Journal of Human Genetics (2014) 59, 66–77

www.nature.com/jhg

REVIEW

The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies

Hans-Jürgen Bandelt[1], Anita Kloss-Brandstätter[2], Martin B Richards[3], Yong-Gang Yao[4] and Ian Logan[5]

Since the determination in 1981 of the sequence of the human mitochondrial DNA (mtDNA) genome, the Cambridge Reference Sequence (CRS), has been used as the reference sequence to annotate mtDNA in molecular anthropology, forensic science and medical genetics. The CRS was eventually upgraded to the revised version (rCRS) in 1999. This reference sequence is a convenient device for recording mtDNA variation, although it has often been misunderstood as a wild-type (WT) or consensus sequence by medical geneticists. Recently, there has been a proposal to replace the rCRS with the so-called Reconstructed Sapiens Reference Sequence (RSRS). Even if it had been estimated accurately, the RSRS would be a cumbersome substitute for the rCRS, as the new proposal fuses—and thus confuses—the two distinct concepts of ancestral lineage and reference point for human mtDNA. Instead, we prefer to maintain the rCRS and to report mtDNA profiles by employing the hitherto predominant circumfix style. Tree diagrams could display mutations by using either the profile notation (in conventional short forms where appropriate) or in a root-upwards way with two suffixes indicating ancestral and derived nucleotides. This would guard against misunderstandings about reporting mtDNA variation. It is therefore neither necessary nor sensible to change the present reference sequence, the rCRS, in any way. The proposed switch to RSRS would inevitably lead to notational chaos, mistakes and misinterpretations.
Journal of Human Genetics (2014) **59,** 66–77; doi:10.1038/jhg.2013.120; published online 5 December 2013

Keywords: error spectrum; mtDNA; notation; rCRS

INTRODUCTION

The purpose of having an mitochondrial DNA (mtDNA) reference sequence is to communicate the variation of a lineage in a compact form by listing only the variants of a new lineage relative to the selected reference sequence. Transmission of entire sequences in, for example, a FASTA format would not fit into the text of a scientific paper or a medical or forensic report because of space limitations—but it is often practical to give a list of only those variants found by comparison against a reference sequence. Since the first determination of the human mtDNA sequence, which was baptized as the Cambridge Reference Sequence (CRS),[1] it became a tradition for researchers in the mtDNA field to report mtDNA variation in a table with the columns showing the sequence positions where the sample lineages differ from the CRS. The CRS was eventually corrected in 1999 by resequencing of the same sample to eliminate the initial sequencing errors. Since then, this improved sequence has been referred to as the revised CRS (rCRS).[2]

In April 2012, *The American Journal of Human Genetics* featured an article[3] as a 'next-generation' approach to understand human matrilineal diversity. The *Journal* proposed that the replacement of the human mtDNA reference sequence, the rCRS,[2] by a reconstructed ancestral sequence, the Reconstructed Sapiens Reference Sequence (RSRS),[3] lays the groundwork for a *new way* of analyzing mtDNA. However, there is widespread doubt as to whether these suggestions have any substance and whether the arguments brought forward by Behar *et al.*[3] are sound and carry enough weight to justify a substantial notational change in the way by which human mtDNA profiles are described.

The authors argue that using the rCRS is inherently flawed because some mutations are presented from *derived* to *ancestral* and that a reference sequence that has the *ancestral* nucleotide states is superior. To underscore the importance of their new reference sequence, the authors cite Darwin and even refer to Copernicus. Do these weighty comparisons stand up? We fear that they are misleading and that

[1]Department of Mathematics, University of Hamburg, Hamburg, Germany; [2]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria; [3]School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield, UK; [4]Key Laboratory of Animal Models and Human Disease Mechanisms of Chinese Academy of Sciences & Yunnan Province, Kunming Institute of Zoology, Kunming, Yunnan, China and [5]Exmouth, Devon, UK
Correspondence: Dr H-J Bandelt, Department of Mathematics, University of Hamburg, Bundesstrasse 55, Hamburg 20146, Germany.
E-mail: bandelt@yahoo.com
Received 8 January 2013; revised 29 September 2013; accepted 25 October 2013; published online 5 December 2013

Behar *et al.*[3] themselves are confounding evolutionary theory with issues of notation and convention. In truth, when adopting a phylogenetic approach to analyzing mitochondrial data, it is rather difficult to imagine a researcher confusing a reference sequence with an ancestral sequence and getting the evolutionary direction of mutations wrong as a result. A notation can be more or less convenient or more or less cumbersome in this or any other respect; however, whichever representation is eventually used has no influence on the phylogenetic analysis. Hence, there is no *new way* of mtDNA analysis created by switching from one reference sequence to another.

We consider that none of the reasons given for rejecting the rCRS is valid or compelling and that a replacement of the rCRS by the RSRS would rather provoke a plethora of documentation errors and misinterpretations in mtDNA studies—the same kind of problems the switch is intended to obviate. In fact, even a minor, seemingly cosmetic, change at position 3107 in presenting the mtDNA reference sequence rCRS[2] by keeping the traditional numbering of nucleotides and yet conforming to GenBank requirements has triggered the submission of erroneous mtDNA sequences to GenBank up to the present day (see below and Supplementary Table S1).

We do however appreciate some of the reasons for the assault on the rCRS by Behar *et al.*[3] as there is, unfortunately, a poor level of standardization in the way in which mtDNA variation is described in many medical papers. Several inconsistencies and ambiguities exist concerning the designation and interpretations of mutations. This is notably true for the new sequencing approaches, where a Yoruban (African) sequence was used in place of the normal reference sequence, the rCRS.

Here, we would like to discuss the factors involved in improving the standard of reporting variants when partial, or complete, mtDNA sequences are available and also in those instances when results may be given for just one, or several, nucleotides, as found in many medical studies. We will see that some separate traditions in representing mtDNA variation can well fit together and coexist with special-purpose presentations in evolutionary direction—relative to rCRS, the unique legitimate reference sequence for human mtDNA.

THE ROLE AND CHOICE A REFERENCE SEQUENCE

In many reports, mtDNA variation is described in a table with the columns showing the sequence positions where the sample lineages differ from the reference sequence. Such a table may be accompanied by an estimated phylogeny, where the rooting is performed using a suitable outgroup, and if the outgroup is not too distant to the sample lineages, one can attempt to reconstruct the sequence evolution and estimate *ancestral* nucleotides at the varied positions. This is the standard way of proceeding in science—it is not in any sense Ptolemaic or non-Darwinian. In evolutionary biology, the difference between a reference sequence and ancestral sequences such as a reconstructed root (for uniparental markers) is widely appreciated and well understood, and this is in fact how things have been working in the human mtDNA evolutionary studies for many years.

It is true that the rCRS (as would any other extant mtDNA sequence) occupies a peripheral position on the human mtDNA phylogenetic tree; however, this fact, in itself, does not diminish the rCRS as a reference sequence *per se*. Nonetheless, it is helpful to ask ourselves how a reference sequence might have been chosen if there had been a free choice at the very beginning. Suppose we had a pioneer study of complete mtDNA variation in some species, where no reference sequence had yet been established. What would be the best choice? The optimality criterion would be to minimize the average length of the known mtDNA haplotype profiles (variants lists)—with the side constraint that the reference sequence should be authentic: either a correctly sequenced extant sequence or a reconstructed ancestral haplotype with a very high likelihood of having once actually existed. A consensus sequence for a sample set, which selects a relative majority nucleotide at each position, does not necessarily meet the latter requirement as it may fall outside an estimated mtDNA tree because of recurrent mutations (see Figure 1 for two hypothetical examples). This phenomenon does occur with some human mtDNA data sets. For example, the consensus sequence obtained for the worldwide data set of Kivisild *et al.*[4] differs from the median node (see below) of the corresponding mtDNA coding-region tree, the root of the globally distributed haplogroup L3 by a single mutation (at position 15301).

A node in the mtDNA tree, which in most applications will be very close to the global consensus haplotype, is a *median* node of the mtDNA tree—that is, a sampled haplotype or an estimated ancestral haplotype that minimizes the average distance to all sampled haplotypes connected by this tree. The median nodes of trees are known under several names in mathematics and economy and are readily identified algorithmically; see Kincaid[5] for a recent survey. Whether a node in an unrooted mtDNA tree qualifies as a median node is easy to determine by counting the numbers of sampled haplotypes in each branch emanating from the same node. If for a particular node none of its emanating branches carry a strict majority of all sampled haplotypes, then it qualifies as a median node. Otherwise, one has to pass from the node just processed to the neighbouring node in the strict majority branch for the next query and continue until a median node is found. The determination of a median node is thus independent of the rooting of the tree and yields either a unique node or a pair of linked nodes (in the case of a 50–50% tie at that link).

Unfortunately, anticipating a median node of a yet unknown mtDNA phylogeny is hampered by the fact that the global mtDNA variation of a species is usually not yet well represented by a small mtDNA tree inferred from a pioneer study. However, this is exactly when a reference sequence usually gets installed and fixed for the future. In retrospect, the median node for the present human mtDNA tree would clearly be the root of haplogroup R with European data and likely be the root of haplogroup L3 (or one of its descendent Australo-Eurasian haplogroups M and N) with worldwide data. This was actually the case with the complete mtDNA tree displaying 53 mtDNA lineages from all over the world (Ingman *et al.*[6]): there were 19 haplogroup N and 13 haplogroup M members, six further L3

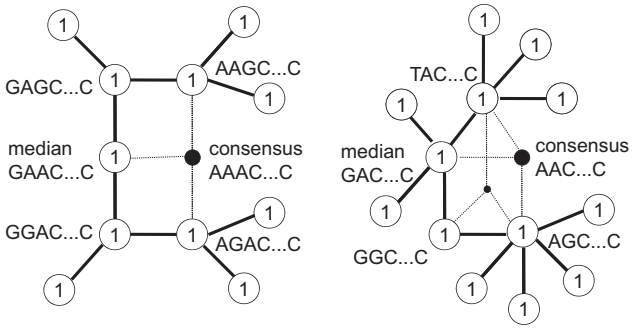

Figure 1 Two hypothetical data sets graphically represented by their (full) quasi-median networks.[100] The relative majority nucleotide is C at each parsimoniously uninformative site and is A at the remaining sites. The unique most parsimonious tree is indicated by bold lines.

lineages and 15 (African) non-L3 lineages (see Bandelt *et al.*[7] for haplogroup notation and a reconstruction of the mutational events).

The global L root of the mtDNA tree, corresponding to a worldwide sample, can never be a median node (in the above sense) because any outgroup (such as Neanderthal mtDNA) would not be included within the sample, and it would be extremely unlikely that both roots of haplogroups L0 and L1′2 (alias L1′2′3′4′5′6) were the median nodes of the mtDNA tree representing some African population samples. For example, in the case of Khoisan mtDNA trees, the median node would either be the L0a′b root or the descendent L0a root (*cf.* Koekemoer[8]). For mtDNA trees of most other African populations, one can expect that the median node is the root of haplogroup L2′3—that is, the smallest haplogroup encompassing haplogroups L2 and L3 (according to the general rules for haplogroup notation first established by Richards *et al.*;[9] see Torroni *et al.*[10]). Note that PhyloTree[11] uses the long form L2′3′4′6, which explicitly indicates that also haplogroups L4 and L6 are embraced by this superhaplogroup.

