Int J Legal Med (2012) 126:901–916
DOI 10.1007/s00414-012-0762-y

ORIGINAL ARTICLE

Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics

Hans-Jürgen Bandelt · Mannis van Oven · Antonio Salas

Received: 24 March 2012 / Accepted: 6 August 2012 / Published online: 1 September 2012

Abstract Haplogrouping refers to the classification of (partial) mitochondrial DNA (mtDNA) sequences into haplogroups using the current knowledge of the worldwide mtDNA phylogeny. Haplogroup assignment of mtDNA control-region sequences assists in the focused comparison with closely related complete mtDNA sequences and thus serves two main goals in forensic genetics: first is the a posteriori quality analysis of sequencing results and second is the prediction of relevant coding-region sites for confirmation or further refinement of haplogroup status. The latter may be important in forensic casework where discrimination power needs to be as high as possible. However, most articles published in forensic genetics perform haplogrouping only in a rudimentary or incorrect way. The present study features PhyloTree as the key tool for assigning control-region sequences to haplogroups and elaborates on additional Web-based searches for finding near-matches with complete mtDNA genomes in the databases. In contrast, none of the automated haplogrouping tools available can yet compete with manual haplogrouping using PhyloTree plus additional Web-based searches, especially when confronted with artificial recombinants still present in forensic mtDNA datasets. We review and classify the various attempts at haplogrouping by using a multiplex approach or relying on automated haplogrouping. Furthermore, we re-examine a few articles in forensic journals providing mtDNA population data where appropriate haplogrouping following PhyloTree immediately highlights several kinds of sequence errors.

Keywords Mitochondrial DNA · Forensics · mtDNA database · Haplogroup · PhyloTree · Multiplex genotyping · mtDNAmanager · MitoTool · HaploGrep · Sequence error · Quality control

Introduction

Mitochondrial DNA (mtDNA) analysis has become a routine target in forensic casework where standard nuclear markers cannot be applied. The EMPOP project (http://www.empop.org) [1] provides the leading mtDNA database of control-region sequences for comparison with partial mtDNA sequences from forensic traces. Release 6 now offers 10,841 entire control-region sequences plus some additional partial sequences. Comparison of a trace sequence with the sequences from that database can provide rough estimates of match probabilities in certain human population groups.

From a given entire control-region sequence or its hypervariable segments I (HVS-I) and/or II (HVS-II) and/or III (HVS-III), one can often determine the approximate location of the mtDNA lineage in the current tree estimate of the global mtDNA phylogeny. This process of allocating partial (or entire) mtDNA genomes to haplogroups is called

Electronic supplementary material The online version of this article (doi:10.1007/s00414-012-0762-y) contains supplementary material, which is available to authorized users.

H.-J. Bandelt (✉)
Department of Mathematics, University of Hamburg,
20146 Hamburg, Germany
e-mail: bandelt@yahoo.com

M. van Oven
Department of Forensic Molecular Biology, Erasmus MC,
University Medical Center Rotterdam,
3000 CA Rotterdam, The Netherlands

A. Salas
Unidade de Xenética, Instituto de Ciencias Forenses, Facultade de Medicina, and Departamento de Anatomía Patolóxica e Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela,
15782 Galicia, Spain

haplogrouping. The term refers to the fact that the basal part of the known mtDNA phylogeny is encoded as a nested system of haplogroups (alias monophyletic clades). A readily accessible global mtDNA tree based on complete mtDNA genomes is provided by the PhyloTree project (http://www.phylotree.org) [2], which was initiated in August 2008 and is updated at regular intervals. In fortunate cases, very few mutational variants alone can unambiguously determine a specific clade of the mtDNA phylogeny; e.g., the combined presence of variants 131C and 183G pinpoint a novel subhaplogroup of B2 (to be named B2j in Build 14) according to [3]. Despite the ease with which mutations and motifs can be searched in PhyloTree, the use of this Web information is apparently not yet routinely used by most forensic laboratories.

In the worst case, no haplogrouping at all is performed. More often, some sort of haplogrouping is aimed at by using fixed partial motifs of mutational variants relative to the revised Cambridge Reference Sequence (rCRS; [4]) that need to be matched entirely and therefore do not allow for occasional back mutations. In other cases, obsolete haplogroup schemes are still in use, which would no longer be supported by up-to-date knowledge of the mtDNA phylogeny. This is also the case for the on-going anthropological Genographic project, which employed a coarse classification tree, which was outdated even at the time (2005) when the project was launched [5]; see Supplementary File 1.

Haplogrouping is useful because it can help to detect sequencing and documentation errors. For instance, optimal haplogrouping of a given mtDNA sequence allows to partition the nucleotide variants into shared and private ones and could therefore indicate the oversight of variants or the presence of phantom mutations and artificial recombination [6–9]. When executing a database search, an "innocent" documentation error that, e.g., transforms a transition into a transversion could convert a very common haplotype into a very rare one. Simple errors could also lead to mismatches when comparing two profiles coming from a forensic stain and a suspect. On the other hand, the potential haplogroup status of a control-region sequence suggests which additional coding-region sites could be screened in order to increase the discrimination power of the mtDNA test [9].

In the present study, we outline strategies for performing haplogroup allocation of control-region sequences via pertinent Google (as performed in [10]) and PhyloTree searches. We then go through the mtDNA population data published in forensic journals (*Legal Medicine*, *Journal of Forensic and Legal Medicine*, *Journal of Forensic Sciences*, *Forensic Science International: Genetics*, and *International Journal of Legal Medicine*) within the period 2007–2011. We show that attempts at haplogrouping, either manual or automated, may fall short of the goal for several reasons. The shortcomings of the various attempts at haplogrouping carried out in those articles are highlighted in brief; see Supplementary File 1. We also demonstrate by way of a case selected from the field of medical genetics that employing HVS-I sequences plus a major portion of the coding-region sequence will normally predict a fine-grained haplogroup status. Similarly, a tailored multiplex SNP analysis of coding-region sites could in principle improve haplogrouping inferred from HVS-I&II variation—provided that the choice of SNPs has been made prudently, that is, by selecting sites as conservative as possible for elucidating pivotal parts of the mtDNA phylogeny that are poorly determined by control-region variants.

Methods

Reference mtDNA tree

The reference tree for the mtDNA phylogeny employed in our study for testing automated classification tools is mtDNA tree Build 13 of 28 December 2011 (http://www.phylotree.org/builds/mtDNA_tree_Build_13.zip). This basal classification tree of haplogroup motifs is now based on 10,627 complete genome sequences and thus appears to be quite robust and detailed. Each haplogroup is characterized by a set of nucleotide variants that are shared by all its members due to common ancestry—unless some sporadic back mutation has partially erased this signature. These shared variants are commonly referred to as the "sequence motif" or "mutation(al) motif". If one wishes to know the motif (expressed as differences from the rCRS) characterizing a particular haplogroup, one has to record the mutations appearing along the branches of the tree all the way from the rCRS to the node of the targeted haplogroup. Supplementary File 2 provides the control-region motifs for all haplogroups included in PhyloTree Build 13.

Mutation weighting

The sequence motif of some mtDNA profile could be partially incomplete (i.e., lacking variants that would be expected from the phylogeny) due to oversight or natural back mutation. Knowledge of positional mutation rates is useful to evaluate how likely a back mutation would occur at a given position. Therefore, in order to explore how frequently a site is hit by a particular mutation, we refer to the analysis carried out by Soares et al. [11] (based on 2,196 selected complete mtDNA sequences). In what follows we will briefly refer to the Soares et al. score as the number of occurrences (back or forth) inferred from the test dataset of [11]. The top ten coding-region mutations in [11] are the transitions at sites 709, 11914, 5460, 15924, 1719, 13708, 13105, 5147, 1598, and 8251, with scores ranging from 59 down to 22.

This may also be compared to the number of occurrences in Build 13 of PhyloTree, where the top ten coding-region mutations are at sites 709, 13708, 11914, 5460, 15924, 1719, 5147, 1598, 10398, and 13105, with scores from 36 to 15, quite well agreeing with the previous list (site 8251 is number 12 here with score 13). The mutations recorded along this tree are shared mutations on basal parts of the estimated mtDNA phylogeny, except for a few singular lineages with their full spectrum of mutations. Mutations beyond the current haplogroup level ("private mutations") are thus not shown. Therefore, the mutations in PhyloTree constitute only a minority of all mutations that could be retrieved from an exhaustive global tree of all published mtDNA sequences that seem to be of reasonable quality.

Mutation weighting is even more important for the control-region which includes the most extreme hotspots. The top nine control-region mutations (excluding the hotspot transition at 16519 and some length polymorphisms) in [11] are the transitions at sites 152, 16311, 146, 195, 16189, 16129, 16093, 16362, and 150 with scores ranging from 157 down to 63. In PhyloTree, the top nine are nearly the same: 152, 16311, 195, 146, 16189, 16129, 16362, 150, and 16172 with scores from 139 to 40.

