Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 333 (2005) 122–130
BBRC
www.elsevier.com/locate/ybbrc

Low "penetrance" of phylogenetic knowledge in mitochondrial disease studies

Hans-Jürgen Bandelt [a,*], Alessandro Achilli [b], Qing-Peng Kong [c], Antonio Salas [d], Sabine Lutz-Bonengel [e], Chang Sun [c], Ya-Ping Zhang [c], Antonio Torroni [b], Yong-Gang Yao [c]

[a] *Department of Mathematics, University of Hamburg, 20146 Hamburg, Germany*
[b] *Dipartimento di Genetica e Microbiologia, Università di Pavia, 27100 Pavia, Italy*
[c] *Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 650223 Kunming, Yunnan, China*
[d] *Unidad de Xenética, Instituto de Medicina Legal, Facultad de Medicina, Universidad de Santiago de Compostela, 15782, Centro Nacional de Xenotipado (CeGen), Hospital Clínico Universitario, 15706 Galicia, Spain*
[e] *Institute of Legal Medicine, Albert Ludwig University Freiburg, 79104 Freiburg, Germany*

Received 11 April 2005
Available online 20 April 2005

Abstract

An up-to-date view of the worldwide mitochondrial DNA (mtDNA) phylogeny together with an evaluation of the conservation of each site is a reliable tool for detecting errors in mtDNA studies and assessing the functional importance of alleged pathogenic mutations. However, most of the published studies on mitochondrial diseases make very little use of the phylogenetic knowledge that is currently available. This drawback has two inadvertent consequences: first, there is no sufficient a posteriori quality assessment of complete mtDNA sequencing efforts; and second, no feedback is provided for the general mtDNA database when apparently new mtDNA lineages are discovered. We demonstrate, by way of example, these issues by reanalysing three mtDNA sequencing attempts, two from Europe and another one from East Asia. To further validate our phylogenetic deductions, we completely sequenced two mtDNAs from healthy subjects that nearly match the mtDNAs of two patients, whose sequences gave problematic results.

Keywords: Mitochondrial DNA; Phylogeny; Mitochondrial disease; Phantom mutation; Documentation error; Mitochondrial encephaloneuromyopathy; Dilated cardiomyopathy; Nonsyndromic deafness

Much of the early understanding of human mtDNA variation in the coding region was owing to the field of medical genetics. Indeed, in the 1990s the only available complete mtDNA sequences were generated in connection with medical studies, although these early sequencing efforts had to be considered with some reservation because of obvious shortcomings [1,2]. When complete sequences accumulated, accompanying phylogenetic analyses were very rare and not fully satisfactory [3–5]. Phylogenetic approaches to mtDNA analysis, relying on the specific characteristics of maternal inheritance of mtDNA, are being exercised to some extent only in the fields of forensic and population genetics but hardly ever in purely medical studies. Worse, the growing body of >1700 published complete mtDNA sequences is notoriously ignored in medical research. A more fluid interplay between medical, forensic, and population genetics could improve the quality of complete mtDNA sequencing results and promote a better understanding of the molecular basis of mitochondrial disorders. In what follows, we critically examine three

* Corresponding author. Fax: +49 40 42838 5190.
E-mail address: bandelt@math.uni-hamburg.de (H.-J. Bandelt).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.04.055

mtDNA complete sequencing attempts that were recently published in this journal, and thereby we exemplify the stringent need of using the currently available phylogenetic knowledge for performing mitochondrial disease studies with better focus.

Conceptually, each mtDNA occupies a place in the worldwide mtDNA phylogenetic tree according to the observed mutations scored relative to the revised Cambridge reference sequence (rCRS) [6]. Because of the large amount of available sequence information both the backbone and younger branches of the world phylogeny are now known to some detail, so that every newly sequenced mtDNA should fit into one of the branches of the mtDNA tree. When an mtDNA sequence does not properly fit, then either it could represent a rare basal branch hitherto unobserved or alternative hypotheses should be taken into consideration. Since there is no biological process that would explain back mutations simultaneously occurring at several positions in the same mtDNA, artefacts would be the only explanation in such a case.

Sequencing of mtDNA and subsequent documentation do not always run smoothly and may result in sequences that deviate from the real mtDNA sequences at several positions. Among sequencing artefacts, the most common ones are contamination or sample mix-up events, which commonly lead to artificial recombinants [7–9], and documentation errors due to defective data handling and sequence documentation [10].

Material and methods

Since the mtDNAs in medical cases considered here are of European and East Asian ancestry, we refer to the up-to-date information provided for Europeans [11–13] and East Asians [14,15], supplemented by coding-region information [16,17]. For easier reference we employ the standard haplogroup nomenclature, which is routinely being updated and refined in order to reflect the current state of phylogenetic knowledge. For the sake of clarity, Fig. 1 displays the branches (haplogroups) of the Eurasian portion of the mtDNA tree that highlight the main errors in the publications commented below.

Not all of the mutations allocated to the branches of this tree are equally informative for classification purposes. For instance, length polymorphisms of polycytosine stretches (e.g., 309+C, 309+CC, or A16183C) and dinucleotide repeats (such as [522–523]del), or some extremely frequent substitutions (T152C, T16519C, for example) constitute mutational hotspots. Ancestral haplotypes cannot always be inferred unambiguously in regard to these hotspots, and therefore some of these mutations are not displayed in large mtDNA trees [11,14]. The hypervariable segments I and II (HVS-I and HVS-II) of the mtDNA control-region harbour several positions that are very frequently hit by transitions [18]. In Fig. 1, we highlighted potential oversights also at those hotspot positions in the control region, because there is clear evidence that part of the control region was systematically left unanalysed in those publications (see below). The coding region (577–16023, including a few intergenic non-coding nucleotides) also