Thus, reporting human mtDNA profiles with respect to the root of the human mtDNA tree would be far from optimal. This is especially true for Eurasian mtDNA data, as can even be seen with just a few representative examples (see Supplementary Table S2 for three control-region sequences).

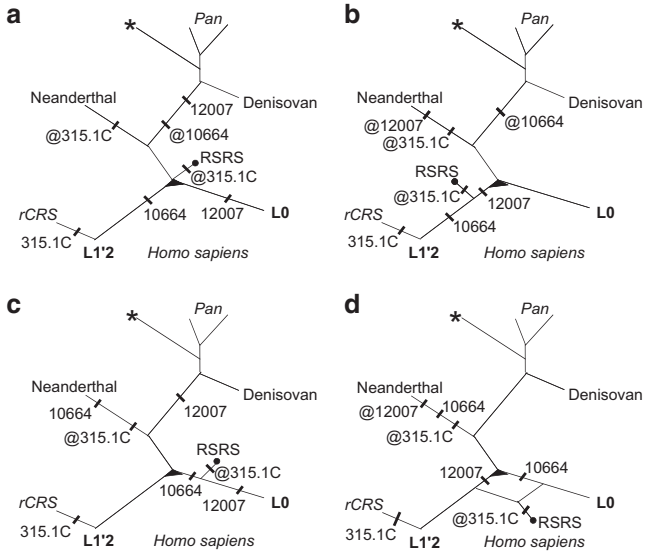

Figure 2 The ambiguous rooting of the human mtDNA tree highlighting the nucleotide variation at positions 10664 and 12007 as well as the indel scored post 315. The estimated root for the *Homo sapiens* mtDNA subtree is indicated by a filled triangle, whereas the rooting of the entire tree is marked with an asterisk. Branches are not drawn to scale.

A BRIEF HISTORY OF THE RCRS AND AN EARLY COMPETITOR

Although correct for the most part, the original CRS of 1981[1] had some minor chimeric features and a few incorrect nucleotides. In the 1990s, it became well known that there were some errors in the CRS; however, it was difficult to distinguish reliably genuine private variants from the errors, and a number of partially corrected versions of the original CRS were in use, even until well after 1999 (see Bandelt *et al.*[12]). In fact, resequencing of the same sample by Andrews *et al.*[2] settled the problem by identifying and correcting 11 errors. Since then this improved sequence has been referred to as the rCRS and can surely be regarded as authentic as it was carefully generated in two parallel sequencing projects.

In the late 1990s, complete mtDNA sequences were few in number (~50) and scattered throughout the literature, mainly in disease contexts. Many of those early mtDNA sequences were fraught with problems.[13] In 1994, the CRS was contrasted with nine (nearly) complete mtDNA sequences from seven Japanese samples and two samples of European matrilineal origin; see the tree presented in Figure 2 of Ozawa[14] (*cf.* Table 1 in Bandelt *et al.*[12]). It was then believed that the root of the Eurasian mtDNA fell in between the CRS and the remaining lineages. The reference sequence could in fact have been adjusted to a more central rooting (in what we would now call the L3 root) with the addition of a haplogroup L0a2d sequence, which was submitted to GenBank under the accession number D38112 in 1994.[15] Thus, in principle, the roots of haplogroups M and R as well as those of the larger haplogroup L3 could all have been inferred (nearly correctly) by 1995.

There was also an early chance to change the reference sequence to the ancestral sequence of haplogroup R when the study of Marzuki *et al.*[16] explicitly set out to determine a consensus sequence from 13 complete mtDNA sequences. However, the bulk of the new lineages added were so poorly sequenced that even with today's knowledge it is hard to determine the haplogroup status in some instances. However, from the table shown in Figure 2 of Marzuki *et al.*,[16] the 13 lineages (read from top to bottom) likely had the following haplogroup status: K1, I1, H3, H1bk (?), U5a2b1a, non-H, A4a1, H1a1, H3 (?), J1c1a, H2a2a1 (CRS), L3e1e and non-H. If these samples had been correctly sequenced, their consensus sequence would have provided an accurate root for haplogroup R.

A prompt switch from the CRS to a reference sequence based on the R root would have resulted in relatively little impact being made on restriction-site (RFLP) analyses and analyses of the two hypervariable segments, which were the main targets of researchers at the time. Differences in scoring would have occurred only at positions 73 and 263 (in the second hypervariable segment) and the *Alu*I site at 7025. However, these opportunities for a change to a *central* position for the reference sequence were missed for good, as to replace the rCRS later than 1999 would have resulted in too much confusion in view of the growing number of completely sequenced mitochondrial genomes becoming available.

It is interesting to see that the new RSRS as proposed by Behar *et al.*[3] continues to use the same numbering of nucleotide positions as in the rCRS; this means that there should be a deletion at nucleotide 3107 (or at 3106). One of the most vexing errors in the old CRS was the scoring CC at positions 3106–3107, where only one cytosine was found after the revision. In order to retain the original numbering of the nucleotides and minimize confusion, Andrews *et al.*[2] sensibly decided that 3106 should constitute a gap in the rCRS. The rCRS in FASTA format was submitted to GenBank (accession number NC_001807.3) on 7 May 2001. Nonetheless, some partially corrected versions of the CRS have stayed in use since 2001—for example, the data of Vives-Bauza *et al.*[17] and Komaki *et al.*,[18] where C3106del (along with a mosaic of other erroneous variants inherited from the old CRS) were scored. In the former paper, eight of the known error positions (3106, 4985, 9559, 11 335, 13 702, 14 199, 14 272 and 14 766) as well as the rather stable insertion 315.1C relative to rCRS were scored quite differently in patients with Parkinson's disease (PD) and in controls, signalling mixed use of the CRS and the rCRS, besides incomplete sequencing (as is also seen at other positions).

Significantly, 3 years later, Maruszak *et al.*[19] in a review referring to this study, stated that '*C3106del (16S rRNA), prevailing in PD individuals had been previously reported as a common polymorphism*'.

As late as 2006, an incorrect reference sequence, still bearing some of the old errors of the CRS such as CC at 3106–3107, was in use (*for example,* Martorell *et al.*[20]), as commented upon in Bandelt *et al.*;[21] see the reply of Martorell *et al.*[22] Unfortunately, even now there is still not a full understanding of the scale of this problem. Even Behar *et al.*[3] have sequences in their recent data set that are mistakenly scored as CC at positions 3106–3107; see, for example, the GenBank sequence JQ706067 published 7 March 2012.

This all goes to show that, even 11 years after the publication of the rCRS by Andrews *et al.*,[2] the difference between the CRS and the rCRS has not yet been fully appreciated by all researchers of mtDNA. From the above examples, one could extrapolate that a radical change of the reference sequence as proposed by Behar *et al.*[3] would have repercussions for more than a generation (>25 years), at least in medical genetics.

It is worth reflecting on the sequence of events that led to the above confusion concerning 3106–3107. It begins, as stated above, with the rCRS sequence NC_001807.3 appearing on 7 May 2001 with the original position 3106 deleted, whereas an incorrect version of the rCRS appeared on GenBank later, on 13 November 2001 (J01415.1) with the outdated 'CC' at 3106–3107. Eventually, this erroneous sequence was replaced (25 August 2006 J01415.2); however, this time the 3106–3107 couplet was scored as 'CN' instead of just the 'C' (which would have caused the position shifted by –1 after 3106 by the automatic numbering). All subsequent revisions, including the latest one dated 30 April 2010 (NC_012920), continue to have 'CN'. This version of the rCRS has thus been promoted to official RefSeq status in GenBank and consequently also been recommended by Human Genome Variation Society (HGVS) (http://www.hgvs.org/mutnomen/refseq.html#wt). One should bear in mind however that this sequence has been stretched on the procrustean bed of GenBank just in order to retain the original numbering of the old CRS. Taken at face value, as apparently carried out by several scholars of mtDNA research, it would not constitute the real rCRS as N stands for one of the nucleotides A, G, C, T according to the IUPAC code and hence *not* for a gap.[23] Thus, the presently featured rCRS is one nucleotide longer than the real rCRS. Confronted with this elongated version, any comparison with real sequences should formally include the scoring 3107del, as was consistently performed—for example, in Taylor *et al.*[24] In Aikhionbare *et al.*[25] one finds a mutation scored as '3107delG', which could only refer to the H-strand.

The complete human mtDNA sequences stored in GenBank now testify to several different scoring methods at 3106–3107. Most submissions show the correct sequencing results—that is, just a 'C' appears. Other submissions reflect the outdated addition of an extra nucleotide C; later submissions show the pattern NC,[26] whereas some recent submissions show CN under the influence of the recent GenBank modification of the rCRS (for example, Chandrasekar *et al.*[27]). The false CC has also been reported in a number of sequences (for example, accession numbers JQ706025, JQ706026, JQ706032, JQ706041, JQ706045-JQ706047, JQ706062 and JQ706067) as originally submitted to GenBank by Behar *et al.*[3] The correct single C as well as the incorrect CN may even co-occur within the same data set (for example, Gonder *et al.*;[28] Pala *et al.*[29]); see Supplementary Table S1. The false CT (that is, N3107T) has been reported in 26 out of 31 partial mtDNA sequences of patients by Houshmand *et al.*[30] With the advent of high-throughput sequencing, the N scoring at 3107 in the reference sequence eventually led to new errors, incurred by suboptimal bioinformatics tools (such as Sequencher) and lack of quality control. Several mtDNA sequences submitted by Barbieri *et al.*[31] show varying patterns of undetermined nucleotides in the region 3105–3113 as well as a false insertion of one T; see Supplementary Table S3.

However, most researchers find the rCRS is easy to use, and the problems with the deletion should not cause any difficulty as long as care is taken. It is highly desirable that all sequences submitted to GenBank should be of the actual nucleotides and not include the dummy nucleotide by default. The gap in the numbering of nucleotides could be generated by automatic tools handling the sequence information. Of course, the optimal solution would be that GenBank permitted reference sequences to contain gaps.

CONSENSUS SEQUENCES IN DISGUISE

The reference sequence has hardly ever been confounded with the ancestral sequence of the entire mtDNA phylogeny; however, it has often been regarded as a sort of 'wild-type (WT)' or consensus sequence. Actually, the 'WT' metaphor in mtDNA research is only appropriate for somatic mtDNA variation with one clearly dominant type and singular aberrations but not so for mtDNA variation typically found in a species or a population group, although this has been quite common practice, see for example, Table 2 of Wallace.[32] As emphasized on the website of the HGVS '*the debate about what is wild type can be unsolvable when variants are very common (near 50%) or differ between populations*' (http://www.hgvs.org/mutnomen/refseq.html#mtDNA). With human mtDNA, there cannot be a solution either by splitting the human mtDNA pool and artificially delineating 'races' (such as 'African', 'Asian' or 'Caucasian', as has been popular in medical and forensic genetics; see Salas *et al.*[33] and Bandelt *et al.*[34]), each possessing their own 'WT' mtDNA sequences.

