Mitochondrion 13 (2013) 360–363

Contents lists available at SciVerse ScienceDirect

Mitochondrion

journal homepage: www.elsevier.com/locate/mito

Short communication

An update to MitoTool: Using a new scoring system for faster mtDNA haplogroup determination

Long Fan [a,b], Yong-Gang Yao [a,*]

[a] Key Laboratory of Animal Models and Human Disease Mechanisms of the Chinese Academy of Sciences & Yunnan Province, Kunming Institute of Zoology, Kunming, Yunnan 650223, China
[b] School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong Special Administrative Region, China

ARTICLE INFO

Article history:
Received 24 September 2012
Received in revised form 5 March 2013
Accepted 19 April 2013
Available online 28 April 2013

Keywords:
mtDNA
Haplogroup
Determination
Scoring system

ABSTRACT

The determination of human mitochondrial DNA (mtDNA) haplogroups is not only crucial in anthropological and forensic studies, but is also helpful in the medical field to prevent establishment of wrong disease associations. In recent years, high-throughput technologies and the huge amounts of data they create, as well as the regular updates to the mtDNA phylogenetic tree, mean that there is a need for an automated approach which can make a speedier determination of haplogroups than can be made by using the traditional manual method. Here, we update the MitoTool (www.mitotool.org) by incorporating a novel scoring system for the determination of mtDNA into haplogroups, which has advantages on speed, accuracy and ease of implementation. In order to make the access to MitoTool easier, we also provide a stand-alone version of the program that will run on a local computer and this version is freely available at the MitoTool website.

1. Introduction

The prominent properties of mtDNA such as maternal inheritance, absence of recombination and the high mutation rate make this molecule widely used in population genetics, forensics and medical genetics. Meanwhile, the assignment of mtDNA into haplogroups has become a routine analysis (even a critical prerequisite in the forensic field), which helps researchers to (i) conduct an *a posteriori* quality control of data, as this recommended analysis is beneficial for detecting five major types of errors in mtDNA data, including base shifts, reference bias, phantom mutations, base misscoring and artefactual recombination (Bandelt et al., 2001); and (ii) avoid potential pitfalls in mtDNA disease studies (Yao et al., 2006, 2009). However, traditional manual haplogroup determination is a daunting task, especially when MitoChip (Maitra et al., 2004) and high throughput next-generation sequencing technologies (Gunnarsdottir et al., 2011) are employed, as these produce a huge amount of mtDNA data. Furthermore, with the aim to provide the most up-to-date version of the mtDNA classification system and avoid some possible conflicts, the mtDNA tree at www.phylotree.org (van Oven and Kayser, 2009) is revised regularly. For the beginner who is not familiar with manual haplogroup determination, the continually expanding mtDNA tree and size of datasets make haplogroup determination ever more challenging.

Hitherto several tools (e.g. mtDNAmanager (Lee et al., 2008) and HaploGrep (Kloss-Brandstatter et al., 2011)) and algorithms (e.g. random forest [RF] and support vector machines [SVM] (Wong et al., 2011)) were developed to solve the haplogrouping problem. As a complementary tool to traditional manual haplogrouping, automated haplogrouping has advantages on speed and is easy to manage, but it does not compete with manual identification due to the following reasons. First, a subsequent manual check is indispensable for ensuring the accuracy, especially when automatic haplogrouping is confronted with artefactual recombinants (Bandelt et al., 2012) and/or finds multiple haplogroup assignments. Second, automated haplogrouping will generate information but it does not have insight, thus the identification of any new haplogroups which do not exist in the existing mtDNA tree requires manual interpretation. In our previous study, we established a web-based platform (MitoTool: http://www.mitotool.org) for the automated determination of mtDNA haplogroups (Fan and Yao, 2011). Here we present an update for this platform, including a novel scoring system for haplogrouping and a stand-alone version of MitoTool. Compared with the existing methods, the updated MitoTool is fast and accurate, and its scoring system is easy to implement in other software.

