Journal of Genetic Genealogy, 3(2):47-62, 2007)

Mitochondrial DNA Control-Region Mutations at Positions 514-524 in Haplogroup K and Beyond

William R. Hurst

Abstract

Long neglected by scientists and mostly excluded from their phylogenetic trees, the variants at positions 00514-00524 in mitochondrial DNA were investigated to determine their usefulness within mtDNA haplogroup K and in the full mtDNA tree. The complex and diverse nomenclature for these variants had to be collected. The percentages of these heteroplasmic variants in the haplogroup K subclades were determined. An attempt was made to establish what, if any, inheritance patterns could be found for these variants in K. How they differ from other mtDNA mutations and how they compare with Y-DNA mutations was investigated. The primary databases used were the mtDNA Haplogroup K Project and the federal GenBank. The few scientific papers on the variants were examined. A less detailed study was made of the variants as they appear in other mtDNA haplogroups. Rules which the variants appear to be following in K were matched against the conclusions of the scientific papers and the observations from the other haplogroups. Finally, areas for further research concerning these variants and other mtDNA mutations were presented.

Introduction

This study began as an investigation of the variants at mitochondrial DNA positions 514 through 524 in sequences from Haplogroup K. The Cambridge Reference Sequence (CRS) variant of these positions consists of alternating cytosine and adenine bases: CACACACACAC. An early observation was that the incidence of variants containing insertions with respect to the CRS (the insertion variants) was significantly higher in K than in other mtDNA haplogroups, while the incidence of the CRS variant in K was somewhat lower, and the deletions variant was significantly lower. Later, using the few scientific papers on the subject, these positions in the other mtDNA haplogroups were investigated. The insertions and deletions at these positions have not been well studied in the past for several reasons. Early mtDNA papers focused on the first hypervariable region, HVR1. Sometimes certain coding-region mutations were investigated, or sometimes HVR2 was included. However, positions 514-524 are in the old HVR3, which has received even less attention. These positions were considered unstable and too variable to be of help in defining subclades. Due to their nature, they, along with certain other less stable mutations, may cause reticulations in phylogenetic trees, so they were usually excluded. Scientists and testing companies could not even agree on what to call them.

The goal in the present study is to rectify the past neglect of these interesting mutations by (1) studying the added resolution that they bring to one mtDNA haplogroup–Haplogroup K, (2) looking at the few scientific papers that focused on them, (3) looking at their role in the mtDNA tree in general, and (4) summarizing what has been learned. Suggestions for future research will follow.

Nomenclature

The first large hurdle that must be dealt with is nomenclature. The CRS is the standard against which all mtDNA sequences are measured. The current version is the Revised Cambridge Reference Sequence, or rCRS, but the common initials CRS will be used here (Anderson et al. 1981; Andrews et al. 1999). The CRS has the sequence CACACACACAC from HVR (*hypervariable region*) positions 514 to 524. Depending on how you look at this sequence it is composed of five CA (cytosine and adenine) or five AC *dinucleotide pairs*. Mutations occur at these positions when one or more pairs of bases are inserted or deleted. In accordance with common practice, insertions and deletions are always measured in reference to the five-CA-pairs found in the CRS sequence (the CRS variant).

Address for correspondence: wrhurst_17@msn.com. W. R. Hurst is the Administrator of the Haplogroup K Project.

Received: July 12, 2007; accepted: August 30, 2007.

Family Tree DNA (FTDNA), whose sequences are used most often in this paper, recently has labeled the mutation when one CA pair is inserted as 524.1C, 524.2A, and 522-, 523- when one CA pair is deleted. Additional insertions are shown as 524.3C, 524.4A, etc. No second pair of deletions has been observed in the FTDNA databases. However, in some older FTDNA test results *indels* (insertions or deletions) are shown as 524.1A, 524.2C, etc. or 523-, 524-. The Sorenson Molecular Genetics Foundation (SMGF) uses the latter set of designations. Other DNA testing companies use different systems of reporting these. Relative Genetics uses 523.1C, 523.2A for the insertions. Argus BioSciences uses 524insA, 524insC and 522delC, 523delA, or even 524insAC for a pair of insertions. Wilson et al. (2002b), representing the Federal Bureau of Investigation forensic unit, recommended 524.1A, 524.2C and 523D, 524D. Ian Logan's mtDNA database commonly uses 523.C, 523.A for the insertions and "C522., A523." for the deletions; but more often they are not listed for each separate sequence, but under "variable changes" at the beginning of a page of sequences. Kivisild et al. (2006), which contains the most recent detailed mtDNA tree, uses 523+CA and 523+2(CA) for one and two pairs of insertions and 522-523d for the deletions. The scientific papers discussed below use a different approach; they simply report the number of repeats. So the CRS variant is "allele 5" or "$(CA)_5$", with one pair of deletions as allele 4 or $(CA)_4$, and one pair of insertions as allele 6 or $(CA)_6$, etc. The Mitomap database lists other scientific papers which refer to insertions and deletions at almost every position between 514 and 524. The main point to remember is that all of these systems are describing exactly the same things. Here the terms CRS variant, one or more pairs of insertions, and one or more pairs of deletions, will be used. Also, the term "position 524" or just "524" will be used instead of 514-524, because there is no way to determine in a string of (CA)n, exactly where any CA insertion or deletion has occurred. "Variants," "insertions" and "deletions" will refer to the position 524 variants, unless otherwise specified. Another point is that the insertions or deletions of C or A never occur individually, but always in CA pairs–except in the rare case of a point mutation (single base change) occurring at one of the positions.

Another nomenclature factor is that the sequence 514-524 is part of the original HVR3 (aka HVS-III) section of the hypervariable region or *hypervariable segment* (HVS), which runs from positions 438 to 534. All of the HVR regions together, plus a few other locations, are also known collectively as the *displacement loop* or *D-loop* or *control region*. FTDNA includes HVR3 as part of its HVR2 test. Argus Biosciences includes HVR3 as part of its HVR package. Relative Genetics offers HVR3 separately (Relative Genetics is being acquired by Ancestry.com, effective by the end of 2007).

References will be made below to sequences in FTDNA's MitoSearch database, the mtDNA Haplogroup K Project (which included 321 high-resolution HVR1+HVR2 sequences as of July 23, 2007), and the federal GenBank database. In MitoSearch, sequences are always labeled as just K, while in the K Project about 10% of the total sequences or 14% of the high-resolution sequences have confirmed subclade (also called subhaplogroup) designations based on full-sequence tests. GenBank sequences vary in how they are labeled, based on their origin. Most subclade designations are those from Behar et al. (2006, Fig. 1), referred to below as the "Behar K tree." Subclade designations of sequences not confirmed by full-sequence tests are as predicted by the author. Additional provisional subclade designations used in this article are those of the author and may change when a new authoritative K tree is published.

Definitions

Mutations

Mitochondrial DNA mutations are most commonly *single nucleotide polymorphisms* (SNPs), in which one of the four *bases* or basic units of DNA, cytosine (C), guanine (G), adenine (A), or thymine (T), is replaced by one of the others. The most common replacements (greater than 95%), C to T and vice versa, and A to G and vice versa, are called *transitions*; all other replacements are called *transversions*. Mutations may also consist of a base being inserted or deleted (indels). Those types of *de novo* mutations are similar to those in nuclear DNA, including Y-chromosome DNA; in fact, SNP mutations are exactly the same in mtDNA as in Y-DNA. Indels at mtDNA locations 514-524 are similar to the *short tandem repeats* (STRs) in Y-DNA. (Chung et al. 2005). Only one copy of Y-DNA is transmitted from father to son, since there is only one copy of the Y chromosome in each cell. For mtDNA, each cell may contain hundreds or even thousands of mitochondria and therefore hundreds or thousands of copies of mtDNA, so that multiple copies of mtDNA are transmitted from mother to child. However, the number of copies transmitted is limited by bottlenecks in egg development (Shoubridge et al. 2007). A *de novo* mutation may occur in just one of the many mtDNA copies, lie undetected for generations, and then by random chance later become the dominant variant. Turner (2006, Fig. 1) has a diagram showing how a variant can go from being undetectable to being either the dominant variant or disappearing completely.

