



Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (*numts*)

Sébastien Calvignac^{a,*}, Lara Konecny^a, Florian Malard^a, Christophe J. Douady^{a,b}

^a Université de Lyon, F-69622, Lyon, France, Université Lyon 1, Villeurbanne, CNRS, UMR5023, Laboratoire d'Ecologie des Hydrosystèmes Fluviaux, France

^b Institut Universitaire de France, France

ARTICLE INFO

Article history:

Received 30 April 2010

Received in revised form 20 September 2010

Accepted 19 October 2010

Available online 1 November 2010

Keywords:

Mitochondrial DNA

numt

Paralog

Enrichment

Long-range PCR

Dilution

ABSTRACT

Molecular tools have become prominent in ecology and evolution. A target of choice for molecular ecologists and evolutionists is mitochondrial DNA (mtDNA), whose many advantages have also convinced broad-scale, pragmatic programmes such as barcode initiatives. Of course, mtDNA is also of interest to human geneticists investigating mitochondrial diseases. Studies using mtDNA are however put at great risk by the inadvertent co-amplification or preferred amplification of nuclear pseudogenes (*numts*). *A posteriori* analysis of putative mtDNA sequences can help in removing *numts* but faces severe limitations (e.g. recently translocated *numts* will most of the time go unnoticed). Counter-measures taken *a priori*, i.e. explicitly designed for avoiding *numt* co-amplification or preferred amplification, are appealing but have never been properly assessed. Here we investigate the efficiency of four such measures (mtDNA enrichment, cDNA amplification, long-range amplification and pre-PCR dilution) on a common set of *numt* cases, showing that mtDNA enrichment is the worst performer while the use of pre-PCR dilution is a simple, yet robust method to prevent the pollution of putative mtDNA datasets with *numts*. Therefore, straightforward recommendations can be made that, if followed, will considerably increase the confidence in the mitochondrial origin of any mtDNA-like sequence.

© 2010 Elsevier B.V. and Mitochondria Research Society. All rights reserved.

1. Introduction

Clonality and near-neutrality have made mitochondrial DNA (mtDNA) a preferred target for molecular ecologists and evolutionists (Galtier et al., 2009). These assumptions now appear overoptimistic in many instances (Balloux, 2009; Galtier et al., 2009; White et al., 2008). Another serious drawback of mitochondrial markers is the risk of co-amplifying, or worse, preferentially amplifying, nuclear paralogs of mtDNA (thereafter called *numts* (Lopez et al., 1994)). Recent scans have indeed revealed that many animal nuclear genomes have captured large numbers of *numts* (e.g. ca. 500 in *Danio rerio* (Antunes and Ramos, 2005) and ca. 1600 in *Apis mellifera* (Pamilo et al., 2007)), which have from time to time proved to be pervasive in mtDNA studies (e.g. gorilla (Thalmann et al., 2004, 2005) and muskoxen (Kolokotronis et al., 2007)). Including *numts* in mtDNA datasets imperils subsequent analyses with major biological consequences because the fundamental principle of orthology is no longer respected. False phylogenies or fake cryptic species can be inferred if relatively ancient (interspecific) events are under study as recently

shown for *Apodemus* (wood mice) systematics (Dubey et al., 2009) or crayfish and grasshoppers (genus *Schistocerca*, *Calliptamus*, *Acrida*, *Locusta* and *Orconectes*) cryptic species definition (Song et al., 2008). Where more recent (intraspecific) events are concerned, *numts* can bias demographic profiles and selective histories.

Fortunately, careful examination and analysis of putative mtDNA alignments, i.e. quality control (Buhay, 2009; Song et al., 2008), will help in removing a number of *numts*. Various strategies have been considered (Bensasson et al., 2001; Buhay, 2009; Song et al., 2008; Sorenson and Quinn, 1998):

- i) *Looking for insertions/deletions (indels) inducing frameshifts and for in-frame stop codons in coding sequences.* *Numts* are non-functional in the nuclear genome and will tolerate coding sequence disruption, which is obviously not the case of severely purified organellar protein coding sequences.
- ii) *Examining codon position substitution bias.* The degeneracy of genetic codes concentrates most organellar variability in 3rd position of codons (synonymous position). This constraint does not exist for *numts* which should deviate from the rule and accumulate mutations at any codon position.
- iii) *Comparing reconstructed mtDNA phylogenies with phylogenies derived from other molecular markers or non-molecular characters.* It can betray aberrant sisterships for *numts* when unambiguous phylogenetic relationships are ascertained elsewhere (e.g. morpho-anatomy or nuclear markers).

* Corresponding author. Present address: Emerging Zoonoses, Robert Koch Institut, Nordufer 20, 13353 Berlin, Germany. Tel.: + 49 30 18754 2592; fax: + 49 30 18754 2181.

E-mail addresses: calvignacs@rki.de (S. Calvignac), konecny@univ-lyon1.fr (L. Konecny), malard@univ-lyon1.fr (F. Malard), douady@univ-lyon1.fr (C.J. Douady).

- iv) *Determining rates of evolution.* MtDNA substitution rates tend to be higher than nuclear ones. Hence *numts* would have relatively slow rates of evolution.

Method (iv) is inherently less reliable than methods (i) to (iii): mtDNA heterotachy is common and *numts* might experience rate acceleration as a consequence of a loss of constraints. Using methods (i) to (iii) would only help detect major departures from expectations. Therefore, the probability of identifying recently translocated *numts* will be low, especially if they are numerous and preferentially amplified (Song et al., 2008). Finally, *numts* stemming from non-translated sequences will obviously be harder to detect since points (i) and (ii) cannot apply to them. In the peculiar case of ribosomal DNA, incorrect stem-loop folding has been suggested as an alternative anti-*numt* diagnostic (Olson and Yoder, 2002).

Careful primer design has also been put forward as potentially able to decrease the risk of *numt* co-amplification (Bensasson et al., 2001; Sorenson and Quinn, 1998). In cases where *numts* have already been identified through quality control, primer design should take it into account and aim at maximizing the number of 3' mismatches between primers and *numts* (Bensasson et al., 2001; Sorenson and Quinn, 1998). Besides, some authors suggested that the use of more specific primers (as opposed to universal ones) could be beneficial (Song et al., 2008). Recent in-depth analyses however disregarded this possibility, revealing *numts* are found very frequently with narrower-spectrum primers as well (Moulton et al., 2010).

Bench methods that do not depend in any way on prior knowledge about mtDNA and *numt* sequences are therefore appealing. Any such strategy will only rely on the distinctive properties of both molecules. A review of the literature allows identifying four main methods:

