GigaScience, 8, 2019, 1–17

doi: 10.1093/gigascience/giz029
Advance Access Publication Date: 0 2019
Technical Note

TECHNICAL NOTE

An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies

Jan Axtner[1,*], Alex Crampton-Platt[1], Lisa A. Hörig[1], Azlan Mohamed[1], Charles C.Y. Xu[2,3,4], Douglas W. Yu[2,5] and Andreas Wilting[1]

[1]Leibniz Institute for Zoo and Wildlife Research, Department of Ecological Dynamics, Alfred-Kowalke-Str. 17, 10315 Berlin, Germany;, [2]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 32 Jiaochang East Rd, Kunming, Yunnan 650223, China;, [3]Groningen Institute for Evolutionary Life Sciences, University of Groningen, P.O. Box 11103, 9700 CC Groningen, The Netherlands;, [4]Redpath Museum and Department of Biology, McGill University 859 Sherbooke Street West, Montreal, PQ, Canada H3A 2K6; and [5]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR47TJ, UK

*Correspondence address. Jan Axtner, Leibniz Institute for Zoo and Wildlife Research, Department of Ecological Dynamics, Alfred-Kowalke-Str. 17, 10315 Berlin, Germany, E-mail: axtner@izw-berlin.de http://orcid.org/0000-0003-1269-5586

Abstract

Background: The use of environmental DNA for species detection via metabarcoding is growing rapidly. We present a co-designed lab workflow and bioinformatic pipeline to mitigate the 2 most important risks of environmental DNA use: sample contamination and taxonomic misassignment. These risks arise from the need for polymerase chain reaction (PCR) amplification to detect the trace amounts of DNA combined with the necessity of using short target regions due to DNA degradation. **Findings:** Our high-throughput workflow minimizes these risks via a 4-step strategy: (i) technical replication with 2 PCR replicates and 2 extraction replicates; (ii) using multi-markers (*12S*, *16S*, *CytB*); (iii) a "twin-tagging," 2-step PCR protocol; and (iv) use of the probabilistic taxonomic assignment method PROTAX, which can account for incomplete reference databases. Because annotation errors in the reference sequences can result in taxonomic misassignment, we supply a protocol for curating sequence datasets. For some taxonomic groups and some markers, curation resulted in >50% of sequences being deleted from public reference databases, owing to (i) limited overlap between our target amplicon and reference sequences, (ii) mislabelling of reference sequences, and (iii) redundancy. Finally, we provide a bioinformatic pipeline to process amplicons and conduct PROTAX assignment and tested it on an invertebrate-derived DNA dataset from 1,532 leeches from Sabah, Malaysia. Twin-tagging allowed us to detect and exclude sequences with non-matching tags. The smallest DNA fragment (*16S*) amplified most frequently for all samples but was less powerful for discriminating at species rank. Using a stringent and lax acceptance criterion we found 162 (stringent) and 190 (lax) vertebrate detections of 95 (stringent) and 109 (lax) leech samples. **Conclusions:** Our metabarcoding workflow should help research groups increase the robustness of their results and therefore facilitate wider use of environmental and invertebrate-derived DNA, which is turning into a valuable source of ecological and conservation information on tetrapods.

Received: 13 June 2018; **Revised:** 12 October 2018; **Accepted:** 7 March 2019

Keywords: metabarcoding; invertebrate-derived -DNA; environmental DNA; leeches

Introduction

Monitoring, or even detecting, elusive or cryptic species in the wild can be challenging. In recent years there has been an increase in the availability of cost-effective DNA-based methods made possible by advances in high-throughput DNA sequencing (HTS). One such method is eDNA metabarcoding, which seeks to identify the species present in a habitat from traces of "environmental DNA" (eDNA) in substrates such as water, soil, or faeces. A variant of eDNA metabarcoding, known as "invertebrate-derived DNA" (iDNA) metabarcoding, targets the genetic material of prey or host species extracted from copro-, sarco-, or haematophagous invertebrates. Examples include ticks [1], blowflies or carrion flies [2–5], mosquitoes [6–9], and leeches [10–13]. Many of these parasites are ubiquitous, highly abundant, and easy to collect, making them an ideal source of biodiversity data, especially for terrestrial vertebrates that are otherwise difficult to detect [10, 14, 15]. In particular, the possibility for bulk collection and sequencing to screen large areas and minimize costs is attractive. However, most of the recent iDNA studies focus on single-specimen DNA extracts and Sanger sequencing and thus are not making use of the advances of HTS and a metabarcoding framework for carrying out larger scale biodiversity surveys.

That said, e/iDNA metabarcoding also poses several challenges, due to the low quality and low amounts of target DNA available, relative to non-target DNA (including the high-quality DNA of the live-collected, invertebrate vector). In bulk iDNA samples comprising many invertebrate specimens, this problem is further exacerbated by the variable time since each individual has fed, if at all, leading to differences in the relative amounts and degradation of target DNA per specimen. This makes e/iDNA studies similar to ancient DNA samples, which also pose the problem of low quality and low amounts of target DNA [16, 17]. The great disparity in the ratio of target to non-target DNA and the low overall amount of the former necessitates an enrichment step, which is achieved via the amplification of a short target sequence (amplicon) by polymerase chain reaction (PCR) to obtain enough target material for sequencing. However, this enrichment step can result in false-positive species detections, either through sample cross-contamination or through volatile short PCR amplicons in the laboratory, and in false-negative results, through primer bias and low concentrations of template DNA. Although laboratory standards to prevent and control for such false results are well established in the field of ancient DNA, there are still no best-practice guidelines for e/iDNA studies, and thus few studies sufficiently account for such problems [18].

The problem is exacerbated by the use of "universal" primers used for the PCR, which maximize the taxonomic diversity of the amplified sequences. This makes the method a powerful biodiversity assessment tool, even where little is known *a priori* about which species might be found. However, using such primers, in combination with low quality and quantity of target DNA, which often necessitates a high number of PCR cycles to generate enough amplicon products for sequencing, makes metabarcoding studies particularly vulnerable to false results [13, 19, 20]. The high number of PCR cycles, combined with the high sequencing depth of HTS, also increase the likelihood that contaminants are amplified and detected, possibly to the same or greater extent as some true-positive trace DNA. As e/iDNA have been proposed as tools to detect very rare and high-priority conservation species such as the saola, *Pseudoryx nghetinhensis* [10], false detection might result in misdirected conservation activities worth several hundreds of thousands of US dollars such as for the ivory-billed woodpecker, where most likely false evidence of the bird's existence has been overemphasized to shore up political and financial support for saving it [21]. Therefore, similar to ancient DNA studies, great care must be taken to minimize the possibility for cross-contamination in the laboratory and to maximize the correct detection of species through proper experimental and analytical design. Replication in particular is an important tool for reducing the incidence of false-negative and false-positive results, but the trade-off is increased cost, workload, and analytical complexity [19].

An important source of false-positive species detections is the incorrect assignment of taxonomies to the millions of short HTS reads generated by metabarcoding. Although there has been a proliferation of tools focused on this step, most can be categorized into just 3 groups depending on whether the algorithm utilizes sequence similarity searches, sequence composition models, or phylogenetic methods [22–24]. The 1 commonality among all methods is the need for a reliable reference database of correctly identified sequences, yet there are few curated databases currently appropriate for use in e/iDNA metabarcoding. Two exceptions are SILVA [25] for the nuclear markers short subunit and long subunit ribosomal RNA (rRNA) used in microbial ecology, and the Barcode of Life Database [26] for the *COI* "DNA barcode" region. For other loci, a non-curated database downloaded from the International Nucleotide Sequence Database Collaboration (INSDC; e.g., GenBank) is generally used. However, the INSDC places the burden for metadata accuracy, including taxonomy, on the sequence submitters, with no restriction on sequence quality or veracity. For instance, specimen identification is often carried out by non-specialists, which increases error rates, and common laboratory contaminant species (e.g., human DNA sequences) are sometimes submitted in lieu of the sample itself. The rate of sequence mislabelling in fungi has been assessed for GenBank, where it was found to be up to 20% [27], and it is an issue that is often neglected [28, 29]. For several curated microbial databases (Greengenes, All-species Living Tree Project [LTP], Ribosomal Database Project, SILVA), mislabelling rates have been estimated at between 0.2% and 2.5% [30]. Given the lack of professional curation it is likely that the true proportion of mislabelled samples in GenBank is somewhere between these numbers. Moreover, correctly identifying such errors is labour intensive, so most metabarcoding studies simply base their taxonomic assignments on sequence-similarity searches of the whole INSDC database (e.g., with BLAST) [3, 10, 12] and thus can only detect errors if assignments are ecologically unlikely. Furthermore, reference sequences for the species that are likely to be sampled in e/iDNA studies are often underrepresented in or absent from these databases, which increases the possibility of incorrect assignment. For instance, <50% of species occurring in a tropical megadiverse rainforest are represented in Genbank (see findings below). When species-level matches are ambiguous, it might still be possible to assign a sequence to a higher taxonomic rank by using an appropriate algorithm such as Metagenome Analyzer's (MEGAN) Lowest Common Ancestor [31] or PROTAX [32].

We present here a complete laboratory workflow and complementary bioinformatics pipeline, starting from DNA extrac-

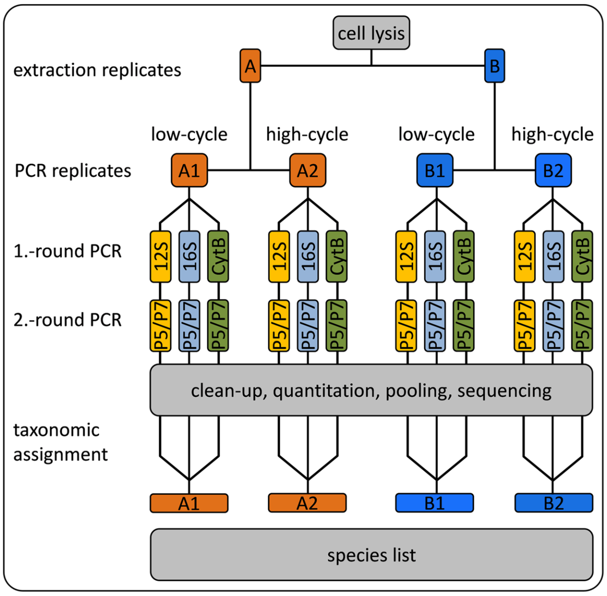

Figure 1: Laboratory scheme; during DNA extraction the sample is split into 2 Extraction Replicates A and B. Our protocol consists of 2 rounds of PCR during which the sample tags, the necessary sequencing primer, and sequencing adapters are added to the the amplicons. For each extraction replicate we ran a low-cycle PCR and a high-cycle PCR for each marker such that we have 12 independent PCR replicates per sample. All PCR products were sequenced and the obtained reads were taxonomically identified with PROTAX.

tion to taxonomic assignment of HTS reads using a curated reference database. The laboratory workflow allows for efficient screening of hundreds of e/iDNA samples. The workflow includes the following steps: (i) 2 ”extraction replicates” are separated during DNA extraction, and each is sequenced in 2 PCR replicates (Fig. 1); (ii) robustness of taxonomic assignment is improved by using up to 3 mitochondrial markers; (iii) a “twin-tagged,” 2-step PCR protocol prevents cross-sample contamination because no unlabelled PCR products are produced (Fig. 2) while also allowing for hundreds of PCR products to be pooled before costly Illumina library preparation; and (iv) our bioinformatics pipeline includes a standardized, automated, and replicable protocol to create a curated database, which allows updating as new reference sequences become available, and to be expanded to other amplicons. We provide scripts for processing raw sequence data to quality-controlled dereplicated reads and for taxonomic assignment of these reads using PROTAX [32], a probabilistic method that has been shown to be robust even when reference databases are incomplete [23, 4] (all scripts are available from GitHub [33]).