The Cambridge Reference Sequence (as the CRS and the rCRS) has, however, repeatedly been mistaken as a consensus sequence. For example, it was described as the 'human mitochondrial Cambridge consensus sequence',[35] 'Anderson mtDNA consensus sequence',[36] 'human mtDNA consensus sequence'[37] and simply as 'Cambridge consensus sequence'.[38] In the latter case, the fact that the CRS (and not the rCRS) was cited and labelled a consensus sequence highlights the lack of phylogenetic understanding in this sub-community, manifest in the way the results are presented and interpreted (see Salas *et al.*[39] for a detailed critique). The consensus mislabelling was eventually handed down to the rCRS, where it still thrives (for example, Lu *et al.*[40] and Yan *et al.*[41]). The labelling of the rCRS as the WT sequence can also still be found in the medical literature; see, for instance, the Supplementary Table of Zhang *et al.*[42]

As a curiosity, not only a reference sequence but also the root of the mtDNA phylogeny has been confused with a consensus sequence derived from some sample set deemed to be representative of worldwide mtDNA variation. This was cited in an article[43]— which actually followed the agenda of creationism[44]— using a strongly Eurasian-biased sample; the root of haplogroup R was then offered as the root of the global mtDNA phylogeny, in a way as a late echo of Ozawa.[14]

Summarizing, the problem that persists is that some scholars of human mtDNA variation evidently believe that there exists a 'WT' mtDNA sequence and do not fully appreciate that the reference sequence is just an arbitrary extant sequence selected for notational purposes. Even so, this cannot be attributed to the rCRS acting as a reference sequence. In the absence of phylogenetic thinking, even an ancestral sequence would run the risk of being mistaken as a consensus or WT sequence. Therefore, replacing the rCRS by the RSRS would not *per se* lead to phylogenetic understanding or exert

'*educational influence on the scientific and public perception of human evolution*'.[3]

ERROR PREVENTION

One of the arguments against the use of the rCRS expressed by Behar *et al.*[3] concerns the errors reflecting a reference bias. It is certainly true that the fact that the vast majority of mtDNA lineages in any population group do not share the nucleotides with the rCRS at the seven positions defining H2a2a1 within haplogroup H (263, 315.1, 750, 1438, 4769, 8860 and 15 326) has led to numerous problems. Early sequencing attempts typically resulted in incomplete sequences and those seven positions were often afflicted. However, a switch of reference sequences from rCRS to RSRS would just aggravate this problem, as all of a sudden >40 nucleotide variants would separate the new 'reference' sequence from almost all Eurasian mtDNA sequences. This would provide ample opportunity to miss several of those variants, paving the way to interpreting some of those differences as being disease-associated.[12]

Most vexing has been the misunderstanding of the role of the transition at 8860, as in virtually all data sets its consensus nucleotide (G) is predominant and different from the rCRS nucleotide (A). Both nonsynonymous mutations A8860G and A15326G are characteristic for virtually all mtDNAs outside the very minor European haplogroup H2a2a, and should therefore always be recorded at frequency close to 100% whenever the two mitochondrial genes *ATPase*6 and *ND*1 were screened in any population; however, for example, Güney *et al.*[45] failed to record a single instance of A8860G or A15326G in 60 samples. In contrast, the authors of a recent study on the presence of mitochondrial mutations in inherited cataracts[46] found it worth stating and referencing that A8860G and A15326G were '*reported in subfertile, abdominal aortic aneurysm, AD, PD, T2DM w/wo angiopathy, LHON, and dystonia*'. Aikhionbare *et al.*[25] claimed that the change at 8860 occured in 39 of 52 colorectal cancer adenomas and cancer tissues but only in 7 of 52 matched surrounding normal tissues. In a most extreme case, the title of an article posed the question: '*Is 8860 variation a rare polymorphism or associated as a secondary effect in* hypertrophic cardiomyopathy *disease?*';[30] the authors went on to claim that A8860G was found in 30 out of 31 patients of hypertrophic cardiomyopathy but in *none* of their 60 controls. In a similar vein, but less extreme, A8860G was highlighted as occurring in 20 out of 23 oligoasthenozoospermic men, without disclosing the corresponding frequency in the control group.[47] Later, the same research group asserted that A8860G was present in 21 out of the 23 infertile men but in only 14 out of the 30 controls.[48] These two papers (and others) were discussed by Palanichamy and Zhang,[49] who emphasized the implausible and inconsistent frequency patterns of haplogroup-specific mutations recorded in those Indian studies.

Most of the rCRS-biased errors could in fact have been prevented by referring to a mtDNA tree, as it is important to distinguish older inherited variants from recent or somatic ones. The presence of this kind of error has often signalled that an mtDNA data set was generally of insufficient quality.[50,51] Tracing an mtDNA lineage in a basal mtDNA classification tree (as offered by PhyloTree) can help to sort out potential oversights. At least, an mtDNA lineage under consideration should be allocated to the finest haplogroup level. Then the number of additional, seemingly private variants can be quite small, especially with well-classified European mtDNA lineages. For instance, the two identical sequences with GenBank accession numbers GQ129152 and GQ129153 (see http://www.ianlogan.co.uk/sequences_by_group/u5b3_genbank_sequences.htm) could optimally be referred to as U5b3a1a and A7403G and G12172A haplotypes (relative to the rCRS), meaning that they share all the variants required for haplogroup U5b3a1a status but in addition bear the two variants listed explicitly. One should bear in mind, however, that the naming of haplogroups may slightly change over time, so that proper reference to a specific article or a dated version of a website (such as PhyloTree[11]) is needed.

THE RECONSTRUCTION OF ANCESTRAL SEQUENCES

Ancestral sequences, *alias* roots of particular haplogroups, serve to determine the evolutionary direction, or polarity, of mutations along the links of the mtDNA phylogeny. For example, ancestral nodes of the mtDNA tree of all (error-free) complete mtDNA sequences have been used for allocating partial mtDNA sequences to their candidate haplogroups. For such comparisons, only the separating mutations (and no additional mutations drawn from the sequence motifs relative to the rCRS) are relevant; this, however, was previously not really put into practice in available automatic tools for haplogrouping.[52]

Typically, ancestral haplotypes cannot be directly observed in samples but have to be reconstructed from derived extant sequences. Standard phylogenetic analyses can be employed to estimate the roots of haplogroups and, in particular, the L root, which is the ancestral sequence for the entire mtDNA phylogeny. The results of such an estimation may differ greatly in regard to the potential authenticity of the reconstructed root. The optimal situation is with a multifurcation node, where many lineages branch off, testifying to a sudden radiation event, as is the case for a number of haplogroup roots between the rCRS and the root of the ubiquitous Afro-Eurasian haplogroup L3—but not further down the (African) mtDNA tree, where only bifurcations occur.

The most difficult case, in fact, is the estimation of the L root of the entire human mtDNA phylogeny, as the outgroup (Neanderthal mtDNA) is not very close. Not only that, but the two deepest branches carrying haplogroups L0 and L1′2 are characterized by as many as 10 and 8 mutations, respectively (according to Build 15 of PhyloTree). The recent augmentation of the existing database on complete mtDNA genomes provided by Behar *et al.*,[3] although certainly having a considerable impact on the fine classification of European mtDNA lineages, has no influence on the reconstruction of the L root. Only new population samples from southern and eastern Africa not yet analyzed for mtDNA could possibly have provided additional evidence; however, the clients of the company which contributed the bulk of the new data presented by Behar *et al.*[3] tend not to come from these parts of the world. The present rooting of the mtDNA tree has actually not changed since Build 6 of PhyloTree (released 28 September 2009). Therefore, this root was not really new in 2012; the only novel features are that it now contains estimated nucleotides (or gaps) at the few positions ignored in PhyloTree (as deemed to be extreme mutational hotspots) and that it has been baptized and promoted as 'the Revised Sapiens Reference Sequence'—really something of a misnomer, as a new *reference* sequence should only be installed through consensus across the entire mtDNA community.

On the 'mtDNA Community' website (http://www.mtdnacommunity.org/about.aspx), it is claimed that '*the extinct complete human mtDNA root sequence was never precisely determined*'. We agree because all builds of PhyloTree since Build 3 have not only ignored a handful of extremely frequent changes but also the variation at position 315.1. The length polymorphism of the majority 6-cytosine tract 311–315.1, however, is not an extreme hotspot, quite in contrast to the one of 303–309.[34] The bulk of length changes attributed to the post-310

cytosine tract are in fact incurred by a T to C transition at 310, which unites the two cytosine tracts. Such a long stretch is highly prone to (somatic) length polymorphism. By current convention, any length change is however scored at the end (in 3′ direction), thus hitting around 315.1. To estimate an upper bound for the frequencies of a 5-cytosine tract (downstream from T310), we have screened the forensic literature data set stored in EMPOP (empop.org) by assuming G at 263 but shared states with rCRS at 310 (T) and 315.1 (deleted). This essentially excludes haplogroup H2a2a1 and the secondary length polymorphism caused by T310C. We then find only six mtDNA sequences with this pattern out of 6757 mtDNA sequences. This number is much lower than the number of mutations (>25) at the 67 top mutational hotspots in a tree connecting 2196 complete mtDNA sequences, as estimated by Soares *et al.*[53] Therefore, position 315.1 cannot be regarded as hypervariable and should not be ignored in future builds of PhyloTree.

As both deepest branches of the human mtDNA phylogeny show the insertion 315.1C compared with rCRS, the most parsimonious solution is to assume the majority state of six cytosines in 311–315.1 for the L root—that is, the root of the entire mtDNA tree. This is also enforced by the available chimp/bonobo and Denisovan mtDNA sequences that (although differing in the pre-310 cytosine tract) support the longer version of the post-310 cytosine tract (although not all of the sequencing results may be sufficiently reliable for this homonucleotide tract). In any case, the RSRS is unlikely to be the real root of the mtDNA phylogeny, as it appears that the number of cytosines flanking the T at position 310 in this hypothetical sequence were just copied from the rCRS by default and without proper analysis.

Before the availability of several complete mtDNA lineages from the extinct Neanderthals, the rooting of the human mtDNA tree was based on the very distant chimp (and bonobo) mtDNA sequences as outgroups (besides a few shorter stretches provided by certain numts); see Kivisild *et al.*[4] Nonetheless, the same 18 mutations separating the roots of haplogroups L0 and L1′2 have been present since Build 1 of PhyloTree in 2008, albeit sorted differently between the intermediate L root and its two descendants. In particular, both Kivisild *et al.*[4] and Logan[54] assumed that a transition at 12007 was part of the L root profile, whereas a transition at 10664 was not, in contrast to what has been asserted for the RSRS. The reason for this discrepancy is that the Neanderthal sequences show the opposite pattern. Now, although, additional support for this aspect of the previous rooting comes from the two Denisovan mtDNA sequences that bear the chimp/bonobo nucleotide states at both positions. The Denisovan sequences were ignored by Behar *et al.*[3] but they testify to three equally parsimonious rootings involving the two positions 10664 and 12007 (Figure 2). We suggest to use the L root estimated in Figure 2d, as this would come closer to previous estimates of the global ancestral sequence (Kivisild *et al.*[4] and PhyloTree, Build 2[11]). The profile of this reconstructed root sequence is shown in Supplementary Table S4.

More profound indeterminacy concerning the reconstruction of the L root emerges when one considers most parsimonious hominin trees of modern human mtDNA lineages including Neanderthal and Denisovans mtDNAs. For instance, the variant 189G relative to rCRS is shared by the two outgroups and haplogroup L0a′b′f′k, whereas the rCRS nucleotide A189 (following the notation of nucleotide states illustrated in Figure 3) is common to L0d and L1′2. Therefore, an equally parsimonious reconstruction of the basal variation at 189 would assume 189G for the L root and thus allocate changes from G to A at 189 to the two links leading to the roots of L0d and L1′2.

