Fishing for barcodes in the Torrent: from COI to complete mitogenomes on NGS platforms

Damien Hinsinger, Régis Debruyne, Maeva Thomas, Gaël Denys, Marion I. Mennesson, Jose Utage, Agnès Dettaï

Research Article **Open Access**

Damien D. Hinsinger, Regis Debruyne, Maeva Thomas, Gaël P. J. Denys, Marion Mennesson, Jose Utge, Agnes Dettai

Fishing for barcodes in the Torrent: from COI to complete mitogenomes on NGS platforms

DOI 10.1515/dna-2015-0019
Received March 31, 2015; accepted September 8, 2015

Abstract: The adoption of Next-Generation Sequencing (NGS) by the field of DNA barcoding of Metazoa has been hindered by the fit between the classical COI barcode and the Sanger-based sequencing method. Here we describe a framework for the sequencing and multiplexing of mitogenomes on NGS platforms that implements (I) a universal long-range PCR-based amplification technique, (II) a two-level multiplexing approach (*i.e.* divergence-based and specific tag indexing), and (III) a dedicated demultiplexing and assembling script from an Ion Torrent sequencing platform. We provide a case study of mitogenomes obtained for two vouchered individuals of daces *Leuciscus burdigalensis* and *L. oxyrrhis* and show that this workflow enables to recover over 100 mitogenomes per sequencing chip on a PGM sequencer, bringing the individual cost down below 7,50€ per mitogenome (as of current 2015 sequencing costs). The use of several kilobases for identification purposes, as involved in the improved DNA-barcode we propose, stress the need for data reliability, especially through metadata. Based on both scientific and economic considerations, this framework presents a relevant approach for multiplexing samples, adaptable on any desktop NGS platform. It enables to extend from the prevalent barcoding approach by shifting from the single COI to complete mitogenome sequencing.

Keywords: DNA barcoding; mitogenome assembly; Next-Generation Sequencing; sample multiplexing; sequence post-processing

***Corresponding author: Agnes Dettai,** Institut de Systématique, Évolution, Biodiversité ISYEB, UMR 7205 CNRS, MNHN, UPMC, EPHE Muséum national d'Histoire naturelle, Sorbonne Universités. 57 rue Cuvier, CP30, 75005 Paris, France, E-mail: adettai@mnhn.fr
Damien D. Hinsinger, Institut de Systématique, Évolution, Biodiversité ISYEB, UMR 7205 CNRS, MNHN, UPMC, EPHE Muséum national d'Histoire naturelle, Sorbonne Universités. 57 rue Cuvier, CP30, 75005 Paris, France
Jose Utage, Regis Debruyne, Outils et Méthodes de la Systématique Intégrative, UMS 2700, MNHN, CNRS, Muséum national d'Histoire naturelle, Sorbonne Universités. 57 rue Cuvier, CP26, 75005 Paris, France
Maeva Thomas, Gaël P. J. Denys, Marion Mennesson, Unité Biologie des organismes et écosystèmes aquatiques (BOREA, UMR 7208), Sorbonne Universités, Muséum national d'Histoire naturelle, Université Pierre et Marie Curie, Université de Caen Basse-Normandie, CNRS, IRD, 57 rue Cuvier, CP26, 75005 Paris, France

1 Introduction

For over a decade now, the DNA-barcoding effort (*sensu* [1] for Actinopterygian fish species has focused on extensively sequencing the barcode region of the mitochondrial Cytochrome Oxidase 1 gene (COI) *via* a classical sanger-based approach [2,3]. This identification and preliminary analysis tool for characterizing fish diversity has been very successful thanks to the variability within the COI barcode sequence coupled with taxonomically efficient PCR primers (*e.g.* [4], and the availability of a dedicated, easy to use database to store and analyze the reference sequences: the Barcode of Life Database (BOLD; [5]. It has also brought methodological improvements through the encouraged use of shared, standardized markers and systematic vouchering of specimens attached to the sequences [1,6,7]. The appropriateness of COI as a marker for fish identification has been questioned, and studies have used other mitochondrial markers instead (for instance [8–10], leading to some dispersion of efforts in the search for a more variable and obtainable marker. While there are limitations to the current COI barcoding strategy within metazoans, we consider that over the last twelve years its benefits have been noticeable. Although not perfect, it has provided a clear and unified framework for biodiversity analysis, especially in Actinopterygians where COI has proven appropriate [2,3], although COI lacks variability in some groups [11]. However, the rise of the second-generation sequencing techniques with their ability to generate large amounts of genomic data has increasingly challenged this framework [12], to the point

where some authors consider it nothing but outdated and irrelevant [13,14].

The current second-generation sequencing platforms were originally designed to deep sequence individual whole-genome DNA libraries. Thus, they are not directly adapted to the multiplexed targeted sequencing of short genomic fragments like COI barcodes. As some other authors point out [15], this contrasts with the simplicity and immediate efficacy of COI barcode sequencing. One of the greatest difficulties on those platforms pertains to attaching sequence reads/consensuses to actual specimen vouchers while sequencing a cost-effective number of samples, without multiplying sequence tags and their added cost and labwork [16–18]. Methods implementing known indexed libraries in order to demultiplex sequences from individual organisms have indeed proven work-intensive (when PCR-based; [19,20] and/or expensive (when ligation-based) and, despite being productive [21], lack the simplicity and appeal of the Sanger-based sequencing of individual PCRs.

To overcome this problem for the Next Generation Sequencing methods (NGS; [22], several pragmatic approaches intend to improve multiplexing by considering an implicit tagging of the samples based on their sequence divergence, as originally proposed by Pollock et al. [23] in the pre-NGS era. Both PCR-based [16] and PCR-free [17,24] approaches have thus been described. They all make use of the throughput of current NGS platforms to sequence not only the COI barcode sequence but the entire mitogenome, with high multiplexing of individuals. In these new NGS frameworks, currently available COI barcode reference data have proved very useful to guide the *a priori* taxon-multiplexing strategy [17] as well as for *a posteriori* assignation of assembled mitogenomes to the samples. The COI and other mitochondrial sequences are used as 'baits' [16,25], and thus the necessary traceability to specimen vouchers is maintained.

In recent years, the number of mitogenome sequence releases in GenBank has considerably risen thanks to increased throughput in new sequencing platforms [7,17,26]. Furthermore, the number of available mitogenomes, and their usefulness for phylogeny at different scales has been particularly well explored in Actinopterygians since the description of the carp mitogenome in 1994 [27]. This is largely due the effort of one team over the last 15 years which is responsible for the publication of more than 1340 fish mitogenomes and the corresponding publications (from [28] to [29] This includes the creation of Mitofish and MitoAnnotator, a dedicated database and its annotation tool for fish mitogenomes [30]. Consequently, Actinopterygians are currently the second best represented vertebrate group after mammals in mitogenome sequence number in NCBI Nucleotide, and the first in number of represented species (figure 1). Since the COI barcode belongs to the mitochondrial genome, it exhibits the same evolutionary characteristics. Thus both have the same advantages: unambiguous orthology and uniparental inheritance facilitating their use within bifurcating phylogenetic approaches [29,31,32], as well as high levels of phylogenetic content with somewhat uniform evolutionary rates [33]. But they also have the same limitations: potential species paraphyly, lineage-specific rates of evolution and base compositions [34].

We promote here an affordable and easy approach to develop actinopterygian mitogenomics still further, by sequencing complete or large fragments of mitogenomes using next generation sequencing technologies and two-level multiplexing. In the line of Timmermans et al. [16], Kane et al. [35], Dettai et al. [17] and Strohm et al. [7], we discuss how this could represent an extension of the original barcode marker for more precise identification of vouchers and metagenomics, while remaining fully compatible with the previously obtained mitochondrial datasets. We describe briefly the mitogenomes derived from the presented workflow for two specimens from species of daces endemic to the South-West of France, *Leuciscus burdigalensis* Valenciennes 1844 and *Leuciscus oxyrrhis* (La Blanchère 1873), as an example of improvements for barcoding, phylogeographic and taxonomic studies. We also evaluate the presence of sequences of mitogenomic origin in 11 complete fish nuclear genomes available in the ENSEMBL database to assess the risk of sequencing NUMTs instead of the desired mitogenomes.

2 Methods

2.1 Dedicated Teleost long-range PCR amplification of mitogenomes

We set out to amplify the complete mitochondrial genome for organisms of various groups of local research interest largely spread within teleosts: Cyprinidae, Esocidae, Cottidae, Nototheniidae, Gobiidae, Eleotridae, Lampridae, and Zoarcidae. We downloaded complete mitochondrial sequences from NCBI Nucleotide for species from these groups and aligned them using ClustalW (with default settings; [36]). Conserved sequence regions (identified by eye) were subsequently imported into Oligo 4.1 Primer Analysis Software (National BioScience Inc., Plymouth).

