A Refined Phylogeny for mtDNA Haplogroup J

J. J. Logan

Abstract

This short report presents an updated version of the phylogeny for mtDNA Haplogroup J based on 253 full genome sequences plus 38 sequences that are complete in the coding region but incomplete in the control region.

Introduction

A preliminary phylogeny for Haplogroup J was presented in "The Subclades of mtDNA Haplogroup J" (Logan, 2008A). It was based on 111 full-genome sequences (FGS) of mitochondrial DNA from GenBank (Benson et al., 2004). That paper confirmed the conclusion of Palanichamy et al (2004) that selected polymorphisms in the first Hyper Variable Region (HVR1) of the control region (also referred to as the D-loop or the major non-coding region), were actually homoplasic within Haplogroup J, thus producing major errors in haplogroup prediction using HVR1-only results. Although coding region data is required for high confidence, the study also showed that a better than 95% identification of J1 and J2 clades can be achieved by including both HVR2 and HVR3 results, or using the expanded HVR2 results as provided by Family Tree DNA.

Having established an initial structure for Haplogroup J, based on 111 full genome sequences, the next step was to develop a broader perspective on the haplogroup and the results of this work were presented in the article, "A Comprehensive Analysis of mtDNA Haplogroup J" (Logan, 2008b). The analysis in that paper was based on an expanded reference database that included 7 full genome sequences that had been added to GenBank since the first analysis as well as 38 sequences that had not been included previously because they did not include control region data, although they were full-sequence otherwise. The inclusion of these sequences, covering 93% of the genome, was considered a significant addition and thus there were 156 sequences in the reference database used to refine the phylogeny as presented in that paper.

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

Received: January 19, 2009; accepted: March 4, 2009

The availability of mtDNA sequences continues to grow with over 5400 human mitochondrial "full genome sequences" currently available in GenBank, of which over 200 sequences are Haplogroup J. In addition a number of full genome sequences were made available through the Haplogroup J testing program at Family Tree DNA (J-mtDNA Project, 2008). Finally, the author's participation in an ISOGG (2008) task force to develop a standard overall mtDNA phylogeny has once again prompted a reexamination of the J phylogeny using the 291 sequences currently in his reference database. This refinement did not change any of the high-level structure of earlier phylogenies but did result in the a more than doubling of the number of subclades previously defined.

Methods

Analysis of sequences and development of a matrix from which the phylogeny was inferred are as described in previous reports (Logan, 2008a, 2008b). Sequences from GenBank were accessed through the Nucleotide search page at National Center for Biological Information (GenBank 2009) and differences with respect to the rCRS were identified using Ian Logan's Greasemonkey scripts (Logan 2009). The sequences from Family Tree DNA were reported directly as polymorphism reports. In both cases, the reports were parsed using an Excel spreadsheet and then the sequences were carefully aligned in a single matrix. Using a maximum parsimony criteria, the rows and columns of the matrix were then rearranged to identify clades and their corresponding definitions.

Results

A formal definition of the clades is given in the Table 1. Each clade name is included in a box which is indented from the left to show the hierarchical level of that clade. It is followed by a list of the polymorphisms relative to

the revised Cambridge Reference Sequence (Andrews et al 1999, Mitomap, 2008). For the convenience of those who wish to use this table to estimate the clade of a sequence using control region test results, the control region polymorphisms are shown in bold and those from the HVR2 region are further italicized. Each polymorphism is shown as a numeric position indicator preceded by a letter indicating the reference sequence allele and followed by a letter indicating the observed allele. Exceptions include insertions where there is no reference value and deletions where a "d" suffix is used indicate the absence of a nucleotide at that position. For back mutations the position number is followed by an "@" rather than repeating the reference value. The underscored polymorphisms indicate that they have significant homoplasic presence in other clades of Haplogroup J. Similarly, parentheses are used to indicate that a given polymorphism is absent in a significant number of sequence, such as due to back mutations. For the convenience of those who may wish to evaluate the support for these definitions, the number of times that the indicated set of polymorphisms occurred in the database is shown near the first column.