To illustrate how the use of those scores can enhance haplogrouping, consider the control-region profile 16221T 16291T 263G 309.1C 315.1C (sample BIO-03 Pa from the Basque province Biskaia [12]). Except for hotspot insertions at site 309, this profile matches the control-region variation of two complete mtDNA sequences (GenBank accession numbers GQ888728 and HQ675034), which are allocated to haplogroup HV4a1a. On the other hand, the closely related control-region profile 16291T 263G 309.1C 315.1C with an extra change 311T is found within haplogroup H2a2b (GenBank accession number AY339426). Nonetheless, the classification of the former profile as belonging to haplogroup HV4a1a is on safe grounds: transitions at 16291 occur very often (26 times in PhyloTree and 34 times recorded in [11]), whereas the 16221T variant seems to be confined to haplogroup H4a (with one back mutation recorded in PhyloTree).

Haplogrouping

As a first attempt towards a formal procedure, one might regard haplogrouping of a single sequence as the minimum (weighted) length extension of the mtDNA tree by adding the new sequence to the reference tree—or much better—to the total tree of complete mtDNA sequences underlying the reference tree. In other words, one would in principle seek for a location in the tree where the new sequence branches off so that the weighted sum of private mutations gets minimized. This process just constitutes the familiar sequential step in the Wagner tree procedure erst proposed in [13]. The obvious drawback is that the new sequence would typically get misplaced due to homoplasy (long-branch attraction) in the rather rare case that it belongs to an entirely new major haplogroup that is not yet represented in the reference tree.

The control-region is the mtDNA segment that is most commonly analyzed in forensic and population genetic studies. Haplogrouping control-region segments is often—but not always—straightforward using PhyloTree, where the optimization criterion can be fulfilled almost by visual inspection, at least guided by some experience. In any case, the strategy would be to screen PhyloTree for the mutational variants characterizing the sequence of interest and using the tools available in the preferred Web browser. The search strategies proposed here extend those previously used in [14, 15] for finding articles in which certain mtDNA mutations were mentioned; see Table 1.

To exemplify these strategies, consider the control-region profiles PK-187, PK-189, PK-249 of [16], all sharing the rare variant 16179del, which were allocated by the authors to haplogroup M30 but not classified further. In order to explore the control-region variation within this haplogroup, one could enter the Website http://www.ianlogan.co.uk/sequences_by_group/m30_genbank_sequences.htm and immediately find a complete mtDNA sequence (GenBank accession number AY922256) within haplogroup M30d1 listed with this particular deletion. This justifies to classify those three samples as M30d1 haplotypes. Alternatively, although more laboriously, one could scan the M node of the mtDNA tree provided by a Russian company (http://mtdna.gentis.ru/tx/hg/M/tx_M_304991.htm#) to find first the M4″67 node ("Taxon M_326510") in which the M30 node ("Taxon M30_326525") is nested and then explore the entire M30 subtree branch by branch until one eventually arrives at the desired sequence bearing 16179del ("Taxon M30d1_81654").

In a similar way, one could fine-classify the control-region profile PK-044 from [16] allocated to haplogroup W3. This sample shows four variants 16209C, 16255A, 119C, and 185A not covered by the motif of haplogroup W3. Ian Logan's Website for haplogroup W lists two complete mtDNA sequences (GenBank accession numbers GU147938 and HM214761) allocated to haplogroup W3a which share the first three of those seemingly private mutations. The same pair of sequences is then eventually retrieved on the Gentis Website under "Taxon N2W3a_322142". For further instances of fine-classification beyond PhyloTree, see "Results" and Supplementary File 1.

Multiplex genotyping design

Since control-region variation alone cannot settle the desired haplogroup assignment in many cases, the most natural

Table 1 Quick Web-based searches for mtDNA haplogroup affiliation

Goal	Website	Action / Query
Determining haplogroup	http://www.phylotree.org/tree/main.htm	Download the entire tree and search site numbers (e.g., "73")
Exploring the subhaplogroup levels	http://mtdna.gentis.ru/	Travel through the tree from its root to the targeted haplogroup (according to PhyloTree) and then scan the nested array of subhaplogroups ("child taxa")
Finding closely related lineages in GenBank	http://www.ianlogan.co.uk/sequences_by_group/haplogroup_select.htm	Click haplogroup and search private mutations (e.g., A73G)
Finding perfect matches in fragments	http://www.google.com/	Enter truncated profile (e.g., "'73G 263G 315.1C 489C 573.1C 573.2C'" or "'A73G A263G 315.1C T489C 573.1C 573.2C'")
Finding related partial profiles	http://www.google.com/	Enter partial profile (e.g., "A73G G94A PhyloTree" or "A73G G94A mtDNA" or "73G 94A mtDNA")
Finding specific variant	http://www.google.com/	Search variant (e.g. "A73G PhyloTree" or "A73G mtDNA" or "'73A >G mtDNA" or "73G mtDNA")

approach is to gain more information from the coding-region, for instance, either by directly sequencing several fragments [17] or by SNP screening via multiplex analysis [18–22]; see Table 1 in [23] for a list of articles on SNP multiplex analyses. The design of a multiplex SNP array would depend on whether it is supposed to (1) be used on a (sub-)continental scale without prior sequence information (e.g., [18–21]) or (2) be restricted to a specific haplogroup (e.g., [24, 25]) or (3) complement control-region information (e.g., [19, 26]). Further, the purpose of a multiplex analysis matters: either haplogroup assignment is aimed at (as in [18]) or discrimination power between samples in forensic casework is to be increased (1) without prior knowledge of haplogroup status (2) or with haplogroup allocation presumed, or (3) with control-region information given [27]. For the purpose of discriminating otherwise identical mtDNA control-region sequences one would preferably choose highly mutable sites, which however are typically ignored or considerably downweighted for haplogroup assignment.

There is no realistic solution for a worldwide universal SNP design, even when only discrimination power is targeted [23], given the limitations of casework samples with respect to availability of mtDNA for such additional analyses. Neither would any worldwide SNP array be of much use even for coarse haplogrouping since there are simply too many basal branches in the Eurasian mtDNA phylogeny (especially within haplogroup M). For the purpose of just identifying African, West Eurasian and Native American matrilineal ancestry, three multiplexes of small or medium sizes could suffice [21]. In contrast, South and Southeast Asian or Oceanian ancestry is difficult to determine through multiplex analysis since the knowledge of complete mitochondrial variation is yet quite limited; see, e.g., [28]. In this situation, one would best proceed with a large number of parallel multiplex analyses, in order to allocate profiles to main branches of the phylogenetic tree, and some subsequent nested multiplexes for refined classification. A multiplex design involving large numbers of SNPs [29] could then assist in general haplogroup assignment but never to the level provided by complete genome sequencing, given that a pre-designed multiplex array can only target known variation (inferred from the known part of the phylogeny and population data).

Then the ideal properties of any multiplex design for haplogrouping are the following: (1) it should either have a sub-continental focus or subclassify only a single major haplogroup, (2) it should shield against mtDNA outliers from other parts of the mtDNA phylogeny, (3) it should use phylogenetically meaningful default categories (such as L3*, M*, N*, R*, etc.), (4) it should avoid including sites of very high mutation rates (if possible). Two further points are that (5) the decision tree for haplogroup assignment should constitute a truncation of the entire mtDNA tree from PhyloTree relative to the tested variants of the entire mtDNA molecule, and that (6) the control-region information should be integrated in order to arrive at a haplogroup allocation at some finer level whenever (partial) control-region sequencing has been carried out.

As to points (1), (2), and (3), every sub-continent-specific multiplex that is designed for allocating mtDNA samples to the according haplogroups should provide information on the important multi-furcation pivots of the mtDNA phylogeny. Therefore, at least the following sites should be included: 1018 (recognizing haplogroup L3), 10400 (haplogroup M), either 9540 or 10873 (haplogroup N), and 12705 (haplogroup R). All of these mutations appear only once in PhyloTree Build 13 and have Soares et al. [11] scores of at most 2, so that recurrent mutations at these sites are not likely to blur a multiplex analysis. In particular, a reliable recognition of Native American mtDNA haplogroups in multi-ethnic populations cannot make do with just five mtDNA coding-region sites [30].

As to point (4), one cannot always avoid fast sites, such as position 3010, which is for instance the only site distinguishing haplogroup H1 from H, or worse, position 709 (relevant for the distinction of D4g1c from D4g1). Even the top-most variable nucleotide substitution at position 16519, which is ignored in PhyloTree, can be of use for particular haplogroups. Enigmatically, this site is very stable within haplogroup F, where 16519T (with the rCRS nucleotide) is virtually fixed in subhaplogroups F3 and F4, whereas 16519C is characteristic for subhaplogroups F1 and F2, so that it could make sense to employ this in a small multiplex design [31].

As to point (5), any "new strategy" for the discrimination of mtDNA haplogroups should really use the most recent update of PhyloTree and preferably employ additional information. A scheme solely derived from a truncated and distorted version of an mtDNA tree that has become obsolete (see, e.g., [32]) is of no help for the practicing forensic geneticist.

