A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies

Emmanuel Corse, Emese Meglecz, Gait Archambaud, Morgane Ardisson, Jean-François Martin, Christelle Tougard, Rémi Chappaz, Vincent Dubut

A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies

Emmanuel Corse[1] | Emese Meglécz[1] | Gaït Archambaud[2] | Morgane Ardisson[3] | Jean-François Martin[4] | Christelle Tougard[5] | Rémi Chappaz[1] | Vincent Dubut[1]

[1]Aix Marseille Univ, Avignon Univ, CNRS, IRD, UMR IMBE, Marseille, France

[2]Irstea, UR RECOVER, Equipe FRESCHCO, Aix-en-Provence, France

[3]INRA, CIRAD, Montpellier SupAgro, UMR AGAP, Montpellier, France

[4]Montpellier SupAgro, INRA, CIRAD, IRD, UMR CBGP, Montferrier-sur-Lez, France

[5]CNRS, Université de Montpellier, IRD, CIRAD, EPHE, UMR ISEM, Montpellier, France

Correspondence
Emmanuel Corse or Vincent Dubut, Aix Marseille Univ, Avignon Univ, CNRS, IRD, UMR IMBE, Marseille, France.
Emails: emmanuel.corse@gmail.com; vincent.dubut@imbe.fr

Funding information
Électricité de France; Syndicat Mixte d'Aménagement du Val Durance (France); Office National de l'Eau et des Milieux Aquatiques (France); Agence de l'Eau Rhône-Méditerranée (France); Conseil Régional de Provence-Alpes-Côte d'Azur (France)

Abstract

The main objective of this work was to develop and validate a robust and reliable "from-benchtop-to-desktop" metabarcoding workflow to investigate the diet of invertebrate-eaters. We applied our workflow to faecal DNA samples of an invertebrate-eating fish species. A fragment of the cytochrome *c* oxidase I (COI) gene was amplified by combining two minibarcoding primer sets to maximize the taxonomic coverage. Amplicons were sequenced by an Illumina MiSeq platform. We developed a filtering approach based on a series of nonarbitrary thresholds established from control samples and from molecular replicates to address the elimination of cross-contamination, PCR/sequencing errors and mistagging artefacts. This resulted in a conservative and informative metabarcoding data set. We developed a taxonomic assignment procedure that combines different approaches and that allowed the identification of ~75% of invertebrate COI variants to the species level. Moreover, based on the diversity of the variants, we introduced a semiquantitative statistic in our diet study, the minimum number of individuals, which is based on the number of distinct variants in each sample. The metabarcoding approach described in this article may guide future diet studies that aim to produce robust data sets associated with a fine and accurate identification of prey items.

KEYWORDS
cytochrome *c* oxidase I, diet studies, HTS data filtering, metabarcoding, taxonomic assignment

1 | INTRODUCTION

In ecology and conservation, reliable diet data sets are critical to the understanding of prey/habitat relationships and feeding habitats. In this perspective, the use of DNA-based approaches in trophic ecology has grown during the last years (e.g., Razgour et al., 2011; Soininen et al., 2015), especially due to the advent of high-throughput sequencing (HTS), leading to the development of metabarcoding (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). However, diet metabarcoding approaches face four main challenges: (i) amplification bias related to the degradation of DNA (Sint, Raso, Kaufmann, & Traugott, 2011); (ii) taxonomic coverage of primers (Gibson et al., 2014); (iii) taxonomic identification and resolution of DNA barcode sequences (Richardson, Bengtsson-Palme, & Johnson, 2017); and (iv) filtering of HTS data to eliminate artefacts (e.g., PCR/sequencing errors, mistagging, contamination) that produce false positives and constitute low-frequency noise (LFN; sensu De Barba et al., 2014). In diet studies, metabarcoding primers should therefore target short regions (i.e., <300 bp) of multicopy DNA to tackle the degradation of DNA (Pompanon et al., 2012). Moreover, binding sites of primers should be sufficiently conserved to minimize biases in taxonomic coverage (Clarke, Soubrier, Weyrich, & Cooper, 2014;

Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). Alternatively, taxonomic coverage can be improved by amplifying several loci (De Barba et al., 2014) or using several sets of primers that target the same locus (Gibson et al., 2014). In both cases, the choice of the PCR primers is critical for optimizing detection of all prey in faeces. As for the taxonomic identification, a considerable trade-off between accuracy and sensitivity should be considered when selecting an assignment procedure (Richardson et al., 2017). Moreover, both the confidence and the resolution of taxonomic classifiers are highly dependent on the richness of reference sequence databases of the targeted loci (Gibson et al., 2014; Porter et al., 2014). Additionally, producing robust data sets is critical for conducting reliable ecological studies. However, most of the current clustering-based methods for filtering HTS need specific (and partly arbitrary) parameterization and often overestimate the real number of taxa in samples (Brown, Chain, Crease, MacIsaac, & Cristescu, 2015; Clare, Chain, Littlefair, & Cristescu, 2016; Flynn, Brown, Chain, MacIsaac, & Cristescu, 2015). In fact, even if clustering reads into operational taxonomic units can partially address the overestimation of taxa, clustering-based methods do not account for some LFNs, such as cross-sample contamination and mistagging.

In this context, and following the recommendations by Murray, Coghlan, and Bunce (2015), we developed a "from-benchtop-to-desktop" metabarcoding workflow to investigate the diet of invertebrate-eaters. We particularly focused on two main objectives. Our first objective was data reliability, which involves both minimizing false-negatives and false-positives. This was achieved by ensuring the efficiency of the combined use of two primer sets, by performing several PCR replicates and by developing a clustering-free filtering method based on control samples and nonarbitrary thresholds. Second, we developed an approach that should maximize the taxonomic resolution of molecular identification of prey by combining four assignment procedures and three reference databases. Our workflow was applied to the biodiversity assessment of prey ingested by *Zingel asper* (Linnaeus, 1758), a critically endangered benthic freshwater fish. Using this model, we demonstrated the application of our workflow on faecal samples and illustrated its interest for a better characterization of feeding habitats.

2 | MATERIAL AND METHODS

2.1 | Faecal sample collection

Thirty-five *Z. asper* specimens from which faeces could be collected were sampled on 5 September 2014 in the Durance River (France: 44°20′14″N, 5°54′46″E). Fishes were caught by electrofishing and their abdomen was squeezed by hand in order to drain their faeces. Collected faeces were stored in a 1.5-ml vial containing 96% ethanol. After sampling, individuals were immediately released within the fishing area. Faeces were stored at −20°C until DNA extraction. Additionally, five faecal samples from a fish species living in brackish-water habitats, *Pomatoschistus microps* (Krøyer, 1838), were also analysed to assess the versatility of the workflow and to control for mistagging in our HTS data set (see below). These individuals were sampled in the Vaccarès Lagoon (French Mediterranean coast: 44°20′14″N, 5°54′46″E).

2.2 | Faecal DNA extraction and controls

All faecal DNA extraction steps were conducted in a room dedicated to the handling of degraded DNA ("Plateforme ADN Dégradé" of the Institut des Sciences de l'Evolution de Montpellier, France) and following the specific safety measures described by Monti et al. (2015). Before DNA extraction, faeces were dried using the Eppendorf Concentrator Plus (Eppendorf, Germany). One volume of dried faeces, one volume of zirconium oxide beads (0.5 mm) and ½ volume of sterile water were mixed to crush samples using a Bullet Blender (Next Advance, USA). The DNeasy® *mericon* Food Kit (QIAGEN, Germany) was used to extract DNA from faecal samples to minimize the level of co-extracted products and improve PCR success (Zarzoso-Lacoste, Corse, & Vidal, 2013). Each extraction series included (i) 23 faecal samples, (ii) a negative control for extraction (T_{ext}) that consisted of 50 μl of DNA-free water subjected to DNA extraction protocol and (iii) a negative control for DNA aerosols (T_{pai}) that consisted of a 1.5-ml vial containing 50 μl of DNA-free water that remained open but otherwise untouched during the extraction protocol. DNA concentrations were quantified using a Qubit Fluorometer (Invitrogen, Darmstadt, Germany) and standardized to max. 20 ng/μl.

