Int J Legal Med (2008) 122:11–21
DOI 10.1007/s00414-006-0151-5

ORIGINAL ARTICLE

Consistent treatment of length variants in the human mtDNA control region: a reappraisal

H.-J. Bandelt · W. Parson

Received: 20 March 2006 / Accepted: 15 December 2006 / Published online: 9 March 2007

Abstract In forensic science, as well as in molecular anthropology and medical genetics, human mitochondrial DNA (mtDNA) variation is being recorded by aligning mtDNA sequences to the revised Cambridge reference sequence (rCRS). This task is straightforward for the vast majority of nucleotide positions but appears to be difficult for some short sequence stretches, namely, in regions displaying length variation. Earlier guidelines for imposing a unique alignment relied on binary alignment to a standard sequence (the rCRS) and used additional priority rules for resolving ambiguities. It turns out, however, that these rules have not been applied rigorously and led to inconsistent nomenclature. There is no way to adapt the priority rules in a reasonable way because binary alignment to a standard sequence is bound to produce artificial alignments that may place sequences separated by a single mutation at mismatch distance larger than 1. To remedy the situation, we propose a phylogenetic approach for multiple alignment and resulting notation.

Keywords Mitochondrial DNA · Haplogroup · Alignment · Phylogeny

H.-J. Bandelt
Department of Mathematics, University of Hamburg,
Bundesstraße 55,
20146 Hamburg, Germany
e-mail: bandelt@math.uni-hamburg.de

W. Parson (✉)
Institute of Legal Medicine, Innsbruck Medical University,
Muellerstrasse 44,
6020 Innsbruck, Austria
e-mail: walther.parson@i-med.ac.at

Introduction

The notation of mitochondrial DNA (mtDNA) variation is reported by reference to the revised Cambridge reference sequence (rCRS; [2]), which is the corrected version of the first fully sequenced mtDNA genome [1]. The forensic community adopted this convention accompanied by guidelines to standardize the mtDNA reporting process [4, 9, 23], in particular by providing general rules for sequence alignment. In a number of cases, however, specifically in the vicinity of the polycytosine stretches of the mtDNA control region, more than one alignment is often conceivable. Therefore, Wilson et al. [24, 25] attempted to standardize nomenclature in length variant regions by introducing formal rules for most parsimonious sequence alignment with respect to the rCRS, which were in principle then applied to the SWGDAM mtDNA database [17].

The increasing body of population data that became available in the past few years improved our understanding of the worldwide mitochondrial phylogeny and revealed that the application of the formal alignment rules could sometimes result in disputable assignments of haplotypes. In this study, we discuss the shortcomings of the Wilson et al. rules and propose a phylogenetic approach to the alignment of mtDNA sequences.

Alignment and nomenclature

In general, the alignment of human mtDNA sequences does not pose serious problems except for the vicinity of polycytosine or dinucleotide tracts that are both prone to considerable length heteroplasmy. Occasionally, however, one faces some real obstacles for unique alignment in some short sequence stretches due to multiple mutations.

To illustrate the difficulties with the alignment in the vicinity of polycytosine tracts, suppose we observe ATCCCCCCCCCCA in some samples where the rCRS shows ACCCCCCA (region 567–574). Most parsimoniously, this new variant would have to be scored relative to the rCRS as 567.1T 573.1C 573.2C 573.3C 573.4C, but alternatively, one could also consider scoring it less parsimoniously as 568T 573.1C 573.2C 573.3C 573.4C 573.5C. There is no reason a priori to reject the latter way of scoring for which actually two circumstances come in support: first, a transition at 568 without subsequent insertions of C has already been observed (e.g., in sample USA.AFR.000555 from the SWGDAM database), but no instance with a T insertion after position 567; second, and more importantly, insertions of C (ranging from 1 to 6) are abundant in this C tract.

Much more frequently, one is confronted with the obstacle that compound mutations generating a long C tract in HVS-I or HVS-II and transforming it further cannot be reconstructed in a unique way even if one strictly followed (unweighted) parsimony. To address this inherent ambiguity of the local alignment, Wilson et al. [24, 25] have attempted to establish formal rules for choosing one and only one alignment and thereby to standardize the notation for mtDNA sequences. To achieve this, they exclusively used binary alignment of every individual sequence to the reference sequence (rCRS) by employing parsimony and additional priority rules to break ties.

Formal alignment

The general recommendations by Wilson et al. [24, 25] are spelled out as the following three formal rules:

(Rule 1) Sequences should be aligned so that the least number of differences from the rCRS is present.

(Rule 2) If there is more than one way to maintain the same number of differences with respect to the rCRS, differences should be prioritized as follows: insertions/deletions (indels) → transitions → transversions.

(Rule 3) Indels should be placed 3′ with respect to the light strand and should be combined when the same number of differences to the rCRS is maintained.

To illustrate the effect of Rule 2 on the alignment of mtDNA sequences outside length variants, consider an HVS-II sequence that has inverted the nucleotide pair CT at positions 151–152 of the rCRS to TC. The normal alignment scores this as two transitions, 151T & 152C. However, the same number of two differences to rCRS is obtained when a deletion and an insertion are imposed instead, expressed as 151d (=151del) & 152.1C. Because by Rule 2 indels are preferred over transitions, the latter alignment would then be the recommended one. Such a recommendation in this case has obviously never been intended, let alone, followed—see USA.CAU.000088 as an example (Table 1).

A necessary assumption seems to be that the rules should only be applied under specific conditions, such as to length variants or polycytosine tracts, as may be illustrated with the following example. The inversion of CT at 16188–16189 to TC changes the lengths of the short C tracts surrounding position 16189 and would therefore be scored as 16188d & 16193.1C if one followed Rule 2. However, the standard notation adhered to in population genetics would be 16188T & 16189C in this case.

