Journal: www.jogg.info
Originally Published: Volume 6, Number 1 (Fall 2010)
Reference Number: 61.007

A REFERENCE DATABASE TO SUPPORT ANALYSIS OF MTDNA HAPLOGROUP N, ITS DESCENDANT HAPLOGROUPS, AND ASSOCIATED CLADES

Author(s): Jim Logan and T. Whit Athey

A Reference Database to Support Analysis of mtDNA Haplogroup N, its Descendant Haplogroups, and Associated Clades

Jim Logan and Whit Athey

Abstract

GenBank (2010) is a collection of publicly available DNA sequences maintained by the National Center for Biotechnology Information. It currently contains about 7000 homo sapiens mtDNA 'full genome sequences', although some of these are complete in the coding region only. Abstracts of these sequences that list differences from the CRS are available through Ian Logan's Checker web site (Logan I, 2010). This paper describes the selection of sequences from this source that belong to Super-Haplogroup N, its subordinate haplogroups, and their clades, and how these sequences were organized into a matrix reflecting the current consensus about the phylogeny of homo sapiens mtDNA. Features of this matrix are described, the content analyzed, and a number of issues articulated relative to the use of various alleles in clade definitions. Illustrated herein is also a scheme for quantitatively assessing the quality of alleles used in the clade definitions. Products of the application of this scheme, in turn, are used as major components of a semi-automated allele encyclopedia. Finally, there is a description of the tools used to develop and maintain the matrix, perform ongoing analysis, and produce the core structure of an allele encyclopedia.

Introduction

The formal proposal for the naming of mtDNA haplogroups and their clades is only a little over a dozen years old (Richards et al, 1998), but it has seen tremendous growth and changes in those dozen years, and some would say it has become exceedingly complex. However, much of this complexity is due o the increased availability of complete mtDNA sequences and the increased coverage of the world's populations; revisions reflect progress in the field. The gathering of data and the associated analyses have taken many paths and thus the product is large and hard to use by nonacademic project administrators who are trying to answer questions from their project members or others who are simply trying to learn more about their personal mtDNA. The availability of a series of haplogroup-oriented handbooks and an encyclopedia of mtDNA SNPs would be of great value to these administrators or to individual researchers. This paper proposes a systems approach to the development and refinement of such products. It is based on the lead author's experience in analyzing and reporting on Haplogroup J and observations made in reviewing the large phylogeny as presented in PhyloTree (van Oven and Kayser, 2009). It also includes some reference material for use in understanding details of the approach. In the current document, considerable space is also devoted to formulating issues related to future work. It is anticipated that as the proposed work proceeds, most of these issues will be resolved and thus converted into developmental or maintenance guidelines. The document as a whole may be edited and ultimately become a handbook useful to anyone who is conducting mtDNA research or simply investigating their own mtDNA ancestry. It could then be supplemented by a number of documents addressing details of specific haplogroups.

Previous Developments

A 1987 study concluded that current human mtDNA stem from one woman who was postulated to have lived about 200,000 years ago, probably in Africa (Cann, 1987). This study used a network analysis technique to analyze mtDNA from 147 people and drawn from five widely dispersed geographic populations. Thus was born the concept of a “Mitochondrial Eve” and its support for the “out of Africa” hypothesis. Many studies since,

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

Received: July 8, 2010; Accepted: Nov. 26, 2010; Published: Dec. 19, 2010.

using a various populations, have come to similar conclusions.

One of the first mtDNA population studies used an early form of SNP testing (restriction fragment length polymorphism, or RFLP) to analyze blood samples from 167 Native American subjects from five widely dispersed populations – three in North America, one in Central America, and one in South America (Toronni, 1992). This study produced the first grouping of mtDNA haplotypes that roughly correspond to what are now known as mtDNA Haplogroups A, B, C, and D. The direct sequencing of mtDNA has since become feasible and more recent studies have compared all or part of the mtDNA sequences, providing much greater resolution. With this increase in detail comes improvement in the ability to classify the DNA itself, leading to the approximately 1800 clade definitions found in the current version of PhyloTree (van Oven M, Kayser M, 2009), based on over 6700 complete or near complete full genome sequences available in GenBank (GenBank, 2010).

Few studies have been undertaken that look at single haplogroups in depth. The first such study was done for Haplogroup J as a Master's thesis (Serk, 2004). Ms Serk worked with 712 samples that had been assayed for eleven loci in the coding region plus sequence data from HVR1. Although she developed a top-level phylogeny, it was later discovered that these eleven loci are not adequate for correctly differentiating major clades of the haplogroup..

More recently a survey of the literature pertaining to Haplogroup J was carried out and published in the Journal of Genetic Genealogy (Logan JJ, 2008a). That paper presented a phylogeny of Haplogroup J based on 111 full genome GenBank sequences (Logan I, 2010). This phylogeny was used as a basis for critiquing the existing classification system based on HVR1-only results and the paper also presented recommendations for changes to the existing Haplogroup J phylogeny. The source of definitions of that phylogeny were presented in a matrix with one column for which was different from the Cambridge Reference Sequence (CRS) and one row for each sequence with their CRS differences carefully aligned. The matrix had been transformed to show the clear patterns in the data from which was inferred evolutionary changes in the sample population to produce the phylogeny. A second Haplogroup J paper was published by the same author with a more comprehensive analysis of the mtDNA itself and presented both initial calculations to estimate the age of origin of the major clades and a discussion of available data about current geographic distribution of the haplogroup (Logan JJ, 2008b). The matrix presented as supplementary data for that paper showed an updated phylogeny based on 156 sequences then available from GenBank, including 118 full genome sequences and 38 that were complete only for the coding region. The third paper in the series (Logan, 2009) further refined the phylogeny based on 291 sequences and was able to more than double the number of defined clades. The supplementary data for that paper has since been updated at the publisher's site; the currently posted matrix is Release 7 dated 28 October 2009.

To ensure as much objectivity as possible, the phylogeny in these three studies were developed directly from sequence data by applying maximum parsimony criteria as a basis for transforming an alignment matrix to reveal relationship patterns. From these relationship patterns, evolutionary sequences were inferred. There were also conservative rules applied in the acceptance of a pattern as defining a clade. For example, the general rule was to require three or more instances of a pattern before it was accepted as a clade. There were a few exceptions using only two instances where the branching was obvious and these two instances came from different researchers or different populations and thus it could be reasonably assumed that these samples were independent, e.g., not from close relatives. Only after developing the structure of the phylogeny, was the literature consulted for prior observations. Thus names were applied to the clades so as to reflect any perceived consensus. That is, clade definitions were developed purely from the sequence data, incorporating all available sequences and not influenced by precedence in the literature.

Concurrently with this phylogeny development for Haplogroup J, van Oven and Kayser (2009) developed PhyloTree, a broad scoped phylogeny based largely on existing literature. Acting both as curators of the mtDNA phylogeny data as presented in the literature as well as researchers identifying new patterns, they have been maintaining PhyloTree and making it publicly available; as of November 2010 it is in its tenth release. The second release of PhyloTree (dated August 2008) included citations to the Spring 2008 paper of Logan. Subsequent to that time there has been a collaboration between the lead author and the PhyloTree team. Although some differences of opinion remain concerning nomenclature the groups agree on the basic structure of the phylogeny. In their own work, the authors of PhyloTree claim to have similar conservative rules as those described here, but they have also tried to incorporate and present all definitions found in the literature, even when some definitions were based on a single sample. Some of the definitions are no doubt incorporating "private mutations." Furthermore, new relationships between existing clade definitions have been identified and named, introducing new nodes, while retaining previous node names in the tree structure; branches were moved as appropriate to be consistent with these new nodes representing these new relationships. This has resulted in some unconventional naming of patterns and the fragmentation of previously defined haplogroups. However, adoption of the approach outline here should facilitate the defining and naming of branches.

Work on Haplogroup J is continuing but the main branches of the J tree have most likely already been discovered. To provide a broader perspective for phylogenetic research, the scope of this research has been expanded from Haplogroup J to include all of its parent, Super-Haplogroup R, and ultimately all of superhaplogroup N. At least for the foreseeable future, Haplogroup M and its descendent haplogroups and all of the original African haplogroups are not being considered, but may be added later or a separate matrix developed. Figure 1 provides an illustration showing the general evolutionary relationships of primary haplogroups within this expanded scope, the approximate location of where they may have originated, and a rough indication of migration paths between them. This chart is intended to show only a very general perspective, and thus the true complexity of migrations are not represented. For example, this chart does not show migrations of Haplogroup N directly north into Siberia or south into Australia, nor does it show the major migrations of Haplogroup B into both North and South America and throughout the Pacific.

The starting point for the expanded research was the 15 March 2010 downloading of 3584 mtDNA sequences abstracted from GenBank which appear to be representative of Haplogroup N and its descendents. A phylogeny matrix was constructed from these records and transformed in conformance to the currently posted PhyloTree. Because of the fragmentation of the idealized haplogroup tree an index of major segments of the tree to their ancestors has been found useful and is presented here as Figure 2.

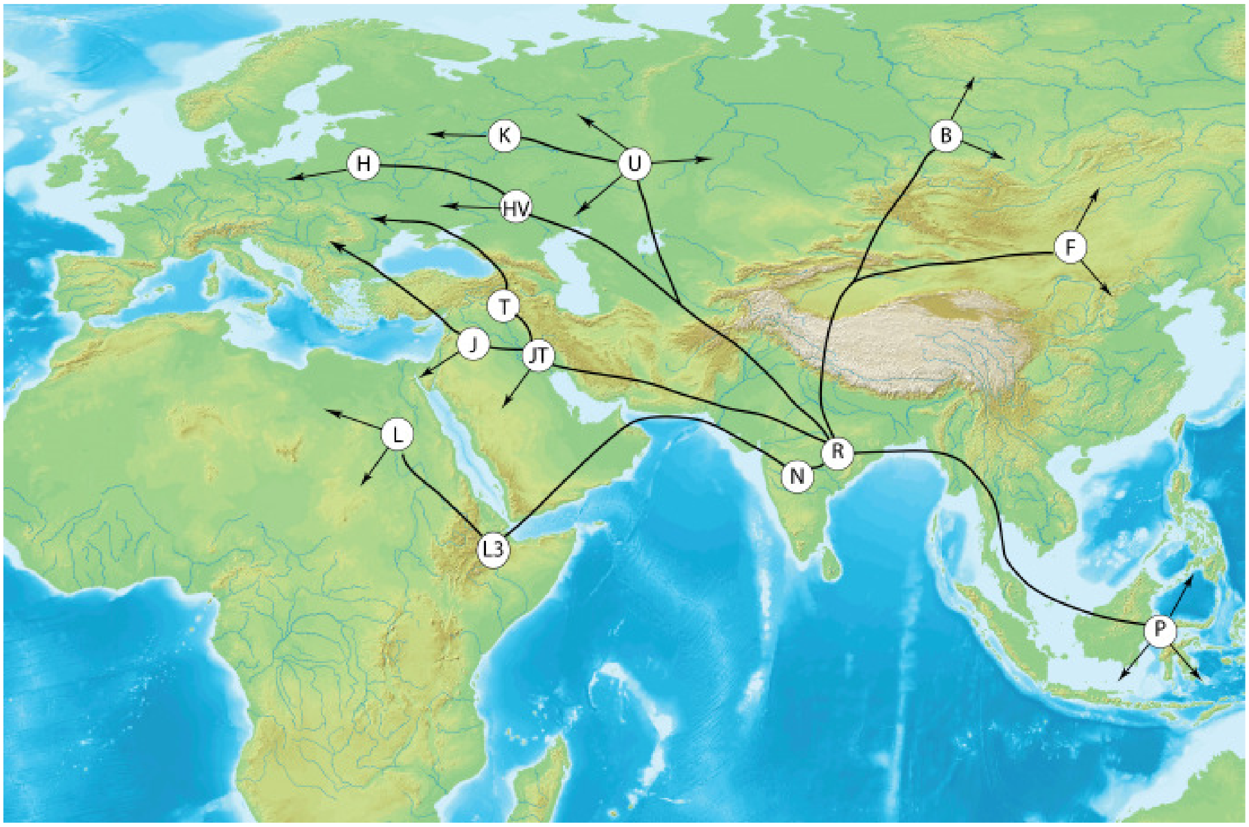


Figure 1: Map showing the major haplogroups derived from superhaplogroup R and illustrating possible regions of their origin. Note that the southern Asia location for the birth and/or maturity of N and R is based on the lead author's interpretation of available research results and that the southern out-of-Africa route is most consistent. There is not yet consensus on this and the chart will be revised as consensus is reached. (Base map used under the terms of the GNU Free Documentation License, Version 1.2 published by Free Software Foundation.)