- 1) *Purification of mtDNA.* Genuine mtDNA is found in mitochondria and *numts* in nuclei. Technically, the objective will therefore be to isolate mtDNA from genomic DNA through the selective sedimentation of mitochondria, from which mtDNA will be extracted. Of note, the method has paved the way to the broad use of mtDNA in the pre-PCR era and a wealth of mtDNA sequences have been determined through cloning of enriched extracts (e.g. Anderson et al., 1981).
- 2) *PCR amplification from RNA derived complementary DNA (cDNA) template.* Several lines of evidence suggest that *numts* do not undergo transcription, contrary to most mtDNA sequences. First, mitochondrial and nuclear genomes have evolved distinct genetic codes, whether vertebrates or invertebrates are under consideration, including only partially overlapping set of start codons (Jukes and Osawa, 1993; Williams and Knowlton, 2001). Second, most *numts* will include neither complete coding sequences nor complete coding sequences accompanied with the unique promoter region of mtDNA (Williams and Knowlton, 2001). Third, *numts* do not seem to be specifically addressed to transcriptionally active parts of the genome (e.g. Pamilo et al., 2007) so that in a number of cases their chance integration will take place into intergenic, untranscribed zones (Collura et al., 1996). Finally, those *numts* integrating into exons will have deleterious effects and therefore be counter-selected (Collura et al., 1996). Technically, the method will obviously rely on RNA, not DNA, extraction from samples. RNA will then be converted in DNA through reverse transcription.
- 3) *Semi-nested/nested PCR amplification on long-range PCR product.* *Numts* tend to be short (with a mean length approaching 100 to 300 bp (Pamilo et al., 2007; Richly and Leister, 2004)) while native mtDNA is a ca. 16 kbp molecule in metazoans. Hence, PCR of standard length (e.g. ca. 650 bp for the primer set LCO1490/HCO2198 (Folmer et al., 1994)) might benefit from being performed on longer, multi kbp amplicons.
- 4) *PCR amplification on diluted DNA extract.* Mitochondrial genomes are typically more abundant than nuclear ones since mononucleated

cells contains several hundreds of mitochondria, each mitochondrion containing itself several mtDNA molecules (Sorenson and Quinn, 1998). Thus, it should be possible to extinguish *numts* by performing PCR assays on extreme dilutions of DNA extracts.

To our knowledge these four methods were only applied in a handful of published cases having dealt with *numts*: method 1 (Collura and Stewart, 1995; Ibarguchi et al., 2006; Kidd and Friesen, 1998; Thalmann et al., 2004, 2005); method 2 (Collura et al., 1996; Williams and Knowlton, 2001); method 3 (Collura and Stewart, 1995; Ibarguchi et al., 2006; Thalmann et al., 2004, 2005); and method 4 (Ibarguchi et al., 2006; Kidd and Friesen, 1998). Moreover, they have never been applied simultaneously to the same *numt* cases, thereby preventing a direct assessment of their respective merits. This article specifically aims to fill this gap by applying these four bench methods to five cases of crustacean *numts*.

2. Material and methods

2.1. Sampling of specimens and selection of *numt* cases

Specimens were collected in the field in the course of a large European sampling scheme of species belonging to the superfamily Aselloidea (Crustacea, Isopoda, Asellota). Immediately after collection, specimens were placed in 96% ethanol at ambient temperature for transportation back to the laboratory (a few hours to a few days), then at 4 °C until being assigned to a species on morpho-anatomical grounds (a few days to a few weeks later), and finally at –20 °C until DNA extraction. In our view, such a storage procedure reflects the daily practice of field sampling, which often precludes immediate state-of-the-art preservation. Ordinary procedure then involved DNA extraction, PCR using a variety of primer pairs targeting cytochrome oxidase I (COI) or 16S mitochondrial rDNA (16S) fragments and direct sequencing.

Co-amplification of at least two mtDNA-like sequences, indicated by apparent polymorphism in chromatograms (double peaks), was common (a short discussion about *numts* in the particular case of invertebrates is provided as Supplementary File 1). Therefore, we amplified and sequenced the same regions twice using different primer pairs to favor identification of mtDNA sequences (Sorenson and Quinn, 1998). Most of the time, this strategy, which was applied to 50% of sampled individuals, did not allow extinguishing co-amplified signal. In some cases where unambiguous chromatograms had been initially determined, we obtained different sequences or observed different behaviors depending on which primer pair was used. We selected five such cases to assess the efficiency of anti-*numt* methods, namely (with the targeted marker between parentheses): *Proasellus aquaecalidae* (COI), *Proasellus ortizi* (COI), *Proasellus aragonensis* (COI), *Proasellus ligusticus* (16S) and *Proasellus banyulensis* (COI). All five instances are described in Table 1 and in Supplementary File 2. A total of four individuals from each species were used for analyses. The DNA of two individuals per species was extracted using a chloroform-based protocol and served for analyses applying methods 1, 3 and 4, while the other two individual's DNA and RNA were extracted using TRI reagent®, which allowed applying method 2 (Fig. 1 and below).

2.2. Methods used to screen for the presence of *numts*

The general pattern of the experiments is summarized in Fig. 1 and detailed thereafter.

2.2.1. Method 1. MtDNA and nuclear DNA enrichment and extraction

Mitochondria isolation and DNA extraction. To enrich DNA extracts in mtDNA, we used the small-scale isolation procedure described in

Table 1
Numt cases under study.

Species	Population Country	Targeted marker	Primer pair	Sequence obtained
<i>Proasellus aquaealidae</i> (Racovitza, 1922)	Cave Petites Eaux Chaudes France	COI	Pair 1: COI_PwaltF1 HCO2198 Pair 2: LCO1490 COI_PstrouR1	<i>numt</i> (frameshift-inducing deletion) Putative mtDNA
<i>Proasellus ortizi</i> Henry and Magniez, 1992	Cave Jivero Spain	COI	Pair 3: COI_PsynaF1 HCO2198 Pair 2	<i>numt</i> (high divergence and short branch in phylogenetic reconstruction) Putative mtDNA
<i>Proasellus aragonensis</i> Henry and Magniez, 1992	River Esca at Burgui Spain	COI	Pair 3 Pair 4: LCO1490 COI_PsynaR1	Co-amplification putative mtDNA and <i>numt</i> (‘messy’ chromatogram profile) Putative mtDNA
<i>Proasellus ligusticus</i> Bodon and Argano, 1982	River Maridanasca at Carrodano Italy	16S	Pair 5: 16S_ProaF1 16S_br Pair 6: 16S_StenaF1 16S_StenaR1	<i>numt</i> (high divergence, deletion and polymorphic sites in chromatogram) Putative mtDNA
<i>Proasellus banyulensis</i> (Racovitza, 1919)	Well Sainte-Marie de Gaja France	COI	Pair 1 Pair 7: LCO1490 COI_PwaltR1	Co-amplification putative mtDNA and <i>numt</i> (polymorphic sites in chromatogram) Putative mtDNA

Beckman et al. (1993), whose efficiency was recently confirmed (Ibarguchi et al., 2006). Briefly, whole organisms were crushed in 1 mL of homogenization buffer (100 mM Tris-HCl, pH7.4, 250 mM sucrose, 10 mM EDTA). Homogenates were centrifuged at 1500 g for 10 min at 4 °C. Supernatants (which presumably contained mitochondria) were removed and placed in fresh tubes while the pellets, containing large cellular debris (including nuclei, thus called there after nuclear pellets), were kept at 4 °C until further analysis.

Supernatants were then centrifuged at 10,000 g for 10 min at 4 °C. The resulting pellets were made of mitochondria (animal mitochondria sediment in sucrose 250 mM). DNA was purified from pellets using the Wizard Miniprep kit, following manufacturer’s instructions (Promega, Mannheim, Germany).

