mtDNA in Western Europe by Richards et al. (1998) proposed that approximately 85% of 757 individuals tested had their origins in the European Upper Paleolithic whereas the ancestors of the other 15% had arrived more recently from the Near East–predominantly from the haplogroup cluster JT. In an earlier paper (Richards, 1996) they provided the observation that their cluster 2B (now known as Haplogroup T) was relatively uniform throughout Europe with a concentration of about 8%, but their cluster 2A (now know as Haplogroup J) varied widely with a range of about 2% in the Basques up to 22% in Cornwall. They made the observation that it is "in the Middle East, but not elsewhere in the world, we find two missing ancestral haplotypes that link the western and central European clusters" and that the ancestral haplotype is found "in the Middle East but in none of the European samples or elsewhere." Thus, they suggest that Haplogroup J originated in the Middle East and that several different lineages migrated into Europe, splitting into the western and central European clusters but having little impact on the Iberian Peninsula, especially the Basque country." They also developed phylogenetic network for each of several haplotype clusters, including Haplogroups J and T and from these concluded the overall age of Haplogroup J was approximately 28,000 years old, originating in the Middle East and arriving in Europe "when the Neolithic economy spread into Europe starting about 10,000 years ago," with components evolving between 8,000 and 6,500 years ago. Although they provided additional detail for subclusters J1 and J2, it has since been determined that their use of HVR1 data for classification is inadequate (Palanichamy et al., 2004; Logan, 2008) and thus their conclusions are not accurate. In a later review, Richards (2003) expressed the opinion that the main Neolithic founders were "likely to have been members of Haplogroups J and T1, but that the "contribution of the Neolithic Near Eastern lineages to the gene pool of modern Europeans was around a quarter or less."

Using a much expanded study group, Richards et al. (2000) "formalized the procedure for founder analysis, investigated the extent of confounding recurrent gene flow between the putative source and derived populations, and developed criteria that take into account the effects of both gene flow and recurrent mutations." Among their results, they refined the overall age of Haplogroup J to 42,400-53,700 years as determined from the Near East samples and to 23,000-27,400 years as determined from European samples. The corresponding ages for Haplogroup T are 41,900-52,000 and 33,100-40,200 respectively. Although these two clades were apparently contemporary in the Near East, they clearly had different migration patterns into Europe. Their origin within the Middle East has not been established.

In an attempt to identify and describe the effects on mtDNA of "demographic phenomena dating back to the Paleolithic, the Mesolithic, or the Neolithic" periods, Simoni et al. (2000) collected 2619 mtDNA sequences for HVR1 distributed over 36 regions of Europe. Although the sample size was relatively small in some regions, they developed an overall of frequencies for the major haplogroups in each of the regions. No occurences of Haplogroup J were identified is several regions such as Norway and Saami. However, the highs of 17% and 15.9% found in Iceland and Cornwall, respectively, are similar to the frequency of 16.7% in the Near East--its presumed origin. Similarly, Haplogroup T was not observed in Norway or Saami, with the highs of 25.5% occurring in Georgia and 21.7% in the Italian Alps. The Near East frequency was 11.9%.

There is not yet available a comprehensive founder analysis for Haplogroups J or T throughout Europe. However, some inferences can be made about a few very specific geographic areas. One such localized study is the one by Alfanso-Sanchez et al. (2006) that studied the Swanetia region of Georgia. Georgia is in the Caucasus region between Asia and Europe and is considered on the major migration routes from the Near East, yet it is interesting that Haplogroup J did not make the top ten but her sister, Haplogroup T, was one of four that appeared greater than 10% in the population. This could be indicative that the two haplogroups took significantly different migration paths from the Near East into Europe. It is also likely that this indicates that there was a considerable geographic separation of the origin of these two haplogroups, thus causing the different routes.