Segment Name	<--------------------------- Ancestors --------------------------------->						
A						A	N
B4				B4	B4'5	R	N
B5				B5	B4'5	R	N
B7				B7	R11'B7	R	N
F				F	R9	R	N
H			H	HV	R0	R	N
HV1'6			HV1'6	HV	R0	R	N
I				I	N1	N1'5	N
J			J	JT	R0	R	N
K	K	U8b'k	U8	U2'3'4'7'8	U	R	N
N1'5							N
N2							N
N5					N5	N1'5	N
N9							N
N13'14'21'22						N13'14'21'22	N
O						O	N
P					P	R	N
R (misc)					R (misc)	R	N
R0a				R0a	R0	R	N
R1					R1	R	N
R11				R11	R11'B7	R	N
R2				R2	R0	R	N
R5'6'7'8					R5'6'7'8	R	N
R9				R9	R9	R	N
S						S	N
T			T	JT	R0	R	N
U1				U1	U	R	N
U2			U2	U2'3'4'7'8	U	R	N
U3			U3	U2'3'4'7'8	U	R	N
U4		U4	U4'9	U2'3'4'7'8	U	R	N
U5				U5	U	R	N
U6				U6	U	R	N
U7			U7	U2'3'4'7'8	U	R	N
U8a		U8a	U8	U2'3'4'7'8	U	R	N
U8b	U8b	U8b'k	U8	U2'3'4'7'8	U	R	N
U9		U9	U4'9	U2'3'4'7'8	U	R	N
V		V	HV0	HV	R0	R	N
W					W	N2	N
X						X	N
Y					Y	N9	N

Figure 2: Index of major components (blocks) of the mtDNA haplogroup tree as they lead back through their ancestral nodes to Haplogroup N.

In the systems engineering approach taken here, we consider the data in a top-down manner. Rooting analyses in the context of an integrated phylogeny oriented database is inherently top down. This contrasts, for example, with the approach of searching the literature to find clade definitions and trying to integrate them into a single set of definitions that define a tree. Since this top down database for all of Super-Haplogroup N and its components is maintained in spreadsheet form, the zoom feature can be used at any time to zoom out to see the forest, or zoom in on the trees down to their individual branches and leaves.

The Phylogeny versus a Classification Tree

An mtDNA phylogeny can be characterized as a set of relationships typically drawn as a tree structure describing inferred evolution of mtDNA from a single most recent common ancestor (sometimes called Mitochondrial Eve) to a diverse set mtDNA sequences that exists in the current human population, where the root of this tree is the inferred most recent common ancestor, the path from one node to the next is labeled by one or more polymorphisms defining differences between nodes, and the terminal points are the individual mtDNA sequences in the sample dataset. Node points on the phylogeny are also each given unique names assigned in hierarchical with the extreme nodes being unique identifiers for the specific sequences. The collection of all mtDNA sequences encompassed by traversing all links from a node out the tree to the end nodes is called a clade, although larger clades are sometimes called haplogroups or even superhaplogroups.

This paper is concerned with the development and maintenance of such an mtDNA phylogeny. On the other hand, if we maintain the topological relationships of this phylogeny and keep the names on the nodes but draw the tree using the CRS as the root, we then have a classification tree. The formal description of any node on that tree (representing a clade in the phylogeny) is then the list of all differences found on the links that we traverse when we start from the CRS move to that node. For Reference an illustrative classification tree is presented in Figure 3. Note that the root of this tree is labeled H2c2a which is the classification of the CRS. The formal definition of a clade is thus the set of differences from this reference sequence to the corresponding node. For example Haplogroup H can be defined as the set of differences at (263, 8860, 15326, 750, 4769, 1438)

The path length to a node (and thus the formal definition of a clade) can become quite long. As an extreme example, the path leading to Mitochondrial Eve (included with African Clades at the lower left) is defined by some fifty polymorphisms plus the absence of a difference at 15301 which has apparently mutated back to the reference value (Logan I, 2007). For many haplogroups or their clades, such as Haplogroup J taken in isolation, these long definitions present no problem since the characterization of the haplogroup as a whole can be taken as a base set of differences and its clades simply characterized by additional differences named on the links. A different kind of problem is illustrated in Figure 3 where the progression from H2a2a to H2a2 to H2a to H2 and finally to H goes up the phylogeny hierarchy rather than down as that phylogeny as commonly presented. Further illustration of possible confusion can be seen in the definition of H2a1 where the straight forward definition is the set (261, 8860, 15326, 750, 951, 16354) but if one starts from the standpoint of the Haplogroup H, the definition clade H2a1 is derived by adding the polymorphisms at 951 and 16354 but removing the 4769 that is included in the definition of H2 above. Extreme care must be taking in our analysis, especially of those clades in Haplogroup H.

Although the organization of mtDNA sequences in this paper are presented in accordance with a consensus phylogeny, the criteria of selection of data to be included in the database and the organization of that data are necessarily based on the classification tree. Presented in this way, it is easy to see that virtually all full mtDNA sequences other than Haplogroup H will have the common set of differences used for defining that haplogroup plus others. For example, all sequences in Haplogroups R, U, N, L3, et. will have these same six Haplogroup H differences plus those at 2706, 7028, 14766, 73, and 11719.

Source of mtDNA Data

It is instructive to take a closer look at the data available to support this research and the manipulation of this data. The process also leads to the identification of a number of issues to be used as the basis for establishing research guidelines.

The direct source of data is Ian Logan's abstracts of mtDNA sequences as they are added to GenBank and maintained in Checker scripts (Logan I, 2010). As seen from the above classification tree and starting with a complete set of full mitochondrial sequences (FMS) from Checker, the criteria for generating a dataset to study superhaplogroup HV (one of the most complex set of relationships in the entire tree), the criteria would be simply to eliminate all sequences that contain the difference 14766. For the broader superhaplogroup N and its descendants the criteria would be to eliminate all sequences that contain the set of differences at 8701, 9540, 10398, 10873, and 15301.

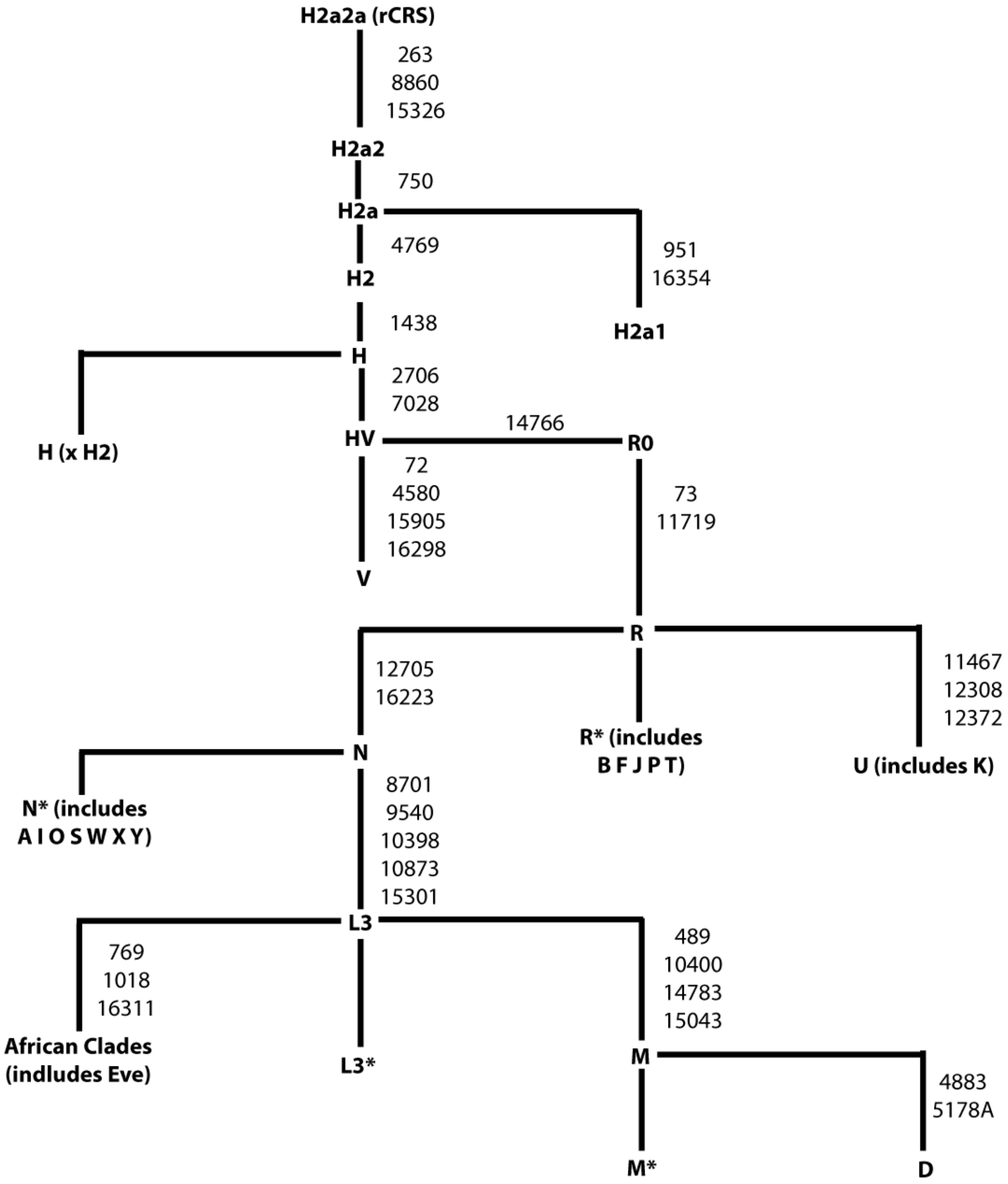


Figure 3: Illustrative nodes and links from the top-level mtDNA classification tree based on current consensus phylogeny (van Oven and Kayser, 2010).

The base dataset for the current working matrix for Haplogroup N was created by running a Python script against the Checker the above criteria for superhaplogroup N. However, since each sequence being examined is subject to random mutation, we relaxed the criteria and conservatively rejected only records that have two or more of these alleles and selected the rest. This is a compromise since there are no doubt a few sequences in the resulting database that are not truly Haplogroup N, but we believed that this was a better alternative than to reject sequences which may be crucial to defining a clade. This process resulted in the selection of 3584 records from the 6716 available in Checker as of 15 March, 2010. A Python script was then used to perform a multiple-sequence alignment to produce a matrix with one column for each record selected and one row for each unique difference from CRS observed in any of the records selected. The base matrix contained 4769 such distinct alleles. The occurrences of these alleles were also totaled and