2.3 | Local DNA library construction

A local (noncomprehensive) DNA library (lcDNA samples) was constructed using invertebrates sampled in the Durance River and invertebrate samples from our laboratory collections. The sample composition and laboratory protocols are detailed in Table S1 and Appendix S1, respectively. We successfully sequenced 301 samples, representing 209 distinct species.

2.4 | Minibarcoding protocol and taxonomic coverage

Two DNA primer pairs were initially selected. They both amplify a short fragment from the 5′ end of the mitochondrial cytochrome *c* oxidase I (COI): ZBJ-ArtF1c and ZBJ-ArtR2c (Arthropods "universal"; Zeale, Butlin, Barker, Lees, & Jones, 2011), hereafter abbreviated as ZF and ZR; and Uni-MinibarF1 and Uni-MinibarR1 (Eukaryotes "universal"; Meusnier et al., 2008), hereafter abbreviated as MF and MR. All four pairwise combinations of the primers were tested for PCR success on lcDNA samples, and only MFZR and ZFZR were selected as a result (see Appendix S2). These two primer pairs produce overlapping amplicons (from ~210 to ~230 bp including primers) with MFZR amplicons being 18 bp longer in the 5′ region. The COI region of the predator (*Z. asper*) was not amplified using either ZFZR or MFZR; however, the MFZR primer pair amplified an unspecific fragment (~850 bp). Therefore, a blocking primer (BlNupRan) was developed to inhibit the amplification of the nontargeted locus (see

Appendix S2). The taxonomic coverage of both primer sets was assessed by in vitro tests using the lcDNA samples and in silico using *EcoPCR*, *EcoTaxStat* and *EcoTaxSpecificity* (Ficetola et al., 2010) (see Appendix S2).

2.5 | PCR-based enrichment of prey DNA and high-throughput sequencing

To track amplicons back to the samples and to avoid flashes of light during HTS, 12–14 nucleotide-long sequence tags were added onto the 5′ end of each primer (eight distinct tags for the ZF and MF primers, and 12 tags for ZR), creating 96 forward and reverse potential tag combinations (Table S2). Three PCR replicates were conducted in a volume of 25 μl for each minibarcode primer pair, resulting in six independent PCR enrichments per sample. Template DNA consisted of 2.5 μl of faecal standardized DNA extracts. For the primer pair MFZR, blocking primer BlNupRan was added in the PCR mix at 400 nM. Additionally, several control samples were subjected to the PCR enrichment step: (i) T_{ext} and T_{pai} (see above) are controls for cross-sample or exogenous contaminations during the DNA extractions; (ii) T_{PCR} indicates the level of cross-contamination during the preparation of the PCR mix and plates (tagged primers but no DNA template in the PCR vial); (iii) one negative control (T_{tag}) was included to assess the level of mistagging due to the recombination of sequences from different samples (see Schnell, Bohmann, & Gilbert, 2015). The T_{tag} consisted of an empty vial on the PCR plate that did not contain any tagged primers or DNA. This creates an extra tag combination that is not used for any samples or controls (as in: Esling, Lejzerowicz, & Pawlowski, 2015); (iv) finally, two mock faecal samples were also amplified (T_{pos1} and T_{pos2}) and they consisted of identical artificial mixes of the DNA obtained from six potential prey (*Caenis pusilla*, *Baetis rhodani*, *Orthocladiinae* sp., *Chironomidae* sp., *Hydropsyche pellucidula*, *Phoxinus* cf. *phoxinus*) and from *Z. asper*. The DNA concentration of T_{pos1} and T_{pos2} was 0.2 ng/μl for each potential prey, and 0.8 ng/μl for *Z. asper*. T_{pos1} and T_{pos2} were used to gauge sequencing or PCR artefacts and evaluate reliability of our metabarcoding analyses (see De Barba et al., 2014). The DNA from the specimens used for our mock samples was extracted in a room free of DNA handling. Moreover, the COI reference sequences of each specimen were originally obtained by amplification using CK4 primer set (see Appendix S1) and Sanger sequencing (except the *Chironomidae* sp. sample that failed to be amplified with CK4).