Chloroform-based DNA extraction of nuclear pellets. We also extracted DNA from nuclear pellets because we wanted (i) to examine the

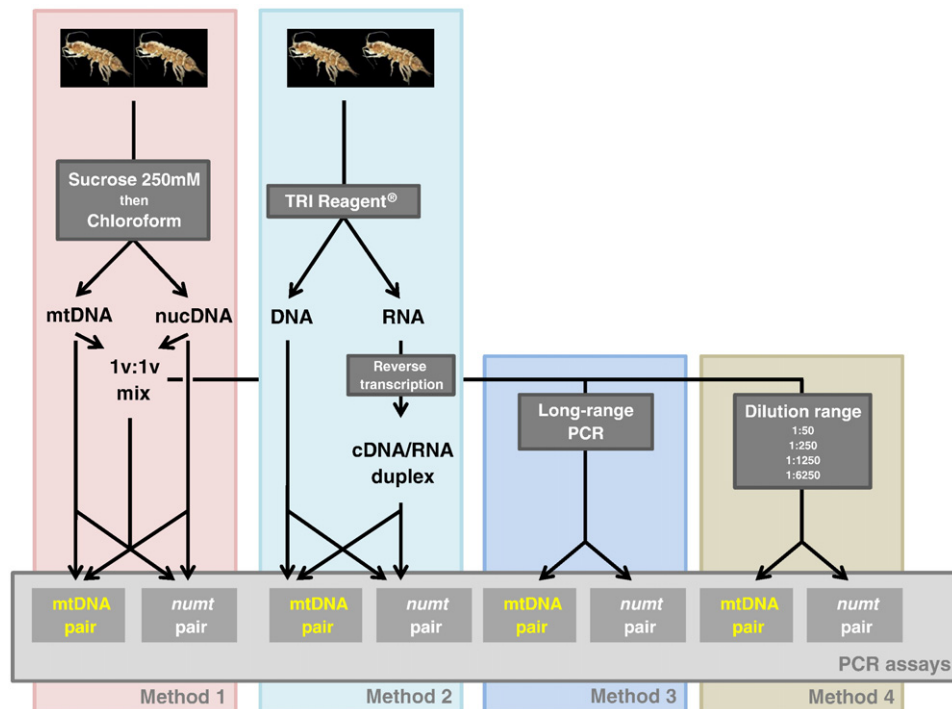


Fig. 1. General organization of the experiments. For each possible *numt* case, 2 individuals were submitted either to RNA/DNA co-extraction (method 2) or to mtDNA/nuclear DNA fractionation (method 1). Then, a mix of both fractions (mimicking a global DNA extraction) provided the substrate for long-range pre-amplification (method 3) and dilution ranges (method 4). PCR assays were finally performed using ‘mtDNA pairs’ and ‘*numt* pairs’. Photo credit: Claire Morvan.

reciprocal purity of mitochondrial/nuclear compartments and (ii) to assess if this protocol would be compatible with studies focusing on both mtDNA and non-*numt* nuclear markers. Besides, extracting nuclear pellet DNA was necessary to allow for the assessment of *numt* presence on an individual basis (developed later).

The chloroform protocol applied here was our routine protocol for DNA extraction from whole organisms (thereafter referred to as chloroform DNA extraction). Nuclear pellets were suspended in 200 μ L of digestion buffer (NaCl 0.1 M, EDTA 0.01 M, SDS 0.5%, Tris-HCl 0.05 M) and digested with proteinase K (final concentration 15 mg/mL) for 1 h at 56 °C. Homogenates were centrifuged for 5 min at 13,200 g at ambient temperature. Supernatants were transferred to fresh tubes and 250 μ L of NaCl 2.6 M were added before vortexing and centrifuging for 15 min at 13,200 g. Again, supernatants were transferred to fresh tubes and 400 μ L of chloroform was added before prolonged vortexing and centrifuging for 5 min at 13,200 g. The resulting aqueous phases were removed and placed in fresh tubes with glycogen (co-precipitant, final concentration 40 μ g/mL) and 600 μ L of 96% ethanol. Tubes were then gently mixed by inversion and placed for at least 45 min at –80 °C. Thereafter, tubes were centrifuged 20 min at 4 °C. Supernatants were thrown away and pellets washed in 70% ethanol. Again, supernatants were removed and pellets were allowed to air dry for up to 2 h. Dry pellets were finally suspended in TE (Tris-HCl 10 mM, EDTA 1 mM).

2.2.2. Method 2. RNA and DNA purification and extraction

RNA extraction, DNA decontamination and cDNA generation from RNA extraction. In order to prevent any contamination with RNAses (especially given the limited amount of RNA expected to have persisted in ethanol-preserved specimens) all tubes used in our experiments were RNase free (Axygen, Union City, U.S.A.) and diethylpyrocarbonate (DEPC) treated water was used to prepare all reagents. The method employed relied on the use of TRI Reagent® (Molecular Research Center, Cincinnati, U.S.A.), a complete, ready-to-use reagent theoretically allowing the isolation of RNA, DNA and protein fractions from a single sample, according to a protocol slightly modified from the manufacturer (Colson-Proch et al., 2009). Whole organisms were first crushed in 300 μ L of TRI reagent® at ambient temperature then 300 μ L of TRI reagent® was added, the homogenate vigorously vortexed and incubated for 5 min at ambient temperature. Two hundred microlitres of chloroform was added before vigorously vortexing the mix and incubating it for 10 min at 37 °C. The homogenate was then centrifuged for 15 min at 12,000 g at 4 °C. At that stage RNA partitions to the upper aqueous phase, DNA to the interphase and proteins to the lower organic phase. RNA-containing upper phases were removed and placed in fresh tubes while interphases and lower phases were kept at 4 °C until being analyzed. Aqueous phases were submitted to a new cycle of TRI reagent® (450 μ L)/chloroform (160 μ L) extraction in the same conditions as above. Following upper phases were placed in fresh tubes and 500 μ L of isopropanol was added before vigorously vortexing the mix and incubating it 10 min at ambient temperature. Tubes were then centrifuged for 10 min at 12,000 g at 4 °C. Supernatants were discarded and pellets were washed in 900 μ L of 75% ethanol, slightly vortexed and centrifuged for 10 min at 12,000 g at 4 °C. Again, supernatants were thrown away before a last centrifugation for 2 min at 12,000 g at 4 °C. Remaining ethanol was discarded by pipetting and allowed to evaporate for a few minutes before pellets were solubilized with DEPC treated water.

RNA extracts were treated with Turbo DNase (Ambion, Austin, U.S.A.) following manufacturer's instructions, in order to eliminate any contaminating DNA. To ascertain that DNA had not contaminated RNA extracts, control PCRs targeting 16S were performed on RNA before reverse transcription, which consistently yielded negative results. Reverse transcription of poly-adenylated RNA was then performed using a Moloney murine leukemia virus reverse transcriptase defective for RNase H activity (M-MLV RT RNase H⁻; Promega, Mannheim,

Germany) and RNasin (Promega, Mannheim, Germany) in the presence of 15-mers of thymine, according to manufacturer's recommendations. Thus, RNA/cDNA duplex served as substrate for any following PCR-based analyses.

DNA co-extraction. We assessed the efficacy of DNA co-extraction when using TRI reagent® because we had to check for the presence of *numts* and because we also wished to keep the option of targeting non-coding regions. Preliminary experiments revealed that applying manufacturer's instructions did result in poor DNA yields, whether co-precipitant (glycogen) was added or not (manufacturer's recommendation for < 10 μ g DNA yields). Better results were obtained by directly submitting interphases to our chloroform DNA extraction method (paragraph 2.2.1).

2.2.3. Method 3. Long-range amplification

A number of commercial kits are now available that allow for very long-range amplification of fragments longer than 20 kbp. However, in our experience, it is common that fragments longer than 4 to 5 kbp are reluctant to amplification, at least when practiced from ethanol-preserved specimen DNA extracts. Therefore, we chose to target one ca. 3 kbp fragment spanning the region between 16S and COI (primers used were HCO2198 and 16Sbr (Folmer et al., 1994; Simon, 1991)). This fragment could be easily amplified using classic *Thermophilus aquaticus* (*Taq*) polymerases, on condition that cycling conditions were adapted (thereafter adaptations with respect to standard PCR conditions described in paragraph 2.3.2 are in bold). **One hundred to 200 ng** of 1v:1v mixture of mitochondrial and nuclear DNA extracts (obtained through method 1) served as template in 40 μ L reactions containing **0.36 mM dNTPs** (Eurogentec, Liege, Belgium), 0.1 mg/mL BSA (New England Biolabs, Ipswich, U.S.A.), **0.5 mM primers**, 0.15 U *Taq* DNA polymerase and 1 \times PCR buffer (New England Biolabs, Ipswich, U.S.A.). PCRs were performed using the following settings: (i) one step of **1 min at 94 °C**, (ii) 40 cycles of 15 s at 94 °C, 20 s at 48–53 °C, **3 min 30 s at 68 °C**, and (iii) one step of **5 min at 68 °C**. (Note that temperature at elongation was of particular importance as most *Taq* DNA polymerases turned out to be inactivated by long expositions to heat: the same protocol with elongation at 72 °C greatly reduced PCR success). Five to 35 μ L of each reaction were used for migration in 1.3% agarose gels. Where amplicons of the expected size were visible, gel slices were cut out from the gel and simply centrifuged through a sterile filtered pipette cone into a fresh tube. The resulting gel-purified solution was kept at 4 °C until secondary PCR amplification.