Heteroplasmy

The phenomenon of different mtDNA variants being found in different mitochondria or in different cells in the same person is known as *heteroplasmy. Point*

heteroplasmy (or *structural heteroplasmy*) is the term used when different SNP variants are found in a cell. *Length heteroplasmy* is the occurrence of any mixture of a CRS variant and insertions or deletions in a given region (Scientific Working Group on DNA Analysis, 2003). Insertions and deletions are often found where there are strings of the same base; most commonly these are *poly-cytosine stretches,* sequences with several C bases together. (Carter, 2007, pp 3-4) The suggested mechanism for this type of mutation is the same as that for the common short tandem repeat*s* (STRs) in Y-DNA: *replication slippage,* where the DNA replication system loses count of the numbers of the same base or combination of bases (Howell, 2000, p. 1596). The multi-CA string of bases at mtDNA position 524, being composed of repeats of two different bases, is the most similar to that of Y-DNA STRs. Not coincidentally, as with Y-DNA, the average mutation rate for mtDNA indels may be higher than that for SNPs, although there are exceptions in haplogroup K. See Dupuy et al. (2004) for an extensive discussion of the effect of repeat counts or allele length, and the number of nucleotides in each repeat, on Y-DNA STR mutation rates.

Strictly speaking, when the term *heteroplasmic mutation* is used, or when heteroplasmy is used as a noun, what is usually meant is a situation where two or more variants for the same position are detected by an mtDNA test. Where the heteroplasmy is due to SNP variants, there is a set of IUPAC (International Union of Pure and Applied Chemistry) codes; 16093Y, for example, would mean that both the mutated version 16093C and the CRS variant 16093T were detected in a sample (Scientific Working Group on DNA Analysis, 2003). FTDNA and most other companies performing mtDNA tests for genetic genealogy do not use the codes; apparently, they simply report the variant with the highest percentage. Reportedly, for full-sequence tests FTDNA reports both bases when heteroplasmy is present; but those results are not normally available to anyone except the test subject. Relative Genetics does use the codes. However, there are no IUPAC codes for length heteroplasmies such as occurs at position 524.

Heteroplasmy and heteroplasmic mutation are often used more loosely to explain why certain mutations, SNPs or indels, occur by the inheritance of different variants between generations. Apparently, even if the mutated variant is not detected in the mother, by the normal random processes of cell division and replication, the mutated version may be passed to the child and sometimes become dominant in the child or a later descendant. Thus, a mutation caused by heteroplasmy may appear without there having been either a recent actual replacement of one base with another or a replication slippage. The child simply inherits a different dominant variant from that dominant in the mother (Turner 2006, Fig. 1). The heteroplasmic mutations in the tree appear to be following their own hidden, seemingly mysterious, inheritance patterns. One may think of them as underground rivers, occasionally popping to the surface and then receding – or a parallel system or a second layer of mutations, or mutations lurking below the radar. Pick your favorite analogy.

Perhaps there was an intermediate step where, using the strict definition of a heteroplasmic mutation, both variants were detectable. The key word here is "detectable," since those heteroplasmies that are not detectable by the direct sequencing method commonly used by testing companies – which would require perhaps 20% for the minority variant to be observed – may be detectable at 5% by other methods (Tully et al. 2000). In fact, detection of heteroplasmies as low as 1-2% has a special name: *microheteroplasmy* (Smigrodzki and Khan, 2005).

The Behar K tree demonstrates the problem which the effects of undetectable heteroplasmy cause with trees created with software such as Fluxus-Engineering's Network program. To prevent reticulations caused by heteroplasmic and other recurrent mutations, Behar excluded our 524 insertions as well as the positions 309 and 315 insertions and certain other HVR and coding-region mutations. And yet, there are patterns involving position 524 in the K subclades. The 524 insertions are found in certain subclades, but not in others; and likewise the deletions. These patterns will be discussed in detail for each subclade below. Even adding them back to the data used for the Fluxus diagram does not always explain the appearances of the 524 indels. Turner (2006) expressed the situation well in the title of an article in this Journal: "Now You See It, Now You Don't: Heteroplasmy in Mitochondrial DNA." We will see below how this system works in the K subclades for the position 524 variants.

We see that a mutation reported for a person may have occurred in two general ways; (1) by a *de novo* mutation similar to a nuclear DNA mutation, either by a base replacement (SNP) or by replication slippage, or (2) inheritance of a heteroplasmic variant. Often it is not obvious by which method a mutation has occurred.

In the context of heteroplasmy, the term "fixed" means that only one heteroplasmic variant is inherited by the founder of a subclade. If a different variant appears later in that subclade or a lower subclade, it may be assumed that there has been a *de novo* mutation. "Fixed out" means that a particular variant is missing from the group of inherited variants. If that variant later appears in that subclade or one of its descendant subclades, it again may be assumed that there has been a *de novo* mutation. Tully et al. (2000) has some discussion of the term "fixed." A related term is

"resolved." If a woman with a strict heteroplasmy (two or more variants detectable) has a descendant with only one variant detectable, the position is said to be resolved at that variant. A progression over many generations might be (1) a woman with only the T or CRS variant detectable at position 16093, (2) a heteroplasmy such as 16093Y – both C and T variants detectable, (3) a descendant with the position resolved at 16093C – the T variant not longer detectable, (4) a descendant with the position fixed at 16093C—with the T variant nonexistent for all practical purposes. Sigurðardóttir et al. (2000, p 1606) stated "Furthermore, the processes by which heteroplasmy is resolved—and, hence, the likely long-term fate, in descendants, at the site that is heteroplasmic—does not seem well understood." The difficulty we face is that, for example, when FTDNA says that a person has 16093C, it is not obvious whether the variant is fixed or resolved, or whether they have just picked the majority variant of a heteroplasmy.

Haplogroup Notation

For any mtDNA haplogroup, there are often several levels of subclades or subhaplogroups. For this article, the major or high-level K subclades are K1, K1a, K1b, K1c and K2. All others will be called "lower" subclades. An example of the full list of subclades down one branch of the K tree is K, K1, K1a, K1a1, K1a1b, K1a1b1 and K1a1b1a. Except when specified below, subclade counts do not include that of their lower subclades. The analogy to a tree trunk with smaller and smaller branches and twigs is not perfect; in the above example K1a1b1a happens to be larger than most of its parent subclades when their lower subclades are not included.

Points of Conundrum

For this article the term *points of conundrum* will be used for certain branching points on the K phylogenetic tree which are clearly defined by coding-region or HVR mutations, but which may *appear* to originate or pass on the length heteroplasmic variants at position 524 between generations and nodes on the tree by the only occasionally visible heteroplasmic system. The reason for using the new term is not that a new method of heredity has been discovered, just that the effects of undetected heteroplasmic mutations has not been widely discussed. Typically, a subclade which has haplotypes with more than one variant, divides into two or more lower subclades with different combinations of the variants. **Table 1** shows the percentages of each type of variant (deletions, CRS and insertions) in Haplogroup K as a whole, along with the same information from the Sorenson Molecular Geneology Foundation (SMGF), representing all haplogroups.

In **Table 1**, the percentages of the position 524 variants for the members of the mtDNA Haplogroup K Project are those of the Family Tree DNA high-resolution (HVR1+HVR2) members as of July 23, 2007. The SMGF (Sorenson Molecular Genealogy Foundation) percentages are from their Top 50 Mutations list as of July 10, 2007. The position 524 insertion variants are 4.6 times *more* likely in the K Project than in the SMGF database. The deletions are 7.6 times *less* likely in K than in SMGF. The CRS variant percentage is roughly the same for the two databases, with that for the SMGF database slightly higher. Both databases are probably over-weighted toward USA and Northern Europe samples, so the worldwide percentage of insertions is probably lower than that shown and the percentage of deletions may be higher.