Amplicons were checked by gel electrophoresis and were then pooled by replicate series, that is, 48 samples for each of the six replicate series (two primer pairs, three replicates each). Electrophoretic migrations were carried out using 20 μl of each pool on an agarose gel at 1.25%. The amplicons with expected sizes were isolated using a sterile scalpel, and the DNA was purified using the PureLink® Quick Gel Extraction Kit (Life Technologies, Germany). About 20 ng of purified DNA from each replicate series was used to generate six Illumina sequencing libraries using the TruSeq® Nano DNA Sample Preparation Kit (Illumina, San Diego, CA, USA). The library preparation included DNA end-repairing, A-tailing, Illumina adapter ligation, and limited amplification (12 PCR cycles). Illumina libraries were distinguished by the ligation of distinct adapters. The libraries were controlled for size and quality using the Agilent Bioanalyzer DNA 1000 Kit (Agilent Technologies, Palo Alto, CA, USA) and for DNA concentrations with the Kapa Library Quantification Kit for Illumina® platforms (KapaBiosystems, Wilmington, MA, USA). The six libraries were pooled at equimolar concentration (4 nM) and sequenced on an ILLUMINA MISEQ v3 platform as paired-end 250-nucleotide reads.

2.6 | Sequence processing and filtering

We developed a pipeline (see Figure 1) to filter the obtained MiSeq data (see Appendix S3 for a detailed version of the filtering pipeline). Unless specified otherwise, bioinformatics processing of reads was performed using custom Perl scripts (Dryad; https://doi.org/10.5061/dryad.f40v5) and statistical analyses were conducted using R software (R Development Core Team 2014). Reads from the different primer pairs and replicate series were sorted according to the Illumina adapter sequences. *PEAR* v0.9.5 (Zhang, Kobert, Flouri, & Stamatakis, 2014) was used to merge read pairs and discard low-quality reads (Figure 1, Step 1). Merged reads were then assigned to samples and replicates using a BLAST-based approach, and tags and primers were trimmed from the reads. Only reads that had a perfect match to tags were accepted. This constituted an additional step to eliminate low-quality reads. Strictly identical trimmed reads within each of the six replicate series were pooled into variants (i.e. dereplicated reads) and singletons (i.e. only one read in a replicate series) were discarded (Figure 1, Step 2). Variants of the three replicate series obtained from the same primer set (MFZR or ZFZR) were pooled. The read number associated with each variant–replicate combination, however, was kept.

The following filtering steps were done separately for MFZR and ZFZR. Variants that did not comply with our BLAST conditions (*E*-value threshold: 1e-10; minimum query coverage: 80%) against our custom COI database (COI-filtering-DB; Dryad, https://doi.org/10.5061/dryad.f40v5) were considered as non-COI sequences and discarded (Figure 1, Step 3).