Methods

Establishment of the tetrapod reference database

Reference database

A custom bash script was written to generate a tetrapod reference database for ≤4 mitochondrial markers—a short 93–base pair (bp) fragment of *16S* rRNA (*16S*), a 389-bp fragment of *12S* rRNA (*12S*), a 302-bp fragment of cytochrome b (*CytB*), and a 250-bp mitochondrial cytochrome c oxidase subunit I amplicon (*COI*) that has previously been used in iDNA studies [2]. An important time-saving step was the use of the FASTA-formatted Midori mitochondrial database [34], which is a lightly curated subset of Genbank. Our script updated the FASTA files with a subset of target species, removed errors and redundancy, trimmed the sequences to include only the amplicon regions, and output FASTA files with species names and GenBank accessions in the headers.

The script accepts 4 data inputs, 2 of which are optional. The required inputs are (i) the Midori sequences (December 2015 “UNIQUE,” downloaded from [35]) for the relevant genes and (ii) an initial reference taxonomy of tetrapods. This taxonomy is needed to find or generate a full taxonomic classification for each sequence because the taxonomies in Midori are from Genbank and thus include incorrect, synonymized, or incomplete taxonomies. Here we used the Integrated Taxonomic Information System (ITIS) classification for Tetrapoda, obtained with the R package ”taxize” version 0.9.0 ([36], functions ”downstream” and ”classification”). The optional inputs are (iii) supplementary FASTA files of reference sequences that should be added to the database and (iv) a list of target species to be queried on GenBank to capture any sequences published since the December 2015 Midori dataset was generated.

For this study, 72 recently published [37] and 7 unpublished partial mitochondrial mammal genomes (accession Nos. MH464789, MH464790, MH464791, MH464792, MH464793, MH464794, MH464795, MH464796, MH464797, MH464798, MH464799, MH464800, MH464801) were added as input (iii). A list of 103 mammal species known to be present in the sampling area plus *Homo sapiens* and our positive control *Myodes glareolus* was added as input (iv).

With the above inputs, the 7 curation steps are as follows: (1) remove sequences not identified to species; (2) add extra sequences from optional inputs (iii) and (iv) above; (3) trim the sequences to leave only the target amplicon; (4) remove sequences with ambiguities; (5) compare species names from the Midori dataset to the reference taxonomy from input (ii) and replace with a consensus taxonomy; (6) identify and remove putatively mislabelled sequences; and (7) dereplicate sequences, retaining 1 haplotype per species.

The script is split into 4 modules, allowing optional manual curation at 3 key steps. The steps covered by each of the 4 modules are summarized in Table 1. The main programs used are highlighted and cited in the text where relevant, but many intermediate steps used common UNIX tools and unpublished lightweight utilities freely available from GitHub (Table 2).

Module 1 The first step is to select the tetrapod sequences from the Midori database for each of the 4 selected loci (input [i] above). This, and the subsequent step to discard sequences without strict binomial species names and reduce subspecies identifications to species level, are made possible by the inclusion of the full National Center for Biotechnology Information (NCBI) taxonomic classification of each sequence in the FASTA header by the Midori pipeline. The headers of the retained sequences are then reformatted to include just the species name and GenBank accession separated by underscores. If desired, additional sequences from local FASTA files are now added to the Midori set (input [iii]). The headers of these FASTA files are required to be in the same format. Next, optional queries are made to the NCBI GenBank and RefSeq databases for each species in a provided list (input [iv]) for each of the 4 target loci, using NCBI’s Entrez Direct [38]. Matching sequences are downloaded in FASTA format, sequences prefixed as “UNVERIFIED” are discarded, the headers are simplified as previously, and those sequences not already in the Midori set are added. Trimming each sequence down to the relevant target marker was car-

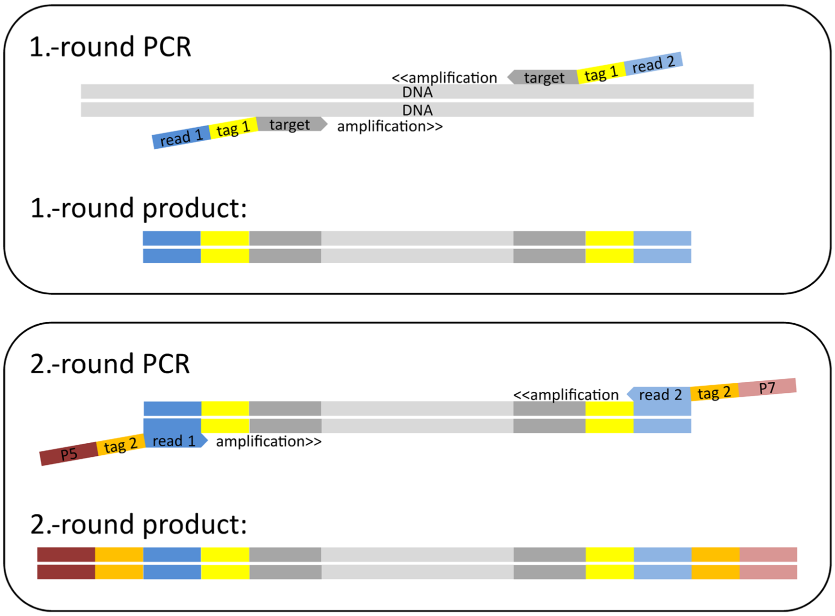

Figure 2: Scheme to build double "twin-tagged" PCR libraries. The first round of PCR uses target-specific primers (*12S*, *16S*, or *CytB*, dark grey) that have both been extended with the same (i.e., "twin") sample-identifying tag sequences Tag 1 (yellow) and then with the different Read 1 (dark blue) and Read 2 (light blue) sequence primers. The second round of PCR uses the priming sites of the Read 1 and Read 2 sequencing primers to add twin plate-identifying tag sequences Tag 2 (orange) and the P5 (dark red) and P7 (light red) Illumina adapters.

Table 1: Main steps undertaken by each module of the database curation script

Module	Steps
Module 1	Extract subset of raw Midori database for query taxon and loci Remove sequences with non-binomial species names, reduce subspecies to species labels Add local sequences (optional) Check for relevant new sequences for list of query species on NCBI (GenBank and RefSeq) (optional) Select amplicon region and remove primers Remove sequences with ambiguous bases Align End of module: optional check of alignments
Module 2	Compare sequence species labels with taxonomy Non-matching labels queried against Catalogue of Life to check for known synonyms Remaining mismatches kept if genus already exists in taxonomy, otherwise flagged for removal End of module: optional check of flagged species labels
Module 3	Discard flagged sequences Update taxonomy key file for sequences found to be incorrectly labelled in Module 2 Run SATIVA End of module: optional check of putatively mislabelled sequences
Module 4	Discard flagged sequences Finalize consensus taxonomy and relabel sequences with correct species label and accession number Select 1 representative sequence per haplotype per species

ried out in a 2-step process in which usearch (–search_pcr) was used to select sequences where both primers were present, and these were in turn used as a reference dataset for blastn to select partially matching sequences from the rest of the dataset [39, 40]. Sequences with a hit length of ≥90% of the expected marker length were retained by extracting the relevant subsequence based on the BLAST hit co-ordinates. Sequences with ambiguous bases were discarded at this stage. In the final step in module 1, a multiple-sequence alignment was generated with MAFFT (MAFFT, RRID:SCR_011811) [41, 42] for each partially curated amplicon dataset (for the SATIVA step below). The script then breaks to allow the user to check for any obviously problematic sequences that should be discarded before continuing.

Module 2 The species labels of the edited alignments are compared with the reference taxonomy (input [ii]). Any species not found is queried against the Catalogue of Life database via "taxize" in case these are known synonyms, and the correct species label and classification is added to the reference taxonomy. The original species label is retained as a key to facilitate sequence renaming, and a note is added to indicate its status as a synonym. Finally, the genus name of any species not found in the Catalogue of Life is searched against the consensus taxonomy, and if found, the novel species is added by taking the higher classification levels from one of the other species in the genus. Orphan species labels are printed to a text file, and the script breaks to allow the user to check this list and manually create classifications for some or all if appropriate.

Module 3 This module begins by checking for any manually generated classification files (from the end of Module 2) and merging them with the reference taxonomy from Module 2. Any remaining sequences with unverifiable classifications are removed at this step. The next steps convert the sequences and taxonomy file to the correct formats for SATIVA [30], which detects possibly mislabelled sequences by generating a maximum likelihood phylogeny from the alignment in Module 1 and comparing each sequence's taxonomy against its phylogenetic neighbors. Sequence headers in the edited MAFFT alignments are reformatted to include only the GenBank accession, and a taxonomy key file is generated with the correct classification listed for each accession number. In cases where the original species label is found to be a synonym, the corrected label is used. Putatively mislabelled sequences in each amplicon are then detected with SATIVA, and the script breaks to allow inspection of the results. The user may choose to make appropriate edits to the taxonomy key file or a list of putative mislabels at this point.

Module 4 Any sequences that are still flagged as mislabelled at the start of the fourth module are deleted from the SATIVA input alignments, and all remaining sequences are relabelled with the correct species name and accession. A final consensus taxonomy file is generated in the format required by PROTAX. Alignments are subsequently unaligned prior to species-by-species selection of a single representative per unique haplotype. Sequences that are the only representative of a species are automatically added to the final database. Otherwise, all sequences for each species are extracted in turn, aligned with MAFFT, and collapsed to unique haplotypes with collapsetypes_4.6.pl (0 differences allowed [43]). Representative sequences are then unaligned and added to the final database.

iDNA samples

We used 242 collections of haematophagous terrestrial leeches from Deramakot Forest Reserve in Sabah, Malaysian Borneo, stored in RNA fixating saturated ammonium sulfate solution as samples. Each sample consisted of 1–77 leech specimens (median, 4). In total, 1,532 leeches were collected, exported under the permit (JKM/MBS.1000-2/3 JLD.2 [8] issued by the Sabah Biodiversity Council), and analysed at the laboratories of the Leibniz Institute for Zoo and Wildlife Research (Leibniz-IZW).

Laboratory workflow

The laboratory workflow is designed both to minimize the risk of sample cross-contamination and to aid identification of any instances that do occur. All laboratory steps (extraction, pre- and post-PCR steps, sequencing) took place in separate laboratories and no samples or materials were allowed to re-enter upstream laboratories at any point in the workflow. All sample handling was carried out under specific hoods that were wiped with bleach, sterilized, and UV irradiated for 30 minutes after each use. All laboratories are further UV irradiated for 4 hours each night.