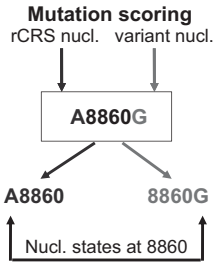

Figure 3 Standard rCRS-based notation for mutations and nucleotide states. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

Likelihood analysis would certainly reveal a whole spectrum of different L roots with similar likelihoods.

The RSRS is therefore a merely hypothetical root and is very likely non-authentic for the reasons given above. Even worse, any future discovery of a novel most basal branch of the human mtDNA phylogeny or yet another cousin group of modern humans and Neanderthals might necessitate a further reassessment of the root of the modern human mtDNA tree. Therefore, equating a reference sequence with some reconstructed root relative to the database of all available modern human mtDNA sequences has the severe disadvantages of non-authenticity and instability in respect to future findings. Any reconstruction of an ancestral mtDNA sequence is inevitably subjected to perpetual revision; however, an ongoing annual update would hardly be a desirable feature for a reference point.

THE 'YORUBAN MTDNA REFERENCE' INTERLUDE

In a cloak-and-dagger operation, a new human mtDNA 'reference' sequence was uploaded in GenBank as NC_001807.4 on 27 December 2001. This sequence was transferred from a Yoruban mtDNA sequence with GenBank accession number AF347015 (submitted earlier in 2001), which now has haplogroup status L3e2b1a1 (with private variants at positions 408, 11017, 11722, 12850, 14580, 15932, and some length variants that are generally not shown on PhyloTree). Who exactly was responsible for this coup remains in the dark; however, it took quite some effort and time by the MITOPMAP team to get this reference sequence removed.

However, the timing here was unfortunate and some commercial single nucleotide length polymorphism (SNP) chip manufactures started to use the Yoruban L3e2b1a1 sequence as their reference point. This affected the SNP-chip arrays used for genome-wide association studies. For example, in 2007, the commercial company deCODEme used 163 mtDNA SNPs, where alignment and variant nucleotides referred to the Yoruban sequence. The complications that arose from the switch to the 'Yoruban reference sequence' can be inferred from the Website http://www.snpedia.com/index.php/MtDNA_Position_-Conversions.[55] Moreover, the SNP sites were not chosen prudently for haplogroup determination. Quite a number of mtDNA haplogroups known at the time could not be recognized, such as the important nested subhaplogroups R0 (also denoted as pre-HV before 2007) and R. In contrast, the peripheral haplogroups H2a and L3e2b1a to which the rCRS and Yoruban sequences belong were targeted with absurd precision, namely, the panel of SNPs contained positions 1438 (H2), 4769 (H2a), 750 (H2a2), 2352 and 14212 (L3e), 9377 (L3e2b1) and 2483 (L3e2b1a).

The Affymetrix MitoChip v.2.0 had been employed in a number of studies since 2007; see references given in Thieme *et al.*[56] results obtained with this chip have suffered from a few miscalls;[57] however,

the full amount was probably underrated in some studies (for example, Rollins *et al.*[58]). Special care and software help to improve sequence accuracy;[56,59,60] however, one would still have to reckon with some spurious N calls.[61]

The main problem with documenting and interpreting the MitoChip results is, however, the reference to the Yoruban sequence. In the case of the results displayed in Table 3 of Rollins *et al.*,[58] the changes at 750 and 4769 appear in as many as 11/77 of the patients and for 1438 a frequency of 21.4% is listed (which could best correspond to 15/70, whence there must have been at least seven ambiguous calls). Note that positions 709 and 4769 are rather conservative and were unvaried among the complete mtDNA sequences from schizophrenic patients in the studies of Ueno *et al.*[62] and Bertolin,[63] except for A4769 in one haplogroup R2 sequence (where this variant relative to rCRS enters the haplogroup motif). In another study,[64] the frequency of A4769 in patients (11/454) was even found to be slightly lower than in controls (18/616), likely indicating the haplogroup H2 status. Inasmuch as the frequency of haplogroup H2 in the study of Rollins *et al.*[58] equals 3, of which at most 2 would belong to H2a2 (in view of Table 1 and Figure 1 in that article), we conclude that nine mtDNA samples likely had their rCRS nucleotide and variant nucleotide inverted at the sites distinguishing the rCRS from the Yoruban sequence. A partly inverted reading is then also to be expected for T195C, which would read as 'C195T' with respect to the Yoruban sequence as reference point. The latter reading was actually used in Rollins *et al.*[58] The extremely high frequency 57.1% (44/77) of 195C in the patients' group has not been supported by a subsequent study.[65]

On the other hand, the polymorphism C12705T can be found on a variety of commercial SNP arrays and is stored in dbSNP as rs2854122 (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=2854122) with the reference 'allele' being a C. A switch of the reference point to the RSRS would also implicate a switch in reporting this polymorphism to the reference nucleotide being a T. Therefore, there is already discordance in data storage between dbSNP and the default version of PhyloTree. As on some SNP arrays (for example, Illumina HumanHap550v3.0) this SNP is targeted on the heavy strand of the mitochondrial genome, even more confusion has been created in reporting this SNP.

SCORING MUTATIONS—NOTATIONAL DISCORD

mtDNA profiles have traditionally been scored relative to the rCRS; however, notation for mtDNA variation differs across the fields of medical genetics, forensic genetics and molecular anthropology, and even within one field several notational styles may coexist. In the past, forensic genetics did not use prefixes for representing mtDNA profiles but consistently used suffixing for indicating the variant nucleotide relative to the rCRS. In molecular anthropology, prefixes were not provided either and suffixes were kept to a minimum. Therefore, a first transitional hit as seen from rCRS at position 15301 is simply recorded as '15301'. Suffixes are only employed to indicate transversions and indels. Recurrent hits along the pathway from rCRS to a node in the mtDNA tree were highlighted by prefixing with @.

In medical genetics, both circumfixing (that is, the simultaneous use of pre- and suffixes, such as A15301G), and the alternative '>' notation (15301A>G) have been used, although in a rather inconsistent way. The HGVS (http://www.hgvs.org/) has specific recommendations for the nomenclature of mtDNA variants. Namely, the HGVS indicates that '*…according to current recommendations variants in the mitochondrial DNA should be described in relation to a full mitochondrial DNA sequence, i.e. for human the Homo sapiens mitochondrion, complete genome (GenBank NC_012920.1). Descriptions should be preceded by 'm.', like m.8993T>C.*' Note that 'm.' abbreviates 'mitochondrial sequence'. This recommendation is not always followed exactly: for example, instead of m.10398A>G one encounters m.10398G>A.[66,67] This notwithstanding, many medical journals (such as *Journal of Medical Genetics*) still publish studies of putatively pathogenic mutations scored in the circumfixing style; see, for instance, Yan *et al.*[41] This style is often used for tables; see, for example, Gil-Borlado *et al.*[68] Some journals (such as *Human Mutation*) strictly adhere to the above recommendation; however, this does not extend to diagrams where this would be unnecessarily lengthy; see, for example, Figure 4 of Bi *et al.*[69]

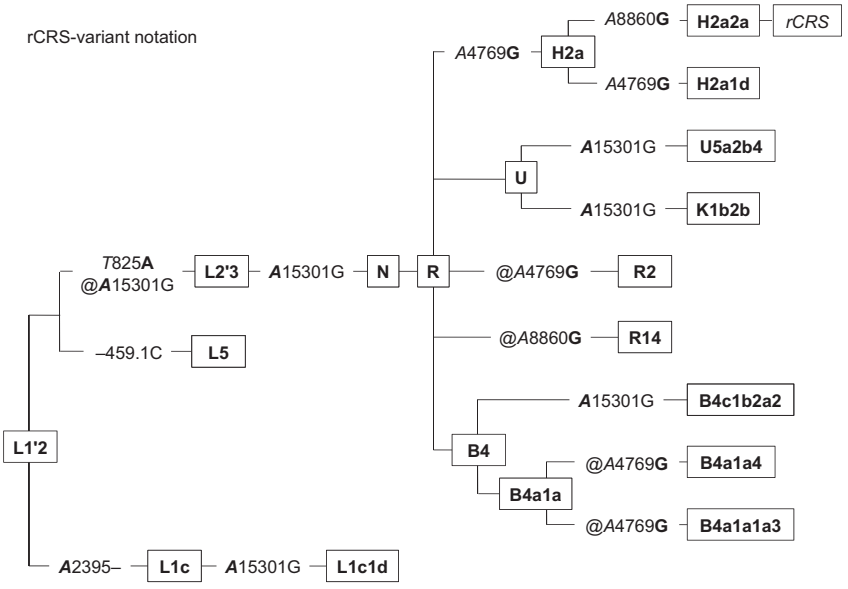

Figure 4 A view on the part of the basal worldwide mtDNA phylogeny in which the nucleotides at positions 825, 4769, 8860, 15301 vary and indels are scored at positions 459 and 2395. For each mutation, the rCRS nucleotide state is indicated as an italic prefix and the variant state as a suffix; the ancestral nucleotide states (relative to the L1′2 root) are shown in boldface. Haplogroups are denoted as in Build 15 of PhyloTree.

In any way, the circumfix notation was not always used consistently in mitochondrial genetic disease association studies. For example, the common polymorphism on 10398 has to be denoted as A10398G following the standard approach of indicating the rCRS nucleotide in front. Nevertheless, a number of papers all dealing with a presumptive genetic predisposition to breast cancer and other diseases reported this polymorphism as G10398A,[70–79] whereas other papers reporting a diversity of disease associations correctly referred to this SNP as A10398G. There are approximately twice as many publications reporting the correct rCRS-based notation of A10398G as papers using the inappropriate notation 'G10398A'. The latter could be interpreted as a G to A mutation in evolutionary direction, which constitutes one of the characteristic changes within haplogroup L3 on the way to the nested haplogroup N. It is not clear, however, whether this interpretation was actually intended by all those who used this scoring. For instance, there is a case in the literature[80] where the circumfixes were interchanged consistently so that the rCRS nucleotides served as suffixes and the variant nucleotides as prefixes. As a consequence of this switch, haplogroups were misdetermined in that five mtDNA lineages (from haplgroups K2a5a, H1c, T2b7a, U5b3a2 and H1bb, respectively) were all allocated to haplogroup H2 (see Table 10 in chapter 6[80]). 'G8860A' can also occur as single typos for example, Holyoake *et al.*[81] (overlooked by Bandelt[82]) and Aikhionbare *et al.*[25,83]

Scoring mutations in the evolutionary direction presupposes that the ancestral nucleotides at all positions in the mtDNA sequence are known. Such scoring has occasionally been applied, for example, by scholars of the Estonian school of mtDNA research, for example, in the tree diagram of Hartmann *et al.*[59] by employing suffixes, and earlier by Kivisild *et al.*[4] for coding-region data labelled in the anthropological style; there position 825 is the only one where a transversion between the then estimated L root of the mtDNA tree and the rCRS was recorded, namely as '825T' meaning that the rCRS nucleotide is the derived state T (whereas standard notation would require 'T825A' or '825A'); in addition, all indicated changes on the amino-acid level conformed to the evolutionary direction. Similarly, '825T' has always been used in PhyloTree. Accordingly, Sequeira *et al.*[65] have presented an mtDNA tree in this style, where the root of haplogroup L1′2 served as the sequence ancestral to 23 complete mtDNA sequences, from which mutations were scored. Note that this worked well without invoking a change in reference sequences or bringing the L root (or the RSRS) into play.