The phylogenetic information can provide us with a clue about which of the alternative alignments would be natural in such cases. For instance, sample CHN.ASN.000169 is a member of haplogroup M13a [14]. The ancestral motif of haplogroup M13 is assumed to bear 16188T but not 16189C, whereas 16189C is characteristic of the subhaplogroup M13a (according to Kong et al. [14]). Thus, the standard alignment appears to be supported phylogenetically (at least, given the present information). Another temporal order of mutations at 16188–16189 may be manifest in CHN.ASN.000303, which clearly belongs to haplogroup D5a2 (see Table 1; compare with related mtDNA sequences from Kong et al. [13] and Tanaka et al. [22]). The 16189C variant in this mtDNA is inherited from the ancestral haplotype of haplogroup D5, whereas 16188T is a private variant that came on top of the variants shared by the haplogroup D5a2 lineages. This suggests the usual scoring 16188T & 16189C instead of the interpretation 16188d & 16193.1C. In terms of the latter alignment, the comparison of CHN.ASN.000303 with other haplogroup D5a2 lineages having an uninterrupted long C tract (such as KOR.ASN.000131) would yield three differences (16188d & 16193.1C vs 16189C) although only a single change (16188T) would separate the two lineages within 16184–16193 (Table 2). Relying on the formal alignment rules, a forensic comparison would erroneously report three differences in such a case although only a single change in the C tract has taken place in otherwise identical sequences.

The effect of a “jumping” alignment as a consequence of Rules 1–3 can also be observed in mtDNA lineages that bear a C tract ranging from 16184 to 16191, which is then long enough to be subject to length variation. This can be found in the members of the Koryak/Inuit/Athabaskan branch of haplogroup A2 [20]; a typical example is the Navajo sample USA.154.000031 (Table 1; see also the work of Budowle et al. [8]). Another sample, USA.154.000067, nearly matches the former in HVS-I: the only difference is a C tract longer by one nucleotide. Yet, the alignment obeying the three rules jumps and thus creates a seeming difference of two

Table 1 Contrasting phylogenetic and formal binary alignments

Sample information		Alignments	
SWGDAM	Haplo-group	Phylogenetic	Formal rules 1–3
USA. CAU.000088	T2	16126C 16189C 16294T 16296T 73G **151T 152C** 263G 315.1C *(SWGDAM)*	16126C 16189C 16294T 16296T 73G **151d 152.1C** 263G 315.1C
USA. CAU.000642	T2	16126C **16188T 16189C** 16278T 16294T 16296T 73G 263G 309.1C 315.1C *(SWGDAM)*	16126C **16188d 16193.1C** 16278T 16294T 16296T 73G 263G 309.1C 315.1C
CHN. ASN.000303	D5a2	16092C 16164G 16172C 16182C 16183T **16188T 16189C** 16223T 16260T 16266T 16362C 73G 150T 263G 315.1C 489C 523d 524d *(SWGDAM)*	16092C 16164G 16172C 16182C 16183T **16188d 16193.1C** 16223T 16260T 16266T 16362C 73G 150T 263G 315.1C 489C 523d 524d
KOR. ASN.000131	D5a2	16092C 16164G 16172C 16182C 16183C **16189C** 16223T 16266T 16362C 73G 150T 263G 309.1C 315.1C *(SWGDAM)*	
CHN. ASN.000169	M13a	16145A 16148T **16188T 16189C** 16223T 16381C 73G 152C 263G 309.1C 315.1C 489C 527N	16145A 16148T **16188d 16193.1C** 16223T 16381C 73G 152C 263G 309.1C 315.1C 489C 527N *(SWGDAM)*
USA. CAU.001450	T1b	16126C 16163G **16184A 16189C** 16243C 16294T 16519C 73G 263G 315.1C *(SWGDAM)*	16126C 16163G **16183.1A 16189d** 16243C 16294T 16519C 73G 263G 315.1C
USA. CAU.000155	T1a	16126C 16163G 16186T **16189C 16193d** 16294T 73G 263G 309.1C 315.1C *(SWGDAM)*	16126C 16163G 16186T **16189d** 16294T 73G 263G 309.1C 315.1C
USA. HIS.000328	T1a	16126C 16163G 16186T **16189C 16193d** 16294T 16519C 73G 263G 309.1C 315.1C	16126C 16163G 16186T **16189d** 16294T 16519C 73G 263G 309.1C 315.1C *(SWGDAM)*
THA. ASN.000057	B4c2	16147T **16183C 16184A 16189C** 16217C 16235G 16294T 16519C 73G 263G 309.1C 309.2C 315.1C 523d 524d *(SWGDAM)*	16147T **16182.1C 16189d** 16217C 16235G 16294T 16519C 73G 263G 309.1C 309.2C 315.1C 523d 524d
SKE. AFR.000050	L3e2b	16172C **16184N 16189C** 16223T 73G 150T 195C 263G 309.1C 315.1C *(SWGDAM)*	
CHN. ASN.000072	D5c	**16189C 16190T 16193.1C 16193.2C** 16362C 16390A 16519C 73G 146C 150T **151T 152C** 182T 217C 263G 295N 309.1C 309.2C 315.1C 489C	**16188.1C 16193.1C** 16362C 16390A 16519C 73G 146C 150T **151d 152.1C** 182T 217C 263G 295N 309.1C 309.2C 315.1C 489C
KOR. ASN.000057	D5c	16093C **16189C 16190T 16193.1C 16193.2C** 16362C 73G 146C 150T **151T 152C** 182T 217C 263G 315.1C	16093C **16188.1C 16193.1C** 16362C 73G 146C 150T **151d 152.1C** 182T 217C 263G 315.1C
USA. HIS.000509	B2	16183C **16189C 16190T 16193.1C 16193.2C** 16217C 16234T 16235G 16261T 16519C 73G 263G 309.1C 309.2C 315.1C 499A	16183C **16188.1C 16193.1C** 16217C 16234T 16235G 16261T 16519C 73G 263G 309.1C 309.2C 315.1C 499A *(SWGDAM)*
USA. AFR.000213	L3e2b	16129A 16172C **16183C 16188T 16189C 16193.1C** 16223T 16320T 16519C 73G 150T 195C 263G 309.1C 315.1C	16129A 16172C **16183d 16193.1C 16193.2C** 16223T 16320T 16519C 73G 150T 195C 263G 309.1C 315.1C
USA. AFR.000537	L3e2b	16147T 16172C **16183C 16188T 16189C 16193.1C** 16223T 16320T 16519C 73 94N 150T 152C 195C 224N 263G 297G 309.1C 315.1C	16147T 16172C **16183d 16193.1C 16193.2C** 16223T 16320T 16519C 73G 94N 150T 152C 195C 224N 263G 297G 309.1C 315.1C *(SWGDAM)*
USA. AFR.000542	L3e2b	16172C **16189C** 16223T 16320T 16519C 73G 150T 195C 263G 315.1C *(SWGDAM)*	
USA.154.000067 (Navajo)	A2	16111T 16189C **16191.1C 16192T** 16223T 16290T 16319A 16362C 73G 146C 153G 235G 263G 309.1C 315.1C	16111T 16189C **16192.1T** 16223T 16290T 16319A 16362C 73G 146C 153G 235G 263G 309.1C 315.1C *(SWGDAM)*
USA.154.000031 (Navajo)	A2	16111T 16189C **16192T** 16223T 16290T 16319A 16362C 73G 146C 153G 235G 263G 315.1C *(SWGDAM)*	
USA.154.000003 (Navajo)	A2	16111T 16189C **16191.1C 16192T** 16223T 16233G 16290T 16319A 16331G 73G 146C 153G 235G 263G 315.1C	16111T 16189C **16192.1T** 16223T 16233G 16290T 16319A 16331G 73G 146C 153G 235G 263G 315.1C *(SWGDAM)*
USA.154.000108 (Navajo)	A2	16111T 16189C **16192T** 16223T 16233G 16290T 16319A 16331G 73G 146C 153G 235G 263G 315.1C *(SWGDAM)*	