Efficient long-range PCR amplifications require a particular attention to primer design, with a few general

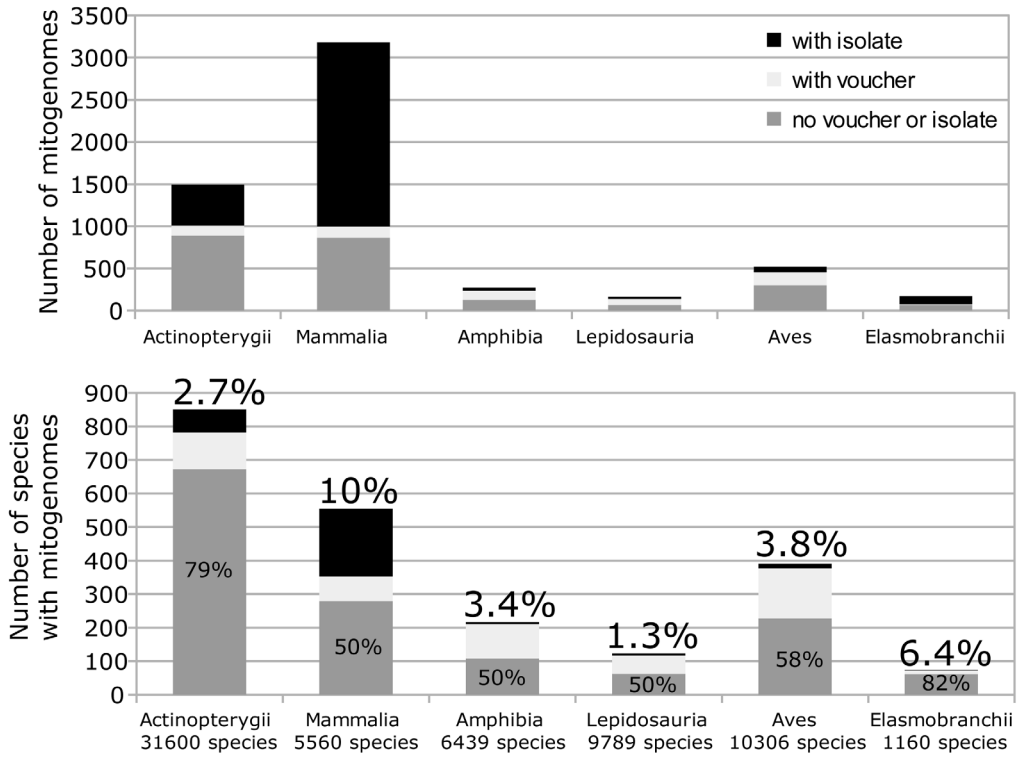

Figure 1: Number of mitogenome sequences and species represented in NCBI GenBank (last checked 02/03/2015). We searched GenBank (excluding Refseq sequences) using 'groupe name complete mitochondrion' and filtering by sequence length between 12,000 and 25,000. For Mammalia *Homo sapiens* was excluded. To evaluate how many listed a reference to a voucher, voucher was added as a search word, and to evaluate the number listing only an isolate reference but no voucher, we used (voucher OR isolate), with the number evaluated for voucher substracted. Total species numbers are given under the Y axis (source: http://www.catalogueoflife.org/col/browse/tree/id/21835362). Percentage of total species represented indicated above the bars, percentage of species with sequence and no voucher or isolate indicated on the bar.

recommendations often being the key to success. We thus took care to select primers with little or no self- or heterodimer hybridization, and closely matching annealing temperatures with primer melting temperatures selected between 65 and 70°C. We also used existing 12S rDNA primers [37]. Primers are provided in table 1, and also available in the BOLD primer database.

In order to complete the mitogenome amplification, we generally amplified 3 overlapping fragments (Figure 2). We used an in-house modified protocol of the HotStart LongAmp® Taq DNA Polymerase (New England Biolabs, Ipswich MA): PCR reactions were performed in 18 µl volume including 5X LongAmp Taq Reaction Buffer, 0.4 ng/ul Bovine Serum Albumin, 3.5% DMSO, 300 nM of each primer, 300 µM of dNTPs, and 1 unit of LongAmp Taq polymerase. After an initial denaturation of 30 s at 94°C, the DNA was amplified through 45 cycles of 20 s at 94°C, 30 s at 62.5°C, and 15 min at 65°C, with a terminal elongation for 15 min at 65°C. Whenever the amplification was not straightforward with pre-selected generic primers, specific primers were designed to fit the specific mutations of the problematic group. Additional primers for Cottidae and Gasterosteidae are listed in table 1, with the modifications highlighted. The PCRs were visualized on ethidium bromide stained agarose gel, and we estimated the amounts for pooling from the intensity of the bands [16], taking into account PCR size and using one of our own PCRs as intensity standard across gels. Within each taxon, a variable number (2 to 20) of PCRs of nuclear markers of phylogenetic interest (Rag1, IRBP, Pkd1, S7...) were also added in the individual pools to maximize sequencing efficiency.

2.2 Two-level multiplexing of individual mitogenomes in libraries

The first level of taxonomic multiplexing in our approach involves the pooling of long-range PCR between species for distantly-related fishes (Figure 3). These pools contained on average 8 mitogenomes for which molarity was roughly equalized based on agarose gel quantitation. We update our library preparation protocol slightly to

Table 1: Primers for the long PCRs, with primers adapted to Gasterosteidae and Cottidae included. For these, the modified bases are in bold black. Annealing temperature is 62.5°C for all.

Primer name	5'-3' sequence	Publication
12S-L1091R	AAACTGGGATTAGATACCCCACTAT	Kocher et al.(1989)
12S-H1478	TGACTGCAGAGGGTGACGGGCGGTGTGT	
12S-H1481-Acanth	**TTGGG**GAGAGTGACGGGCGRTGTGT**GCG**	
MtH20	GCCCATTTCAACATCTTCAGTGTTA	This study
MtL8	TTAACTAAAGCATAACACTGAAGATGT	
MtL5231	TAGATGGGAAGGCTTCGATCCTACA	
MtL5231-Cottus	TAGATGGG**T**AGGC**C**TCGATCCTACA	
MtL5231-Gast	TAG**C**TAGG**C**AGGC**C**TCGATCCT**G**CA	
MtL5247	GATCCTACAAACTCTTAGTTAACAGCTA	
MtH7045	CTTGAAACCAGTTTATGGGGGTTC	
MtH7061	GGGTTATGTGGCTGGCTTGAAAC	
MtH7061-Acanth	**C**GGTTATGTGG**T**TGGCTTGAAAC	
MtL9563	TCTACGTCTCCATCTACTGATGAGG	
MtH9621	CAAGACCGGGTGATTGGAAGTCA	
MtH11944	CATAGCTTTTACTTGGATTTGCACCA	
MtH11944-Gast	CATTTYTACTCCTACTTGGATTTGCACCA	
MtL11910	CAGCTCATCCATTGGTCTTAGGAAC	
MtL11910-Gast	CAGCTCATCCATTGG**A**CTTAGGA**T**C	
MtH15027	GGCTTACAAGACCGGCGCTC	
MtL15023	GCCAGAGCGCCGGTCTTGTAA	

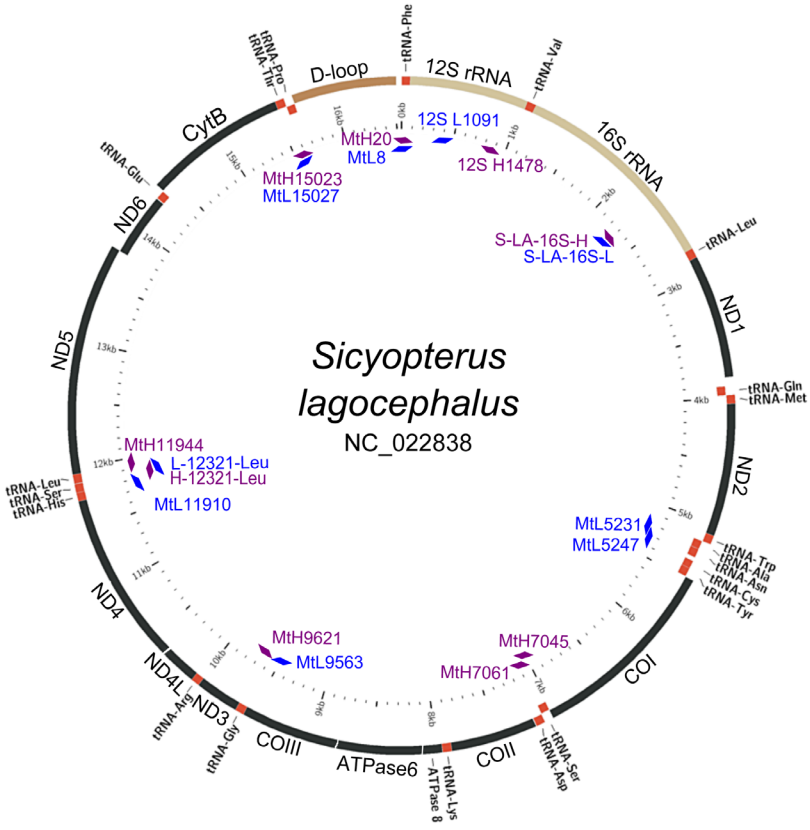

Figure 2: Position of the primers on the mitogenome of *Sicyopterus lagocephalus* (Chiang et al. 2013). Modified from a figure generated by Mitofish using Circos.