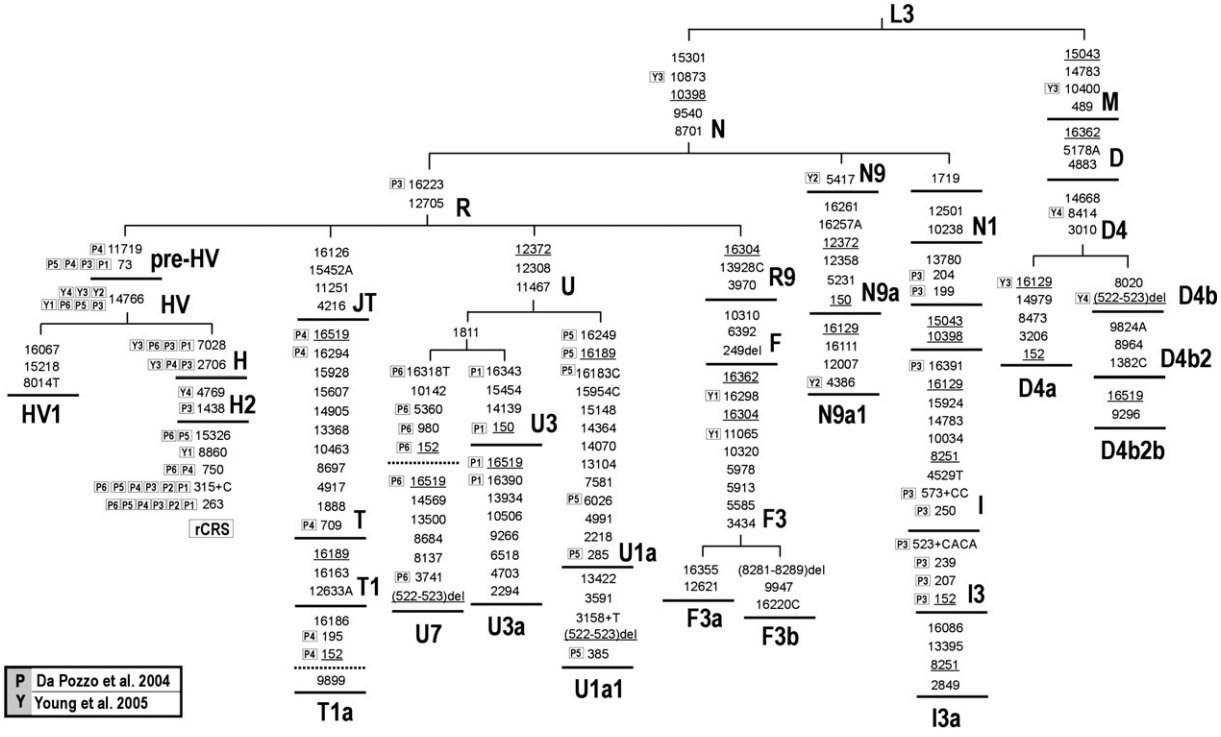

Fig. 1. Section of the European and East Asian mtDNA phylogeny highlighting subhaplogroups discussed in the text. Position numbers refer to the rCRS [6] and designate transitions unless a suffix indicates a transversion (A, C, or T), an insertion (+), or a deletion (del). Recurrent mutations are underlined. Squares to the left of nucleotide positions indicate potential oversights in published sequences, where inscribed P1, P2, P3, P4, P5, and P6 refer to Patient 1, 2, 3, 4, 5, and 6, respectively, from [22], whereas Y1, Y2, Y3, and Y4 refer to Pedigree 101, 102, 103, and 104, respectively, from [37].

has a handful of mutational hotspots (such as position 8251), as can be inferred from counting recurrent events in Eurasian mtDNA trees.

A compiled mtDNA database encompassing the profiles of more than 30,000 combined HVS-I/HVS-II sequences from the forensic and population genetics literature is employed for comparison. In addition, we make use of the SWGDAM database [19] with recent amendments [20,21] that responded in part to the alarming findings of [8,9].

The complete mtDNA genomes of two healthy donors (one Chinese and one Italian) were determined according to [11,12], respectively.

Results

Snapshot of Italian mtDNA variation

Da Pozzo et al. [22] claimed to have analysed the complete mitochondrial genome in six patients with mitochondrial encephaloneuromyopathies. The mtDNA lineages as recorded in their Table 1 all belong to familiar European mtDNA haplogroups, but lack an astonishing number of expected nucleotide variants. This becomes evident when one compares the listed polymorphisms with the phylogenetic tree displayed in (Fig. 1 of [11])—as we will now do, case by case.

Patient 1 (haplogroup U3a)

The majority of mutations for haplogroup U3a status are in place, except for C7028T and all mutations in the control region (such as the expected A73G, C150T, A263G, 315+C, A16343G, and G16390A). Hence, only the coding region was effectively sequenced. Of the three mtDNA mutations, G3423T, A1530G, and A5498G, observed in this patient that come on top of the U3a motif, the first one is not real because, by mistake, this refers to an erroneous base in the original reference sequence (CRS [23]) which is corrected in the revised Cambridge Reference Sequence (rCRS [6]). It is then useful to contrast the coding-region information of published U3a sequences with the patient's mtDNA; see Table 1, which in fact reveals a number of problems with the published record.

Table 1
Comparison of mtDNA coding-region sequences from haplogroup U3a

Source	Sample code	Coding region
[22]	P1	A1530G **A2294G** A5498G @7028
[24]	$U3_1$	G3010A C7256T T10724C G15734A
[24]	$U3_2$	C867T C7256T A10352G
[16,17]	#52	A723C **A2294G** G3010A @6518
[16,17]	#118	A723C **A2294G** G3010A C6496A A6716G
[16,17]	#271	**A2294G** G3010A T4703C T10724C G15734A
[13]	#5	G1719A **A2294G** T6050C A11050G T13215C

Mutations are listed with respect to the putative root of haplogroup U3a according to [11], which is determined by the coding-region mutations A750G, A1438G, A2706G, T4703C, A4769G, T6518C, C7028T, A8860G, G9266A, A10506G, A11467G, G11719A, A12308G, G12372A, C13934T, A14139G, C14766T, A15326G, and T15454C. Mutation A2294G (in bold) likely belongs to the motif of haplogroup U3a as well. Positions with back mutations to the corresponding rCRS nucleotide are prefixed with @.