2. Approach

2.1. Classification standard

Phylotree (van Oven and Kayser, 2009) is well annotated and is now the *de facto* standard mtDNA tree, so we follow Phylotree to name and classify haplogroups. Since mtDNA sequences belonging to the same haplogroup share the same combination of a group of (ancestral) variants, we extracted the variant list of each haplogroup by traversing the tree structure of Phylotree recursively and we stored the entire Phylotree as a text file. During this tree transformation, deletions of one or more bases, insertions of one or more bases and point variants were treated equally as the generalized variants,

* Corresponding author. Tel./fax: +86 871 5180085.
E-mail addresses: mitotool@gmail.com (L. Fan), ygyaozh@gmail.com (Y.-G. Yao).

1567-7249/$ – see front matter © 2013 Elsevier B.V. and Mitochondria Research Society. All rights reserved.
http://dx.doi.org/10.1016/j.mito.2013.04.011

e.g. the deletion 523–524d will be recorded as one generalized variant, and not two single deletions (523d and 524d). For a back mutation, we record only its current status in a certain haplogroup. For instance, position 263 of haplogroup L0 mutates toward a base identical-by-state to the revised Cambridge Reference Sequence (rCRS) (Andrews et al., 1999) in the rCRS-oriented version of mtDNA tree Build 14, so the variant list of L0 does not contain the A263G variant. In contrast, position 263 in subhaplogroup L0a′b′f of L0 mutates toward a different base relative to the rCRS, for this reason the variant list of L0a′b′f does contain the A263G variant.

2.2. Scoring system

When we perform any kind of classification of real life, we need to assess not only the similarity among objects but also their differences. Inspired by this simple and rational idea, we have designed a novel scoring system for mtDNA determination.

Let symbol M of the formula represent the similarity between a queried mtDNA sequence and a tested haplogroup, while $N - M$ implies the insufficiency that prevents the query sample being assigned to the tested haplogroup. Subtracting the insufficiency $N - M$ from the sufficiency M gives the equation: $S = M - (N - M) = 2M - N$. Therefore, for each tested haplogroup, only a similarity parameter S needs to be calculated according to $S = 2M - N$, where M is the number of variants shared by the query sample and the currently tested haplogroup, and N is the total number of variants expected in currently tested haplogroup. Then, all of the tested haplogroups can be sorted according to the descending order of their S. The haplogroup with the highest S value can then be taken as the best result for this mtDNA sequence.

As a demonstration, we give an example here. Suppose the query sample is a mtDNA control region sequence which contains variants 263, 16354 and 16519, in terms of mtDNA tree Build 14 (5 Apr 2012) at www.phylotree.org, H2a1 is expected to have 263 and 16354, and H2a1e is expected to have 263, 575 and 16354 in the corresponding control region sequence. According to the core formula, S_{H2a1} is equal to $2 \times M_{H2a1} - N_{H2a1}$, namely $2 \times 2 - 2 = 2$, while S_{H2a1e} is equal to $2 \times M_{H2a1e} - N_{H2a1e}$, namely $2 \times 2 - 3 = 1$. Thus, classifying the queried mtDNA to H2a1 is better than to H2a1e.

In practical usage, if several haplogroups with a far pairwise phylogenetic distance have the same highest S and all of them become the best hits, this situation means that (i) the queried mtDNA needs quality checking, or (ii) the information is not sufficient for haplogrouping, such as when using partial mtDNA sequence. In these cases of ambiguous assignment of haplogroups, it is necessary that the user performs a manual inspection. In addition, if the number of variants of one query is larger than the number of variants expected by its best hit, this situation indicates that (i) many private variants exist in this query; or (ii) the data quality of the query is too low, and the query has a high possibility of having an artefactual recombination. For instance, the whole mtDNA sequence DQ418488 (GenBank accession number) has 28 useful variants except for 309 + CC and 315 + C. In terms of our scoring system, its best automated determination relative to mtDNA tree Build 14 is the haplogroup M which only expects 22 variants, and there is a big gap (6 variants) between the observed and expected variants. Indeed, according to our previous manual check (Yao et al., 2009), sequence DQ418488 has a problem with artefactual recombination. Consequently, when similar cases occur, manual checking of the original data is indispensable. Evidently, this scoring system can be used for both determining haplogroups and reminding users of data quality problems.