DNA extraction

DNA was extracted from each sample in bulk. Leeches were cut into small pieces with a fresh scalpel blade and incubated in lysate buffer (proteinase K and ATL buffer at a ratio of 1:10; 0.2 ml per leech) overnight at 55°C (≥12 hours) in an appropriately sized vessel for the number of leeches (2- or 5-ml reaction tube). For samples with >35 leeches, the reaction volume was split in 2 and recombined after lysis.

Each lysate was split into 2 extraction replicates (A and B; maximum volume, 600 µl) and all further steps were applied to these independently. We followed the DNeasy 96 Blood & Tissue protocol for animal tissues (Qiagen, Hilden, Germany) on 96 plates for cleanup. DNA was eluted twice with 100 μl Tris-ethylenediaminetetraacetic acid buffer. DNA concentration was measured with PicoGreen dsDNA Assay Kit (Quant-iT, ThermoFisherScientific, Waltham, MA, USA) in 384-well plate format using an appropriate plate reader (200 PRO NanoQuant, Tecan Trading AG, Männedorf, Switzerland). Finally, all samples were diluted to a maximum concentration of 10 ng/μl.

Two-round PCR protocol

We amplified 3 mitochondrial markers—a short 93-bp fragment of *16S* rRNA (*16S*), a 389-bp fragment of *12S* rRNA (*12S*), and a 302-bp fragment of cytochrome b (*CytB*). For each marker, we ran a 2-round PCR protocol (Figs 1, 2). The primers were chosen on the expectation of successful DNA amplification over a large number of tetrapod species [44, 45], and we tested the fit of candidate primers on an alignment of available mitochondrial sequences of 134 southeast Asian mammal species. Primer sequences are in Table 3.

Primer modification We modified primers of the 3 markers to avoid the production of unlabelled PCR products, to allow the detection and deletion of tag-jumping events [51], and to reduce the cost of primers and library preparation. We used 2 rounds of PCR. The first round amplified the target gene and attached 1 of 25 different "twin-tag" pairs (Tag 1), identifying the sample within a given PCR. By "twin-tag," we mean that both the forward and reverse primers were given the same sample-identifying sequence ("tags") added as primer extensions (Fig. 2). The tags differed with a pairwise distance of ≥3 nucleotides ([51]; Supplemental Table S1). These primers also contained different forward and reverse sequences (Read 1 and Read 2 sequence

Table 2: GNU core utilities and other lightweight tools used for manipulation of text and sequence files

Tool	Function	Source
awk, cut, grep, join, sed, sort, tr	Processing text files	GNU core utilities
seqbuddy	Processing FASTA/Q files	[46]
seqkit	Processing FASTA/Q files	[47]
seqtk	Processing FASTA/Q files	[48]
tabtk	Processing tab-delimited text files	[49]

Table 3: Sequence motifs that compose the 25 different target primers for the first and the second PCR. First PCR primers consist of target-specific primer followed by an overhang out of sample-specific Tag 1 and Read 1 and Read 2 sequencing primer, respectively. The second PCR primers consist of the Read 1 or the Read 2 sequencing primer followed by a plate-specific Tag 2 and the P5 and P7 adapters, respectively (see also Fig. 2).

Name	Sequence	Reference
Tag A	TGCAT	[50]
Tag B	TCAGC	[50]
Tag C	AAGCG	[50]
Tag D	ACAAG	[50]
Tag E	AGTGG	[50]
Tag F	TTGAC	[50]
Tag G	CCTAT	[50]
Tag H	GGATG	[50]
Tag I	CTAGG	[50]
Tag K	CACCT	[50]
Tag L	GTCAA	[50]
Tag M	GAAGT	[50]
Tag N	CGGTT	[50]
Tag O	ACCGA	[50]
Tag P	ACGTC	[50]
Tag Q	AGACT	[50]
Tag R	AGGAA	[50]
Tag S	ATTCC	[50]
Tag T	CAATC	[50]
Tag V	CATGA	[50]
Tag W	CCACA	[50]
Tag X	GCTTA	[50]
Tag Y	GGTAC	[50]
Tag Z	AACAC	[50]
Tag Control	ATCTG	[50]
CytB-fw	AAAAAGCTTCCATCCAACATCTCAGCATGATGAAA	[44]
CytB-rv	AAACTGCAGCCCCTCAGAATGATATTTGTCCTCA	[44]
16S-fw	CGGTTGGGGTGACCTCGGA	[45]
16S-rv	GCTGTTATCCCTAGGGTAACT	[45]
12S-fw	AAAAAGCTTCAAACTGGGATTAGATACCCCACTAT	[44]
12S-rv	TGACTGCAGAGGGTGACGGGCGGTGTGT	[44]
Read 1 sequence primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	Illumina Document 1000000002694 v03
Read 2 sequence primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina Document 1000000002694 v03
P5 adapter	AATGATACGGCGACCACCGAGATCTACAC	Illumina Document 1000000002694 v03
P7 adapter	CAAGCAGAAGACGGCATACGAGAT	Illumina Document 1000000002694 v03

primers) (Supplemental Table S1) to act as priming sites for the second PCR round (Fig. 2).

The second round added the Illumina adapters for sequencing and attached 1 of 20 twin-tag pairs (Tag 2) identifying the PCR, with a pairwise distance of ≥3 [50]. These primers also contained the Illumina P5 and P7 adapter sequences (Fig. 2). Thus, no unlabelled PCR products were ever produced, and the combination of Tags 1 and 2 allowed the pooling of ≤480 (= 24 × 20) samples in a single library preparation step (one Tag 1 was reserved for controls). Twin tags allowed us later to detect and delete tag-jumping events [51] (Fig. 2).

Cycle number considerations Because we know that our target DNA is at low concentration in the samples, we are faced with a trade-off between (i) using fewer PCR cycles (e.g., 30) to minimize amplification bias (caused by some target DNA binding better to the primer sequences and thus outcompeting other target sequences that bind less well [52]) and (ii) using more PCR cycles (e.g., 40) to ensure that low-concentration target DNA is sufficiently amplified in the first place. Rather than choose between these 2 extremes, we ran both low- and high-cycle protocols and sequenced both sets of amplicons.

Thus, each of the 2 Extraction Replicates A and B was split and amplified using different numbers of cycles (PCR Replicates 1 and 2) for a total of 4 (= 2 extraction replicates x 2 PCR replicates → A1/A2 and B1/B2) replicates per sample per marker (Fig. 1). For PCR Replicates A1/B1, we used 30 cycles in the first PCR round to minimize the effect of amplification bias. For PCR Replicates A2/B2, we used 40 cycles in the first PCR round to increase the likelihood of detecting species with very low levels of input DNA (Fig. 1).

PCR protocol The first-round PCR reaction volume was 20 μl, including 0.1 μM primer mix, 0.2 mM dNTPs, 1.5 mM $MgCl_2$, 1x PCR buffer, 0.5 U AmpliTaq Gold (Invitrogen, Karlsruhe, Germany), and 2 μl of template DNA. Initial denaturation was 5 minutes at 95°C, followed by repeated cycles of 30 seconds at 95°C, 30 seconds at 54°C, and 45 seconds at 72°C. Final elongation was 5 minutes at 72°C. Samples were amplified in batches of 24 plus a negative (water) and a positive control (bank vole, *Myodes glareolus* DNA). All 3 markers were amplified simultaneously in individual wells for each batch of samples in a single PCR plate. Non-target by-products were removed as required from some 12S PCRs by purification with magnetic Agencourt AMPure beads (Beckman Coulter, Krefeld, Germany).

In the second-round PCR, we used the same PCR protocol as above with 2 μl of the product of the first-round PCR and 10 PCR cycles.

Quality control and sequencing

Amplification was visually verified after the second-round PCR by means of gel electrophoresis on 1.5% agarose gels. Controls were additionally checked with a TapeStation 2200 (D1000 ScreenTape assay, Agilent, Waldbronn, Germany). All samples were purified with AMPure beads, using a bead-to-template ratio of 0.7:1 for *12S* and *CytB* products, and a ratio of 1:1 for *16S* products. DNA concentration was measured with PicoGreen dsDNA as described above. Sequencing libraries were made by equimolar pooling of all positive amplifications; final concentrations were between 2 and 4 nmol. Because of different amplicon lengths and therefore different binding affinities to the flow cell, *12S* and *CytB* products were combined in a single library, whereas positive *16S* products were always combined in a separate library. *12S/CytB* libraries were sequenced independently from *16S* libraries. Apart from our negative controls, we did not include samples that did not amplify because this would have resulted in highly diluted libraries. Up to 11 libraries were sequenced on each run of Illumina MiSeq, following standard protocols. Libraries were sequenced with MiSeq Reagent Kit V3 (600 cycles, 300 bp paired-end reads) and had a final concentration of 11 pM spiked with 20–30% of PhiX control.

Bioinformatics workflow

Read processing

Although the curation of the reference databases is our main focus, it is just 1 part of the bioinformatics workflow for e/iDNA metabarcoding. A custom bash script was used to process raw basecall files into demultiplexed, cleaned, and dereplicated reads in FASTQ format on a run-by-run basis. All runs and amplicons were processed with the same settings unless otherwise indicated. bcl2fastq (Illumina) was used to convert the basecall file from each library to paired-end FASTQ files, demultiplexed into the separate PCRs via the Tag 2 pairs, allowing $\leq$1 mismatch in each Tag 2. Each FASTQ file was further demultiplexed into samples via the Tag 1 pairs using AdapterRemoval (AdapterRemoval, RRID:SCR_011834) [53], again allowing $\leq$1 mismatch in each tag. These steps allowed reads to be assigned to the correct samples.

In all cases, amplicons were short enough to expect paired reads to overlap. For libraries with >1,000 reads pairs were merged with usearch (–fastq_mergepairs [54, 55]), and only successfully merged pairs were retained. For libraries with >500 merged pairs the primer sequences were trimmed away with cutadapt (cutadapt, RRID:SCR_011841) [56], and only successfully trimmed reads $\geq$90% of expected amplicon length were passed to a quality-filtering step with usearch (–fastq_filter). Finally, reads were dereplicated with usearch (–derep_fulllength), and singletons were discarded. The number of replicates that each unique sequence represented was also added to the read header at this step (option –sizeout). The number of reads processed at each step for each sample is reported in a standard tab-delimited txt-file.

Taxonomic assignment

The curated reference sequences and associated taxonomy were used for PROTAX taxonomic assignment of the dereplicated reads [24, 32]. PROTAX gives unbiased estimates of placement probability for each read at each taxonomic rank, allowing assignments to be made to a higher rank even when there is uncertainty at the species level. In other words, and unlike other taxonomic assignment methods, PROTAX can estimate the probability that a sequence belongs to a taxon that is not present in the reference database. This was considered an important feature owing to the known incompleteness of the reference databases for tetrapods in the sampled location. As other studies have compared PROTAX with more established methods, e.g., MEGAN [31] (see [4, 24]), it was beyond the scope of this study to evaluate the performance of PROTAX.