We suggest to resolve the rather confusing picture of the variation within haplogroup U3a as follows. Inasmuch as the data of [24] are most unreliable (see [11]), it is realistic to assume that A2294G was simply overlooked in both $U3_1$ and $U3_2$ (Table 1). Moreover, the three mutations G3010A, T10724C, and G15734A were probably missed in $U3_2$ as well. It is then also possible that C6518T might have been overlooked in #52. In any case, we should revise the U3a motif proposed in [11] by including the mutation A2294G.

Patient 2 (haplogroup HV1)

This mtDNA is also recorded with the false G3423T. The mutations falling into the coding region and HVS-I clearly indicate haplogroup HV1 membership [12]. An exact match for HVS-I is found in USA.CAU.000157 from the SWGDAM database [19]. It thus becomes evident that the patient's mtDNA lacks all mutations (at least A263G, 315+C, and perhaps 249del plus 309+CC) to be expected in HVS-II. Of the three private mutations, G709A, A13098G, and C14082G, the latter might very well be a phantom mutation of the type frequently found in the original data set of [16], where GGC often changed to GGG artificially in a first portion of the original data set [17].

Patient 3 (haplogroup I3a)

Again, the false G3423T is listed. We are seeing here an mtDNA that is evidently a member of haplogroup I (despite the absence of G8251A) but that does not belong to the two subhaplogroups, I1 and I2, described so far. Equally clear is that at least the mutations A1438G, A2706G, C7028T, and C14766T were overlooked and that the whole variation in the major part of the control region, say, 16194–576, was missed in the mtDNA analysis of this patient. A quick search for a perfect match of the motif T16086C, G16129A, C16223T for the range 16086–16223 in the SWGDAM database yields two sequences. They both bear T239C (and further mutations) plus the haplogroup I marker T250C. Then another search for a perfect match of the motif T239C, T250C for the range 239–250 gave another hit, so that we filtered out the three HVS-I/HVS-II sequences from the SWGDAM database that are potentially related to the patient's mtDNA (Table 2). Further near-matches are obtained by searching through >30,000 published HVS-I sequences for the motif T16086C, G16129A, and C16223T. This suggests that the mutations T152C, G207A, and T239C, together with some of the seemingly private coding-region mutations G2849A, A13395G, and T14783C of Patient 3, may characterize a new subhaplogroup of haplogroup I. Searching for entries in the SWGDAM database having

Table 2
mtDNA sequences from haplogroup I3

Source	Sample code	HVS-I	HVS-II	HVS-III	Coding region [partial screening][a]
[22]	P3	T16086C	n.d.	n.d.	@1438 @2706 G2849A @7028 @8251 A13395G @14766 T14783C
This study	RM258	T16086C (T16519C)	T152C G207A T239C 309+C	523+CACA 573.2+CCC	G709A G2849A @8251 A13395G
[42]	#114	T16086C	T152C G207A T239C 309+C	n.d.	n.d.
[43]		T16086C	T152C G207A T239C	n.d.	n.d.
[19]	USA.CAU.000637	T16086C	T152C G207A T239C	n.d.	n.d.
[42]	#115	T16086C	G207A T239C 309+C	n.d.	n.d.
[19]	USA.CAU.000366	T16086C T16263C T16311C	T152C G207A T239C 309+C	n.d.	n.d.
[42]	#119	—	T152C G207A T239C 309+C	n.d.	n.d.
[42]	#120	—	T152C G207A T239C 309+CC	n.d.	n.d.
[43]		—	T152C G207A T239C 309+CC	n.d.	n.d.
[19]	USA.CAU.000473	A16235G C16260T	C150T T152C G207A T239C 309+C	n.d.	n.d.
[34]	#54	A16235G C16260T	C150T T152C G207A T239C 309+C	n.d.	n.d.
[31]/this study[b]	#7	— (T16519C)	T152C G207A T239C 309+C	523+CACA 573.2+C	[—]
[31]/this study[b]	#15	— (T16519C)	T152C G207A T239C	C456T 523+CACA 573.2+CCCC[c]	[T8260C]

Mutations are listed with respect to the root of haplogroup I, which is determined relative to rCRS by mutations A73G, T199C T204C, T250C, A263G, 315+C, 573+CC (= 573.2), A750G, A1438G, G1719A, A2706G, A4529T, A4769G, C7028T, G8251A, A8860G, T10034C, T10238C, A10398G, G11719A, G12501A, C12705T, A13780G, C14766T, G15043A, A15326G, A15924G, G16129A, C16223T, and G16391A [11]. Bracketed mutation T16519C was only screened in this study; @ designates a back mutation; n.d., not determined.

[a] Results (relative to the root of haplogroup I) of sequencing the fragments 577–725, 2543–3455, 7989–8509, 13370–13674, and 14780–14950 are in square brackets.

[b] HVS-I, II, III were amplified and resequenced with primers described in [31], but using BigDye terminator chemistry (version 1.1) and capillary electrophoresis (310 Genetic Analyzer), leading to a revision of the HVS-III parts.

[c] With length heteroplasmy.

the three mutations T204C, G207A, and T250C yields another ten sequences without T239C; all ten sequences share the additional mutations G16129A, C16223T, A73G, T152C, T199C, A263G, and 315+C, so that we are possibly seeing a more basal subhaplogroup lacking T239C.

In order to confirm our expectation, we have completely sequenced the mtDNA of a (healthy) subject (RM258) from Rome, for which the HVS-I sequence yielded a perfect match with USA.CAU.000637 (Table 2). As a result we detected three of the seemingly private coding-region mutations of Patient 3 (relative to the root of haplogroup I) also in the Roman subject. Leaving aside four back mutations as potential oversights, only a single coding-region change appears to be a private mutation of Patient 3, namely, the synonymous mutation T14783C. This proves that we have indeed found an mtDNA very closely related to the patient's mtDNA. We therefore tentatively identified I3—a new subhaplogroup of I, with the characteristic mutation array T152C, G207A, T239C, 523+CACA, and 573.2+C. It then seems that G2849A, @8251, A13395G, and T16086C are diagnostic for a subhaplogroup, I3a, of I3.