3. Speed and accuracy benchmark

At present, the state-of-the-art automated haplogrouping algorithms and tools (Table 1) are at different stages of development:

(i) Machine learning methods for mtDNA haplogrouping (e.g. RF and SVM) are still not fully realized in any straightforwardly accessible online tools and their accuracies are too low, e.g. for the haplogrouping of 60 samples with N* status, the accuracy rate is approximately 10% (Wong et al., 2011);
(ii) mtDNAoffice (Soares et al., 2012) is a stand-alone software for determining macro-haplogroups through clustering. It is restricted to protein coding region and is not suitable for haplogrouping at a high resolution (e.g. subhaplogroups);
(iii) mtDNAmanager (Lee et al., 2008) stopped its update more than a year ago and only allows the haplogrouping based on mtDNA control region;
(iv) some mtDNA databases (e.g. HmtDB (Rubino et al., 2012) and MitoVariome (Lee et al., 2009)) also provide limited function for haplogrouping and both can only list the match percentages between different haplogroups and queries;
(v) HaploGrep (Kloss-Brandstatter et al., 2011) is well maintained, and it was reported that HaploGrep is the most accurate algorithm among the current automated haplogrouping tools (Bandelt et al., 2012).

In the following discussion, we compare MitoTool with its new scoring system with HaploGrep's to show its performance:

(i) In the preprocessing stage, our scoring system needs no determination of the phylogenetic weights for each variant, therefore it has no bias against sequences phylogenetically close to the rCRS. The scoring system of HaploGrep involves the weighting of variants according to their occurrence frequencies in mtDNA tree, which reflects a rationale that was used during manual haplogrouping. However, this weighting induces a bias, which means that the score drop from the perfect haplogroup allocation to the slightly imperfect allocation is quite drastic for mtDNA sequence phylogenetically close to the rCRS, but this kind of score drop is very minor for mtDNA sequence phylogenetically distant from the rCRS (Bandelt et al., 2012). This phenomenon means that the omission and inclusion of a private variant with heavy weighting may lead to various effects on different haplogroups. For instance, variant 16303 occurs once in haplogroup O1 and has a relatively heavier weight, the inclusion of 16303 as a private variant has a considerable effect to the scores of HV subhaplogroups in contrast to the scores of L0 and L1 subhaplogroups (Bandelt et al., 2012). On the contrary, the presence of 16303 has equal weight as other variant and has an identical effect on diverse haplogroups in our scoring system.
(ii) Our scoring system can perform haplogroup determination for each mtDNA no matter whether rCRS (Andrews et al., 1999) or Reconstructed Sapiens Reference Sequence (RSRS) (Behar et al., 2012) is used as the reference sequence to export the sequence variations in the lineage. Currently, Phylotree offers two versions of mtDNA tree (the rCRS-oriented version and the RSRS-oriented version) encoded with respect to different reference sequences. Accordingly, a researcher may choose the rCRS as the reference sequence to score the genetic variants in certain mtDNA sequences, and then use the corresponding mtDNA tree that refers to the rCRS-oriented version, or vice versa. For instance, the whole mtDNA sequence J01415.2 (GenBank accession number) has no variant compared to the rCRS, while it contains 50 variants compared to the RSRS. With the selection of the correct version of mtDNA tree, our scoring system can identify precisely the haplogroup status of J01415.2 under both conditions. Hence, by using the correct version of the mtDNA tree, a researcher who uses our scoring system can employ the rCRS or the RSRS with the aim to record the list of variants in mtDNA sequences. The recently updated MitoTool using the new scoring system supplies this choice for the users as well. Note that the inclusion of RSRS as the reference at the MitoTool does not necessarily mean that

Table 1
Summary of methods available for automated mtDNA haplogrouping.