Results

Quality control enhanced by haplogrouping

Oversights or misdocumentation, phantom mutations, and artificial recombination are the main kinds of errors in mtDNA datasets; for a fine-classification of errors, see [33]. When (nearly) the entire mtDNA genome of a sample is sequenced, then at least with West Eurasian mtDNA samples haplogrouping is quite pedestrian, so that potential oversights could be seen immediately. With short stretches of mtDNA sequences such as HVS-I&II it may be difficult to distinguish oversights from natural back mutations since many of the variable sites are mutational hot spots. Nonetheless, if several expected mutations were not recorded, then one can predict incomplete documentation of nucleotide variants. This is the case with the two haplogroup U2e1 sequences (VP64 and VP65) from [34]: the former lacks four expected variants (16051G, 16362C, 73G, and 340T) and the latter three (16051, 16362C, 217C), which seems to reflect a reading problem at the beginning and end of the two sequenced parts rather than the effect of natural back mutations.

Phantom mutations are still an issue in forensic population studies despite the fact that they should be practically absent if standard sequencing protocols were followed. Private or seemingly private mutational variants of a given mtDNA profile are those variants which are not matched by the associated haplogroup motif. One could then search for additional matches in mtDNA profiles belonging to the targeted haplogroup. Otherwise, rare mutations indicate potential phantom mutations when they are repeatedly recorded as private variants on the background of different haplogroups within a population sample set [6]. Notorious control-region phantom mutations (artificially generated under suboptimal sequencing conditions) have been investigated in [7]. In particular, extension of the C stretch preceding site 310 typically led to seeming mutations further down in the 3′ direction at sites 317, 320, 330, 343, 345, etc. (see Table 5 of [7]). For instance, among 15,195 sequences recorded in EMPOP Release 6 that cover these sites, there is not a single instance of 320T. In [11] this variant received two hits, both coming from the study [35] that had a few problems with phantom mutations [7]. In the GenBank entries of complete mtDNA sequences, one additionally finds this mutation in EF184595 (from the problematic dataset [36]; see [37]) and in FJ383190 [38] as well as in HM238198.

Now, in the dataset provided by [34] we are seeing five samples (VP24, VP45, VP57, VP79, and VP90) of haplogroup status R0, J1c2, T1a, HV0, and U8b, respectively, for which 320T was recorded. Moreover, in two cases, this is followed by another unknown transversion variant, 324G or 328C, respectively, for which the scores in EMPOP and [11] are all zero. In a study of Japanese mtDNA variation [39], 320T was recorded seven times on different haplogroup backgrounds (B4, B4a, B4b1b, B4b1b′c, B4f, D5a2a1, and N9a2a). A common cause for phantom mutations in the HVS-II segment is the time-saving approach of reading only the light strand, which however is not recommended if one wishes to achieve optimum sequence quality in forensic DNA analysis [40].

Artificial recombination between HVS-I and HVS-II has haunted large databases as well as single studies since many years [41–43]. In the Japanese study [39], the classification of HVS-I&II sequences incorporated the knowledge of the mtDNA phylogeny at the time. Now, with a deeper knowledge of the mtDNA phylogeny, one can be more specific in view of Build 13 from PhyloTree. For instance, the HVS-I&II profile no. 94 (16136C 16183C 16189C 16217C 16284G 73G 146C 199C 202G 207A 263G 315.1C), classified as B4b1 by the authors of [39], would now best be traced in PhyloTree. The rare variant 202G (occurring as a unique event in both PhyloTree and [11]) immediately points to haplogroup B4b1a1, and in fact, all other variants are captured by this classification as well, with no variant from the profile left unmatched. So, in this case, the profile exactly constitutes that particular haplogroup motif. However, confronted with the unclassified sample no. 739 (16189C 16190T 16193.1C 16193.2C 16261T 16362C 73G 146C 199C 202G 207A 263G 309.1C 315.1C), we would be led to exactly the same haplogroup slot with HVS-II alone, but the HVS-I part is at odds with this classification, as it would need to postulate three back mutations plus additional mutations. The HVS-I part alone would rather

point to haplogroup B4a (http://www.ianlogan.co.uk/sequences_by_group/b4a_genbank_sequences.htm). There is a second case of artificial recombination in [39]. The HVS-I&II profile no. 92 (16136C 16183C 16189C 16217C 16284G 73G 103A 263G 315.1C) gets allocated to B4b1a1 according to its HVS-I part, but its HVS-II part is incompatible with this clear-cut haplogroup allocation and suggests B5b status instead (e.g., either B5b1 or B5b2c; see http://www.ianlogan.co.uk/sequences_by_group/b5_genbank_sequences.htm).

As for another instance of artificial recombination from [34], the recorded profile 16126C 16270T 16294T 16304C 73G 242T 263G 295T 309.1C 315.1C (sample VP61) testifies to J1b1a status when just HVS-II is consulted, but the HVS-I part is definitely incompatible with this assignment and suggests haplogroup T2b status instead. Further subhaplogrouping is not possible as 16270T is not found among 87 complete T2b mtDNA sequences drawn from GenBank as listed on Ian Logan's Website for haplogroup T2b. Finally, a classic case of sample crossover has occurred in [44] between two samples from haplogroup B2 (similar to haplotype LPAZ008) and haplogroup C1 (similar to haplotype LPAZ080), which has given rise to the artificial hybrid mtDNA sequences LPAZ092 and LPAZ094; see Table 2.

In conclusion, haplogrouping is an invaluable tool for narrowing the focus on potential sequencing or documentation errors. Coarse-grained haplogrouping may be sufficient for comparison of the mtDNA pools of different geographic or ethnic populations but not so for forensic databases. For example, if one wished to investigate the profile 16129A 16180G 16223T 16234T 16288C 16298C 16327T 16518T 16519C 16527T 54A 73G 249del 263G in the sequence range 16020–300 (drawn from [45]), the obvious assignment to haplogroup C would not be of much help since this would leave us uninformed about the rare variants 54A and 16518T, which do not occur in the list of Soares et al. [11] and therefore constitute candidates for scrutinizing. A search of 16518T in PhyloTree immediately leads to haplogroup C5c, and Ian Logan's Website for C5 subsequently provides us with a near-match (in subhaplogroup C5c1, GenBank accession FJ951452) of the entire profile with differences only at positions 16093, 16129, and 16180, all of which are unsuspicious.

Multiplex haplogrouping

In order to demonstrate the haplogrouping problems via multiplexing, we reassessed in detail the haplogrouping performed with the 33 mtSNP typing assay (33-plex) designed by [46]. The same assay seems to have been used later in [47] (without reference to [46]). None of the major West Eurasian superhaplogroups (M, N, and R) were explicitly targeted by these mtSNPs. As a result, one arrives at odd default categories, so that a mixed bag of unassigned lineages in- and outside haplogroup N arises, misleadingly referred to as "R". This in part explains suboptimal results in the case of the Turkish sample since mtDNA lineages of ultimate African or East Asian ancestry could not be identified (apart from some Near-Eastern haplogroups which are rare in Europe). Moreover, a number of minor basal West Eurasian haplogroups (such as M1, R1, R2, R0a, HV1, HV2, etc.) were not interrogated. Due to this lack of basal distinction, some specific lineages outside haplogroup R would be erroneously allocated to particular R subhaplogroups because of recurrent mutations occurring at the selected SNP sites. This is the case with sample mt150*, where the HVS-I&II variation points to the Southeast Asian haplogroup E1a1a. The multiplex analysis of [47] can also yield misleading results when dealing with haplogroup HV. Thus, their multiplex design does not distinguish between subhaplogroups E of M and HV of R, but E1a1 has 14766C as a single characteristic variant within haplogroup E, just as haplogroup HV does within R0. As for point (4) (see "Methods"), five of the top ten mutational hotspots in the coding-region (viz. sites 709, 5460, 1719, 13708, and 8251) were actually selected in [46, 47], among which is the top hypervariable coding-region site 709. Finally, the single default category named "R" is too ambiguous because it embraces most African and Asian mtDNA lineages from various major haplogroups. Since 709 was chosen as the single characteristic site for "N2 status" every complete mtDNA profile with haplogroup defined according to the decision tree would automatically fall into the category "N2" whenever it bears the 709 variant and is outside haplogroups HV, JT, U, and X. For example, most of the African haplogroup L1b would thus be shunted into "N2".

As for point (5) (see "Methods"), the decision tree of [46] and its apparent update in [47] (where "2709" should read "2706" throughout) as a truncation of Build 8 of PhyloTree may miss minor haplogroups due to mutations that cause a primer mismatch for a targeted SNP. Consider, for instance, the haplogroup U5b2a1b sample mt180 (see no. 6 in Table 2), which was allocated to haplogroup R by the HVS-I&II analysis [48] and received the status "Not full profile" by the multiplex analysis [47]. However, the latter analysis shows that two SNP sites positively supported U and U5 status and only the information for site 1719 was missing. The latter site is rather irrelevant for the allocation to haplogroup U5. On the other hand, the failure of the 1719 reaction (i.e., due to a primer mismatch) in this case even gives a hint as to which subhaplogroup of U5 can best be searched for near-matches with the HVS-I&II profile: site 1721 is characteristic for haplogroup U5b2. In fact, the optimal near-match is found within a particular subhaplogroup of U5b2.