Patient 4 (haplogroup T1a)

We should disregard G3423T as before. Haplogroup status T1 is well supported, but G709A, A750G, A2706G, G11719A, and all mutations within, say, 16194–576, have been missed. The coding-region mutation C12633A characteristic of subhaplogroup T1a has been misscored as C12633T. The patient's mtDNA does not possess the mutation T9899C previously regarded as characteristic of T1a. Intriguingly, mtDNA sample #87 from [16,17] does not show this mutation either but bears two of the patient's mutations, G7853A and C15295T, so that G15110A appears to be the single private mutation in the patient's mtDNA.

Patient 5 (haplogroup U1a1)

The mtDNA sequence belongs to haplogroup U1a but lacks the mutations G6026A, C14766T, A15326G, and the whole variation in, say, 16194–513. The (hotspot) mutations A16183C and T16189C typically seen in this haplogroup might be misscored here as A16183G and C16188T. A comparison with sequence #250 from [16,17], sequence K4b from [25], and sequence #18 from [13] allows us to conclude that

A385G plus the four mutations (522–523)del, 3158+T, G3591A, and A13422G recorded for the patient's mtDNA define a subhaplogroup, U1a1, of U1a.

Patient 6 (haplogroup U7)

This mtDNA sequence qualifies as a haplogroup U7 member, albeit with nearly the whole control region left unanalysed and the necessary mutations A750G, C7028T, C14766T, and A15326G unrecorded. Unfortunately, the uncertainty whether any further mutation might have been missed in this sequence does not help to resolve the conflicts between the haplogroup U7 sequences published so far. In particular, the lineages from [24,26] might lack some mutation(s), whereas we contend that the U7 sequence from [13] and the consensus of three U7 sequences from [11] are quite reliable (Table 3).

We conclude that the six mtDNA sequences provided by [22] essentially constitute incomplete coding-region sequences, missing numerous mutations, especially from the nine coding-region mutations that distinguish rCRS from the root haplotype of the superhaplogroup R (to which all six sequences but one belong). Although rare instances of single back mutations at these sites have been observed among >2100 coding-region sequences from human population genetics, such a massive mutation loss must be generated by oversight and non-stringent use of the correct reference sequence; cf. [27]. An error load of ~4 mutations per coding-region sequence is certainly unacceptable for medical studies.

Palimpsest of German mtDNA variation

Ruppert et al. [28] asserted that they "*analysed the whole mitochondrial genome in a series of 45 patients with DCM* [dilated cardiomyopathy] *for alterations and compared the findings with those of 62 control subjects*." These authors, however, performed mutation screening of the mtDNA with dideoxy fingerprinting of the PCR products. Their Table 1 is intended to list all "*point mutations in tRNA/rRNA genes, in the D-loop region as well as in non-coding regions of the mtDNA*." However, less than a handful of mutations in HVS-I among 107 subjects were found (G16110C, T16126C, and T16189C). The familiar 9-bp deletion (8281–8289)del and the 9-bp insertion 8289 + CCCCCTCTA in the intergenic region between COII and $tRNA^{Lys}$ are miscoded as 11 bp indels. Nearly all mtDNAs outside haplogroup H2 possess A1438G, but this mutation was reported in only a single control subject in their Table 2. Therefore, A and G at 1438 were effectively interchanged. The rRNA mutation A2706G signifying non-membership to haplogroup H is expected to occur with a relative frequency of 50–60% in most parts of Europe (Table 1 of [12]), but this polymorphism is not recorded in [28]. The mutation G15928C characteristic of haplogroup T occurs six times in the data set and thus nearly with the expected frequency, but none of the other three companion T markers A750G, G1888A, and T10463C in rRNA/tRNA genes were found to be polymorphic in this data set. This indicates a massive oversight of polymorphic sites in the coding region.

The variation detected in the third hypervariable segment (HVS-III, with range 438–576) also seems to be dubious. The frequent occurrence of the mutation C570T (12%), which is not found at all in the SWGDAM database of 1905 mtDNA sequences covering this part of the control region, is most suspicious. On the other hand, the expected haplogroup J mutation T489C is absent from (Table 1 of [28]), although, among the listed mutations, the motif A188G, G228A, and C295T together with the coding-region mutation C6464T clearly point to the presence of subhaplogroup J1c1 (in two controls). Poor sequencing results for HVS-III are not uncommon, especially in the region ~530–570, which may be difficult to read under certain circumstances (A. Brandstätter, unpubl. data), as is, e.g., testified by the numerous ambiguous nucleotides recorded in the SWGDAM database for that region.

Table 2 of [28] is supposed to list only "new" polymorphisms, but at least 20 of those mutations have already been published, some of which even date back to the data set published in [29]. Interestingly, A3360G, A5378G, and A10825G, each found with fre-

Table 3
Comparison of mtDNA coding-region sequences from haplogroup U7

Source	Sample code	Coding region
[22]	P6	@750 @980 @3741 @5360 @7028 **C8137T C8684T** T10084C A13395G **T13500C G14569A** @14766 @15326
[26]	#146/147	T9480C T10084C
[24]	U7	**C3741T C8137T** T10084C A10382G **T13500C** T14798C T15601C
[11]	B19	**C3741T C8137T C8684T T13500C G14569A** C5486T T13281C
[11]	B81	**C3741T C8137T C8684T T13500C G14569A** C5486T T13281C A9476G
[11]	C22	**C3741T C8137T C8684T T13500C G14569A** A7673G C14131T A15244G
[13]	#13	T961C 965 + CCC **C3741T C8137T C8684T** T10084C A11065G **T13500C G14569A** A15671G

Mutations are listed with respect to the putative root of haplogroup U7 according to [11], which is determined by the coding-region mutations A750G, T980C, A1438G, A1811G, A2706G, A4769G, C5360T, C7028T, A8860G, C10142T, A11467G, G11719A, A12308G, G12372A, C14766T, and A15326G; mutations in bold likely belong to the root motif, too; @ designates a back mutation.

quency 1 in DCM patients, appear together in the single complete sequence from haplogroup R1 known to date [11]. It is then conceivable that the mutation G5585T recorded in the table has actually been base-shifted and would conform to C5586T, which is observed in the R1 sequence, too.