2.2.4. Method 4. Dilution ranges

Dilution ranges were conceived so as to ultimately reduce the number of nuclear genome per reaction to less than one copy. In standard conditions (no dilution) about 12.5 ng of genomic DNA is used per 35 μ L reaction. Taking into account the 5 C-values presently available from *Asellota* (*Proasellus meridianus* 1.92 pg, *Proasellus coxalis* 1.3 pg, *Asellus aquaticus* 2.52 pg (Rocchi et al., 1989), *Caecidota forbesi* 1.9 pg and *Caecidota racovitzae* 1.71 pg (Gregory, 2010)) and recent results suggesting that C-values were indeed approximately comprised between 1.5 and 2 pg in the genus *Proasellus* (Lara Konecny and Christophe Douady, unpublished results), we made the assumption that the diploid genome size in all species examined here was ca. 3.5 pg. As a consequence, standard reactions could be considered as starting in the presence of ca. 3500 complete nuclear genomes. Therefore, we chose to dilute as follows: 1:50, 1:250, 1:1250 and 1:6250, thus ensuring that any PCR reaction performed on the most extreme dilution would contain less than one nuclear genome. Here also 1v:1v mixture of mitochondrial and nuclear DNA extracts (obtained through method 1) were used as the starting material for dilution ranges.

2.3. Downstream molecular methods

2.3.1. DNA and RNA quantitation

Between 1 and 2 μL of DNA or RNA extract were used to quantify DNA using a Nanodrop® ND1000 device (Labtech International, Ringmer, U.K.) according to manufacturer's instructions.

2.3.2. PCR assays

DNA/cDNA extracts and gel-purified amplicons were submitted to PCR amplification attempts as described in Fig. 1 using primer pairs detailed in Table 1 and Supplementary File 2. Generally, 0.5 to 1 μL of DNA/cDNA extract or gel-purified amplicon was used as template in 35 μL reactions otherwise containing 0.17 mM dNTPs, 0.1 mg/mL BSA, 0.16 mM primers (primer sequences are available in Supplementary File 2), 0.15 U *Taq* DNA polymerase and 1 \times PCR buffer. PCR was performed using the following settings: (i) one step of 2 min at 94 °C, (ii) 40 cycles of 15 s at 94 °C, 20 s at 48–53 °C, 30 s at 72 °C, and (iii) one step of 5 min at 72 °C. PCR assays were considered positive whatever the intensity of the band (comprising very faint bands). Negative results were replicated once. In most cases, no band was visible when replicating negative results. When results were conflicting (first assay negative, second assay positive) amplicons were sent to sequencing: when sequencing failed the end result was considered negative, when it succeeded the overall result was considered positive.

2.3.3. DNA sequencing and sequence analysis

We reasoned that direct sequencing of PCR products would both better represent daily practice and yield to a clear yes/no answer about the ability of each method to bring putative mtDNA sequences up to the level of “most frequent mtDNA-like sequences” present in PCR products (which is the final objective of any method devised to counter *numts*). Therefore, we did not clone amplicons but directly sequenced them. Sequencing was performed by service providers using Sanger's methodology (GATC Biotech; Konstanz, Germany; Eurofins MWG Operon, Ebersberg, Germany). Chromatograms were visualized using FinchTV and sequences aligned in SeaView v4 (Gouy et al., 2010). The latter was also used where sequence manipulation (e.g. amino acid translation) or phylogenetic analyses were needed (Supplementary File 2).

3. Results and discussion

Our purpose was to assess the relative performance of four bench methods which could prove valuable as anti-*numt* strategies. To this end, we applied these methods to five species cases where *numt* amplification was strongly suspected (Tables 1 and 2 and Supplementary File 2). It should be noted here that in the following paragraphs, mtDNA sequences identified by our analyses will consistently be referred to as putative mtDNA sequences. This choice should not be understood as reflecting experimental uncertainties (evidences of mitochondrial or nuclear origin are provided in Supplementary File 2) but rather as highlighting the objective impossibility to reach a 100% confidence about the mitochondrial origin of any mtDNA-like sequence (i.e. it is a conservative wording).

3.1. Individual assessment regarding *numts*

When assessing anti-*numt* methods, one is faced with a major challenge: *numts* have to be identified at the individual scale. For instance, Song et al. (2008) reported that only 18 of 95 *Orconectes barri* and *Orconectes australis* individuals allowed amplification and clean direct sequencing of *numts*, highlighting a strong individual heterogeneity. This can affect methodological investigations, e.g. if one observes that a given *numt* is preferentially amplified from a given individual and if this individual DNA has been extracted using a

chloroform protocol, the same individual is no longer accessible to other extraction methods (particularly when small invertebrates are under study) and other individuals cannot be guaranteed to behave in the same way. Our experimental system took this constraint into account by performing *de novo* DNA extractions for all individuals, thereby allowing their status regarding *numts* to be checked on a case-per-case basis (Fig. 1).

Our experimental system notably allowed three of the four methods (methods 1, 3 and 4) to be tested on the same 10 individuals (two individuals per species) (Fig. 1). Method 1 indeed ended up with generating two DNA extracts: one from the mitochondrial compartment, the other from the nuclear compartment. By mixing volume per volume (1v:1v) part of these two DNA extracts, the original mix of DNA (mtDNA and nuclear DNA) in the organism could be mimicked (even though imperfectly). Results of PCR assays led on these 1v:1v mixtures of mitochondrial and nuclear fractions informed individual situations with respect to *numts* (Tables 1 and 2 and Supplementary File 2).

For *P. aquacalidae*, individuals 1 and 2 yielded an identifiable *numt* sequence (frameshift-inducing deletion) with primer pair 1 and a putative mtDNA sequence with primer pair 2. For *P. ortizi*, a putative mtDNA sequence was obtained with primer pair 2 from both individuals while primer pair 3 amplified a very divergent sequence (ca. 17%), whose branch length in a phylogenetic reconstruction was much shorter than that of the putative mtDNA sequence, thereby suggesting a possible nuclear origin. For *P. aragonensis* only one primer pair (primer pair 4) allowed for examination of a clean sequence from individuals 1 and 2, which was considered of putative mitochondrial origin. By contrast, primer pair 3 consistently came to ‘messy’ chromatogram profiles. Those resembled examples shown in Buhay (2009) and Collura et al. (1996) which were interpreted as resulting from *numt* co-amplification. We retained this interpretation and assumed that primer pair 3 co-amplified *numts* and possibly mtDNA. Both *P. ligusticus* specimens led to the amplification of a putative mtDNA sequence with primer pair 6 while primer pair 5 likely co-amplified putative *numts* and possibly mtDNA, with the resulting sequence exhibiting a five base pair deletion as well as double peaks in chromatograms. Finally *P. banyulensis* individuals allowed the amplification of a putative mtDNA sequence with primer pair 7 while polymorphic sites appeared in chromatograms when primer pair 1 was employed, which suggested that the putative mtDNA sequence was co-amplified with one or several *numts*. All evidences are shown and further discussed in Supplementary File 2.