Table 1 (continued)

Sample information		Alignments	
SWGDAM	Haplo-group	Phylogenetic	Formal rules 1–3
USA. AFR.000477	L2a1	16189C **16191.1C 16192T** 16223T 16278T 16294T 16309G 16390A 16519C 73G 143A 146C 152C 263G 315.1C 534T	16189C **16192.1T** 16223T 16278T 16294T 16309G 16390A 16519C 73G 143A 146C 152C 263G 315.1C 534T *(SWGDAM)*
USA. AFR.000905	L2a1	16189C **16192T** 16223T 16278T 16294T 16309G 16390A 16519C 73G 143A 146C 152C 195C 263G 315.1C 534T *(SWGDAM)*	
USA. AFR.000831	L4	16179T **16183C 16189C 16191T 16193.1C** 16223T 16239T 16311C 16320T 16362C 16519C 73G 150T 199C 204C 260C 263G 309.1C 315.1C **513A 523d 524d**	16179T **16183C 16189C 16190.1T** 16223T 16239T 16311C 16320T 16362C 16519C 73G 150T 199C 204C 260C 263G 309.1C 315.1C **513d 514d** *(SWGDAM)*
USA. AFR.000458	L4	16179T **16182C 16183C 16187T 16189C 16193d** 16223T 16239T 16311C 16320T 16362C 16519C 73G 150T 199C 204C 207A 263G 309.1C 315.1C **513A**	16179T **16182d 16183d 16193.1C** 16223T 16239T 16311C 16320T 16362C 16519C 73G 150T 199C 204C 207A 263G 309.1C 315.1C **513A** *(SWGDAM)*
USA. AFR.000084	L4	16179T **16183C 16189C 16193.1C** 16223T 16239T 16311C 16320T 16362C 73G 150T 199C 204C 263G 309.1C 315.1C *(SWGDAM)*	
USA. AFR.001199	L4	16179T **16183C 16189C** 16223T 16239T 16311C 16320T 16362C 16519C 73G 146C 150T 199C 204C 263G 309.1C 315.1C **513A 523d 524d** 527N	16179T **16183C 16189C** 16223T 16239T 16311C 16320T 16362C 16519C 73G 146C 150T 199C 204C 263G 309.1C 315.1C **513d 514d** 527N *(SWGDAM)*
USA. AFR.000832	L2c	16214T 16223T 16274A 16278T 16390A 73G 93G 146C 150T 152C 153G 182T 195C 198T 207A 263G 315.1C 325T **513A 523d 524d** *(SWGDAM)*	16214T 16223T 16274A 16278T 16390A 73G 93G 146C 150T 152C 153G 182T 195C 198T 207A 263G 315.1C 325T **513d 514d**
USA. AFR.001041	L2c	16214T 16223T 16264T 16278T 16390A 73G 89C 93G 146C 150T 152C 182T 195C 198T 263G 309.1C 315.1C 325T **523d 524d** *(SWGDAM)*	
USA. HIS.000204	A2	16111T 16182C 16183C 16189C 16223T 16290T 16319A 16362C 73G 146C 153G 235G 263G **310C 315d 316d 317d** *(SWGDAM)*	16111T 16182C 16183C 16189C 16223T 16290T 16319A 16362C 73G 146C 153G 235G 263G **310d 311d 312d**
USA.008.000169 (Apache)	B2	16129A 16147T 16183C 16193.1C 16217C 16325C 73G 146C 154C 263G **310C 314d 315d**	16129A 16147T 16183C 16193.1C 16217C 16325C 73G 146C 154C 263G **310d 311d** *(SWGDAM)*
USA. HIS.000134	B2	16152C 16182C 16183C 16189C 16217C 16325C 73G 263G **310C** *(SWGDAM)*	
USA. HIS.000332	T2	16126C 16294T 16296T 16324C 16519C 73G 92C 103N 125N 263G **310C 312d 313d 314d 315d**	16126C 16294T 16296T 16324C 16519C 73G 92C 103N 125N 263G **310d 311d 312d 313d** *(SWGDAM)*
USA. CAU000445	J2a	16069T 16126C 16145A 16231C 16261T 73G 150T 152C 195C 215G 263G 295T **310.1T 315.1C** 319C *(SWGDAM)*	16069T 16126C 16145A 16231C 16261T 73G 150T 152C 195C 215G 263G 295T **310.1T 310.2C**
USA. CAU.000600	J2a	16069T 16126C 16145A 16183C 16189C 16231C 16249C 16261T 73G 150T 152C 195C 215G 263G 295T **310C** 317N 319C 320N *(SWGDAM)*	
CHN. ASN.000451	B4a1a1	16182C 16183C 16189C 16217C 16247G 16261T 16519C 73G 146C 263G **308d 309d 315.1C** 523d 524d	16182C 16183C 16189C 16217C 16247G 16261T 16519C 73G 146C 263G **308T 310d** 523d 524d *(SWGDAM)*
CHN. ASN.000432	B4a1a1	16183C 16189C 16217C 16247G 16261T 16519C 73G 146C 263G **315.1C** 318C 523d 524d *(SWGDAM)*	
CHN. ASN.000439	B4a1a1	16182C 16183C 16189C 16217C 16247G 16261T 16519C 73G 146C 151T 263G **309.1C 315.1C** 523d 524d *(SWGDAM)*	
CHN. ASN.000469	B4a1a1	16182C 16183C 16189C 16217C 16247G 16261T 16519C 73G 146C 263G **309.1C 309.2C 315.1C** 523d 524d *(SWGDAM)*	
THA. ASN.000021	B4g	16181C 16182C 16183C 16189C 16213A 16217C 16242T 16261T 16292T 16301T 16519C 61A 62A 73G 150T 263G **308d 309d 315.1C** 523d 524d *(SWGDAM)*	16181C 16182C 16183C 16189C 16213A 16217C 16242T 16261T 16292T 16301T 16519C 61A 62A 73G 150T 263G **308T 310d** 523d 524d