Little is known about haplogroup R1, which is absent or very rare in most parts of Europe. The ancestral HVS-I motif seems to comprise only a single mutation, T16311C, but a subhaplogroup, which we will name R1a, has 16278 in addition. Unfortunately, both HVS-I motifs occur in haplogroup H as well [12]. Further unpublished data suggest that the restriction sites +4914*Bfa*I (caused by A4917G), −5584*Alu*I/−5586*Dde*I (= C5586T), −5823*Alu*I (= A5823G), and +16517*Hae*III (= T16519C) are characteristic of haplogroup R1 and that +15494*Dde*I (= G15497A) is diagnostic for the subhaplogroup R1a [30]. Since G15497A is not a "novel" mutation, we cannot know whether the specific mtDNA from [28] actually belongs to R1a in particular. It thus seems that a chance has been missed to supplement the meagre information for the rather rare haplogroup R1.

The relatively large number of "novel" mutations (not yet listed in Mitomap in protein coding genes) of this rare mtDNA contributed to the seeming mutation load of the DCM patients, which was deemed to be *statistically increased* [28]. Contrasting the mutational counts in DCM patients and controls is misleading anyway because the applied test statistic (χ^2) would stipulate independence of single mutations. However, a considerable number even of the "novel" mutations detected in any single mtDNA relative to the rCRS are not private as they are in fact inherited en bloc, which is testified by the corresponding pathway of the complete mtDNA tree known to date. There is hence no justification for claiming a significantly elevated number of mutations in DCM patients.

The *total number of detected sequence alterations in the whole mtDNA of patients with DCM and controls* [28] was found to be 458. Then 137 of these fall into the control region (16024–576) according to their Table 2. Thus, exactly three mutations per sample on average were detected in the coding region. For this count, the reference point is left open by the authors: it cannot be the rCRS (in view of the mutation count at position 1438) but would rather constitute some consensus sequence, which is most likely identical to the root haplotype of haplogroup R. However, the expected number of coding-region mutations relative to the R root is known to be about 12; cf. [11]. Therefore only one out of four mutations on average was actually detected in [28] with the method of dideoxy fingerprinting.

Even, if in the case of a single position the polymorphism happened to be determined correctly, a comparison between the small groups of DCM patients and controls can be misleading. For instance, the T16189C variant *that may be associated with a susceptibility to dilated cardiomyopathy could be detected in seven patients with DCM (15.6%) as well as in six control persons (9.7%)*. However, the expected relative frequency is about 15.5% in Germany, as is, for example, attained by the southwest German data set ($n = 200$) of [31].

A similar case could be made for the UK, where the T16189C variant is found in approximately 12% of the individuals [32,33]; e.g., 11% in the data set ($n = 100$) of [34], a value which is intermediate between the corresponding percentages for DCM patients and controls from England in the study of [35]. It has been claimed that founder effects could be excluded to have influenced the 16189 polymorphism since the variant T16189C has arisen multiple times on different haplogroup backgrounds [35,36]. However, this is not a valid argument because this variant belongs to the inherited motif of some haplogroups, which may have quite varying frequencies due to population structure. Such an effect can well be seen in the populations from the Caucasus, for example, where the frequencies T16189C are in general twice as high compared to western Europe (with 24% on average, but within a wide range 3–70%, especially in Daghestan). Since the 16189 polymorphism is virtually ubiquitous, it has often been regarded in association with a number of diseases [36]. Closer examination, however, using more extensive and balanced sampling should reveal that such associations are rather spurious [32,33].

Snapshot of Chinese mtDNA variation

Whereas the previous mtDNA sequences constitute a snapshot of European mtDNA variation, the four complete mtDNA sequences published by Young et al. [37] allow us to take a glimpse at East Asian mtDNA variation. These authors analysed the complete mtDNA in four Chinese pedigrees with aminoglycoside-induced and non-syndromic hearing impairment. There are problems with these data similar to the mitochondrial encephaloneuromyopathies case, although these sequences were in fact roughly allocated to mtDNA haplogroups according to the classification scheme of [38]. We will compare the reported mtDNA variation, mutation by mutation, to the East Asian mtDNA tree shown in (Fig. 1 of [14]) and employ additional information from [15]. Before discussing the cases in detail, one needs to correct some typos in (Table 1 of [37]). Site 5885 should read 5585, and the nucleotides at 10640 and 12358 in the "CRS" column should be corrected to T and A, respectively. The entry G in column BJ101 likely is a typo and should be substituted by T.

Table 4
mtDNA sequences from haplogroup F3

Source	Sample code	Haplogroup	Control region [alternatively 16,001–407]	Coding region [alternatively 10,525–11,490]
[37]	BJ101	**F3b**	T152C C495T A16220C A16265G	C5076T T5442C C8270G **(8281–8289)del** @8860 **G9947A** G10427A T10873C, C10980G A10988C @11065 @14766 T15784C
This study	QJ383	**F3b**	C151T T152C 309.1+C **A16220C** A16227G	T2392C T3535C T4802C **(8281–8289)del G9947A** A10398G G11914A C13044T C15910T
This study	Yi361	**F3b**	[T152C **A16220C** A16227G]	[—]
[14]	XJ8451	**F3a**	T195C T16209C T16311C **C16355T**	A3390G T7094C **C12621T**

Mutations are listed with respect to the root of haplogroup F3, which is determined relative to rCRS by mutations A73G, 249del, A263G, 309+C (309.1), 315+C, A750G, A1438G, A2706G, A3434G, C3970T, A4769G, G5585A, A5978G, G5913A, T6392C, C7028T, A8860G, G10310A, G10320A, A11065G, G11719A, G13928C, C14766T, A15326G, T16298C, and T16362C. Mutations diagnostic for subhaplogroups (F3a and F3b, respectively) are in bold; @ designates a back mutation.