Method	Preprocessing	Application scope	Optional reference sequence	Online	Latest update	Automated haplogrouping	Reference
Machine learning	Data training	Tested in control region	rCRS	No	–	Yes	Wong et al. (2011)
mtDNAoffice	–	Protein coding region	–	No	–	Yes	Soares et al. (2012)
mtDNAmanager	Needs backend database	Control region	rCRS	Yes	October 2, 2011	Yes	Lee et al. (2008)
MitoVariome	–	Whole mtDNA	rCRS	Yes	July 8, 2009	Limited query	Lee et al. (2009)
HmtDB	–	Complete range	rCRS	Yes	June 19, 2011	Limited query	Rubino et al. (2012)
HaploGrep	Weight calculation	Complete range	rCRS	Yes	October 4, 2012	Yes	Kloss-Brandstatter et al. (2011)
Updated MitoTool	–	Complete range	rCRS and RSRS	Yes	October 27, 2012	Yes	Fan and Yao (2011)

we have agreed with the proposal of the replacement of the rCRS by the RSRS (Behar et al., 2012). We believe that the proposed switch to RSRS will inevitably lead to notational chaos, mistakes and misinterpretations in the field. But we have included the RSRS-oriented version of mtDNA tree in the MitoTool just to meet the needs of those users who would like to score mtDNA variation relative to the RSRS.

(iii) During the main calculation stage, our scoring system has a faster speed and occupies less computer memory as it utilizes a simpler core formula. In addition to the scoring system itself, computational equipment, programming language, compilation optimization and parallel computation can also influence executive time. Furthermore, online testing will be usually affected by network speed. We do not directly perform speed comparison, but utilize a theoretical analysis to illustrate the difference. Here, we list the scoring equation of HaploGrep below:

$$S_{HaploGrep} = \frac{1}{2} \times \left(\frac{\sum_{i=1^{w_i}}^{M}}{\sum_{i=1^{w_i}}^{N}} + \frac{\sum_{i=1^{w_i}}^{M}}{\sum_{i=1^{w_i}}^{Q}} \right) \quad (1)$$

where w_i is the phylogenetic weight for the ith variants, M is the number of variants shared by the query sample and currently tested haplogroup, N is the total number of variants expected in currently tested haplogroup, and Q is the total number of variants of query sample. From this formula, we can see that HaploGrep requires more calculations for each sample, and then these retrievals of preprocessed weights of the variants and the summing operations cost more time and memory, although the space and computation complexity of the two scoring systems increase linearly with the size of imported samples, in another word, their big-O notations are both $O(n)$.

(iv) Our scoring system has slightly better performance when tested for real data and artefactual mtDNA recombinants. Considering the cases mentioned in the study of Bandelt et al. (2012) as examples, the updated MitoTool using the new scoring system and HaploGrep had the same high accuracy to classify all 26 real samples (Supplementary Table 1). However, in the 7 cases of artefactual recombination (Supplementary Table 2), the updated MitoTool detected the major components of artefactual recombination of 5 cases, whereas HaploGrep only successfully identified the major components of 3 cases. For instance, the artefactual mtDNA with sample ID 739 is generated by mixing two haplogroups B4a and B4b1a1, HaploGrep classified this sample into haplogroup R31, while updated MitoTool at least detected the major component B4a. Besides, for three artefactual mtDNAs (sample ID: USA.AFR.000942, FRA.CAU.000084 and LPAZ094), the updated MitoTool directly gave the uncertain determinations by listing the haplogroups with far phylogenetic distance, e.g. L1b3 and M31a1 are displayed for USA.AFR.000942. These ambiguous results should prompt users to check any results where there is a doubt over the quality.

(v) Finally, our scoring system can be easily implemented in other software. As each variant is treated equally in this system, no matter which type the variant is and the frequency with which the variant appears in the whole mtDNA tree, our scoring system does not need to (i) calculate the "phylogenetic weights" like HaploGrep (Kloss-Brandstatter et al., 2011), or (ii) perform data training like RF and SVM (Wong et al., 2011), or (iii) construct a backend database like mtDNAmanager (Lee et al., 2008).