This data is also presented here in the form of a two-part graphic. Figure 1 shows the overall structure of mtDNA Haplogroup J, and details of the various subclades, except for the J1c subclade that is detailed within its context in Figure 2.

A matrix of the aligned and analyzed haplotypes used in the development of this phylogeny is available in the supplementary material. Note that selected columns of the matrix are lightly shaded to indicate those sequences that are complete in the coding region but not in the control region. Thus, empty cells that are shaded and correspond to control regions polymorphisms should be considered as "not known" rather than "no polymorphism." This matrix, along with the table and figures, will be periodically updated in the supplementary data files as new information becomes available.

Discussion

The purpose of this brief report is to make the updated phylogeny presented here freely available to all interested parties. However, the results must be considered a work in progress and further refinements may take place as new data is acquired and analyzed. In particular, some of the clade definitions at the end of limbs and branches must be considered tentative because they are based on small sample sizes and will be confirmed or restructured with further analysis incorporating additional data. Furthermore, the nomenclature is subject to change as a result of harmonization with other researchers. For example, an active effort is underway to harmonize this work with the updates to the tree of van Oven and Kayser (2009).

As of this writing there are three clusters that have been flagged for possible future definition as subclades. Each of these are clearly identifiable in the supplementary matrix and all three of them are marked on the graphic version of the phylogeny. However, these are not included in Table 1, which includes the definitions of the subclades.

The first issue is the apparent further subdivision of clade J1c8 that is clearly visible in the supplementary matrix. Upon closer examination it was determined that some sequences were reported to have heteroplasmic results at T16092, whereas others reported simple substitutions. It is probable that the difference is caused by different testing and/or reporting standards. For this release, this polymorphism has simply been ignored. This is of little overall significance since this occurs at the extreme of the phylogeny.

The second indication of possible refinement is the possible addition of a J1c10 based on a 16188 insertion and includes three sequences. Closer examination show that two of these three sequence are identical and the third one differs from these two at a single nucleotide position. Since they all came from the same study, it is possible that they are all from the same family and thus are not independent samples suitable for defining a clade. Thus, this clade is held in abeyance pending confirmation from additional data.

The third is an unresolved reticulation at J2a1a. It appears that G513A and A3447G could be used to define a new branch, but so could T1850C, together with the T insertion at position 310. However these two potential definitions have a substantial overlap making clear definition impossible. This situation also occurs at the extreme of the tree and will likely resolve itself as additional data is gathered.

Acknowledgements

I wish to thank Mannis van Oven for his thorough review of various forms of the phylogeny presented in this paper and pointing out the back mutation at A2706

Journal of Genetic Genealogy, 5(1):16-22, 2009

Table 1
A Phylogeny for mtDNA Haplogroup J

	R	***A73G*** , ***A263G*** , ***315.1C*** , A750G, A1438G, A2706G, A4769G, C7028T, A8860G, G11719A, C14766T, A15326G								
		JT	T4216C, A11251G, C15452A, **T16126C**							
291			J	C295T, T489C, A10398G, A12612G, G13708A, **C16069T**						
234				J1	***C462T,*** G3010A					
38					J1b	G8269A, (**C16222T**), **G16145A**, **C16261T**				
29						J1b1	G5460A, T13879C			
23							J1b1a	***C242T*** , T2158C, G8557A, (**G12007A**), **T16172C**		
8								J1b1a1	T15067C	
2									J1b1a1a	***C264T*** , T8286C, 8287.1C
5								J1b1a2	C5463T, T6911C, **C16192T**	
6							J1b1b	***C271T***		
5								J1b1b1	**C522d**, **A523d**, 10410A, **C16222@**,	
2									J1b1b1a	A2707C, **C16290T**
7						J1b2	C1733T			
4							J1b2a	T6719C, A14927G		
2						J1b3	A8460G, **A16235G**, **T16271C**			
189					J1c	***(G185A)*** , ***(G228A)*** , T14798C				
57						J1c1	***A188G***			
4							J1c1a	T6293C		
2								J1c1a1	A7245G, G8839A	
3							J1c1b	T4454C		
12							J1c1c	G10685A, T13281C, A13933G		
2								J1c1c1	***C222T***	
3								J1c1c2	G8865A	
3							J1c1d	A13032G, T14325C		
3							J1c1e	**C16366T**		
31						J1c2	***T482C*** , T3394C			
9							J1c2a	A9635C, C11623T, T13899C		
12							J1c2b	A7184G		
7								J1c2b1	G5773A	
4									J1c2b1a	T10463C
2									J1c2b1b	A5411G, G13368A, T14200C
2							J1c2c	T10454C		
30						J1c3	C13934T			
7							J1c3a	G9548A		
4								J1c3a1	T7711C	
3								J1c3a2	T9836C	
2									J1c3a2a	G14323A, G15355A
5							J1c3b	C15367T		
3								J1c3b1	G5237A, G6261A	
3							J1c3c	A2706@		
2							J1c3d	A1811G		
10						J1c4	A9632G, T12083G			
3							J1c4a	G11778A, G11914A		
2							J1c4b	A9120G		
16						J1c5	A5198G			