Mutations were scored (root-upward) relative to the root of haplogroup R by Hudson *et al.*[84] A root-upward notation of mutations relative to the root of haplogroup JT was aimed at by Pala *et al.*,[29] except for the length polymorphism at 315.1, which was annotated rather in the rCRS-based style and variably with or without the @ prefix. There exist only few other instances scattered through the literature where mutations between the L root and the rCRS were deliberately scored in evolutionary direction using the circumfix style (leaving aside singular typos) but alas without indicating the deviating interpretation of circumfixes—for example, 'A11719G',[85] 'T7028C', 'A11719G', 'T14766C'.[86] Hence, scoring mutations in evolutionary direction is not an invention of Behar *et al.*[3] (or van Oven *et al.* [87] and Achilli *et al.*[88]) but was used earlier in several studies *without* invoking a change of reference sequences. The acute problem manifest in PhyloTree, from Build 14 onwards, is that circumfix codes are now being used and read from ancestral (prefix) to derived (suffix), which deliberately supersedes the traditional scoring system.

Virtually all possible combinations of allele/variant or mutation/polymorphism with position numbers and a single nucleotide or pairs of nucleotides seem to occur in the medical literature. For instance, m.10398A>G (p.T114A) has been referred to as a nonsynonymous *variant*,[89] or m.10398G as a common *polymorphism*[90] or an *allele*.[91] More precision and some standardization of wording would be desirable. Especially the terms 'mutation' and 'polymorphism' have been used inconsistently (for example, Shin *et al.*[92] and Rosson *et al.*[93]). An explicit but unsatisfactory definition that is often applied is gleaned from nuclear DNA: a mutation is defined as a change in a DNA sequence from the normal variant that is prevalent in the population to a rare and abnormal variant, where the (arbitrary) cutoff point between a mutation and a polymorphism is 1 per cent (http://genome.wellcome.ac.uk/doc_WTD020780.html).

STANDARD NOTATION OF MTDNA VARIATION

The rCRS-based nomenclature suffices for the needs of present biomedical applications (including medical, population and forensic ones), although each field has its own traditions and conventions. In order to unify previous notational systems, we propose expanding the traditional variant notation to a common system that enables one to read off the mtDNA profiles directly from the tree without additional consultation of the rCRS for particular nucleotide positions. Each field could then resort to various forms of shorthand for tables or diagrams whenever it was felt necessary. We will make a clear distinction between reporting an mtDNA profile and focussing on the evolutionary features of either the entire mtDNA phylogeny (in regard to selection) or of particular positions deemed to have a role in a disease context.

The options for standardization of notation for bringing the slightly different conventions under a common roof are quite restricted when for each field only minor notational changes appear acceptable. Forensic genetics has hitherto employed suffixing with the variant nucleotide relative to rCRS. This has the disadvantage that transversions can no longer be distinguished from transitions without consulting the rCRS nucleotide. Transversions are quite rare, and if they occur more frequently than to be expected, this could indicate documentation errors and would prompt one to consult the primary data again. Molecular anthropology has applied suffixes sparingly, just for transversions and indels. Circumfixing has been the traditional approach in medical genetics. Therefore, the common roof for a long form can only amount to circumfixing with the rCRS nucleotide in front and the variant nucleotide at the end.

Specifically, for reporting a profile, a change at 15301 from the rCRS nucleotide to its transitional variant would be denoted as A15301G with the understanding that the prefix nucleotide refers to the rCRS nucleotide. If proceeding further along the down-up pathway from the rCRS to some point in the mtDNA tree we encounter a second transitional change at 15301, thus back to the rCRS nucleotide, this gets recorded as @A15301G. If further on, eventually yet another transition at 15301 is to be recorded, then one writes A15301G again. Thus, by traversing from rCRS towards the targeted node of the mtDNA tree one erects the mtDNA profile step by step, whereby later changes at the same site replace the earlier changes; see Figure 4 for a view at the mtDNA phylogeny through the lenses of positions 459, 825, 2395, 4769, 8860 and 15301. Note that with respect to these positions, the L1′2 root would be annotated relative to rCRS as T825A A4769G A8860G (thus, in particular, with no change at 459, 2395 and 15301 compared with the rCRS).

The @ notation has been viewed critically by Behar *et al.*[3] in the way that it has often been referred to as a 'back mutation'. This is in fact problematic for two reasons: first, this might allude to polarity and thus evolutionary direction, which was not intended (at least, by most authors who used this symbol) and in fact may be violated, and

second, if one had (in evolutionary direction) the chain of changes A→C→T→A at some position, then the final mutation from T to A would not really be a back mutation in a strict sense, as a direct mutation from A to T has not occurred before in that line of descent. Therefore, @ is defined here as an arrival at the rCRS nucleotide from another nucleotide when traversing the down-up pathway (without imposing any evolutionary direction). This has simply the technical function as highlighting a cancellation rule for profile construction in that—for example, both A15301G and a subsequent @A15301G cancel out each other.

A deletion (relative to the rCRS) is recorded in a similar manner as a substitution, with the only difference that the missing nucleotide is denoted by a short dash '–' or preferably by a (slightly longer) minus sign '–' such as A290– A291–, for example. Insertions are treated analogously: a first new inserted position is indicated by '.1' and a second inserted position by '.2', and so on. For example, three inserted cytosines following position 573 are denoted by –573.1C –573.2C –573.3C, where the prefix '–' reminds one that this position is absent in the rCRS. For better emphasis in a running text, indels may be reported in short forms by using the acronyms 'del' and 'ins'/ '+' (plus the necessary specification). For long indels, the listing position by position would become a bit cumbersome so that a short form such as (290–291)del or 573+CCC is preferable. The well-known nine-base-pair deletion is usually abbreviated by the acronym 9bpdel. For any indel, a return to the rCRS state would get prefixed with @ as in the case of substitutions.

Further short cuts may be performed by using auxiliary reference nodes. For instance, if control-region sequences were compared, then it would be helpful to assume A73G A263G –315.1C from the outset (as in part stated in Table 1 of Kivisild *et al.*[94] and Supplementary Table S2 of Kong *et al.*[95])—that is, effectively using the R root as the reference point for the control region (see for example, Kivisild *et al.*[94]); one should then list any returns to the rCRS nucleotide by repeating the corresponding code (preferably prefixed with @ for clarity), so that the initial default setting at that position would get cancelled. If, for example, the complete mtDNA lineage stored in GenBank under accession number AY739001 with profile A263G –309.C –315.C A750G A1438G A4769G A8860G A15326G T16362C T16519C relative to the rCRS was to be reported in a table where all profiles were annotated relative to the R root (A73G A263G –315.1C A750G A1438G A2706G A4769G C7028T A8860G G11719A C14766T A15326G), then this particular sequence would be simply cited as @A73G –309.C @A2706G @C7028T @G11719A @C14766T T16362C T16519C relative to the R root. This down-up profile notation could help to represent profile listings of West Eurasian mtDNA data in an optimally compact form.

A distinction should be made between variant and mutation: the emphasis of the former is on the actual nucleotide state in a particular sequence, whereas the latter highlights the change of nucleotides. *Mutations* are thus allocated to links of an mtDNA tree, whereas the *variants* (at the corresponding positions) label the nodes. To express that a sequence has a certain nucleotide, say, at position 10398, we could write that this sequence bears the rCRS-variant A10398 (as for example, Wong *et al.*[96] did) or in the other case that it bears the variant 10398G (as do virtually all haplogroup M lineages). That is, the nucleotide gets either prefixed or suffixed depending on the corresponding nucleotide in the rCRS (Figure 3). This is more compact than the *ad hoc* notation A10398G(G) and A4769G(A) (for expressing 10398G and A4768) used in Table 1 of Mosquera-Miguel *et al.*[64] For mtDNA sequences, (nucleotide) *variant* is a more appropriate term than *allele*, which would be better used on the amino-acid level. In the preceding example, one would use the circumfixed A10398G (or just 10398 as its short form) to label the links in the mtDNA tree where A changes to G, such as, for example, between the roots of haplogroups JT and J.

Although the standard reference-variant notation will serve most purposes for the analysis of human mtDNA, there is some need for a root-upwards notation. For instance, if one focuses on a particular position within the entire mtDNA phylogeny or a large haplogroup, the root-upwards notation directly shows which variant at that

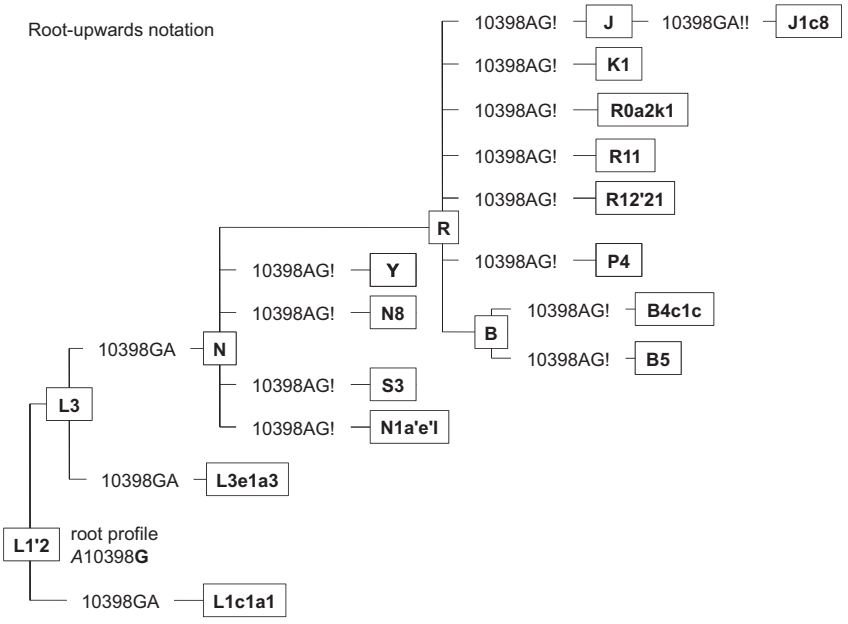

Figure 5 A view on the part of the basal worldwide mtDNA phylogeny in which position 10398 is varied. Mutations are scored in evolutionary direction following the double-suffix style used by Bermisheva *et al.*[97] The profile at the operating root is scored relative to the rCRS; haplogroups are denoted as in Build 15 of PhyloTree.

position is seen in which parts of the mtDNA tree. The operating root is not necessarily the L root but could be any descendent haplotype ancestral for the mtDNA lineages under study. If, say, haplogroups J and T are targeted, then the closest possible operating root is the JT root. This particular root would come along with its profile relative to rCRS. In order to distinguish the reference-variant notation for profiles from actual changes in evolutionary direction, we recommend using the double-suffix style pioneered by Bermisheva *et al.*[97] When, for instance, at a link of the mtDNA tree position 10398 changes from A to G, this is expressed by '10398AG' and when the reverse change is observed at another link this gets labelled by '10398GA'. As with the rCRS-variant style, one could also use a shorthand here by suppressing all first suffixes and the second suffixes unless they refer to transversions or indels. The distinction between the rCRS-variant style and roots upwards notation would then become visible only at positions 315.1, 523–524 and 825, so that a tree diagram in shorthand notation could easily be misread if the reader was not informed about the style used.