Table 1. 524 Variants in Haplogroup K

	Deletions %	CRS %	Insertions %
K Project	2.2	68.4	29.4
SMGF	16.8	76.8	6.4

Table 2 illustrates the percentages of each variant in most K subclades. The subclades listed include those from the Behar K tree which have examples in the K Project confirmed by full-sequence tests or known examples in GenBank, plus provisional subclades used by the author: K1a10, K1a11, Pre-K1a9 and Pre-K1a10. Those with plus signs, K1a+, K1b+, K1c+ and K2+, include not only samples which have been assigned high-level subclade designations after full-sequence tests; but also samples from the K Project that have not been tested adequately to determine their possible membership in a lower subclades. These may eventually move into one of the more specific lower subclades listed. The Counts column lists the number of examples of each subclade from the K Project and GenBank. The GenBank examples include the 121 full-sequence used in the Behar K tree except for those marked "H" (for Herrnstadt) which, until recently, were not in GenBank. Even now the published Herrnstadt sequences do not include HVR mutations. Added are several other K examples listed on Ian Logan's website. A very few lower subclades on the Behar K tree do not have confirmed examples in the K Project or known examples in GenBank–with HVR mutations–and so are not listed. The next six columns are percentages for the 524 variants. The last column is the combined percentage for the insertion pairs. Deletion variant counts are marked in tan, with yellow used for the CRS variant and blue used for the insertion variants. Sequences from FTDNA's MitoSearch database were also examined, but for the sake of consistency and avoidance of duplications, only the K Project sequences were counted

Table 2. Percentages of Position 524 Heteroplasmic Variants in Haplogroup K Subclades

Subclade	Counts	522-,523-%	CRS %	524.1,524.2 %	524.3,524.4 %	524.5,524.6 %	524.7,524.8 %	524 Total Inserts %
Repeats		4	5	6	7	8	9	
K2+	11-KP		100					0
K2a	34-KP,19-GB	6	94					0
K2a1a	1-GB		100					0
K2a2	1-GB		100					0
K2a2a	8-KP,2-GB		100					0
K2a3	2-GB		100					0
K2a4	1-GB		100					0
K2c	1-GB		100					0
K1	1-KP		100					0
K1c+	14-KP	21	79					0
K1c1	1-KP,8-GB		100					0
K1c1a	1-GB	100						0
K1c1b	4-GB		100					0
K1c2	26-KP,1-GB		96	4				4
K1a+	67-KP	1	70	24	4	1		29
K1a1	1-KP,1-GB		100					0
K1a1a	1-KP		100					0
K1a1b	1-KP,1-GB		100					0
K1a1b1	2-KP,1-GB	33	67					0
K1a1b1a	30-KP,7-GB		97	3				3
K1a6	2-GB		100					0
K1a7	1-GB		100					0
K1a8	3-GB		100					0
K1a11	8-KP		100					0
K1a3	1-GB		100					0
K1a3a	1-KP,1-GB		100					0
K1a3a1	1-GB			100				100
K1a3a1a	1-GB		100					0
K1a2	4-GB		25	75				75
K1a5	1-GB			100				100
K1a4	6-GB			67	33			100
K1a4a1	7-KP,3-GB		30	70				80
K1a4b	1-GB			100				100
K1a4c	1-GB			100				100
Pre-K1a9	6-KP,2-GB		100					0
K1a9	10-KP,4-GB	7	93					0
Pre-K1a10	23-KP,1-GB			42	50	4	4	100
K1a10	24-KP			25	71	4		100
K1b+	3-KP			100				100
K1b1a	7-KP,1-GB		75	25				25
K1b1b	2-GB			50	50			100
K1b1c	1-GB		100					0
K1b2	16-KP,4-GB		35	40	20	5		65

in **Table 2**. Also, although MitoSearch has more K entries than the K Project, none of those are labeled with subclade designations.

The first point of conundrum for Haplogroup K was probably at the founding of K itself, since some of the variants–perhaps only the CRS and one pair each of deletions and insertions–are assumed to have been inherited by the K founder, rather than all of them originating as *de novo* mutations within K.[1] Haplogroup K is divided into K1 and K2. K2, defined by mutations 9716C and 146C, and its lower subclades, apparently did not inherit any of the insertion variants and do not have them now. All of the examples in K2+ in **Table 2** are presumed to be unconfirmed members of K2b, since the other two divisions of K2 have defining HVR mutations. All the K2+ samples here only have the CRS variant. No known K2b examples are found in GenBank. K2a, defined by 709A, 4561C and 152C, has only 6% of haplotypes with the deletion variants, with two examples in the K Project and one on GenBank. Since there are no known examples in the K2, K2a1a, K2a2, K2a2a, K2a3, K2a4, K2b or K2c subclades with the deletions, there is some chance that the ones in K2a were created by a *de novo* mutation at some point after the founding of that subclade.

K1 is defined by mutations 1189C and 10398G and is divided into three subgroups, K1a, K1b and K1c. In the Haplogroup K Project there is only one example, which was determined by a full-sequence test, of a K1 not assigned to a lower subclade.[2] There is also one ancient example, Ötzi the Iceman (Rollo et al., 2006: Endicott et al., 2007). The K1 founder potentially had several of the 524 indel variants, including at least the CRS and one each of the deletion and insertion variants.

K1c is defined by HVR mutations only: 146C, 152C and 498-. Only three exactly matching individuals in K1c+, or probably K1c1, since that lower subclade is defined by coding-region mutations only, have the deletions. Until more examples in different haplotypes are found, there is the possibility that those three came from one *de novo* mutation. The one known available example of K1c1a, from GenBank, has the deletions. The confirmed examples of K1c1 and K1c1b have only the CRS. Only in K1c2, which adds 16320T and three coding-region mutations, is there an example of a pair of insertions in the K1c group. However, that example (actually two closely related individuals) has other rare mutations, which suggests that the insertions may represent a *de novo* mutation.

The major subclade K1a includes over 60% of the K Project, and 80% of all of Behar's K subjects (Behar et al. 2006), although the latter number is probably just a characteristic of the particular population that Behar was studying. K1a is defined by one HVR2 mutation, 497T, which appears no other place in K and perhaps in no other haplogroup. But the founder must have received the CRS variant plus at least one deletion variant and one insertion variant from its parent. Within K1a there is great variety; currently there are examples of six different variants that each appear in at least one K1a sequence. K1a+ mostly consists of those K Project examples which can't be assigned to lower subclades, plus a small number designated as K1a* or just K1a (no lower subclade) by FTDNA (FTDNA has now dropped the use of the asterisk in subclade designations). It has almost every possible variant, missing only the four-insertion-pair variant. There are ten lower subclades of K1a which have only the CRS variant. Four of those, K1a1, K1a1a, K1a1b and K1a3a, are currently represented in the K Project by single examples in subclades requiring full-sequence tests to confirm coding-region results. K1a3, K1a3a1a, K1a6, K1a7 and K1a8 are found only in the GenBank examples. An unlabeled cluster (provisionally called K1a11 by the author) with HVR mutations 16T, 150T, 199C and 16129A, which apparently is the same as the unlabeled sequence on the Behar K tree located between K1a9 and K1a1, also has only the CRS variant. There are three confirmed examples of this cluster in the K Project and five others that are unconfirmed. **Figure 1** below shows just the K1a section of the K tree with suggested additions including K1a11.