All in all, these 5 cases covered much of the possible outcomes of *numt* co-amplification: identified, unambiguous *numt* (*P. aquacalidae* (COI)), suspected but not identified *numts* (covering translated or non-translated mtDNA portions; *P. ortizi* (COI) and *P. ligusticus* (16S)), failure to properly sequence amplicons due to a mixture of *numts* and possibly mtDNA sequences (*P. aragonensis* (COI)), ambiguous chromatograms (*P. banyulensis* (COI) and *P. ligusticus* (16S)). In the rest of this article, primer pairs amplifying putative mtDNA sequences will be called ‘mtDNA pairs’ while those presumably (co-)amplifying *numts* will be called ‘*numt* pairs’.

It was not possible to apply method 2 (based on RNA extraction) to the same individuals as those examined for the other three methods because the small size of the organisms precluded multiple extractions from the same individuals. Therefore, two additional individuals per species were used to investigate its efficiency. TRI Reagent® is depicted as allowing the sequential isolation of macromolecules, including DNA and RNA. Accordingly, it theoretically renders feasible to confirm *numt* cases on an individual basis, using DNA extracts.

PCR assays were led on such DNA extracts. Unfortunately, confirmation of *numt* presence was only possible for three individuals: one *P. aquacalidae* specimen (individual 3) and both *P. aragonensis* individuals (Table 3). In other cases, it was either

Table 2

MtDNA purification performance. +: amplification was positive but amplicon was not sequenced, -: amplification was unsuccessful, mt: amplification was positive; sequence corresponded to the putative mtDNA, numt: amplification was positive; sequence corresponded to the numt, mt + numt: amplification was positive; sequence corresponded to a mixture of putative mtDNA and numt (double peaks at polymorphic sites), no seq.: amplification was positive but amplicon yielded no interpretable sequence. Cases where fractionation would have avoided inadvertent identification of a numt sequence as an mtDNA one are indicated in green and bold.

Species	Individual	DNA concentration (ng/μL)		PCR assay 'mtDNA pair'			PCR assay 'numt pair'		
		Mt fraction	Nuclear fraction	Mt fraction	Nuclear fraction	1v:1v mixture	Mt fraction	Nuclear fraction	1v:1v mixture
<i>Proasellus aquaealidae</i>	1	3.2	47.9	+	+	mt	mt+numt	numt	numt
	2	1.1	41.4	+	+	mt	–	numt	numt
<i>Proasellus ortizi</i>	1	4.0	35.1	mt	–	mt	numt	–	no seq.
	2	3.5	41.8	mt	mt	mt	numt	numt	numt
<i>Proasellus aragonensis</i>	1	9.0	41.1	mt	mt	mt	no seq.	no seq.	no seq.
	2	9.4	68.4	mt	mt	mt	no seq.	no seq.t	no seq.
<i>Proasellus ligusticus</i>	1	2.9	21.8	mt	+	mt	numt	numt	numt
	2	1.9	30.9	mt	+	mt	numt	numt	numt
<i>Proasellus banyulensis</i>	1	5.3	68.0	mt	mt	mt	mt+numt	mt+numt	mt+numt
	2	4.0	37.0	mt	mt	mt	mt+numt	mt+numt	mt+numt

impossible to obtain amplicons from DNA extracts or numts were not apparent anymore (i.e. 'numt pairs' amplified mtDNA). This was in sharp contrast with results presented in Table 2 where all individuals confirmed numt problems for the same species and populations. Poor DNA yield (as seen in Table 3) could not explain by itself why PCR-based methods were so inefficient. When deposited on agarose gels, it became evident that not only was DNA rare in these extracts, it was also very fragmented when compared to chloroform-based nuclear fraction extracts, at such a point that only faint low-sized smears of DNA were visible (Fig. 2). Therefore, the low number of confirmed numt cases with this method probably found its origin in the nature of DNA molecules co-extracted using TRI Reagent®.

3.2. MtDNA enrichment (method 1)

It is expected that if mtDNA enrichment is sufficient to favor mtDNA over numt amplification the two fractions (nuclear and mitochondrial) would show differential behaviors over PCR assays. Indeed, 'mtDNA pairs' should yield PCR products when run on mitochondrial, but not nuclear, fractions (i.e. assuming nuclear fraction would not be contaminated with mtDNA). 'Numt pairs' should exhibit the reciprocal behavior, or if amplifying from mitochondrial fraction, yield putative mtDNA sequence. From the 5 cases under scrutiny only one, that of *P. aquaealidae*, led to such an observation (individual 1: nucDNA fraction = numt, mtDNA fraction = polymorphism, i.e. mtDNA and numt; individual 2: nucDNA fraction = numt; mtDNA fraction = failure to amplify; Table 2).

Table 3

RNA extraction performance. PCR success and results of sequencing are indicated in the same way as in Table 2. nd: not determined, * indicates that the value is the mean of two measures. Grayed lines stand for individuals for which the original problem could not be confirmed. Cases where cDNA amplification would have avoided inadvertent identification of a numt sequence as an mtDNA one are indicated in green and bold.

Species	Individual	Concentration (ng/μL)		PCR assay 'mtDNA pair'		PCR assay 'numt pair'	
		RNA	DNA	RNA	DNA	RNA	DNA
<i>Proasellus aquaealidae</i>	3	nd	6.9	+	+	mt	numt
	4	142.2*	2.6	–	–	–	–
<i>Proasellus ortizi</i>	3	179.3	53.5	–	–	–	–
	4	65.0	26.7	–	–	–	–
<i>Proasellus aragonensis</i>	3	80.0	17.1	mt	mt	no seq.	no seq.
	4	30.8	7.9	+	mt	–	no seq.
<i>Proasellus ligusticus</i>	3	57.1*	7.4	mt	mt	mt	mt
	4	147.6*	3.8	+	mt	mt	mt
<i>Proasellus banyulensis</i>	3	45.3*	13.8	+	+	–	mt
	4	157.3*	58.0	+	–	–	mt

MtDNA enrichment has been used relatively rarely to attempt unmasking numts over the last two decades whereas it was a standard procedure in the pre-PCR era (e.g. Anderson et al., 1981). Interestingly, the few results available are consistent with our findings which suggest a moderate efficiency of the method. MtDNA enrichment failed to identify the genuine mtDNA sequence in Sumatran orangutans (Collura and Stewart, 1995), before being used in conjunction with other methods to allow identifying mitochondrial copies in guillemots (Kidd and Friesen, 1998) and gorillas (Thalmann et al., 2004, 2005). More recently, Iburguchi et al. (2006) investigated the efficacy of various mtDNA purification protocols by targeting mitochondrial or nuclear markers, but not necessarily numts. These authors found that in roughly 45% cases, one or another method would come to identifiable segregation of nuclear and mitochondrial DNA, even though reciprocal contamination could not be excluded (Iburguchi et al., 2006). However, this proportion integrated variable success rates depending on the tissue sources and mtDNA enrichment methods. For the method of purification used in the present study, four of four samples resulted in visible enrichment, but all samples were taken from avian liver or heart, which are mitochondria-rich (Iburguchi et al., 2006). When, in the same study, two small invertebrates were submitted to another mitochondrial enrichment protocol, it performed poorly, only allowing moderate enrichment in one case, while both samples had been fresh-frozen at –80 °C.