Table 1 continued on next page

Table 1 (continued)
A Phylogeny for mtDNA Haplogroup J

5							J1c5a	T2387C, C10192T				
4								J1c5a1	A10598G			
3							J1c5b	T11087C				
2								J1c5b1	T6681C, C12239T			
6						J1c6	C4025T					
13						J1c7	C6554T, G12127A					
11							J1c7a	C6464A, A13681G				
8						J1c8	T10084C					
2							J1c8a	A9052G				
2						J1c9	C6887T					
6					J1d	***T152C***, **C522d**, **A523d**, G7789A, A7963G						
2						J1d1	A16300G					
2						J1d2	T689C, G9123A, T10166C, G14040A, A14280G					
56			J2	(***C150T***), C7476T, G15257A								
24				J2a	***T195C***, A10499G, G11377A							
18					J2a1	***T152C***, ***A215G***, G7789A, (A13722G), A14133G, T16231C, **G16145A**, **C16261T**						
15						J2a1a	***T319C***					
5					J2a2	T6671C, A11002G, A12570G, A15679G						
3						J2a2a	C8386T					
2							J2a2a1	A235G				
33				J2b	***T152C***, C5633T, G10172A, G15812A, **C16193T**							
23					J2b1	**C16278T**						
7						J2b1a	G14569A					
4						J2b1b	T6216C					
2							J2b1b1	A3348G				
2						J2b1c	T199C					
3				J2b2	T2404C, G6962T, T10389C							
2					J2b2a	G7211A						

that I had missed. Not only have his comments improved this paper but also his collaboration has brought our respective work into basic harmonization and established a firm basis for developing a worldwide consensus for the phylogeny of Haplogroup J.

Web Resources

http://www.familytreedna.com/ Home page for Family Tree DNA.

http://www.familytreedna.com/public/J-mtDNA/ Public access page for J-mtDNA Project at Family Tree DNA.

http://www.ianlogan.co.uk/. Website that includes description of Greasemonkey scripts used to extract polymorphisms associated with full genome mtDNA sequences.

http://www.isogg.org/. Home page for the International Society of Genetic Genealogy.

http://www.mitomap.org/mitoseq.html. Reference page for the revised Cambridge Reference Sequence provided by the MitoMap organization.

http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide Search page for retrieving mtDNA sequences from GenBank.

http://www.phylotree.org/ A web page that is periodically updated to show the entire mtDNA phylogeny as it