The largest Ashkenazi subclade, K1a1b1a, has 30 probable examples in the K Project, with six confirmed by full-sequence tests. All those have the CRS 524 variant. However, one of the seven GenBank examples has one pair of insertions. For that insertion pair to show up so far down the tree from any other sequence with insertions surely means that a *de novo* mutation has occurred, rather than one inherited from the K1a level. It might be noted that this is the only K1a1b1a example on the Behar K tree which doesn't have either 16223T or 114T. The one example of K1a3a1, from GenBank, also has an insertion pair. (Note that its "daughter" K1a3a1a, mentioned above, is CRS; so the insertion pair probably wasn't fixed).

One lower subclade, K1a1b1, has only two confirmed examples in the K Project; one has the CRS variant, one the deletions. The one GenBank example has the CRS variant.

K1a4 is defined by coding-region mutation 11485C. K1a4a adds mutation 6260A. So far, there are no confirmed examples of those subclades, or K1a4b or

[1] See MitoSearch entry ARFHH for an interpretation of the ancestral haplotype for K, but this assumes only the 524 CRS variant.
[2] Also, Behar (2006, Table 6) lists one Iranian K1*.

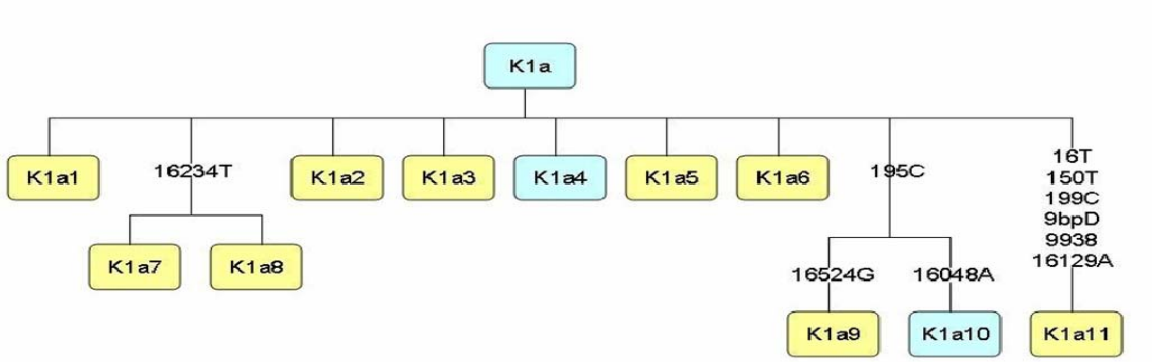

Figure 1. Simplified K1a Tree with Additional Provisional Subclades K1a10 and K1a11

K1a4c, in the K Project. K1a4 is represented in six GenBank sequences; four of which have one insertion pair, while two have two pairs. K1a4a1, which is defined by the addition of coding-region mutations 11840T and 13740C, and is represented in the K Project by seven confirmed examples and three in GenBank, so far either has the CRS variants (three examples) or one pair of insertions (seven examples). K1a4b and K1a4c each have only one GenBank example with one insertion pair. So, K1a4 in general appears to have either inherited three variants, with one insertion pair being dominant; or it inherited only the one-pair variant with the CRS and the second insertion pair mutating later by replication slippage.

K1a2, found only in four GenBank samples, has one CRS and three with one insertion pair. K1a5 has only one example with a pair of insertions in GenBank.

The last K1a lower subclade on the Behar K tree is K1a9, but it is not at the same level as most of the other eight. Instead, it is below a branching point at 195C, which is perhaps the most interesting point of conundrum. The examples below this point on the Behar K tree include only 8.3% of the 121 total sequences used to create that tree; while in the K Project 19.6% of 321 sequences appear to fall below that point. The founder with 195C must have transmitted the CRS, the deletion variant, and at least one pair of insertions. K1a9, defined by 16524G, is thus three levels down from the last defining coding-region mutation at the K1 point. All ten of the K1a9 sequences in the K Project (and all those in MitoSearch) only have the CRS variant, but one of the four on GenBank has a deletion variant. That single example is probably due to a *de novo* replication slippage.

Although it is not on the Behar K tree, probably because Behar found only one example–a non-Jewish Moroccan (Behar et al., 2006, Table 4)–there is a second large cluster under 195C defined by 16048A, provisionally called K1a10. This cluster never has deletions or the CRS variant. Instead it has from one to three pairs of the insertions. However, K1a9 and K1a10 may not have branched directly from the 195C point. There are many K1a sequences which have 195C, but not 16524G or 16048A. One was confirmed as a K1a4a1, another confirmed as a K1a1b1a; but two have been reported as just K1a*–not in a known lower subclade. Such examples will be provisionally called here either "Pre-K1a10" or "Pre-K1a9" depending on whether or not they have 524 insertions. A Fluxus Network diagram is available at the K Project News tab[3] which would indicate that Pre-K1a9 branched off 195C; from there descended Pre-K1a10. Then K1a9 and K1a10 each descended from a sequence in their "Pre" version. It further appears that K1a10 was founded from a Pre-K1a10 haplotype which had predominantly two pairs of insertions, which might explain why two pairs is the modal value for K1a10. Pre-K1a10 has, by definition, no deletions or CRS variants; it has one to four pairs of insertions. The involvement of several steps in the process to the two end subclades may help explain why they have become almost fixed, one at the CRS variant and one with insertion pairs. An alternate tree structure which might be adopted in the future would have the 195C point labeled "K1a9," the existing K1a9 relabeled as "K1a9a," and the provisional K1a10 relabeled as "K1a9b." Pre-K1a9 and Pre-K1a10 would not fit into a K tree such as Behar's, since that one does not use the 524 insertions. The provisional K1a11 discussed above could be relabeled K1a10. While logical, this structure might cause confusion among those accustomed to the current and provisional designations. There is precedent for having two subclades below an unlabeled node; K1a7 and K1a8 are below an unlabeled point with the

[3] http://www.familytreedna.com/public/mtDNA_K

16234T mutation. However, in that case there are apparently no examples of sequences below the 16234T mutation that are not in either K1a7 or K1a8. An alternate theory might be that the 195C mutation occurred twice, once with and once without insertions. However, the original theory is more likely to be correct because K1a9 and K1a10 have two additional factors in common. Neither has defining coding-region mutations and the known examples rarely have private coding-region mutations. Also, neither subclade has any of the position 309 insertions which are common in other K subclades – an excellent example of a heteroplasmic variant being "fixed out." **Figure 1** shows the proposed placement of K1a10 and K1a11 on the K1a segment of the K tree.

Figure 1 is restricted to the section of the K tree under major subclade K1a which is defined by HVR2 mutation 497T. For the defining mutations of lower and 195C. Unlike the split below 195C in K1a where one main branch only has the CRS and deletions while the other one has only insertions, the K1b daughters both may have the CRS or insertions, but never the deletions. Since K1b2 has the two HVR mutations, it's much easier to identify. Its most common variant is one pair of insertions at 40%, closely followed by the CRS at 35%, with two insertions pairs third at 20%. There is one sequence with three insertion pairs. Altogether, insertions are found in 65% of K1b2.

Of K1b1's lower subclades, K1b1a is easily identified by HVR mutations 16319A and 152C, with a good representation in the K Project. Its majority variant is the CRS; 25% have one insertion pair. However, the MitoSearch examples were equally divided. K1b1b requires two additional coding-region mutations. There are no confirmed examples in the K Project, but the two on GenBank have one and two insertion pairs. K1b1c has several coding-region mutations and two HVR mutations, 94A and 16266T; but there are no examples of this so far in FTDNA's public databases. The one example on GenBank has CRS. There are three sequences from the K Project which can't be assigned to a lower subclade; they are shown as K1b+; all three have one insertion pair. Thus there are multiple points of conundrum below K1b. The CRS and one- and two-pair variants were inherited, with no need for *de novo* mutations to explain the existing haplotypes, except possibly for the three-insertion-pair example in K1b2.