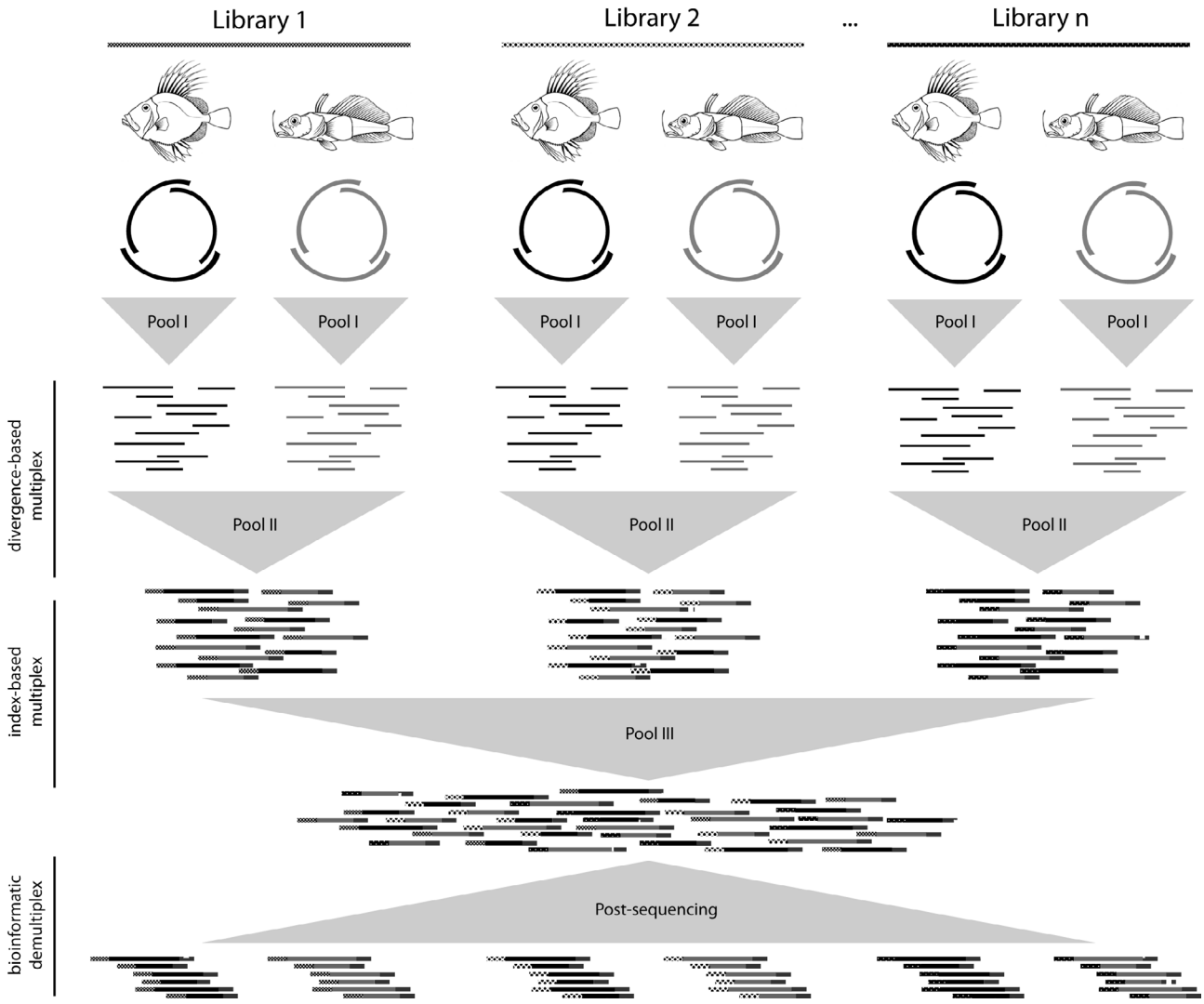

Figure 3: Synoptic figure of the global workflow. Tagged libraries are represented by a specific motif corresponding to their barcoded adapter. Divergent fish groups display DNA in light and dark shades of grey. The three pool stages and two multiplexing stages are shown.

improve quality, facilitate preparation and minimize the cost of individual libraries. We describe here the latest, most successful iteration of the protocol. Pooled PCRs were equalized to a final volume of 100 µl and sheared on a bioruptor standard (Diagenode, Liège, Belgium) for 20 minutes (with 30 seconds on/off intervals, on the HI setting, under constant cooling at 4°C). Initial attempts to fragment DNA of lower pool volumes (30-40 µl) yielded poorly reproducible results. Forty microliters of the fragmented DNA were processed in the next steps.

The second level of multiplexing relies on the indexing of DNA libraries prior to sequencing. We used the NEBNext fast DNA library prep protocol (ver 4.1) scaling down the reaction volumes to 50% to reduce the cost, in parallel with the Ion Xpress barcodes (Life Technologies, Carlsbad California). We performed the final library size selection with a double SPRI protocol using the NucleoMag paramagnetic separation beads (Macherey-Nagel, Düren, Germany) first with 0.55 bead/DNA ratio and then 0.15, to select fragments compatible with the 400 bp sequencing kit of the PGM platform. After separate quantitation of indexed libraries using the Ion Taqman quantitation kit (with reagents volumes scaled down to 1:2; Life technologies), an equimolar pool of 20 tagged libraries was amplified and sequenced on a 316v2 sequencing chip for the Ion Torrent PGM platform (with an expected throughput of at least 1M reads), so that, on average, a total of 160 mitogenomes were targeted on each chip (Figure 3).

2.3 Demultiplexing and assembly of individual mitogenomes

In order to automate most of the sequence reads post-processing, we produced a set of linux shell scripts (available on request; Figure 4). The main script (MitoPip_v1.sh) was called with the folder containing the raw sequences (either in .fastq or .bam output format from the Torrent server) and the first and last library to analyze (allowing partial analysis of a partially filled chip) as parameters. All the parameters required for steps invoked in the script were read from the "main_config.txt" file. We used relaxed criteria (Kmer in 9-55, coverage cut-off of 5) in Oases 0.2.08 [38], (Kmer in 47-127 with a step of 10, a max read length of 1000, average insert size of 200, reverse_seq, asm_flags, rank and map_en set to 0, 3, 1 and 128, respectively) in SOAPdenovo2 r240 [39], and (Ion Torrent set as the sequencing technology with minimum read length = 100, no quality check, minimum relative score = 95) in MIRA 4.0.2 [40] for the first round

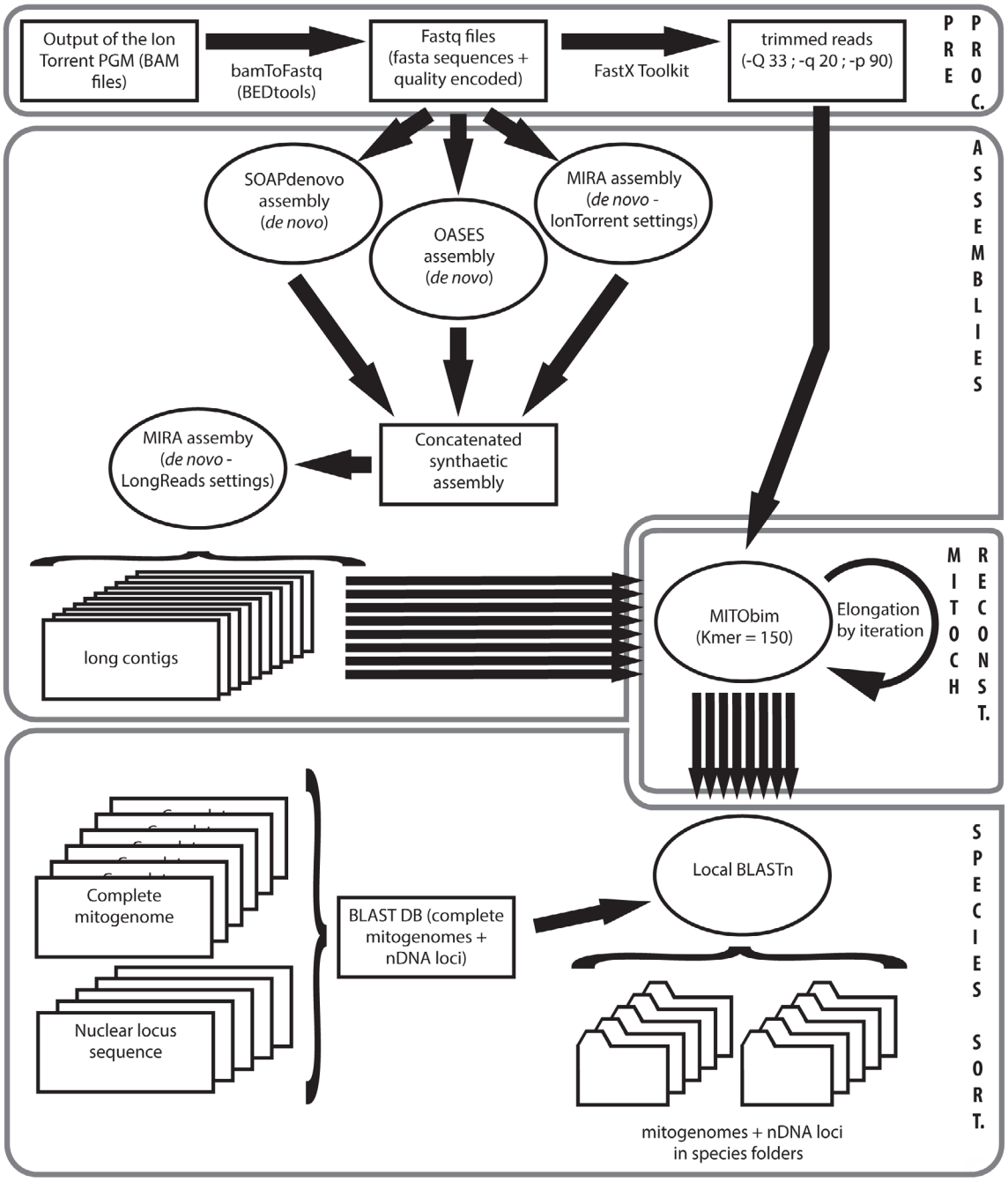

Figure 4: Schematic illustration of the steps of the post-sequencing pipeline.

Table 2: Sequencing cost breakdown. Prices are given in € as of early 2015 reagents costs.

long-range PCR (NEB)	per chip	per library	per mitogenome
PCR	108	5,4	0,6
DNA library prep	180	9	1
DNA Library quantitation	100	5	0,6
Template preparation and 316v2 sequencing (400 bp)	750	37,5	4,5
TOTAL (rounded cost)	1138	56,9	7,1

of assembly. We kept only the resulting contigs comprised between 200 and 20,000 bp for the subsequent analyses. Then, we produced long contigs (LC) performing a second round of MIRA assembly (Sanger sequencing technology, -NW:cmrnl = no, -HS:nrr = 5, -SK:mmhr = 10, -AS:epoq = no, a minimum relative score of 90 and a minimum_ read_length of 100). We implemented a slightly modified version of MITObim v1.7 [25] for Ion Torrent data (k_bait = 150, readlength = 250, insertsize = 400, available on request) to elongate the LC as multiple starting seeds (using the –quick option in MITObim), using reads cleaned with FastX toolkit 0.0.13 (>80% of bp with quality >20) as input data. We included other options for MITObim analysis, allowing the use of up to three short sequences as seeds (such as COI, 16S or cytB with the –quick option) or a related complete available mitogenome as a reference for mapping. As a final check for taxonomic assignation, we BLASTed the elongated LC against a customized local BLAST database including all available mitogenomes for the studied groups using BLASTn 2.2.29+ [41] ; www.BLAST.ncbi.nlm.nih.gov/BLAST.cgi), keeping only the best hit (e-value < 10-3 and identity > 85% threshold).