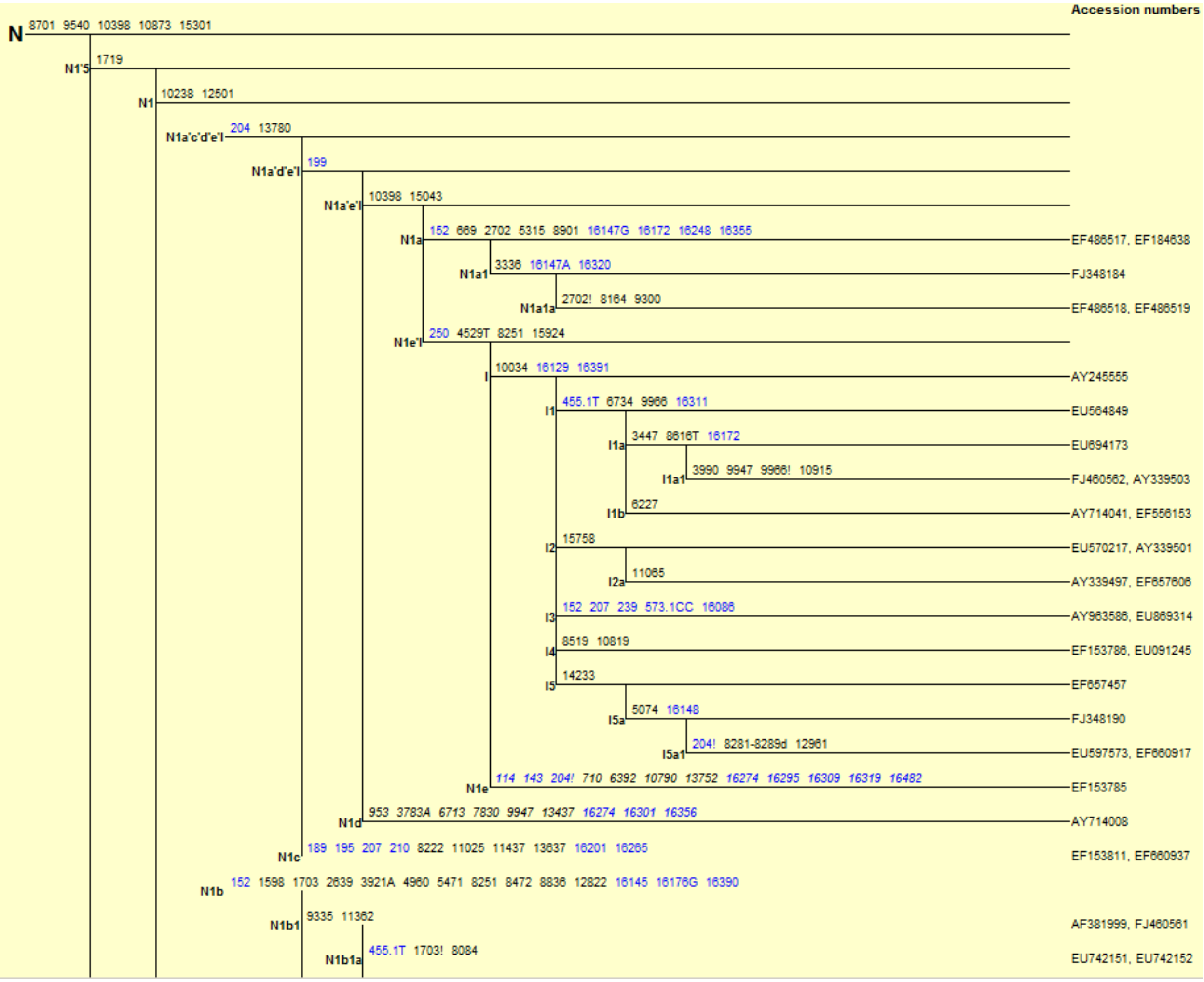


Figure 4: A screenshot showing the beginning of the PhyloTree definitions for the clades of mtDNA Haplogroup N.

provided in a separate column and the rows were then sorted in rank order based on the counts to facilitate review.

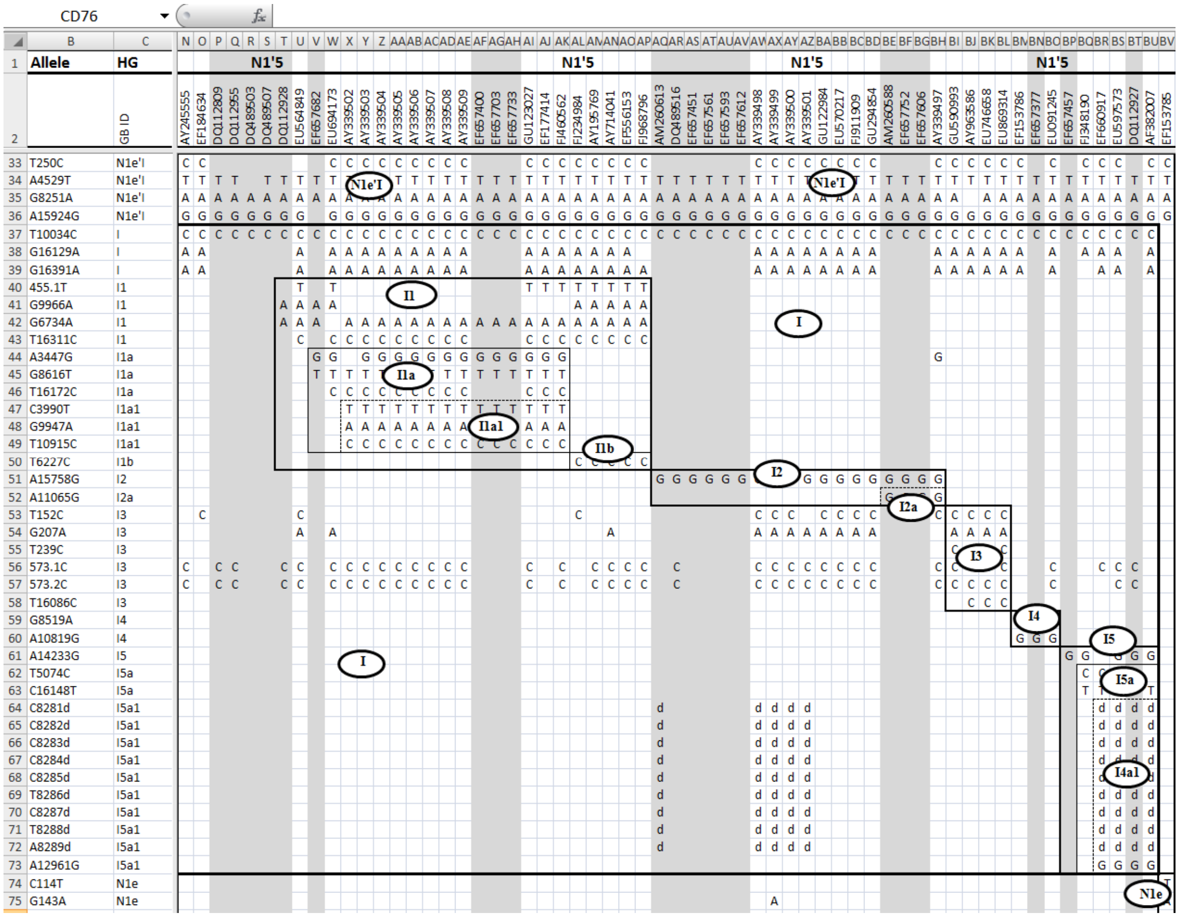
This base matrix was then transformed to infer and explicitly show the phylogenetic structure within the data. The matrix transformations were all linear (i.e., moving whole rows and whole columns) so as to preserve all relationships within and between haplotypes. Clades have been labeled in general conformance of consensus naming as documented in PhyloTree (van Oven and Kaiser, 2009). Figure 4 is a segment of PhyloTree that will be used in the discussion below.

Description of the mtMatrix-N

From this point forward, the assumption is made that the master matrix will be maintained and periodically updated as new data become available and as revisions are made to the phylogeny. The name "mtMatrix-N" has thus been coined and will be used designate this evolving database.

mtMatrix-N currently has nearly 4000 columns and over 7000 rows. The clade definitions start in the upper left hand corner and generally proceed in a hierarchical fashion from top down and left to right with alleles shown in rows (repeated as necessary) to provide definition of the clades and the columns represent GenBank sequences grouped within these clades.

A small segment of mtMatrix-N is shown in Figure 5. This segment shows Haplogroup I in context of Haplogroup N1 and will be used to illustrated the features of the matrix as well as provide a base for illustration of some of the issues that will be discussed



	B	C	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV
1	Allele	HG						N1'5																			N1'5														N1'5															N1'5							
2		GB ID	AY245555	EF184634	DQ112809	DQ112955	DQ489503	DQ489507	DQ112928	EU564849	EF657682	EU694173	AY339502	AY339503	AY339504	AY339505	AY339506	AY339507	AY339508	AY339509	EF657400	EF657703	EF657733	GU123027	EF177414	FJ460562	FJ234984	AY195769	AY714041	EF556153	FJ968796	AM260613	DQ489516	EF657451	EF657561	EF657593	EF657612	AY339498	AY339499	AY339500	AY339501	GU122984	EU570217	FJ911909	GU294854	AM260588	EF657752	EF657606	AY339497	GU590993	AY963586	EU746658	EU869314	EF153786	EF657377	EU091245	EF657457	FJ348190	EF660917	EU597573	DQ112927	AF382007	EF153785
33	T250C	N1e'I	C	C								C	C	C	C	C	C	C	C	C				C	C	C	C	C	C	C	C							C	C	C	C	C	C	C	C				C	C	C	C	C	C		C		C	C	C		C	C
34	A4529T	N1e'I	T	T	T	T		T	T	T	T	T	T	[illegible]	[illegible]	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	[illegible]	[illegible]	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
35	G8251A	N1e'I	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A		A	A	A	A	A	A	A	A	A	A	A	A
36	A15924G	N1e'I	G	G	G	G	G	G	G	G		G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
37	T10034C	I	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
38	G16129A	I	A	A						A		A	A	A	A	A	A	A	A	A				A	A	A	A	A	A	A								A	A	A	A	A	A	A	A				A	A	A	A	A	A		A		A	A	A		A	
39	G16391A	I	A	A						A		A	A	A	A	A	A	A	A	A				A	A	A	A	A	A	A	A							A	A	A	A	A	A	A	A				A	A	A	A	A	A		A			A	A		A	
40	455.1T	I1								T		T												T	T	T	T	T	T	T	T																																
41	G9966A	I1							A	A	A	A															A	A	A	A	A																																
42	G6734A	I1							A	A	A		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A																																
43	T16311C	I1								C		C	C	C	C	C	C	C	C	C				C	C	C	C	C	C	C	C																																
44	A3447G	I1a									G	G		G	G	G	G	G	G	G	G	G	G	G	G	G																							G														
45	G8616T	I1a									T	T	T	T	T	[illegible]	[illegible]	T	T	T	T	T	T	T	T	T																																					
46	T16172C	I1a										C	C	C	C	C	C	C	C	C				C	C	C																																					
47	C3990T	I1a1											T	T	T	T	T	T	T	T	T	T	T	T	T	T																																					
48	G9947A	I1a1											A	A	A	A	A	A	A	A	[illegible]	[illegible]	[illegible]	A	A	A																																					
49	T10915C	I1a1											C	C	C	C	C	C	C	C	C	C	C	C	C	C																																					
50	T6227C	I1b																									C	C	C	C	C																																
51	A15758G	I2																														G	G	G	G	G	G	[illegible]	[illegible]	[illegible]	G	G	G	G	G	G	G	G	G														
52	A11065G	I2a																																												G	[illegible]	[illegible]	G														
53	T152C	I3		C						C																	C											C	C	C		C	C	C	C				C	C	C	C	C										
54	G207A	I3								A		A																	A											A	A	A	A	A	A	A	A					A	A	A	A										
55	T239C	I3																																																C	[illegible]	[illegible]	C										
56	573.1C	I3	C		C	C			C	C		C	C	C	C	C	C	C	C	C				C		C		C	C	C	C		C					C	C	C	C	C	C	C	C				C	C	[illegible]	[illegible]	C			C			C	C	C		
57	573.2C	I3	C		C	C			C	C		C	C	C	C	C	C	C	C	C				C		C		C	C	C	C		C					C	C	C	C	C	C	C	C				C	C	C	C	C			C				C	C		
58	T16086C	I3																																																	C	C	C										
59	G8519A	I4																																																													
60	A10819G	I4																																																				G	G	G							
61	A14233G	I5																																																						G	G	[illegible]	G	G	G	
62	T5074C	I5a																																																							C	C	[illegible]	[illegible]		
63	C16148T	I5a																																																							T	T	[illegible]	[illegible]	T	
64	C8281d	I5a1																													d						d	d	d	d																		d	d	d	d	
65	C8282d	I5a1																													d						d	d	d	d																		d	d	d	d	
66	C8283d	I5a1																													d						d	d	d	d																		d	d	d	d	
67	C8284d	I5a1																													d						d	d	d	d																		d	d	d	d	
68	C8285d	I5a1																													d						d	d	d	d																		d	[illegible]	[illegible]	d	
69	T8286d	I5a1																													d						d	d	d	d																		d	d	d	d	
70	C8287d	I5a1																													d						d	d	d	d																		d	d	d	d	
71	T8288d	I5a1																													d						d	d	d	d																		d	d	d	d	
72	A8289d	I5a1																													d						d	d	d	d																		d	d	d	d	
73	A12961G	I5a1																																																							G	G	G	G	
74	C114T	N1e																																																											T
75	G143A	N1e																																				A																								A

Figure 5: An annotated screen shot of the Haplogroup I portion of a matrix produced by organizing 3584 mtDNA sequences selected to represent Haplogroup N, its subordinate haplogroups, and their clades.

below. Annotations in this illustration include boxes around details of selected clades with the weight of the lines on these boxes indicating the hierarchy of definition. The symbols in the ovals are the names of the clades. This illustration shows Haplogroup I bounded by the boldest lines. I1 is seen to lie within I and bounded by medium weight lines. Similarly, I1a shown within I1 and bounded by a light weight line with I1a1 further bounded by a light broken line. Ovals and boxes are not included in the matrix itself because as new sequences become available and further analysis is carried out, revisions to the boxes and possibly the labels will likely become necessary.