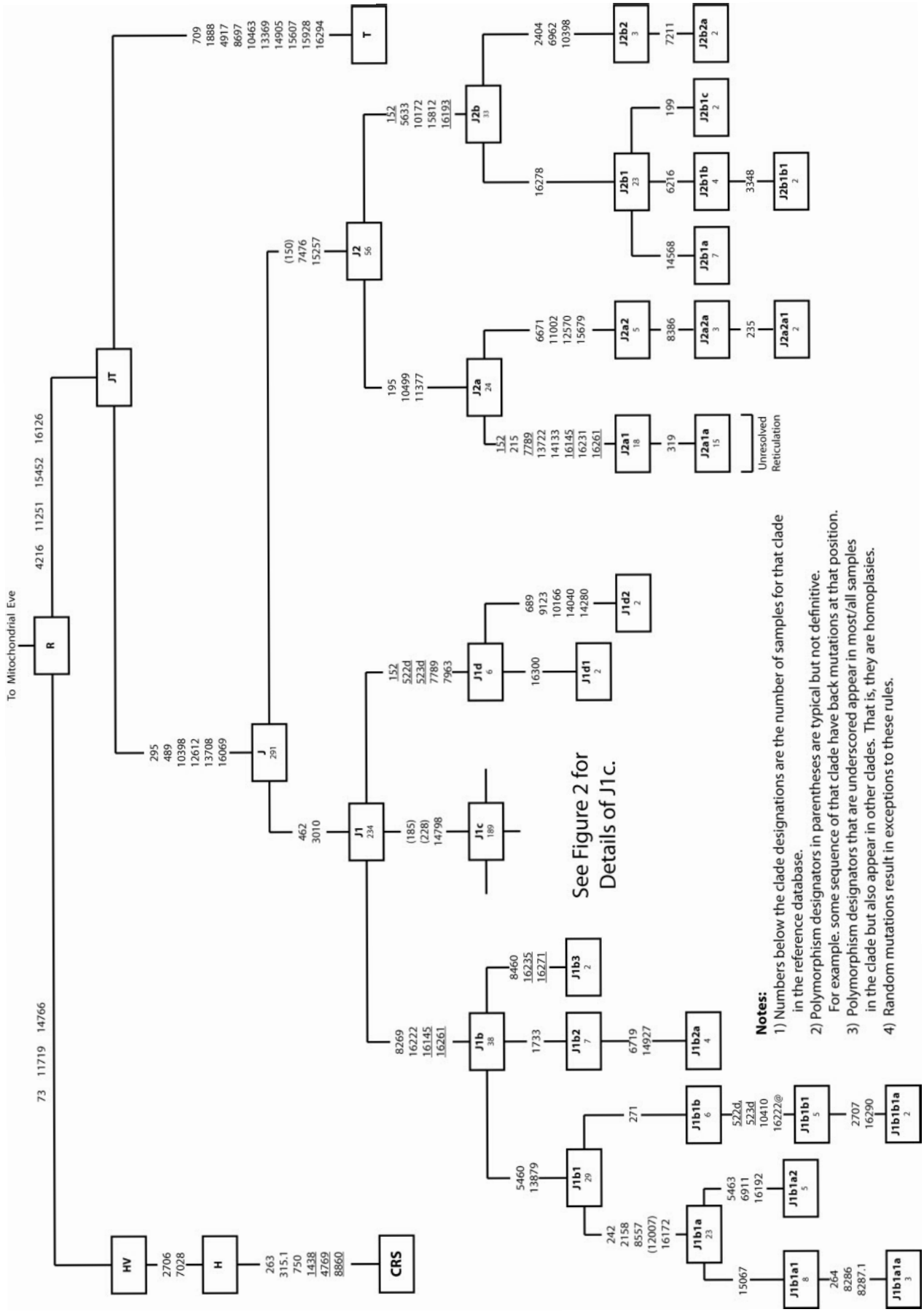

Figure 1 The phylogeny of mtDNA Haplogroup J in tree format.