We assume that "316 Ins C" actually means the (nearly omnipresent) mutation 315+C in the standard notation, where an insertion is scored at the last possible position in the rCRS after which the nucleotide can be inserted. Similarly, we would score the 9-bp deletion at sites 8281–8289 rather than at 8271–8279.

Pedigree BJ101 (haplogroup F3b)

This mtDNA sequence clearly belongs to haplogroup F3, although three (A11065G, C12621T, and C16355T) of the eleven F3-specific mutations listed in [14] are absent in this pedigree. As explained in [39], the single representative of haplogroup F3 in [14] actually belongs to a specific subhaplogroup, now called F3a. The sister subhaplogroup, referred to as F3b, is characterized by the combined HVS-I and HVS-II motif A16220C, T16298C, T16362C, A73G, 249del, A263G, 309+C, and 315+C, but no coding-region information has been available so far. With respect to HVS-I and HVS-II the pedigree's mtDNA nearly matches (except for T16311C) the mtDNA F3b lineage USA.335.000127 from the SWGDAM database. Therefore, A11065G, C12621T, and C16355T are potentially the only diagnostic mutations for haplogroup F3a, whereas the mutations characteristic of haplogroup F3b are to be sought among the private mutations of the pedigree's mtDNA, namely, C495T, C5076T, T5442C, C8270G, (8281–8289)del, @8860, G9947A, G10427A, T10873C, C10980G, A10988C, @14766, T15784C, and A16220C. Three mutations are suspicious here: @8860 and @14766 may constitute an oversight of the mutations A8860G and C14766T, which are generally shared by all sequences outside haplogroup HV, whereas T10873C is extremely rare within superhaplogroup N [9].

In order to confirm the potential motif of haplogroup F3 and identify the characteristic mutations of its subhaplogroup F3b, we have completely sequenced a sample from Qijiang (Chongqing, southwest China), harbouring the 16227 transition. The comparison with the complete F3a sequence from [14] and the somewhat incomplete F3b sequence BJ101 then suggests that A11065G was overlooked in the latter and would be a characteristic mutation for the whole haplogroup F3 (Table 4). We conclude that subhaplogroup F3a has only C12621T and C16355T as its diagnostic mutations, whereas F3b has A16220C, the 9-bp deletion, and G9947A (which, of course, hinges on the hypothesis that no further mutations were overlooked in Pedigree BJ101).

Pedigree BJ102 (haplogroup N9a1)

This mtDNA sequence belongs to haplogroup N9a but misses the N9-diagnostic mutation G5417A and, as in the preceding case, shows the back mutation @14766 and the suspicious T10873C. It belongs to a particular subhaplogroup, N9a1 [15], which may have the four characteristic mutations, T4386C, G12007A, C16111T, and G16129A, despite the fact that the first one was not recorded in the pedigree's mtDNA.

Pedigree BJ103 (haplogroup D4a)

Although mutation G16129A is not present, we can safely infer haplogroup status D4a in this case. Five coding-region mutations are missing: A2706G, C7028T, C10400T, T10873C, and C14766T.

Pedigree BJ104 (haplogroup D4b2b)

This mtDNA sequence qualifies as a member of haplogroup D4b2b [15], despite the lack of the expected (522–523)del, since it bears the mutations A1382C, G8020A, C8964T, C9296T, T9824A, and T16519C. Three mutations were overlooked, namely, A4769G, C14766T, and the D4-specific marker C8414T.

In summary, the data of [37] fit well into the East Asian mtDNA phylogeny, despite an error load of ~4 mutations per sequence. In particular, site 14766 was probably read with the original CRS rather than the revised CRS, and the variation at site 10873 was possibly not recorded correctly. This notwithstanding, the nearly

complete mtDNA sequence representing haplogroup F3b filled a lacuna in the previously known mtDNA phylogeny.

Conclusion

We have demonstrated that the sequencing efforts of [22,37] were more or less incomplete based on the present knowledge of the mtDNA phylogeny. The samples stem from distant mtDNA pools (Italy and China), but the artefacts bear a similar signature of documentation errors and non-stringent use of rCRS as the proper reference sequence. An incomplete and flawed cumulative mtDNA mutation list, which disentangles and distorts the relevant haplotype information, as provided in [28], is of little help for understanding the genetic basis of a certain disease (in this case, dilated cardiomyopathy). In particular, mtDNA screening by dideoxy fingerprinting analysis seems to be a rather unreliable method for pinpointing all mutations in the whole mitochondrial genome. It would be most desirable to see corrections of all these results, especially because some of the (corrected) sequences would valuably supplement our knowledge about the global mtDNA phylogeny.

Although the main target of a disease study are non-synonymous mutations, usually a complete record of all mutations is needed in order to allocate the mtDNA under examination to its place in the mtDNA phylogeny. This also includes the variation in the control region, especially HVS-I, because the worldwide HVS-I database is huge and can assist in focussing the search for closely related mtDNAs from controls to be compared to the patients' mtDNAs. In many cases, samples are stored in deep freezers so that DNA is still available for screening of specific sites in the coding region. We have proven that the near-matching strategy can readily sort out closely related mtDNAs, which renders the search for potentially pathogenic mutations much more effective than, say, a Mitomap query.

It goes without saying that any discussion about "novel" or "pathogenic" mutations has to be based on a complete knowledge of the worldwide mtDNA phylogeny—and on reasonably correct mtDNA sequences from patients and controls (see, e.g. [40]). With the emerging finer details of the continental mtDNA phylogenies, every mutation, if not private, can be positioned in the current mtDNA tree. This helps to prevent premature claims about the pathogenicity of a certain mutation. Namely, if a mutation defines a whole major or minor subhaplogroup, then it would be very unlikely that such a mutation may be a primary disease mutation. But for establishing the status of secondary disease mutation one would need a well-designed study with a phylogenetic focus that evaluates mtDNA lineages from patients and controls in different populations [41].