Now, with more recent updates of PhyloTree, one could discern more haplogroups than those that were listed in Table 1 of [47]. For instance, the haplogroup J1c samples

Table 2 Instances of (partial) control-region sequences with haplogroup classification according to original publication, mtDNAmanager, MitoTool, HaploGrep, and manual Web searches

No.	Sample ref.	Range[a]	Haplotype[b]	HG status in article	"Expected" vs "Estimated HG" by mtDNAmanager [50][c]	MitoTool [52][d]	HaploGrep [53][e]	Likely HG [2]	Related sample from GenBank
1	[61]	CR	16086 16129 16209 16223 16272 73 152 225 249d 263 315 + C 316 489 523-524d	M5	L3f* vs L3f* (Afr, WeuA, Ad); M20 vs M20 (EA, Oc)	M20	M20 √	M20	HM030505
2	[62]	CR	16172 16183C 16189 16209 16223 16258T 16311 16362 73 185A 189 195 234 263 309 + C 315 + C 523-524d	N9a5	L3f2 vs L3f2 (Afr, WeuA, Ad, EA); M10a (Oc)	N10a	N10a ~	N10a	HM030542
3	[62]	CR	16069 16172 16223 16278 16291A 16298 16362 73 150 152 199 263 309 + CC 315 + C	N*	L3b1a vs L3b1a (Afr, Ad, WeuA); J2 vs L3b1a (EA); N10b vs N10b (Oc)	N10b	N10b √	N10b	HM030500
4	[48]	HVS-I/II	16114A 16126 16218 16223 16275 16291 16356 16390 16391 73 263 309 + C 315 + C	L3	M3 vs M3 (Afr); E1a1a vs M3 (OC, Ad); M3 vs M3 (WeuA, EA)	H2a2b; H2a5; H34; H20;	M52 ?	M52a	EF093557
5	[48]	HVS-I/II	16223 16291 16362 16390 73 263 309 + C 315 + C	L3	E1a1a vs E1a1a	E1a1a	E1a1a √	E1a1a	EF093544
6	[48]	HVS-I/II	16189 16319 16325 73 150 152 263 315 + C	R	U5b2a1 vs U5b2a1 (Afr, WeuA); G1a1 vs U5b2a1 (EA, Oc, Ad)	U5b1a1b	U5b2a1b √	U5b2a1b	GU296545
7	[48]	HVS-I/II	16051 16162 16213 16266 73 146 263 315 + C	U2	H1a vs H1a	H1a3; H1h; H1n; U2b	H1a3 ~	H1a3	EU979418
8	Family Tree DNA	CR	16069 16126 16145 16231 16261 73 150 152 195 215 263 295 310 + T 315 + C 319 489 513	n.a.	J2a vs J2a	J2a1a1	J2a1a1 √	J2a1a1	GU903270
9	Family Tree DNA	CR	16086 16222 16224 16270 16311 16519 73 146 263 315 + C	n.a.	K2b1 vs K2b1	K2b1a	K2b1a √	K2b1a	EU770310
10	[63]	CR	16086 16239 16311 16320 73 150 263 315 + C	n.a.	L3e2 vs ?	HV9; H13a2b; HV7; HV10; HV11	U5b2a1a2 ?	U5b2a1a1	GU296544
11	[64]	CR	16224 16519 73 152 204 263 315 + C 497 524 + AC	?	K1a vs ?	H16a; H9; H3g; K1a3a1b	K1a4c ~	K1a *(K1a4a1)*	EU597496
12	[64]	CR	16069 16261 73 185 189 263 295 315 + C 462 489	?	J1c7 vs R8a2	J1c7	J1c + 16261 ~	J1c6	AY495209
13	[64]	CR	16224 16519 73 152 204 263 272 315 + C 497 524 + AC	?	K1a vs ?	H16a; H9; H3g; K1a3a1b	K1a4c ~	K1a *(K1a4a1)*	EU597496
14	[64]	CR	16183C 16189 16193 + C 73 262 263 285 309 + CC 315 + C 323 385 523-524d	?	U1a1 vs ?	H1g; H1q; U1a1	U1a1 √	U1a1	AY882396
15	[64]	CR	16069 16145 16207 16222 16231 16261 73 150 152 195 215 246 263 295 309 + CC 315 + C 319 489 513	?	J2a vs ?	J2a1a	J2a1a ~	J2a1a	FJ348157
16	[64]	CR	16183C 16189 16193 + C 73 262 263 285 309 + C 315 + C 323 385 523-524d	?	U1a1 vs ?	H1g; H1q; U1a1	U1a1 √	U1a1	AY882396
17	[64]	CR	16153 16298Y 72 73 93 95C 263 309 + C 315 + C	?	? vs ?	H2a2; H2a; H2; H; HV0; V; V7a; R0	V7a ?	V7a	AF347006
18	[64]	CR	16153 16298 72 73 93 95C 263 309 + C 315 + C	?	? vs ?	H2a2; H2a; H2; H; HV0; V; V7a; R0	V7a ~	V7a	AF347006
19	[64]	CR	16183C 16189 16193 + C 16217 16519 73 263 309 + CCC 315 + C 498d 499	B4b	B4b vs B4b	B4b; B2	B4b √	B4b *(B2d)*	EU095550
20[f]	[64]	CR		D4/G	D4/G vs D4/G	M74; D4j11	D4e2a √	M *(M43a)*	FJ770954

Table 2 (continued)

No.	Sample ref.	Range[a]	Haplotype[b]	HG status in article	"Expected" vs "Estimated HG" by mtDNAmanager [50][c]	MitoTool [52][d]	HaploGrep [53][e]	Likely HG [2]	Related sample from GenBank
			16223 16311 16362 16519 73 263 315 + C 489 573 + CCC						
21	[64]	CR	16093 16172 16223 16297 16311 16362 16400 16519 73 146 185 263 315 + C 489	D4/G	N10a vs D4/G (Afr, WeuA , Ad) M7e vs M7e (Oc, EA)	M7e	M7e ?	M74	HM030520
22	[64]	CR	16153 16298 72 93 95C 263 309 + C 315 + C 523-524d	HV0	HV0 vs HV0	H2a2; H2a; H2; H; HV0; V; V7a; R0	V7a √	V7a	AF347006
23	[64]	CR	16298 16519 263 315 + C	HV0*	HV0* vs HV0*	H2a2; H2a; H2; H; HV0; V; R0	HV0 ~	HV *(V)*	AY495306
24	[64]	CR	16067 16311 152 195 263 315 + C	HV1 H11	HV1 H11 vs HV1 H11	H11; H2b	HV1b + 152 ~	HV1 *(HV1b)*	HQ165756
25	[64]	CR	16129 16185 16223 16224 16260 16298 16519 73 151 152 249d 263 315 + C 489	M37a	Z vs Z	Z1a	Z1a √	Z1a1a	AY339515
26	[64]	CR	16069 16126 16214 16311 16362 73 150 195 235 263 295 309 + C 315 + C 326 489	R0a J	A4 vs J2	J2a2a	J2a2a ~	J2a2a	EF660967

[a] Reported sequence range: *CR* entire control-region; *HVS-I/II* HVS-I and HVS-II

[b] Mutations are recorded with respect to the rCRS [4]. All mutations are transitions unless a letter specifies a transversion (A, C, G, T) or a deletion (d); an insertion is signaled as "+" followed by the inserted nucleotide(s)

[c] The haplogroup status is identical for the five metapopulations considered by mtDNAmanager with the exception of those cases specified in round brackets where population codes are as follows: *Afr* African; *WeuA* West Eurasian, *EA* East Asian, *Oc* Oceanian, *Ad* Admixed

[d] The primary classification provided by MitoTool reflects completely matched motifs of specific haplogroups

[e] HaploGrep provides three levels of security for the haplogroup classification (designated here by symbols ?, ~, √: haplogroup assignment not reliable probably due to insufficient data (?), haplogroup assignment critical (~), and high amount of polymorphisms explained by the sample's haplogroup affiliation (√))

[f] If 573 + CCC is changed to 573 + CC in haplotype #20, then mtDNAmanager yields D4/G vs D4/G

mt183 was allocated to J on the basis of the HVS-I&II but received status "Hg not defined" (because the presence of the rCRS-variant at site 2706) according to the multiplex analysis [47]. However, PhyloTree highlights the recurrent mutation at site 2706 in the definition of haplogroup J1c3c. Therefore, the SNP analysis is consistent with the HVS-I&II haplotype and in conjunction now yields this finer classification.

For the Danish sample, haplogroup assignments based on HVS-I&II [48] was poorly contrasted with the multiplex analysis [47]. The use of the same sample numbering/codes permits the reader to combine the results from the three studies [46–48]. In some cases, the multiplex analysis of [47] would support the haplogroup allocation obtained from PhyloTree. For instance, consider sample mt098 (no. 7 in Table 2) which bears the motif 73G 16051G 16163G characteristic of haplogroup H1a3: in fact, the multiplex analysis correctly identifies this sample as belonging to H1, so that an initial guess of haplogroup status "U2" on the basis of HVS-I/II can firmly be rejected.