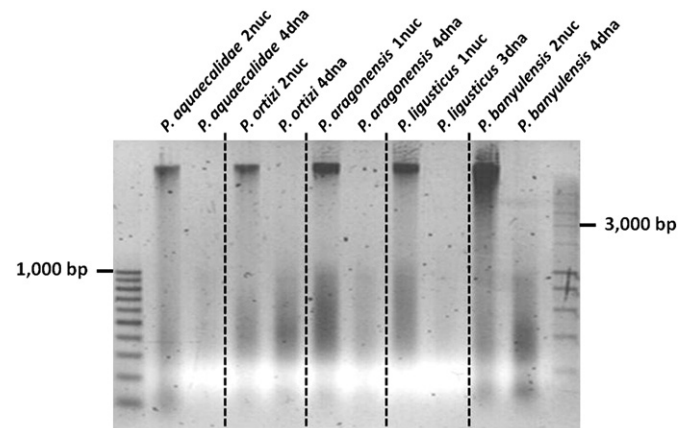


Fig. 2. Total DNA extracted thanks to chloroform or TRI reagent® protocols. Ten μL of total DNA was deposited on a 1.3% agarose gel which was run for 30 min at 100 V. The gel was then soaked into a solution of 10 mg/mL ethidium bromide and photographed under UV exposure. Names of specimens (i.e. species names and individual numbers) are followed by terms: nuc: chloroform extraction of nuclear fraction; dna: TRI reagent® extraction.

Our results further highlight that mtDNA enrichment is not an adequate strategy to remove *numts*, at least when working on non-ideal sources such as small invertebrates preserved in ethanol. One possible explanation could lie into the disruption of mitochondrial membranes during freezing which results in reciprocal contamination of the two fractions, precluding any attempt to identify mtDNA-like sequences identified from nuclear fractions as *numts* and from mitochondrial fractions as mtDNA. The phenomenon might however be less deleterious (from the standpoint of the enrichment itself) when the starting material is mitochondria-rich (e.g. liver from vertebrates or possibly brain/neural tissue from invertebrates). Then, the probability of conserving a significant number of entire, 'sedimentable' mitochondria would presumably be higher.

3.3. RNA extraction (method 2)

RNA extraction was rather successful as shown by the deposition and migration of total RNA extracts on agarose gels (Fig. 3). However, RNA extracts from ethanol-preserved specimens revealed poor quality as compared to RNA extracted from fresh-frozen individuals. RNA generally appeared as a low-sized smear (<1500 bp), when RNA from fresh-frozen individuals appeared as high-sized smear with intense bands corresponding to rRNAs (Fig. 3). Quantitation with a Nanodrop device yielded aberrant estimates, unrelated to trends observed in agarose gels (Table 3). This could either relate to faulty interpretation of gel smears (e.g. wider smears could be misleadingly interpreted as reflecting lesser RNA content) or to long soaking in ethanol, though we have no biochemical explanation to this last phenomenon.

Reverse transcribed RNA extracts allowed successful amplifications from seven individuals from four species (except *P. ortizi*) with the 'mtDNA pair' and from four individuals from three species (except *P. ortizi* and *P. banyulensis*) with the 'numt pair' (Table 3). However, the efficiency of the strategy could only be assessed from results obtained on *P. aquaealidae* individual 3 and on *P. aragonensis* individuals (the only ones for which *numt*-related problems were confirmed from DNA extracts). In those cases, the 'numt pair' recovered the putative mtDNA sequence when performed on RNA/cDNA duplex from *P. aquaealidae* individual 3 and failed to yield any PCR product when performed on RNA/cDNA duplex from *P. aragonensis* individual 4 (Table 3). For the remaining individual (*P. aragonensis* individual 3), the same results were obtained from DNA or RNA/cDNA templates: mtDNA sequences were amplified with the "mtDNA pair" and no sequence could be determined using the "numt pair" (Table 3). If one makes the assumption that all individuals

were undetected cases of *numt* co-amplification, then it is worth noticing that RNA/cDNA duplex-based PCR amplification with 'numt pairs' never yielded *numt* sequences (Table 3).

RNA extraction was first suggested by Collura et al. (1996) who identified the proper mtDNA sequence of proboscis monkeys. More recently, it was used on *Alpheus* shrimps, also allowing the recovery of unique putative mtDNA sequences (Williams and Knowlton, 2001). Our results further exemplify the efficiency of the method, with two well-characterized instances and four favorable cases. They are also interesting in that they were obtained from animals preserved in ethanol (proboscis monkey and *Alpheus* shrimp samples were fresh-frozen (Collura et al., 1996; Williams and Knowlton, 2001)). To our knowledge, this is the first report of successful RNA extraction from that kind of samples. Given that ethanol is commonly used for cryo-collection preservation, this could considerably broaden the possibility of using RNA extraction as a means to get rid of *numts*, though the quality of such extracts would certainly deserve further investigation. It should also be noted that, if one has interest in non-transcribed sequences, the question will be raised regarding the suitability of co-extracted DNA for PCR analyses, at least in the case of small invertebrates.

3.4. Long-range amplification (method 3)

We tested the possibility to change the ratio of *numt* to mtDNA templates in reaction through preliminary long-range PCR, based on the observation that *numts* are generally much shorter than the whole mtDNA (notable exceptions exist (Lopez et al., 1994)). Three kbp-long amplicons were obtained for all individuals but bands were of variable intensity (Table 4). Results of PCR assays showed that in seven cases 'numt pairs' revealed either unable to amplify from three kbp amplicons or, where amplification was successful, led to obtaining putative mtDNA sequences or a mixture of putative mtDNA and *numt* sequences (Table 4). Therefore, inadvertent identification of *numts* as mtDNA would not have been possible here.

Long-range pre-amplification was used by Collura and Stewart (1995) on the orangutan case but did not allow for ruling out *numts*. A variant method was used by Thalmann et al. (2004) but it essentially relied on the circularity of mtDNA with respect to nuclear DNA. Recently, the method was applied to seedsnipes and found to be inefficient at excluding *numts* (Ibarguchi et al., 2006). Finally, attempts to amplify longer fragments than usual (slightly >1 kbp) have been relatively frequent and unsuccessful (e.g. Kolokotronis et al., 2007; Pereira and Baker, 2004). This relatively poor performance could find its roots in an initial overestimation of the benefits of the method. Nevertheless, our results, together with the

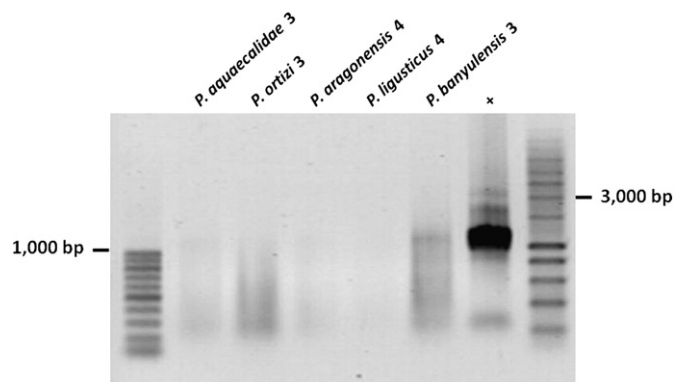


Fig. 3. Total RNA extracted from ethanol-preserved specimens. One and a half μ l of total RNA extract was deposited on a 1.3% agarose gel which was run on ice for 30 min at 100 V. The gel was then soaked into a solution of 10 mg/mL ethidium bromide and photographed under UV exposure. For comparison, a living *Proasellus meridianus* (Rhône basin, Balan, France) was submitted to RNA extraction too and its RNA deposited on the same gel (lane +).