Table 1 (continued)

Sample information		Alignments	
SWGDAM	Haplo-group	Phylogenetic	Formal rules 1–3
THA. ASN.000061	B4g	16181C 16182C 16183C 16189C 16213A 16217C 16261T 16292T 16519C 61A 62A 73G 263G **309.1C 315.1C** 514N 523d 524d *(SWGDAM)*	
CHN. ASN.000206	B4g	16181C 16182C 16183C 16189C 16213A 16217C 16242T 16261T 16292T 16301T 16519C 61A 62A 73G 183G 263G **309.1C 309.2C 309.3C 315.1C** 323N 324N 523d 524d *(SWGDAM)*	
CHN. ASN.000241	B4	16182C 16183C 16189C 16217C 16219G 16261T 16286T 16519C 73G 263G **309d 315.1C** 523d 524d *(SWGDAM)*	
CHN. ASN.000076	Z	16166C 16185T **16189C 16193d** 16223T 16260T 16298C 73G 152C **247A 249d** 263G 309.1C 315.1C 489C 505N 527N	16166C 16185T **16189d** 16223T 16260T 16298C 73G 152C **247d** 263G 309.1C 315.1C 489C 505N 527N *(SWGDAM)*
CHN. ASN.000061	Z	16136C 16185T 16223T 16260T 16298C 73G 152C **249d** 263G 309.1C 315.1C 489C *(SWGDAM)*	
USA. AFR.000624	L1c1	16086C 16129A 16169T 16172C 16189C 16192T 16223T 16278T 16293G 16311C 16360T 73G **151T 152C** 182T 186A 189C 195C 198T **247A 249d** 263G 297G 315.1C 316A	16086C 16129A 16169T 16172C 16189C 16192T 16223T 16278T 16293G 16311C 16360T 73G **151d 152.1C** 182T 186A 189C 195C 198T **247d** 263G 297G 315.1C 316A
SKE. AFR.000005	L1c1	16086C 16129A 16187T 16189C 16223T 16278T 16293G 16294T 16360T 73G **151T 152C** 182T 186A 189C 195C 204C **247A** 263G 297G 315.1C 316A *(SWGDAM)*	1608C 16129A 16187T 16189C 16223T 16278T 16293G 16294T 16360T 73G **151d 152.1C** 182T 186A 189C 195C 204C **247A** 263G 297G 315.1C 316A

mismatches (Table 2). Analogous results are obtained when samples USA.154.000108 and USA.154.000003 are compared (Table 1).

A similar situation occurs with an African profile (USA. AFR.000477) from a specific subhaplogroup of haplogroup L2a1 with the characteristic variant 534T, which displays the same sequence variation within 16183–16194 as those latter two Navajo sequences. The SWGDAM profiles obeying the three rules are, however, artificially aligned to their corresponding mtDNA relatives in these cases because the variant 16192T appears to be ancestral to those particular subhaplogroups of A2 and L2a1, respectively (Table 1).

To give another example, three changes each are needed to derive the variation within 16183–16194 observed in the two haplogroup D5c profiles CHN.ASN.000072 and KOR. ASN.000057 from the ancestral motif with 16189C [22]. This can be realized in two ways: namely, either by reversing the 16189C change and inserting two C nucleotides, 16188.1C & 16193.1C, or introducing the changes 16189C & 16190T & 16193.1C & 16193.2C. The formal rules require the former scoring, whereas we find the latter scenario more plausible because it would invoke only two C insertions extending the long C tract plus a subsequent transition interrupting the long C tract. An analogous situation occurs in haplogroup B2 with USA.HIS.000509 (Table 1).

A similar case can be made for haplogroup L3e2b, where the profile USA.AFR.000542 represents the ancestral sequence (bearing 16189C). Then three additional changes within 16183–16194 explain the variation (16183C & 16188T & 16189C & 16193.1C) of USA.AFR.000213 and USA.AFR.000537 in 16183–16194 (Table 2). This derivation, invoking an elongation of the long C tract with subsequent interruption through a C-to-T mutation, corresponds with the phylogenetic approach in contrast to the more parsimonious one positing 16183C & 16187.1T & 16189C instead. However, this latter scoring, as represented by the SWGDAM entry USA.AFR.000213, violates Rule 2 because 16183d & 16193.1C & 16193.2C (as found in USA.AFR.000537) has higher priority (Table 1).