The origin of Haplogroups J and T in the Middle East and their Neolithic expansion into Europe are well known but the exact origins and specific migration patterns have yet to be established. There is little doubt that the Jordan area was a significant element in the story. In a study of 101 samples from Amman, Jordan and 44 samples from the area of the Dead Sea in the Jordan Valley, Gonzalez et al. (2008) analyzed haplogroup frequencies for these locations and compared them with 23 other Middle Eastern populations. The Amman population was found to have typical haplogroup diversity with the frequencies of Haplogroups J and T at about 6% and 10% respectively, whereas the

diversity of the Dead Sea population was found to be very limited–it was devoid of Haplogroup T and had only a few J1 samples. The final conclusion of the authors of the study was that "although the Levant is a proven crossroad of bi-directional migrations between Africa and West Asia, some geographic areas, such as the Dead Sea area, and social isolates, such as the Druze, have generally resisted that human traffic."

Malyarchuk and his associates did a series of studies of Eastern European populations relating to the origin of the Slavs: Russians and Ukrainians (Malyarchuk and Derenko, 2001), Poles and Russians (Malyarchuk et al., 2002), Bosnians and Slovenians (Malyarchuk et al., 2003), and Czechs (Malyarchuk et al., 2006). In each of these studies they found that most of the mtDNA found belonged to western haplogroups (H, HV, J, T, U, N1, W, and X). Within this broad similarity, they did find heterogeneity between regions with a very broad north-south correlation between their test populations and the corresponding regions to the west. The overall frequencies of Haplogroups J and T found in each region are shown in Table 6.

Malyarchuk and his associates also investigated the origin of the Roma (Gypsies) in Poland (Malyarchuk et al., 2005) and Slovakia (Malyarchuk et al., 2008). The most interesting result of these two studies relative to Haplogroups J and T was the complete absence of T in the Polish population, but 13 of the 69 samples (18.8%) were Haplogroup J. All 13 J samples had the same HVR1/HVR2 signature–an obvious founder effect. The Slovak percentages were more typical with 9.2% concentration of Haplogroup J in the Roma population compared to 9.6% in the general population, but for Haplogroup T, the frequency of 10.6% was nearly twice that of the general population.

Technology of extraction and analysis of mtDNA has progressed to the point where studies of ancient DNA (aDNA) are increasingly reported. Iazgirre and de la Rue (1999) reported on the extraction and coding-region RFLP classification of mtDNA from 121 dental samples from four prehistoric Basque sites. Radiocarbon dating places these samples between approximately 3400 and 5000 years before present. Ignoring the site with a very small sample size, they found that one site had no Haplogroup J or T but the other two sites had a frequency of 15.9% and 16.7% for Haplogroup J and 4.8% and 16.7% for a Haplogroup T-X combination. Some years later, an expanded team (Alzualde et al., 2005) then looked at a different site in the Basque area, dated to the sixth and seventh century AD. They were able to sequence the HVR1 for 48 of the 67 that they classified using RFLP analysis. They found frequencies of 16.7 % for Haplogroup J and 10.8% for the Haplogroup T-X combination. These were comparable to the values for the previous aDNA but significantly higher than the 2.4% and 6% that had been previously reported for the extant population. They suggest that based on lineage J as a mark of migration of Neolithic population from the Near East, that this heterogeneity within the Basque regions shows that "adoption of Neolithic culture followed different paths within the same region," and that certain inferences based solely on the frequencies in present day populations do not appear to be correct." From further analysis of the DNA and associated archaeology, Azuslde (2006) concluded by stressing "the importance of ancient NDA data for reconstructing the biological history of human populations, rendering it possible to verify certain hypotheses based solely on current population data." Further they questioned "the generally accepted belief that, since ancient times, the influence of other human groups has been very scarce in the Basque Country."

However, a recent study was conduced to provide "a more complete characterization of the mitochondrial genome variability of the Basques" (Alfonso-Sanchez et al., 2008). They sequenced HVR1 and HVR2 of 55 healthy men selected to be non-related based on a three-generation pedigree charts. The most interesting result from that study was the high frequency of J, especially J1c and J2a with frequencies of 10.9% and 3.6% respectively. This 14.5% total J is in sharp contrast to the 2.4% commonly referenced for the Basques. On the other hand, it is in line both with the results from ancient DNA and what would be expected when compared with other local regions from the north of Iberia. The complex pattern of spatial heterogeneity is likely to be the result of "restricted gene flow, and accordingly, population fragmentation and reproductive isolation."