Finally, the species-specific long contigs and reconstructed mitogenomes were sorted in folders according to species for users analyses. The mitogenomes and contigs are then imported in Geneious v7.0 [42], checked against the sequence libraries they were extracted from for coverage and quality, and the smaller contigs are user-assembled.

2.4 Coverage analysis

In order to analyse the effect of the sequence coverage over the quality of the assembled mitogenomes, we compared actual datasets obtained for 9 *Esox* specimens sequenced on differently tagged libraries, with several read subsets of our best covered *Esox* mitogenome: specimen BRO-506, from the same sequencing run. From the sequence reads obtained for this specimen, we used 14,022 reads to generate the reference sequence: we assembled a single mitogenomic contig that we subsequently trimmed down to 14,034 bp by excluding the highly variable control region holding sequence repeats.

Read subsamples were generated using the Downsample SAM/BAM tool of NGS Picard Tools (Broad Institute) as implemented within Galaxy (ver 15.02; [43] with 5 samples each (with random seed) for probabilities ranging from 0.01 up to 0.50 of the original 14,022 reads. Reads subsamples were subsequently mapped onto the *Esox* reference sequence, to derive several metrics: breadth and depth of coverage, and error rates statistics in the assembled consensus sequence. In practice, the error rate was calculated as the percentage of changes in the consensus sequence of the covered fraction of the mitogenome for each read subsample when compared to the reference sequence built from all the reads.

2.5 Case study of two *Leuciscus* species

To document the actual efficiency of our global workflow, we selected two adult specimens from French endemic *Leuciscus* species as a practical example. The vouchers from the Languedoc Roussillon region corresponding to vouchers *Leuciscus oxyrrhis* MNHN-IC-2010-1839 from Tarnon stream at Florac (Lozère, Garonne drainage) and *Leuciscus burdigalensis* MNHN-IC-2010-1830 from the Agly river at Latour-de-France (Pyrénées Orientales) were identified morphologically with the criteria given by Kottelat and Freyhof [44]. Fish finclips (for other species, fin clips/tissue from newly collected to six years old were used instead) preserved in 95% ethanol were used to extract DNA. Several types of extraction were tested on the other samples, including CTAB extraction [45] and automated extraction using an Eppendorf epMotion 5075 with NucleoSpinR 96 Tissues kits (Macherey–Nagel) following the instructions of the manufacturer.

The sequences were checked with the assemblies using Geneious v7.0 and compared with closely-related published mitogenomes and the COI sequences corresponding to the specimens [46]. We controlled the protein translations of the coding genes and performed automatic annotation using MitoAnnotator and comparison with published *Leuciscus* mitogenomes. The complete mitogenomes are deposited in GenBank

(accession numbers KT223567 and KT223568) and in the BOLD (Project MtBA, sample numbers FFFtag4075 and FFFtag4059).

2.6 Search for long NUMTs in fish mitogenomes

The osteichthyan nuclear genome sequences available in the ENSEMBL database were blastN-queried through the BLAT/BLAST portal, using the mitogenome sequence for the corresponding species retrieved from NCBI nucleotide, and search sensitivity adjusted for "Distant homologies". The only fish species not queried was *Poecilia formosa*, for which no mitogenome was available for the same species. As we are trying to evaluate the presence of long NUMTs, only sequences longer than 1000 bp were retrieved and aligned to the query sequences using MUSCLE [47]. Genomic location was recorded, and the sequences were checked through blast search in NCBI (nucleotide nr database).

3 Results

3.1 Primers and Long Range Amplification success

The three main PCR amplifications generated fragments around 6.7 kilobases (Figure 2). The three fragments all included coding genes with good reference datasets: Mt1 (best primer pair: 12SL1091 with MtH7061) and Mt2 (best primer pair: MtL5231 with MtH11944) overlapped over the whole COI sequence, while Mt3 (MtL11910 with 12SH1478) included the complete Cytochrome b.

DNA extracted by two methods and extractions up to six years old yielded long PCRs, with better success rates for CTAB extracts than for robot and kit extractions. Amplification was successful across Teleostei, but some taxa were problematic for some primer pairs. However, the primer adaptation approach proved both fast and effective. The large reference data available for Actinopterygii (Figure 1) provides easy primer verification by comparison with closely-related taxa and modification when and where necessary before ordering or PCR-testing. The new primers are presented in table 1, with primers adapted to Gasterosteidae and Cottidae included as examples of the primer adaptation. The Mt1 and the Mt3 fragments overlap by more than 400 bp in the 12S, Mt1 and Mt2 overlap over the whole COI coding sequence, and Mt2-Mt3 still overlap over 10 bp once the primers are removed. Species where gene order was modified need adjustment of the primers used together so they bracket a reasonably-sized sequence fragment. In some groups, the control region was an obstacle, including for PCR. For instance, pike mitogenomes control regions include a low complexity, 100% AT 157 bp fragment that prevented correct PCR and sequencing. This region is also missing in two out of the three pike mitogenomes available in GenBank. Primers surrounding the control region helped amplify the rest of the mitogenome.

3.2 Sequence production and post-processing using our pipeline

The latest iteration of sequencing yielded 1,351,350 usable reads on a 316v2 chip for a total of 255M bp. Once demultiplexed per index, tagged libraries provided between 22,118 and 156,300 reads each (7-fold difference). The median read length for these libraries was consistently around 200 bp.

As an example, considering library 5 of chip 3, the 156,159 raw reads (mean length = 226 bp, max length = 997 bp) resulted in 115,523 cleaned reads after trimming (average length = 235 bp). The contigs number and mean length for the Oases, SoapDenovo and MIRA assemblies were 12,984, 1,550 and 2,224 and the length 617 bp (range 100-25657 bp), 177 bp (range 58-3156 bp) and 395 bp (range 101-9107 bp) respectively. The composite assembly comprised 2092 long-contigs of 711 bp mean length (range 100-14910 bp), of which 1000 were subsequently analysed. MITObim increased this length to 11296 bp (164-36,704 bp), with 182 contigs longer than 14 kb. These elongated LC were assigned to 18 species (326 sequences unambiguously assigned, 10 species comprising at least 5 sequences).

3.3 Coverage analysis

We used simulated subsamples of *Esox* to investigate the change of three parameters: breadth of coverage, depth of coverage and consensus error rates, according to the number of reads per individual mitochondrial library (Figure 5). The simulations reveal that the breadth of coverage of the mitogenome sequence increases rapidly from about 75% to beyond 99% when the number of sequence reads increases between 100 and 1,000 (*i.e.* subsamples of 1 to 10% of the original reads; Figure 5A), and the depth of coverage increases from around 2X up to 20X (Figure 5B). This increase is accompanied by a drastic reduction in consensus error rates: from almost 3% down to 0.1-0.2% (Figure 5A). Further increase in the number of reads assembled only produced a marginal improvement for the contig length and quality. The reference coverage

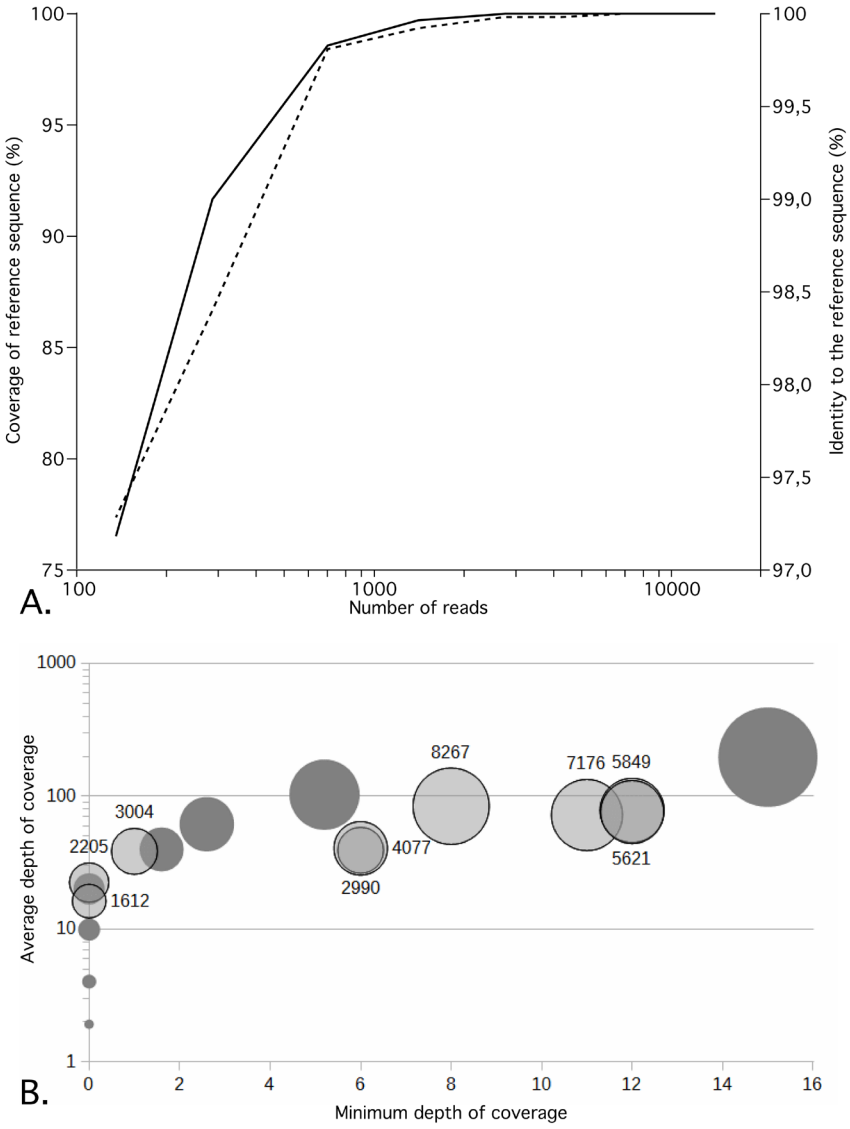

Figure 5: Results of the coverage analysis based on simulated re-samples of BRO-506 and actual *Esox* data. Evolution of the breadth of coverage (solid line; left scale) and of the sequence identity to the *Esox* reference sequence (dashed line; right scale) based on the amount of subsampled reads (in abscissa). B: Bubble plot for the depth of coverage (minimum x-axis, average in y-axis). the size of the bubble reflects the number of assembled reads: 1, 2, 5, 10, 20, 30, 50 and 100% of the 14,022 reads for the *Esox* reference displayed in dark grey; the total amount of assembled reads for the other *Esox* specimens is displayed next to their light grey bubble.

is always complete for about 3,000 reads (about 40X coverage), while 4,000 reads (60X) are necessary to remove the remnant errors from the consensus. Deeper sequencing of the mitogenome then only affects the coverage depth which reaches a minimum value of 16X and an average of 160.9X when using 100% of the *Esox* reads.