The mtDNA lineages USA.CAU.001450 and THA. ASN.000057 are particular members of large haplogroups (T1 and B, respectively) for which the long C tract (incurred by 16189C) is characteristic but not the variant 16184A, so that the modification of the short A tract constitutes a private variant or defines a minor subhaplogroup (such as B4c2). This also applies to the two haplogroup T1a lineages USA.CAU.000155 and USA. HIS.000328, where the private variant constitutes a mere length variation of the long C tract from 16187 to 16193, which we would denote by 16193d in this case (Table 1).

The problem with 16184A is exacerbated when this variant is heteroplasmic and competes with the reference nucleotide, that is, when 16184M (using the IUPAC code) is observed. The SWGDAM database does not document ambiguities (incurred by real heteroplasmy or reading difficulties in one of the paired electropherograms) in this detailed way but always reports "N" in such ambiguous cases. Then suppose that the undetermined position in the sample SKE.AFR.000050

Table 2 Jumping alignment in HVS-I

Position	16182	16183	16184	16185	16186	16187		16188	16189	16190	16191		16192		16193			16194
rCRS	A	A	C	C	C	C	–	C	T	C	C	–	C	–	C	–	–	A
D5a2 consensus	C	C							C									
D5a2 variant	C	T						T	C									
D5a2 variant (Rules 1–3)	C	T						–								C		
A2-subclade consensus									C				T					
A2-subclade variant									C			C	T					
A2-subclade variant (Rules 1–3)									C					T				
L3e2b consensus									C									
L3e2b variant		C						T	C							C		
L3e2b variant (Rules 1–3)		–														C	C	
L3e2b variant *(SWGDAM)*		C					T		C									

having a long C tract (Table 1) was in fact 16184M and reflected true heteroplasmy. According to the formal rules, one variant bearing 16184A, when homoplasmic, would have to be scored as 16183.1A & 16189d (within 16183–16193), whereas the other homoplasmic variant showing 16184C (as the rCRS) would be recorded as 16189C in this region. Thus, ignoring the undetermined position for forensic purposes would still yield a recognizable match only with the latter homoplasmic variant but not with the former one to which it would appear to be at distance 2—at least when the formal rules were strictly followed.

When a long C tract occurs in HVS-II (indicated by 310C or 310d), the formal alignment according to Rules 1–3 may also jump depending on the number of inserted C molecules; see USA.008.000169, USA.HIS.000134, USA.HIS.000332, and USA.CAU.000600 as examples following the above-mentioned principle (Table 1 and Table 3). In this context, Rule 3 effectively demands that contiguity of an indel has higher priority than unrestricted 3′ placement; see example 17 of the work of Wilson et al. [24]. There is a problem though when it comes to insertions, namely, an insertion of TC after position 310 should then be encoded as 310.1T & 310.2C, but in practice, this change is recorded as 310.1T & 315.1C (e.g., USA.CAU000445; Table 1).

Even when the nucleotides in the immediate vicinity of position 309 are apparently unaltered relative to the rCRS, the alignment may jump depending on the lengths of the surrounding C tracts. This can be seen with mtDNA sequences belonging to a particular subhaplogroup of B4a: the typical length (7) of the C tract preceding 310 is represented in haplogroup B4a1a1 [14] by sample CHN.ASN.000432 in the SWGDAM database (Table 1). A loss of two cytosines there, as seen in CHN.ASN.000451, causes a jump of the formal alignment to three changes when adhering to Rules 1–3. Relative to a putative intermediate haplotype with six cytosines in the C tract (as, e.g., seen in sample CHN.ASN.000241 from a different B4 subhaplogroup), the jump would go from one mismatch (between C tract lengths 6 and 5) to four mismatches, viz., at positions 308, 309, 310, and 315.1. This would also jeopardize comparisons between heteroplasmic mtDNAs with respect to these two length variants and the corresponding two homoplasmic mtDNAs, so that this would directly affect forensic comparisons (Table 3).

The preceding alignment switch that takes place when comparing C-tract lengths of 6/5 and 5/5 cytosines (pre/post 310) is directly incurred by the unfortunate choice of a distant reference sequence for the binary alignment. A consensus sequence for the region 303–315 would rather have the C-tract lengths 7/6 instead of 7/5 as in the rCRS. Suppose one would choose the consensus sequence with motif 73G 263G 315.1C (relative to the rCRS) as the standard sequence to which all other sequences get aligned formally. Then all sequences with C-tract lengths ranging from 5/6 to 10/6 would get properly aligned. Thus, the idiosyncratic features of the rCRS in HVS-II exacerbate the problems with the formal binary alignment.

Another notational complication occurs in the region 515–524 with length variation of the AC repeat. A number of mtDNA lineages bear a G-to-A change at position 513, which effectively prolongs the AC repeat because position 514 bears C. This single nucleotide change would then be expressed as 513A (as in USA.AFR.000458). If, however, such a lineage also had a shortened AC repeat (which on its own is scored as 523d & 524d), then this compound change would be scored as 513d & 514d following the formal rules. This can well be demonstrated with mtDNA samples from a side branch of haplogroup L4 [12]; see Table 1. Also, there is an analogous instance in haplogroup L2c where a transition to 513A is accompanied by deletion events in the dimeric repeat, e.g., in USA.AFR.000832. When comparing this sequence to a somewhat related sequence (USA.AFR.001041) that lacks the 513 change, one could well envision that the 513A occurred after the deletion of an AC repeat in this particular branch of haplogroup L2c.