Classification with PROTAX is a 2-step process. First, PROTAX selected a subset of the reference database that was used as training data to parameterize a PROTAX model for each marker, and second, the fitted models were used to assign 4 taxonomic ranks (species, genus, family, order) to each of the dereplicated reads, along with a probability estimate at each level. We also included the best similarity score of the assigned species or genus, mined from the LAST results (see below) for each read. This was helpful for flagging problematic assignments for downstream manual inspection, i.e., high probability assignments based on low similarity scores (implying that there are no better matches available) and low probability assignments based on high similarity scores (indicates conflicting database signal from several species with highly similar sequences).

Fitting the PROTAX model followed Somervuo et al. [32] except that 5,000 training sequences were randomly selected for each target marker due to the large size of the reference database. In each case, 4,500 training sequences represented a mix of known species with reference sequences (conspecific sequences retained in the database) and known species without reference sequences (conspecific sequences omitted, simulating species missing from the database), and 500 sequences represented previously unknown lineages distributed evenly across the 4 taxonomic levels (i.e., mimicked a mix of completely novel species, genera, families, and orders). Pairwise sequence similarities of queries and references were calculated with LAST [57] following the approach of Somervuo et al. [32]. The models were weighted towards the Bornean mammals expected in the sampled area by assigning a prior probability of 90% to these 103 species and a 10% probability to all others ([32]; Supplemental Table S2). In cases of missing interspecific variation, this helped to avoid assignments to geographically impossible taxa, especially in the case of the very short 93-bp fragment of 16S. Maximum *a posteriori* parameter estimates were obtained following the approach of Somervuo et al. [24], but the models were parameterized for each of the 4 taxonomic levels independently, with a total of 5 parameters at each level (4 regression coefficients and the probability of mislabelling).

Dereplicated reads for each sample were then classified using a custom bash script on a run-by-run basis. For each sample, reads in FASTQ format were converted to FASTA, and pairwise similarities were calculated against the full reference sequence database for the applicable marker with LAST (LAST, RRID:SCR_006119). Assignments of each read to a taxonomic node based on these sequence similarities were made using a Perl script and the trained model for that level. The taxonomy of each node assignment was added with a second Perl script for a final table including the node assignment, probability, taxonomic level, and taxonomic path for each read. Read count information was included directly in the classification output via the size annotation added to the read headers during dereplication. All Perl scripts to convert input files into the formats expected by PROTAX, R code for training the model following Somervuo et

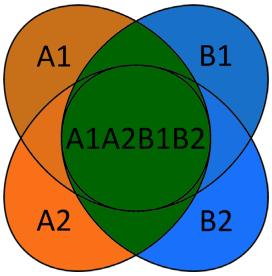

Figure 3: For the stringent acceptance criterion we only accepted taxonomic assignments that were positively detected in both Extraction Replicates A and B (green). The numbers 1 and 2 refer to the 2 PCR replicates for each extraction replicate.

al. [32], and Perl scripts for taxonomic assignment were provided by P. Somervuo (personal communication).

Acceptance criteria

In total we had 12 PCR reactions per sample: 2 Extraction Replicates A and B X 2 PCR Replicates 1 and 2 per extraction replication × the 3 markers (Fig. 1). We applied 2 different acceptance criteria to the data with different stringency regimes: a more naive criterion that accepted any 2 positives out of the 12 PCR replicates (hereafter referred to as ”lax”) and a stringent criterion that only accepted taxonomic assignments that were positively detected in both extraction replicates (A and B, Fig. 3). Our lax approach refers to one of the approaches of Ficetola et al. [19], in which they evaluated different statistical approaches developed to estimate occupancy in the presence of observational errors, and has been applied in other studies (e.g., [13]). The reason for conservatively omitting assignments that appeared in only 1 extraction replicate was to rule out sample cross-contamination during DNA extraction. In addition, we only accepted assignments with ≥10 reads per marker, if only 1 marker was sequenced. If a species was assigned in >1 marker (e.g., *12S* and *16S*), we accepted the assignment even if in 1 sequencing run the number of reads was <10.

Owing to the imperfect PCR amplification of markers (the small *16S* fragment amplified better than the longer *CytB* fragment) and missing reference sequences in the database or shared sequence motifs between species, reads sometimes were assigned to species level for 1 marker but only to genus level for another marker. Thus, the final identification of species could not be automated, and manual inspection and curation was needed. For each assignment, 3 parameters were taken into consideration: number of sequencing reads, the mean probability estimate derived from PROTAX, and the mean sequence similarity to the reference sequences based on LAST.

Shotgun sequencing to quantify mammalian DNA content

Because the success of the metabarcoding largely depends on the mammal DNA quantity in our leech bulk samples, we quantified the mammalian DNA content in a subset of 58 of our leech samples using shotgun sequencing. Extracted DNA was sheared with a Covaris M220 focused-ultra-sonicator to a peak target size of 100–200 bp and rechecked for size distribution. Double-stranded Illumina sequencing libraries were prepared according to a ligation protocol designed by Fortes and Paijmans [58] with single 8-nucleotide indices. All libraries were pooled equimolarly and sequenced on the MiSeq using the v3 150-cycle kit. We demultiplexed reads using bcl2fastq and cutadapt for trimming the adapters. We used BLAST (NCBI BLAST, RRID:SCR_004870) search to identify reads and applied Metagenome Analyzer MEGAN (MEGAN, RRID:SCR_011942) [31] to explore the taxonomic content of the data based on the NCBI taxonomy. Finally we used KRONA (Krona, RRID:SCR_012785) [59] for visualization of the results.

Findings and Discussion

Database curation

The Midori UNIQUE database (December 2015 version) contains 1,019,391 sequences across the 4 mitochondrial loci of interest (*12S*: 66,937; *16S*: 146,164; *CytB*: 223,247; *COI*: 583,043), covering all Metazoa. Of these, 258,225 (25.3%) derive from the 4 tetrapod classes (Amphibia: 55,254; Aves: 51,096; Mammalia: 101,106; Reptilia: 50,769). The distribution of these sequences between classes and loci, and the losses at each curation step are shown in Fig. 4. In 3 of the 4 classes, there is a clear bias towards *CytB* sequences, with >50% of sequences derived from this locus. In both Aves and Mammalia, the *16S* and *12S* loci are severely underrepresented at <10% each, while for Reptilia, *COI* is the least sequenced locus in the database.

The numbers of sequences and rates of loss due to our curation steps varied among taxonomic classes and the 4 loci, although losses were observed between steps in almost all instances. The most substantial losses followed amplicon trimming and removal of non-unique sequences. Amplicon trimming led to especially high losses in Amphibia and *16S*, indicating that data published on GenBank for this class and marker do not generally overlap with our amplicons. Meanwhile, the high level of redundancy in public databases was highlighted by the substantial reduction in the number of sequences during the final step of removing redundant sequences—in all cases >10% of sequences was discarded, with some losses exceeding 50% (Mammalia: *COI*, *CytB*, *16S*; Amphibia: *16S*).

Data loss due to apparent mislabelling ranged between 1.9% and 7.4% and was thus generally higher than similar estimates for curated microbial databases [30]. SATIVA flags potential mislabels and suggests an alternative label supported by the phylogenetic placement of the sequences, allowing the user to make an appropriate decision on a case-by-case basis. The pipeline pauses after this step to allow such manual inspection to take place. However, for the current database, the number of sequences flagged was large (4,378 in total) and the required taxonomic expertise was lacking, so all flagged sequences from nontarget species were discarded to be conservative. The majority of mislabels were identified at species level (3,053), but there were also substantial numbers at genus (788), family (364), and order (102) level. In each amplicon 2–3 sequences from Bornean mammal species were unflagged to retain the sequences in the database. This was important because in each case these were the only reference sequences available for the species. Additionally, *Muntiacus vaginalis* sequences that were automatically synonymized to *Muntiacus muntjak* on the basis of the available information in the Catalogue of Life were revised back to their original identifications to reflect current taxonomic knowledge.

Database composition

The final database was skewed even more strongly towards *CytB* than was the raw database. It was the most abundant locus for each class and represented >60% of sequences for both Mammalia and Reptilia. In all classes, *16S* made up <10% of the final database, with Reptilia *COI* also at <10%.

Figure 5 shows that most species represented in the curated

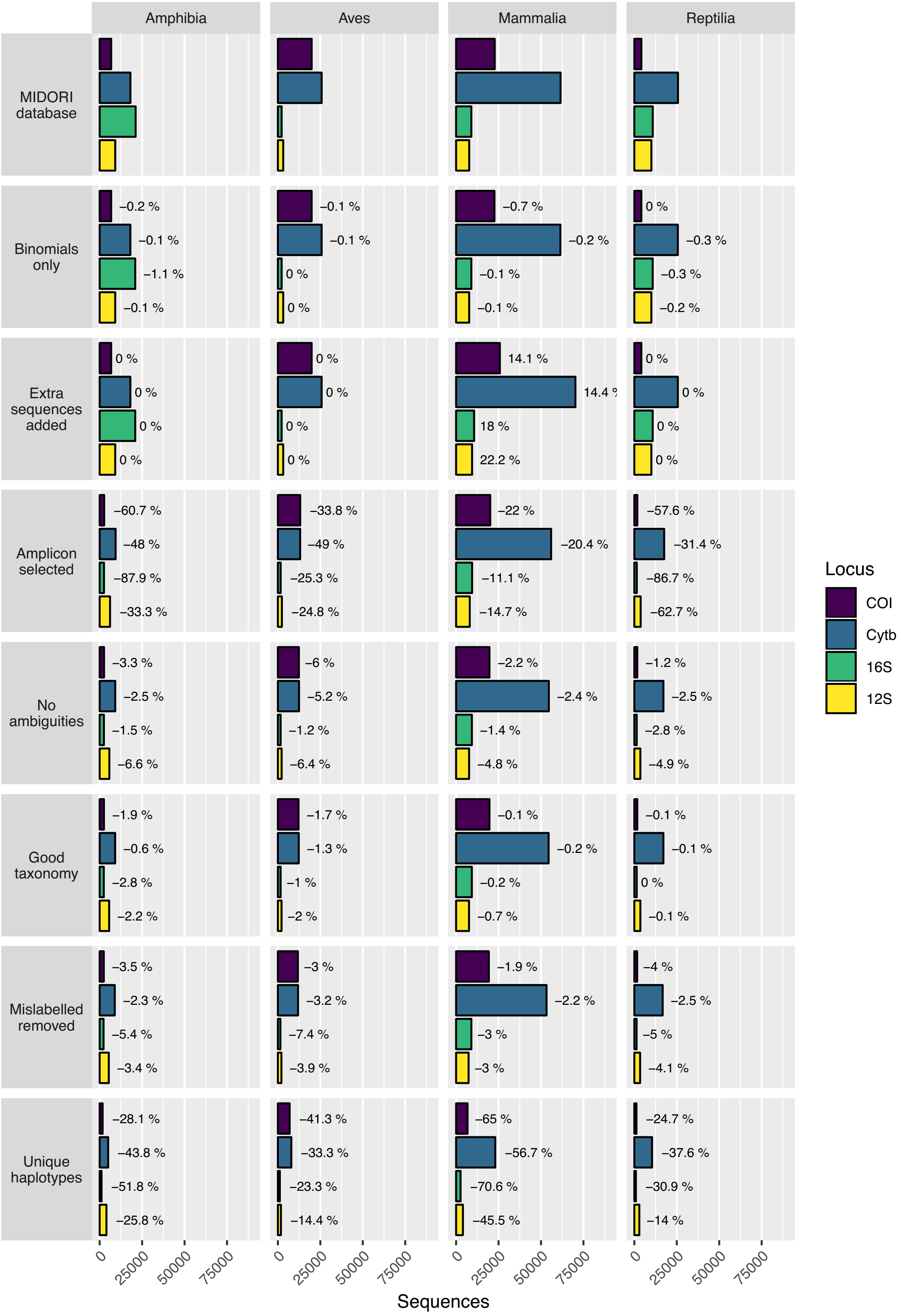

Figure 4: Data availability and percentage loss at each major step in the database curation procedure for each target amplicon and class of Tetrapoda. The number of sequences decreases between steps except "Extra sequences added," where additional target sequences are included for Mammalia and there is no change for the other 3 classes.