Figure 6 presents the same structural data in a tree format. There are distinct advantages and disadvantages to both formats.

Understanding the features of mtMatrix-N makes it easier to understand the observations concerning the current synthesis phylogeny as represented in mtMatrix-N and the issues derived from these observations. Thus, the next few paragraphs are dedicated to a more detailed description. mtMatrix-N is maintained in Microsoft Excel and references to row and column identifiers are to those provided by Excel. Description of the observations are generally relative to Figure 5.

Row 2 provides the GenBank identifier for the mtDNA sequences abstracted; columns N, O, P, Q, etc. represent allele differences found in these sequences. The description of these sequences as provided by Ian Logan are in row 3, but are not shown in the illustrations here. Some of these descriptions also contain a geographic location of the sample as provided by the researcher; where available this information has been extracted and placed in row 4. However, this descriptive data is there only for future use and will not be needed in describing either mtMatrix-N or the issues.

Each row below the row scroll break corresponds to an allele observed in one or more records in the database;

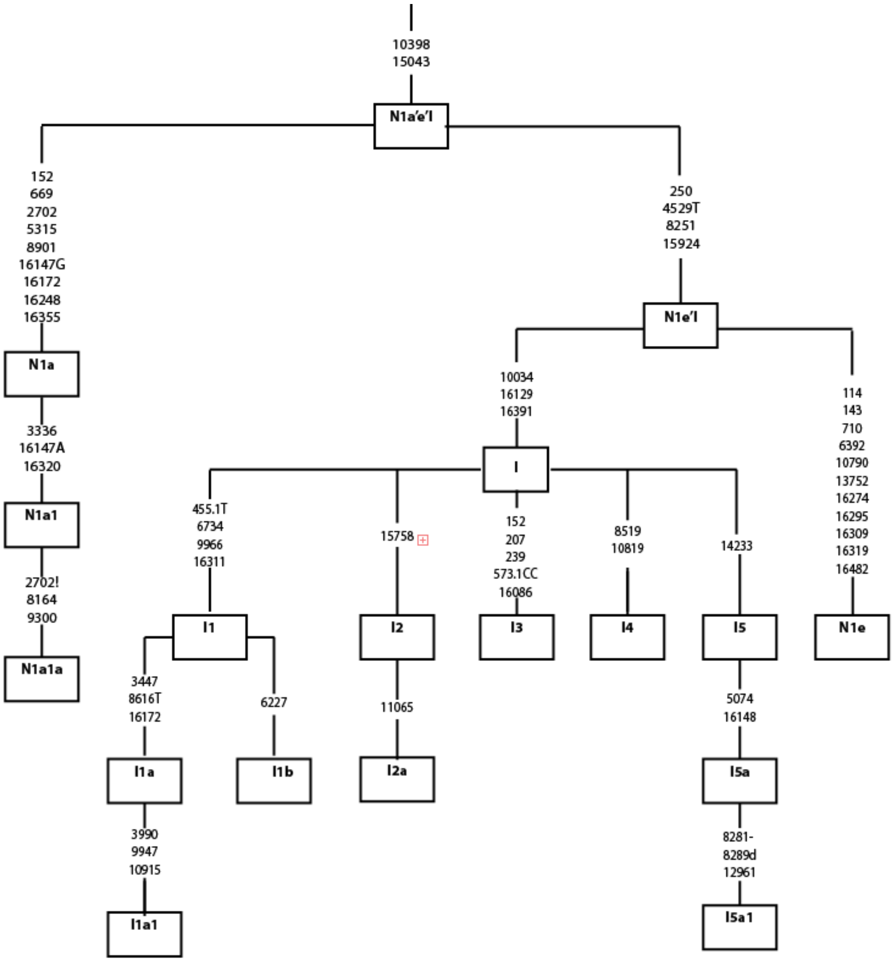


Figure 6: A phylogeny of Haplogroup I in tree format showing its context within Haplogroup N1

the name of that allele is given in column B. The allele names are formatted as follows: except for insertions, the first character is the base or nucleotide value (typically A, G, C, or T) found in the CRS. The next one to five numerical characters represent the numerical position in the mtDNA molecule, and the final character is the base or nucleotide to which the initial base has mutated. For deletions, the letter d is used instead for the final character. For insertions, there is no initial alphabetic character and the trailing allele value indicator is preceded by a decimal point and a number to indicate the position of the insertion relative to the insertion point.

Note that the alleles from the various reports have all been carefully aligned and that a symbol has been placed in the cell at the intersection of the record column and the allele row if that allele was observed in the sequence but left blank otherwise. Thus mtMatrix-N started out as a very sparse matrix until it was transformed by moving rows and columns to make the phylogenetic patterns visible as illustrated Figure 5. The identifiers in the ovals are hierarchical names for the clades 'discovered' by this process.

Many readers will have noted by now that several of the columns have a light grey background. This is a warning to the reviewer/analyst that although these mtDNA sequences were categorized in GenBank as "full genome sequences", they are incomplete in some aspect -- typically they are complete in the coding region but were not tested in the control region. These reports are included in mtMatrix-N for any information value they might contain, but sometimes there are significant gaps elsewhere in the test results. At some future time it might be decided to remove them, perhaps selectively, such as when there are so many others within a given clade that their information value is negligible. When such sequences are present it is important to know how to interpret them. As an illustration, look at columns P, Q, R, S, and T; each of these sequences have the light grey background. Now note that for each of these sequences the cells for rows 33 are empty whereas the cells for rows 34, 35, and 36 all contain a symbol. If the four alleles represented by these rows are used to define the clade named N1e'I, why are the cells for 33 empty? The explanation is relatively straight forward. Position 250 for allele T250C is in the control region of mtDNA and this is the only one of the four alleles that is so located. A check on the five sequences in question reveals they were incomplete and the control region was not tested. Thus, although it is not known for sure whether T250C would be revealed with further testing, it is a reasonable assumption in light of the fact that presence of G8251A and A15924G was observed in all five sequences in question and that A4529T was in four of the five. Furthermore, further examinations shows that these five sequences each have all the alleles that define N1 of which this chart is a segment. For the current example, four of these sequences have no apparent information value, but the fifth one (column T, ID DQ112928) could be significant in the analysis of clade I1.

Another feature of mtMatrix-N (not shown in this illustration) is the ability to assess the utilization of each allele in defining the phylogenetic structure. This feature is described further in the section on quality metrics below.

Observations from Haplogroup I

This section describes some of the observations made in developing the current version of mtMatrix-N and the subsequent ongoing review. The observations are then used as the basis for formalizing a number of issues presented in the following section for consideration in the development of any plan for future work. These observations can easily be illustrated in the charts already presented.

Although the original naming system made sense, some strange nomenclature has crept in where I is coupled with N1e to produce the N1e'I clade that is shown as the parent of I. But this is just a step away (off the page in the illustration) from coupling N1e'I with N1a to produce N1a'e'I. This in turn, is coupled with N1d and ultimately with N1c to produce N1a'c'd'e'I. Such constructs are awkward but probably necessary for complete and accurate representation of the phylogeny. However, this prompts questions as to when and under what conditions should individual haplogroups and associated clades be renamed to provide a nomenclature that is easier to follow.

The definition of Haplogroup I1 needs to be challenged. As presented (see Figure 4), it is defined by the insertion at 455.1T, and the transitions at 6734, 9966, and 16311. As synthesized from GenBank data into mtMatrix-n, Haplogroup I1 has 22 sample sequences as shown by the bounding box for I1 which is 23 columns wide. Included in its top-level definition are the insertion 455.1T (row 40) and the transition G9966A, but there are only ten occurrences of the insertion and nine occurrences of the transition in the 23 examples. Nine or ten out of 23 is hardly definitive of a clade. However, in addition to the ten samples where the insertion at 455.1 was reported, there are four more samples that are incomplete and could have this allele if they had been tested in the control region. Even so, 14 out of 23 is hardly definitive either. Removal of these two alleles from the definition of I1 would not reduce the membership defined by the other two alleles. Furthermore, the G6734A transition appears to be a very good indicator of the clade, in that it is found only 7 other times outside the I1 clade anywhere in the 3891 sequences in the current working database extracted from GenBank to represent the N superhaplogroup and all of its clades. T16311C appears to be a good indicator within the N1 block, but not outside the immediate block with occurrences in each of a total of 511 sequences. In fact, review of T16311T reveals that this allele is actually rather homoplasic and is used in 22 other clade definitions including clades in H, R, F, B, P, U, and K. Thus I1, as defined, includes G6734A as a relatively good defining allele, T10034C as an 'also present' allele, and two others that should not be included as defining.

A contrasting observation relates to occasional missing alleles and 'strays.' Row 36 presents an excellent example where the missing A4529T for sequence DQ489503 (column R) is the single exception out of 59 sequences classified as N1e'I. This is probably phylogenetically insignificant and could be simply a reading error in the test or a random back mutation. Another example is row 44 where there is a single exception in the 15 sequences classified as I1a. This allele also illustrates the occurrence of a stray allele A3447G in sequence AY339497 (column BH). This stray also appears to be phylogenetically insignificant and could be a random mutation or a reading error in the test.

The T152C (row 53) used in the definition of I3 is an extreme example of an allele present in all the sequences declared to be in the clade but having little information value. Not only does T152C occur in the four sequences of I3, it occurs 16 other times within the N1 block but also is found 607 times elsewhere in the database. This is an extreme case of homoplasy. Nevertheless it is found in 54 definitions within PhyloTree, including10 times in N exclusive of R, 33 times in R exclusive of U, and 11 times in U. With this track record one could ask why it is used in any definition. Staying with I3, G207A (row 54) presents a similar problem but to a lesser degree. The inclusion of the two insertions in the definition of I3 also presents severe problems in that it does not sufficiently differentiate I3, even within the N1 block. The T16086C (row 58) appears innocuous enough for inclusion in the definition of I3 (or perhaps a subclade) until the appearance in other clades is investigated; it appears in 9 sequences as part of the definition of a B4 clade and twice in a U3 clade, with other occurrences scattered throughout the database. On the other hand it appears that the T239C (row 55) is a near perfect allele for inclusion in the definition of I3 in that all four of the occurrences in the N1 block are in I3 and all four of the occurrence outside that block are used in the definition of Haplogroup H6. I3 is probably a valid clade but only one of the six alleles used in its definition is of high quality.

Next we consider Haplogroup N1e. This clade is at the edge of Figure 5 in the lower right corner, but it can be seen clearly in Figure 6. There are 59 GenBank records identified in these charts as belonging to either N1e or I in the composite clade named N1e'I. N1e and I were apparently identified as clades before the discovery that they had four alleles in their common ancestry (see the top of Figure 5). But is this complication reasonable? In fact, it can be questioned whether N1e is even a clade, since there is only a single sequence that qualifies for inclusion by definition of the clade, which itself requires 11 alleles. Perhaps the single haplotype currently used as a basis of the N1e definition should be reassigned and the four alleles now defining N1e'I node be returned to the basic definition of Haplogroup I. Looking slightly forward in mtMatrix-N, N1d is similarly found to contain a single sequence and N1c contains only two. Similar situations occur throughout the current phylogeny, with the extreme example being Haplogroup R14, a clade with a single member under R. This clade is defined by 22 alleles, even where three C insertions at one locus are counted as one allele.

Finally, it has been observed that some definitions include the absence of an allele as a criterion. An example occurs in the definition of clade N1e definition, but has not been implemented in mtMatrix-N. Specifically, the third element for the N1e definition as seen in Figure 4 is "204!" where the ! indicates the allele is identical to the CRS reference. It is recognized that there are occasions of back mutation where such use is justified, but the example for N1e does not seem justified since the clade contains only a single example. In analyzing this example, someone apparently noticed that this example did not contain T204C. But review reveals that 3 of the 34 N1e'I sequences with control region results also did not contain this allele. What is the justification for singling this one out for inclusion in the definition of N1e?

Issues

A fundamental issue relative to reference phylogenies is the purpose of their presentation and what should be included in the charts. For example, there is a tension between completeness for the analyst and clarity for use by the broader community. The insertion of intermediate alleles in clade definitions after they have been named often produces awkward constructions such N1a'c'd'e'I. It is certainly valid information for the analyst who may be computing estimates of age of origin, but it adds little if anything to a basic chart for the average member of the mtDNA Haplogroup I project.