The entire K haplogroup seems to be divided between two groups of subclades. One group, including K2, K1c, K1a1, K1a3, K1a6, K1a7, K1a8 and K1a11, almost never have the 524 insertions, but sometimes have the deletions. The other group, including K1b, K1a2, K1a4, K1a5 and the subclades under K1a+195C, include examples with each variant. In addition, there subclades K1a1 through K1a8 see the Behar K tree. There are two new provisional lower subclades added here. K1a10, defined by HVR1 mutation 16048A, is a sister subclade to K1a9 under HVR2 mutation 195C. K1a11 is defined by several HVR and coding-region mutations. "9bpD" is a sequence of nine coding-region deletions from 8281 to 8289. K1a7 and K1a8 are shown here next to K1a1 just to demonstrate the sharing of HVR1 mutation 16234T in all three subclades. The subclades in turquoise generally have a significant percentage of examples with position 524 insertions, while those in yellow do not.

The remaining major subclade is K1b, which is defined by the coding-region mutation 5913A, at another point of conundrum. K1b is split between K1b1, defined by three coding-region mutations 9962A, 10289G and 15946T, and K1b2 which is defined by coding-region transversion 12738G and two HVR mutations, 146C are many examples in K1a whose lower subclade cannot be predicted; including some with each of the variants.

Several possible general rules may be observed from the above table and discussion. One rule is that no lower subclade has examples with both deletions and insertions. Many subclades have only the CRS variant. Subclades K1a3a1, K1a4, K1a5, and K1b1b – all with few examples so far – and K1a10 (and its "Pre" cluster) have just insertions. Only one single-example subclade, K1c1a, has only deletions. The major rule observed is that in no subclade can we observe a missing variant between two numbers of repeats; that is, in no subclade do we find only the CRS and two pairs of insertions as the alternatives, or one pair and three pairs of insertions, etc. When there are three or more variants in a subclade, the highest percentage is always in the middle, with a drop-off on either side. If there are only two variants including the CRS, the CRS variant is usually the most frequent, except in K1a2 and K1a4a1, in which one pair of insertions is the dominant variant. Of the four major subclades, K1a, K1b, K1c and K2, the insertions are found primarily in the first two, while deletions are found in all of them except K1b. The higher the number of insertion repeats seen in the majority variant of a subclade, the greater the number of variants found in that subclade. The single subclade, K1c1a, which has deletions as the majority variant, has no other variant form. Subclades with the CRS as the majority variant have an average of 1.6 different variants; for one pair of insertions the average is 1.8; for two pairs it's 3.5. Apparently the mutation (replication slippage) rate increases along with the number of repeats, as it does for Y-STRs. With only one deletion variant and four insertion variants, and with the insertion variants appearing lower down the K tree, the trend over time appears to be with an increase in the number of repeats.

Insights from Previously Published Scientific Research Articles

Three articles have been published dealing specifically with the variants at position 524: Szibor et al. (1997) reported on three European populations and an African Bantu population; Chung et al. (2005) reported on 500 Koreans; and Szibor et al. (2007) reported on a study of 2,458 Germans. All three articles focused on the possible use of the position variants for forensic identification. None of these studies attempted to identify the haplogroups involved, so they are not specific to haplogroup K or any other haplogroup.

Szibor et al. (1997) stated that repeat polymorphisms are found in only two places in mtDNA. Other than position 524, there is a nine base pair deletion in the coding region, which has a lower variability than position 524. This set of deletions at position 8281-8289 appears in at least two places on the Behar K tree, but so far has only shown up in the K1a11 subclade in the K Project (See **Figure 1**). The position 524 variants are not as useful as chromosomal STRs for forensic purposes, being described as "moderately informative." The usefulness of 524 for forensic purposes comes from the high mtDNA count per cell as compared to chromosomal DNA. The populations examined were 396 Germans, 100 Hungarians, 191 Russians and 105 Cameroon Bantus. The three European populations had slight differences in the combination of variants found, each having a majority of the CRS variant with one pair of deletions and up to three pairs of insertions. The Cameroon population was significantly different, a majority having one deletion pair and two individuals having two deletion pairs; no insertions were found.

Chung et al. (2005) noted the similarity of these 524 positions to nuclear DNA STRs. In the study of 500 unrelated Koreans, the CRS variant was the most common. A few transitions or transversions were found within the repeats or in the flanking region (One example of such a flanking-region transition, 513A, has been observed in a K Project sequence which has one pair of insertions). They found three examples of length heteroplasmy – individuals with more than one variant. One had both the CRS variant and the deletions; one had the CRS and one pair of insertions; and the third had the CRS and one and two pairs of insertions. Again, FTDNA and GenBank have no method available to report length heteroplasmy. Chung refers to the "high slippage rate of dinucleotide repeats in STRs" and the "resultant stutter production . . . correlated to the length of repeat stretches . . . ." Also, the authors said that PCR "did not produce stutters up to 6 CA repeat units [one pair of insertions], while clones with 7 CA repeats [two insertion pairs] showed traces of stutter in the electropherograms." "Taken together, polymerase slippage is the driving force not only in stutter artifact genesis during PCR but also causing new mutations (plus/minus one repeat) in dinucleotide repeats." The authors acknowledged that these repeats ". . . help to understand the mechanism of mitochondrial evolution . . . ." That's as close as these papers get to discussing these positions as inherited DNA rather than their use for forensic purposes.

Szibor et al. (2007) studied 2,458 German samples, also focusing on forensic identity testing. They also found the CRS to be the most common variant, with up to three pairs of insertions, but only one pair of deletions. They found 34 individual with heteroplasmy, a higher rate than found in the Korean samples. The reason given for the higher rate was that the German population had more of the variants with higher repeats—"such alleles seemed to be prone to develop heteroplasmy." Also, "Heteroplasmy seems to be preferably bound to alleles with higher repeat numbers. This observation may reflect that $(CA)_4$ and $(CA)_5$ seem to be more stable than longer alleles." However, when studying one five-member family, they found three brothers with $(CA)_4/(CA)_5$ heteroplasmies, but their two cousins had just $(CA)_5$. A second family exhibited similar results. From these pedigree studies they concluded that even for these shorter repeats "the inheritance of $(CA)_n$ repeat heteroplasmy seems to be unstable. . . ." The study noticed the "distinctive racial differences concerning the $(CA)_n$ frequencies . . ." without attributing the differences to the ages of the mtDNA haplogroups. They concluded that "$(CA)_n$ heteroplasmy can appear or disappear during a few generations."

In **Table 3,** the Korean samples are from the Chung et al. (2005) article. The other samples are from the Szibor et al. (1997, 2007) articles. As above, tan denotes deletions, yellow denotes CRS, and turquoise denotes insertions.

Combined, the three studies found the same 524 variants in roughly the same percentages as is found in the SMGF database, except for the two deletion pairs in the Cameroon samples. What is not apparent from studying haplotypes in the K Project and GenBank databases is the degree of strict heteroplasmy. The finding in Szibor et al. (2007) that heteroplasmy is more common when the insertions have a higher number of repeats, leads one to believe that the insertions also have a higher mutation rate. That theory is supported by the fact that in contrast to the numerous K subclades with only the CRS variant, a high number of variants are found in subclades with the multiple insertion pairs. The unstable inheritance of the 524 variants noted by Szibor et al. (2007) in two pedigrees is somewhat disturbing. However, of the very few examples of related persons in the K Project, there are no known cases of differences at 524. It should be noted that the

Table 3. Position 524 Heteroplasmic Variant Percentages in World Populations

CA Repeats Country, Year	Sample Count	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	8 (%)	% With 524 Insertions
Korea 2005	500		37	61	1	1		2
Cameroon 1997	105	2	53	44				0
Hungary 1997	100		19	75	5	1		6
Russia 1997	191		1	81	8	1		9
Germany 1997	396		11	79	8	2		10
Germany 2007	2458		11	80	8	2	0.4	10

current paper only refers to the insertions being fixed in a subclade, such as K1a10–not any one particular insertions variant being fixed.

