

## **Comparison of Mitotyper Rules and Phylogenetic-based mtDNA Nomenclature System**

Deborah Polansky<sup>1</sup>, Bobi K. Den Hartog<sup>2</sup>, John W. Elling<sup>2</sup>, Constance L. Fisher<sup>1</sup>,  
Russell B. Kepler<sup>2</sup>, Bruce Budowle<sup>1</sup>

1. FBI Laboratory, 2501 Investigation Parkway, Quantico, VA 22135, USA

2. Mitotech LLC, 590 Monte Alto, Santa Fe, NM 87501, USA

## **ABSTRACT**

A consistent nomenclature scheme is necessary to characterize a mitochondrial DNA (mtDNA) haplotype. A standard nomenclature, called the Mitotyper Rules™, has been developed that applies typing rules in a hierarchical manner reflecting the forensic practitioner's nomenclature preferences. In this work, an empirical comparison between the revised hierarchical nomenclature rules and the phylogenetic approach to mtDNA type description has been conducted on 5173 samples from the European Mitochondrial DNA Population database (EMPOP) in order to identify the degree and significance of any differences. The comparison of the original EMPOP types and the results of retyping the sequences using the Mitotyper Rules demonstrate a high degree of concordance between the two alignment schemes. Differences in types resulted primarily because the Mitotyper Rules selected an alignment with the fewest number of differences compared with the rCRS. In addition, several identical regions were described in more than one way in the EMPOP dataset but were described consistently by the Mitotyper Rules, demonstrating a limitation of a solely phylogenetic approach in that it may not consistently type non-haplogroup-specific sites. Using a rules-based approach, commonly occurring as well as private variants are subjected to the same rules for naming, which is particularly advantageous when typing partial sequence data.

An alignment strategy is required to be able to search and compare mitochondrial DNA (mtDNA) sequences contained within a reference database. The current forensic mtDNA analysis is based on sequencing of at least the hypervariable regions of the non-coding portion of the mtDNA genome. These approximately 600 bases of sequence are highly informative for differentiating individuals. However, it is difficult to convey the results simply as a string of bases. Instead nomenclature strategies have been developed that identify a limited number of select bases as opposed to recording all bases in these regions. These limited bases are listed as differences with respect to the standard mitochondrial DNA reference sequence, the revised Cambridge Reference Sequence (rCRS) (1). It is essential that the alignment and resulting naming of variant sites be consistent for searching a population database for statistical inferences of the rarity of an evidentiary mtDNA sequence (2). The complication confronting any nomenclature scheme is to consistently select among multiple possible alignments that can occur for some mtDNA sequences.

Two nomenclature approaches have been proposed to describe mtDNA variation: a rule-based hierarchical approach and a system based on phylogenetic relationships. Wilson et al (3) proposed a hierarchical approach comprised of a set of rules for consistently selecting one alignment over other possible alignments. The main criterion of the Wilson et al. approach is parsimony, the selection of an alignment that has the fewest number of differences compared with the rCRS. If there are still multiple alignments with an equal number of differences, additional rules are executed to select a preferred alignment. The principles of a hierarchical approach have been endorsed by the

Scientific Working Group on DNA Analysis Methods (SWGDM) and incorporated into the manually typed SWGDAM mtDNA population database (4,6). However, the Wilson et al (3) recommendations were not strictly followed which is a point that Bandelt and Parson (6) did not consider in their analysis of the robustness of the SWGDAM database nomenclature.. Primarily, the historical manually-derived alignments did not abide by the rule favoring insertions and deletions (indels) over substitutions.

A more intuitive hierarchical nomenclature system has been developed (5) that still is based on the parsimony criterion, has several constraint criteria but favors substitutions over indels. Within the SWGDAM database 99.92% of the haplotypes were consistent with this enhanced hierarchical approach. The remaining 0.08% of sequence differences was converted to be consistent with this more effective nomenclature approach (JOFS-09-010 - In Press).