What is the criteria for including alleles in the description/definition of a clade? Should there be a strict definition of a clade separated from a broader description? For example, as described above, the only high quality allele present in the current I3 definition is T239C. The double C insertion at 573 could be included in a description as "also present' but probably should not be included in the definition for the clade. In the past, underscores and parentheses have been used to qualify some alleles used in clade definitions. The underscore was used when the allele was found as defining in other clades in the primary haplogroup and the parentheses were used when it was only present in most but not all sequences in the clade. Perhaps some quality metric can be developed and included as another column in mtMatrix-N. Another approach would be to use a very strict definition for the clade but then also provide a clade encyclopedia that would include broader descriptions to include listing of alleles in the ancestry, listing of alleles that are defining, listing of alleles that are "also present," and list of alleles that were also considered but rejected and the reason for their rejection.

Under what conditions should a new clade be defined and added to the tree? Adding a clade before it is well established may have led to the N1e'I construction since the current N1e definition still has only a single example sequence in the database. Should a criterion be established that two example sequences are required? Three? Even if a number is established, what precautions should be taken to assure that the samples are not close relatives and thus represent a single lineage. Should genealogical data be part of the criteria? Should the samples be required to come from different locations and/or from independent researchers? Should it be required that the test results differ in at least one nucleotide position?

When should redefining/renaming of haplogroups and associated clades be proposed in order to clean up the presentation? This has been done periodically for Y-DNA, but how about mtDNA? Is there an alternative naming system for mtDNA clades that can be applied such as using the combination of a simple haplogroup name and one or more defining alleles, similar to constructions used in Y haplogroups such as G-M406?

What should be included in the allele encyclopedia? There should at least be a listing of all alleles observed in mtMatrix-N and a count of their occurrences. Perhaps there should be cross-references to any use in any clade definition with the quality metric for that clade. Development of such metric would be much more meaningful in the context of mtMatrix-N where we can look across all of Super-Haplogroup N as distinct isolated lettered haplogroups.

Should incomplete data be included in mtMatrix-N as illustrated above? Should we include non-GenBank data, such as that available from mtDNA projects? If so, how to we make sure we don't have duplicates if that data is later included in GenBank?

What criteria should be used to solve apparent reticulations or other conflicts where there are two patterns that solve a particular problem?

Should we have a separate guide/matrix for use when only HVR1/HVR2 results are available? If so, another set of criteria is probably required.

There will undoubtedly be other issues surface as our work proceeds.

Quality Metrics

Throughout the development and maintenance of any phylogeny and its supporting data, one goal is to be as objective as possible. As illustrated in the issues section, many factors go into deciding when to name an apparent clade and what alleles are to be used in the definition of that clade. This section illustrates an approach to assigning a quality metric to the alleles in the context of the various clades. It is not expected that the metrics illustrated here will be accepted as a standard, but rather they are presented to stimulate discussion for refinement or complete replacement of the individual metrics or the reformulation of the overall quality metric from the components.

As motivation of the illustration, consider the following qualitative scheme for evaluating alleles as an element of a clade definition. Assume a quality scale of zero to five, where an allele in a specific context is assigned a five if it is ideal in all aspects, and is assigned a zero if use of that allele in a given clade definition would be of no value or even be misleading. It is easy to imagine the assignment of values between these extremes where the use of an allele in a definition would be of some value but would also present one or more less than perfect situations. The issue is the development of a scheme for assignment of such a number to the alleles when used within various contexts of clade definitions.

There are two primary facets in evaluating the quality of an allele for defining a clade in the phylogeny -- a general one relating to the type of polymorphism that the allele represents and a specific one relating to the contexts of one or more clades. Historically, researchers have differentiated alleles that derive from the primary non-coding region (positions 16024 through 16469 and 1 through 574) from the remainder of the molecule, typically referred to as the coding region. Since the coding region predominantly codes for genes or functional RNA (i.e., tRNA or rRNA), a change in this region has the potential for significantly affecting the functioning of the mitochondrial mechanisms. This contrasts with the non-coding region where a simple mutation generally has no known impact. It is proposed that the difference in significance of the various regions be accounted for by defaults weights of 5 for the segments that code for proteins, 3 for segments that code for RNA, and 1 for the non-coding regions. However, since the coding of each component of a protein uses three nucleotides and further since that code has some redundancy, many of the changes actually have no effect (i.e., the codes are synonyms). Thus, it is proposed that these "synonymous mutations" be assigned a weight of 1, comparable to changes in the non-coding regions. Note that mutations that do change the composition of the protein (i.e., the non-synonymous ones) tend to be selected against in the evolutionary processes; thus a 5 to one ratio appears reasonable.

The above facet is based entirely on the characteristics of the allele and does not consider the context of its use in defining clades in a phylogeny. The following describes an approach for computing several factors that may be applied to adjust for use of the various alleles in specific contexts or scopes of review. For the approach illustrated here, these contexts are hierarchical and are dependent on making a variety of counts of occurrence of the alleles in these context. At the lowest level, the context is limited to the clade being defined where a completeness factor may be defined as simply the ratio of the number of occurrences of the allele in the clade divided by the number of haplotypes in the clade -- a number between 0 and 1 with 1 being ideal. At a context higher in the hierarchy, such as a named haplogroup that is parent to the clade being defined, then a discrimination factor may be defined as the ratio of the number of alleles in the clade being defined by the number of alleles within that next higher level, e.g., the named haplogroup. Again this is a number between 0 and 1 with 1 being ideal. From a broader perspective, homoplasies should also be taken into consideration. For example, a given allele may be used in the definition of one or more clades, either within the same major haplogroup or elsewhere in the database. A homoplasy factor could be computed simply as the reciprocal of the number of such definitions where the allele is used, but this should probably be tempered by some transformation function such as taking a square root. An overall quality metric for a given allele being used in a clade definition can then be derived by multiplying each of these factors together and applying them to the weight described in the previous paragraph. This approach has been demonstrated in the current version of mtMatrix-N for the clades in Haplogroups I and U and selected other contexts and can be seen in conjunction with the description of an allele encyclopedia described below.

Allele Encyclopedia

It has been recognized for some time that some kind of consolidation of data about the various alleles would be useful. Now that mtMatrix exists, it is feasible to automatically produce the foundation for such an encyclopedia as illustrated in Figure 7. The illustration is from an Excel spreadsheet with the first three rows as header descriptions and the following rows containing a variety of "facts" about the various alleles and the contexts in which they occur.

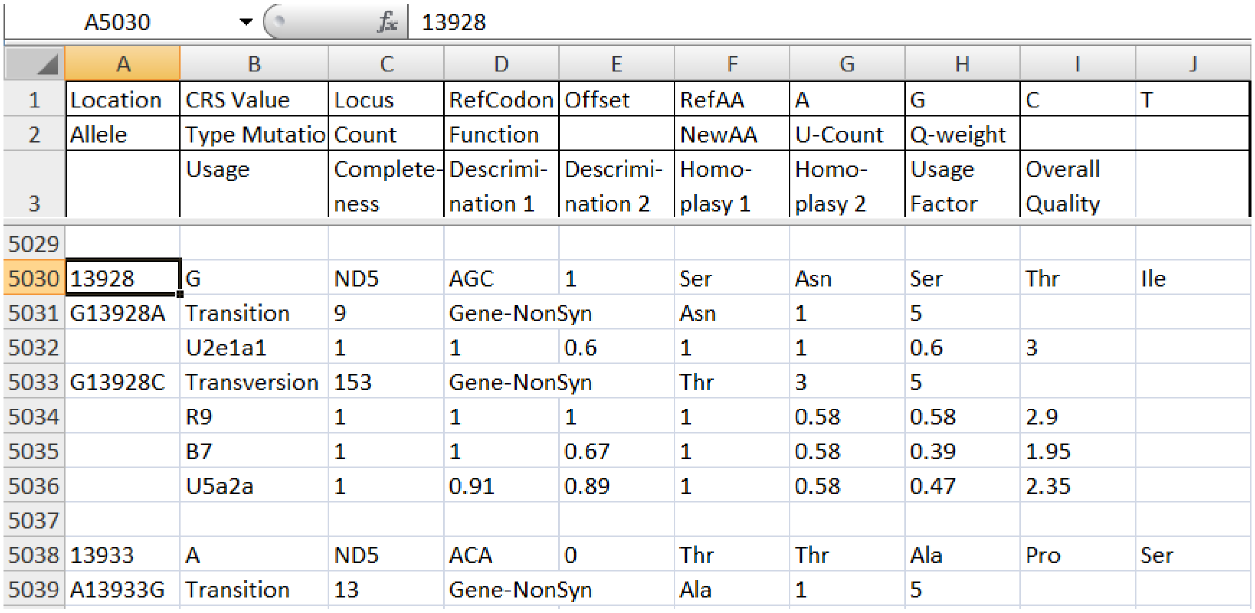
The rows of data are grouped, as illustrated by rows 5030 through 5036 with blank rows before and after the group. In this case the group concerns location 13928 of the mtDNA molecule and as described by the first row in the data group (row 5030). A guide for interpreting this data row is in row 1 of the header. Specifically this reference nucleotide for position 13928 is a G (Guanine), which falls in the structural locus for the ND5 gene. Furthermore it is part of an AGC codon and is offset by one position from the beginning of that codon, i.e. in the middle position of that codon. As shown in the RefAA column (short for reference amino acid), AGC codes for Ser (abbreviation for Serine) in the ND5 protein to be produced. The last four fields of this allele description show what would result for various substitution. In the case of 13928, instead of Ser, a substitution for that G in the AGC codon, would produce the amino acid Asn (Asparagine), Thr (Threonine), or Ile (Isoleucine), depending on whether the substitution was an A, C, or T.

Following the allele description in each group are one or more rows, each providing data on a specific polymorphism at the location being considered. The header from these polymorphism descriptions is row 2. This is illustrated in row 5031 which describes a transition from G to A and row 5033 which describes the transversion from a G to a C. As shown in the count column (column C) the transition was observed 9 times among the haplotypes in the matrix and the transversion occurred 153 times. Both produced non-synonymous changes within the gene and thus a Q-weight 5 as described in the Quality Metric section above. In column G there is a U-count which is the number of times this allele is used in a clade definition within the database. In the case of the transition, there is a single usage that occurs in U2e1a1 as described in the immediately following row (row 5032). On the other hand, the transversion is used in the definition of 3 clades -- R9, B7, and U5a2a that are described in rows 5034, 5035, and 5036, respectively. Associated with each of the four rows that describe where an allele is used in a clade definition, there are five context qualifiers as described in the Quality Metrics section. These quality factors have all been multiplied together to produce a Usage Factor given in Column H. This Usage Factor, in turn, has been multiplied by the Q-weight of the corresponding allele to product an Overall Quality for the specific allele as used in the definition of the respective clades. For the quality metrics currently illustrated, the maximum possible value is a 5.0. The four examples here are all quite reasonable, but there are some contexts that produce quite low quality scores indicating that they should probably not be included in some or all definitions where they are now used. For example, the 455.1T included in the clade definition of I1 produced a score of only 0.06 for the present implementation. Note that this is the same allele described in the previous description of mtMatrix where it was pointed out that the 455.1T allele is used in definition of the I1 clade, but it occurs in only 9 of the 18 haplotypes in that clade that are complete in both the coding and non-coding regions.

Tools Supporting the Approach

With only small differences (i.e., due to the insertion and deletion polymorphisms) each mitochondrial DNA (mtDNA) in humans is 16,569 base pairs in length. Rather than working with the entire sequence, however, it is common to compare each sequence to a standard (the CRS), and work with differences from that standard. A typical mtDNA report of a sequence recently added to GenBank is illustrated in Figure 8.