Do the results from the three papers observe the rules suggested by our study of Haplogroup K? Lower subclades – or any subclades of haplogroups at all – were not covered in these papers, so the rule that a lower subclade will not have both insertions and deletions does not apply. Certainly none of the populations showed a variant that skipped a step between different numbers of repeats. The highest percentage variant is always in the middle. No population had only two variants, so the rule that CRS would be dominant doesn't apply. However, in the Cameroon population, the deletions together outnumber the CRS. The rule that insertion variants have a higher mutation rate is discussed in the previous paragraph.

524 Variants in Other mtDNA Haplogroups

The population studies in the three articles above are interesting, but genetic genealogists are used to thinking more in terms of haplogroups instead of populations. So the next area of investigation was the comparable distribution of the 524 variants in the other mtDNA haplogroups. MitoSearch was used as a data source, as were websites of other mtDNA haplogroup projects. However, study of the lower subclades of the other haplogroups was generally beyond the scope of this article.

The CRS variant with five CA pairs probably was *not* the ancestral variant because all of the most deeply rooting haplogroups have deletion variants. All the admittedly few L0 examples on MitoSearch have the deletions. L1 has 88% deletions. L2, usually shown as a side-branch to the main line of the mtDNA tree, has 66% CRS, 28% deletions, with a small percentage of insertions. On the main trunk leading to the other haplogroups, only when L3 is reached does the CRS variant become the majority, but with a large minority of the deletions remaining. The insertions start making a small 2% appearance in L3. The major branch M is majority CRS. One of its North American sub-branches, C, is majority CRS, while another, D, is all deletions. Arriving at macro-haplogroup N, the CRS reaches 87%. In the haplogroup I side-branch the insertions reach one of their highest frequencies at 23%. Side-branch haplogroup A is back to 98% deletions, while the other side-branches from N are about 98% CRS.

When macro-haplogroup R is reached, the CRS variant is seen in 91% of samples, with small amounts of insertions and deletions. From R, side-branch F is back to 100% deletions; but others, B, J and T, are about 90% CRS. The main line down to H and the branch to V logically have a large majority of CRS, since the CRS is in haplogroup H. When U is reached directly from R, things get complicated. The plain U (or U*) on MitoSearch is 87% CRS, with a good representation of the insertions at 10%; but the range of variants in the subgroups of U is great. U3 is all CRS. U5 is 91%

CRS, while the small U7 is 100% deletions. U1 has 47% CRS, 40% deletions, and a respectable 14% insertions. U2 has a plurality in CRS and a few deletions; but its 41% insertions are among the highest of all haplogroups. U4 is 38% CRS, but has the highest insertion frequency of all at 60%. And last, K is about 69% CRS and 29% insertions.

So the deletions were probably, in fact, the ancestral variant in "mitochondrial Eve"; and the CRS became the majority by the time L3 was reached. (There is probably not enough evidence to determine whether the two-deletion-pair variant – only seen so far in the Cameroon sample – was even more ancient or was a replication slippage from an individual with one pair of deletions). With the interesting exception of haplogroup I, the insertions did not become prominent until U was reached on a side branch. Even then only a few of U's lower haplogroups have them in large numbers. Only in I and K does the second pair of insertions reach double-digit percentages. Oddly, U4 has a double-digit frequency in three pairs of insertions; but not of two pairs. Those two haplogroups are also the only ones known to have examples of four insertion pairs, while U* also has three pairs.

Again, the results from other haplogroups can be measured against the rules derived from haplogroup K. No lower subclades were studied, so it was not determined if any had both deletions and insertions. Four haplogroups only had one variant, two each for deletions and CRS. Two haplogroups, W and HV, skipped a variant – one insertion pair. When there were only two variants, L1, C and A had majority deletions, while V and U6 had majority CRS – reflecting probably the age of the haplogroups. If there were three or more variants, the majority or plurality one was toward the middle. I and K had the majority variant, CRS, off the center of the distribution due to the number of insertion variants.

Table 4 shows the 524 variant percentages in the major mtDNA haplogroups. The data for the table were collected, if available, from mtDNA haplogroup projects as listed on the World Families website. For haplogroups A, B, C and D, data were from the Amerind Founder Project. Data for haplogroups with "ms," such as L1ms%, was collected from MitoSearch. Haplogroups and certain major subclades are listed approximately in order of founding from mitochondrial Eve, with haplogroups in side-branches kept together. For example, haplogroups C and D are listed under M before the main line resumes with N. Again, the insertions are in tan, the CRS variant is in yellow, and the insertions variants are in turquoise. **Figure 2** presents the same data in a phylogenetic tree format, but with the same colors used only for the most common variant in each haplogroup.

Areas for Future Research

Position 524 Variants

The 524 insertions, in combination with other mutations, might be useful in determining the age of some subclades. It might be especially helpful if the other mutation were one of the more stable ones. For example, assume that the original provisional K1a10 with 16048A, which only appears at one place in K, had two pairs of 524 insertions. Therefore, the 29% of K1a10 with one pair or three pairs of insertions had to have mutated since the founding of the subclade. However, there is still the question of whether the one and three pair variants were inherited as minority variants which later became dominant or were due to *de novo* replication slippages.

Where possible, attention should be given to any differences in the 524 variants found in related persons in K and other haplogroups to see if there are any similarities to the differences between family members found in Szibor et al. (2007).

Other Haplogroup K Heteroplasmic Mutations

Of course, 524 is not the only position which presents a problem in determining the method and origin of mutations. There are other examples within haplogroup K; a few will be discussed below. Possible examples in other haplogroups have not been investigated; that is beyond the scope of this study.

A group of such mutations is found below K1a1. Mutation 16234T defines K1a1b1a, where it appears to have become fixed. The Behar K tree also shows this mutation in a parallel example in K1a1b1. However, the one sample from this branch in the K Project does not have 16234T, so the mutation is probably not fixed there. 16234T is also the defining mutation connecting K1a7 and K1a8; perhaps there is a point of conundrum connecting those two subclades with K1a1. Mutation 16223T is most commonly found two steps down from the K1a1b1a modal, but also other places in K1a1. The third mutation of the group, 114T, is found in most examples of K1a1b1a. It is also found in some examples of K1a1a and K1a1b; leading to the possibility that it was a *de novo* mutation at the K1a1 point. A less common fourth mutation, 16092C, may be another member of this group; it probably mutated *de novo* below K1a1 and now appears in various combinations in its lower subclades and branches.

Mutation 16051G appears in several examples of K1a10. It also appears in a few K1a sequences which have not yet been given a lower assignment. Perhaps future results will help determine at what point this mutation originated.