The preceding discussion of alignment problems could evoke the wrong impression that the formal rules—when

Table 3 Alignment of C-tract variants in HVS-II

Position	302	303	304	305	306	307	308	309				310	311	312	313	314	315		316	317
Type[a] 7/5 rCRS	A	C	C	C	C	C	C	C	–	–	–	T	C	C	C	C	C	–	G	C
Type 10/6									C	C	C							C		
Type 9/6									C	C								C		
Type 8/6									C									C		
Type 7/6																		C		
Type 6/6								–										C		
Type 5/6							–	–										C		
Type 5/6 (Rules 1–3)							T										–			
Type 13												C								
Type 12												C					–			
Type 12 (Rules 1–3)												–								
Type 11												C				–	–			
Type 11 (Rules 1–3)												–	–							
Type 10												C			–	–	–			
Type 10 (Rules 1–3)												–	–	–						
Type 9												C		–	–	–	–			
Type 9 (Rules 1–3)												–	–	–	–					
Type 12^{del}												C					–		–	–
Type 12^{del} (Rules 1–3)												–							–	–

[a]Type *i*/*j* signifies two short C tracts of lengths *i* and *j*; type *k* (k^{del}) indicates a single long C tract of length *k* (with subsequent deletions, respectively).

properly adjusted—would only fail in the vicinity of C tracts and AC repeats. This is not the case. Problems may arise at any position of the mtDNA sequence where some indel events occurred in the mtDNA phylogeny. Typical examples in this regard are occasional changes in the region 247–249, where the rCRS has GAA. Haplogroups CZ and F bear an independent characteristic deletion of one adenine there, which is recorded 3′-most as 249d. Now, a subsequent transition at 247 leads to a sequence with motif 247A & 249d, which, when directly aligned to the rCRS, would incorrectly be interpreted as 247d just as in CHN. ASN.000076 (Table 1). Thus, again, application of Rules 1–3 leads to a jumping alignment, turning a single change into seeming two changes. In contrast, the scoring 247d in sample USA.AFR.000624 corresponds to a different evolutionary pathway: the ancestral haplogroup L1c1 sequence has the motif 247A in this stretch, so that the subsequent loss of one adenine needs to be scored as 247A & 249d (Table 1). There is only a single candidate (SKE. AFR.000054) in the SWGDAM database where the scoring 247d might be justified (given the present knowledge).

In conclusion, the three formal rules have the disadvantage that (1) they cannot be applied meaningfully as stated, (2) the potential range of their application can hardly be determined beforehand, and that (3) even under careful handling, a mere length variation of the long C tracts or AC repeats can yield artificial alignment switches and thereby increase the distance from one mismatch up to four apparent mismatches. The SWGDAM database very well testifies to the inherent ambiguity in the application of Rules 1–3, as each of these rules is found to be violated there in several instances (Tables 1, 2 and 3).

To define "length variants", some initial alignments of at least two sequences must be given beforehand. However, if alignment rules were supposed to be conditional to the presence of length variation, then the requirements for alignment would become circular. This is well reflected by the examples put forward by Wilson et al. [25]: it is stated that "no insertions or deletions are present; thus, there are no alternative alignments..." for example 1—which expresses the fact that those sequence stretches under comparison were of equal length; on the other hand, the partial sequences of their example 18 also have equal length, but an alignment combining one deletion with one insertion and thus postulating length variation is derived from the rules. It seems to be impossible to establish meaningful rules for unambiguous alignment in the absence of an evolutionary perspective: "Good alignments of related sequences are ones that better reflect the evolutionary relationship between them." ([10], p. 184).

Phylogenetic alignment

Whether a long C tract around positions 16189 and 310 is always created by a T-to-C transition or by a deletion in vivo is not known a priori and is difficult to ascertain. Uninterrupted polycytosine tracts usually display considerable length heteroplasmy, where a population of two to six different length variants [5, 15, 16, 18, 19] can be discerned in the tissues and body fluids of an individual provided that low-level heteroplasmic variants can be distinguished from background noise with appropriate sequencing technology [6, 7]. The demonstration of length

heteroplasmy in a sequence electropherogram depends on the applied sequencing chemistry and primers and thus may vary between laboratories. Differences can already be observed between forward and reverse sequencing reactions within the same sample. Therefore, it is difficult, if not impossible, to suggest a general notation of heteroplasmic length variants. For mtDNA databases, we envision that the standard encoding would seek to identify and report a dominant variant, which can be confirmed in the majority of cases.

For the evolution of the C tracts, we can nonetheless envision a default scenario, which would capture most events reflected in the phylogeny. For instance, the short C tracts around T at position 16189 seem relatively stable. Then only their joining incurred by 16189C would destabilize this region. The first kind of mutations that would subsequently respond to further lengthening of the long C tract is the loss of one or two nucleotides of the preceding short A tract. The resulting compound changes have traditionally been scored as A to C transversions at 16181–16183, with the understanding that these changes play a role quite different from any other transversion in the control region. A long C tract would then typically have a length between 9 to 13 nucleotides. Although such long homopolymeric regions are tolerated over long evolutionary periods, as testified by many haplogroups with an ancestral 16189C motif, there seems to be a tendency to interrupt such a long C tract by some subsequent C to T changes somewhere close to its middle (e.g., compare the haplogroup L4 sequences listed in Table 1). We suggest that the notational system follows this default scenario by taking 16189C as an indicator for the union of the two short C tracts, then using 16183C, 16182C, and 16181C for the changes of the length of the preceding A tract, and finally a transition that again splits the long C tract into two shorter ones.

There are slight differences in the behavior between the C tracts around positions 310 and 16189. The typical C tract post 310 is one nucleotide longer (scored as 315.1C) than that of the rCRS and relatively stable, with only occasional losses of 315.1C or gains of 315.2C, whereas the standard seven-nucleotide C tract from 303 to 309 is quite variable in length and may have sizes ranging from five to ten; for example, all these different lengths occur in haplogroup B4 (Table 1). Length variants in HVS-II develop either with insertions of C residues between positions 302 and 310 or with a transition or deletion event at position 310. There is a tendency that the latter leads to a reduction of C molecules in the C tract, whereas in HVS-I, the transition at 16189C is usually accompanied by the gain of C molecules with respect to the rCRS. In analogy to the above C-tract convention for HVS-I, we suggest to take 310C as an indicator for the union of the two C tracts and designate the reduction of the C-tract length with deletions 3′ to the homopolymeric region (Table 3).