A particularly interesting case in this regard constitutes the outlier sample mt055* from [48] which received the status "Hg not defined" in [47]. In fact, the HVS-I&II profile points to haplogroup M52a (of North Indian and Nepalese provenance) in view of the presence of transitions 16275G and 16390A (despite the lack of 16327A). The first hint at this unexpected allocation comes from the rare transition at site 16275 (with Soares et al., score 1), which occurs only once in PhyloTree Build 13, viz. as one of the three control-region mutations defining M52a. A Google search with "A16275G PhyloTree" provides (via Ian Logan's Website) two additional GenBank sequences that can be allocated to other haplogroups (W1f and G1a1a, respectively). A further search with "16275G mtDNA" points to Family Tree DNA (FTDNA) control-region profiles from haplogroup T2b and from haplogroup M52b (matching the complete mtDNA sequence with Genbank accession number FJ383488). More interestingly, the latter Google search also directs one to the Release history for EMPOP 2, where the user gets informed that the sequence information for the Danish sample has been expanded to the entire control-region, so that the variants 16519C 489C 573.1C 573.2C 573.3C 573.4C 573.5C 573.6C come on top of the HVS-I&II profile. This is in line with haplogroup status M and covers also the insertion of three cytosines at site 573, which is characteristic of haplogroup M52a. An EMPOP query for the two variants 16275G and 16390A together, however, yields yet another hit with a forensic sample from Thailand, which was previously published by [49] and has haplogroup R9b1b status. The partial profile 16223T 16275G 16390A 489C 573.1C 573.2C 573.3C is distinctive enough to guarantee M52a status, although this entails that the definition of M52a may need some modification in the future by not necessarily requiring 16327A, for example. If one would yet be in slight doubt about this haplogroup allocation, one could of course consult the coding-region. The study [47] has actually done that: one of the two characteristic sites of the larger haplogroup M52′58 is 5460, which is exactly the site that gave rise to the status "Hg not defined" in [47].

Automated haplogrouping

The Web-tool mtDNAmanager [50] is supposed to deliver for any partial (HVS-I/HVS-II/HVS-III) or entire control-region sequence two kinds of haplogroup status: "expected" and "estimated". Any discrepancy between expected and estimated status is supposed to be regarded as dubious, warranting some verification by the user (if possible). The exact processes by which these inferences are made are just vaguely indicated in the accompanying article [50]. Therefore the implemented algorithm can only be treated as a black box, for which some features might be uncovered through some series of focused tests. To this end we took some published entire control-region sequences and modified some of them by single changes.

To begin with, consider the haplogroup J2a1a1 sequence no. 15 of Table 2. This haplogroup, as well as its nested superhaplogroups J2a1a and J2a1, are all unknown to mtDNAmanager, although each of them has diagnostic control-region sites and J2a1a1 has essentially been in place since PhyloTree Build 3 (1 March 2009). Thus, when the entire control-region is offered as input, then "J2a" is returned for both expected and estimated haplogroup. Now, if one removes the variant nucleotide at 152, then "J2a" versus "J2" will be the answer to the corresponding query, and if one drops 150 instead, then "J2a" versus "J" will be returned. Finally, if just 295 is removed, then "J2a" versus no haplogroup status proposed (to be interpreted as "?") will be the output. This shows that the "estimated haplogroup" status does not tolerate single back mutations at any of the diagnostic sites listed in the protocol of mtDNAmanager; note also that in this particular example, the mentioned transitions are mutational hotspots, so sporadic reversions may not be infrequent within these haplogroups. If this feature was properly understood by the user in a laboratory, then the analyst could go back to the electropherograms and check. But if everything appeared to be correct in any of these hypothetical cases, then the only reasonable output would be J2a1a1 in view of the wealth of mutations left that signalize this quite specific haplogroup status.

The strict requirement handled by mtDNAmanager that all variants from a perceived haplogroup motif must be present can be quite misleading. In particular, PhyloTree itself or the listing of all complete mtDNA sequences from

GenBank (see Ian Logan's Website) falling into a particular subhaplogroup may clearly demonstrate the reversal of some basal nucleotide variant (relative to rCRS) or the parallel gain of 73G which is apparently given a special treatment by mtDNAmanager. For example, haplogroup V7a bears the HVS-I&II motif 16153A 16298C 72C 93G 263G (and 309.1C 315.1C). The sequences nos. 17, 18, and 22 of Table 2 are therefore clearly members of this haplogroup but are partially left unclassified by mtDNAmanager (Table 2). Moreover, the variant 95C also present in these three sequences is confirmed to occur in some haplogroup V7a member (GenBank accession number AF347006).

The next test case is a haplogroup K2b1a sequence (no. 9 in Table 2). This haplogroup has also been present in PhyloTree since Build 3, but mtDNAmanager recognizes only the next superhaplogroup K2b1. Now, if one queries only the HVS-I part (or together with the HVS-III part which does not contain any variant here), then expected and estimated haplogroup differ: "K2b1" versus "U5 K". Although the variant 16270T is diagnostic for haplogroup U5, its parallel presence within haplogroup K2b is well confirmed (giving rise to K2b1) and the additional mutation (16222T for K2b1a) renders the assignment K2b1 quite reliable.

It is also instructive to see how mtDNAmanager deals with obvious instances of artificial recombination. First, the SWGDAM database [51] curated by the FBI has never been fully revised and still contains two very clear cases of hybrid profiles; see entries nos. 1 and 2 of Table 3. The SWGDAM profiles were added to the reference data on which mtDNAmanager operates. Table 3 also includes a few other instances we have found in the recent forensic datasets (see above and Supplementary File 1). In no case mtDNAmanager gave a warning about the hybrid status of these artificial recombinants, essentially because no precautions were taken to contrast the haplogroup predictions from the separate segments HVS-I and HVS-II. And in no case were the two different haplogroup sources identified. In one case (no. 5) "expected" and "estimated" haplogroup status even coincided, so that the user would not be alerted at all. This shows that mtDNAmanager is unhelpful for discovering clear instances of artificial recombination.

In summary, (1) mtDNAmanager works with a restricted and not properly updated list of haplogroup motifs, (2) variants are not weighted according to their mutability but instead according to polymorphism counts (in a cryptic fashion), (3) the distinction between "expected" and "estimated" haplogroup is not well justified and can rather mislead the average user, (4) mtDNAmanager does not incorporate the information that can be drawn from the complete mtDNA sequences stored in GenBank, (5) it handles ambiguous "meta-population" categories (e.g., "Admixed groups" includes Dubaians, Egyptians, "Hispanics", Northern Africans, and

Table 3 Artificial mtDNA recombinants from forensic databases and their haplogroup classification according to mtDNAmanager, MitoTool, HaploGrep, and manual Web searches

No.	Sample ID, source	HVS-I[a]	HVS-II[a]	HG status [2]	"Expected" vs "Estimated" 'HG'" [50]	Best HG according to MitoTool [52]	Best HG according to HaploGrep [53]
1	USA.AFR.000942 [51]	16126 16187 16189 16223 16264 16270 16278 16293 16311 16519	73 249d 263 290d 291d 309 + C 315 + C 489	L1b × C1	C1c vs X M31a1	M31a1	M31a1 ?
2	FRA.CAU.000084 [51]	16298	73 185 188 228 263 295 315 + C	HV0×J1c	? vs ?	H2a2, H2a, H2, H, HV0, V, R0	J1c2 ~
3	VP61 [34]	16126 16270 16294 16304	73 242 263 295 309 + C 315 + C	T2b × J1b1a	T2b vs T U5	T2b4	T2b + 16296! ~
4	739 [39]	16189 16190 16193 + CC 16261 16362	73 146 199 202 207 263 309 + C 315 + C	B4a × B4b1a1	M7c vs R30a*	H1b; H1g; H1q; H1h; H1n; H7a1; B4a1a; B4a1c4	R31 ?
5	92 [39]	16136 16183C 16189 16217 16284	73 103 263 315 + C	B4b1a1 × B5b	B4b1 vs B4b1	H1g; H1q; B4; B4b1	B4b1 ~
6	LPAZ092 [44]	16181 16189 16217	73 185 249d 263 290d 291d 309 + C 315 + C	B2 × C1	B4 vs ?	H1g; H1q; B4	B4 ?
7	LPAZ094 [44]	16182C 16183C 16189 16223 16298 16325 16327 16344	73 263 309 + CC 315 + C	C1 × B2	U5b2a1 vs M8	H1g; H1q; N9b	C4c1b ?

[a] Mutations are recorded with respect to the rCRS [4]. All mutations are transitions unless a letter specifies a transversion (A, C, G, T) or a deletion (d); an insertion is signaled as "+" followed by the inserted nucleotide(s). A reversal of a mutation along a branch of the mtDNA tree is highlighted by the suffix "!" (as in PhyloTree)

Venezuelans) which can misdirect otherwise clear-cut haplogroup allocation, and (6) does not shield against serious errors in the employed reference database or input sequences.

The MitoTool software [52] has a wider scope than mtDNAmanager since it provides several applications of interest in the biomedical field and it allows dealing with entire genomes (not merely control-region segments). Here, we only consider the module that allows for haplogroup classification. The strategy followed by this web interface application seems to be based on a score system that takes into account the number of exact matches between the tested profile and the diagnostic variants that define the different haplogroups of the worldwide phylogeny, in a similar way as mtDNAmanager does, but here too the user is not informed about the exact optimality criterion. The classification tool performs substantially better than mtDNAmanager as can be concluded from the results of Table 2. However, like mtDNAmanager, the procedure implemented in MitoTool does not allow any detection of signals of artificial recombinants (Table 3).