Then, LFN filters were used to discard variants with low read counts, likely originating from contamination, mistagging or sequencing/PCR errors bias. All variant–replicate combinations were considered independently. For each replicate of a sample, all the variants associated with a read count or a relative frequency that were under one of the LFN thresholds (i.e., not distinguishable from the noise inherent to the HTS data) were removed. Three different thresholds were considered. In fact, a variant can be rare in a given replicate (i) compared to the total number of reads in the replicate (N_{repl}), (ii) compared to its total number of reads in the run (N_{var}) or (iii) if it has few reads in a replicate in absolute terms ($N_{var\text{-}repl}$). The first threshold (LFN_{pos} = 0.3%) was based on the relative frequency ($N_{var\text{-}repl}/N_{repl}$) of the least frequent expected variant in all mock community replicates. This threshold helps to eliminate most low-

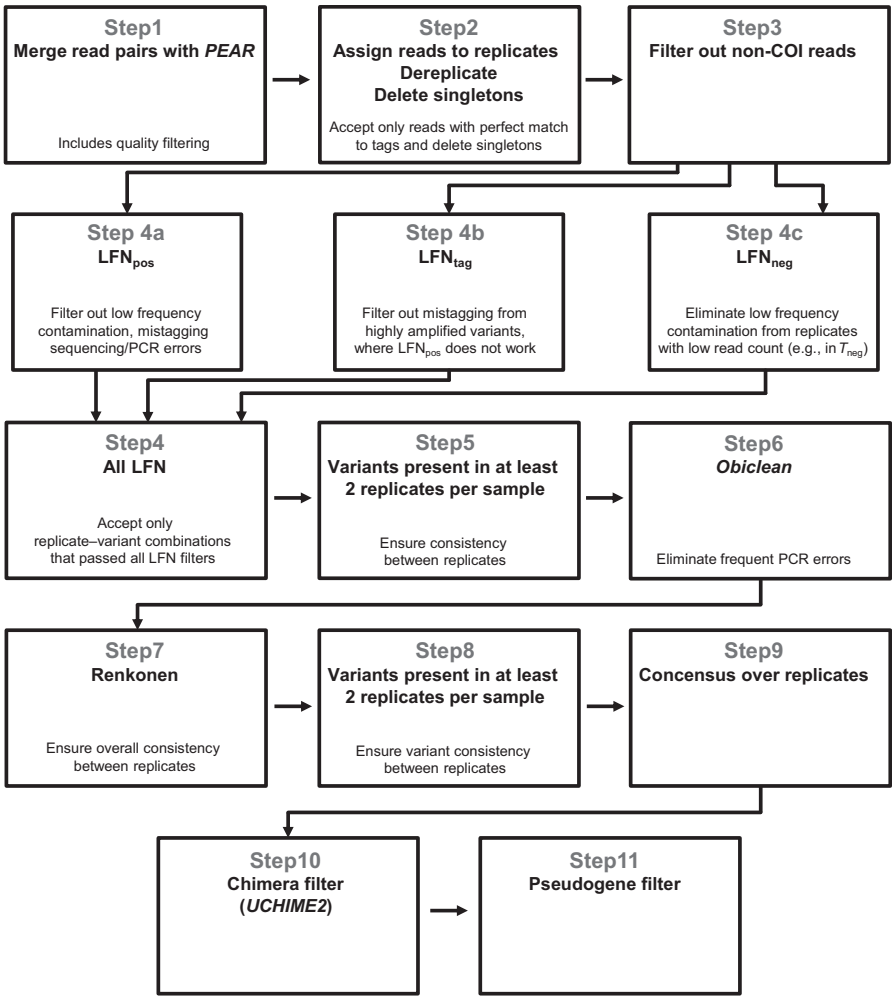

FIGURE 1 High-throughput sequencing filtering pipeline

frequency variants due to the bias cited above (Figure 1, Step 4a). The second threshold (LFN_{tag} = 0.25%) was determined based on (i) freshwater prey variants present in samples of brackish-water fish samples (*P. microps*), (ii) brackish-water prey variants present in samples of freshwater fish samples (*Z. asper*) and (iii) unexpected variants present in the mock samples. The LFN_{tag} allowed discarding variants that appear by mistagging or cross-sample contamination due to their high frequency in the run (Figure 1, Step 4b). Finally, for replicates that had a low number of reads, contamination could not necessarily be discarded with LFN_{pos} and LFN_{tag} filters. Therefore, a last threshold (LFN_{neg}) based on the maximum read count ($N_{var\text{-}repl}$) among all variants of all negative controls replicates was used (LFN_{neg}: 31 and 53 for MFZR and ZFZR, respectively; Figure 1, Step 4c). The three LFN filters were run in parallel on the merged and dereplicated COI variants and only variant–replicate combinations that passed all three LFN filters were retained (Figure 1, Step 4), except those that were present in only one replicate within a sample (Figure 1, Step 5).

When occurring early in PCR cycles, errors can generate variants with quite high frequency compared to the error-free sequences. We therefore used the *Obiclean* program (*Obitools* package; Boyer et al., 2016) to further filter out variants resulting from PCR and/or sequencing errors. Based on the known variants of mock samples, a 20% threshold was determined, and all variants classified as "internal" (see: Shehzad et al., 2012; De Barba et al., 2014; Appendix S3) were discarded (Figure 1, Step 6).