3.4 Two mitogenomes of *Leuciscus* spp.

The amplifications worked with the recommended three pairs of primers, for three fragments of 6.7 kb. After primers and quality trimming (Geneious by default trim parameters, 100% bp with Q > 40), 6951 (mean depth 59.1, range 10-480) and 4703 (mean depth 39.4, range 4-200) reads remained in *L. oxyrrhis* and *L. burdigalensis* assembly, respectively.

The consensus mitogenomic sequences from both dace samples exhibit the exact same length (16,006 base pairs) and positions of start and end for all coding and non-coding regions. Of the 37 differences observed between the two specimens, only three are reflected in the amino acid sequence (supplemental figure SF1). All but two of the 37 differences between the two mitogenome

sequences are shared with at least another individual from the same drainage (data not shown). The Folmer region of COI had been sequenced previously using Sanger sequencing for both specimens [46], and the sequences are identical with the sequences from our assemblies. This is also the case for all other mitogenomes sequenced for which we already had the barcode region (N>150). For this fragment, the two sequences are nested within a *Leuciscus burdigalensis* cluster on a BOLD COI distance tree. There are no SNPs in this region between the individuals of the two species. We performed a BLAST search using the cytochrome b sequences. The closest hits were from unvouchered samples identified as *Leuciscus leuciscus* (Linnaeus 1758) from the area [48]. However, there were no samples identified as *Leuciscus burdigalensis* or *oxyrrhis* in GenBank for *Cytochrome b*. The direction and order of genes is identical to those of most fish mitochondrial genomes, and the size of the coding genes is identical to the two Leuciscus sequences available currently in the RefSeq (NC_018825.1 and NC_024528.1; ST1). The sequence available for the most closely-related species according to Costedoat et al. [48], *L. idus* (Linnaeus 1758) has an unverified status in GenBank (KF913024.1, [49]). It displays a frame-shift at the end of the ND2 gene that is absent in our sequences and the RefSeq sequences of the two more distant species. The ML phylogenetic tree showed a high similarity between the two specimens (0.0021 substitutions per site). However, considering the two specimens together, the group they formed displayed a branch length (0.0329) similar to that exhibited by the other species (ranging from 0.0174 - L. idus - to 0.0541 - L. waleckii; supplemental figure SF2). The tree displays the relationships expected from previous publications on the group, and has the same topology as the COI fragment alone.

3.5 NUMTs in complete fish genomes

The results of the NUMT searches are presented in additional table ST1. While the "complete" fish genomes are in fact not fully sequenced nor assembled, we expect to find some sequences corresponding to NUMTs, with varying sizes and sequence divergence from the current mitogenome sequence for the species, depending on the size of the integrated sequence size and the age of the integration. These sequences are expected to have, at least in some cases, flanking sequences of non-mitochondrial origin, and to be part of larger scaffolds containing also other sequences. For *Astyanax mexicanus*, *Oreochromis niloticus*, *Lepisosteus oculatus*, and *Danio rerio*, all blast hits are shorter than 261 bp. For the other species, there are longer results (ST1), but always constituted exclusively of sequences that can be aligned with mitochondrial sequences. There were no flanking sequences in any of these, these sequences are not integrated within larger scaffold containing other sequences.

4 Discussion

4.1 General applicability and specific efficiency for teleosts

Some of the limitations in the classical COI-based barcoding hinge on the definition of working PCR primers for COI amplification. Sanger sequencing, because of the limited sequence length, imposes the definition of primers every 600-800 base pairs, whether the region is conserved or not across taxa. Combining long PCRs and next-generation sequencing supports selecting primers in interestingly conserved regions because the sequencing size restriction is relaxed.

While our primers work largely from Cyprinids to Carangids through pikes, the variability in the mitogenome is high, and some groups like Gasterosteidae or Cottidae could not be amplified using the original set. There are additional mitochondrial long PCR primers available from previous publications [50–52], for instance). However, our primer pairs amplify smaller fragments, and make it possible to use older or lower quality, more fragmented DNA, as well as cheaper brands of long fragment amplification Taq polymerase (i.e. three PCR reactions using our protocol cost less than one euro). Within minutes, based on the wealth of mitogenome sequences available in NCBI nucleotide for Actinopterygians, new primers adapted to specific groups can be designed by slightly modifying the described primers to better fit the mitogenome sequences from the group of interest at the same position. This basic adaptation does not require primer design expertise. It was sufficient for Gasterosteidae or Cottidae, and can probably be applied successfully to groups of local interest.

Most Actinopterygian mitogenomes present the same gene order [29], but some have undergone gene order changes, as first described by Miya and Nishida [53]. They can be quite frequent in some families like Myctophidae [54]. These gene order changes are prone to problems for PCR-based approaches, although changing primer combinations can be effective. Moreover, in some species, parts of the mitogenome, and especially the control region, contain patterns within the sequence like low complexity areas, composition biases or duplications.

These can impact PCR and sequencing success [55,56]. However these issues can be anticipated when sequences are available by checking the mitogenome description publications and the sequences themselves, like with the long low-complexity region in pike or the duplicated control region in Nototheniidae [57]. Obtaining the complete mitogenome might reveal very difficult for some species, and might not be worth the effort, as long mitogenomic fragments might already provide sufficient information. We designed primers flanking the control region for such cases.

NUMTs can be a serious issue in mitochondrial-focused molecular studies, although relatively few have been found in the sequenced genomes of teleost fishes ([58] but see [59,60]), and empirical records of NUMTs are thus limited (e.g. 1/242 spp. in [61]). The available complete genome data do not provide a complete panel of NUMTs that might be present in those genomes, for various reasons (i.e. assembly problems, incomplete coverage etc.). However, it is striking that in none of the complete genomes queried, we could recover any sequence showing some similarity to those mitogenomes while assembled to a larger scaffold of nuclear sequences, or even included between non mitochondrial flanking sequences. Despite there were hits similar to mitochondrial stretches longer than 1000 bp in six of the 10 nuclear genomes analyzed, none of them was ever assembled within larger scaffolds, suggesting fragments not integrated in the nuclear genomes or even artefacts. Blast searches for these sequences also shared characteristics that suggest they are not integrated in a nuclear genome: their similarity is higher to mitogenomes from other, divergent individuals of the same species, or even different species (*e.g.* a 6000 bp mitochondrial sequence of squirrel was actually recovered from the *Oreochromis niloticus* genome). Divergences are concentrated in the sequence ends, where most sequencing artefacts happen. Ambiguous base strings (N) separated sequence regions with different results in BLAST searches (Takifugu for instance). In agreement with previous publications, these results still suggest that longer NUMTs are rare in actinopterygian genome.

Therefore this approach, like others approaches based on long PCRs and mitogenome sequencing, considerably limits the risk of integrating unrecognized nuclear insertions of mitochondrial DNA (NUMTs) into the datasets [62]. Indeed, carefully checking our primers pairs against the "known" NUMTs in the three species studied in Antunes & Ramos [58] resulted in no possible amplification, due to the lack of similarity for both the forward and reverse primers into the nuclear inserted fragment (results not shown). Moreover, MITOBim was designed to properly handle NUMTs potentially found in genome skimming approach, and our approach including a MITOBim iterative mapping benefit from this robustness. Multi-bands PCR [63], sequence ambiguities in the assembled genome (especially where the PCR fragment overlaps), or SNP resulting in frameshift or stop codons are signs of potential problems, whether or not prior occurrence of NUMTs has been suspected in the group of study. Distance or phylogenetic trees are precious tools to detect possible problems. An unexpected position in the topology of a phylogenetic reconstruction, or different positions in trees derived from the three different PCRs [64] should also be carefully checked. If NUMTs are suspected, the primers pairs producing the longest fragment should be preferentially used [63,65,66].

4.2 Demultiplexing

Very few methods have been proposed for *in silico* demultiplexing of pooled mitogenome (but see [18,67]), although several approaches were designed for reconstructing mitogenome from directly sequenced genomic DNA (*i.e.* without PCR for specific sorting of the mtDNA nor mtDNA enrichment): *e.g.* MITObim [25], MIA [68], or IMAGE [69]. To our knowledge, the method we propose herein is the first to make use of both pooled and PCR-amplified mitogenome. Contrary to Timmermans et al. (2010), MIRA was not able to retrieve complete or nearly complete mitogenomes from the cleaned reads. This was expected: the reads from the Timmermans study were from 454 sequencing (374 bp mean length for the second run), whereas we dealt with shorter reads (on average 235 bp for the presented *Leuciscus* species) using Ion Torrent, and read length have an impact on mitogenome reconstruction [17]. In order to counterbalance this smaller read lengths, and to avoid assembler-specific bias, we used several assemblers designed for transcriptomics, as the copy number profile for each PCR fragment from each mitogenome resembles locus expression variation more closely than genomic DNA [24]. The process explores a large range of K-mer, and finally combines the output, similarly to the assembly of Sanger shotgun reads. In addition, we tuned MITObim to more stringency by increasing the K-mer length to 150, following to Dettai et al. (17). This resulted in very few chimeric mitogenomes (generally longer than 20 kb) that could be easily identified.

