Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 335 (2005) 891–899
BBRC
www.elsevier.com/locate/ybbrc

A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics

Antonio Salas [a,*], Ángel Carracedo [a], Vincent Macaulay [b], Martin Richards [c], Hans-Jürgen Bandelt [d]

[a] *Unidade de Xenética, Instituto de Medicina Legal, Facultade de Medicina, 15782 Universidade de Santiago de Compostela, Centro Nacional de Xenotipado (CeGen), Hospital Clínico Universitario, 15706 Galicia, Spain*
[b] *Department of Statistics, University of Glasgow, Glasgow, UK*
[c] *School of Biology, University of Leeds, Leeds, UK*
[d] *Fachbereich Mathematik, Universität Hamburg, Hamburg, Germany*

Received 26 July 2005
Available online 8 August 2005

Abstract

Several suggestions have been made for avoiding errors in mitochondrial DNA (mtDNA) sequencing and documentation. Unfortunately, the current clinical, forensic, and population genetic literature on mtDNA still delivers a large number of studies with flawed sequence data, which, in extreme cases, damage the whole message of a study. The phylogenetic approach has been shown to be useful for pinpointing most of the errors. However, many geneticists, especially in the forensic and medical fields, are not familiar with either effective search strategies or the evolutionary terminology. We here provide a manual that should help prevent errors at any stage by re-examining data fresh from the sequencer in the light of previously published data. A fictitious case study of a European mtDNA data set (albeit composed from the literature) then demonstrates the steps one has to go through in order to assess the quality of sequencing and documentation.

Keywords: Human mtDNA; HVS-I; HVS-II; Error detection; Networks; Phylogenetic analysis; Haplogroup; SWGDAM

During the last decade, the evolution of PCR technology, in parallel with the development of new automatic sequencers and sequencing chemistries, has allowed the improvement of electropherogram reliability and the accuracy of the DNA sequence data. Large volumes of mtDNA sequence data are nowadays produced automatically and reported in the literature or in databases. However, many errors still routinely occur, most of which could have been avoided if careful checking strategies had been applied and lab conditions critically re-examined. A lack of knowledge in this regard seems to persist even in population genetics and forensics, as, for example, occasionally expressed by misrepresentation of the error issue [1,2].

Numerous papers have investigated the error issue to date [3–9], but this effort has not yet been assimilated by the scientific community. Most of the previous publications about errors in mitochondrial data dealt with data sets from population genetics and forensics, but the medical field seems to be most strongly affected by missequencing and misdocumentation. For instance, a list of control region mutations found in colorectal cancer patients was displayed in [10]; according to the authors, position 71 was altered in three out of seven cases (deletion of one G). Taking into account that the majority of mutations found to be unstable in tumors are also common polymorphisms in human populations [11], it is at least surprising to see this highly stable site in human

* Corresponding author. Fax: +34 981 580336.
E-mail address: apimlase@usc.es (A. Salas).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.07.161

studies as extremely unstable in [10]. A readability problem in the G stretch (66–71) is a possible explanation for this result. In [12], the authors sought to find a G to C mutation at site 73 in the mitochondrial databank, which is an obvious misdocumentation, since position 73 shows an 'A' in the revised Cambridge reference sequence (rCRS). In the medical field, incomplete and incorrect recording of mutations in total sequencing attempts and sample mix-up seem to be committed on a routine basis; see, e.g., the cases reanalyzed in [13,14].

Many studies (particularly in the medical field) just target the coding region of the mtDNA genome, since there is a general expectation that only mutations in this region could be responsible (as a risk factor) for certain (complex) diseases. In many of these studies, the control region is therefore not treated with much care (and at best semi-sequenced) or not even analyzed at all. However, it is important to screen the control region, or at least the first two hypervariable segments (HVS-I and HVS-II), in order to (i) link the coding-region information to potential motifs in the HVS-I/II database and (ii) have a feedback as to quality assessment.

As to the feature (i) of linking motifs, the worldwide database of published HVS-I sequences is quite enormous, currently comprising ~40,000 sequences, albeit scattered through publicly available databases and articles. It is not prudent, however, to employ a database (on the web) that simply compiled data from tables of published articles, without comparing with the original sources. For example, the data entries in the mtDNA database "HVRbase" were found to be highly unreliable [15]—and they still are, for several reasons (Bandelt et al., unpublished manuscript). We therefore advise against employing "HVRbase" to circumvent the more laborious own compilation of published data, table by table. When a sufficient number (say, more than 10,000) of HVS-I sequences (plus corresponding HVS-II sequences in some cases) are ready for screening, it will in general be quite easy to connect a particular complete mtDNA sequence from a patient with a particular control-region motif in the database. One can then focus the search for potentially phylogenetically related mtDNAs to those which (nearly) match the patient's HVS-I motif or full control-region motif. As exemplified in [13], complete sequencing of such candidate samples can then help to distinguish shared polymorphisms from private mutations, so that the search of potentially pathogenic mutations becomes much more focused.

As to the feature (ii) of quality assessment, sequencing of the control region with its hypervariable segments demands high standards of sequencing (because of homopolymeric tracts) and documentation (because of numerous variant nucleotides to be recorded correctly). The large number of HVS-I sequences available for direct comparison aids in pinpointing principal problems in sequencing and documentation. To give a motivating example, the authors of [16] have aimed at analyzing the entire mtDNA genomes of three LHON patients (of Chinese origin). One of the mtDNA sequences obtained (allocated to mtDNA haplogroup F1) is reported to bear the transversion T16304G in HVS-I (relative to the rCRS). However, consultation of previous publications [17–19] reveals that the transition T16304C is shared by haplogroup F1 mtDNAs. Moreover, among all published HVS-I sequences we found only two instances of T16304G in the data sets [20,21], which seem to be problematic anyway [3,4]. The best explanation for seeing T16304G is a quite typical documentation error, viz. the confusion of the nucleotides C and G [3]. Thus, being warned by this HVS-I finding, one would then further search for coding region mutations for which C and G seem to have been interchanged by mistake; and indeed, we have observed such instances (Bandelt et al., unpublished manuscript) in the three mtDNA sequences reported by [16]. Moreover, the sequence allocated to haplogroup M10 lacks the haplogroup M marker mutation T489C in the third hypervariable segment (HVS-III)—a very unusual finding. This could suggest that some coding-region mutations in that data set might be overlooked as well, which in fact seems to be the case; compare with the data of [22] re-analyzed by [13].

Site-specific mutational rates and error detection

Conceptually, the method of error detection we follow is very simple: any single mtDNA sequence necessarily needs to fit into a specific part of the mtDNA phylogeny that is characterized by specific mutations. We often detect some pieces that do not quite fit the pattern, where we may expect that an artifact—and not a biological process—could explain the profile. For instance, the deletion 249d in HVS-II is almost always connected with one or other of two different HVS-I motifs, viz. either the mutational pair C16223T T16298C or the single change T16304C relative to the rCRS [23]. We then learn from a systematic phylogenetic study that the two motifs C16223T T16298C A73G 249d A263G and T16304C A73G 249d A263G constitute ancestral HVS-I and II sequences for basal branches of the mtDNA phylogeny [19] and are thus inherited as a single haplotype block rather than having arisen de novo multiple times (although it should be borne in mind that the mutational stability of either HVS-I motif is not infallible). Therefore, it would at least be somewhat unusual (but not implausible or unbelievable) to detect a profile that combines a typical West Eurasian HVS-I lineage, such as C16069T T16126C, with 249d, especially when, on the other hand, an expected mutation (here, say, C295T) was missing. It is the combination of several of such unusual findings that would point to flawed re-