For highlighting subsequent returns to a nucleotide seen just before on the ascending pathway from the operating root, an extra suffix '!' gets added, just as in Builds 14 and 15 of PhyloTree; see Figure 5 for an illustration regarding position 10398. According to Behar *et al.*,[3] 'an exclamation mark (!) at the end of a labelled position denotes a reversion to the ancestral state'. This, however, appears to be somewhat misleading since '*the* ancestral state' cannot refer to the RSRS (or any more likely globally ancestral sequence), since then a double exclamation mark could not occur because once the ancestral nucleotide is reached, any subsequent mutation necessarily establishes a variant nucleotide relative to the globally ancestral sequence. The updated definition in Build 15 of PhyloTree is also imprecise: 'back mutations to an ancestral state are indicated with an exclamation mark (!), two exclamation marks for a double back mutation (!!) and so on'. Here, it is not exactly clear what '*an* ancestral state' means but obviously the *directly* (and not any earlier) ancestral state must be referred to.

In order to avoid potential confusion between rCRS-variant and root-upwards notation, it could be helpful to distinguish the rCRS nucleotide from the operating root nucleotide when presenting diagrams or tables: the former nucleotide could be set in italics and the latter in bold, and if coincident in bold italics; see Figure 4. This could remind the reader about the distinction of root and reference sequence and get the profile as well as the evolutionary direction of nucleotide changes right. In any case, it is important to be as explicit as possible about the way mutations have to be read and understood in a table or a tree or network diagram—something that was all too often rather hidden away in past publications.

CONCLUSION

The uses of a reference sequence, such as the rCRS, and some presumed global ancestral sequence, such as an estimate of the L root, are different since they serve different purposes. A replacement of the rCRS by a sequence close to the L root, such as the RSRS, or some other sequence, will in most circumstances hardly have any merit. On the contrary, it will complicate communication and lead to problems in medical and forensic casework as existing protocols and in-house databases and so on will all have to be transformed, in part even manually, and yet the published record will continue to be at odds with the transformed profiles. The field of molecular anthropology would be somewhat less affected by a switching of reference sequences because most 'big' studies handle their analyses *in silico*, and the change of reference sequence can be made relatively easily depending on any particular requirement.

Hence, there is no advantage in using a reference sequence based on a reconstructed root of the human phylogenetic tree, and we anticipate a broad consensus that any replacement of the rCRS, with its >30 years tradition, is unwarranted. Even worse is the availability of more than one reference sequence at the same time, with users being able to choose between them *ad libitum*. The PhyloTree project (www.phylotree.org) has become an important reference Web tool for mtDNA scholars during the last few years; however, the last version, Build 14, was first generated exclusively with the RSRS as the new reference sequence (on 5 April 2012), and only later (May 30) accompanied by the rCRS-based version offered 'for those used to the 'old' notation style'. The circumfixing of mutations in evolutionary direction as given by the primary version of Build 14 directly conflicts with the convention of medical genetics where either circumfixing with the rCRS nucleotide in front or strict post-fixing with the > notation has been employed. If Web projects such as the university-based PhyloTree and the new, commercial mtDNA Community (www.mtdnacommunity.org) continue to follow this radical change of notation rather than serve the needs of medical and forensic genetics on the long run, then a parallel Web project guaranteeing the standard notation should be developed to prevent many years of confusion.

We certainly agree with Behar *et al.*[3] that mtDNA nomenclature needs further standardization; however, the past has shown that this goal cannot be accomplished—and will indeed be undermined—by switching reference sequences. In contrast, the use of the rCRS should be solidified and accompanied by unambiguous notation, where a nucleotide position would always be prefixed by the variant in the rCRS and suffixed by the variant nucleotide in a lineage under consideration. Some traditional short cuts are still feasible, for example, deleting parts of the circumfixes. When the mutational process is being studied, mutations along a tree could well be scored in a root-upwards way by showing the mutations in evolutionary direction with two suffixes, provided that the deepest root in the respective tree is presented by its profile relative to the rCRS. This root profile should be either shown in the figure or stated in the corresponding legend. No matter which representation is chosen, it needs some simple steps in order to translate the tree information into the mtDNA profile of the tree nodes. These steps can either be executed by eye in the case of small trees or with bioinformatic tools for larger trees representing major portions of the entire mtDNA database.

The way in which the goals of the Behar *et al.*[3] article have been formulated and promoted has further fostered common misconceptions about the use and role of an mtDNA reference sequence. The reconstruction of ancestral sequences, as they have commonly been used for the estimation of haplogroup ages and positional mutation rates[53] is not influenced by the preferred reference sequence. As we have made clear above, there is no way to arrive at a unique reconstruction of a global root because of the inherent ambiguity. The estimation of mutation rates and haplogroup ages will, however, depend only marginally on a particular choice of a global ancestor from any of the possible candidates. Moreover, the choice of a reference sequence for reporting human mtDNA variation (including heteroplasmies) has nothing to do with the approaches of Next Generation Sequencing, no matter whether SNPs or (nearly) complete mtDNA sequences are being produced. If an accompanying bioinformatics tool outputs variations relative to a reference sequence different from rCRS, then an additional automatic tool is needed that translates this output into the standard rCRS-based motif lists before this information is reported in an article. In principle, alignment should not depend on any reference sequence for documentation. Unfortunately, alignment to a single sequence (whether ancestral or

extant) would ignore the well-known fact that alignment and phylogeny estimation cannot be separated as independent and subsequent tasks.[34,98] It remains to be investigated to what extent the current Next Generation Sequencing tools indeed violate this principle and hence may give suboptimal results.

In summary, the past has clearly shown that any manipulation of the reference sequence leads to innumerable mistakes and misinterpretations. The problems due to misunderstandings about human mtDNA in the context of new sequencing approaches for mtDNA are vexing enough, and a further change of the reference sequence is the very last thing medical genetics needs. Once a reference sequence has been accepted by the vast majority of mtDNA researchers and been in use for many years, it should not be changed. The rCRS will be retained as the unique reference point in forensic genetic research[99] and also in medical genetics as guided by MITOMAP: '*Douglas Wallace, Marie Lott and the MITOMAP team strongly support the retention of the rCRS as the human mtDNA reference sequence*' (private communication). We hope that the notational standardization for reporting mtDNA variation along the lines we have indicated is flexible enough to meet different goals and may help to guard against misunderstandings.

ACKNOWLEDGEMENTS

We thank Antonio Salas and Walther Parson for their help and support. Y-GY was supported by the National Science Foundation of China (30925021), Top Talents Program of Yunnan Province (2009CI119) and the Chinese Academy of Sciences. We are grateful to Marie Lott and Douglas Wallace for providing information about the case of the 'Yoruban mtDNA reference'.