Table 6
Frequency of Haplogroups J and T within Eastern Europe

Haplogroup	Bosnia	Slovenia	Czech	Poland	Russian	Ukranian	Polish Roma	Slovak Roma
J	6.7	9.6	11.7	7.8	8.0	7.0	18.8	9.2
T	4.9	5.8	12.3	11.5	10.9	17.0	0.0	10..6

Table 7
Summary of Geographic Analysis of Haplogroup Data Extracted from Macaulay (2000)

Population	Sample Size	Haplo-group J Count	Haplogroup J Frequency	Average J Length	Haplo-group T Count	Haplogroup T Frequency	Average T Length
Nubia	80	3	3.75	2.333	3	3.75	6.67
Egypt	67	4	5.97	3.000	15	22.39	5.93
Bedouin	29	6	20.69	5.000	2	6.90	6.00
Yemen	43	12	27.91	3.417	0	0.00	
Iraq	116	15	12.93	3.533	9	7.76	5.67
Iran	12	1	8.33	5.000	1	8.33	5.00
Syria	69	7	10.14	2.429	7	10.14	4.83
Palestine	117	11	9.40	3.364	15	12.82	5.07
Druze	45	3	6.67	2.667	2	4.44	7.00
Turkey	218	22	10.09	2.955	26	11.93	5.12
Kurd	53	5	9.43	4.600	4	7.55	5.00
Armenia	191	16	8.38	3.563	22	11.52	5.45
Azeri	48	2	4.17	4.000	8	16.67	5.38
N. Caucasus	208	16	7.69	3.188	20	9.62	4.00
Med. East	167	21	12.57	1.571	10	5.99	4.00
Southeast Eur.	233	19	8.15	1.722	26	11.16	6.12
Med. Central	302	23	7.62	2.957	30	9.93	4.33
Alpine Eurpe	218	25	11.47	2.640	18	8.26	4.61
North Cen Eur.	332	31	9.34	2.290	34	10.24	4.88
Med. West	217	16	7.37	2.188	13	5.99	5.23
Basque	156	4	2.56	2.750	8	5.13	3.50
Northwest Eur.	456	63	13.82	2.095	33	7.24	4.76
Scandinavia	316	27	8.54	2.519	26	8.23	4.42
Northeast Eur.	407	32	7.86	2.032	31	7.62	4.84
Overall	**4100**	**384**	**9.37**	**2.992**	**363**	**8.85**	**5.12**

Richards et al. (2000) were cited above as the team that formalized founder analysis of populations using mtDNA data. Thirty-five team members were represented as co-authors of that paper and the supplementary data they produced deserves a more detailed review. Their database (Macaulay, 2001) includes results of HVR1 analysis of 4100 samples from 24 widely distributed regions of the Near East and Europe. The team found 1451 different haplotypes and assigned them to haplogroups using the then existing mtDNA motif classification system. The results included 384 samples of Haplogroup J, or 9.37% of the total, and 363 samples for Haplogroup T, or 8.85%.

For the present study sample sizes and counts for Haplogroups J and T were extracted from the Macaulay database for each geographical region, and the frequencies of the corresponding haplogroups were computed. The results are shown in Table 7. The average length for each of the haplogroups is also shown.

Maps showing the frequency by regions for Haplogroups J and T are presented in Figures 3 and 4, respectively. However, caution must be exercised to avoid reading too much into these maps. In addition to dealing with relative small numbers for some regions, many other factors must be taken into consideration before actual migration paths can be drawn. More specifically, the current frequency in any given region is affected by many factors: population movements do not necessarily produce smooth gradients, but may instead represent movements for relative long distances in an irregular manner; there are back migrations; a popula-

Figure 3. Relative frequency (in percent) of Haplogroup J as derived from (Macaulay 2001)

tion may be decimated by natural disasters or diseases; etc. For example, a casual glance at the map for J might suggest that it originated in what is now Yemen and expanded through the Balkans and on to northwest Europe with spurs on both sides of this line. However, a look at the length of the corresponding haplotype branches suggests that the oldest populations are in the Bedouin, Iranian, and Kurdish regions, supporting the idea of a Middle East origin, but further north.