Timmermans et al. (16) and Dettai et al. (17) proposed somewhat symmetrical methods to use the COI barcodes in mitogenome multiplex sequencing. While Dettai et al. make use of the divergence of existing COI barcodes as a

guide prior to sample multiplexing, Timmermans et al. used COI barcodes generated from the actual samples as baits to identify *a posteriori de novo* assembled genomes. Clearly, these two strategies are not mutually exclusive and can be used sequentially to improve the overall efficiency of the multiplex/demultiplex operational stages.

This strategy yields a lot of sequences, and can be upgraded still by pooling nuclear PCRs with the mitochondrial ones at the first stage of library preparation. This does not impact the overall process, and locus demultiplexing remains obvious because of the high divergence between mitochondrial and nuclear phylogenetic markers. However, when pooling multiple taxa for a single nuclear marker, the lower variability of the nuclear markers must be taken into account, and species that could be safely combined for the highly divergent mitochondrial markers might be too similar for some nuclear markers. It is safer to perform a preliminary analysis using available sequences for the groups considered for multiplexing with sliding window analysis [17]. As PCR length is not limiting for this type of sequencing, it is particularly interesting to amplify larger fragments for nuclear markers too. For instance, using the outermost primers for IRBP and Pkd1 [70] allows the easy sequencing of respectively 2.5 and 2.3 kilobases for these markers, and the Rag1 fragment of 1.6 kb popular for teleost phylogenetics can be sequenced without internal primers.

Finally, it is interesting to highlight that other sequence manipulation and analysis programs, alongside or instead of Geneious, could be used for the final step of sequence checking, like CodonCode Aligner (CodonCode Corporation). Most of the biodiversity-related labs already have licences for such assemblers/aligners, so the post-sequencing migration from COI to complete mitochondrial is far easier than trying to adapt COI to NGS [16].

4.3 Coverage

Our current experimental setup relies on the multiplexing of 20 indexed libraries per 316v2 PGM chip. This chip has a guaranteed throughput of 1M reads, so provides an average 50,000 sequence reads per indexed library. Each indexed library being composed of on average 8-10 mitogenomes, a single mitogenome can be covered roughly with 5,000 reads, just with the guaranteed minimum. This approach is therefore a good fit for sequencers generating a small volume of longer sequences, like 454 GSjunior, Ion Torrent PGM, or single Illumina Mi-Seq lanes.

Our strategy for multiplexing "only" ~160 mitogenomes per sequencing chip can appear conservative. Based on our coverage analyses, it appears that an average of 3,000 reads per mitogenome is enough to complete correct mitogenome sequences, so that up to 400 mitogenomes could be sequenced on a single 316v2 chip in theory. However, our approach relies on a rough quantitation of PCR concentrations to save time and reagents, and therefore has to take into account potential larger discrepancies in the sequence coverage both within and between organisms. There are indeed three potential levels where equimolarity of samples is needed: (I) when pooling separate PCRs for single individuals, (II) when pooling different organisms within indexed libraries, and (III) when performing the final pooling of indexed libraries prior to emPCR. The evaluation using intensity of the PCR bands on gels is sufficient for the first two (Timmermans et al. (16), this study), in addition to being fast and cheap. For the third, we rely on a taqman quantitation of the 20 individual libraries prior to pooling which, in practice, only maintains the variation in reads per library within a 10-fold factor. Therefore, our conservative approach warrants that a maximum number of (if not all) pooled mitogenomic sequences can be reconstructed safely with a high level of confidence. The low cost of the present approach makes re-sequencing a PCR with a too low coverage a better alternative than a precise quantitation of the hundreds of combined PCRs.

Furthermore, coverage can be impacted by sequence composition, and especially extreme composition biases [55,56]. Such biases are sometimes present in mitochondrial genomes, the *Esox lucius* control region contains a 157 base pairs long 100% AT stretch (genbank accession: NC_004593.1). For such cases, sequencer choice is important, as some platforms are less sensitive to some types of biases and patterns [56].

4.4 Adapting barcode sequencing to NGS one step further

Adapting the targeted Sanger sequencing of COI barcodes to NGS platforms does not take advantage of the actual design and benefits of these platforms, and would be an expensive and work-intensive endeavour. Extension to mitogenomes and large genomic fragments is a logical step forward [16–18,23] and does considerably increase identification precision. As we show here, mitogenomes are easily obtainable for Actinopterygians. COI barcode sequence variability can also be too low between populations, or even species pairs [71,72], *Leuciscus* spp in this study). In these low divergence cases, having access to a larger number of informative sites limits stochastic effects and has beneficial effects on the robustness

of the results [73]. For our *Leuciscus* sequences, the Folmer region was identical for the two individuals. The complete mitochondrial genome presented 37 differences, of which 35 were informative when compared to three additional fishes from the same drainage. This higher variability adds support for both identifications and phylogenetic reconstruction, especially for species with very little intra and interspecific variability, like cod [71], antarctic Nototheniidae [72] or tuna [74]. However, optimal exploitation of this variability also requires more information about the origin of the samples (i.e. specimen metadata).

4.5 Data reliability

Sequence data reliability, and especially specimen identification, has been a longstanding problem, and GenBank's reliability in this matter is notoriously low. Strohm et al. (2015) recently evaluated the number of possible mis-identifications of mitogenomes. Thirteen percents of the Perciformes sequences checked clustered with another species in BOLD, and therefore represent possible misidentifications. The presence in the mitogenomes of many popular sequence markers makes them easily comparable to the existing reference datasets, and identification evaluation is fast and generally very efficient [7,31]. This should be standard verification for all mitogenome sequences, although the results can be difficult to interpret if there are taxonomic problems in the group. Our PCR fragments contain either COI (Mt1 and Mt2) or Cytochrome b (Mt3), and there is overlap between the fragments to check the consistency between the different PCRs from an individual.

The difficulty to obtain a correct identification of specimens is compounded in Actinopterygians because of cryptic and unidentified diversity, including in "well known" faunas like European or North American freshwater fishes [46,75–78]. The Barcode of Life standards have popularized the consistent [6] use of voucher specimens and specimen metadata attached to the sequences. While the conservation of whole specimens as vouchers requires long-term, adapted storage facilities, the benefits are on multiple levels [1,2,79]. Voucher specimens support re-evaluation of morphological identification, and can be used as a base for a new taxonomic study if the systematics of the group are in doubt, as is the case in a large number of fish species. Our *Leuciscus* sequences provide an example of this. BLAST search for Cytochrome b with our sequences retrieved sequences from the same drainages, but identified as *Leuciscus leuciscus* and non-vouchered. These sequences were published in 2006 [48], and the authors concluded that there were several lineages of dace. In 2007, Kottelat and Freyhoff [44] published a revision based on morphology only of the freshwater fish of Europe. Their revision reestablished *L. burdigalensis, L. oxyrrhis* and *L. bearnensis* (Blanchard 1866), and we used their criteria to identify the specimens included here. A revision in an integrative framework of French daces is still needed using an integrative taxonomy approach [80], and it might result in additional changes. Thus even acting on the best knowledge at a given time point, attribution to a species can be questioned later on with new, more precise data. This is not an isolated case: in the study of Strohm et al. (2015), 42% of the sequences matched to more than one species, and therefore require further evaluation. Yet corrections in GenBank are difficult at best [81], and do not encourage the rectification of identifications to keep up with current taxonomic knowledge. The data access and visualization processes in BOLD are much more conducive to a dynamic correction.

Currently, only a small proportion of the sequences and the mitogenomes available in GenBank are linked to a voucher specimen (Figure 1). Worse, for 79% of mitogenome sequences neither a voucher nor an isolate number are given, indications of geographical origin, and even more complete metadata, are also rare [7]. This limits their use for other applications like phylogeography and multi-marker analyses, as it is impossible to know which individual has been sequenced for a marker, and the various alleles of nominal species can be non-monophyletic for some loci [82]. Some mitogenome sequences have even been assembled from several individuals, sometimes without indications that this is the case [7]. Aggregating sequences of unknown source and potentially different individuals (some of which might be misidentified) to "represent" the species for different markers is problematic theoretically and even practically for smaller evolutionary scales, and when an insufficient number of individuals per species is studied [82], whether for identification or systematics. The mitogenome presents a very high number of SNPs between individuals [21,29,71], and all the markers are physically linked, so composite sequences should be avoided even carefully. The mitogenome is a single unit with one history, that can therefore be considered a single marker, and as has been shown repeatedly, one marker is not enough for species delineation [32] nor reconstructing species history [82]. To move beyond this single marker requires being able to reliably put data together, and that means knowing whether the sequences come from the same individuals or not. Transposing the Barcoding of Life standards [6] for link to voucher, locality, and sample number would be a considerable improvement, especially

if the sequences are to be used for smaller scale studies where incomplete lineage sorting and coalescence need to be taken into account. It is also crucial that the voucher or sample identification is clearly listed both in the paper and in the molecular databases.

5 Conclusion

The search for relevant, multi-locus, genome-wide based barcodes are certainly key in the near future. Yet the continuous documentation of the ever-growing COI barcode database is highly relevant for identification and biological inventories, and are today the base for further in-depth molecular analysis of taxa. The extension of barcode to mitogenomes is now technically easy, compatible with the previous barcoding efforts, and can build on more than a thousand mitogenome sequences across actinopterygians. This dataset can now be densified at smaller scale [29] with an unprecedented precision level for a single and largely comparable marker.