By comparing replicates within samples, we were able to assess the repeatability of the experimental procedure in two different ways: (i) only variants present in at least two replicates of a sample were retained (Figure 1, Step 5), and (ii) distance between replicates was taken into account (Figure 1, Step 7). To this latter end, we followed the strategy developed by De Barba et al. (2014) and used the Renkonen distance (RD) to compare distances between replicates of the same sample. The threshold was set at the 10% upper tail of the distribution of the RDs, which corresponds in our data set to RD = 0.157 for MFZR, and RD = 0.088 for ZFZR (Fig. S1). All those replicates separated from the other replicates by RD above the defined threshold were discarded. Furthermore, samples with only one replicate were excluded (Figure 1, steps 7–8).

After the filtering, consensus samples were created by averaging read counts over replicates (Figure 1, Step 9). The pool of all remaining variants and contigs was filtered for chimeras using *UCHIME2* (Edgar, 2016a; Figure 1, Step 10) and for potential pseudogenes

(Figure 1, Step 11). Variants obtained from ZFZR and MFZR primer pairs that perfectly matched on their overlapping regions were merged into contigs. At this stage, the read count values associated with variants were disregarded, and only the presence/absence of the variants/contigs was considered in a given sample.

In parallel, the data set that passed filtering steps 1–3 was also analysed by *UNOISE2* (Edgar, 2016b) for denoising reads. *UNOISE2* intends to eliminate all noise coming from sequencing and PCR errors. However, *UNOISE2* does not deal with mistagging and contamination. We therefore further filtered the data by *UNCROSS* (Edgar, 2016c). The *UNOISE2*/*UNCROSS* filtering procedure is roughly comparable to our filtering steps 4 and 6. To make the results of this second filtering approach comparable to our pipeline, we further applied our filtering steps 5, 7 and 8 to ensure repeatability within samples, as well as filtering steps 10 and 11 to eliminate chimeras and pseudogenes variants, respectively.

2.7 | Taxonomic assignation of variants

Four different approaches were used for the taxonomic assignment of variants and contigs. More detailed descriptions of the methods are given in Appendix S4. First, the phylogenetic approach implemented in the Statistical Assignment Package (SAP; Munch, Boomsma, Huelsenbeck, Willerslev, & Nielsen, 2008) was used to build phylogenetic trees for each of the variants/contigs and their homologues in GenBank using ≥95%, ≥85% and ≥70% sequence identity thresholds in three consecutive runs. The posterior probability (pp) for the query sequence to belong to a clade was estimated at different taxonomic levels.