1 Anderson, S., Bankier, A. T., Barrell, B. G., Debruijn, M. H. L., Coulson, A. R., Drouin, J. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).
2 Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23,** 147–147 (1999).
3 Behar, D. M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E. L., Silva, N. M. *et al.* A 'Copernican' reassessment of the human mitochondrial DNA tree from its toot. *Am. J. Hum. Genet.* **90,** 675–684 (2012).
4 Kivisild, T., Shen, P. D., Wall, D. P., Do, B., Sung, R., Davis, K. *et al.* The role of selection in the evolution of human mitochondrial genomes. *Genetics* **172,** 373–387 (2006).
5 Kincaid, R. K. in *Foundations of Location Analysis* (eds Eiselt, H. A. & Marianov, V.) Ch. 14, 315–334 (Springer, 2011).
6 Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408,** 708–713 (2000).
7 Bandelt, H.-J., Kong, Q.-P., Richards, M. & Macaulay, V. in *Human Mitochondrial DNA and the Evolution of Homo sapiens* (eds Bandelt, H.-J., Macaulay, V. & Richards, M.) 47–90 (Springer-Verlag, Berlin-Heidelberg, 2006).
8 Koekemoer, M. Construction of a mitochondrial consensus sequence for the Khoi-San population of Southern Africa. PhD Thesis. North-West University, Potchefstroom Campus (2010).
9 Richards, M. B., Macaulay, V. A., Bandelt, H.-J. & Sykes, B. C. Phylogeography of mitochondrial DNA in western Europe. *Ann. Hum. Genet.* **62,** 241–260 (1998).
10 Torroni, A., Achilli, A., Macaulay, V., Richards, M. & Bandelt, H.-J. Harvesting the fruit of the human mtDNA tree. *Trends Genet.* **22,** 339–345 (2006).
11 van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30,** E386–E394 (2009).
12 Bandelt, H.-J., Yao, Y.-G., Bravi, C. M., Salas, A. & Kivisild, T. Median network analysis of defectively sequenced entire mitochondrial genomes from early and contemporary disease studies. *J. Hum. Genet.* **54,** 174–181 (2009).
13 Macaulay, V., Richards, M. & Sykes, B. Mitochondrial DNA recombination—no need to panic. *Proc. Biol. Sci.* **266,** 2037–2039 (1999).
14 Ozawa, T. Mitochondrial cardiomyopathy. *Herz* **19,** 105–118 (1994).
15 Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. & Takahata, N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl Acad. Sci. USA* **92,** 532–536 (1995).
16 Marzuki, S., Noer, A. S., Lertrit, P., Thyagarajan, D., Kapsa, R., Utthanaphol, P. *et al.* Normal variants of human mitochondrial DNA and translation products: the building of a reference data base. *Hum. Genet.* **88,** 139–145 (1991).
17 Vives-Bauza, C., Andreu, A. L., Manfredi, G., Beal, M. F., Janetzky, B., Gruenewald, T. H. *et al.* Sequence analysis of the entire mitochondrial genome in Parkinson's disease. *Biochem. Biophys. Res. Commun.* **290,** 1593–1601 (2002).
18 Komaki, H., Akanuma, J., Iwata, H., Takahashi, T., Mashima, Y., Nonaka, I. *et al.* A novel mtDNA C11777A mutation in Leigh syndrome. *Mitochondrion* **2,** 293–304 (2003).
19 Maruszak, A., Gaweda-Walerych, K., Sołtyszewski, I. & Kekanowski, C. Mitochondrial DNA in pathogenesis of Alzheimer's and Parkinson's diseases. *Acta Neurobiol. Exp.* **66,** 153–176 (2006).
20 Martorell, L., Segués, T., Folch, G., Valero, J., Joven, J., Labad, A. *et al.* New variants in the mitochondrial genomes of schizophrenic patients.. *Eur. J. Hum. Genet.* **14,** 520–528 (2006).
21 Bandelt, H.-J., Olivieri, A., Bravi, C., Yao, Y.-G., Torroni, A. & Salas, A. 'Distorted' mitochondrial DNA sequences in schizophrenic patients. *Eur. J. Hum. Genet.* **15,** 400–402 (2007).
22 Martorell, L. & Vilella, E. 'Distorted' mitochondrial DNA sequences in schizophrenic patients—Reply to Bandelt *et al.*. *Eur. Hum. Genet.* **15,** 402–404 (2007).
23 Bandelt, H.-J., Salas, A., Taylor, R. W. & Yao, Y.-G. Exaggerated status of 'novel' and 'pathogenic' mtDNA sequence variants due to inadequate database searches. *Hum. Mutat.* **30,** 191–196 (2009).
24 Taylor, R. W., Jobling, M. S., Turnbull, D. M. & Chinnery, P. F. Frequency of rare mitochondrial DNA mutations in patients with suspected Leber's hereditary optic neuropathy. *J. Med. Genet.* **40,** e85 (2003).
25 Aikhionbare, F. O., Khan, M., Carey, D., Okoli, J. & Go, R. Is cumulative frequency of mitochondrial DNA variants a biomarker for colorectal tumor progression? *Mol. Cancer* **3** (2004).
26 Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E. *et al.* The dawn of human matrilineal diversity. *Am. J. Hum. Genet.* **82,** 1130–1140 (2008).
27 Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B. N., Urade, B. P., Mallick, S. *et al.* Updating phylogeny of mitochondrial DNA macrohaplogroup M in India: dispersal of modern human in south Asian corridor. *PLoS One* **4,** e7447 (2009).
28 Gonder, M. K., Mortensen, H. M., Reed, F. A., de Sousa, A. & Tishkoff, S. A. Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* **24,** 757–768 (2007).
29 Pala, M., Olivieri, A., Achilli, A., Accetturo, M., Metspalu, E., Reidla, M. *et al.* Mitochondrial DNA signals of late glacial recolonization of Europe from Near Eastern refugia. *Am. J. Hum. Genet.* **90,** 915–924 (2012).
30 Houshmand, M., Montazeri, M., Kuchekian, N., Noohi, F., Nozar, G. & Zamani, A. Is 8860 variation a rare polymorphism or associated as a secondary effect in HCM disease? *Arch. Med. Sci.* **7,** 242–246 (2011).
31 Barbieri, C., Whitten, M., Beyer, K., Schreiber, H., Li, M. K. & Pakendorf, B. Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. *Mol. Biol. Evol.* **29,** 1213–1223 (2012).
32 Wallace, D. C. & Mitochondrial, D. N. A. sequence variation in human evolution and disease. *Proc. Natl Acad. Sci. USA* **91,** 8739–8746 (1994).
33 Salas, A., Bandelt, H.-J., Macaulay, V. & Richards, M. B. Phylogeographic investigations: the role of trees in forensic genetics. *Forensic Sci. Int.* **168,** 1–13 (2007).
34 Bandelt, H.-J. & Parson, W. Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int. J. Legal Med.* **122,** 11–21 (2008).
35 Prezant, T. R., Agapian, J. V. & Fischelghodsian, N. Corrections to the human mitochondrial ribosomal RNA sequences. *Hum. Genet.* **93,** 87–88 (1994).
36 Zeller, M., Mirghomizadeh, F., Wehner, H. D. & Blin, N. Frequent D-loop polymorphism in mtDNA enables genotyping of 1400-year-old human remains from Merowingian graves. *J. Appl. Genet.* **41,** 285–292 (2000).
37 Sternberg, D., Chatzoglou, E., Laforêt, P., Fayet, G., Jardel, C., Blondy, P. *et al.* Mitochondrial DNA transfer RNA gene sequence variations in patients with mitochondrial disorders. *Brain* **124,** 984–994 (2001).
38 Zarrouk Mahjoub, S., Mehri, S., Ourda, F., Boussaada, R., Mechmeche, R., Ben Arab, S. *et al.* Transition m.3308T>C in the ND1 gene is associated with left ventricular hypertrabeculation/noncompaction. *Cardiology* **118,** 153–158 (2011).
39 Salas, A. & Elson, J. L. Raising doubts about the pathogenicity of mitochondrial DNA mutation m.3308T>C in left ventricular hypertraveculation/noncompaction. *Cardiology* **122,** 113–115 (2012).
40 Lu, Z. Q., Chen, H., Meng, Y. Z., Wang, Y., Xue, L., Zhi, S. C. *et al.* The tRNAMet 4435A>G mutation in the mitochondrial haplogroup G2a1 is responsible for maternally inherited hypertension in a Chinese pedigree. *Eur. J. Hum. Genet.* **19,** 1181–1186 (2011).
41 Yan, X. K., Wang, X. J., Wang, Z. M., Sun, S., Chen, G. L., He, Y. Z. *et al.* Maternally transmitted late-onset non-syndromic deafness is associated with the novel heteroplasmic T12201C mutation in the mitochondrial tRNAHis gene. *J. Med. Genet.* **48,** 682–690 (2011).
42 Zhang, R. X., Zhang, F. B., Wang, C. J., Wang, S. X., Shiao, Y. H. & Guo, Z. J. Identification of sequence polymorphism in the D-Loop region of mitochondrial DNA as a risk factor for hepatocellular carcinoma with distinct etiology. *J. Exp. Clin. Cancer Res.* **29,** 155 (2010).
43 Carter, R. W. Mitochondrial diversity within modern human populations. *Nucleic Acids Res.* **35,** 3039–3045 (2007).
44 Carter, R. W., Criswell, D. & Sanford, J. in *Proceedings of the Sixth International Conference on Creationism* (eds Snelling, A. A.) 111–116 (TX: Institute for Creation Research, Pittsburgh, PA, USA, Creation Science Fellowship and Dallas, 2008).
45 Güney, A. I., Javadova, D., Kirac, D., Ulucan, K., Koc, G., Ergec, D. *et al.* Detection of Y chromosome microdeletions and mitochondrial DNA mutations in male infertility patients. *Genet. Mol. Res.* **11,** 1039–1048 (2012).