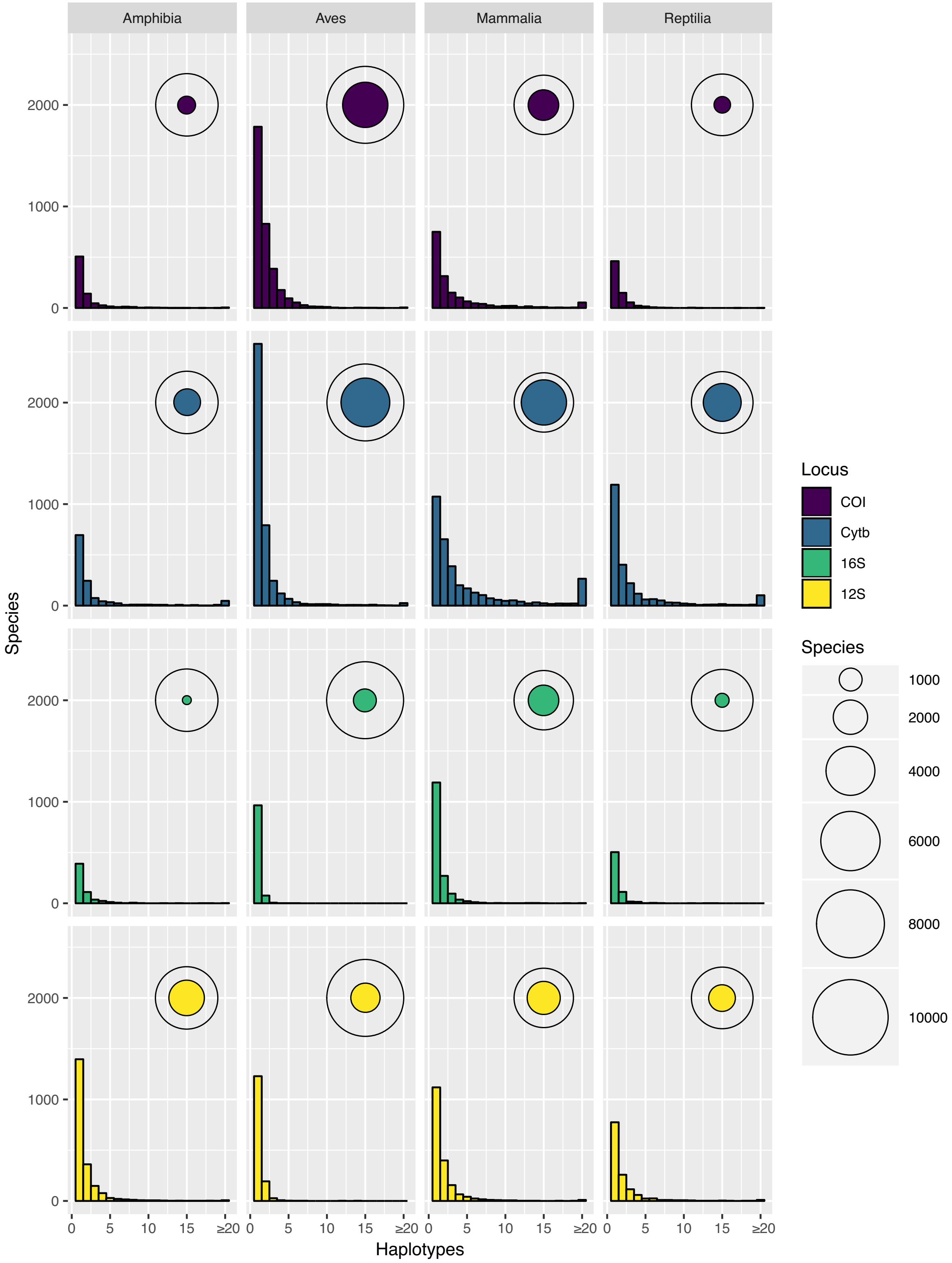

Figure 5: Haplotype number by species (frequency distribution) and the total number of species with ≥1 haplotype, shown relative to the total number of species in the taxonomy for that category (bubbles), shown for each marker and class of Tetrapoda. The proportion of species covered by the database varies between categories, but in all cases a majority of recovered species are represented by a single unique haplotype.

database for any locus have just 1 unique haplotype against which HTS reads can be compared; only a few species have many haplotypes. The prevalence of species with ≥20 haplotypes is particularly notable in *CytB*, where the 4 classes have between 25 (Aves) and 265 (Mammalia) species in this category. The coloured circles in Fig. 5 also show that the species of the taxonomy are incompletely represented across all loci and that coverage varies substantially between taxonomic groups. In spite of global initiatives to generate *COI* sequences [26], this marker does not offer the best species-level coverage in any class and is a poor choice for Amphibia and Reptilia (<15% of species included). Even the best performing marker, *CytB*, is not a universally appropriate choice because Amphibia is better covered by *12S*. These differences in underlying database composition will affect the likelihood of obtaining accurate taxonomic assignment for any 1 species from any single marker. Further barcoding campaigns are clearly needed to fill gaps in the reference databases for all markers and all classes to increase the power of future e/iDNA studies. As the costs of HTS decrease, we expect that such gap-filling will increasingly shift towards sequencing of whole mitochondrial genomes of specimen obtained from museum collections, trapping campaigns, etc. [37], reducing the effect of marker choice on detection likelihood. In the meantime, however, the total number of species covered by the database can be increased by combining multiple loci (here, ≤4) and thus the impacts of database gaps on correctly detecting species can be minimized ([60]; Fig. 6).

In the present study, the primary target for iDNA sampling was the mammal fauna of Malaysian Borneo, and the 103 species expected in the sampling area represent an informative case study highlighting the deficiencies in existing databases (Fig. 7). Nine species are completely unrepresented, while only slightly over half (55 species) have ≥1 sequence for all of the loci. Individually, each marker covers more than half of the target species, but none achieves >85% coverage (*12S*: 75 species; *16S*: 68; *CytB*: 88; *COI*: 66). Equally striking is the lack of within-species diversity because most of the incorporated species are represented by only a single haplotype per locus. Some of the species have large distribution ranges, so it is likely that in some cases the populations on Borneo differ genetically from the available reference sequences, possibly limiting assignment success. Only a few expected species have been sequenced extensively, and most are of economic importance to humans (e.g., *Bos taurus*, *Bubalus bubalis*, *Macaca* spp., *Paradoxurus hermaphroditus*, *Rattus* spp., *Sus scrofa*), with as many as 100 haplotypes available (*Canis lupus*). Other well-represented species (≥20 haplotypes) present in the sampling area include several Muridae (*Chiropodomys gliroides*, *Leopoldamys sabanus*, *Maxomys surifer*, *Maxomys whiteheadi*) and the leopard cat (*Prionailurus bengalensis*).

Laboratory workflow

Shotgun sequencing of a subset of our samples revealed that the median mammalian DNA content was only 0.9% (range, 0%–98%). These estimates are approximate, but with >75% of the samples being <5%, this shows clearly the scarcity of target DNA in bulk iDNA samples. The generally low DNA content and the fact that the target DNA is often degraded make enrichment of the target barcoding loci necessary. We used PCR with high cycle numbers to obtain enough DNA for sequencing. However, this second step increases the risk of PCR error: artificial sequence variation, non-target amplification, and/or raising contaminations up to a detectable level.

We addressed these problems by running 2 extraction replicates, 2 PCR replicates, and a multi-marker approach. The need for PCR replicates has been acknowledged and addressed extensively in ancient DNA studies [16] and has also been highlighted for metabarcoding studies [19, 20, 61, 62]. Despite this, many e/iDNA studies do not carry out multiple PCR replicates to detect and omit potential false sequences. In addition, extraction replicates are seldom applied, despite the evidence that cross-sample DNA contamination can occur during DNA extraction [63–65]. We only accepted sequences that appeared in ≥2 independent PCRs for the lax and for the stringent criterion, where it has to occur in each Extraction Replicate A and B (Fig. 1). The latter acceptance criterion is quite conservative and produces higher false-negative rates than, e.g., accepting occurrence of ≥2 positives. However, it also reduces the risk of accepting false-positive results compared with it (see Supplemental Fig. S1 for a simulation of false-positive and false-negative result rates within a PCR), especially with increasing risk of false-positive occurrence in a PCR, e.g., due to higher risk of contamination. Metabarcoding studies are very prone to false-negative results, and downstream analyses such as occupancy models for species distributions can account for imperfect detection and false-negative results. However, methods for discounting false-positive detections are not well developed [66]. Thus, we think it is more important to avoid false-positive results, especially if the results will be used to make management decisions regarding rare or endangered species. In contrast, it might be acceptable to use a relaxed acceptance criterion for more common species, as long as the ratio of false-positives to true-positives is small and does not affect species distribution estimates. Employing both of our tested criteria researchers could flag unreliable assignments and management decisions can still use this information, but now in a forewarned way. An alternative to our acceptance criteria could be use the PCR replicates themselves to model the detection probability within a sample using an occupancy framework [20, 66–68].

We used 3 different loci to correct for potential PCR amplification biases. We were, however, unable to quantify this bias in this study due to the high degradation of the target mammalian DNA, which resulted in much higher overall amplification rates for *16S*, the shortest of our PCR amplicons. For *16S*, 85% of the samples amplified, whereas for *CytB* and *12S*, only 57% and 44% amplified, respectively. Also the read losses due to trimming and quality filtering were substantially lower for the *16S* sequencing runs (1.3% and 5.3% on average, Supplemental Table S3) compared with the sequencing runs for the longer fragments of *12S* and *CytB* (65.3% and 44.3% on average, Supplemental Table S3). Despite the greater taxonomic resolution of the longer *12S* and *CytB* fragments, our poorer amplification and sequencing results for these longer fragments emphasize that e/iDNA studies should generally focus on short PCR fragments to increase the likelihood of positive amplifications of the degraded target DNA. In the case of mammal-focussed e/iDNA studies, developing a shorter (100 bp) *CytB* fragment would likely be useful.