These easily obtained complete or almost complete mitogenome sequences open new venues for applications in phylogenetics, identification, and metagenomics. It is obvious that this approach and its variations can be transposed to any other biological entity of interest for mitogenome sequencing, as shown by previous publications [16,18] and tests using the same protocols on insects and birds performed at the MNHN. When analyzed using appropriate concepts and methods, the mitogenome data can also be used for larger scale phylogenetics, but more interestingly for small scale biogeography and integrative taxonomy studies as well [21,29,71]. It provides an interesting complement to nuclear sequence data because of its contrasting maternal inheritance, higher mutation rate, and lack of recombination leading to a lower effective population size (Ne). However it still represents only the maternal lineage, so groups where introgression is frequent like temperate freshwater fishes would benefit even more from standardized nuclear markers for identification, and these can easily be sequenced using the present approach too. For instance, *Leuciscus idus* shares mitochondrial sequences with dace *Leuciscus leuciscus* [48], and some pickerel species also share mitochondrial sequences (*Esox niger* Lesueur 1818 and *E. americanus americanus* Gmelin 1789; [75]. No doubt we will discover other such groups as the range of fish studied in depth grows. However, all these uses rely on avoiding creating chimeric sequences from several individuals, as well as accurate metadata including sample reference, locality of capture and date, and preferably vouchering of the specimens [7].

With the growing popularity of metagenomics, the high number of copies per cell and high resolutive power of a mitogenome reference dataset are very promising. Instead of relying on happenstance to recover COI sequences during PCR-free metagenome sequencing, any sequence from the mitogenome represented in the tested sample can do [16–18]. The mitogenome reference datasets are therefore a promising tool at all levels of biodiversity: within species, between species, and to study ecosystems. Simple techniques for acquisition of multiple sequences for alow cost helps to move from the deluge of the simple description of mitogenomes publications [26]

Currently, the Barcode of Life database does not accommodate full mitogenomes. It does accept a maximum of 13 mitochondrial markers deposited separately, though. From our experience, the sample description format in BOLD promotes completion and correction of sample metadata through the available, pre-listed categories better than the less user-friendly entry and correction process in NCBI nucleotide. A reference mitogenomic dataset would perfectly complement existing resources in BOLD, and probably ameliorate accompanying metadata deposition.

Acknowledgments: This work was supported by the Action Thématique du Muséum (ATM) “Génomique & Collections” and by the Fondation TOTAL (project ‘‘Classification des poissons marins: les téleostéens acanthomorphes’’). We thank the Service de Systématique Moléculaire (part of the Plateforme Analytique du Muséum, Muséum national d’Histoire naturelle) for granting access to its molecular platform. We are grateful to the lab engineers (Céline Bonillo, Delphine Gey and Josie Lambourdière) in the continuous support and help to implement this work in the new routines of the SSM. We thank also Henri Persat (Lyon I University), Frédéric Melki and Benjamin Adam (Biotope), the French National Agency for Water and Aquatic Environments (Onema) as well as the Associations agréées de pêche et de protection des milieux aquatiques (AAPPMA) of Florac and Saint-Paul-de-Fenouillet for the fish sampling. We are grateful to JF Dejouannet for illustrations used in figure 3. We thank Gael Lancelot for language corrections of the manuscript.

Contributions of the authors: AD has designed the theoretical framework of this project. AD, MT, MM and GD have produced the original PCR data and produced the DNA libraries. JU and RD have performed template preparation and sequencing. DH has designed, programmed and tested the post-processing bioinformatics and scripts.

DH, RD, AD, GD analyzed the data presented. DH, RD and AD produced the figures and tables. All the authors participated to the writing and editing of the manuscript.

Conflict of interest: Authors declare nothing to disclose.

References

[1] Hebert P.D.N., Cywinska A., Ball S.L., deWaard J.R., Biological identifications through DNA barcodes, Proc. Biol. Sci., 2003, 270, 313–21

[2] Ward R.D., Zemlak T.S., Innes B.H., Last P.R., Hebert P.D.N., DNA barcoding Australia's fish species, Philos. Trans. R. Soc. Lond. B Biol. Sci., 2005, 360, 1847–57

[3] Ward R.D., Hanner R., Hebert P.D.N., The campaign to DNA barcode all fishes, FISH-BOL, J. Fish Biol., 2009, 74, 329–56

[4] Becker S., Hanner R., Steinke D., Five years of FISH-BOL: brief status report, Mitochondrial DNA, 2011, 22 Suppl 1, 3–9

[5] Ratnasingham S., Hebert P.D.N., BOLD: The Barcode of Life Data System (http://www.barcodinglife.org), Mol. Ecol. Notes., 2007, 7, 355–64

[6] Hanner R., Data standards for BARCODE records in INSDC (BRIs). 2009. http://www.barcoding.si.edu/PDF/Guidelines%20for%20non-CO1%20selection%20FINAL.pdf

[7] Strohm J.H.T., Gwiazdowski R.A., Hanner R., Mitogenome metadata: current trends and proposed standards, Mitochondrial DNA.\, 2015, 1–7

[8] Garcia-Vasquez E., Perez J., Martinez J.L., Pardinas A.F., Lopez B., Karaiskou N., et al., High level of mislabeling in spanish and greek hake markets suggests the fraudulent introduction of African species, J. Agric. Food Chem., 2011, 59, 475–80

[9] Von der Heyden S., Barendse J., Seebregts A.J., Matthee C.A., Misleading the masses: detection of mislabeled and substituted frozen fish products in South Africa, ICES J. Mar. Sci., 2010, 176–85

[10] Naylor G.J.P., Caira J.N., Jensen K., Rosana K.A.M., White W.T., Last P.R., A DNA sequence-based approach to the identification of shark and ray species and its implications for global elasmobranch diversity and parasitology, Bull. Am. Mus. Nat. Hist., 2012, 2012

[11] Dettai A., Adamowizc S.J., Allcock L., Arango C.P., Barnes D.K.A., Barratt I., et al., DNA barcoding and molecular systematics of the benthic and demersal organisms of the CEAMARC survey, Polar Sci., 2011, 5, 298–312

[12] Pompanon F., Samadi S., Next generation sequencing for characterizing biodiversity: promises and challenges, Genetica., 2015, 143, 133–8

[13] Taylor H.R., Harris W.E., An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding, Mol. Ecol. Resour., 2012, 12, 377–88

[14] Dowton M., Meiklejohn K., Cameron S.L., Wallman J., A preliminary framework for DNA barcoding, incorporating the multispecies coalescent, Syst. Biol., 2014, 63, 639–44

[15] Collins R.A., Cruickshank R.H., Known Knowns, Known Unknowns, Unknown Unknowns and Unknown Knowns in DNA Barcoding: A Comment on Dowton et al., Syst. Biol., 2014, 63, 1005–9

[16] Timmermans M.J.T.N., Dodsworth S., Culverwell C.L., Bocak L., Ahrens D., Littlewood D.T.J., et al., Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics, Nucleic Acids Res., 2010, 38, e197

[17] Dettai A., Gallut C., Brouillet S., Pothier J., Lecointre G., Debruyne R., Conveniently Pre-Tagged and Pre-Packaged: Extended Molecular Identification and Metagenomics Using Complete Metazoan Mitochondrial Genomes, PLoS One, 2012, 7, e51263

[18] Tang M., Tan M., Meng G., Yang S., Su X., Liu S., et al., Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics, Nucleic Acids Res., 2014, gku917

[19] Meyer M., Stenzel U., Hofreiter M., Parallel tagged sequencing on the 454 platform, Nat. Protoc., 2008, 3, 267–78

[20] Bybee S.M., Bracken-Grissom H., Haynes B.D., Hermansen R.A., Byers R.L., Clement M.J., et al., Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics, Genome Biol. Evol., 2011, 3, 1312–23

[21] Feutry P., Kyne P.M., Pillans R.D., Chen X., Naylor G.J., Grewe P.M., Mitogenomics of the Speartooth Shark challenges ten years of control region sequencing, BMC Evol. Biol., 2014, 14, 232

[22] Shendure J., Ji H., Next-generation DNA sequencing, Nat. Biotechnol., 2008, 26, 1135–45

[23] Pollock D.D., Eisen J.A., Doggett N.A., Cummings M.P., A case for evolutionary genomics and the comprehensive examination of sequence biodiversity, Mol. Biol. Evol., 2000, 17, 1776–88

[24] Rubinstein N.D., Feldstein T., Shenkar N., Botero-Castro F., Griggio F., Mastrototaro F., et al., Deep Sequencing of Mixed Total DNA without Barcodes Allows Efficient Assembly of Highly Plastic Ascidian Mitochondrial Genomes, Genome Biol. Evol., 2013, 5, 1185–99

[25] Hahn C., Bachmann L., Chevreux B., Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach, Nucleic Acids Res., 2013, gkt371

[26] Smith D.R., The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs?, Brief. Funct. Genomics, 2015, elv027

[27] Chang Y.S., Huang F.L., Lo T.B., The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome, J. Mol. Evol., 1994, 38, 138–55

[28] Miya M., Kawaguchi A., Nishida M., Mitogenomic exploration of higher teleostean phylogenies: a case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences, Mol. Biol. Evol., 2001, 18, 1993–2009

[29] Miya M., Nishida M., The mitogenomic contributions to molecular phylogenetics and evolution of fishes: a 15-year retrospect, Ichthyol Res., 2015, 62, 29–71

[30] Iwasaki W., Fukunaga T., Isagozawa R., Yamada K., Maeda Y., Satoh T.P., et al., MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline, Mol. Biol. Evol., 2013, 30, 2531–40

[31] Botero-Castro F., Delsuc F., Douzery E.J.P., Thrice better than once: quality control guidelines to validate new mitogenomes, Mitochondrial DNA, 2014