Table 4

Long-range PCR performance. PCR success and results of sequencing are indicated in the same way as in Table 2. \pm : 3 kbp amplification was not as intense as in the other cases; nd: not determined (as POJ5 semi-nested amplification did not yield sequenceable amplicons, POJ4, whose 3 kbp amplicon was much fainter, was not tested). Cases where long-range amplicon-based amplification would have avoided inadvertent identification of a *numt* sequence as an mtDNA one are indicated in green and bold.

Species	Individual	Long-range PCR	PCR assay 'mtDNA pair'	PCR assay 'numt pair'
<i>Proasellus aquaealidae</i>	1	+	+	mt+numt
	2	+	+	–
<i>Proasellus ortizi</i>	1	\pm	nd	nd
	2	+	no seq.	–
<i>Proasellus aragonensis</i>	1	\pm	–	–
	2	+	mt	–
<i>Proasellus ligusticus</i>	1	+	+	mt
	2	+	+	mt
<i>Proasellus banyulensis</i>	1	\pm	+	–
	2	\pm	+	–

observation that some genomes seem to lack large *numts* (e.g. no *numt* larger than 1280 bp was found in *A. mellifera* genome (Pamilo et al., 2007)), show that the method should not be dismissed *a priori* and appeal at larger censuses of its efficiency. Such censuses would certainly benefit from investigating the question of the “appropriate” size of long-range amplifications on a broad taxonomic spectrum.

3.5. Dilution (method 4)

PCR assays were performed on dilution ranges so as reactions were expected to start in the presence of less than 1 to more than 3500 nuclear genomes per reaction (corresponding to the use of 6250-fold diluted to non-diluted DNA extracts). Fig. 4 summarizes the outcome of these experiments. ‘MtDNA pairs’ and ‘*numt* pairs’ exhibited strikingly different behaviors: with the former, amplicons were obtained in all cases down to maximum dilution (d6250), with the latter, PCR success progressively decreased, no amplicon being visible at maximum dilution (Fig. 4). Sequencing of dilution amplicons was then performed on an individual basis, the maximum dilution from which a sequence could be obtained being indicated in Table 5. In nine cases, ‘mtDNA pairs’ allowed recovering sequences deeper into the dilution range than ‘*numt* pairs’. In seven instances the maximum dilution (d6250) yielded interpretable sequences with ‘mtDNA pairs’ while with ‘*numt* pairs’ any dilution impeded obtaining any interpretable sequence for six individuals (Table 5).

Dilution ranges were first used by Kidd and Friesen (1998) on guillemot DNA extracts. However, these assays were led on enriched mitochondrial and nuclear fractions so that the observed success could only be linked marginally to dilution. Ibaruchi et al. (2006) actually produced the first estimate of the efficiency of the method by itself. They diluted two seedsnipe DNA extracts down to d500 and still observed mitochondrial and nuclear amplicons. As birds exhibit average genome sizes close to some crustaceans (Gregory, 2005), our dilution ranges are roughly comparable. At d250 more than half of our samples yielded amplicons, which is consistent with Ibaruchi et al.'s (2006) findings. Thus, the inefficiency reported by these authors might result from not having diluted enough DNA extracts. By contrast in our case, only ‘mtDNA pairs’ amplified most extreme dilutions and led to interpretable sequences from dilutions higher than d50, which supports the view that dilution would enhance dramatic behavioral differences between any *numt* and its mtDNA counterpart. An important confounding factor might have been that of the performance of primer pairs *per se*. However, we are confident that primer properties did not influence notably the efficiency of dilution because we used a total of seven primer pairs. Therefore, dilution appears as a simple and robust means to avoid *numt* amplification and sequencing.

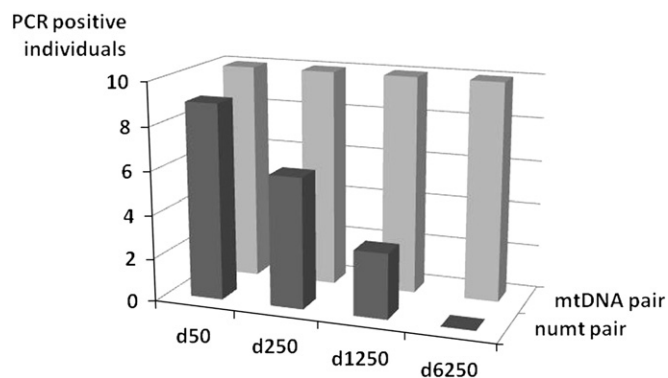


Fig. 4. PCR success on dilution ranges using ‘mtDNA pairs’ and ‘*numt* pairs’. Individuals were considered PCR positive when bands were visible (whatever their intensity).

Table 5

Successful sequencing on dilution range amplicons. The last successful sequencing is indicated for each individual (i.e. the last dilution to have provided an interpretable sequence). dX: diluted X-fold (d1 is non-diluted). Results of sequencing are indicated in the same way as in Table 2. *: band was extremely faint and accordingly was not sent to sequencing. Cases where dilution would have avoided inadvertent identification of a *numt* sequence as an mtDNA one are indicated in green and bold.

Species	Individual	PCR assay 'mtDNA pair'	PCR assay ' <i>numt</i> pair'
<i>Proasellus aquaecalidae</i>	1	d1250 mt	d50*
	2	d6250 mt	d1 <i>numt</i>
<i>Proasellus ortizi</i>	1	d50 mt	d50 <i>numt</i>
	2	d250 mt	d50 <i>numt</i>
<i>Proasellus aragonensis</i>	1	d6250 mt	d1 no seq.
	2	d6250 mt	d1 no seq.
<i>Proasellus ligusticus</i>	1	d6250 mt	d1 <i>numt</i>
	2	d6250 mt	d1 <i>numt</i>
<i>Proasellus banyulensis</i>	1	d6250 mt	d50 mt+<i>numt</i>
	2	d6250 mt	d1 mt+<i>numt</i>

3.6. Conclusions

Numts pose a very concrete and serious challenge to those interested in determining mtDNA sequences, i.e. virtually all molecular ecologists and evolutionists but also, to a lesser extent, human geneticists (Yao et al., 2008). Two categories of strategies are implemented for managing this risk. Quality control of sequences and possibly subsequent PCR primer modifications rely on *a posteriori* analyses of sequences (Buhay, 2009; Song et al., 2008). Experience has proven that, even though largely publicized, it remains a neglected tool (see the impressive amount of COI *numts* inadvertently deposited in GenBank for crustaceans only; Buhay, 2009). Counter-measures taken *a priori* (to prevent the amplification of *numts*) constitute another possible leverage. In our point of view, such measures should be applied to any published dataset as a minimal assessment of its likelihood to be polluted by pseudogenes. According to our results and literature, mtDNA enrichment will not yield satisfying results unless high quality sources of mitochondria are available, which will often not be the case. The three other methods seem to present better efficiencies. However, preliminary long-range amplification will not be applicable to every taxon and will require case-per-case assessment (in our case, it proves rather satisfying). RNA extraction is more easily applicable, including in the case of ethanol-preserved cryo-collections. But for samples expected to yield low quantities of DNA, sequential extraction of RNA and DNA will be unsatisfying if one intends to target non-transcribed DNA sequences. Dilution is by far the easiest method and shows the best efficiency. Therefore, we advise to systematically proceed to control PCR amplifications of dilution ranges and subsequent sequencing of some PCR products from at least a sub-sample of any dataset. Of course, dilution ranges should be adapted to genome size estimates. Sequences determined from PCR products obtained from most extreme dilutions (down to less than a single copy per reaction) will likely be of organellar origin. In our view, this should be regarded as an addition, not a substitute, to the strategy consisting in using several independent primer pairs for any marker of interest. The fact that ‘*numt* pairs’ are often successful in amplifying mtDNA when placed in more favorable conditions (Tables 2–5) even suggests that confirming mtDNA-like sequences using multiple primer pairs (including those with a known preference for *numts*) in the context of *a priori* methods might be a robust way to ascertain their mitochondrial origin.