Another major precaution was the use of twin-tagging for both PCRs (Fig. 2). This ensures that unlabelled PCR products are never produced and allows us to multiplex a large number of samples on a single Illumina MiSeq run. Just 24 sample Tags 1 and 20 plate Tags 2 allow the differentiation of up to 480 samples with matching tags on both ends. The same number of individual primers would have needed longer tags to maintain enough distance between them and would have resulted in an even longer adapter-tag overhang compared with primer length. This would have most likely resulted in lower binding efficiencies as a result of steric hindrances of the primers. Furthermore, this would have resulted in increased primer costs. Thus, our approach reduced sequencing and primer purchase costs while at the same time largely eliminating sample misassignment via tag jumping, because tag-jump sequences have non-matching forward and reverse Tag 1 sequences [51]. We estimated the rate of tag jumps producing non-matching Tag 1 sequences to be 1–

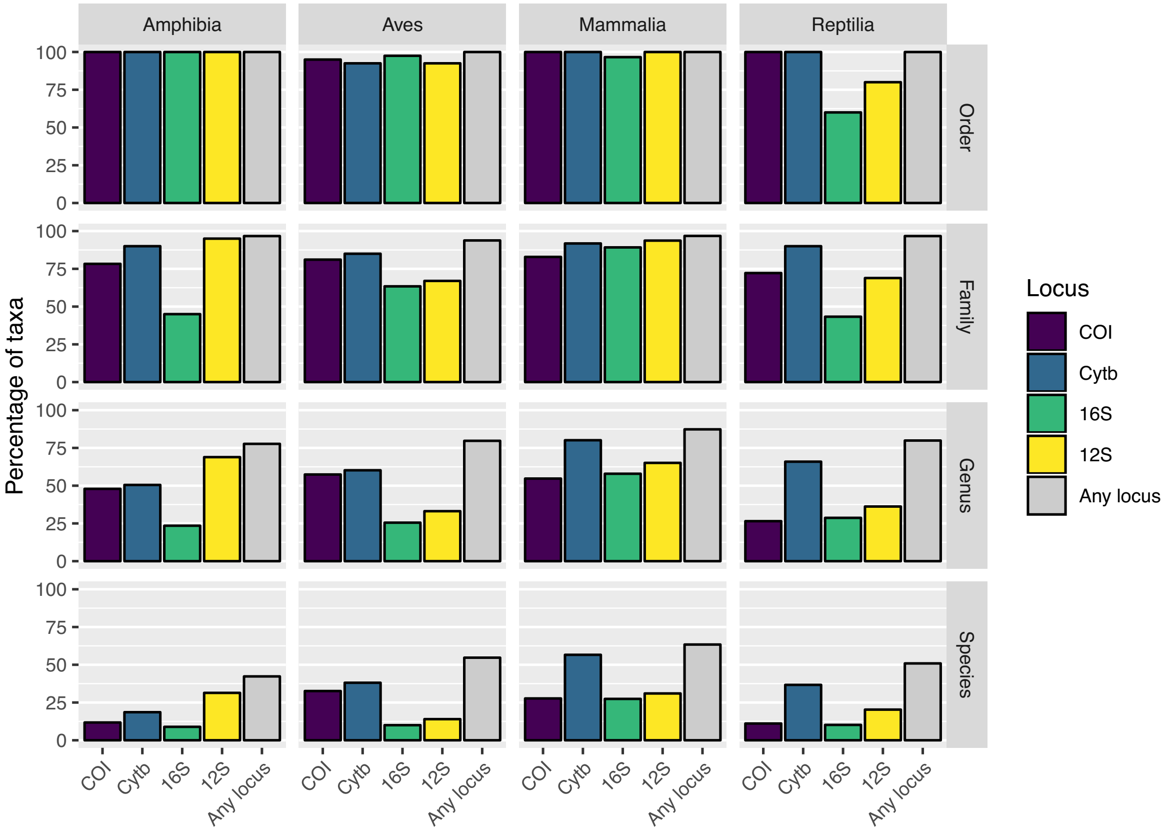

Figure 6: The percentage of the full taxonomy covered by the final database at each taxonomic level for each class of Tetrapoda. Includes the percentage of taxa represented by each marker and all markers combined. In all cases taking all 4 markers together increases the proportion of species, genera, and families covered by the database, but it remains incomplete when compared with the full taxonomy.

5%, and these were removed from the dataset (Table 4). For our sequenced PCR plates, the rate of non-matching Tag 2 tags was 2%. These numbers are smaller than data from Zepeda-Mendoza et al. [62], who reported on sequence losses of 19–23% due to unused tag combinations when they tested their DAMe pipeline on different datasets built using standard blunt-end ligation technique. Although their numbers might not be 1-to-1 comparable to our results because they counted unique sequences and we report on read numbers, our PCR libraries with matching barcodes seem to reduce the risk of tag jumping compared with blunt-end ligation techniques. For the second PCR round, we used the same tag pair Tag 2 for all 24 samples of a PCR plate. To reduce cost we tested pooling these 24 samples prior to the second PCR round, but we detected a very high tag-jumping rate of >40% (Table 4), which ultimately would increase cost through reduced sequencing efficiency. Twin-tagging increases costs because of the need to purchase a larger number of primer pairs, but at the same time it increases confidence in the results.

Tagging primers in the first PCR reduces the risk of cross-contamination via aerosolized PCR products. However, we would not be able to detect a contamination prior to the second PCR from one plate to another because we used the same 24 tags (Tag 1) for all plates. Nevertheless such a contamination is very unlikely to result in any accepted false-positive finding as it would be improbable to contaminate both the A and B replicates, given the exchange of all reagents and the time gap between the PCRs. Previous studies have shown that unlabelled volatile PCR products pose a great risk of false detections [69], a risk that is greatly increased if a high number of samples are analysed in the laboratories [13]. Also, in laboratories where other research projects are conducted, this approach allows the detection of cross-experiment contamination. Therefore, we see a clear advantage of our approach over ligation techniques when it comes to producing sequencing libraries because the Illumina tags are only added after the first PCR, and thus the risk of cross-contamination with unlabelled PCR amplicons is very low.

Assignment results

A robust assignment of species is an important factor in metabarcoding because an incorrect identification might result in incorrect management interventions. The reliability of taxonomic assignments is expected to vary with respect to both marker information content and database completeness, and this is reflected in the probability estimates provided by PROTAX. In a recent study, <10% of the mammal assignments made at species level against a worldwide reference database were considered reliable with the short *16S* amplicon, but this increased to 46% with full-length *16S* sequences [32]. In contrast, in the same study >80% of insect assignments at species level were considered reliable with a more complete, geographically restricted database of full-length *COI* barcodes. A similar pattern was observed in our data during manual curation of the assignment results—there was more ambiguity in the results for the short *16S* amplicon than for other markers. However, due to the limited amount of often degraded target DNA in e/iDNA samples, short amplicons amplify much better. In our case, this had

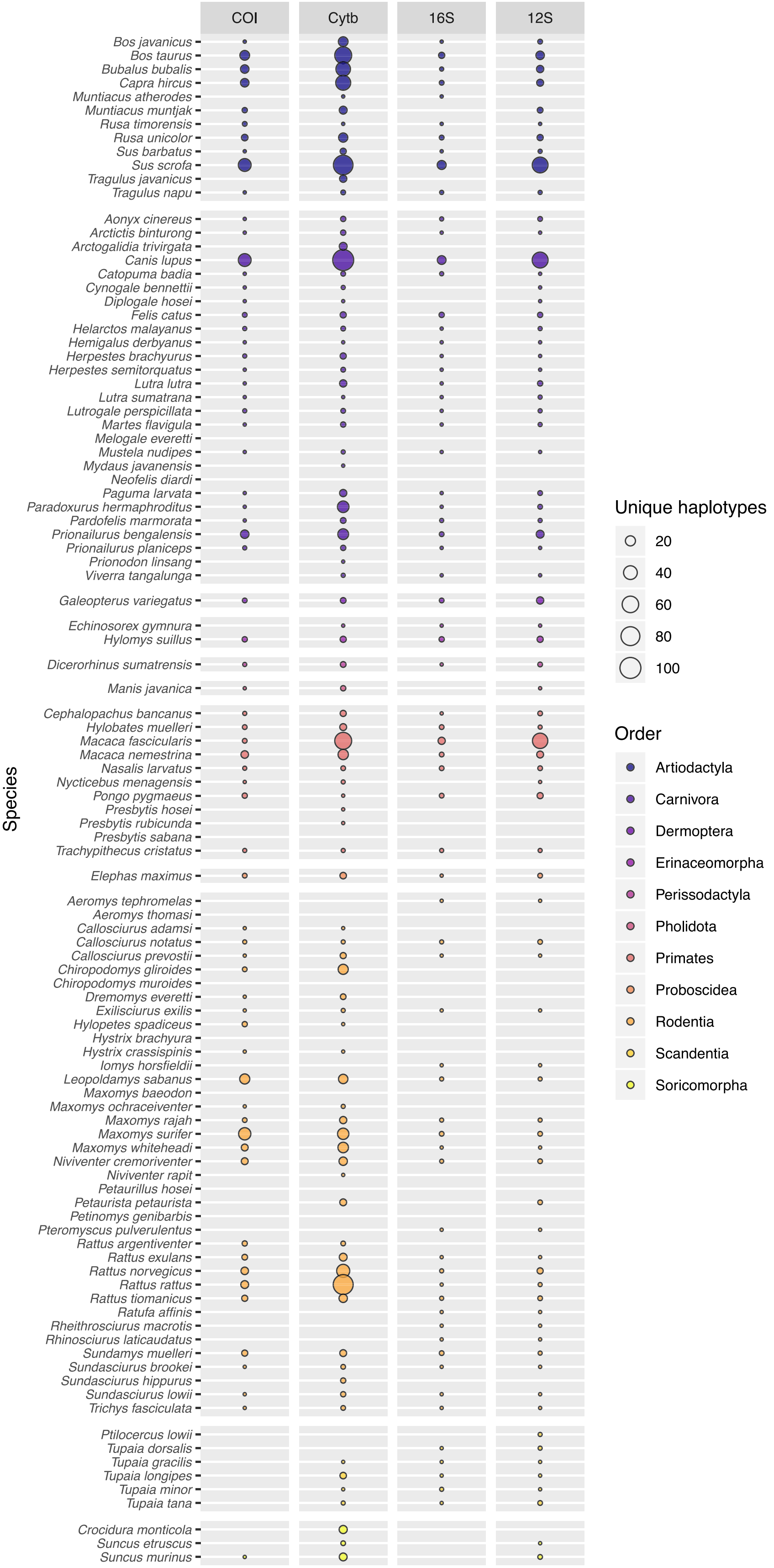

Figure 7: The number of unique haplotypes per marker for each of the 103 mammal species expected in the study area. Bubble size is proportional to the number of haplotypes and varies between 0 and 100. Only 55 species have ≥1 sequence per marker and 9 species are completely unrepresented in the current database.