4. Stand-alone version of MitoTool

The response time of a web-based software for analyzing each query is usually limited by the status of network and the configuration of the server. Also, some users take the security of their own original data seriously. For these reasons, we have developed a stand-alone version of MitoTool which provides a user-friendly interface and the major functions of the web-based version except for some information retrieval, and can be run on a local computer using the Windows, Mac or Linux operating systems. This stand-alone version is implemented using C++, in which Qt library (http://qt-project.org/), Boost Math Toolkit (http://www.boost.org) and SeqAn library (Doring et al., 2008) are used for the construction of user interface, the statistical tests of haplogroup distribution and sequence alignment, respectively. Compared to the web-based version, it does not ask for online data transmission between local computer and the server, so it has advantage on speed and security, which is helpful for improving access of MitoTool.

5. Conclusion

The updated MitoTool (http://www.mitotool.org) with the embedding of the novel scoring system is accurate and fast, and where the mtDNA tree will be updated synchronously with Phylotree. We also provide a stand-alone version of the MitoTool for those who would like to perform haplogroup determination on a local computer instead of via online web server. We hope that more and more users will employ this platform.

A limitation on the available tools for the automated mtDNA haplogroup determination is that most of them rely on PhyloTree which limits the identification of any new haplogroup(s). Similarly, MitoTool is implemented using the same starting resource and thus has not been able to address this issue. Therefore, the performance of our updated MitoTool depends on the completeness and incompleteness of the queried mtDNA genome sequence and the starting mtDNA tree resource. For those mtDNAs without haplogroup characteristic variants in (partial) control-region sequences, it has the ability to assign those mtDNAs into respective subhaplogroups given the presence of a subhaplogroup motif in the queried sequence and the presence of this subhaplogroup in the Phylotree. It is unlikely that automated mtDNA haplogroup determination can completely replace the traditional manual haplogroup determination, unless we obtain a complete mtDNA tree of the world population. A fast (and maybe rough) haplogroup determination and error detection of bulk of mtDNA sequences, facilitated with manual checking continues to be the best way to handle mtDNA data.

Acknowledgments

We thank Ian Logan for helpful comments and language editing. This work was supported by the National Science Foundation of China (30925021), the Top Talents Program of Yunnan Province (2009CI119) and the Chinese Academy of Sciences.

Conflicts of interest: none declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mito.2013.04.011.

References

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., Howell, N., 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 23, 147.

Bandelt, H.J., Lahermo, P., Richards, M., Macaulay, V., 2001. Detecting errors in mtDNA data by phylogenetic analysis. Int. J. Legal Med. 115, 64–69.

Bandelt, H.J., van Oven, M., Salas, A., 2012. Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. Int. J. Legal Med. 126, 901–916.

Behar, D.M., van Oven, M., Rosset, S., Metspalu, M., Loogvali, E.L., Silva, N.M., Kivisild, T., Torroni, A., Villems, R., 2012. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am. J. Hum. Genet. 90, 675–684.

Doring, A., Weese, D., Rausch, T., Reinert, K., 2008. SeqAn an efficient, generic C++ library for sequence analysis. BMC Bioinformatics 9, 11.

Fan, L., Yao, Y.G., 2011. MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA sequence variations. Mitochondrion 11, 351–356.

Gunnarsdottir, E.D., Li, M., Bauchet, M., Finstermeier, K., Stoneking, M., 2011. High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res. 21, 1–11.

Kloss-Brandstatter, 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, 25–32.

Lee, H.Y., Song, I., Ha, E., Cho, S.B., Yang, W.I., Shin, K.J., 2008. mtDNAmanager: a web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. BMC Bioinforma. 9, 483.

Lee, Y.S., Kim, W.Y., Ji, M., Kim, J.H., Bhak, J., 2009. MitoVariome: a variome database of human mitochondrial DNA. BMC Genomics 10 (Suppl. 3), S12.

Maitra, A., Cohen, Y., Gillespie, S.E., Mambo, E., Fukushima, N., Hoque, M.O., Shah, N., Goggins, M., Califano, J., Sidransky, D., Chakravarti, A., 2004. The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. Genome Res. 14, 812–819.