We recommend staying as close as possible to a reconstructed phylogeny using all available data. Nevertheless, we strive for some canonical notations that are in agreement with the broad forensic tradition. Namely, we suggest replacing Rules 1–3 by the following guidelines:

- (Phylogenetic law) *Sequences should be aligned with regard to the current knowledge of the phylogeny. In the case of multiple equally plausible solutions, one should strive for maximum (weighted) parsimony. Variants flanking long C tracts, however, are subject to extra conventions in view of extensive length heteroplasmy.*
- (C tract conventions) *The long C tracts of HVS-I and HVS-II should always be scored with 16189C and 310C, respectively, so that phylogenetically subsequent interruptions by novel C to T changes are encoded by the corresponding transition. Length variation of the short A tract preceding 16184 should be notated in terms of transversions.*
- (Indel scoring) *Indels should be placed 3′ with respect to the light strand unless the phylogeny suggests otherwise.*

Phylogenetic alignment is a necessary prerequisite for comparing two or more mtDNA profiles. To determine the likelihood that two profiles differing by exactly one mutation, either in heteroplasmic or in homoplasmic state, could coexist in one individual is, however, not straightforward. The site-specific mutational spectrum gleaned from an estimated mtDNA phylogeny [3] cannot necessarily be equated with the corresponding spectrum for somatic mutations or de novo mutations in the germline because some mutations (such as the transition at position 215, for example; [11]) may have a reduced chance to transcend from heteroplasmic to homoplasmic state and get fixed in the matriline. On the other hand, there is no positive evidence to date that the somatic mutation process in the control region would strongly deviate from the corresponding long-term process that has been inferred from the mtDNA phylogeny, let alone that the somatic mutation spectrum within an individual would conform to Rules 1–3 of the formal binary alignment to the rCRS.

Discussion

When Wilson et al. [24] highlighted the problems about notational ambiguity and published their recommendations, they anticipated that not all investigators would agree with the proposed rules: “These rules as described herein may be accepted, or other proposed approaches may be considered.

At least the issues are raised, and discussion can begin." [25]. At first sight, the rules appeared to be convenient, simple, and plausible—so, the necessary discussion in the forensic community did not really gain momentum. However, in practical casework, the rules were handled more flexibly, and sometimes, the regions with puzzling length variants were ignored altogether when the application of the three formal rules would have led to jumping alignment.

We have shown that the recommendations based on binary alignment to a standard sequence (rCRS) and formal priority rules have the following serious drawbacks that impact on their forensic use:

(a) The formal rules need at least to be restricted to regions with notorious length variation as they cannot be applied meaningfully to other parts of the sequences.
(b) The resulting alignment is biased toward the reference sequence rCRS and does not take the phylogenetic background of the sample into consideration.
(c) The rules can lead to "jumping" alignments that place sequences separated by a single mutation at larger distance, which is in conflict with the evolutionary history of the sequences.
(d) In forensic casework, jumping alignment introduces more differences than necessary to explain the mutations between the mtDNA sequences of stain and culprit and thus leads to unjustified exclusion of the culprit as the donor of the stain.
(e) The formal rules can distort the estimation of site-specific mutational rates that are necessary to interpret differences between sequences.

Sequence alignment and nomenclature is a difficult task in and around homopolymeric regions of the mtDNA control region, simply as it is sometimes impossible to identify the exact mutational position and event a priori. Moreover, unforeseen situations with complex rearrangements and length variation may arise outside the C tracts that resist unambiguous alignment. Consider, for example, the short stretch GTATTTTC from positions 54 to 61 in the rCRS. In the Indian haplogroup M39 [21], one observes the two variants GTTTTTTC and GTTATTC. Then one alignment poses the change 56T for the first variant and 56del & 58A for the second one (see IND.CAU.000109 and IND.CAU.000117 from the SWGDAM database). Alternatively, one could hypothesize a common insertion 55.1T for both variants with different subsequent losses of thymines, namely, 56del and 59del & 60del, respectively [21]. The latter alignment would be less parsimonious than the first one if the loss of TT in the second case was scored as two events but would be equally parsimonious when the double loss was counted as a single change.

Thus, the alignment problem cannot be dealt with in a straightforward formal way. Phylogenetic estimation is important for a meaningful evolutionary interpretation of mtDNA haplotypes, which is also fundamental to forensic considerations. Alignment cannot be split from phylogeny conceptually, and binary alignment to a standard sequence would fall short of the goal anyway. Therefore, multiple sequence alignment with a phylogenetic perspective has to be performed. This, seemingly, has the disadvantage that the notational system capturing the alignment would not always be stable over time when finer details of the phylogeny emerge. However, nothing could really stay unaltered because knowledge about the genetic marker, human mtDNA, keeps growing.

Any forensic comparison of an mtDNA sequence to the mtDNA sequences stored in a database needs to take the most plausible reconstruction of potential evolutionary pathways into consideration. In general, this pathway will not directly link the mtDNA sequence with the rCRS but rather with the most similar mtDNA sequences from the smallest recognizable subhaplogroup the targeted mtDNA belongs to. Therefore, the notational devices for recording mtDNA sequences relative to the rCRS, which would reflect an alignment with all other mtDNA sequences, have to depart from realistic estimates of the whole mtDNA phylogeny. One can therefore never totally exclude ambiguity because one may never know the exact relationship in some intricate cases. With the best of current knowledge, one would then have to consider all plausible alternative pathways expressed by different alignments. To facilitate such comparisons, some standardized nomenclatures are helpful, which harmonize the notation of mtDNA profiles across databases, but this should not seduce the user of a database to rely on a single "optimal" alignment in all cases. Actually, Wilson et al. [25] already pointed to this kind of caveat by suggesting that one may wish to "...query a long string of bases rather than a set of differences from a reference. Such an alternative to the current method might be explored in an effort to avoid inconsistencies caused by optional alignments when applied to forensic applications."

Conclusions

We conclude that sequence alignment and the corresponding notational system should not employ binary alignment to a standard sequence. This simplistic alignment procedure would not only distort phylogenetic considerations between closely related sequences but also directly impact forensic comparison: frequently observed insertions of C molecules in length variants that typically occur in tissues with higher mutability (e.g., hair and muscle) may lead to jumping alignment with respect to the sequence

determined from the reference material of the same person, such as blood. At this point, one either needs to reject the rules or ignore the polymorphisms in the entire region affected by length variation to avoid a false exclusion scenario.