Table 4: Number of reads per sequencing run and the numbers of reads with matching, non-matching, or unidentifiable tags for 7 of the 8 sequencing runs*

Run	Total reads	Matching Tag 2 reads	Non-matching Tag 2 reads	Matching Tag 1 reads (%)[1]	Non-matching Tag 1 reads	Erroneous Tag 1 reads	Erroneous Tag 1 reads (%)[2]	Reads	Reads (%)[2]
SeqRun01	18,438,517	18,102,702	282,419	1.5	17,514,515	451,028	2.5	137,159	0.8
SeqRun02	25,385,558	24,596,380	626,245	2.5	23,426,084	612,045	2.5	558,251	2.3
SeqRun03	14,875,796	14,393,884	343,528	2.3	13,766,187	426,181	3.0	201,516	1.4
SeqRun04	2,027,794	1,935,149	56,077	2.8	1,806,655	88,307	4.6	40,187	2.1
SeqRun05	18,221,504	17,500,366	421,588	2.3	16,793,851	482,365	2.8	161,458	0.9
SeqRun06	20,718,202	19,874,913	429,048	2.1	19,317,305	371,048	1.9	81,422	0.4
SeqRun07	24,604,610	23,746,938	663,730	2.7	22,446,187	497,366	2.1	803,385	3.4
Total	124,271,981	120,150,332	2,822,635	2.3	115,070,784	2,928,340	2.5	1983,378	1.7
IndexRun	10,276,093	10,116,808	NA	NA	5,841,190	4,186,688	41.4	88,930	0.9

[1] refers to the total number of reads.

[2] refers to the number of reads with matching Tag 2.

*Sequencing run SeqRun08 run contained libraries of another project; thus, we were unable to provide a number of raw reads. NA: not applicable.

the drawback that some species lacked any interspecific variation, and thus sequencing reads shared 99–100% identity for several species. For example, our only *16S* reference of *Sus barbatus* was 100% identical to *S. scrofa*. But because the latter species does not occur in the studied area, we could assign all reads manually to *S. barbatus*. In several cases we were able to confirm *S. barbatus* by additional *CytB* results, highlighting the usefulness of multiple markers.

Another advantage of multiple markers is the opportunity to fill gaps in the reference database. For example, we lacked *16S* reference sequences for *Hystrix brachyura*, and reads were assigned by PROTAX only to the unknown species *Hystrix* sp. In 1 sample, however, almost 5,000 *CytB* reads could be confidently assigned to *H. brachyura*, and thus we used the *Hystrix* sp. *16S* sequences in the same sample to build a consensus *16S* reference sequence for *H. brachyura* for future analyses. In another example we had *CytB* reads assigned to *Mydaus javanicus*, the Sunda stink badger, in 1 sample but *12S* reads assigned to *Mydaus* sp. in another one. Because we lacked a *12S Mydaus* reference and because there is only 1 *Mydaus* species on Borneo, we could assume that this second sample is most likely also *M. javanicus*.

We also inferred that PCR and sequencing errors resulted in reads being assigned to sister taxa. We observed that a high number of reads of a true sequence were assigned to a species and a lower number of noise sequences were assigned to a sister taxon. Such a pattern was observed for ungulates, especially deer, which showed little variance in *16S*. It is hard to identify and control for such a pattern automatically, and this highlights the importance of visual inspection of the results.

For the more lax criterion (2 positive PCR replicates) we accepted 190 species assignments out of 109 leech samples. Under the stringent criterion (i.e., having positive detections in both Extraction Replicates A and B) we accepted $\sim$14% fewer assignments: in total 162 vertebrate detections within 95 bulk samples (Table 5). For 48% of the species frequencies did not change and almost half of the not accepted assignments were from the most frequent species *Rusa unicolor* and *S. barbatus*. However, with the more stringent criterion we did not accept 2 species (1× *Macaca fascicularis* and 2× *Mydaus javanensis*). In 3 cases the stringent criterion would not accept assignments that could be made only to unknown species (*Macaca* sp.) (Table 5). For this genus we have 2 occurring species in the area. Because the true occurrence of species within our leeches was unknown, we can-

Table 5: Number of accepted species assignments with 2 different acceptance criteria: the more stringent criterion accepting only assignments occurring in both extraction replicates (A and B) and the more lax criterion accepting assignments with $\geq$2 positive results in any of the 12 PCR replicates

Species	Stringent	Lax	Change
Aonyx cinereus	1	1	0
Arctictis binturong	1	1	0
Bos javanicus	9	11	+2
Echinosorex gymnura	5	6	+1
Felis catus	2	2	0
Helarctos malayanus	5	6	+1
Hemigalus derbyanus	3	3	0
Hystrix brachyura	4	5	+1
Kalophrynus pleurostigma	1	1	0
Macaca fascicularis		1	+1
Macaca nemestrina	1	2	+1
Macaca sp.		3	+3
Manis javanica	2	2	0
Muntiacus atherodes	6	6	0
Muntiacus muntjak	2	2	0
Muntiacus sp.	10	10	0
Mydaus javanensis		2	+2
Pongo pygmaeus	5	5	0
Rusa unicolor	59	67	+8
Sus barbatus	17	22	+5
Tragulus javanicus	4	6	+2
Tragulus napu	10	11	+1
Trichys fasciculata	5	5	0
Viverra tangalunga	11	11	0
Total accepted assignments	162	190	+28

not evaluate how many of the additional 27 detections in the lax criterion are false-positive results and how many might be false-negative results for the stricter criterion. However, by accepting only positive AB assignment results, we increase the confidence of species detection, even if the total number of reads for that species was low. When it comes to rare or threated species this outweighs the risk of reporting false-positive results in our opinion. A total of 48% of the assignments with the stringent criterion were present in all 4 A1, A2, B1, and B2. A total of 35% were present in ≥3 replicates (e.g., A1, A2, B1).

The mean number of reads per sample used for the taxomomic assignment varied from 162,487 16S reads for SeqRun01 to 7,638 *CytB* reads for SeqRun05 (Supplemental Table S4). In almost all cases, however, the number of reads of an accepted assignment was high (median = 52,386; mean = 300,996; SD = 326,883). PCR stochasticity, primer biases, multiple species in individual samples, and pooling of samples exert too many uncertainties that could bias the sequencing results [70, 71]. Thus, we do not believe that raw read numbers are the most reliable indicators of tetrapod DNA quantity in iDNA samples. Replication of detection is inherently more reliable. In contrast to our expectation that higher cycle number might be necessary to amplify even the lowest amounts of target DNA, our data do not support this hypothesis. Although we observed an increase in positive PCRs for A2/B2 (the 40-cycle PCR replicates), the total number of accepted assignments in A1/B1 and A2/B2 samples did not differ. This indicates first that high PCR cycle numbers mainly increased the risk of false-positive results and second that our multiple precautions successfully minimized the acceptance of false detections.

Conclusion

Metabarcoding of e/iDNA samples will certainly become a very valuable tool in assessing biodiversity because it allows species to be detected non-invasively without the need to capture and handle the animals [72] and because sampling effort can often be greatly reduced. However, the technical and analytical challenges linked to sample types (low quantity and quality of DNA) and poor reference databases have so far been insufficiently recognized. In contrast to ancient DNA studies for which standardized laboratory procedures and specialized bioinformatics pipelines have been established and are followed in most cases, there is limited methodological consensus in e/iDNA studies, which reduces rigour. In this study, we present a robust metabarcoding workflow for e/iDNA studies. We hope that the provided scripts and protocols facilitate further technical and analytical developments. The use of e/iDNA metabarcoding to study the rarest and most endangered species such as the Saola is exciting, but geneticists bear the heavy responsibility of providing correct answers to conservationists.

Availability of supporting data and materials

Project Name: screenforbio-mbc
Project home page: https://github.com/alexcrampton-platt/screenforbio-mbc [33]
Operatin systems: Pipeline was tested o Mac OSX (10.13) and Scientific Linux release 6.9 (Carbon) and Ubuntu Server 18.04 LTS
Programming language: Bash, R
Sequencing data are available in the European Bioinformatics Institute via bioproject No. PRJEB27367. All other supporting data are also available via the *GigaScience* GigaDB repository [73].

Additional files

Supplemental Figure 1: The rates of accepted false-negative results (upper graph) and false-positive results (lower graph) for both our used acceptance criteria for varying PCR detection probabilities. The red line always denotes the stringent acceptance criterion that a positive is only accepted if it is present in ≥1 A and 1 B replicate. The lax criterion (blue) accepted at any 2 positives out of the 12 replicates. The stringent criterion poses a higher risk of accepting a false-negative result, but it reduces clearly the risk of false-positive results, especially with increasing detection probability due to higher risk of contamination.
Supplemental table 1: Complete list of all used primer sequences in 5′-3′ direction.
Supplemental table 2: List of Bornean species that were weighted in the PROTAX assignment.
Supplemental table 3: Summary of the read losses of each sample during the read processing steps for each sequencing run seperately. The first line gives the raw read number per sample. The losses are given as percentage of each step; 1. merging of the R1/R2 reads of the Illumina sequencing done by *usearch* [51, 50], 2. clipping of primers and trimming of reads using *cutadapt* [52], 3. quality filtering and 4. dereplication, both using usearch.
Supplemental table 4: Number of merged R1/R2 reads per sample that were used for the taxonomic assignment for each of the 8 sequencing runs. Displayed are the median, minimum, maximum read numbers per PCR replicate, the mean and its standard deviation as well as the number of PCR replicates with <500 reads.

Abbreviations

12S: 12S rRNA; 16S: 16S rRNA; amplicon: amplification of a short target sequence; bp: base pair; COI: cytochrome c oxidase subunit I amplicon; CytB: cytochrome b; eDNA: environmental DNA; HTS: high-throughput DNA sequencing; iDNA: invertebrate-derived DNA; INSDC: International Nucleotide Sequence Database Collaboration; ITIS: Integrated Taxonomic Information System; Leibniz-IZW: Leibniz Institute for Zoo and Wildlife Research; MEGAN: Metagenome Analyzer; NCBI: National Center for Biotechnology Information; PCR: polymerase chain reaction; rRNA: ribosomal RNA.

Competing interests

The authors declare that they have no competing interests.

Funding

All authors thank the German Federal Ministry of Education and Research (BMBF FKZ: 01LN1301A) and the Leibniz-IZW for funding this study. C.C.Y.X. was also supported by the MEME Erasmus Mundus Programme in Evolutionary Biology, and by the Groningen University Fund and the Marco Polo Fund from the University of Groningen. D.W.Y. and C.C.Y.X. were supported by the National Natural Science Foundation of China (41661144002, 31670536, 31400470, 31500305), the Key Research Program of Frontier Sciences, CAS (QYZDY-SSW-SMC024), the Bureau of International Cooperation project (GJHZ1754), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA20050202, XDB31000000), the Ministry of Science and Technology of China (2012FY110800), and the State Key Laboratory of Genetic Resources and Evolution at the Kunming Institute of Zoology.

Authors' contributions

A.W. conceived the project. J.A. developed the laboratory workflow. A.C.P. developed bioinformatics pipeline. A.C.P., C.C.Y.X., D.W.Y implement PROTAX into the bioinformatics pipeline. A.M. collected samples and provided taxonomic expertise. J.A., L.A.H. did the laboratory work. J.A., A.C.P. processed bioinformatics and analysed data. J.A., L.A.H., A.W. analysed species assignments. J.A., A.C.P., D.W.Y, A.W. wrote the article. All authors discussed the project and data. All authors read and approved the final manuscript.