DNA sequences are deposited to GenBank by researchers from around the world, and thus GenBank currently has about 7000 complete (or near complete) mtDNA sequences from a wide variety of sources. FASTA files for these sequences can be accessed directly from GenBank. However, for purposes of this project it is more convenient to simply access the CRS differences using Checker, a tool developed and



A5030 fx 13928

	A	B	C	D	E	F	G	H	I	J
1	Location	CRS Value	Locus	RefCodon	Offset	RefAA	A	G	C	T
2	Allele	Type Mutatio	Count	Function		NewAA	U-Count	Q-weight		
3		Usage	Completeness	Descrimination 1	Descrimination 2	Homoplasy 1	Homoplasy 2	Usage Factor	Overall Quality	
5029										
5030	13928	G	ND5	AGC	1	Ser	Asn	Ser	Thr	Ile
5031	G13928A	Transition	9	Gene-NonSyn		Asn	1	5		
5032		U2e1a1	1	1	0.6	1	1	0.6	3	
5033	G13928C	Transversion	153	Gene-NonSyn		Thr	3	5		
5034		R9	1	1	1	1	0.58	0.58	2.9	
5035		B7	1	1	0.67	1	0.58	0.39	1.95	
5036		U5a2a	1	0.91	0.89	1	0.58	0.47	2.35	
5037										
5038	13933	A	ND5	ACA	0	Thr	Thr	Ala	Pro	Ser
5039	A13933G	Transition	13	Gene-NonSyn		Ala	1	5		

Figure 7: Screen shot of a segment of computer generated foundation for an mtDNA allele encyclopedia with the cursor placed on position A5030 -- the beginning of a report for mtDNA position 13028.

```
HM047061 FTDNA Haplogroup N1a1 13-APR-2010
A73G    T152C   T199C   T204C   G207A   A263G   315.1C  573.1C  573.2C  T669C
A750G   A1438G  G1719A  G2702A  A2706G  T3336C  A4769G  A5315G  C7028T  G8485A
A8860G  A8901G  T10238C A10398G C10473Y A11641G G11719A G12501A C12705T G13477A
A13780G C14766T G15043A T15299C A15326G T15697C T16086C C16147A C16223T C16248T
C16320T C16355T T16519C
```

Figure 8: Typical report from an mtDNA sequence recently added to GenBank

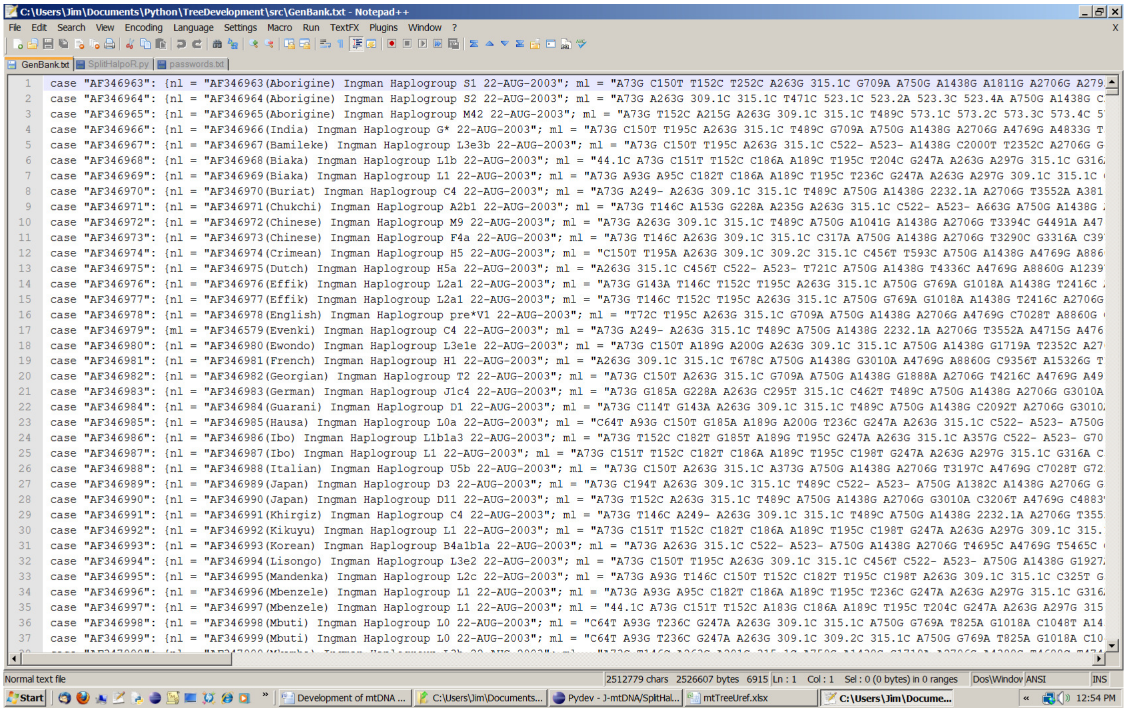
maintained by Ian Logan (2010) in the UK. A segment of this Checker script was used as the primary source of data in development of the current version of mtMatrix-N is shown in Figure 9.

To facilitate the implementation of the approach that has been presented here, a Python script has been developed and used to accept Checker scripts as illustrated from Figure 9, select only those reports satisfying given criteria based on presence or absence of certain alleles (polymorphism values) in the individual reports, align the alleles on each of these reports (producing a sparse matrix), count the number of occurrences of each allele, sort the rows in order of the frequency of occurrence of the various alleles, and create an output in a CSV format that can be read directly by an Microsoft Excel spreadsheet. A screenshot of a segment of this output as read by Excel is shown in Figure 10.

Unfortunately, the next step is manual. Keeping all relationships fixed, the rows and columns are rearranged to identify and display the evolutionary patterns implicit in the data. Figure 11 shows obvious patterns identified in this manner. The first round of matrix transformation was specifically to organize the data in conformance with PhyloTree (van Oven and Kayser, 2009) and thus uses the PhyloTree designations. As illustrated in the observations section above, mtMatrix-N, in turn, can now be used to identify errors in PhyloTree and suggest refinements to its structure.

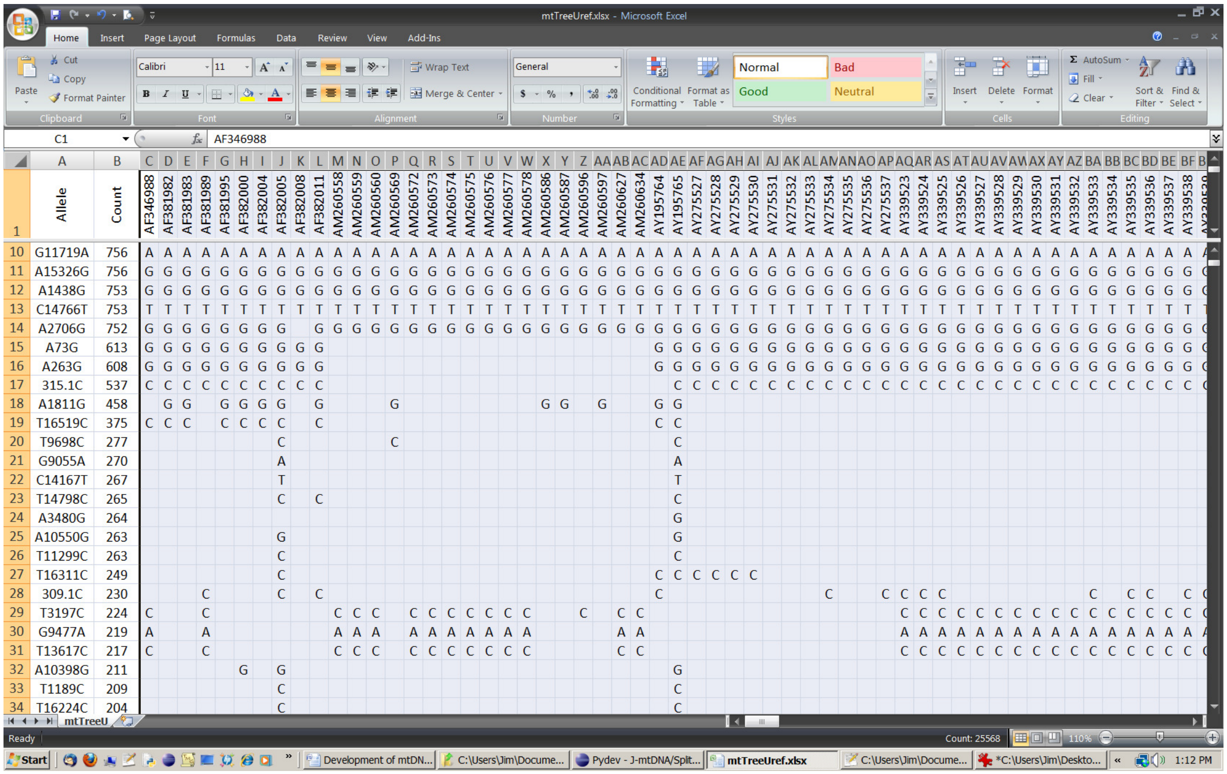
Once data is available in matrix format as illustrated, it can be used for clade definitions and presented in tree format as illustrated in Figure 12.

The download from GenBank (via Checker) described above was made on 15 March 2010. Since that time there have been approximately 250 full genome mtDNA sequences added to GenBank. Ian Logan has accordingly sent out email announcements giving reports as illustrated before; approximately 200 of them satisfy the selection criteria being used in the project. As part of database maintenance, the content of these emails have been compiled into batches and Python has been used for