This review of the literature concerning the origins of the clades is representative but is certainly not exhaustive. More work is required to integrate results, but more importantly, new research is required to provide more data and more complete data. There are several reasons for this.

First, studies have not kept up with the technology. For some geographical areas, the only results available are from RFLP analysis. In other studies the sequence data was limited to HVR1, sometimes complemented with RFLP typing and selective sequencing. Very few results are available for the entire mitochondrial genome.

Second, knowledge of the general phylogeny of Haplogroup J is still evolving and consensus has not yet been reached. The HVR1 motifs used by most of the available studies are not adequate for high-resolution classi-

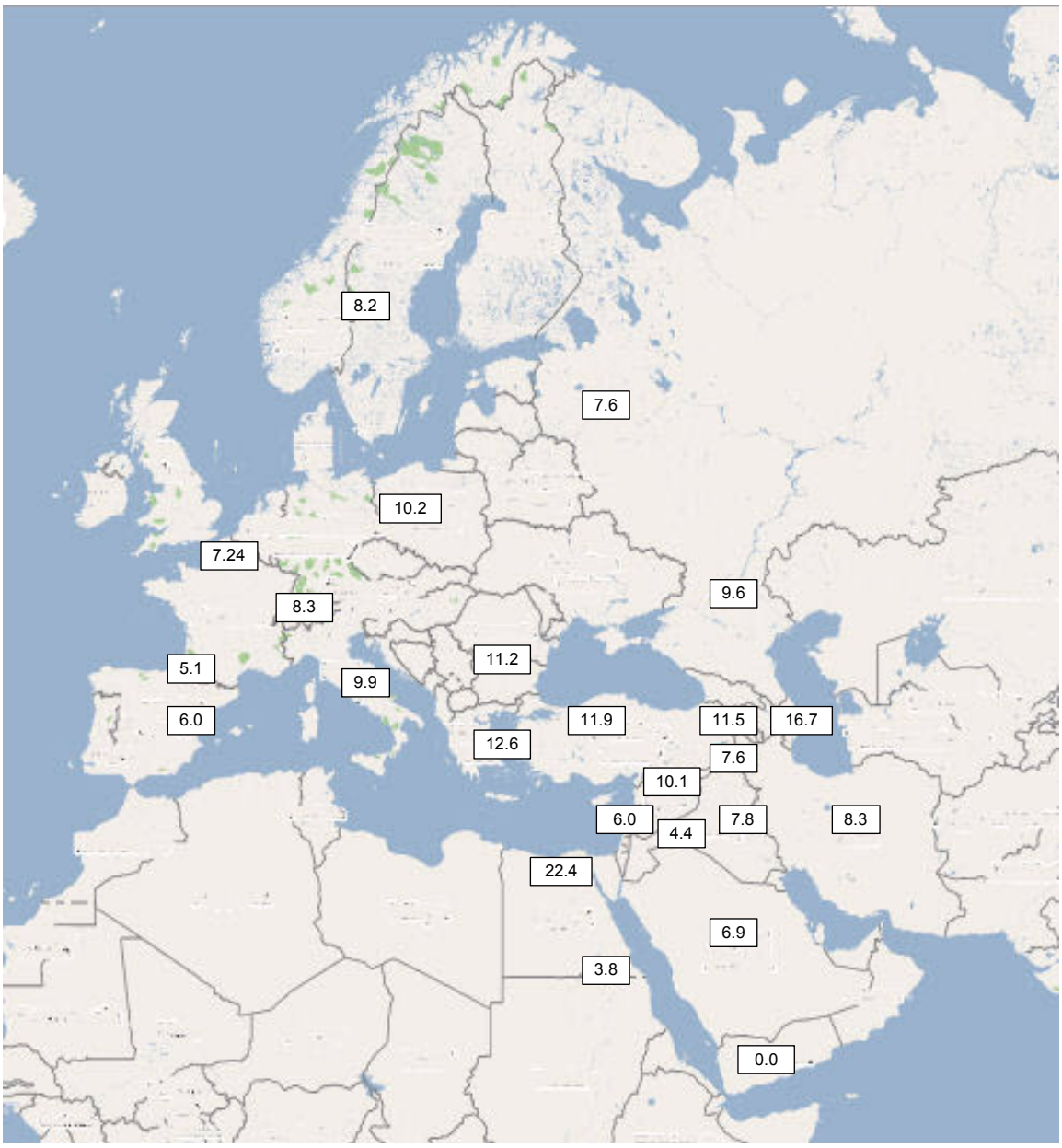

Figure 4. Relative frequency (in percent) of Haplogroup T as derived from (Macaulay 2001)

fication of Haplogroup J, only for identifying a haplogroup as a whole. Most studies are not even distinguishing J2 from J1. Furthermore, errors have been identified, but not all later studies have recognized these errors, or at least have not taken them into consideration consistently.

Third, most basic research is geographically very limited in scope, but then comparisons are made with data from studies of other geographic areas--studies that may be inconsistent in purpose.

Fourth, global databases (such as GenBank) are a great asset for comparing sequences, but are not structured to capture context data beyond literature citations. Founder analysis, for example, requires location. Supplementary databases are needed to cross reference each DNA sample to geographic location of source, any associated archaeological context (e.g., dating of skeletal remains), significant pedigree data (including location and dates), as available, etc.

With time, the improvements will be made, but of course, the technology will have moved on. Nevertheless, the author, for one, expects to continue to review the literature for data relative to a better understanding of Haplogroup J and performing analyses toward that understanding, including refinement of the classification structure, development of expanded databases, and integration of pieces into a global anthropology.

Conclusions