A posteriori methods will necessarily benefit from applying *a priori* ones. Sequences passed through the filters of *a priori* methods will indeed be more reliable (i.e. more likely to be of mitochondrial origin) and could be established as reference mtDNA sequences. Other methods could also help in establishing such a reference catalogue. In that order of idea, high-throughput DNA sequencing holds great

promise (Margulies et al., 2005), on condition that it is properly implemented. The recent publication of a new method of “fishing” within libraries prepared for high-throughput sequencing paves the way to large-scale characterization of authentic mitochondrial genomes. Primer extension capture (PEC) was successfully applied to Neanderthal remains, allowing for the recovery of five complete mtDNA genomes (Briggs et al., 2009). Since it is not dependent on specific PCR (though unspecific library amplification is part of the protocol), mtDNA-directed PEC should benefit from the favorable ratio of mtDNA versus nuclear DNA: any mtDNA-like sequence determined by mtDNA-directed PEC will have a very high probability to be of mitochondrial origin (on condition that probes are designed properly and target ultraconserved regions of all available mitogenomes of the group of interest). Assuming mtDNA content would be increased to just 10% of the total sequenced DNA (a reasonable assumption given Briggs et al. (2009) data), a single run on the 454 FLX system would thus allow generating the impressive amount of 250 reference mitogenomes at a 10× coverage. Such effort – which could of course only be supported within the framework of large research programmes – would likely help making the most-used mtDNA molecule an even more reliable tool.

Supplementary materials related to this article can be found online at doi:10.1016/j.mito.2010.10.004.

Acknowledgements

We are very grateful to Ludovic Orlando for his careful reading of a first draft of this manuscript as well as to all members of the group Groundwater Hydrobiology and Ecology for stimulating discussions. We also thank Jennifer Buhay and one anonymous reviewer for constructive comments about the latest versions of this manuscript. This research was funded by the Agence Nationale de la Recherche (ANR-08-JCJC-0120-01, «DEEP») and by the Institut Universitaire de France.

References

- Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R., Young, I.G., 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290, 457–465.
- Antunes, A., Ramos, M.J., 2005. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. *Genomics* 86, 708–717.
- Balloux, F., 2009. The worm in the fruit of the mitochondrial DNA tree. *Heredity* 104, 419–420.
- Beckman, K.B., Smith, M.F., Orrego, C., 1993. Purification of mitochondrial DNA with Wizard Minipreps DNA purification system. *Promega Notes* 43, 10–13.
- Bensasson, D., Zhang, D., Hartl, D.L., Hewitt, G.M., 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.* 16, 314–321.
- Briggs, A.W., Good, J.M., Green, R.E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Schmitz, R., Doronichev, V.B., Golovanova, L.V., de la Rasilla, M., Fortea, J., Rosas, A., Pääbo, S., 2009. Targeted retrieval and analysis of five Neanderthal mtDNA genomes. *Science* 325, 318–321.
- Buhay, J.E., 2009. “COI-like” sequences are becoming problematic in molecular systematic and DNA barcoding studies. *J. Crustacean Biol.* 29, 96–110.
- Collura, R.V., Stewart, C.B., 1995. Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. *Nature* 378, 485–489.
- Collura, R.V., Auerbach, M.R., Stewart, C.B., 1996. A quick, direct method that can differentiate expressed mitochondrial genes from their nuclear pseudogenes. *Curr. Biol.* 6, 1337–1339.
- Colson-Proch, C., Morales, A., Hervant, F., Konecny, L., Moulin, C., Douady, C.J., 2009. First cellular approach of the effects of global warming on groundwater organisms: a study of the HSP70 gene expression. *Cell Stress Chaperones* 3, 259–270.
- Dubey, S., Michaux, J., Brunner, H., Hutterer, R., Vogel, P., 2009. False phylogenies on wood mice due to cryptic cytochrome-b pseudogene. *Mol. Phylogenet. Evol.* 50, 633–641.
- Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3, 294–299.
- Galtier, N., Nabholz, B., Glemin, S., Hurst, G.D., 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol. Ecol.* 18, 4541–4550.
- Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–224.
- Gregory, T.R., 2005. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann. Bot.* 95, 133–146.
- Gregory, T.R., 2010. Animal Genome Size Database.
- Ibarguchi, G., Friesen, V.L., Lougheed, S.C., 2006. Defeating numts: semi-pure mitochondrial DNA from eggs and simple purification methods for field-collected wildlife tissues. *Genome* 49, 1438–1450.
- Jukes, T.H., Osawa, S., 1993. Evolutionary changes in the genetic code. *Comp. Biochem. Physiol. B* 106, 489–494.
- Kidd, M.G., Friesen, V.L., 1998. Sequence variation in the guillemot (*Alcidae: Cepphus*) mitochondrial control region and its nuclear homolog. *Mol. Biol. Evol.* 15, 61–70.
- Kolokotronis, S.O., Macphee, R.D., Greenwood, A.D., 2007. Detection of mitochondrial insertions in the nucleus (Numts) of Pleistocene and modern muskoxen. *BMC Evol. Biol.* 7, 67.
- Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., O'Brien, S.J., 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J. Mol. Evol.* 39, 174–190.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.
- Moulton, M.J., Song, H., Whiting, M.F., 2010. Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta). *Mol. Ecol. Res.* 10, 615–627.
- Olson, L.E., Yoder, A.D., 2002. Using secondary structure to identify ribosomal numts: cautionary examples from the human genome. *Mol. Biol. Evol.* 19, 93–100.
- Pamilo, P., Viljakainen, L., Vihavainen, A., 2007. Exceptionally high density of NUMTs in the honeybee genome. *Mol. Biol. Evol.* 24, 1340–1346.
- Pereira, S.L., Baker, A.J., 2004. Low number of mitochondrial pseudogenes in the chicken (*Gallus gallus*) nuclear genome: implications for molecular inference of population history and phylogenetics. *BMC Evol. Biol.* 4, 17.
- Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21, 1081–1084.
- Rocchi, A., Lanza, V., Di Castro, M., 1989. Surface spreading of synaptonemal complexes in three isopod crustacean species. *Genetica* 78, 125–132.
- Simon, C., 1991. Appendix 3. In: Hewitt, G.M., Johnston, A., Young, J.P. (Eds.), *Molecular Techniques in Taxonomy*. Springer-Verlag, pp. 345–355.
- Song, H., Buhay, J.E., Whiting, M.F., Crandall, K.A., 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci. USA* 105, 13486–13491.
- Sorenson, M.D., Quinn, T.W., 1998. Numts: a challenge for avian systematics and population biology. *Auk* 115, 214–221.
- Thalman, O., Hebler, J., Poinar, H.N., Paabo, S., Vigilant, L., 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. *Mol. Ecol.* 13, 321–335.
- Thalman, O., Serre, D., Hofreiter, M., Lukas, D., Eriksson, J., Vigilant, L., 2005. Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. *Mol. Ecol.* 14, 179–188.
- White, D.J., Wolff, J.N., Pierson, M., Gemmill, N.J., 2008. Revealing the hidden complexities of mtDNA inheritance. *Mol. Ecol.* 17, 4925–4942.
- Williams, S.T., Knowlton, N., 2001. Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. *Mol. Biol. Evol.* 18, 1484–1493.
- Yao, Y.G., Kong, Q.P., Salas, A., Bandelt, H.J., 2008. Pseudomitochondrial genome haunts disease studies. *J. Med. Genet.* 45, 769–772.