A rule-based alignment scheme is very desirable for characterizing short fragments derived from highly degraded samples such as bone and teeth or telogen hairs which are commonly analyzed in forensic mtDNA casework. In addition, the hierarchical approach is robust because it provides for a stable nomenclature which will not change as new mtDNA types are discovered (5). New rules can be created to address novel polymorphic regions without reanalysis of previously typed sequences in the dataset.

In contrast to the hierarchical approach, an evolutionary phylogenetic approach has been proffered (6). Some limitations of this phylogenetic approach are: the nomenclature is not stable in that novel mtDNA types will require a reanalysis of previously typed samples in a dataset, not all variant sites are addressed (only phylogenetic signatures), and the routine user is required to have substantial intimate

knowledge of phylogenetic relationships of mtDNA which is not operationally practical or likely (5). Variants that occur in one to only a few individuals may not correlate to those sites that define specific haplogroups, either because the sampling is too small or they are homoplastic. For these variants there are no consistent phylogenetic rules applied for nomenclature and variation in nomenclature can occur (see examples below).

Forensic scientists, unlike evolutionary biologists, require exact matching sequences when searching a database, regardless of the evolutionary events that generated the mtDNA sequence of interest. The merits of each nomenclature system can be debated but nomenclature consistency should be one of the primary criteria for a database construct. In this work, an empirical comparison between the revised hierarchical nomenclature rules and the phylogenetic approach to mtDNA type description has been conducted on the same data set in order to identify the degree and significance of any differences.

To this end, the 5173 phylogenetically-aligned haplotypes contained within the European Mitochondrial DNA Population database (EMPOP) (7) were re-aligned using the Mitotyper Rules implemented in the Mitotyper Software and the results compared.

## **Results**

Only 51 of the 5173 samples, representing 23 different Mitotyper regions, differed between the two systems, most at a single polymorphic region (Table 1).

There were 44278 polymorphic regions in the EMPOP data set as determined by Mitotyper. The overall difference at these 44278 regions between the EMPOP types and the Mitotyper Rules is only 0.12%. As expected, most of the differences between the two nomenclatures resided within the homopolymeric and dinucleotide repeat regions located

at np 16180-16199 (9 regions observed in 14 samples), np 300-318 (6 regions observed in 11 samples), np 513-526 (2 regions in 7 samples) and np 568-582 (2 regions in 5 samples). These differences are primarily due to Mitotyper selection of an alignment with the least number of differences with respect to the rCRS (Table 2).

In the EMPOP data there were a few examples of inconsistent naming of identical regions. For instance there was a difference at np16180-16199 in which samples USA0600921, SVN0600003, ITA0500260 and USA0601079 were described one way and sample POL0600375 was named differently. Neither type corresponded to the Mitotyper result. In the np 300-318 region there were two EMPOP named sequences that were named differently; SVN0600092 and HUN0500202, the latter being consistent with Mitotyper. These and two additional examples of differences are shown in Table 3.

Outside the homopolymeric regions, several samples aligned by Mitotyper were consistently discordant with the EMPOP phylogenetic approach. Such differences in EMPOP (0.06%) occur because these nucleotide changes are variants that reside primarily at regions that do not define haplogroups (private variants) (Table 3). For example, sample POL0600267 was typed with a 42.1C alignment and samples HUN0600207, AUT0500255, KEN0500083, KEN0500041, KEN0500073, and USA0600739 were typed as 44.1C by the EMPOP approach (the 44.1C type is consistent with the Mitotyper result). Thus, if a forensic evidence sequence of approximately 100 bases in length were typed as a 42.1C only one match would be observed, even though there should be a total of 7 matching sequences. The same would hold true for the 57.1C (GRC0500319 and GRC0500123) and 58C, 60.1T (HUN0500135, GRC0500023, GRC0500024, HUN0500250, HUN0500193, ITA0500300, GRC0500047, KEN0500066,

and HUN0500045) EMPOP typings where only two matches would be found if the more parsimonious 57.1C type were used (consistent with Mitotyper). The Mitotyper software does not suffer the potential for nomenclature inconsistency at private mutations, because it names with a set of rules all variants, not just phylogenetic signature ones.