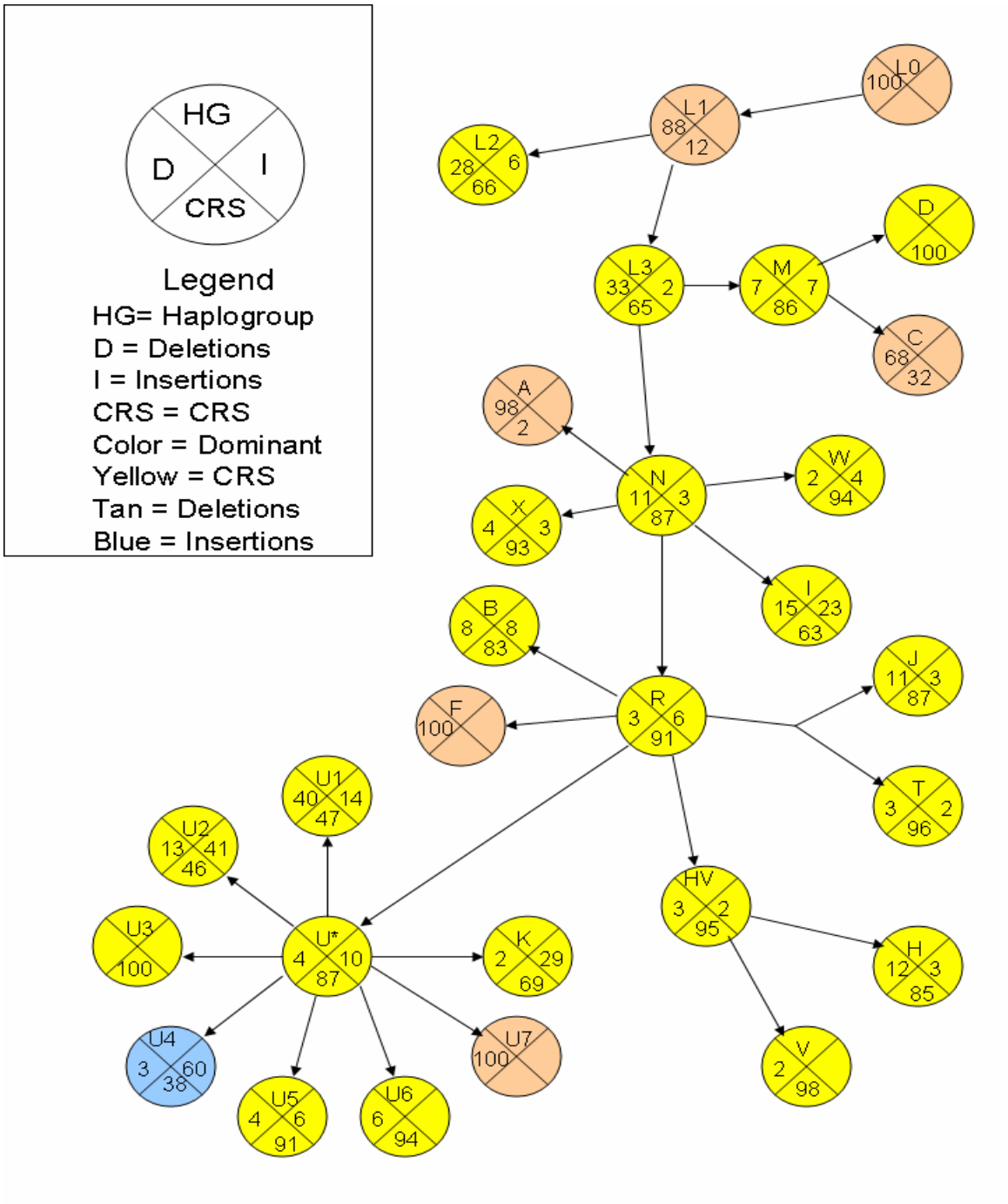

Figure 2. Mitochondrial Phylogenetic Tree with 514-524 Status

The Behar K tree has one sequence with 16266T in K1b1c, while there is one in K1b2 in the K Project. Did these mutations occur at the K1b split and have phylogenetic significance for these subgroups, or are these mutations just independent *de novo* mutations? Since the same mutation occurs in a K1a sequence in the K Project, this may just be a mutation that frequently occurs *de novo.*

Categories of Mutations

In general, the various types of mutations need further study. In addition to positions such as 524 and those just discussed, there are several other categories. Some, such as 497T, which can only be found once in the phylogenetic tree, or 498- and 16048A, which might be thought of as *unique event polymorphisms within K,* are at one end of the scale. Those mutations are at least the equal of coding-region mutations in their ability to define subclades. At the opposite end of the scale, insertion 309.1C shows up in virtually every K subclade and thus has no use in defining subclades – the notable exceptions being K1a9 and K1a10, where it is fixed out. 16093C is also common in many subclades. Tully et al. (2000) discusses position 16093 in great detail. Most other mutations fall somewhere between these two extremes. Recurrent mutations 146C, 152C and 195C are used to define more than one subclade and also appear in other subclades. Each mutation affected by heteroplasmy has its own independent system or parallel layer. For many such mutations, the point of its origin may often be determined. As an example, 114T probably was a *de novo* mutation at the K1a1 node, since it appears only below that point.

Position 524 just happens to be the most complicated of all, mostly due to its STR-like qualities. Only when an individual haplotype is sequenced is the exact combination of mutations revealed. A flat, two-dimensional phylogenetic tree does not do justice to the complex heteroplasmic inheritance pattern of mtDNA. Perhaps a tree with all the reticulations left in would be preferable to the standard published trees. With several parallel systems in operation, the idea of simply adding up the number of mutations to attempt to determine the age of a subclade or K itself – or any other haplogroup – may not be very useful.

As mentioned above, strict heteroplasmies are not reported by FTDNA and some other testing companies. However, the recent paper "The Genographic Project Public Participation Mitochondrial DNA Database" (Behar et al. 2007) reveals that FTDNA's testing laboratory is detecting heteroplasmies. The 1,759 HVR1-only K entries in "Dataset S1" show 40 haplotypes with one heteroplasmy, always marked "N" instead of one of the IUPAC codes, which is an acceptable practice according to Scientific Working Group on DNA Analysis (2003). Fourteen different positions are involved, mostly with one haplotype each. 16093N is the most common at 19 entries (1.1%). There are 400 (22.7%) with the mutation 16093C; the rest are CRS. There are no known reports of the Genographic Project actually listing 16093N on a user personal page. Many of those with 16093N no doubt have transferred their results to FTDNA where they are either reported as 16093C, or if CRS not listed at all (There has been a recent report of an FTDNA customer receiving an e-mail stating that a heteroplasmy had been found at one position). Tully et al. (2000, pp 435, 439) noticed that almost all (11 of 13) heteroplasmies at 16093 had C as the majority variant. So perhaps most of the 19 Genographic 16093 heteroplasmies in K would transfer to FTDNA as 16093C. However, there may be some variation by haplogroup. In the K Project 22% of the entries have 16093C, while Tully et al. (2000, Table 2) found only 6% and SMGF has 5.5%. Tully also noted that 16093 was the most heteroplasmic position in the original HVR1.

A recent paper (Irwin et al. 2007, Tables S1 and S2) reported the HVR sequences for 400 Northern Greeks and Greek Cypriots. This paper found 32 examples of heteroplasmies, all denoted by the IUPAC codes. In haplogroup K examples they reported 146Y (C/T combination), 16189Y and 16093Y twice. They also found examples of 16311Y, 16129R (A/G combination), 195Y and 152Y, but in other haplogroups, not K.

Of course, research similar to the above for haplogroup K could be performed on any of the other haplogroups.

Conclusions

Position 524 has not been well-studied in the past, partly due to its location in the old HVR3, which has not often been tested. Even very recent mtDNA studies, such as that of the National Genographic Project (Behar et al., 2007), often only report on HVR1 and perhaps haplogroup-defining control-region SNPs.

The study of 524 variants is hindered somewhat by the great variety of naming systems, a situation that is not likely to improve anytime soon.

More than any other mtDNA position, 524 mutates in a manner similar to that of nuclear DNA (including Y-DNA) STRs: replication slippage. As in Y-DNA, this mtDNA STR mutates at a faster rate than SNPs. But unlike Y-DNA, 524 and other mtDNA positions may also "mutate" by the seemingly random changes caused by heteroplasmy. "Mutation rates" for mtDNA must take into consideration *de novo* mutations and inherited heteroplasmic variants. The random selection process

leads to complications beyond the capabilities of standard two-dimensional phylogenetic trees.