Electronic-database information

Mitomap, http://www.mitomap.org/ SWGDAM database, http://www.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm GenBank, http://www.ncbi.nlm.nih.gov/Genbank/; for the two complete sequences provided here, see Accession Nos.: AY963586 and AY972053.

Acknowledgments

This research received support from Progetto CNR-MIUR Genomica Funzionale-Legge 449/97 (to A.T.), the Italian Ministry of the University (Progetti Ricerca Interesse Nazionale 2003) (to A.T.), and Fondo Investimenti Ricerca di Base 2001 (to A.T.).

References

[1] V. Macaulay, M. Richards, B. Sykes, Mitochondrial DNA recombination—no need to panic, Proc. R. Soc. London Ser. B 266 (1999) 2037–2039.
[2] T. Kivisild, H.V. Tolk, J. Parik, Y. Wang, S.S. Papiha, H.-J. Bandelt, R. Villems, The emerging limbs and twigs of the East Asian mtDNA tree, Mol. Biol. Evol. 19 (2002) 1737–1751, erratum in: Mol. Biol. Evol. 20 (2002) 162.
[3] T. Ozawa, M. Tanaka, H. Ino, K. Ohno, T. Sano, Y. Wada, M. Yoneda, Y. Tanno, T. Miyatake, T. Tanaka, S. Itoyama, S. Ikebe, T. Kondo, Y. Mizuno, Distinct clustering of point mutations in mitochondrial DNA among patients with mitochondrial encephalomyopathies and with Parkinson's disease, Biochem. Biophys. Res. Commun. 176 (1991) 938–946.
[4] T. Ozawa, M. Tanaka, S. Sugiyama, H. Ino, K. Ohno, K. Hattori, T. Ohbayashi, T. Ito, H. Deguchi, K. Kawamura, Y. Nakane, K. Hashiba, Patients with idiopathic cardiomyopathy belong to the same mitochondrial DNA gene family of Parkinson's disease and mitochondrial encephalomyopathy, Biochem. Biophys. Res. Commun. 177 (1991) 518–525.
[5] T. Ozawa, Mechanism of somatic mitochondrial DNA mutations associated with age and diseases, Biochem. Biophys. Acta 1271 (1995) 177–189.
[6] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (1999) 147.
[7] H.-J. Bandelt, L. Quintana-Murci, A. Salas, V. Macaulay, The fingerprint of phantom mutations in mitochondrial DNA data, Am. J. Hum. Genet. 71 (2002) 1150–1160.
[8] H.-J. Bandelt, A. Salas, C. Bravi, Problems in FBI mtDNA database, Science 305 (2004) 1402–1404.
[9] H.-J. Bandelt, A. Salas, S. Lutz-Bonengel, Artificial recombination in forensic mtDNA population databases, Int. J. Legal Med. 118 (2004) 267–273.
[10] H.-J. Bandelt, P. Lahermo, M. Richards, V. Macaulay, Detecting errors in mtDNA data by phylogenetic analysis, Int. J. Legal. Med. 115 (2001) 64–69.

