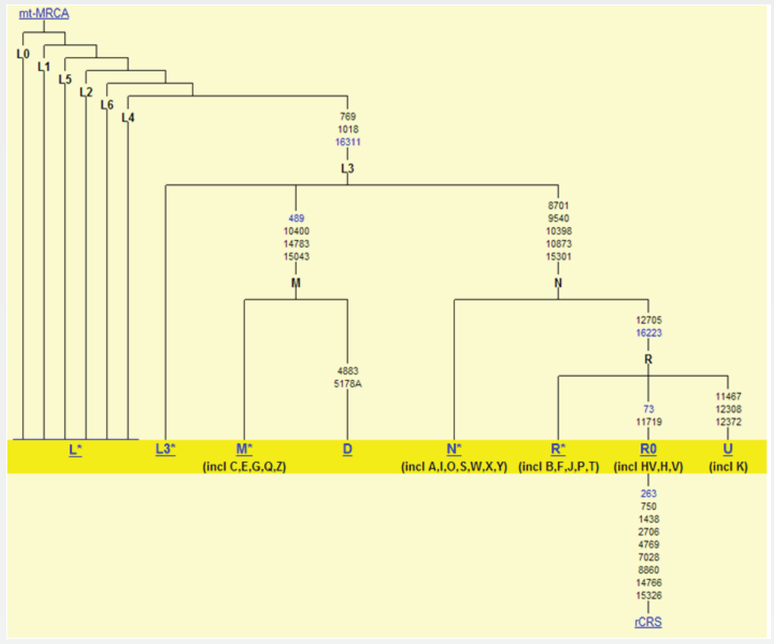
An automated method for deriving mitochondrial DNA (mtDNA) haplogroups based on changes within the Hypervariable Regions

V.L. Vance, J.J. Bryan, M.R. Szczepanski, A.B. Carter, C.L. Mouritsen
Sorenson Genomics, Salt Lake City, Utah

Abstract

In 2009, van Oven and Kayser described a comprehensive phylogenetic tree of human mtDNA variation which has been made accessible at www.Phylotree.org. This phylogenetic tree is based on both coding and control (hypervariable) regions. Van Oven and Kayser identified the variances from the revised Cambridge Reference Sequence (rCRS) which define an individuals' haplotype and corresponding mtDNA haplogroup. A new computer-based method has been developed for assigning mtDNA haplogroups using the variances and haplogroup nomenclature described by van Oven and Kayser. This new method makes use of Structured Query Language (SQL) and a mathematical algorithm that allows for the reliable determination of one's haplogroup based solely on mtDNA sequence from the Hypervariable Regions (HVR). The SQL-based algorithm combines a database search process with a method that walks stepwise through the phylogenetic tree, which is rooted with rCRS at the first position. Using a novel scoring method to account for the number and stability of the markers that define each haplogroup, an individual's HVR differences from rCRS are compared with the haplogroup designations defined in the mtDNA Phylotree. The algorithm has a high degree of reliability even when potential "back-mutations" and/or recent mutations are observed at key haplogroup defining positions, in which case a haplogroup is assigned based on likelihood and match criteria thresholds defined within the algorithm. In instances of ambiguous calls, the algorithm has the ability to select the nearest parental haplogroup in the tree. This new method was validated by comparing the haplogroups assigned by our method to the haplogroups assigned by van Oven and Kayser for samples with mtDNA haplotypes published in Phylotree. This comparison showed concordance of our method to be greater than 95%. Use of our system can accurately and quickly estimate over 800 different mtDNA haplogroups across the mtDNA tree using only rCRS differences within Hypervariable Region 1, over 1000 haplogroups using Hypervariable Regions 1 and 2, and nearly 1100 haplogroups using Hypervariable regions 1, 2 and 3. Since Phylotree is continually being updated as new data are published, we have incorporated a parsing tool that allows the program to be updated as the science progresses.