C:\Users\Jim\Documents\Python\TreeDevelopment\src\GenBank.txt - Notepad++
File Edit Search View Encoding Language Settings Macro Run TextFX Plugins Window ?
GenBank.txt
SplitHalpoR.py
passwords.txt
Normal text file
2512779 chars 2526607 bytes 6915 Ln : 1 Col : 1 Sel : 0 (0 bytes) in 0 ranges
Dos\Window ANSI INS
Start
Development of mtDNA ...
C:\Users\Jim\Documents...
Pydev - J-mtDNA/SplitHal...
mtTreeUref.xlsx
C:\Users\Jim\Docume...
12:54 PM

```
case "AF346963": {nl = "AF346963(Aborigine) Ingman Haplogroup S1 22-AUG-2003"; ml = "A73G C150T T152C T252C A263G 315.1C G709A A750G A1438G A1811G A2706G A279
case "AF346964": {nl = "AF346964(Aborigine) Ingman Haplogroup S2 22-AUG-2003"; ml = "A73G A263G 309.1C 315.1C T471C 523.1C 523.2A 523.3C 523.4A A750G A1438G C
case "AF346965": {nl = "AF346965(Aborigine) Ingman Haplogroup M42 22-AUG-2003"; ml = "A73G T152C A215G A263G 309.1C 315.1C T489C 573.1C 573.2C 573.3C 573.4C 5
case "AF346966": {nl = "AF346966(India) Ingman Haplogroup G* 22-AUG-2003"; ml = "A73G C150T T195C A263G 315.1C T489C G709A A750G A1438G A2706G A4769G A4833G T
case "AF346967": {nl = "AF346967(Bamileke) Ingman Haplogroup L3e3b 22-AUG-2003"; ml = "A73G C150T T195C A263G 315.1C C522- A523- A1438G C2000T T2352C A2706G G
case "AF346968": {nl = "AF346968(Biaka) Ingman Haplogroup L1b 22-AUG-2003"; ml = "44.1C A73G C151T T152C C186A A189C T195C T204C G247A A263G A297G 315.1C G316
case "AF346969": {nl = "AF346969(Biaka) Ingman Haplogroup L1 22-AUG-2003"; ml = "A73G A93G A95C C182T C186A A189C T195C T236C G247A A263G A297G 309.1C 315.1C
case "AF346970": {nl = "AF346970(Buriat) Ingman Haplogroup C4 22-AUG-2003"; ml = "A73G A249- A263G 309.1C 315.1C T489C A750G A1438G 2232.1A A2706G T3552A A381
case "AF346971": {nl = "AF346971(Chukchi) Ingman Haplogroup A2b1 22-AUG-2003"; ml = "A73G T146C A153G G228A A235G A263G 315.1C C522- A523- A663G A750G A1438G
case "AF346972": {nl = "AF346972(Chinese) Ingman Haplogroup M9 22-AUG-2003"; ml = "A73G A263G 309.1C 315.1C T489C A750G A1041G A1438G A2706G T3394C G4491A A47
case "AF346973": {nl = "AF346973(Chinese) Ingman Haplogroup F4a 22-AUG-2003"; ml = "A73G T146C A263G 309.1C 315.1C C317A A750G A1438G A2706G T3290C G3316A C39
case "AF346974": {nl = "AF346974(Crimean) Ingman Haplogroup H5 22-AUG-2003"; ml = "C150T T195A A263G 309.1C 309.2C 315.1C C456T T593C A750G A1438G A4769G A886
case "AF346975": {nl = "AF346975(Dutch) Ingman Haplogroup H5a 22-AUG-2003"; ml = "A263G 315.1C C456T C522- A523- T721C A750G A1438G T4336C A4769G A8860G A1239
case "AF346976": {nl = "AF346976(Effik) Ingman Haplogroup L2a1 22-AUG-2003"; ml = "A73G G143A T146C T152C T195C A263G 315.1C A750G G769A G1018A A1438G T2416C
case "AF346977": {nl = "AF346977(Effik) Ingman Haplogroup L2a1 22-AUG-2003"; ml = "A73G T146C T152C T195C A263G 315.1C A750G G769A G1018A A1438G T2416C A2706G
case "AF346978": {nl = "AF346978(English) Ingman Haplogroup pre*V1 22-AUG-2003"; ml = "T72C T195C A263G 315.1C G709A A750G A1438G A2706G A4769G C7028T A8860G
case "AF346979": {ml = "AF346579(Evenki) Ingman Haplogroup C4 22-AUG-2003"; ml = "A73G A249- A263G 315.1C T489C A750G A1438G 2232.1A A2706G T3552A A4715G A476
case "AF346980": {nl = "AF346980(Ewondo) Ingman Haplogroup L3e1e 22-AUG-2003"; ml = "A73G C150T A189G A200G A263G 309.1C 315.1C A750G A1438G G1719A T2352C A27
case "AF346981": {nl = "AF346981(French) Ingman Haplogroup H1 22-AUG-2003"; ml = "A263G 309.1C 315.1C T678C A750G A1438G G3010A A4769G A8860G C9356T A15326G T
case "AF346982": {nl = "AF346982(Georgian) Ingman Haplogroup T2 22-AUG-2003"; ml = "A73G C150T A263G 315.1C G709A A750G A1438G G1888A A2706G T4216C A4769G A49
case "AF346983": {nl = "AF346983(German) Ingman Haplogroup J1c4 22-AUG-2003"; ml = "A73G G185A G228A A263G C295T 315.1C C462T T489C A750G A1438G A2706G G3010A
case "AF346984": {nl = "AF346984(Guarani) Ingman Haplogroup D1 22-AUG-2003"; ml = "A73G C114T G143A A263G 309.1C 315.1C T489C A750G A1438G C2092T A2706G G3010
case "AF346985": {nl = "AF346985(Hausa) Ingman Haplogroup L0a 22-AUG-2003"; ml = "C64T A93G C150T G185A A189G A200G T236C G247A A263G 315.1C C522- A523- A750G
case "AF346986": {nl = "AF346986(Ibo) Ingman Haplogroup L1b1a3 22-AUG-2003"; ml = "A73G T152C C182T G185T A189G T195C G247A A263G 315.1C A357G C522- A523- G70
case "AF346987": {nl = "AF346987(Ibo) Ingman Haplogroup L1 22-AUG-2003"; ml = "A73G C151T T152C C182T C186A A189C T195C C198T G247A A263G A297G 315.1C G316A C
case "AF346988": {nl = "AF346988(Italian) Ingman Haplogroup U5b 22-AUG-2003"; ml = "A73G C150T A263G 315.1C A373G A750G A1438G A2706G T3197C A4769G C7028T G72
case "AF346989": {nl = "AF346989(Japan) Ingman Haplogroup D3 22-AUG-2003"; ml = "A73G C194T A263G 309.1C 315.1C T489C C522- A523- A750G A1382C A1438G A2706G G
case "AF346990": {nl = "AF346990(Japan) Ingman Haplogroup D11 22-AUG-2003"; ml = "A73G T152C A263G 315.1C T489C A750G A1438G A2706G G3010A C3206T A4769G C4883
case "AF346991": {nl = "AF346991(Khirgiz) Ingman Haplogroup C4 22-AUG-2003"; ml = "A73G T146C A249- A263G 309.1C 315.1C T489C A750G A1438G 2232.1A A2706G T355
case "AF346992": {nl = "AF346992(Kikuyu) Ingman Haplogroup L1 22-AUG-2003"; ml = "A73G C151T T152C C182T C186A A189C T195C C198T G247A A263G A297G 309.1C 315.
case "AF346993": {nl = "AF346993(Korean) Ingman Haplogroup B4a1b1a 22-AUG-2003"; ml = "A73G A263G 315.1C C522- A523- A750G A1438G A2706G T4695C A4769G T5465C
case "AF346994": {nl = "AF346994(Lisongo) Ingman Haplogroup L3e2 22-AUG-2003"; ml = "A73G C150T T195C A263G 309.1C 315.1C C456T C522- A523- A750G A1438G G1927
case "AF346995": {nl = "AF346995(Mandenka) Ingman Haplogroup L2c 22-AUG-2003"; ml = "A73G A93G T146C C150T T152C C182T T195C C198T A263G 309.1C 315.1C C325T G
case "AF346996": {nl = "AF346996(Mbenzele) Ingman Haplogroup L1 22-AUG-2003"; ml = "A73G A93G A95C C182T C186A A189C T195C T236C G247A A263G A297G 315.1C G316
case "AF346997": {nl = "AF346997(Mbenzele) Ingman Haplogroup L1 22-AUG-2003"; ml = "44.1C A73G C151T T152C A183G C186A A189C T195C T204C G247A A263G A297G 315
case "AF346998": {nl = "AF346998(Mbuti) Ingman Haplogroup L0 22-AUG-2003"; ml = "C64T A93G T236C G247A A263G 309.1C 315.1C A750G G769A T825A G1018A C1048T A14
case "AF346999": {nl = "AF346999(Mbuti) Ingman Haplogroup L0 22-AUG-2003"; ml = "C64T A93G T236C G247A A263G 309.1C 309.2C 315.1C A750G G769A T825A G1018A C10
```

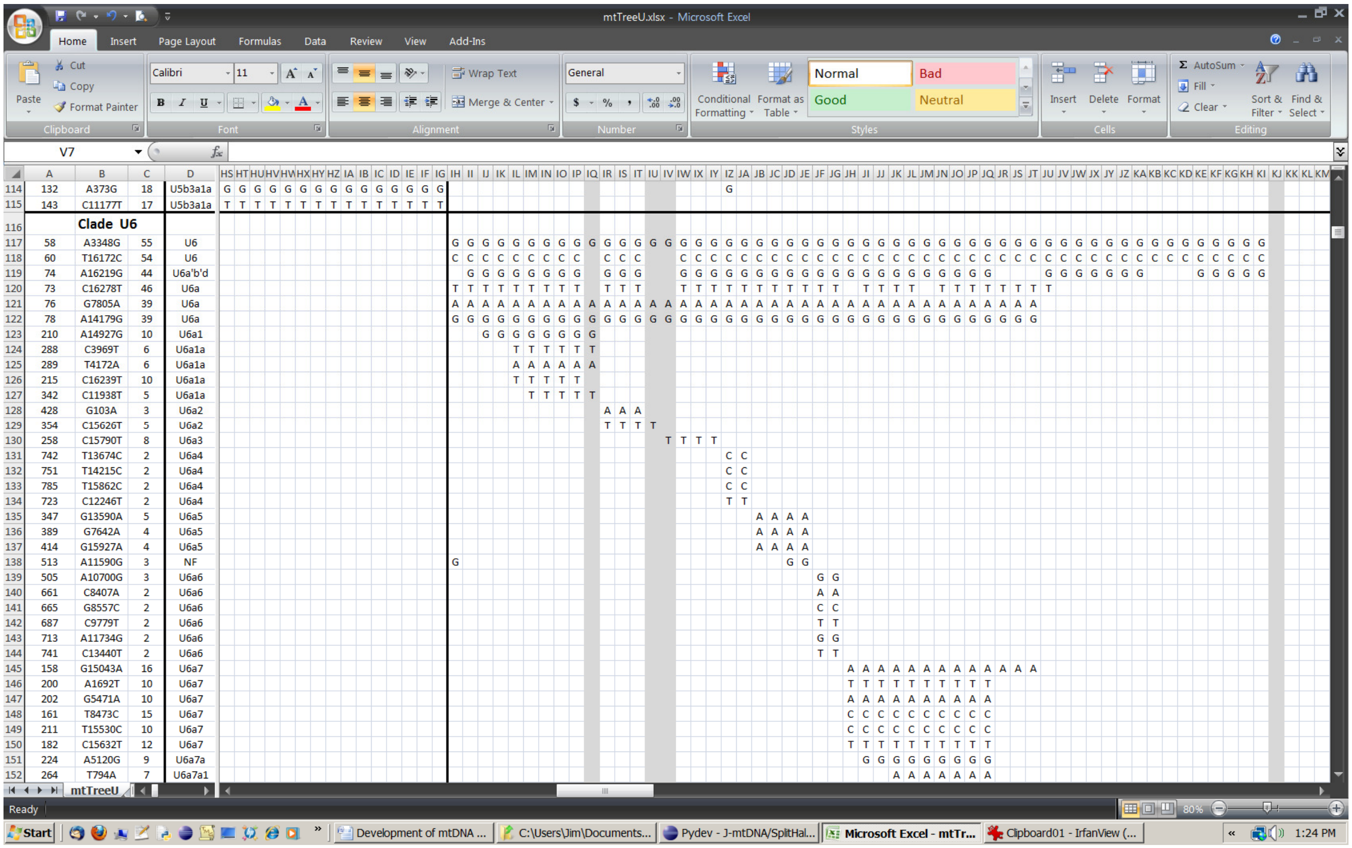
Figure 9: Screenshot of several records of Ian Logan's script that contains the haplogroup descriptions of 6683 mtDNA full (or near full) genome sequences contained in GenBank.


mtTreeUref.xlsx - Microsoft Excel
Home Insert Page Layout Formulas Data Review View Add-Ins
Paste Cut Copy Format Painter Clipboard
Calibri 11 Font
Wrap Text Merge & Center Alignment
General Number
Conditional Formatting Format as Table
Normal Bad Good Neutral Styles
Insert Delete Format Cells
AutoSum Fill Clear Sort & Filter Find & Select Editing
mtTreeU
Ready
Count: 25568
110%
Start
Development of mtDN...
C:\Users\Jim\Docume...
Pydev - J-mtDNA/Split...
mtTreeUref.xlsx
C:\Users\Jim\Docume...
*C:\Users\Jim\Deskto...
1:12 PM

C1 AF346988

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF
1	Allele	Count	AF346988	AF381982	AF381983	AF381989	AF381995	AF382000	AF382004	AF382005	AF382008	AF382011	AM260558	AM260559	AM260560	AM260569	AM260572	AM260573	AM260574	AM260575	AM260576	AM260577	AM260578	AM260586	AM260587	AM260596	AM260597	AM260627	AM260634	AY195764	AY195765	AY275527	AY275528	AY275529	AY275530	AY275531	AY275532	AY275533	AY275534	AY275535	AY275536	AY275537	AY339523	AY339524	AY339525	AY339526	AY339527	AY339528	AY339529	AY339530	AY339531	AY339532	AY339533	AY339534	AY339535	AY339536	AY339537	AY339538
10	G11719A	756	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
11	A15326G	756	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
12	A1438G	753	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
13	C14766T	753	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
14	A2706G	752	G	G	G	G	G	G	G	G		G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
15	A73G	613	G	G	G	G	G	G	G	G	G	G																		G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
16	A263G	608	G	G	G	G	G	G	G	G	G	G																		G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
17	315.1C	537	C	C	C	C	C	C	C	C	C	C																			C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
18	A1811G	458		G	G		G	G	G	G		G				G								G	G		G			G	G																											
19	T16519C	375	C	C	C		C	C	C	C		C																		C	C																											
20	T9698C	277								C						C															C																											
21	G9055A	270								A																					A																											
22	C14167T	267								T																					T																											
23	T14798C	265								C		C																			C																											
24	A3480G	264																													G																											
25	A10550G	263								G																					G																											
26	T11299C	263								C																					C																											
27	T16311C	249								C																				C	C	C	C	C	C																							
28	309.1C	230				C				C		C																		C									C			C	C	C	C									C		C	C		C
29	T3197C	224	C			C							C	C	C		C	C	C	C	C	C	C			C		C	C														C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
30	G9477A	219	A			A							A	A	A		A	A	A	A	A	A	A					A	A														A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
31	T13617C	217	C			C							C	C	C		C	C	C	C	C	C	C					C	C														C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
32	A10398G	211						G		G																					G																											
33	T1189C	209								C																					C																											
34	T16224C	204								C																					C																											

Figure 10: Screenshot of segment of spreadsheet produced by Python software that selected and aligned GenBank entries that satisfied criteria for mtDNA Haplogroup U.