## **Conclusions**

This study shows that the two nomenclature approaches, the hierarchical Mitotyper Rules and the phylogenetic approach used in EMPOP, show a high degree of concordance. The two systems generally name haplotypes in the same manner even though they use different strategies. In the majority of cases, searching a reference database for matching sequences with a type derived from either nomenclature will not result in missed sequences that could have aligned similarly. However, since the EMPOP phylogenetic approach does not have formal rules for typing private variant sites, the possibility of inconsistent alignments for concordant regions exists. This inconsistency may exacerbate the problem of searching a database for concordant types with a partial sequence derived from a forensic sample. The Mitotyper Rules does not have these limitations because it is designed for forensic applications. The Mitotyper Rules and the software implementing them provides consistency of nomenclature for the practitioner and provides a common functional system within and among laboratories. Therefore, we strongly recommend that all databases be reviewed with an automated system such as the Mitotyper software as a quality control to minimize inconsistencies.

Lastly, Wilson et al (3) recommended full text searching of a mtDNA type within a reference database to eliminate the effect of nomenclature/alignment inconsistencies (2). We continue to endorse the concept of full text for searching a database for matching

sequences as it does not require any nomenclature or alignment strategy. However, full text searching may not account for alignments that are one or more bases away from the searched sequence. Such alignment information may be useful to the forensic community when considering potential matches in missing person cases. In such cases, these searches are unlikely to occur through an EMPOP database search. Instead, the search will be in a database constructed of mtDNA sequences from personal items and maternal family members of the missing person. Alternate hypotheses to consider could be that the two sequences originate from two maternally related individuals and the difference is due to a mutation or that these two sequences arise from two unrelated individuals. In this scenario, one could argue that an evolutionary approach would be a better model. However, as the sequence information is reduced, as might occur for degraded and contaminated samples, it may not be possible to derive a phylogenetic signature from the limited information. Therefore, consistency is still a guiding principle for forensic evidence. Regardless, the study herein demonstrates that both approaches, hierarchical and phylogenetic, are highly concordant and results would likely be the same for the majority of haplotypes, even for kinship analyses.

### ***Acknowledgments***

This is publication number 08-16 of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only, and inclusion does not imply endorsement by the Federal Bureau of Investigation. The authors wish to thank Patricia Aagaard, Dr. Alice Isenberg and Dr. Leslie McCurdy for their review and editorial contributions.

## References

1. Andrews S, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, and Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 1999;23:147.
2. Budowle B, Wilson MR, DiZinno, JA, Fasano C, Holland MM, Monson KL. Mitochondrial DNA regions HVI and HVII population data. *For Sci Int* 1999;103:23-35.
3. Wilson MR, Allard MW, Monson KL, Miller WP, Budowle B. Recommendations for consistent treatment of length variants in the human mtDNA control region. *For Sci Int* 2002;129:35-42.
4. Scientific Working Group on DNA Analysis, (April 2003) Guidelines for mitochondrial DNA (mtDNA) nucleotide sequence interpretation. *For Sci Comm*.  
<http://www.fbi.gov/hq/lab/fsc/backissu/april2003/swgdammitodna.htm>
5. Budowle B, Fisher CL, Polanskey D, Den Hartog BK, Kepler RB, Elling JW. Stabilizing mtDNA sequence nomenclature with an operationally efficient approach. *Forensic Sci Int Genetics Supplement Series* 2008;1:671-673.
6. Bandelt HJ, Parson W. Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int J Legal Med* 2008;122:11-21.
7. Mitochondrial DNA Control Region Database. Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria. Accessed April 25, 2008. <http://www.empop.org>

TABLE 1 - Results of types in EMPOP versus types generated by the Mitotyper Rules™.

rCRS range (np)	# of samples with EMPOP types different from Mitotyper types	# of regions with EMPOP types different from Mitotyper types	# of regions with more than one EMPOP type for the same mtDNA region
16180-16199	14	9	1
37-51	1	1	1
57-62	9	1	1
242-255	3	1	0
300-318	11	6	1
455-461	2	1	0
513-526	7	2	0
568-582	5	2	0
Totals	52	23	4

Note: KEN0500056 is counted twice because it has differences in two regions. The total number of unique samples shown is 51.