Nowadays, we can take advantage of the emerging mtDNA sequence information that allows us to compare sequences to their phylogenetic neighbors. This is both relevant to the quality management of published and newly generated sequence data and also for establishing a reliable notational system for length variants. The guidelines as expressed in this study are applicable to all mtDNA sequences in their entire range. They put sequences in a phylogenetically meaningful relationship and aid forensic comparison. In some cases, these guidelines may not suffice to unambiguously determine a single most-plausible alignment, whereas in other instances, the alignment may even change when additional information on the corresponding portion of the mtDNA phylogeny becomes available. Such limitations, however, should not invalidate a phylogenetic approach to sequence alignment as this constitutes the only scientific way to deal with comparison of mtDNA sequences in the light of the current knowledge about human mtDNA variation. In consequence, the new forensic mtDNA database, EMPOP, will strive for a strictly phylogenetic alignment.

Acknowledgment We are indebted to Anita Brandstätter, Nina Duftner, Cordula Eichmann, Anna König, Daniela Niederwieser, and Bettina Zimmermann (Institute of Legal Medicine, Innsbruck Medical University) for their excellent technical work and sequence analysis. Thomas J. Parsons (International Commission on Missing Persons, Sarajevo, Bosnia-Herzegovina) and Jodi A. Irwin (Armed Forces DNA Identification Laboratory) are greatly acknowledged for their helpful discussions. We thank Claudio Bravi for drawing our attention to the alignment problems in the Navajo mtDNA dataset. Finally, we appreciated the constructive criticism of one reviewer that helped improve the text.

References

1. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465
2. 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 23:147
3. Bandelt H-J, Richards M, Macaulay V (eds) (2006) Human mitochondrial DNA and the evolution of *Homo sapiens*. Series: Nucleic Acids and Molecular Biology, vol. 18. Springer, Berlin Heidelberg New York
4. Bär W, Brinkmann B, Budowle B, Carracedo A, Gill P, Holland M, Lincoln PJ, Mayr W, Morling N, Olaisen B, Schneider PM, Tully G, Wilson M (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Int J Legal Med 113:193–196
5. Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am J Hum Genet 57:248–256
6. Brandstätter A, Parson W (2003) Mitochondrial DNA heteroplasmy or artefacts—a matter of the amplification strategy? Int J Legal Med 117:180–184
7. Brandstätter A, Niederstätter H, Parson W (2004) Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed basecalls in the entire human mitochondrial DNA control region. Int J Legal Med 118:47–54
8. Budowle B, Allard MW, Fisher CL, Isenberg AR, Monson KL, Stewart JE, Wilson MR, Miller KW (2002) HVI and HVII mitochondrial DNA data in Apaches and Navajos. Int J Legal Med 116:212–215
9. Carracedo A, Bär W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G, Wilson M (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int 110:79–85
10. Ewans WJ, Grant GR (2001) Statistical methods in bioinformatics. Springer, Berlin Heidelberg New York
11. Forster L, Forster P, Lutz-Bonengel S, Willkomm H, Brinkmann B (2002) Natural radioactivity and human mitochondrial DNA mutations. Proc Natl Acad USA 99:13950–13954
12. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75:752–770
13. Kong Q-P, Yao Y-G, Sun C, Bandelt H-J, Zhu C-L, Zhang Y-P (2003) Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet 73:671–676; erratum (2004) 75:157
14. Kong Q-P, Bandelt H-J, Sun C, Yao Y-G, Salas A, Achilli A, Wang C-Y, Zhong L, Zhu C-L, Wu S-F, Torroni A, Zhang Y-P (2006) Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15:2076–2086
15. Lutz S, Weisser HJ, Heizmann J, Pollak S (1997) A third hypervariable region in the human mitochondrial D-loop. Hum Genet 101:384
16. Lutz S, Weisser H-J, Heizmann J, Pollak S (1998) Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany. Int J Legal Med 111:67–77; errata (1998) 111:286 and (1999) 112:145–150
17. Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B (2002) The mtDNA population database: an integrated software and database resource for forensic comparison. Forensic Sci Commun 4:#2
18. Parson W, Parsons TJ, Scheithauer R, Holland MM (1998) Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: application of mtDNA sequence analysis to a forensic case. Int J Legal Med 111:124–132
19. Pfeiffer H, Lutz-Bonengel S, Pollak S, Fimmers R, Baur MP, Brinkmann B (2004) Mitochondrial DNA control region diversity in hairs and body fluids of monozygotic triplets. Int J Legal Med 118:71–74
20. Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718–726
21. Sun C, Kong Q-P, Palanichamy Mg, Agrawal S, Bandelt H-J, Yao Y-G, Khan F, Zhu C-L, Chaudhuri TK, Zhang Y-P (2006) The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. Mol Biol Evol 23:683–690

22. Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, Fuku N, Guo LJ, Hirose R, Fujita Y, Kurata M, Shinoda K, Umetsu K, Yamada Y, Oshida Y, Sato Y, Hattori N, Mizuno Y, Arai Y, Hirose N, Ohta S, Ogawa O, Tanaka Y, Kawamori R, Shamoto-Nagai M, Maruyama W, Shimokata H, Suzuki R, Shimodaira H (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14:1832–1850
23. Tully G, Bär W, Brinkmann B, Carracedo A, Gill P, Morling N, Parson W, Schneider P (2001) Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. Forensic Sci Int 124:83–91
24. Wilson MR, Allard MW, Monson KL, Miller KWP, Budowle B (2002a) Recommendations for consistent treatment of length variants in the human mtDNA control region. Forensic Sci Int 129:35–42
25. Wilson MR, Allard MW, Monson KL, Miller KWP, Budowle B (2002b) Further discussion of the consistent treatment of length variants in the human mitochondrial DNA control region. Forensic Sci Commun 4:#4