Recently, a more sophisticated tool for automated haplogrouping has been proposed [53]: HaploGrep allows for the classification of mtDNA segments of any length into haplogroups using the current build of PhyloTree as the reference classification tree and an algorithm that involves the weighting of diagnostic sites according to their relative positional mutation rates. The weighting scheme employs a simple linear-affine scaling of the absolute number of occurrences of a particular mutation in the current build of PhyloTree. For example, given that the maximum score of a mutation recorded in PhyloTree Build 13 equals 139, all (and only those) mutations that occur in Build 13 get weighted by 140 minus their respective numbers of occurrences. Thus the top mutation has weight 140−139=1, whereas the next hotspot mutation has weight 140−89=51. A factor of 51 however does not properly reflect the relative likelihoods of these two hotspot mutations. At the other extreme, the difference in weights (139 versus 134) between a mutation recorded only once in Build 13 and a medium-frequent control-region mutation with PhyloTree score 6, say, is far too small in order to have discriminative power.

Table 2 includes the results of HaploGrep (using Build 13 of Phylotree) for the same test haplotypes used for mtDNAmanager and MitoTool. A total of 19 out of the 26 test samples (73 %) from this table were classified as one could achieve by using PhyloTree and scanning individual complete mtDNA sequences manually. An interesting feature of HaploGrep is that it implements a system that alerts the user about the performance of the haplogroup classification, indicating for instance that the haplogroup assignment might not be reliable probably due to insufficient data, or that the haplogroup assignment was uncertain. HaploGrep however has not been designed to detect artificial recombinants; nonetheless, it seems to perform better than mtDNAmanager or MitoTool in the sense that HaploGrep detects at least in some instances the major component of the artificial mixture. In these cases however, the user is not alerted about the presence of artificial recombination. For instance, the first case sample #1 in Table 3 is a recombinant L1b × C1, while HaploGrep (Build 13) indicates a haplogroup M31a1 status.

Given that HaploGrep performs better than other available bioinformatics tools for haplogrouping, it is worth to investigate it more deeply. By way of examples (see Tables 2 and 3), we evaluate the algorithm implemented in HaploGrep. From the parsimony point of view, the mismatch distance between a sample sequence and a haplogroup motif would be relevant; whereby a moderate scaling relative to the phylogenetic position of a haplogroup and the number of seemingly private mutations in the sample may not be unreasonable. However, in HaploGrep, the phylogenetic position of a haplogroup is measured by the distance to a particular extant (contemporary) mtDNA sequence, viz. the rCRS. As a consequence, the drop from a 100 % rank for perfect haplogroup allocation to the rank for a slightly imperfect allocation can be quite drastic for sample sequences phylogenetically close to the rCRS but very minor for African haplogroup L0 mtDNA sequences. To illustrate this reference-associated bias, we have selected six peripheral haplogroups from the mtDNA phylogeny, each characterized by exactly two control-region mutations. We have then deleted exactly one of the two characteristic variants, first the one with the higher mutational score and then the other one. The deletion of one variant had no or very little effect on the haplogroup assignment returned by HaploGrep (Build 13) but a considerable effect on ranks (alias quality scores) of the HV subhaplogroups in contrast to the L0 and L1 subhaplogroups; see Table 4. A similar effect is caused by the addition of a private variant (16303A), which occurs in Build 13 exactly once, namely in the motif of the Australian haplogroup O1. However, the addition of a variant (e.g., 16307G) that is not recorded in PhyloTree has no effect as such mutations are generally ignored in HaploGrep (data not shown).

Apart from the examples given in Tables 2 and 3, one can also test how the software implemented in HaploGrep behaves when dealing with more ambiguous profiles. For instance, if one tries to classify the profile 16093C (sequence range: 16024 to 16400), which could perfectly fit within, e.g., haplogroup H or many other closely related haplogroups, it is classified by HaploGrep (Build 13) in the first place as belonging to the South Asian haplogroup "R8a1+16093" or equally probable as R8a1b, both with quality rank 100 %. One can indeed find profiles 16093C 16519C in India (recall that the variation at 16519 is

Table 4 Quality scores (ranks) of haplogroup assignment in HaploGrep [53] for modified haplogroup control-region motifs

Profile with HG motif	Length of HG motif string[a]	Loss of one motif mutation		Gain (and loss) of mutation	
		Variant	Rank	Variant(s)	Rank
H1b	4	+16356!	90.1	+16303	86.9
		+16362!	90.9	+16362! +16303	75.8
HV12a	3	+16220C!	82.8	+16303	87.2
		+16292!	84.2	+16292! +16303	67.5
R0a1a2	9	+95C!	93.7	+16303	94.1
		+93!	94.0	+93! +16303	87.4
N1b1d	10	+188!	94.0	+16303	94.5
		+185!	94.4	+185! +16303	88.3
L3f1b2	10	+16266A!	95.0	+16303	94.7
		+16172!	95.6	+16172! +16303	89.9
L2b1a	17	+16355!	98.4	+16303	96.5
		+146!!	98.5	+ 146!! +16303	95.2
L1b1a1	20	+264!	97.0	+16303	97.0
		+16215!	97.0	+16215! +16303	94.0
L0a1c	20	+16287! +1612	98.1	+16303	97.1
		9!!	98.2	+16129!! / +16303	95.5

[a]Relative to the rCRS according to PhyloTree Build 13 (see Supplementary File 2)

disregarded in PhyloTree and hence in HaploGrep), but 24 out of 3+29 matches of either profile 16093C or 16093C 16519C in the EMPOP database are from West Eurasia and therefore not belonging to haplogroup R8. Since 16093 is a mutational hotspot one avoids the formal erection of subhaplogroup status within haplogroups HV or H or H1 etc. on the basis of this mutation alone, so that 16093 is usually recorded only as a co-motif mutation which cannot stand by its own. In fact, in no case is 16093 the single motif mutation of a named haplogroup in PhyloTree; quite often it is not even clear whether 16093 is part of the motif as indicated by the bracketed notation "(16093)". For this reason, R8a1 + 16093 cannot be regarded as a haplogroup; the mtDNA sequence(s) allocated to this point of the mtDNA phylogeny would still await characterization of a subhaplogroup; it is then not clear a priori whether or not 16093C will enter the corresponding motifs (as 16093C appears on virtually all haplogroup backgrounds). A 100 % match with a motif can therefore not be interpreted as a successful classification of high quality without further considerations.

The classification tasks that are most difficult (if not frustrating) are (partial) control-region profiles nearly identical to the ancestral (partial) control-region profile of a haplogroup which has dozens of subhaplogroups without characteristic mutation in that segment. In these cases, one would definitely need some coding-region information for any successful haplogroup allocation. For instance, the ancestral motif of haplogroup L3 has only rCRS-variant 16223T within the (HVS-I) range 16024–16400 and this motif was retained in the descendant haplogroups M and N. In fact, there are several subhaplogroups of the latter ones with the same ancestral motif. If no extra mutations were observed in a sample sequence, then classification would become a lottery, whether automated or not. The same is true for the two descendant motifs 16223T 16362C (matching, e.g., haplogroups D, G, M6, and M9) and the rCRS profile (matching haplogroups R, HV, H, etc.) within that HVS-I range. When these three ancestral motifs are entered into HaploGrep, then huge lists of potential haplogroups are returned for 16223T and 16223T 16362C, but the empty input motif in the case of the rCRS is rejected, although this real sequence is found 956 times among 13467 sequences in EMPOP. In the case of input sequences without any scored variant, rank calculations are infeasible or meaningless because of the problem of denominator zero in the formula employed by HaploGrep.

A general problem with most (if not all) approaches to automated haplogroup assignment is the restriction to the current build of PhyloTree, so that the closely related complete mtDNA sequences that are allocated to a specific haplogroup do not get screened one by one for additionally matched variants, which would strengthen the proposed haplogroup assignment. That is, the near-matching strategy (see, e.g., [54]) is in general not employed with automated approaches. For instance, the incorrect assignment of sample no. 10 of Table 2 to haplogroup U5b2a1a2 by HaploGrep (Build 13) is due to the conflicting signals at the two hypervariable sites 16192 and 16311 in the sample for subclassification within the larger haplogroup U5b2a1a.

Scanning the dozen complete mtDNA sequences from U5b2a1a stored in GenBank and conveniently displayed on the U5b2 Website of Ian Logan reveals a perfect match with the control-region profile of a member of U5b2a1a1 (GenBank accession number GU296544), thus solving the issue.

A semi-automated approach is not per se a guarantee for achieving an appropriate haplogroup allocation, especially when the initial algorithm sorts mtDNA samples belonging to single minor haplogroup into distinct slots. This might have happened, for example, to the three haplogroup R2 members in the dataset of [55], which were differently labeled as "R", "R0", and "R2", despite the fact that the characteristic haplogroup R2 variant 16071T has score 1 in both Soares et al. [11] and Build 13 of PhyloTree (and is additionally accompanied by the hotspot variant 152C). The coarse haplogrouping of several haplogroup J2a2a samples to "J" and one haplogroup L3h1a2a sample to "L3" also indicates that the manual search strategies were possibly not yet sufficiently elaborate.