Rubino, F., Piredda, R., Calabrese, F.M., Simone, D., Lang, M., Calabrese, C., Petruzzella, V., Tommaseo-Ponzetta, M., Gasparre, G., Attimonelli, M., 2012. HmtDB, a genomic resource for mitochondrion-based human variability studies. Nucleic Acids Res. 40, D1150–D1159.

Soares, I., Amorim, A., Goios, A., 2012. mtDNAoffice: a software to assign human mtDNA macro haplogroups through automated analysis of the protein coding region. Mitochondrion 12, 666–668.

van Oven, M., Kayser, M., 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 30, E386–E394.

Wong, C., Li, Y., Lee, C., Huang, C.H., 2011. Ensemble learning algorithms for classification of mtDNA into haplogroups. Brief Bioinform. 12, 1–9.

Yao, Y.G., Salas, A., Bravi, C.M., Bandelt, H.J., 2006. A reappraisal of complete mtDNA variation in East Asian families with hearing impairment. Hum. Genet. 119, 505–515.

Yao, Y.G., Salas, A., Logan, I., Bandelt, H.J., 2009. mtDNA data mining in GenBank needs surveying. Am. J. Hum. Genet. 85, 929–933.

Supplementary Table 1. Real data of (partial) mtDNA control-region sequences

ID	Haplotype [a]	HaploGrep [b]	Updated MitoTool [c]	Likely Haplogroup [d]	Related sample from GenBank	Reference
1	16086, 16129, 16209, 16223, 16272, 73, 152, 225, 249d, 263, 315+C, 316, 489, 523-524d	M20	M20	M20	HM030505	(Nur Haslindawaty et al., 2010)
2	16172, 16183C, 16189, 16209, 16223, 16258T, 16311, 16362, 73, 185A, 189, 195, 234, 263, 309+C, 315+C, 523-524d	N10a	N10a	N10a	HM030542	(Irwin et al., 2009)
3	16069, 16172, 16223, 16278, 16291A, 16298, 16362, 73, 150, 152, 199, 263, 309+CC, 315+C	N10b	N10b	N10b	HM030500	(Irwin et al., 2009)
4	16114A, 16126, 16218, 16223, 16275, 16291, 16356, 16390, 16391, 73, 263, 309+C, 315+C	M52a	M52a	M52a	EF093557	(Mikkelsen et al., 2010)
5	16223, 16291, 16362, 16390, 73, 263, 309+C, 315+C	E1a1a	E1a1a	E1a1a	EF093544	(Mikkelsen et al., 2010)
6	16189, 16319, 16325, 73, 150, 152, 263, 315+C	U5b2a1b	U5b2a1b	U5b2a1b	GU296545	(Mikkelsen et al., 2010)
7	16051, 16162, 16213, 16266, 73, 146, 263, 315+C	H1a3c	H1a3c	H1a3c	EU979418	(Mikkelsen et al., 2010)
8	16069, 16126, 16145, 16231, 16261, 73, 150, 152, 195, 215, 263, 295, 310+T, 315+C, 319, 489, 513	J2a1a1a	J2a1a1a	J2a1a1a	GU903270	Family Tree DNA
9	16086, 16222, 16224, 16270, 16311, 16519, 73, 146, 263, 315+C	K2b1a1	K2b1a1	K2b1a1	EU770310	Family Tree DNA
10	16086, 16239, 16311, 16320, 73, 150, 263, 315+C	U5b2a1a2	U5b2a1a1; U5b2a1a2; U3b1a	U5b2a1a1	GU296544	(Malyarchuk et al., 2010)
11	16224, 16519, 73, 152, 204, 263, 315+C, 497, 524+AC	K1a4a1e	K1a4a1e	K1a4a1e	EU597496	(Tillmar et al., 2010)
12	16069, 16261, 73, 185, 189, 263, 295, 315+C, 462, 489	J1c+16261+189	J1c12	J1c6	AY495209	(Tillmar et al., 2010)
13	16224, 16519, 73, 152, 204, 263, 272, 315+C, 497, 524+AC	K1a4a1e	K1a4a1e	K1a4a1e	EU597496	(Tillmar et al., 2010)