Acknowledgements

We thank the Sabah Forestry Department, especially Johnny Kissing, Peter Lagan, and Datuk Sam Mannan, for supporting the fieldwork and the Sabah Biodiversity Council for providing research, collection, and export permits for this work. We are grateful to John Mathai and Seth Timothy Wong for conducting the fieldwork and collecting the leeches. We are also grateful to Jörns Fickel, head of the Department of Evolutionary Genetics of the Leibniz-IZW, for continuous support and collaboration. Furthermore, we would like to thank Sebastian Wieser for laboratory support and Dorina Lenz and Anke Schmidt for their help and fruitful discussions.

References

1. Gariepy TD, Lindsay R, Odgen N, et al. Identifying the last supper: utility of the DNA barcode library for bloodmeal identification in ticks. Mol Ecol Res 2012;**12**:646–52.
2. Lee P-S, Gan HM, Clements GR, et al. Field calibration of blowfly-derived DNA against traditional methods for assessing mammal diversity in tropical forests. Genome 2016;**59**:1008–22.
3. Calvignac-Spencer S, Merkel K, Kutzner N, et al. Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity. Mol Ecol 2013;**22**:915–24.
4. Rodgers TW, Xu CCY, Giacalcone J, et al. Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: evidence from a known tropical mammal community. Mol Ecol Res 2017;**17**(6):1–13.
5. Hoffmann C, Merkel K, Sachse A, et al. Blow flies as urban wildlife sensors. Mol Ecol Res 2018;**18**(3):502–10.
6. Schönberger AC, Wagner S, Tuten HC, et al. Host preferences in host-seeking and 632 blood-fed mosquitoes in Switzerland. Med Vet Entomol 2015;**30**(1):39–52.
7. Townzen JS, Brower AVZ, Judd DD. Identification of mosquito bloodmeals using mitochondrial cytochrome oxidase subunit I and cytochrome b gene sequences. Med Vet Entomol 2008;**22**:386–93.
8. Kocher A, Thoisy B, Catzeflies F, et al. iDNA screening: disease vectors as vertebrate samplers. Mol Ecol 2017;**26**(22):6478–86.
9. Taylor L, Cummings RF, Velten R, et al. Host (avian) biting preference of southern California Culex mosquitoes (Diptera: Culicidae). J Med Entomol 2012;**49**(3):687–96.
10. Schnell IB, Thomsen PF, Wilkinson N, et al. Screening mammal biodiversity using DNA from leeches. Curr Biol 2012;**22**(8):R262–3.
11. Tessler M, Weiskopf SR, Berniker L, et al. Bloodlines: mammals, leeches, and conservation in southern Asia. Syst Biodivers 2018;**16**:1–9.
12. Weiskopf SR, McCarthy KP, Tessler M, et al. Using terrestrial haematophagous leeches to enhance tropical biodiversity monitoring programmes in Bangladesh. J Appl Ecol 2018;**55**(4):2071–81.
13. Schnell IB, Bohmann K, Schultze SE, et al. Debugging diversity - a pan-continental exploration of the potential of terrestrial blood-feeding leeches as a vertebrate monitoring tool. Mol Ecol Res 2018;**18**:1282–98.
14. Calvignac-Spencer S, Leendertz FH, Gilbert MT, et al. An invertebrate stomach's view on vertebrate ecology: certain invertebrates could be used as "vertebrate samplers" and deliver DNA-based information on many aspects of vertebrate ecology. BioEssays 2013;**35**(11):1004–13.
15. Schnell IB, Sollmann R, Calvignac-Spencer S, et al. iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool – prospects, pitfalls and avenues to be developed. Front Zool 2015;**12**:24.
16. Pääbo S, Poinar H, Serre D, et al. Genetic analyses from ancient DNA. Annu Rev Genet 2004;**38**:645–79.
17. Hofreiter M, Paijmans JL, Goodchild H, et al. The future of ancient DNA: technical advances and conceptual shifts. BioEssays 2015;**37**(3):284–93.
18. Cristescu ME, Hebert PDN. Uses and misuses of environmental DNA in biodiversity science and conservation. Annu Rev Ecol Evol Syst 2018;**49**:209–30.
19. Ficetola GF, Pansu J, Bonin A, et al. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Mol Ecol Res 2014;**15**(3):543–56.
20. Ficetola GF, Taberlet P, Coissac E. How to limit false positives in environmental DNA and metabarcoding? Mol Ecol Res 2016;**16**(3):604–7.
21. Dalton R. Still looking for that woodpecker. Nature 2010;**463**:718–9.
22. Bazinet AL, Cummings MP. A comparative evaluation of sequence classification programs. BMC Bioinformatics 2012;**13**(1):92.
23. Richardson RT, Bengtsson-Palme J, Johnson RM. Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Mol Ecol Res 2017;**17**(4):760–9.
24. Somervuo P, Koskela S, Pennanen J, et al. Unbiased probabilistic taxonomic classification for DNA barcoding. Bioinformatics 2016;**32**(19):2920–7.
25. Quast C, Pruesse E, Gerken J, et al. SILVA databases. In: Nelson KE , ed. Encyclopedia of Metagenomics. Boston: Springer; 2015:626–35.
26. Ratnasingham S, Hebert PDN. BOLD: the barcode of life data system. Mol Ecol Notes. 2007;**3**:355–64.
27. Nilsson RH, Ryberg M. Taxonomic reliability of DNA sequences in public sequence databases: a fungal persepctive. PLoS One 2006;**1**:e59.
28. Forster P. To err is human. Ann Hum Gen 2003;**67**:2–4.
29. Harris JD. Can you bank on GenBank? Trends Ecol Evol 2003;**18**:317–9.
30. Kozlov AM, Zhang J, Yilmaz P, et al. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic Acids Res 2016;**44**(11):5022–33.
31. Huson DH, Auch AF, Qi J, et al. MEGAN analysis of metagenomic data. Genome Res 2007;**17**(3):377–86.
32. Somervuo P, Yu DW, Xu CC, et al. Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. Methods Ecol Evol 2017;**8**(4):398–407.
33. Scripts for the ScreenForBio metabarcoding pipeline. https://github.com/alexcrampton-platt/screenforbio-mbc, Ac-

cessed 19th Mar 2019.
34. Machida RJ, Leray M, Ho SL, et al. Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Sci Data 2017;**4**:170027.
35. Midori Reference. http://www.reference-midori.info/download.php#, Accessed 1st Feb 2019
36. Chamberlain SA, Szöcs E. Taxize: taxonomic search and retrieval in R. Version 2. F1000Res 2013;**2**:191.
37. Salleh FM, Ramos-Madrigal J, Peñaloza F, et al. An expanded mammal mitogenome dataset from Southeast Asia. GigaScience 2017;**6**(8):1–8.
38. Kans J. Entrez Direct: E-utilities on the UNIX Command Line. Bethesda, MD: National Center for Biotechnology Information (US). 2010, https://www.ncbi.nlm.nih.gov/books/NBK179288/, Accessed 19 Mar 2019.
39. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;**215**(3):403–10.
40. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;**10**(1):421.
41. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;**30**(4):772–80.
42. Katoh K, Misawa K, Kuma KI, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;**30**(14):3059–66.
43. Chesters D. collapsetypes.pl. 2013. http://sourceforge.net/projects/collapsetypes/, Accessed 1 February 2019.
44. Kocher TD, Thomas WK, Meyer A, et al. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc Natl Acad Sci U S A 1989;**86**(16):6196–200.
45. Taylor PG. Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. Mol Biol Evol 1996;**13**(1):283–5.
46. Bioinformatics toolkits for manipulating sequence, alignment, and phylogenetic tree files. https://github.com/biologyguy/BuddySuite, Accessed 1st Feb 2019
47. A cross-platform and ultrafast toolkit for FASTA/Q file manipulation in Golang. https://github.com/shenwei356/seqkit. , Accessed 1st Feb 2019
48. Toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk, Accessed 1st Feb 2019
49. Toolkit for processing TAB-delimited format. https://github.com/lh3/tabtk, Accessed 1st Feb 2019
50. Faircloth BC, Glenn TC. Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. PloS One 2012;**7**(8):e42543.
51. Schnell IB, Bohmann K, Gilbert MTP. Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. Mol Ecol Res 2015;**15**(6):1289–303.
52. Murray DC, Coghlan ML, Bunce M. From benchtop to desktop: important considerations when designing amplicon sequencing workflows. PLoS One 2015;**10**(4):e0124671.
53. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes 2016;**9**:88.
54. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;**26**(19):2460–1.
55. Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 2015;**31**(21):3476–82.
56. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 2011;**17**(1):10–12.
57. Kielbasa SM, Wan R, Sato K, et al. Adaptive seeds tame genomic sequence comparison. Genome Res 2011;**21**(3):487–93.
58. Fortes GG, Paijmans JLA. Analysis of whole mitogenomes from ancient samples. In: Kroneis T , ed. Whole Genome Amplification: Methods and Protocols. New York: Springer; 2015:179–95.
59. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 2011;**12**:385.
60. Evans NT, Li Y, Renshaw MA, et al. Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. Can J Fish Aquat Sci 2017;**74**(9):1362–74.
61. Bonin A, Taberlet P, Zinger L, et al. Environmental DNA: For Biodiversity Research and Monitoring. 1st ed. Oxford University Press; 2018.
62. Zepeda-Mendoza ML, Bohmann K, Baez A, et al. DAMe: a toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. BMC Res Notes 2016;**9**:255.
63. Racimo F, Renaud G, Slatkin M. Joint estimation of contamination, error and demography for nuclear DNA from ancient humans. PLoS Genet 2016;**12**(4):e1005972.
64. Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and epigenomes. Nat Rev Genet 2015;**16**(7):395.
65. Laurin-Lemay S, Brinkmann H, Philippe H. Origin of land plants revisited in the light of sequence contamination and missing data. Curr Biol 2012;**22**(15):R593–4.
66. Lahoz-Monfort JJ, Guillera-Arroita G, Tingley R. Statistical approaches to account for false-positive errors in environmental DNA samples. Mol Ecol Res 2015;**16**:673–85.
67. Dorazio RM, Erickson RA. Ednaoccupancy: anR package for multiscale occupancy modelling of environmental DNA data. Mol Ecol Res 2018;**18**:368–80.
68. Guillera-Arriota G, Lahoz-Monfort JJ, van Rooyen AR, et al. Dealing with false-positives and false-negative errors about species occurrence at multiple levels. Methods Ecol Evol 2017;**8**:1081–91.
69. Kwok S, Higuchi R. Avoiding false positives with PCR. Nature 1989;**339**:237–8.
70. Bush A, Sollmann R, Wilting A, et al. Connecting Earth observation to high-throughput biodiversity data. Nat Ecol Evol 2017;**1**(7):0176.
71. Piñol J, Mir G, Gomez-Polo P, et al. Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Mol Ecol Res 2014;**15**:819–30.
72. Nichols RV, Vollmers C, Newsom L, et al. Minimizing polymerase biases in metabarcoding. Mol Ecol Res 2018;**18**:927–39.
73. Axtner J, Crampton-Platt A, Hörig L, et al. Supporting data for "An efficient and improved laboratory workflow and tetrapod database for larger scale environmental DNA studies." GigaScience Database 2019. http://dx.doi.org/10.5524/100570.