46 Roshan, M., Kabekkodu, S. P., Vijaya, P. H., Manjunath, K., Graw, J., Gopinath, P. *et al.* Analysis of mitochondrial DNA variations in Indian patients with congenital cataract. *Mol. Vis.* **18,** 181–193 (2012).
47 Kumar, R., Bhat, A., Bamezai, R. N. K., Shamsi, M. B., Kumar, R., Gupta, N. P. *et al.* Necessity of nuclear and mitochondrial genome analysis prior to assisted reproductive techniques/intracytoplasmic sperm injection. *Indian J. Biochem. Biophys.* **44,** 437–442 (2007).
48 Kumar, R., Venkatesh, S., Kumar, M., Tanwar, M., Shasmsi, M. B., Kumar, R. *et al.* Oxidative stress and sperm mitochondrial DNA mutation in idiopathic oligoasthenozoospermic men. *Indian J. Biochem. Biophys.* **46,** 172–177 (2009).
49 Palanichamy, m. G. & Zhang, Y.-P. Identifying potential pitfalls in interpreting mitochondrial DNA mutations of male infertility cases. *Indian J. Med. Res.* **134,** 447–451 (2011).
50 Yao, Y.-G., Macaulay, V., Kivisild, T., Zhang, Y.-P. & Bandelt, H.-J. To trust or not to trust an idiosyncratic mitochondrial data set. *Am. J. Hum. Genet.* **72,** 1341–1346 (2003).
51 Yao, Y.-G., Salas, A., Logan, I. & Bandelt, H.-J. mtDNA Data mining in GenBank needs surveying. *Am. J. Hum. Genet.* **85,** 929–933 (2009).
52 Bandelt, H. J., van Oven, M. & Salas, A. Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. *Int. J. Legal Med.* **126,** 901–916 (2012).
53 Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A. *et al.* Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84,** 740–759 (2009).
54 Logan, I. A suggested genome for 'Mitochondrial Eve'. *J. Genet. Geneal.* **3,** 72–77 (2007).
55 Cariaso, M. & Lennon, G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.* **40,** D1308–D1312 (2012).
56 Thieme, M., Lottaz, C., Niederstätter, H., Parson, W., Spang, R. & Oefner, P. J. ReseqChip: automated integration of multiple local context probe data from the MitoChip array in mitochondrial DNA sequence assembly. *BMC Bioinformatics* **10,** 440 (2009).
57 Vallone, P. M., Jakupciak, J. P. & Coble, M. D. Forensic application of the affymetrix human mitochondrial resequencing array. *Forensic Sci. Int. Genet.* **1,** 196–198 (2007).
58 Rollins, B., Martin, M. V., Sequeira, P. A., Moon, E. A., Morgan, L. Z., Watson, S. J. *et al.* Mitochondrial variants in schizophrenia, bipolar disorder, and major depressive disorder. *PLoS One* **4,** e4913 (2009).
59 Hartmann, A., Thieme, M., Nanduri, L. K., Stempfl, T., Moehle, C., Kivisild, T. *et al.* Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum. Mutat.* **30,** 115–122 (2009).
60 Xie, H. B. M., Perin, J. C., Schurr, T. G., Dulik, M. C., Zhadanov, S. I., Baur, J. A. *et al.* Mitochondrial genome sequence analysis: A custom bioinformatics pipeline substantially improves Affymetrix MitoChip v2.0 call rate and accuracy. *BMC Bioinformatics* **12,** 402 (2011).
61 Zamzami, M. A., Price, G. R., Taylor, R. W., Blakely, E. L., Oancea, I., Bowling, F. *et al.* Insights into N-calls of mitochondrial DNA sequencing using MitoChip v2.0. *BMC Res. Notes* **4,** 426 (2011).
62 Ueno, H., Nishigaki, Y., Kong, Q.-P., Fuku, N., Kojima, S., Iwata, N. *et al.* Analysis of mitochondrial DNA variants in Japanese patients with schizophrenia. *Mitochondrion* **9,** 385–393 (2009).
63 Bertolin, C., Magri, C., Barlati, S., Vettori, A., Perini, G. I., Peruzzi, P. *et al.* Analysis of complete mitochondrial genomes of patients with schizophrenia and bipolar disorder. *J. Hum. Genet.* **56,** 869–872 (2011).
64 Mosquera-Miguel, A., Torrell, H., Abasolo, N., Arrojo, M., Paz, E., Ramos-Ríos, R. *et al.* No evidence that major mtDNA European haplogroups confer risk to schizophrenia. *Am. J. Med. Genet. Part B.* **159B,** 414–421 (2012).
65 Sequeira, A., Martin, M. V., Rollins, B., Moon, E. A., Bunney, W. E., Macciardi, F. *et al.* Mitochondrial mutations and polymorphisms in psychiatric disorders. *Front. Genet.* **3,** 103 (2012).
66 Scaglia, F. The role of mitochondrial fysfunction in psychiatric disease. *Dev. Disabil. Res. Rev.* **16,** 136–143 (2010).
67 Kaipparettu, B. A., Ma, Y. W. & Wong, L.-J. C. Functional effects of cancer mitochondria on energy metabolism and tumorigenesis: utility of transmitochondrial cybrids. *Ann. N. Y. Acad. Sci.* **1201,** 137–146 (2010).
68 Gil Borlado, M. C., Moreno Lastres, D., Gonzalez Hoyuela, M., Moran, M., Blazquez, A., Pello, R. *et al.* Impact of the mitochondrial henetic background in Complex III deficiency. *PLoS One* **5,** e12801 (2010).
69 Bi, R., Zhang, A.-M., Zhang, W., Kong, Q.-P., Wu, B.-L., Yang, X.-H. *et al.* The acquisition of an inheritable 50-bp deletion in the human mtDNA control region does not affect the mtDNA copy number in peripheral blood cells. *Hum. Mutat.* **31,** 538–543 (2010).
70 Canter, J. A., Kallianpur, A. R., Parl, F. F. & Millikan, R. C. Mitochondrial DNA G10398A polymorphism and invasive breast cancer in African-American women. *Cancer Res.* **65,** 8028–8033 (2005).
71 Mims, M. P., Hayes, T. G., Zheng, S. Y., Leal, S. M., Frolov, A., Ittmann, M. M. *et al.* Mitochondrial DNA G10398A polymorphism and invasive breast cancer in African-American women. *Cancer Res.* **66,** 1880–1880 (2006).
72 Bhat, A., Koul, A., Sharma, S., Rai, E., Bukhari, S. I. A., Dhar, M. K. *et al.* The possible role of 10398A and 16189C mtDNA variants in providing susceptibility toT2DM in two North Indian populations: a replicative study. *Hum. Genet.* **120,** 821–826 (2007).
73 Darvishi, K., Sharma, S., Bhat, A. K., Rai, E. & Bamezai, R. N. K. Mitochondrial DNA G10398A polymorphism imparts maternal haplogroup N a risk for breast and esophageal cancer. *Cancer Lett.* **249,** 249–255 (2007).
74 Setiawan, V. W., Chu, L. H., John, E. M., Ding, Y. C., Ingles, S. A., Bernstein, L. *et al.* Mitochondrial DNA G10398A variant is not associated with breast cancer in African-American women. *Cancer Genet. Cytogenet.* **181,** 16–19 (2008).
75 Velez, D. R., Menon, R., Simhan, H., Fortunato, S., Canter, J. A. & Williams, S. M. Mitochondrial DNA variant A4917G, smoking and spontaneous preterm birth. *Mitochondrion* **8,** 130–135 (2008).
76 Kulawiec, M., Owens, K. M. & Singh, K. K. mtDNA G10398A variant in African-American women with breast cancer provides resistance to apoptosis and promotes metastasis in mice. *J. Hum. Genet.* **54,** 647–654 (2009).
77 Rohan, T. E., Wong, L.-J., Wang, T., Haines, J. & Kabat, G. C. Do alterations in mitochondrial DNA play a role in breast carcinogenesis? J. *Oncology* **2010,** 604304 (2010).
78 Tanwar, M., Dada, T., Sihota, R. & Dada, R. Mitochondrial DNA analysis in primary congenital glaucoma. *Mol. Vis.* **16,** 518–533 (2010).
79 Sultana, G. N. N., Rahman, A., Karim, M. M., Shahinuzzaman, A. D. A., Begum, R. & Begum, R. A. Breast cancer risk associated mitochondrial NADH-dehydrogenase subunit-3 (ND3) polymorphisms (G10398A and T10400C) in Bangladeshi women. *J. Med. Genet. Genomics* **3,** 131–135 (2011).
80 Kuhnt, T. Optimierung der Strahlentherapie von Tumoren im Kopf-Hals-Bereich: Biologische und technische Entwicklungen. *Habilitationsschrift. Universität Halle* (http:/sundoc.bibliothek.uni-halle.de/habil-online/08/08H043/habil.pdf) (2008)
81 Holyoake, A. J., McHugh, P., Wu, M., O'Carroll, S., Benny, P., Sin, I. L. *et al.* High incidence of single nucleotide substitutions in the mitochondrial genome is associated with poor semen parameters in men. *Int. J. Androl.* **24,** 175–182 (2001).
82 Bandelt, H.-J. Misanalysis gave false association of mtDNA mutations with infertility. *Int. J. Androl.* **31,** 450–453 (2008).
83 Aikhionbare, F. O., Mehrabi, S., Thompson, W., Yao, X., Grizzle, W. & Partridge, E. mtDNA sequence variants in subtypes of epithelial ovarian cancer stages in relation to ethnic and age difference. *Diagn. Pathol.* **3,** 32 (2008).
84 Hudson, G., Carelli, V., Spruijt, L., Gerards, M., Mowbray, C., Achilli, A. *et al.* Clinical expression of Leber hereditary optic neuropathy is affected by the mitochondrial DNA-haplogroup background. *Am. J. Hum. Genet.* **81,** 228–233 (2007).
85 Torroni, A., Campos, Y., Rengo, C., Sellitto, D., Achilli, A., Magri, C. *et al.* Mitochondrial DNA haplogroups do not play a role in the variable phenotypic presentation of the A3243G mutation. *Am. J. Hum. Genet.* **72,** 1005–1012 (2003).
86 Behar, D. M., Rosset, S., Blue-Smith, J., Balanovsky, O., Tzur, S., Comas, D. *et al.* The genographic project public participation mitochondrial DNA database. *PLoS Genet.* **3,** 1083–1095 (2007).
87 van Oven, M., Vermeulen, M. & Kayser, M. Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution. *Invest. Genet.* **2,** 6 (2011).
88 Achilli, A., Olivieri, A., Pala, M., Kashani, B. H., Carossa, V., Perego, U. A. *et al.* Mitochondrial DNA backgrounds might modulate diabetes complications rather than T2DM as a whole. *PLoS One* **6,** e21029 (2011).
89 Yu, D., Jia, X., Zhang, A.-M., Li, S., Zou, Y., Zhang, Q. *et al.* Mitochondrial DNA sequence variation and haplogroup distribution in Chinese patients with LHON and m.14484T>C. *PLoS One* **5,** e13426 (2010).
90 Niemi, A. K., Moilanen, J. S., Tanaka, M., Hervonen, A., Hurme, M., Lehtimäki, T. *et al.* A combination of three common inherited mitochondrial DNA polymorphisms promotes longevity in Finnish and Japanese subjects. *Eur. J. Hum. Genet.* **13,** 166–170 (2005).
91 Bayona-Bafaluy, M. P., López-Gallardo, E., Montoya, J. & Ruiz-Pesini, E. Maternally inherited susceptibility to cancer. *Bioenergetics* **1807,** 643–649 (2011).
92 Shin, M. G., Kajigaya, S., Levin, B. C. & Young, N. S. Mitochondrial DNA mutations in patients with myelodysplastic syndromes. *Blood* **101,** 3118–3125 (2003).
93 Rosson, D. & Keshgegian, A. A. Frequent mutations in the mitochondrial control region DNA in breast tissue. *Cancer Lett.* **215,** 89–94 (2004).
94 Kivisild, T., Tolk, H.-V., Parik, J., Wang, Y. M., Papiha, S. S., Bandelt, H.-J. *et al.* The emerging limbs and twigs of the East Asian mtDNA tree. *Mol. Biol. Evol.* **19,** 1737–1751 (2002).
95 Kong, Q.-P., Bandelt, H.-J., Sun, C., Yao, Y.-G., Salas, A., Achilli, A. *et al.* Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. *Hum. Mol. Genet.* **15,** 2076–2086 (2006).
96 Wong, L.-J. C. & Senadheera, D. Direct detection of multiple point mutations in mitochondrial DNA. *Clin. Chem.* **43,** 1857–1861 (1997).
97 Bermisheva, M. A., Kutuev, I. A., Korshunova, T. Y., Dubova, N. A., Villems, R. & Khusnutdinova, E. K. Phylogeographic analysis of mitochondrial DNA in the Nogays: a strong mixture of maternal lineages from eastern and western Eurasia. *Mol. Biol.* **38,** 516–523 (2004).
98 Parson, W. & Bandelt, H.-J. Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Sci. Int. Genet.* **1,** 13–19 (2007).
99 Salas, A., Coble, M., Desmyter, S., Grzybowski, T., Gusmão, L., Hohoff, C. *et al.* A cautionary note on switching mitochondrial DNA reference sequences in forensic genetics. *Forensic Sci. Int. Genet.* **6,** E182–E184 (2012).
100 Bandelt, H.-J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16,** 37–48 (1999).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

Table S1. Occurrences of 3107N in GenBank

First author of submission	Example
Gonder	EF184581
Hartmann	EU597486
Bhat	EU872033
Pello	FJ178379
Ricault	FJ543101
Eaaswarkhanth	FJ157838
Rao2009	FJ383174
Razafindrazaka	FJ767910
D'Aurelio	GQ891609
Malhi	GU215075
GeneTree	GU722598
Harich	GU445415
Maksum	HM436814
Ji	FJ748704
Jardel	GQ404809
Gil-Borlado	HM046248
Elango	HM750218
Gomez-Duran	HM103354
Gomez-Duran	JN635298
Sharma	GU480001
Illeperuma	HQ153103
Pena	HQ268504
Delgado-Sanchez	JF318958
Ngili	JN048455
Schuenemann	HE576978
Barbieri	JQ044904
Fernandes	JQ245723
Scholes	JF739535
Pala	JQ797760
Perego	JQ965976
Martinez-Romero	JX401416
Reiff	HQ700839
Trejaut	JX987462

Table S2. Contrasting notation for most common control-region haplotypes in three population data sets recently added to EMPOP[1)]

Population	Size	Code	Control-region profile relative to the rCRS[2)]	Control-region profile relative to the RSRS[2)]
Germany	223	EMP00514	*T*16519**C** *A*263**G** 315.1**C**	**A**16129*G* **T**16187*C* **C**16189*T* **T**16223*C* **G**16230*A* **T**16278*C* **C**16311*T* **G**73*A* **C**146*T* **C**152*T* **C**195*T* **A**247*G* **del**523A **del**524C
Laos	214	EMP00083	*G*16129**A** ***A***16162G ***T***16172C ***T***16304C ***A***16399G *T*16519**C** *A*73**G** ***A***249del *A*263**G** 315.1**C** *A*523**del** *C*524**del**	***A***16162G ***T***16172C **T**16187*C* **C**16189*T* **T**16223*C* **G**16230*A* **T**16278*C* ***T***16304C **C**16311*T* ***A***16399G **C**146*T* **C**152*T* **C**195*T* **A**247*G* ***A***249del
Ghana	191	EMP00081	***T***16126C *C*16187**T** *T*16189**C** *C*16223**T** ***C***16264T ***C***16270T *C*16278**T** ***A***16293G *T*16311**C** *T*16519**C** *A*73**G** *T*146**C** *T*152**C** ***C***182T ***G***185T *T*195**C** *G*247**A** *A*263**G** 315.1**C** ***A***357G *A*523**del** *C*524**del**	***T***16126C **A**16129*G* **G**16230*A* ***C***16264T ***C***16270T ***A***16293G ***C***182T ***G***185T ***A***357G

[1)] See http://empop.org/modules/publications/
[2)] The rCRS nucleotide is indicated as a prefix in italics and the L root nucleotide in bold

Table S3. Variable recording of the stretch 3104-3114 in data of Barbieri et al.

GenBank accession no.	Sequence range 3104-3114
rCRS	TA–CTTCAAAT
JQ045125.1	TACTTCAAAT
JQ044800.1	TACNTTCAAAT
JQ044853.1	TACTTNCAAAT
JQ044921.1	TACTTNNAAAT
JQ044923.1	TACTNNNAAAT
JQ044791.1	TACTTTCAAAT
JQ044816.1	TACNNTCAAAT
JQ044922.1	TANNTTCAAAT
JQ044962.1	TACTTNNNNT
JQ044966.1	TACTTNNNAT
JQ044872.1	TACNNNNNNAT
JQ044888.1	TACNNNNNAAT
JQ045022.1	TANNNNNAAAT
JQ044904.1	TNNNNNNAAAT
JQ045046.1	TNNNNNNNAAT

Table S4. Profile of the estimated human ancestral sequence (L root)[1)] relative to rCRS based on Figure 2d

A73G T146C T152C T195C G247A A263G –315.1C A523– C524– A750G G769A T825A G1018A A1438G A2706G G2758A T2885C C3594T A4104G C4312T A4769G C7028T A7146G C7256T G7521A C8468T C8655T A8701G A8860G T9540C A10398G G10688A T10873C T10915C G11719A G11914A G12007A C12705T A13105G A13276G C13506T T10810C C13650T C14766T A15326G G16129A C16187T T16189C C16223T A16230G C16278T T16311C T16519C

[1)] Differences to RSRS at positions 315.1, 12007, and 10664