[11] M.g. Palanichamy, C. Sun, S. Agrawal, H.-J. Bandelt, Q.-P. Kong, F. Khan, C.-Y. Wang, T.K. Chaudhuri, V. Palla, Y.-P. Zhang, Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia, Am. J. Hum. Genet. 75 (2004) 966–978.
[12] A. Achilli, C. Rengo, C. Magri, V. Battaglia, A. Olivieri, R. Scozzari, F. Cruciani, M. Zeviani, E. Briem, V. Carelli, P. Moral, J.M. Dugoujon, U. Roostalu, E.L. Loogvali, T. Kivisild, H.-J. Bandelt, M. Richards, R. Villems, A.S. Santachiara-Benerecetti, O. Semino, A. Torroni, The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool, Am. J. Hum. Genet. 75 (2004) 910–918.
[13] A. Achilli, C. Rengo, V. Battaglia, M. Pala, A. Olivieri, S. Fornarino, C. Magri, R. Scozzari, N. Babudri, A.S. Santachiara-Benerecetti, H.-J. Bandelt, O. Semino, A. Torroni, Saami and Berbers—an unexpected mitochondrial DNA link, Am. J. Hum. Genet. 76 (2005) 883–886.
[14] Q.-P. Kong, Y.-G. Yao, C. Sun, H.-J. Bandelt, C.-L. Zhu, Y.-P. Zhang, Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences, Am. J. Hum. Genet. 73 (2003) 671–676, erratum in: Am. J. Hum. Genet. 75 (2004) 157.
[15] M. Tanaka, V.M. Cabrera, A.M. González, J.M. Larruga, T. Takeyas, N. Fuku, L.J. Guo, R. Hirose, Y. Fujita, M. Kurata, K. Shinoda, K. Umetsu, Y. Yamada, Y. Oshida, Y. Sato, N. Hattori, Y. Mizuno, Y. Arai, N. Hirose, S. Ohta, O. Ogawa, Y. Tanaka, R. Kawamori, M. Shamoto-Nagai, W. Maruyama, H. Shimokata, R. Suzuki, H. Shimodaira, Mitochondrial genome variation in eastern Asia and the peopling of Japan, Genome Res. 14 (2004) 1832–1850.
[16] C. Herrnstadt, J.L. Elson, E. Fahy, G. Preston, D.M. Turnbull, C. Anderson, S.S. Ghosh, J.M. Olefsky, M.F. Beal, R.E. Davis, N. Howell, Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups, Am. J. Hum. Genet. 70 (2002) 1152–1171, erratum in: Am. J. Hum. Genet. 71 (2002) 448–449.
[17] C. Herrnstadt, G. Preston, N. Howell, Errors, phantom and otherwise, in human mtDNA sequences, Am. J. Hum. Genet. 72 (2003) 1585–1586.
[18] B.A. Malyarchuk, I.B. Rogozin, Mutagenesis by transient misalignment in the human mitochondrial DNA control region, Ann. Hum. Genet. 68 (2004) 324–339.
[19] K.L. Monson, K.W.P. Miller, M.R. Wilson, J.A. DiZinno, B. Budowle, The mtDNA population database: an integrated software and database resource for forensic comparison, Forensic Sci. Comm. 4 (2002) no. 2.
[20] B. Budowle, D. Polanskey, M.W. Allard, R. Chakraborty, Addressing the use of phylogenetics for identification of sequences in error in the SWGDAM mitochondrial DNA database, J Forensic Sci. 49 (2004) 1–6.
[21] D. Polanskey, B. Budowle, Summary of the findings of a quality review of the Scientific Working Group on DNA analysis methods mitochondrial DNA database, Forensic Sci. Comm. 7 (2005) no. 1.
[22] P. Da Pozzo, E. Cardaioli, E. Radi, A. Federico, Sequence analysis of the complete mitochondrial genome in patients with mitochondrial encephaloneuromyopathies lacking the common pathogenic DNA mutations, Biochem. Biophys. Res. Commun. 324 (2004) 360–364.
[23] S. Anderson, A.T. Bankier, B.G. Barrell, M.H. de Bruijn, A.R. Coulson, J. Drouin, I.C. Eperon, D.P. Nierlich, B.A. Roe, F. Sanger, P.H. Schreier, A.J. Smith, R. Staden, I.G. Young, Sequence and organization of the human mitochondrial genome, Nature 290 (1981) 457–465.
[24] N. Maca-Meyer, A.M. González, J.M. Larruga, C. Flores, V.M. Cabrera, Major genomic mitochondrial lineages delineate early human expansions, BMC Genet. 2 (2001) 13.
[25] M. Ingman, U. Gyllensten, Mitochondrial genome variation and evolutionary history of Australian and New Guinean Aborigines, Genome Res. 13 (2003) 1600–1606.
[26] S. Finnilä, M.S. Lehtonen, K. Majamaa, Phylogenetic network for European mtDNA, Am. J. Hum. Genet. 68 (2001) 1475–1484.
[27] Y.-G. Yao, V. Macauley, T. Kivisild, Y.-P. Zhang, H.-J. Bandelt, To trust or not to trust an idiosyncratic mitochondrial data set, Am. J. Hum. Genet. 72 (2003) 1341–1346.
[28] V. Ruppert, D. Nolte, T. Aschenbrenner, S. Pankuweit, R. Funck, B. Maisch, Novel point mutations in the mitochondrial DNA detected in patients with dilated cardiomyopathy by screening the whole mitochondrial genome, Biochem. Biophys. Res. Commun. 318 (2004) 535–543.
[29] M. Ingman, H. Kaessmann, S. Pääbo, U. Gyllensten, Mitochondrial genome variation and the origin of modern humans, Nature 408 (2000) 708–713.
[30] V. Macaulay, M. Richards, E. Hickey, E. Vega, F. Cruciani, V. Guida, R. Scozzari, B. Bonné-Tamir, B. Sykes, A. Torroni, The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs, Am. J. Hum. Genet. 64 (1999) 232–249, erratum in: Am. J. Hum. Genet. 64 (1999) 918–919.
[31] S. Lutz, H.-J. Weisser, J. Heizmann, S. Pollak, Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany, Int. J. Legal Med. 111 (1998) 67–77, errata in: Int. J. Legal Med. 111 (1998) 286, and 112 (1999) 145–150. [Corrected data table available from the author.]
[32] A.M. Gibson, J.A. Edwardson, D.M. Turnbull, I.G. McKeith, C.M. Morris, P.F. Chinnery, No evidence of an association between the T16189C mtDNA variant and late onset dementia, J. Med. Genet. 41 (2004) e7.
[33] S.M. Keers, A.M. Gibson, D.M. Turnbull, P.F. Chinnery, No evidence of an association between the mtDNA 16184-93 polyC tract and late onset dementia, J. Med. Genet. 41 (2004) 957–958.
[34] R. Piercy, K.M. Sullivan, N. Benson, P. Gill, The application of mitochondrial sequence DNA typing to the study of white Caucasian genetic identification, Int. J. Leg. Med. 106 (1993) 85–90.
[35] S.S. Khogali, B.M. Mayosi, J.M. Beattie, W.J. McKenna, H. Watkins, J. Poulton, A common mitochondrial DNA variant associated with susceptibility to dilated cardiomyopathy in two different populations, Lancet 357 (2001) 1265–1267.
[36] J. Poulton, S. Das, Correction: no evidence of an association between the T16189C mtDNA variant and late onset dementia (Gibson et al), J. Med. Genet. 41 (2004) 957.
[37] W.Y. Young, L. Zhao, Y. Qian, Q. Wang, N. Li, J.H. Greinwald Jr., M.X. Guan, Extremely low penetrance of hearing loss in four Chinese families with the mitochondrial 12S rRNA A1555G mutation, Biochem. Biophys. Res. Commun. 328 (2005) 1244–1251.
[38] Y.-G. Yao, Q.-P. Kong, H.-J. Bandelt, T. Kivisild, Y.-P. Zhang, Phylogeographic differentiation of mitochondrial DNA in Han Chinese, Am. J. Hum. Genet. 70 (2002) 635–651.
[39] Y.-G. Yao, C.M. Bravi, H.-J. Bandelt, A call for mtDNA data quality control in forensic science, Forensic Sci. Int. 141 (2004) 1–6.
[40] H.-J. Bandelt, Y.-G. Yao, T. Kivisild, Mitochondrial genes and schizophrenia, Schizophr. Res. 72 (2005) 267–269.
[41] C. Herrnstadt, N. Howell, An evolutionary perspective on pathogenic mtDNA mutations: haplogroup associations of clinical disorders, Mitochondrion 4 (2004) 791–798.
[42] M. Poetsch, H. Wittig, D. Krause, E. Lignitz, Corrigendum to "Mitochondrial diversity of a northeast German population sample", Forensic Sci. Int. 145 (2004) 73–77.
[43] B.A. Malyarchuk, T. Grzybowski, M.V. Derenko, J. Czarny, M. Wonźiak, D. Miścicka-Śliwka, Mitochondrial DNA variability in Poles and Russians, Ann. Hum. Genet. 66 (2002) 261–283.