14	16183C, 16189, 16193+C, 73, 262, 263, 285, 309+CC, 315+C, 323, 385, 523-524d	U1a1	U1a1	U1a1	AY882396	(Tillmar et al., 2010)
15	16069, 16145, 16207, 16222, 16231, 16261, 73, 150, 152, 195, 215, 246, 263, 295, 309+CC, 315+C, 319, 489, 513	J2a1a1	J2a1a1	J2a1a	FJ348157	(Tillmar et al., 2010)
16	16183C, 16189, 16193+C, 73, 262, 263, 285, 309+C, 315+C, 323, 385, 523-524d	U1a1	U1a1	U1a1	AY882396	(Tillmar et al., 2010)
17	16153, 16298Y, 72, 73, 93, 95C, 263, 309+C, 315+C	V7a	V7a	V7a	AF347006	(Tillmar et al., 2010)
18	16153, 16298, 72, 73, 93, 95C, 263, 309+C, 315+C	V7a	V7a	V7a	AF347006	(Tillmar et al., 2010)
19	16183C, 16189, 16193+C, 16217, 16519, 73, 263, 309+CCC, 315+C, 498d, 499	B4b; B2	B4b; B2	B4b; B2	EU095550	(Tillmar et al., 2010)
20	16223, 16311, 16362, 16519, 73, 263, 315+C, 489, 573+CCC	M74; D4j11	M74; D4j11	M74; D4j11	FJ770954	(Tillmar et al., 2010)
21	16093, 16172, 16223, 16297, 16311, 16362, 16400, 16519, 73, 146, 185, 263, 315+C, 489	D4j1b2	M7e	M74	HM030520	(Tillmar et al., 2010)
22	16153, 16298, 72, 93, 95C, 263, 309+C, 315+C, 523-524d	V7a	V7a	V7a	AF347006	(Tillmar et al., 2010)
23	16298, 16519, 263, 315+C	V+@72	R0; HV; HV0; V; V21; HV0f; H	HV; V	AY495306	(Tillmar et al., 2010)
24	16067, 16311, 152, 195, 263, 315+C	HV1b3	HV1b3	HV1b3	HQ165756	(Tillmar et al., 2010)
25	16129, 16185, 16223, 16224, 16260, 16298, 16519, 73, 151, 152, 249d, 263, 315+C, 489	Z1a	Z1a	Z1a	AY339515	(Tillmar et al., 2010)
26	16069, 16126, 16214, 16311, 16362, 73, 150, 195, 235, 263, 295, 309+C, 315+C, 326, 489	J2a2a	J2a2a	J2a2a	EF660967	(Tillmar et al., 2010)

Note: The dataset was adapted from Table 2 of Bandelt et al. (2012) and mtDNA tree Build 14 at Phylotree.org (van Oven and Kayser, 2009) was used

[a] Variants are recorded with respect to the rCRS (Andrews et al., 1999)

[b] Haplogroup classification provided by HaploGrep (Kloss-Brandstatter et al., 2011)

[c] Haplogroup classification provided by updated MitoTool (Fan and Yao, 2011) integrated with new score system