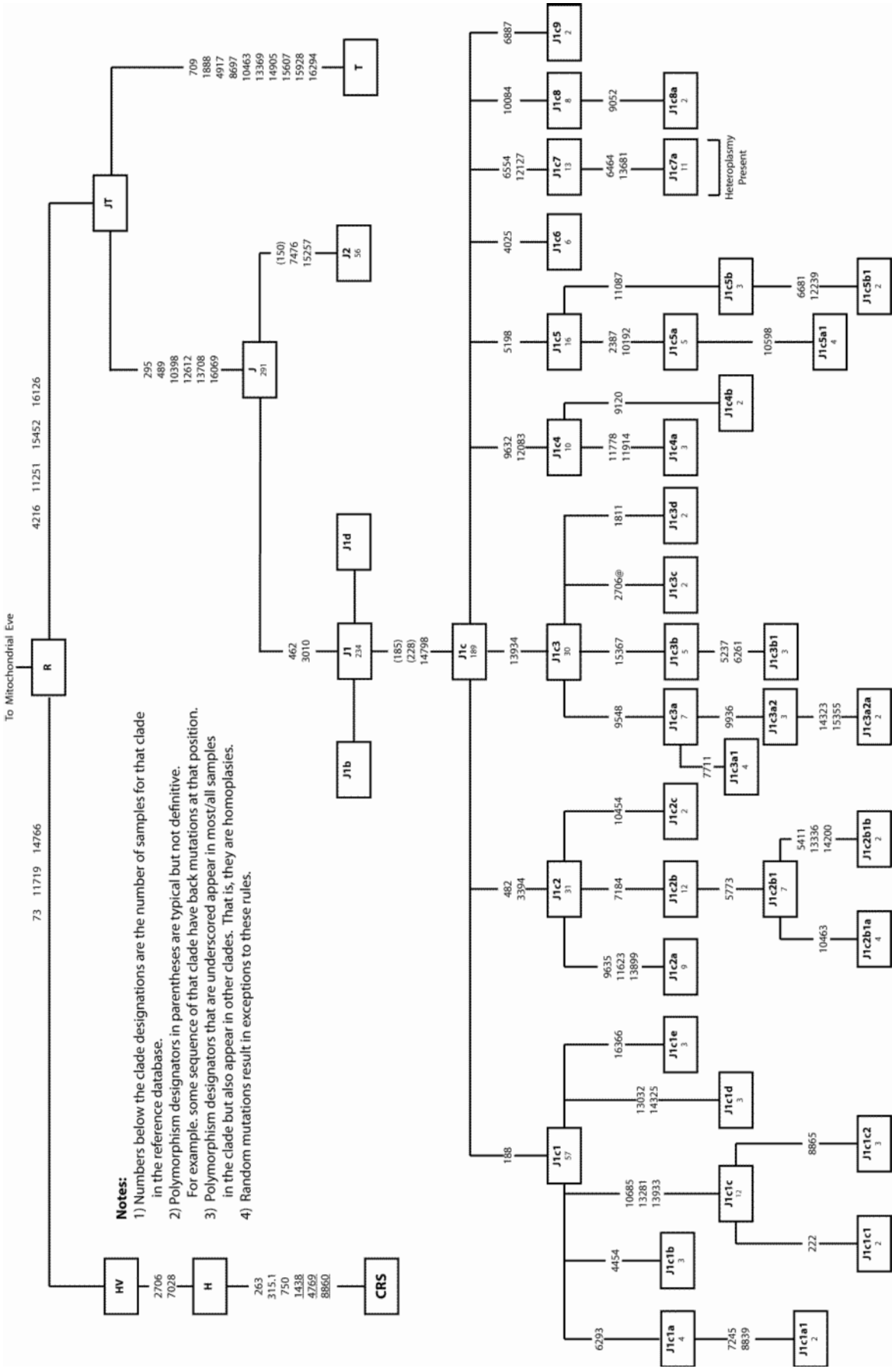

Figure 2 Details of mtDNA subclade J1c, shown in context of overall Haplogroup J.

is developed. It is maintained by Mannis van Oven at Erasmus University Medical Center, Rotterdam, The Netherlands.

Note: Corrections added 31 May 2009 and 15 July 2009.

Supplementary Material

Supplementary data is available at:

http://www.jogg.info/51/files/logansuppl.htm

References

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

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

FTDNA (2008) Family Tree DNA. See Web Resources.

GenBank (2009). A database that contains publicly available DNA sequence, including full genome sequences from human mtDNA. See Web Resources.

ISOGG (2008) International Society of Genetic Genealogy, See Web Resources.

J-mtDNA Project (2008) The J-mtDNA Project at Family Tree DNA. See Web Resources.

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 Ian (2009) Ian Logan website. See Web Resources.

MitoMap (2008) Revised Cambridge Reference Sequence (rCRS) of the Human Mitochondrial DNA. See Web Resources.

Palanichamy M, Sun C, Agrawal S, Bandelt HJ, et al (2004) Phylogeny of mitochondrial DNA Macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet*, 75:966-975.

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30:E386-E394. See also Phylotree.org under Web Resource.