[32] Dupuis J.R., Roe A.D., Sperling F.H., Multi-locus species delimitation in closely related animals and fungi: one marker is not enough, Mol. Ecol., 2012, 21, 4422–36
[33] Papadopoulou A., Anastasiou I., Vogler A.P., Revisiting the Insect Mitochondrial Molecular Clock: The Mid-Aegean Trench Calibration, Mol. Biol. Evol., 2010, 27, 1659–72
[34] Li H., Shao R., Song N., Song F., Jiang P., Li Z., et al., Higher-level phylogeny of paraneopteran insects inferred from mitochondrial genome sequences, Sci. Rep., 2015, 5
[35] Kane N., Sveinsson S., Dempewolf H., Yang JY., Zhang D., Engels J.M.M., et al., Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA, Am. J. Bot., 2012, 99, 320–9
[36] Thompson J.D., Higgins D.G., Gibson T.J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res., 1994, 22, 4673–80
[37] Kocher T.D., Thomas W.K., Meyer A., Edwards S.V., Pääbo S., Villablanca F.X., et al., Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers, Proc. Natl. Acad. Sci. USA, 1989, 86, 6196–200
[38] Schulz M.H., Zerbino D.R., Vingron M., Birney E., Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, Bioinforma. Oxf. Engl., 2012, 28, 1086–92
[39] Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, GigaScience, 2012, 1, 18
[40] Chevreux B., Wetter T., Suhai S., Genome Sequence Assembly Using Trace Signals and Additional Sequence Information., Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma, GCB 99., 1999, 45–56
[41] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., Basic local alignment search tool, J Mol Biol., 1990, 215, 403–10
[42] Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., et al., Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data, Bioinforma. Oxf. Engl., 2012, 28, 1647–9
[43] Goecks J., Nekrutenko A., Taylor J., Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol., 2010, 11, R86
[44] Kottelat M., Freyhof J., Handbook of European freshwater fishes. Publications Kottelat., Berlin: Kottelat, Cornal &Freyhof;, 2007.
[45] Winnepenninckx B., Backeljau T., De Wachter R., Extraction of high molecular weight DNA from molluscs, Trends Genet., 1993, 9, 407
[46] Geiger M.F., Herder F., Monaghan M.T., Almada V., Barbieri R., Bariche M., et al., Spatial heterogeneity in the Mediterranean Biodiversity Hotspot affects barcoding accuracy of its freshwater fishes, Mol. Ecol. Resour., 2014, 14, 1210–21
[47] Edgar R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res., 2004, 32, 1792–7
[48] Costedoat C., Chappaz R., Barascud B., Guillard O., Gilles A., Heterogeneous colonization pattern of European Cyprinids, as highlighted by the dace complex (Teleostei: Cyprinidae), Mol. Phylogenet. Evol., 2006, 41
[49] Wang F., Niu J., Hu S., Xie P., Liu C., Li H., et al., The complete mitochondrial genome of *Leuciscus idus* (Cypriniformes: Cyprinidae), Mitochondrial DNA., 2014
[50] Jun G Inoue M.M., Complete mitochondrial DNA sequence of the Japanese eel *Anguilla japonica*, Fish. Sci., 2001, 67, 118–25
[51] Kawaguchi A., Miya M., Nishida M., Complete mitochondrial DNA sequence of *Aulopus japonicus* (Teleostei: Aulopiformes), a basal Eurypterygii: longer DNA sequences and higher-level relationships, Ichthyol. Res., 2001, 48, 213–23
[52] Kim I.-C., Kweon H.-S., Kim Y.J., Kim C.-B., Gye M.C., Lee W.-O., et al., The complete mitochondrial genome of the javeline goby *Acanthogobius hasta* (Perciformes, Gobiidae) and phylogenetic considerations, Gene, 2004, 336, 147–53
[53] Miya M., Nishida M., Organization of the Mitochondrial Genome of a Deep-Sea Fish, *Gonostoma gracile* (Teleostei: Stomiiformes): First Example of Transfer RNA Gene Rearrangements in Bony Fishes, Mar. Biotechnol. N. Y. N., 1999, 1, 416–0426
[54] Poulsen J.Y., Byrkjedal I., Willassen E., Rees D., Takeshima H., Satoh T.P., et al., Mitogenomic sequences and evidence from unique gene rearrangements corroborate evolutionary relationships of myctophiformes (Neoteleostei), BMC Evol. Biol., 2013, 13, 111
[55] Dohm J., Lottaz C., Borodina T., Himmelbauer H., Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, Nucleic Acids Res., 2008, 36
[56] Ross M.G., Russ C., Costello M., Hollinger A., Lennon N.J., Hegarty R., et al., Characterizing and measuring bias in sequence data, Genome Biol., 2013, 14, R51
[57] Zhuang X., Cheng C.H., ND6 gene "lost" and found: evolution of mitochondrial gene rearrangement in Antarctic notothenioids, Mol. Biol. Evol., 2010, 27, 1391–403
[58] Antunes A., Ramos M.J., Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes, Genomics, 2005, 86, 708–17
[59] Venkatesh B., Dandona N., Brenner S., Fugu genome does not contain mitochondrial pseudogenes, Genomics, 2006, 87, 307–10
[60] Hazkani-Covo E., Zeller R.M., Martin W., Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes, PLoS Genet., 2010, 6, e1000834
[61] Zhang J., Hanner R., Molecular Approach to the Identification of Fish in the South China Sea, PLoS One, 2012, 7, e30621
[62] Kawahara R., Miya M., Mabuchi K., Near TJ., Nishida M., Stickleback phylogenies resolved: evidence from mitochondrial genomes and 11 nuclear genes, Mol. Phylogenet. Evol., 2009, 50, 401–4
[63] Sorenson M.D., Quinn T.W., Numts: A challenge for avian systematics and population biology, The Auk, 1998, 115, 214–21
[64] Collura R.V., Stewart C.B., Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids, Nature, 1995, 378, 485–9
[65] Sato A., O'hUigin C., Figueroa F., Grant P.R., Grant B.R., Tichy H., et al., Phylogeny of Darwin's finches as revealed by mtDNA sequences, Proc. Natl. Acad. Sci. U. S. A., 1999, 96, 5101–6

[66] Hwang U.W., Park C.J., Yong T.S., Kim W., One-step PCR amplification of complete arthropod mitochondrial genomes, Mol. Phylogenet. Evol., 2001, 19, 345–52
[67] Pons J., Bauzà-Ribot M.M., Jaume D., Juan C., Next-generation sequencing, phylogenetic signal and comparative mitogenomic analyses in Metacrangonyctidae (Amphipoda: Crustacea), BMC Genomics, 2014, 15, 566
[68] Green R.E., Malaspinas A.-S., Krause J., Briggs A.W., Johnson P.L.F., Uhler C., et al., A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing, Cell, 2008, 134, 416–26
[69] Tsai I., Otto T., Berriman M., Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps, Genome Biol., 2010, 11, R41
[70] Dettai A., Lecointre G., New insights into the organization and evolution of vertebrate IRBP genes and utility of IRBP gene sequences for the phylogenetic study of the Acanthomorpha (Actinopterygii: Teleostei), Mol. Phylogenet. Evol., 2008, 48, 258–69
[71] Carr S.M., Marshall H.D., Intraspecific Phylogeographic Genomics From Multiple Complete mtDNA Genomes in Atlantic Cod (*Gadus morhua*): Origins of the "Codmother," Transatlantic Vicariance and Midglacial Population Expansion, Genetics, 2008, 180, 381–9
[72] Dettai A., Lautredou A.-C., Bonillo C., Goimbault E., Busson F., Causse R., et al., The actinopterygian diversity of the CEAMARC cruises: Barcoding and molecular taxonomy as a multi-level tool for new findings, Deep Sea Res. Part II Top. Stud. Oceanogr., 2011, 58, 250–63
[73] Rokas A., Carroll S.B., More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy, Mol. Biol. Evol., 2005, 22, 1337–44
[74] Miya M., Friedman M., Satoh T.P., Takeshima H., Sado T., Iwasaki W., et al., Evolutionary Origin of the Scombridae (Tunas and Mackerels): Members of a Paleogene Adaptive Radiation with 14 Other Pelagic Fish Families, PLoS ONE., 2013, 8, e73535
[75] April J., Mayden R.L., Hanner R.H., Bernatchez L., Genetic calibration of species diversity among North America's freshwater fishes, Proc. Natl. Acad. Sci. U. S. A., 2011, 108, 10602–7
[76] Denys G.P.J., Dettai A., Persat H., Hautecœur M., Keith P., Morphological and molecular evidence of three species of pikes *Esox* spp. (Actinopterygii, Esocidae) in France, including the description of a new species, C. R. Biol., 2014, 337, 521–34
[77] Knebelsberger T., Dunz A.R., Neumann D., Geiger M.F., Molecular diversity of Germany's freshwater fishes and lampreys assessed by DNA barcoding, Mol. Ecol. Resour., 2014
[78] Hubert N., Hanner R., Holm E., Mandrak N.E., Taylor E., Burridge M., et al., Identifying Canadian freshwater fishes through DNA barcodes, PloS One., 2008, 3, e2490
[79] Brodersen J., Seehausen O., Why evolutionary biologists should get seriously involved in ecological monitoring and applied biodiversity assessment programs, Evol. Appl., 2014, 7, 968–83
[80] Padial JM., Miralles A., De la Riva I., Vences M., The integrative future of taxonomy, Front. Zool., 2010, 7, 1–16
[81] Karp P.D., What we do not know about sequence analysis and sequence databases, Bioinformatics, 1998, 14, 753–4
[82] Funk D.J., Omland K.E., Species level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA, Annu. Rev. Ecol. Evol. Syst., 2003, 34, 397–423

Supplemental Material: The online version of this article (DOI: 10.1515/dna-2015-0019) offers supplementary material.