[d] Manual classification according to mtDNA tree Build 14 at Phylotree.org

Supplementary Table 2. Artificial mtDNA recombinants from forensic databases

Sample ID (source)	**HVS-I**	**HVS-II**	**Real haplogroup status** [a]	**HaploGrep** [b]	**Updated MitoTool** [c]
USA.AFR.000942 (Monson et al., 2002)	16126 16187 16189 16223 16264 16270 16278 16293 16311 16519	73 249d 263 290d 291d 309+C 315+C 489	L1b × C1	M31a1	L1b2'3; L1b3; M31a1 *
FRA.CAU.000084 (Monson et al., 2002)	16298	73 185 188 228 263 295 315+C	HV0 × J1c	J1c2	R; H3d; H4a1a1a; H7i; J1c2; P; U
VP61 (Zgonjanin et al., 2010)	16126 16270 16294 16304	73 242 263 295 309+C 315+C	T2b × J1b1a	T2b+@16296	T2b4; T2b21
739 (Sekiguchi et al., 2008)	16189 16190 16193+CC 16261 16362	73 146 199 202 207 263 309+C 315+C	B4a × B4b1a1	R31	B4a1a; B4a1c4 *
92 (Sekiguchi et al., 2008)	16136 16183C 16189 16217 16284	73 103 263 315+C	B4b1a1 × B5b	B4	B4; B4b1
LPAZ092 (Afonso Costa et al., 2010)	16181 16189 16217	73 185 249d 263 290-291d 309+C 315+C	B2 × C1	B4	B4
LPAZ094 (Afonso Costa et al., 2010)	16182C 16183C 16189 16223 16298 16325 16327 16344	73 263 309+CC 315+C	C1 × B2	N9b	C4c1b; C7a2; N9b

Note: The dataset was adapted from Table 3 of Bandelt et al. (2012) and mtDNA tree Build 14 at Phylotree.org (van Oven and Kayser, 2009) was used. Variants are recorded with respect to the rCRS (Andrews et al., 1999).

[a] Artificial mtDNA recombinants proposed by Bandelt et al. (2012)

[b] Haplogroup classification provided by HaploGrep (Kloss-Brandstatter et al., 2011)

[c] Haplogroup classification provided by the updated MitoTool (Fan and Yao, 2011) integrated with the new score system

* The updated MitoTool could detect major components of the artificial recombination, whereas HaploGrep failed to do that.

Supplementary Reference

Afonso Costa, H., Carvalho, M., Lopes, V., Balsa, F., Bento, A.M., Serra, A., Andrade, L., Anjos, M.J., Vide, M.C., Pantoja, S., Vieira, D.N., Corte-Real, F., 2010. Mitochondrial DNA sequence analysis of a native Bolivian population. J Forensic Leg Med 17, 247-253.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., Howell, N., 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23, 147.

Bandelt, H.J., van Oven, M., Salas, A., 2012. Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. Int J Legal Med 126, 901–916.

Fan, L., Yao, Y.G., 2011. MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA sequence variations. Mitochondrion 11, 351-356.

Irwin, J.A., Saunier, J.L., Beh, P., Strouss, K.M., Paintner, C.D., Parsons, T.J., 2009. Mitochondrial DNA control region variation in a population sample from Hong Kong, China. Forensic Sci Int Genet 3, e119-125.

Kloss-Brandstatter, 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, 25-32.

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, e10285.

Mikkelsen, M., Sorensen, E., Rasmussen, E.M., Morling, N., 2010. Mitochondrial DNA HV1 and HV2 variation in Danes. Forensic Sci Int Genet 4, e87-88.

Monson, K.L., Miller, K.W.P., Wilson, M.R., DiZinno, J.A., Budowle, B., 2002. The mtDNA Population Database: an integrated software and database resource for forensic comparison. Forensic Sci Commun 4:no 2.

Nur Haslindawaty, A.R., Panneerchelvam, S., Edinur, H.A., Norazmi, M.N., Zafarina, Z., 2010. Sequence polymorphisms of mtDNA HV1, HV2, and HV3 regions in the Malay population of Peninsular Malaysia. Int J Legal Med 124, 415-426.

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, 284-286.

Tillmar, A.O., Coble, M.D., Wallerstrom, T., Holmlund, G., 2010. Homogeneity in mitochondrial DNA control region sequences in Swedish subpopulations. Int J Legal Med 124, 91-98.

van Oven, M., Kayser, M., 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30, E386-394.

Zgonjanin, D., Veselinovic, I., Kubat, M., Furac, I., Antov, M., Loncar, E., Tasic, M., Vukovic, R., Omorjan, R., 2010. Sequence polymorphism of the mitochondrial DNA control region in the population of Vojvodina Province, Serbia. Leg Med (Tokyo) 12, 104-107.