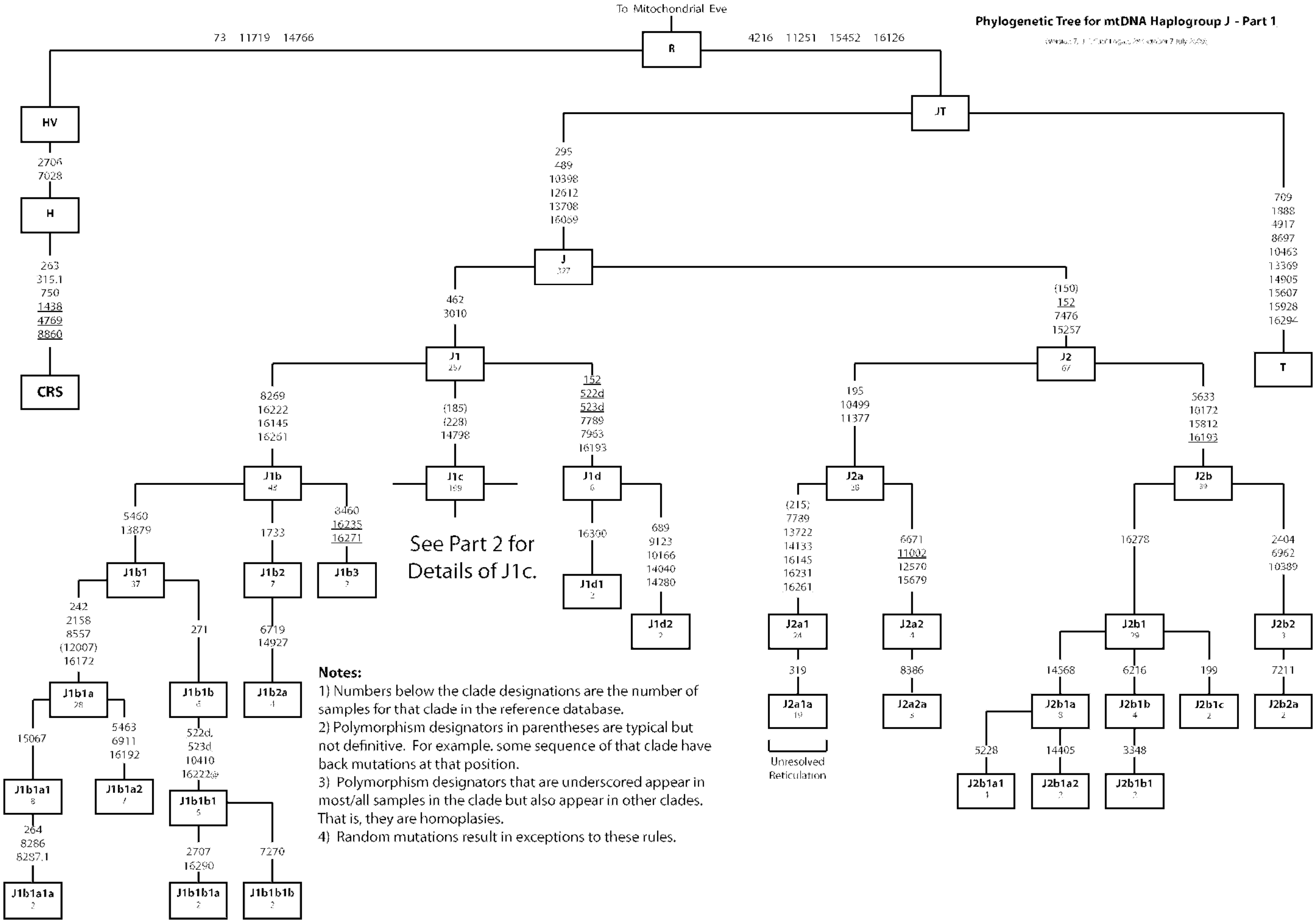
	A	B	C	D	HS	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP	IQ	IR	IS	IT	IU	IV	IW	IX	IY	IZ	JA	JB	JC	JD	JE	JF	JG	JH	JI	JJ	JK	JL	JM	JN	JO	JP	JQ	JR	JS	JT	JU	JV	JW	JX	JY	JZ	KA	KB	KC	KD	KE	KF	KG	KH	KI	KJ	KK	KL	KM
114	132	A373G	18	U5b3a1a	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G																			G																																								
115	143	C11177T	17	U5b3a1a	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T																																																												
116		Clade U6																																																																											
117	58	A3348G	55	U6																G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G				
118	60	T16172C	54	U6																C	C	C	C	C	C	C	C	C		C	C	C			C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C				
119	74	A16219G	44	U6a'b'd																	G	G	G	G	G	G	G	G		G	G	G			G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G				G	G	G	G	G	G	G				G	G	G	G	G				
120	73	C16278T	46	U6a																T	T	T	T	T	T	T	T	T		T	T	T			T	T	T	T	T	T	T	T	T	T	T		T	T	T	T		T	T	T	T	T	T	T	T																		
121	76	G7805A	39	U6a																A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A																			
122	78	A14179G	39	U6a																G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G																			
123	210	A14927G	10	U6a1																		G	G	G	G	G	G	G	G																																																		
124	288	C3969T	6	U6a1a																				T	T	T	T	T	T																																																		
125	289	T4172A	6	U6a1a																				A	A	A	A	A	A																																																		
126	215	C16239T	10	U6a1a																				T	T	T	T	T																																																			
127	342	C11938T	5	U6a1a																					T	T	T	T	T																																																	
128	428	G103A	3	U6a2																										A	A	A																																																
129	354	C15626T	5	U6a2																										T	T	T	T																																														
130	258	C15790T	8	U6a3																														T	T	T	T																																													
131	742	T13674C	2	U6a4																																		C	C																																																				
132	751	T14215C	2	U6a4																																		C	C																																																			
133	785	T15862C	2	U6a4																																		C	C																																																			
134	723	C12246T	2	U6a4																																		T	T																																																			
135|347|G13590A|5|U6a5|																																		|A|A|A|A|																																																	
136|389|G7642A|4|U6a5|																																		|A|A|A|A|																																																	
137|414|G15927A|4|U6a5|																																		|A|A|A|A|																																																	
138|513|A11590G|3|NF|														|G|																					|G|G|																																																					
139|505|A10700G|3|U6a6|																																						|G|G|																																																
140|661|C8407A|2|U6a6|																																						|A|A|																																																
141|665|G8557C|2|U6a6|																																						|C|C|																																																
142|687|C9779T|2|U6a6|																																						|T|T|																																																
143|713|A11734G|2|U6a6|																																						|G|G|																																																
144|741|C13440T|2|U6a6|																																						|T|T|																																																
145	158	G15043A	16	U6a7																																										A	A	A	A	A	A	A	A	A	A	A	A	A																			
146	200	A1692T	10	U6a7																																										T	T	T	T	T	T	T	T	T	T																						
147	202	G5471A	10	U6a7																																										A	A	A	A	A	A	A	A	A	A																						
148	161	T8473C	15	U6a7																																										C	C	C	C	C	C	C	C	C	C																						
149	211	T15530C	10	U6a7																																										C	C	C	C	C	C	C	C	C	C																						
150	182	C15632T	12	U6a7																																										T	T	T	T	T	T	T	T	T	T																						
151	224	A5120G	9	U6a7a																																											G	G	G	G	G	G	G	G	G																						
152	264	T794A	7	U6a7a1																																													A	A	A	A	A	A	A																						

Figure 11: Screenshot of a segment of alignment of mtDNA Haplogroup U alleles showing the derived phylogenetic structure.

allele alignment and formatting suitable for merging the new data into the mtMatrix-N. This time, however, the alignments are not only to each other, but the alleles have also been reorder to conform to order of presentation in the last version mtMatrix-N. But there is a complication. Since some polymorphisms are homoplasic (i.e., branch mutations have occurred more than once within the genetic history of modern humans), they occur in multiple patterns defining the evolutionary clades. To accommodate this phenomenon, rows in the matrix have been duplicated and thus some alleles must be duplicated and included in each of the corresponding rows. These requirements were easily satisfied with a Python script.

Within Haplogroup N, the mtDNA phylogenies for Haplogroup J have been developed and presented (Logan, 2008; Logan 2009) over the past two years (from whence came Figure 12) and have since been cited in PhyloTree. This work is the subject of ongoing collaboration with the developers of PhyloTree. In fact, for the past seven releases of PhyloTree, the portion representing Haplogroup J is essentially identical to this earlier work.

In these earlier publications on Haplogroup J, formatting of input data was done manually and then copied into an Excel spreadsheet. From there, allele alignment was also done manually. one allele at a time to produce a starting matrix. The cell entries were also manually reduced from the full allele name to a single character representation, before pattern searching could begin. All this was very time consuming and subject to small clerical errors including dropping alleles, or putting them in incorrect cells. Computer processing using Python scripts has eliminated these particular errors. Clerical errors are still possible, however, as the matrix was subjected to a transformation to tease out the evolutionary patterns. More important than small errors that get lost in the stochastic nature of the data, is the labor involved. The largest matrix ever developed in the earlier work contained 323 mtDNA sequences and under 600 unique alleles or less than 20 thousand cells used in the matrix. Development of the matrix probably consumed more than 400 hours spread out over two years time. Contrast that with the current version of mtMatrix-N of 3891 sequences by 4984 unique alleles giving over 18 million cells. It is simply not feasible to



Notes:
1) Numbers below the clade designations are the number of samples for that clade in the reference database.
2) Polymorphism designators in parentheses are typical but not definitive. For example. some sequence of that clade have back mutations at that position.
3) Polymorphism designators that are underscored appear in most/all samples in the clade but also appear in other clades. That is, they are homoplasies.
4) Random mutations result in exceptions to these rules.

Figure 12: Illustrative phylogenetic tree for mtDNA Haplogroup J. Such trees can be developed from matrix data described above. This chart comes from a earlier study using a very selective set of available data.

manually prepare such a large matrix in a timely fashion, yet mtMatrix-N was developed by one person in less than four weeks.

Maintenance and Use of mtMatrix-N

Drafts of this document have been made available to the members of the mtTree committee of the International Society of Genetic Genealogists (ISOGG) for consideration of the issues presented here. It is hoped that ISOGG will accept the challenge and provide guidance for use in future work relative to mtDNA phylogenies. It is also expected that mtMatrix-N (for all of Haplogroup N) will be maintained in a fashion similar to the way the earlier matrix for Haplogroup J has been maintained, incorporating new mtDNA sequences as they become available, and can be made available to ISOGG committee members and other researchers who may wish to use it in support of their work. Constructive criticism is always appreciated and should be sent directly to Jim Logan to JJLNV @ comcast.net.

Acknowledgments

mtMatrix-N would not be possible without the contributions made to GenBank by the many researchers and willing participants; these contributions are greatly appreciated. The availability of PhyloTree that made manageable an otherwise onerous task of organizing patterns of DNA into a meaningful structure representing a phylogeny is also gratefully acknowledged. A very special thanks goes to Ian Logan and his Checker database that provided the preprocessed GenBank data used as raw data to build the original matrix and for the updates he so diligently provides on a regular basis to facilitate its maintenance.

References

Rebecca L. Cann, Mark Stoneking & Allan C. Wilson (1987), Mitochondrial DNA and human evolution, *Nature*, 325:31-36.

GenBank (2010). A nucleotide database maintained by the National Center for Biotechnology Information, at the National Institute of Health, Washington, DC. website at http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide

Logan I (2007). A Suggested Mitochondrial Genome for 'Mitochondrial Eve'. *J Genet Geneol*, 3(20):72-77.

Logan I (2010) Checker -- an online utility providing CRS differences for human mtDNA sequences available in GenBank. A database script built into Checker contains these preprocessed differences. An earlier Greasemonkey script that served a similar function no longer being maintained but rather has been replaced with Checker scripts available through http://ianlogan.co.uk/checker/genbank.htm.

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

Logan JJ (2008b) A comprehensive analysis of mtDNA Haplogroup J. *J Genet Geneol,* 4:104-124.

Logan JJ (2009a) A Refined Phylogeny for mtDNA Haplogroup J. *J Genet Geneol*, 5:16-22.

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

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

Antonio Torroni, Theodore G. Schurr, Chi-Chuan Yang, Emoke J. E. Szathmary, Robert C. Williams, Moses S. Schanfield, Gary A. Troup, William C. Knowler, Dale N. Lawrence, Kenneth M. Weiss and Douglas C. Wallace (1992), Native American Mitochondrial DNA Analysis Indicates That the Amerind and Nadine Populations Were Founded by Two Independent Migrations, *Genetics*, 130(1):153-162.

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30:E386-E394. (See also http://www.phylotree.org/)