The work described in this paper is a work-in-progress. It provides a broad review of available data concerning mtDNA Haplogroup J and tries to contribute to the evolving knowledge by developing a phylogeny and associated age estimates. It must be noted, however, that the quality of the product is limited by the techniques employed. For example, as stated above, a single mutation rate cannot adequately represent the entire genome. Future analyses should consider both the differences across the various types of gene (e.g., coding for protein versus RNA) and even specific genes. For genes that encode proteins, the analysis should differentiate those polymorphisms that affect amino acid sequences from those that do not. Currently, neither the size of the database, nor knowledge of various mutation rates, were adequate to take these issues into consideration.

Furthermore, as illustrated by the discussion of Origins and Migrations, just the tip of the iceberg has been addressed. Much work is needed to bring together and integrate the many ongoing relevant studies. For example, no attempt has yet been made to analyze population size growth for Haplogroup J. The potential for such analysis can be seen in the study by Atkinson et al. (2008). They employed the Bayesian skyline plot (BSP) with simulation (Drummond et al., 2005) to "simultaneously estimate a posterior probability distribution for the ancestral genealogy, branch lengths, substitution model parameters, and population parameters through time. Such analyses can then be integrated with the archaeological record, legend, and recorded history to develop a more complete story of Haplogroup J.

New studies are required, with the data needs to be developed and integrated. A single project that is both focused on Haplogroup J (and T) and of broad geographic scope may not be feasible at this time. However, it is hoped that a consortium might develop to permit multiple researchers to contribute to an appropriately designed comprehensive project. The author is currently administrating a public discussion group and associated file exchange to further the cause. Interested persons may join through the link to the mtDNA Haplogroup J Project shown under Web Resources, below.

Supplementary Material

Supplementary data is available at:

http://www.jogg.info/42/logansuppl.xls

Web Resources

http://tech.groups.yahoo.com/group/J-mtDNA/
mtDNA Haplogroup J Project

http://www.mitomap.org
Human Mitochondrial Genome Database

http://www.genpat.uu.se/mtDB/
MtDB: Human Mitochondrial Genome Database

References

Alfanso-Sanchez MA, Martinez-Bouzas C, Castro A, Pena JA, Fernandez-Fernandez I, Herrera RJ, de Pancorbo MM (2006) Sequence polymorphisms in the mtDNA control region in a human isolate: the Georgians from Swanetia. *J Hum Genet*, 51:429-439.