TABLE 2- Differences between types generated by EMPOP and by the Mitotyper Rules.

SAMPLE NAME	EMPOP ALIGNMENT: rCRS AND TYPE rCRS np 16180-16199	MITOTYPER ALIGNMENT: rCRS AND TYPE rCRS np 16180-16199
<b>ITA0500260</b>	AAAACCCCTCC-CCATGCTT	AAAACCCCTCCC-CATGCTT
<b>SVN0600003</b>	AAAACCCCTCCCTCATGCTT	AAAACCCCTCCCTCATGCTT
<b>USA0600921</b>	16189 T C	16189 T C
<b>USA0601079</b>	16191.1 - C 16192 C T	16192.1 - T
<b>ESP0600264</b>	AAAACCCCTCCCCATGCTT	AAAACCCCTCCCCATGCTT
<b>HUN0600051</b>	AAAACCTCCCCC-ATGCTT 16186 C T 16189 T C 16193 C -	AAAACCTCC-CCCCATGCTT 16186 C T 16189 T -
<b>USA0600888</b>	AAAACCCCTCCCC-ATGCTT	AAAACCCCTCCCCATGCTT
<b>USA0600966</b>	AAAACCCCTCCCCCATGCTT 16183 A C 16188 C T 16189 T C 16193.1 - C	AAAACCCCTCCCCCATGCTT 16183 A C 16187.1 - T 16189 T C
<b>DNK0600006</b>	AAAACCCCTCCCC-A AACCCTCCCCCA 16182 A C 16183 A C 16188 C T 16189 T C 16193.1 - C	AAAACCCCTCCCCATGCTT AACCCTCCCCCATGCTT 16182 A C 16183 A C 16187.1 - T 16189 T C
<b>ESP0600272</b>	AAAACCCCTCCCC-ATGCTT AAAACCCCTCCCCATGCTT 16183 A C 16189 T C 16191 C T 16193.1 - C	AAAACCCCTC-CCCATGCTT AAAACCCCTCCCCATGCTT 16183 A C 16189 T C 16190.1 - T
<b>ESP0600235</b>	AAAACCCCTCCCCATGCTT AAAACCTCCCCC-ATGCTT 16185 C T 16189 T C 16193 C -	AAAACCCCTCCCCATGCTT AAAACCTCC-CCCCATGCTT 16185 C T 16189 T -
<b>POL0600375</b>	AAAACCCCTCCCC-ATGCTT AAAACCCCTCCCTCATGCTT 16189 T C 16193 C T 16193.1 - C	AAAACCCCTCCC-CATGCTT AAAACCCCTCCCTCATGCTT 16189 T C 16192.1 - T
<b>VNM0500147</b>	AAAACCCCTCCCCATGCTT ACCCCTCCCCC-ATGCTT 16181 A C 16182 A C 16183 A C 16189 T C 16193 C -	AAAACCCCTCCCCATGCTT A-CCCCCTCCCCATGCTT 16181 A - 16182 A C 16183 A C 16189 T C
<b>JPN0500063</b>	AAAACCCCTCCCC-ATGCTT AAAACCCCTCCCCCGCTT 16183 A C 16189 T C 16193.1 - C 16194 A C 16195 T C	AAAACCCCTCCCCAT-GCTT AAAACCCCTCCCCCGCTT 16183 A C 16189 T C 16194 A C 16195 T C 16195.1 - C

TABLE 2- Differences between types generated by EMPOP and by the Mitotyper Rules. - cont.

SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE np 37-51	MITOTYPER ALIGNMENT WITH rCRS AND TYPE np 37-51
POL0600267	AGCTCT-CCATGCATT AGCTCTCCCATGCATT 42.1 - C	AGCTCTCC-ATGCATT AGCTCTCCCATGCATT 44.1 - C

  

SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE np 57-62	MITOTYPER ALIGNMENT WITH rCRS AND TYPE np 57-62
GRC0500023	TTTT-CG	T-TTTCG
GRC0500024	TCTTTCG	TCTTTCG
GRC0500047	58 T C	57.1 - C
HUN0500045	60.1 - T	
HUN0500135		
HUN0500193		
HUN0500250		
ITA0500300		
KEN0500066		

  

SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE np 242-255	MITOTYPER ALIGNMENT WITH rCRS AND TYPE np 242-255
KEN0500065	CAATTGAATGTCTG	CAATTGAATGTCTG
KEN0500094	CAATTAA-TGTCTG	CAATT-AATGTCTG
USA0600925	247 G A 249 A -	247 G -

  

SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE np 300-318	MITOTYPER ALIGNMENT WITH rCRS AND TYPE np 300-318
GRC0500158		
USA0600843	AAACCCCCCTCCCCCGCT	AAACCCCCCTCCCCCGCT
POL0600388	AAACCCCCCCCC--GCT	AAACCCCCC-CCCCCGCT
HUN0600362	310 T C 315 C -	310 T -
VNM0500039	AAACCCCCCTCCCCCGCT	AAACCCCCCTCCCCCGCT
VNM0500097	AAACCCCCCCCC--GCT 310 T C 314 C - 315 C -	AAACCCCCC-CCCC-GCT 310 T - 315 C -
SVN0600092	AAACCCCCC-TCCCC-GCT AAACCCCCCTCCCCCGCT 309 C T 309.1 - C 315.1 - C	AAACCCCC-CTCCCC-GCT AAACCCCCCTCCCCCGCT 308.1 - T 315.1 - C
BEL0600071	AAACCCCCC--TCCCC-GCT AAACCCCTCCCTCCCCCGCT 307 C T 309.1 - C 309.2 - C 315.1 - C	AAACCCC-CCC-TCCCC-GCT AAACCCCTCCCTCCCCCGCT 306.1 - T 309.1 - C 315.1 - C
GRC0500193	AAACCCCCCTCCCCCGCT	AAACCCCCCTCCCCCGCT
GRC0500219	AAACCCCCCCCC--CT 310 T C 315 C - 316 G -	AAACCCCCC-CCCC-CT 310 T - 316 G -
HUN0500313	AAACCCCCCTCCCCCGCT AAACCCCCCCCC--CT 310 T C 314 C - 315 C - 316 G -	AAACCCCCCTCCCCCGCT AAACCCCCC-CCCC--T 310 T - 316 G - 317 C -

TABLE 2- Differences between types generated by EMPOP and by the Mitotyper Rules. - cont.

SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE np 455-461	MITOTYPER ALIGNMENT WITH rCRS AND TYPE np 455-461
<b>KEN0500015</b>	T--CCCC-TC	T---CCCCTC
<b>KEN0500056</b>	TTTCCCCCTC 455.1 - T 455.2 - T 459.1 - C	TTTCCCCCTC 455.1 - T 455.2 - T 455.3 - C
SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE np 513-526	MITOTYPER ALIGNMENT WITH rCRS AND TYPE np 513-526
<b>HUN0600387</b>	GCACACACACACCG	GCACACACACACCG
<b>KEN0500047</b>	ACACACACAC--CG	--ACACACACACCG
<b>KEN0500056</b>	513 G A	513 G -
<b>KEN0500074</b>	523 A -	514 C -
<b>USA0600909</b>	524 C -	
<b>USA0600964</b>		
<b>GRC0500033</b>	GCACACACACAC----CG GCGCACACACACACACCG 515 A G 524.1 - A 524.2 - C 524.3 - A 524.4 - C	GC----ACACACACACCG GCGCACACACACACACCG 514.1 - G 514.2 - C 514.3 - A 514.4 - C
SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE np 568-582	MITOTYPER ALIGNMENT WITH rCRS AND TYPE np 568-582
<b>DEU0600108</b>	CCCCC--ACAGTTTAT	CCCCCACC--AGTTTAT
<b>GRC0500242</b>	CCCCCCCCCAGTTTAT	CCCCCCCCCAGTTTAT
<b>GRC0500260</b>	573.1 - C	574 A C
<b>HUN0600168</b>	573.2 - C 573.3 - C 574 A C	575.1 - C 575.2 - C 575.3 - C
<b>GRC0500316</b>	CCCCC--ACAGTTTAT CCCCCCCCCAGTTTAT 573.1 - C 573.2 - C 574 A C	CCCCCACC--AGTTTAT CCCCCCCCCAGTTTAT 574 A C 575.1 - C 575.2 - C

TABLE 3 - Sequences found in EMPOP with multiple types.

SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE	SAMPLE NAME	EMPOP ALIGNMENT WITH rCRS AND TYPE	MITOTYPER ALIGNMENT WITH rCRS AND TYPE
rCRS np 16180-16199				
<b>USA0600921</b>	CTCC-CCATGCTTACAAGCAA (rCRS)	<b>POL0600375</b>	CTCCCC-ATGTTACAAGCAA (rCRS)	AAAACCCCCTCCC-CATGCTT (rCRS)
<b>SVN0600003</b>	CCCCCTCATGCTTACAAGCAA		CCCCCTCATGCTTACAAGCAA	AAAACCCCCCCCCTCATGCTT
<b>IT A0500260</b>	16189 T C		16189 T C	16189 T C
<b>USA0601079</b>	16191.1 - C		16193 C T	16192.1 - T
	16192 C T		16193.1 - C	
rCRS np 300-318				
<b>SVN0600092</b>	CAAACCCCCC-TCCCC-GCT (rCRS)	<b>HUN0500202</b>	CAAACCCCCC-CTCCCC-GCT (rCRS)	CAAACCCCCC-CTCCCC-GCT (rCRS)
	CAAACCCCCCTCTCCCCCGCT		CAAACCCCCCTCTCCCCCGCT	CAAACCCCCCTCTCCCCCGCT
	309 C T		308.1 - T	308.1 - T
	309.1 - C		315.1 - C	315.1 - C
	315.1 - C			
rCRS np 37-51				
<b>POL0600267</b>	CTCACGGGAGCTCT-CCATGC (rCRS)	<b>HUN0600207</b>	CTCACGGGAGCTCTCC-ATGC (rCRS)	CTCACGGGAGCTCTCC-ATGC (rCRS)
	CTCACGGGAGCTCTCCCATGC	<b>AUT0500255</b>	CTCACGGGAGCTCTCCCATGC	CTCACGGGAGCTCTCCCATGC
	42.1 - C	<b>KEN0500083</b>	44.1 - C	44.1 - C
		<b>KEN0500041</b>		
		<b>KEN0500073</b>		
		<b>USA0600739</b>		
rCRS np 57-62				
<b>GRC0500319</b>	T-TTTC (rCRS)	<b>HUN0500135</b>	TTTT-C (rCRS)	T-TTTC (rCRS)
<b>GRC0500123</b>	TCTTTC	<b>GRC0500023</b>	TCTTTC	TCTTTC
	57.1 - C	<b>GRC0500024</b>	58 T C	57.1 - C
		<b>HUN0500024</b>	60.1 - T	
		<b>HUN0500250</b>		
		<b>HUN0500193</b>		
		<b>IT A0500300</b>		
		<b>GRC0500047</b>		
		<b>KEN0500066</b>		
		<b>HUN0500045</b>		