Figure 1

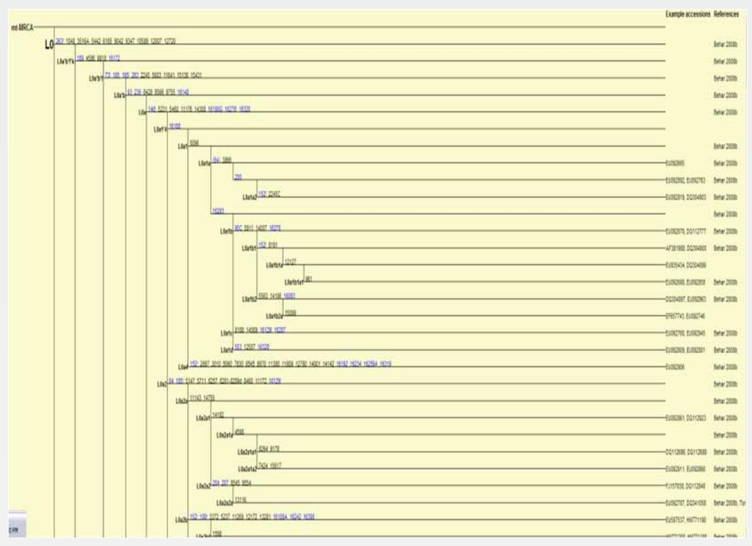


Global human mtDNA phylogenetic tree from www.phylotree.org. The tree is further subdivided into eight trees that can be accessed by clicking on links associated with the haplogroups on the branches of the tree. The differences from the rCRS that distinguish these subtrees are listed with those from the HVR being in blue.

Background

The mitochondrial DNA (mtDNA), being 16,569 bases in length, was the first complete genetic unit of the human genome to be fully sequenced. Due to it maternal, non-recombining inheritance pattern and its high mutation rate, it has been an integral part of understanding human anthropology and phylogenetics. The mtDNA is commonly broken into 2 main parts, the 'coding region' that contains sequence for protein synthesis and the 'control region', which does not. The control region of the mitochondria makes up only a small percentage of the mtDNA genome and has been documented to be more highly polymorphic than the coding regions, particularly in the hypervariable region (HVR) (Wakeley 1993; Meyer et al. 1999). Thus this region tends be used more predominantly for identification and research purposes and why researchers tend to sequence portions of this region for their studies, minimizing cost in providing information. Base pair differences when comparing a sample sequence data to the revised Cambridge Reference Sequence (rCRS) (Anderson et al., 1981; Andrews et al., 1999) can be used to predict haplogroups. Most types of prediction systems have traditionally used the HVR motifs and conducting a database search of haplotypes with designated haplogroups. Using these approaches, only higher level haplogroup assignments can be made. In 2009, a comprehensive phylogenetic tree (Fig 1), based on both coding and control region mutations, was published (van Oven and Kayser, 2009). Previous to this publication the only phylogenetic available were either based only on the coding region (Ruiz-Pesini et al., 2007) or only on HVR. This phylotree also includes haplogroup nomenclature and is accessible on the internet. It is updated every six months as new information is found. Haplogroup predictions using HVR mutations can be made with more confidence when their assignments are being based off of this phylotree.

Figure 2



A portion of the build 10 L subtree from www.phylotree.org. The haplogroup assigned to a branch is listed on the right and the mutational difference defining the branch are above. To the left of each branch is listed GenBank accessioned numbers for representative samples as well as references to publications where the branch was described or the haplogroup nomenclature proposed.

Methods

We designed a system using a SQL based algorithm that combines a database search method with a method that walks through the comprehensive phylogenetic tree from www.phylotree.org. The tree is first inverted with the rCRS haplogroup as the root to simplify walking through the tree. Each haplogroup on the tree is given a marker count score based on the number of defining markers and whether those markers are considered stable, unstable or preliminary. Back mutations are counted as well. The unknown individual's markers are compared with those of the haplogroups' and are given a marker count score based on how many markers they share. A haplogroup is determined by looking at the total marker score as well as the percent. Possibilities are eliminated if they are below a 70% match or if the defined marker score is only 8 and they are less than a 100% match. The parent haplogroup is selected if the calls are ambiguous between subclades. The predictor is designed so that it can easily be updated as updates to the phylotree are made. In order to determine the accuracy of this predictor, we used the HVR differences of the sequences listed on tips of the branches of the phylotree build 10 (see Fig2). We selected 1000 sequences which represented 833 unique haplogroups some of which could not be predicted exactly because they had no defining differences from the HVR (see Table 1). We chose sequences that would represent the majority of the clades of the tree. The sequences were downloaded from the phylotree website and then compared to the rCRS in Sequencher in order to obtain the differences. We ran the profiles of these sequences through our predictor three validation sets with varying amounts of information. We ran one set using only HVR 1(16000-16579) differences, another set using both HVR 1 (16000-16579) and 2 (1-390) differences, and then the last set using differences from HVR 1 ((16000-16579), 2 (1-390), and 3(391-590) as show in Table 2. We then compared the haplogroups predictions from each validation set with those assigned to the sequences by their placement on the tree.