Alfanso-Sanchez MA, Cardoso S, Martinez-Bouzas C, Pena JA, Herrera RJ, Castro A, Fernandez-Fernandez I, de Pancorbo MM (2008) Mitochondrial DNA haplogroup diversity in Basques: a reassessment based on HVI and HVII polymorphisms. *Am J Hum Bio*, 20:154-164.

Alzualde A, Izagirre N, Alonso S, Alonso A, de la Rua C (2005) Temporal Mitochondrial DNA Variation in the Basque Country: Influence of Post-Neolithic Events. *Ann Hum Genet*, 69:665-679.

Alzualde A, Izagirre N, Alonse S, Alonso A, Albarran C, Azkarate A, de la Rua C (2006) Insights Into the isolation of the Basques: mtDNA lineages from the historical site of Aldaieta. *Am J Phys Anthropol*, 130:394-404.

Anderson S., Bankier AT, Barrell BG, de Bruijn MHL, Coulson AC, 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. *Nat Genet*, 23:147.

Annunen-Basilla J, Finnila S, Mykkanen K, Moilanen JS, Veijola J, Poyhonen M, Viitanen M, Kalimo H, Majamaa K (2006) Mitochondrial DNA sequence variation and mutation rate in patients with CADASIL. *Neurogenetics*, 7:185-194.

Atkinson QD., Gray RD, Drummond AJ (2008) mtDNA variation predicts population size in humans and reveals a major southern Asian chapter in human prehistory. *Mol Biol Evol*, 252:468-474.

Bandelt HJ, Kong QP, Richards M, Macaulay V (2006) Estimates of mutation rates and coalescence times. In: Bandelt HJ, Macaulay V, Richards M (Eds.) *Nucleic Acids and Molecular Biology*, Vol. 18, Springer-Verlag.

de Benedictis G, Rose G, Carreiri G, de Luca M, Falcone E, Passarino G, Bonafe M, Monti D, Baffio G, Bertolini S, Mari D, Mattace R, Franceschi C (1999) Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans. FESEB J, 13:1532-1536.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2007) GenBank. *Nuc Acids Res*, 35:D21-D25 (Database Issue). The database is available at the following URL: http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore.

Brown MD, Starikovskaya E, Derbeneva O, Seyed Hosseini S, Allen JC, Mikhailovskaya IE, Sukernik RI, Wallace DC (2002) The role of mtDNA background in disease expression: a new primary LHON mutation associated with Western Eurasian Haplogroup J. *Hum Genet*, 110:130-138.

Brown WM, George M, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proc Nat Acad Sci* (USA), 764:1967-1971.

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

Carelli V, Giordano C, d'Amati G (2003) Pathogenic expression of homoplasmic mtDNA mutations needs a complex nuclear-mitochondrial interaction. *Trends Genet*, 19:257-262.

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.

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

Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bosch E, Bertranpetit J (1997) Mitochondiral DNA variation and the origin of the Europeans. *Hum Genet*, 99:443-449.

Dato S, Passarino G, Rose G, Altomare K, Bellizzi D, Mari V, Feraco E, Franceschi C, de Benedictis G (2004) Association of mitochondrial DNA haplogroup J with longevity is population specific. *Eur J Hum Gen*, 12:1080-1082.

Detjen K. A., S. Tinschert, D, Kaufmann, B. Algermissen, P. Nurnberg, and M. Schuelke (2006) Identical mitochondrial DNA between monozygous twins with discordant neurofibromatosis type 1 phenotype, unpublished.

Drummond AJ, Rambaut F, Shapior B, Pybus OG (2005) Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Mol Biol Evol*, 22(5):1185-1192.