Conclusions

For a non-expert or a novice in mtDNA analysis, it may be tempting to employ automated tools for haplogroup assignment as an all-in-one solution without carrying out additional analyses. None of the tools tested in the present study are free of shortcomings in design and scope. In the case of mtDNAmanager, the classification is based on an algorithm that was not published in detail and whose performance is very poor. There is no justification for using this software as it returns results that are clearly misleading. MitoTool performs better than mtDNAmanager but worse than HaploGrep; in contrast to mtDNAmanager, MitoTool can run complete mtDNA genomes. When entire genomes are used in MitoTool instead of partial control-region sequences, the classification algorithm performs better. The more sophisticated algorithm of HaploGrep performs reasonably well with a number of control-region profiles but it is currently not exempt of problems because it delivers haplogrouping ranks as quality scores that show a strong bias against sequences phylogenetically close to the rCRS.

Therefore, given the in part questionable performance of present algorithms for haplogroup classification when using difficult short-segment profiles, one cannot yet recommend automated classification as the single strategy for sorting partial mtDNA sequences; instead, for the time being, a combined automated and manual approach is more successful. A general drawback of most implemented classification tools is the restriction to matching haplogroup motifs in one way or the other. An alternative could be a formal near-matching strategy operating on a large database of allocated partial mtDNA sequences. For HVS-I sequences this has been realized within the Genographic Project but is limited by insufficient information content of the short HVS-I region and a far too coarse haplogroup scheme [5]. A combination of both approaches that incorporates reconstructed ancestral mtDNA sequences (haplogroup motif sequences) together with near-matching to a reliable expanded database would be desirable.

In any case, the user should go through PhyloTree and Ian Logan's Web pages manually and check if the haplogroup status delivered by an automated software or some initial manual sorting is plausible. Although there may be some discussion about the way how branching patterns are being rearranged and haplogroups renamed in PhyloTree, this resource serves as the gold standard to which any attempts of haplogrouping are to be compared. Although the large database of published complete sequences is not totally error-free, the most problematic datasets have been disregarded by PhyloTree [56, 57]. If necessary, single dubious sequences in otherwise unproblematic datasets can be excluded for the near-matching search [37]. Any missing mutations or hybrid motifs can be highlighted with the help of this classification tree and the strategies outlined in [10, 58].

The main difficulty is then to decide whether an allocation of the given sequence to some branch of the mtDNA tree is either sufficiently convincing or somewhat ambiguous or infeasible as some novel haplogroup may be involved which has not yet been described. A fully automated haplogroup allocation could hardly map expert knowledge into simple rules in an adequate way, given the fact that available mtDNA databases, though constantly expanding, just provide a portion of all published mtDNA variation, which anyway constitutes only a micro-extract of the real-world mtDNA variation. Nonetheless, expert knowledge should gradually evolve into guided semi-automated expert systems that will assist the user in exploring complete mtDNA databases more skillfully and evaluating the phylogenetic position of given partial mtDNA sequences more thoroughly.

The forensic geneticist should check if all the diagnostic variants of a targeted haplogroup are present in the sequence (as highlighted in the output of HaploGrep) that is being classified; if some are but others are not, one should either reconsider the haplogroup allocation (in the case that the positive evidence was weak) or go back to the raw data (electropherograms) and check if these variants are actually present or were omitted by mistake or wrongly documented. If more than one diagnostic mutation is absent, even when the electropherograms are apparently correct, one should also suspect a mistake at some step of the analysis. For instance, artificial recombination cannot be detected by just inspecting the original electropherograms, because the cause of the error could be sample mix-up that occurred before the sequencing stage [8]. If two different segments of the profile

(such as HVS-I and HVS-II) are allocated to different parts of the phylogeny, then artificial recombination has most likely occurred [42, 43], and therefore, re-analysis of the sample from the very beginning (DNA extraction) with forward and reverse sequencing is mandatory. Separate and non-overlapping sequencing of HVS-I and HVS-II is prone to sample mixing which results in hybrid profiles, as is again testified by our re-analysis of two forensic studies. Therefore such a sequencing scheme should be avoided in forensic casework [9]. The incidence of artificial recombination could be reduced by analyzing the same sample independently by two analysts, and also by using a posteriori phylogenetic monitoring as explained above.

Haplogrouping in the forensic context thus serves two goals. First, it enables the forensic scientist to extrapolate from the control-region motif to the expected coding-region motif which, if necessary, could be tested site by site for verification. Therefore, it enables wider comparisons with other closely related sequences, for which more coding-region information is available. Second, but perhaps most importantly for the average forensic laboratory, haplogrouping serves as an element of an a posteriori quality control. In particular, it assists in highlighting private mutations, among which phantom mutations would accumulate and thereby could come into focus for re-inspection. Even a single minor error can have important consequences for database searches or for the comparison of mtDNA profiles coming from evidentiary samples and a suspect. Laboratories should not ignore haplogrouping as a tool that can eventually help avoiding wrong decisions in court.

Two recent editorials [59, 60] in the International *Journal of Legal Medicine* and *Forensic Science International Genetics* have set guidelines for future submission of papers with mtDNA population data. These rules are important to guarantee a minimum level of quality. Our reassessment of studies in low-rank forensic journals stresses that the notorious errors, which could easily be avoided, are committed over and over again. Unfortunately, in neither editorial in these journals was the importance of haplogrouping stressed. This may also reflect the fact that only a minority of articles from these (and other forensic) journals have explicitly cited PhyloTree in the past 4 years.

Acknowledgements The Ministerio de Ciencia e Innovación (SAF2008-02971 and SAF2011-26983) and the European project "A European Initial Training Network on the history, archaeology, and new genetics of the Trans-Atlantic slave trade (EUROTAST)" (EU project: 290344) gave support to AS. MvO was supported in part by the Netherlands Forensic Institute (NFI) and by a grant from the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN).

References

1. Parson W, Dür A (2007) EMPOP—a forensic mtDNA database. Forensic Sci Int Genet 1(2):88–92
2. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30 (2):E386–E394
3. Gómez-Carballa A, Ignacio-Veiga A, Álvarez-Iglesias V, Pastoriza-Mourelle A, Ruiz Y, Pineda L, Carracedo Á, Salas A (2012) A melting pot of multicontinental mtDNA lineages in admixed Venezuelans. Am J Phys Anthropol 147(1):78–87
4. 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
5. Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, Comas D, Mitchell RJ, Quintana-Murci L, Tyler-Smith C, Wells RS (2007) The Genographic Project public participation mitochondrial DNA database. PLoS Genet 3(6):e104
6. Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71(5):1150–1160
7. Brandstätter A, Sanger T, Lutz-Bonengel S, Parson W, Beraud-Colomb E, Wen B, Kong Q-P, Bravi CM, Bandelt HJ (2005) Phantom mutation hotspots in human mitochondrial DNA. Electrophoresis 26(18):3414–3429
8. Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt HJ (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335(3):891–899
9. Salas A, Bandelt HJ, Macaulay V, Richards MB (2007) Phylogeographic investigations: the role of trees in forensic genetics. Forensic Sci Int 168(1):1–13
10. Bandelt HJ, Salas A, Taylor RW, Yao YG (2009) The exaggerated status of "novel" and "pathogenic" mtDNA sequence variants due to inadequate database searches. Hum Mutat 30(2):191–196
11. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84(6):740–759
12. Cardoso S, Villanueva-Millan MJ, Valverde L, Odriozola A, Aznar JM, Pineiro-Hermida S, de Pancorbo MM (2012) Mitochondrial DNA control region variation in an autochthonous Basque population sample from the Basque Country. Forensic Sci Int Genet 6(4): e106–e108
13. Farris JS (1970) Methods for computing Wagner Trees. Syst Biol 1 (19):83–92
14. Bandelt H-J, Salas A, Bravi CM (2006) What is a 'novel' mtDNA mutation—and does 'novelty' really matter? J Hum Genet 51 (12):1073–1082
15. Bandelt HJ, Yao YG, Salas A (2008) The search of 'novel' mtDNA mutations in hypertrophic cardiomyopathy: MITOMAPping as a risk factor. Int J Cardiol 126(3):439–442
16. Rakha A, Shin KJ, Yoon JA, Kim NY, Siddique MH, Yang IS, Yang WI, Lee HY (2011) Forensic and genetic characterization of mtDNA from Pathans of Pakistan. Int J Legal Med 125(6):841–848
17. Yao YG, Kong QP, Man XY, Bandelt HJ, Zhang YP (2003) Reconstructing the evolutionary history of China: a caveat about inferences drawn from ancient DNA. Mol Biol Evol 20(2):214–219
18. Umetsu K, Tanaka M, Yuasa I, Adachi N, Miyoshi A, Kashimura S, Park KS, Wei YH, Watanabe G, Osawa M (2005) Multiplex amplified product-length polymorphism analysis of 36 mitochondrial single-nucleotide polymorphisms for haplogrouping of East Asian populations. Electrophoresis 26(1):91–98

19. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 1:44–55
20. Quintáns B, Álvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo Á (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140(2–3):251–257
21. van Oven M, Vermeulen M, Kayser M (2011) Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution. Investig Genet 2(1):6
22. Álvarez-Iglesias V, Barros F, Carracedo Á, Salas A (2008) Mini-sequencing mitochondrial DNA pathogenic mutations. BMC Med Genet 9:26
23. Salas A, Amigo J (2010) A reduced number of mtSNPs saturates mitochondrial DNA haplotype diversity of worldwide population groups. PLoS One 5(5):e10218
24. Brandstätter A, Salas A, Niederstätter H, Gassner C, Carracedo Á, Parson W (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27(13):2541–2550
25. Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, Cuscó I, Lareu MV, García O, Pérez-Jurado L, Carracedo Á, Salas A (2009) New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS One 4(4):e5112
26. Álvarez-Iglesias V, Salas A, Cerezo M, Ramos-Luis E, Jaime JC, Lareu MV, Carracedo Á (2006) Genotyping coding region mtDNA SNPs for Asian and Native American haplogroup assignation. Int Congress Series 11(1288):4–6
27. Crespillo M, Paredes MR, Prieto L, Montesino M, Salas A, Albarrán C, Álvarez-Iglesias V, Amorin A, Berniell-Lee G, Brehm A, Carril JC, Corach D, Cuevas N, Di Lonardo AM, Doutremepuich C, Espinheira RM, Espinoza M, Gómez F, González A, Hernández A, Hidalgo M, Jimenez M, Leite FP, López AM, López-Soto M, Lorente JA, Pagano S, Palacio AM, Pestano JJ, Pinheiro MF, Raimondi E, Ramon MM, Tovar F, Vidal-Rioja L, Vide MC, Whittle MR, Yunis JJ, Garcia-Hirschfel J (2006) Results of the 2003-2004 GEP-ISFG collaborative study on mitochondrial DNA: focus on the mtDNA profile of a mixed semen-saliva stain. Forensic Sci Int 160(2–3):157–167
28. Ballantyne KN, van Oven M, Ralf A, Stoneking M, Mitchell RJ, van Oorschot RA, Kayser M (2012) MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania. Forensic Sci Int Genet 6(4):425–436
29. Cerezo M, Černý V, Carracedo Á, Salas A (2009) Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies. Electrophoresis 30(21):3665–3673
30. Zuccarelli G, Alechine E, Caputo M, Bobillo C, Corach D, Sala A (2011) Rapid screening for Native American mitochondrial and Y-chromosome haplogroups detection in routine DNA analysis. Forensic Sci Int Genet 5(2):105–108
31. Köhnemann S, Pfeiffer H (2011) Application of mtDNA SNP analysis in forensic casework. Forensic Sci Int Genet 5(3):216–221
32. Yang Y, Zhang P, He Q, Zhu Y, Yang X, Lv R, Chen J (2011) A new strategy for the discrimination of mitochondrial DNA haplogroups in Han population. J Forensic Sci 56(3):586–590
33. Bandelt HJ, Parson W (2004) Fehlerquellen mitochondrialer DNS-Datensätze und Evaluation der mtDNS-Datenbank D-Loop-BASE. Rechtsmedizin 14:251–257
34. Zgonjanin D, Veselinović I, Kubat M, Furač I, Antov M, Lončar E, Tasić M, Vuković R, Omorjan R (2010) Sequence polymorphism of the mitochondrial DNA control region in the population of Vojvodina Province, Serbia. Leg Med (Tokyo) 12(2):104–107
35. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100(1):171–176
36. Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24(3):757–768
37. Yao Y-G, Salas A, Logan I, Bandelt H-J (2009) mtDNA data mining in GenBank needs surveying. Am J Hum Genet 85 (6):929–933; author reply 933
38. Chandrasekar A, Kumar S, Sreenath J, Sarkar BN, Urade BP, Mallick S, Bandopadhyay SS, Barua P, Barik SS, Basu D, Kiran U, Gangopadhyay P, Sahani R, Prasad BVR, Gangopadhyay S, Lakshmi GR, Ravuri RR, Padmaja K, Venugopal PN, Sharma MB, Rao VR (2009) Updating phylogeny of mitochondrial DNA macrohaplogroup M in India: dispersal of modern human in South Asian corridor. PLoS One 4(10):e7447
39. Sekiguchi K, Imaizumi K, Fujii K, Mizuno N, Ogawa Y, Akutsu T, Nakahara H, Kitayama T, Kasai K (2008) Mitochondrial DNA population data of HV1 and HV2 sequences from Japanese individuals. Leg Med (Tokyo) 10(5):284–286
40. Parson W, Bandelt HJ (2007) Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet 1:13–19
41. Bandelt HJ, Kong QP, Parson W, Salas A (2005) More evidence for non-maternal inheritance of mitochondrial DNA? J Med Genet 42:957–960
42. Bandelt HJ, Salas A, Bravi CM (2004) Problems in FBI mtDNA database. Science 305(5689):1402–1404
43. Bandelt HJ, Salas A, Lutz-Bonengel S (2004) Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118(5):267–273
44. Afonso-Costa H, Carvalho M, Lopes V, Balsa F, Bento AM, Serra A, Andrade L, Anjos MJ, Vide MC, Pantoja S, Vieira DN, Corte-Real F (2010) Mitochondrial DNA sequence analysis of a native Bolivian population. J Forensic Leg Med 17(5):247–253
45. Karachanak S, Carossa V, Nesheva D, Olivieri A, Pala M, Hooshiar Kashani B, Grugni V, Battaglia V, Achilli A, Yordanov Y, Galabov AS, Semino O, Toncheva D, Torroni A (2012) Bulgarians vs the other European populations: a mitochondrial DNA perspective. Int J Legal Med 126(4):497–503
46. Mikkelsen M, Rockenbauer E, Sørensen E, Rasmussen M, Børsting C, Morling N (2008) A mitochondrial DNA SNP multiplex assigning Caucasians into 36 haplo- and subhaplogroups. Forensic Sci Int: Genet Supplement Series 1 (1):287–289
47. Mikkelsen M, Rockenbauer E, Demir H, Børsting C, Morling N (2011) Frequencies of 33 coding region mitochondrial SNPs in a Danish and a Turkish population. Forensic Sci Int Genet 5(5):559–560
48. Mikkelsen M, Sørensen E, Rasmussen EM, Morling N (2010) Mitochondrial DNA HV1 and HV2 variation in Danes. Forensic Sci Int Genet 4(4):e87–e88
49. Zimmermann B, Bodner M, Amory S, Fendt L, Röck A, Horst D, Horst B, Sanguansermsri T, Parson W, Brandstätter A (2009) Forensic and phylogeographic characterization of mtDNA lineages from northern Thailand (Chiang Mai). Int J Legal Med 123 (6):495–501
50. Lee HY, Song I, Ha E, Cho SB, Yang WI, Shin KJ (2008) mtDNAmanager: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. BMC Bioinforma 9:483
51. 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:no 2
52. Fan L, Yao YG (2010) MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA sequence variations. Mitochondrion 11(2):351–356
53. Kloss-Brandstätter A, Pacher D, Schonherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2011) HaploGrep: a fast and

reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32(1):25–32
54. Kong QP, Bandelt HJ, Sun C, Yao YG, Salas A, Achilli A, Wang CY, Zhong L, Zhu CL, Wu SF, Torroni A, Zhang YP (2006) Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15(13):2076–2086
55. Scheible M, Alenizi M, Sturk-Andreaggi K, Coble MD, Ismael S, Irwin JA (2011) Mitochondrial DNA control region variation in a Kuwaiti population sample. Forensic Sci Int Genet 5(4):e112–e113
56. van Oven M (2010) Revision of the mtDNA tree and corresponding haplogroup nomenclature. Proc Natl Acad Sci U S A 107(11): E38–E39, author reply e40-31
57. Zhao M, Kong QP, Wang HW, Peng MS, Xie XD, Wang WZ, Jiayang DJG, Cai MC, Zhao SN, Cidanpingcuo TYQ, Wu SF, Yao YG, Bandelt HJ, Zhang YP (2009) Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. Proc Natl Acad Sci U S A 106(50):21230–21235
58. Kong QP, Salas A, Sun C, Fuku N, Tanaka M, Zhong L, Wang CY, Yao YG, Bandelt HJ (2008) Distilling artificial recombinants from large sets of complete mtDNA genomes. PLoS One 3(8):e3016
59. Parson W, Roewer L (2010) Publication of population data of linearly inherited DNA markers in the International Journal of Legal Medicine. Int J Legal Med 124(5):505–509
60. Carracedo Á, Butler JM, Gusmão L, Parson W, Roewer L, Schneider PM (2010) Publication of population data for forensic purposes. Forensic Sci Int Genet 4(3):145–147
61. Nur Haslindawaty AR, Panneerchelvam S, Edinur HA, Norazmi MN, Zafarina Z (2010) Sequence polymorphisms of mtDNA HV1, HV2, and HV3 regions in the Malay population of Peninsular Malaysia. Int J Legal Med 124(5):415–426
62. Irwin JA, Saunier JL, Beh P, Strouss KM, Paintner CD, Parsons TJ (2009) Mitochondrial DNA control region variation in a population sample from Hong Kong, China. Forensic Sci Int Genet 3 (4):e119–e125
63. Malyarchuk B, Derenko M, Grzybowski T, Perkova M, Rogalla U, Vanecek T, Tsybovsky I (2010) The peopling of Europe from the mitochondrial haplogroup U5 perspective. PLoS One 5(4):e10285
64. Tillmar AO, Coble MD, Wallerstrom T, Holmlund G (2010) Homogeneity in mitochondrial DNA control region sequences in Swedish subpopulations. Int J Legal Med 124(2):91–98