Table 1

	# Haplogroups	# Haplogroups with direct HVR definition	# Haplogroups with NO direct HVR definition	# Representative Samples
Phylotree	1945	1095	850	2926
Validation Set	833	713	120	1000
%	42.8	65.1	14.1	33.9

Table 2

Validation Set	Sequence Region
1	HVR I (16000-16579)
2	HVR I and II (16000-16579; 1-390)
3	HVR I, II, and III (16000-16579;1-590)

Results

Validation Set 1

	X	P	C	I	U	Total		Total Concordant Call (X+P+C)
HVR Defined	543	187	24	36	53	843		754
%	64.4	22.2	2.8	4.3	6.3			89.4
non HVR-Defined	N/A	130	9	5	13	157		144
%	N/A	82.8	5.7	3.2	8.3			88.5
Combined	543	317	33	41	66	1000		893
%	54.3	31.7	3.3	4.1	6.6			89.3

Validation Set 2

	X	P	C	I	U	Total		Total Concordant Call (X+P+C)
HVR Defined	743	74	16	19	0	843		833
%	87.1	8.8	1.9	2.3	0			97.7
non HVR-Defined	N/A	133	14	10	0	157		147
%	N/A	84.7	8.9	6.4	0			93.6
Combined	743	207	30	29	0	1000		971
%	74.3	20.7	3	2.9	0			97.1

Validation Set 3

	X	P	C	I	U	Total		Total Concordant Call (X+P+C)
HVR Defined	754	62	14	13	0	843		830
%	89.4	7.4	1.7	1.5	0			98.5
non HVR-Defined	N/A	138	16	3	0	157		154
%	N/A	87.9	10.2	1.9	0			98.1
Combined	754	200	30	16	0	1000		984
%	89.4	20	3	1.6				98.4

X Exact Haplogroup Match
P Parent Haplogroup Match
C Close within Clade Haplogroup Match
I Incorrect Haplogroup Match
U Unassigned--More than one prediction possible

Conclusion

This system has the capability of successfully assigning/predicting roughly 1,100 different mtDNA haplogroups using hypervariable region sample data alone, and does so with ~98% accuracy. The amount of HVR data imputed into the system is directly connected to the number of haplogroups that can be predicted as well as the accuracy of the assignment. With only HVR I data, 846 haplogroups can be predicted with an 89% accuracy. The majority of inaccurate or unassigned calls can be resolved simply by adding HVR II data. 1,057 haplogroups can be predicted with 97% accuracy if data from both HVR I and HVR II are present. Having HVR III data in addition to HVR I and HVR II only improves accuracy to 98% and increases the number of haplogroups that can be predicted to 1,095. Most of the cases where a wrong assignment is made using this system would be hard to call accurately using any method without additional data from the coding regions. This system also includes a parsing function that allows it to be updated and validated very quickly when new data to the tree becomes available and allows fast haplogroup reassignment of even the largest mtDNA databases of samples.

Citations

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F. 1981. Sequence and organization of the human mitochondrial genome. Nature 290(5806):457-465.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23(2):147.

Meyer S, Weiss G, von Haeseler A. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hyper variable regions I and II of human mtDNA. Genetics 152(3):1103-1103-10.

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

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutation 30(2):E386-E394. doi:10.1002/humu.20921

Wakeley J. 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 3 7(6):613-623

Sorenson Genomics™
2495 S. West Temple, Salt Lake City, Utah 84115
SorensonGenomics.com