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

Endicott P, Ho SYW (2008) A Bayesian evaluation of human mitochondrial substitution rates. *Am J Hum Genet*, 82:895-902.

Franceschi C, and 23 coauthors (2007) Genetics of healthy aging in Europe: the EU-integrated project GEHA. *Ann NY Acad Sci*, 1100:21-45.

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, Bonara E, Pennisi LF, Toller M, Iommarini TL, 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), 104:9001-9008.

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.

Gonzalez AM, Karadsheh N, Maca-Meyer N, Glores G, Cabrera VM, Larruga HM (2008) Mitochondrial DNA variation in Jordanians and their genetic relationship to other Middle East populations. *Ann Hum Bio*, 35:212-231.

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

Hagelberg E (2003) Recombination or mutation rate heterogeneity? Implications for Mitochondrial Eve. *Trends in Genetics*, 19:84-90.

Hartmann A., M. Thieme, L. K. Nanduri, T. Stempfl, C. Moehle, T. Kivisild, Oefner PJ (2008) Validation of microarray-based sequencing of 93 worldwide mitochondrial genomes. Unpublished.

Herrnstadt C, Elson JL, Fahy D, Preston G, Turnbull DM, Anderson C, Ghosh SS, Jolefsky 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.

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Hernstadt C (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am J Hum Genet*, 72:659-670.

Hurst R (2007) Mitochondrial DNA control-region mutations at position 514-524 in Haplogroup K and beyond. *J Genet Geneal*, 3:47-62.

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 M, Gyllensten U (2006) MtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids* Res, 34:D749-D751. The database is available at http://www.genpat.uu.se/mtDB/.

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

Izagirre N, de la Rua C (1999) An mtDNA analysis in ancient Basque populations: implications for Haplogroup V as a marker for the major Paleolithic expansion from southwestern Europe. *Am J Hum Genet*, 65:199-207.

John Hopkins University (2008) Leber Hereditary Optic Neuropathy, LHON. *Online Mendelian Inheritance in Man*.

Jobling MA, Hurles ME, Tyler-Smith C (2004) *Human Evolutionary Genetics*, Garland Publishing, New York and Oxford.

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ (2006) The role of selection in the evolution of human mitochondrial genomes. *Genetics*, 172:272-287.

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

Logan J (2008) The subclades of mtDNA Haplogroup J and proposed motifs for assigning control-region sequences into these clades. *J Genet Geneol*, 4:12-26.

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 (2001) Supplementary data from Richards et al. (2000), http://www.stats.gla.ac.uk/~vincent/founder2000/.

Malyarchuk BA, Derenko MV (2001) Mitochondrial DNA variability in Russian and Ukrainians: implication to the origin of the Eastern Slavs. *Ann Hum Genet*, 65:63-78.

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Wozniak M, Miscicka-Sliwka D (2002) Mithchondrial DNA variability in Poles and Russians. *Ann Hum Genet*, 66:261-283.

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Drobnic K, Miscicka-Sliwka D (2003) Mitochondrial DNA variability in Bosnians and Slovenians. *Ann Hum Genet*, 67:412-427.

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Miscicka-Sliwka D (2005) Mitochondrial DNA diversity in the Polish Roma. *Ann Hum Genet*, 70:195-206.

Malyarchuk BA, Perkova MA, Derenko MV, Vanecek T, Lazur J, Gomolcak P (2008) Mithchondrial DNA variability in Slovaks, with application to the Roma origin. *Ann Hum Genet*, 72:228-240.

Malyarchuk BA, Venecek T, Perkova MA, Derenko MV, Sip M (2006) Mitochondrial DNA variability in the Czech population, with application to the ethnic history of Slavs. *Hum Biol*, 78:581-696.

Mishmar D, and 12 others (2003) Natural selection shaped regional mtDNA variations in humans. *Proc Nat Acad Sci* (USA), 100:171-176.

Mitomap – See Web Resources.

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.

Mulligan CJ, Kitchen A, Miyamoto MM (2008) Updated three-stage model for peopling of the Americas. *PLOS One*, 3:e3199.

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.