One pair of deletions–a total of four CA repeats–was probably the original 524 variant, as shown by its predominance in the oldest L haplogroups. Working along the mtDNA tree, the CRS variant eventually became dominant, reaching its highest overall levels logically at haplogroup H. The insertion variants only came into prominence tens of thousands of years later on branches from N and U. Insertions are only dominant in a few haplogroups and subclades, while a variant with two pairs of deletions has not been reported except in one African population. The lower the 524 repeat number, the lower the apparent mutation rate. The rate of replication slippage appears to accelerate for the higher insertion variants, with the trend being toward more repeats as the distance from mtDNA Eve increases. However, even younger haplogroups such as U3 and U7 may become fixed at either the CRS or deletions. The result of this bias toward upward replication slippage is that are very few examples (Cameroon Bantus) of an extra deletion pair from the probable original single deletion pair; but there are many examples of the CRS and up to four insertion pairs. However, the provisionally-named subclade K1a10, which apparently had two insertion pairs as the dominant variant, has many examples where either there has been a *de novo* decrease in repeats or that it also inherited a substantial minority of one- and three-insertion-pair variants. The fact that no examples with the CRS have been found may be evidence of the second alternative.

Many K subclades are fixed at one variant, most commonly the CRS. Rare mutations are almost certainly *de novo*, such as the single known examples of one insertion pair in K1a1b1a and K1c2 where there are no insertion pairs in higher or sister subclades.

Position 524 is useful for predicting subclades from HVR-only results, since there are subclades which almost always have or almost never have a particular variant. Often this is most helpful in conjunction with other listed mutations. It should be noted that almost exactly half of the haplogroups in **Table 4** have greater than 90% in one variant. It is difficult to say that 524 should not be part of the definition of a haplogroup when 100% of the examples are fixed at the same variant as appears to have happened in five of the haplogroups.

Each mtDNA haplotype is composed of mutations mostly inherited from its ancestors, some in a heteroplasmic manner, with *de novo* mutations rarer than might first appear. In the case of the 524 variants, both the CRS variant and at least one pair of variants each with deletions and insertions were probably inherited by the founder of K. Almost all K haplotypes received their 524 variant from this inheritance rather than from *de novo* replication slippages. Even subclades with higher levels of insertion repeats have inheritance histories. The origin of each repeat level might be determined from greater study of more data. If only one variant was inherited by a person, that variant will be fixed in that person and in her descendants. One the other hand, if a variant is not among a combination of variants inherited by a person, it will be "fixed out" in the descendants. Further mutations would require *de novo* replication slippages.

Electronic Database Information

Argus BioSciences LLC
http://www.argusbio.com/

FamilyTreeDNA MitoSearch database
http://www.mitosearch.org/

GenBank
http://www.ncbi.nlm.nih.gov/Genbank/

Ian Logan mtDNA website
http://www.ianlogan.co.uk/mtDNA.htm

Mitomap: A human mitochondrial genome database
http://www.mitomap.org/

mtDNA Haplogroup K Project website
http://www.familytreedna.com/public/mtDNA_K/

Relative Genetics, Inc.
http://relativegenetics.com/

Sorenson Molecular Genealogy Foundation
http://www.smgf.org/

World Families links to other mtDNA haplogroup project websites:
http://worldfamilies.net/reference_mtDNA.html

References

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

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

Behar DM, Metspalu E, Kivisild T, Achilli A, Hadid Y, et al. (2006) The matrilineal ancestry of Ashkenazi Jewry: Portrait of a recent founder event. *Am J Hum Genet*, 78:487–497.

Table 4. Position 524 Heteroplasmic Variant Percentages from FamilyTreeDNA Data by Haplogroup

Haplogroup	Counts	522-, 523- %	CRS %	524.1, 524.2 %	524.3, 524.4 %	524.5, 524.6 %	524.7, 524.8 %	524 Insertions Total %
CA Repeats		4	5	6	7	8	9	
L0ms%	6	100						0
L1ms%	60	88	12					0
L2ms%	109	28	66	1	5			6
L3%	54	33	65	2				2
M%	14	7	86	7				7
D%	5		100					0
C%	31	68	32					0
N%	38	11	87	3				3
I%	40	15	63	3	15	3	3	23
W%	50	2	94		4			4
X%	67	4	93	3				3
A%	47	98	2					0
Rms%	34	3	91	6				6
B%	12	8	83	8				8
Fms%	5	100						0
J%	164	11	87	2	1			3
T%	112	3	96	2				2
HVms%	105	3	95		2			2
V%	64	2	98					0
H%	510	12	85	3				3
U*ms%	53	4	87	6	2	2		10
U1ms%	15	40	47	7	7			14
U2%	39	13	46	36	5			41
U3ms%	27		100					0
U4%	37	3	38	41	8	11		60
U5%	222	4	91	5	1			6
U6ms%	18	6	94					0
U7%	6	100						0
K%	321	2	69	16	12	1	0.3	29

Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, et al. (2007) The Genographic Project public participation mitochondrial DNA database. *PLoS Genet* 3:e104.

Carter, R. (2007) Mitochondrial diversity within modern human populations, *Nucleic Acids Research*, 35:3039-45.

Chung U, Lee HY, Yoo JE, Park MJ, Shin KJ (2005) Mitochondrial DNA CA dinucleotide repeats in Koreans: the presence of length heteroplasmy. *Int J Legal Med*, 119:50-53.

Dupuy BM, Stenersen M, Egeland T, Olaisen B (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum Mutat*, 23(2):117-24.

Endicott P, Sanchez J, Metspalu E, Behar D, Kivisild T (2007) The unresolved location of Ötzi's mtDNA within haplogroup K. *Am J Phys Anthropol*, 132:590-591, discussion 591-3.

Irwin J, Saunier J, Strouss K, Paintner C, Diegoli T, Sturk K, Kovatsi L, Brandstätter A, Cariolou MA, Parson W, Parsons TJ. (2007) Mitochondrial control region sequences from northern Greece and Greek Cypriots. *Int J Legal Med.* [Epub ahead of print]

Kivisild T, Shen P, Wall DP, Do B, Sung R, et al. (2006) The role of selection in the evolution of human mitochondrial genomes. *Genetics*, 172:373–387.

Rollo F, Ermini L, Luciani S, Marota I, Olivieri C, Luiselli L (2006) Fine characterization of the Iceman's mtDNA haplogroup. *Am J Phys Anthropol*, 130:557-564.

Scientific Working Group on DNA Analysis Methods (SWGDAM) (2003) Guidelines for mitochondrial DNA (mtDNA) nucleotide sequence interpretation. Forensic Science Communications.

Shoubridge EA, Wai T (2007) Mitochondrial DNA and the mammalian oocyte. *Curr Top Dev Biol*, 77:87-111.

Sigurðardóttir S, Helgason A, Gulcher JR, Stefansson K, and Donnelly P. (2000) The mutation rate in the human mtDNA control region. *Am J Hum Genet*, 66:1599-1609.

Smigrodzki RM, Khan SM (2005) Mitochondrial microheteroplasmy and a theory of aging and age-related disease. *Rejuvenation Research*, 8:172-198.

Szibor R, Michael M, Spitsyn VA et al (1997) Mitochondrial D-loop 3' (CA)n repeat polymorphism: optimization of analysis and population data. *Electrophoresis* 18:2857-2860.

Szibor R, Plate I, Heinrich M, Michael M, Schöning R, Wittig H, Lutz-Bonengel S (2007) Mitochondrial D-loop (CA)(n) repeat length heteroplasmy: frequency in a German population sample and inheritance studies in two pedigrees. *Int J Leg Med*, 121:207-213.

Tully LA, Parsons TJ, Steighner RJ, Holland MM, Marino MA, Prenger VL (2000) A sensitive denaturing gradient-gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *Am J Hum Genet*, 67:432.

Turner A (2006) `Satiable curiosity: Now you see it, now you don't: Heteroplasmy in mitochondrial DNA. *J Genet Geneal*, 2(1):iv-v.

Wilson, MR, Allard MW, Monson KI, Miller KWP, Budowle B, (2000) Further discussion of the consistent treatment of length variants in the human mitochondrial DNA control region. *Forensic Science Comm*, 4:#4.