Rand DM. (2001) The units of selection of mitochondrial DNA. *Ann Rev Ecol Syst*, 32:415-448.

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 Cavalli-Sforza LL and Minch E, along with the authors' reply, in *Am J Hum Genet*, 61:247-254.

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, and 35 others (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet*, 67:1251-1276.

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

Ross OA, McCormack R, Curran MD, Duquid RA, Barnett YA, Rea IM, Middleton D (2001) Mitochondrial DNA polymorphism: the role in longevity of the Irish population. *Exp Gerentol*, 36:1161-1178.

Ross OA, McCormack R, Maxwell LD, Dugrud RA, Quinn DJ, Barnett YA, Rea IM. El-Agnaf OMA, Gibson JM, Wallace A, Middleton D, Curran MD (2003) mt4216C variant in linkage with the mtDNA TJ cluster may confer a susceptibility to mitochondrial dysfunction resulting in an increased risk of Parkinson's disease in the Irish. *Exp Gerentol*, 38:397-405.

Rosset S, Wells RS, Soria-Hernanz DF, Tyler-Smith C, Royyuru AK, Behar DM, Genographic Consortium (2008) Maximum likelihood estimation of site-specific mutation rates in human mitochondrial DNA from partial phylogenetic classification. *Genetics*, E-Published Articles Ahead of Print. doi:10.1534/genetics.108.091116 (Sep 14, 2008).

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. Nucleic Acids Res, 35:D823-D828 (Database Issue).

Santoro A, Salvioli S, Raule N, Carpi M, Sevina F, Valensin S, Monti D, Bellizzi D, Passarino G, Rose G, Benedictic GD, Franceschi C (2006) Mitochondrial DNA involvement in human longevity. *Biochimica et Biophysica Acta*, 1757:1388-1399.

Santos C, Montiel R, Arruda A, Alverez L, Aluja MP, Lima M (2008) Mutation patterns of mtDNA: empirical inferences for the coding region. *BMC Evol Biol*, 8:167.

Sarich VM, Wilson AC (1967) Inummunological time scale for hominid evolution. *Science* 158:1200-1203.

Shlush LI, Atzmon G, Weisshof R, Behar D, Yudkovsky G, Barzilai N, Skorecki K (2008) Ashkenazi Jewish Centenarians Do Not Demonstrate Enrichment in Mitochondrial Haplogroup J. *PlosOne*, 3(10): e3425.

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patters of mtDNA diversity in Europe. *Am J Hum Genet*, 66:262-278. and Erratum. *Am J Hum Genet*, 66:1785. See also the comments on the article, along with the authors' response, inTorroni, *et al.* (2000) Letter to the editor. *Am J Hum Genet* 66:1173-1179.

Stoneking M,Bharia K, Wilson AC (1986) Rate of sequence divergence estimated from restriction maps of mitochondrial DNAs from Papua New Guinea. *Cold Spring Harbor Symposia on Quantitative Biology,* Vol LI. (On-line access available through JSTOR for members or through member organizations.)

Stoneking M, Sherry ST, Reed AJ, Vigilant L (1992) New approaches to dating suggest a recent age for the human mtDNA ancestor. *Phil Trans R Soc. Lond*, 337:167-175.

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 Nadine populations were founded by two independent migrations. *Genetics*, 130:153-162.

Wallace DC, Singh G, Lott MT, Hodge JA, Shurr TG, Lezza AMS, Elsas LJ, Nikoskelainen EK (1988) Mitochondrial DNA mutations associated with Leber's hereditary optic neueropathy. *Science*, 242:1427-1430.

Wills C (1995) When did Eve live? An evolutionary detective story. *Evolution*, 49:593-607.

Zhang J, Asin-Cayuela J, Fish J, Michikawa Y, Bonafe M, Olivieri F, Passarine G, Benedictis GD, Franceschi C, Atardi G (2003) Strikingly higher frequency in centenarians and twins of mtDNA mutation causing remodeling of replication origin in leukocytes. *Proc Nat Acad Sci* (USA), 100:1116-1121.

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.