Second, we used an automatic procedure that assigned each variant/contig to the lowest taxonomic group (LTG) of BLAST hits against a custom-built local database (Taxassign-DB; Appendix S4). The principle of our LTG approach is similar to the lowest common ancestor developed by Huson, Mitra, Ruscheweyh, Weber, and Schuster (2011). All variants/contigs were BLASTed against the Taxassign-DB (BLASTN, E-value 1e-10, minimum coverage of the query sequence: 90%, subject is annotated to a family or lower level). The LTG was determined based on different similarity cut-offs (S) from 100% to 70%. At each S, the LTG was defined as the lowest taxonomic group that contained at least 90% of the selected hits and contains at least three taxa for $S < 97\%$. The LTG corresponding to the highest S was selected.

Third, we used the Identification Request of BOLD Systems Tools (www.boldsystems.org; Ratnasingham & Hebert, 2007) using the COI "All Barcode Records" database (performed in September 2015). This approach involved all the sequences of the BOLD Systems, even those not published yet.

Fourth, when a cross-comparison between the assignment analyses revealed inconsistencies or when the taxonomic assignments had low resolution, we performed complementary phylogenetic analyses of the variants and contigs. For each of these ambiguous variants/contigs, we selected ten sequences from the Taxassign-DB by retaining: (i) the best BLAST hits (BLASTN, *E*-value 1e-10, minimum coverage of the query sequence: 90%); (ii) a maximum of three sequences per taxon; and (iii) only sequences annotated to at least the Family level. Neighbor-joining trees were constructed based on K2P distances with 1,000 bootstraps using MEGA 6.06 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013). The topology of the phylogenetic trees, especially the apices of branches, was used to resolve the taxonomy of variants and contigs. When the tree topologies did not resolve the taxonomic ambiguities, biogeographical data were then considered (Tachet, Richoux, Bournaud, & Usseglio-Polatera, 2010; OPIE-Benthos database: www.opie-benthos.fr). The combination of the cross-comparison between SAP, LTG and BOLD assignment analyses and the phylogenetic and biogeographical approaches led to a final taxonomic assignment (FTA). To assess the efficiency of the different assignment methods, we focused on the invertebrates' variants and contigs identified in *Z. asper* faecal samples. An Identification Resolution index (IR; Zarzoso-Lacoste et al., 2016) was calculated for each *Z. asper* sample and for each assignment analysis. A score was attributed to each variant/contig according to its taxonomic rank assignment, where the maximal value is given to the species level (i.e., Species = 6, Genus = 5, Family = 4, Order = 3; Class = 2, Phylum = 1, Kingdom or NA = 0) and the IR represented the mean score among the variants of a given sample. Identification Resolution indices of the different assignment methods were compared by pairwise nonparametric Wilcoxon rank sum, controlled for multiple comparisons with the Benjamini & Hochberg (1995) procedure.

2.8 | Additional analyses

The variants and contigs validated in the *Z. asper* and *P. microps* samples were used to further evaluate the coverage of MFZR and ZFZR primer pairs. To standardize this evaluation, a complete-linkage clustering (Sneath & Sokal, 1973) was used to delineate molecular operational taxonomic units (MOTUs; Blaxter et al., 2005) among validated variants and contigs. A maximum of 3% divergence between variants/contigs was allowed within a MOTU. This maximal divergence was previously used as a proxy for delineating invertebrates' species (e.g., Clare et al., 2014; Vesterinen et al., 2016). For taxa that present a higher within-species sequence divergence, however, this 3% threshold will allow taking into account possible differential intraspecific coverage.

In diet studies, the quantification of individuals per prey would add considerable ecological information. However, due to PCR biases, read counts are generally considered as inappropriate to assess the relative abundance of prey (Elbrecht & Leese, 2015). Alternatively, we used a variant-centred approach: we assumed that the number of distinct variants/contigs represented the Minimal Number of Individuals (MNI; White, 1953) ingested by the predator, providing a semiquantitative estimation of the diet composition. The MNI was used to summarize the proportion of prey items or MOTUs in each faecal sample as well as at the population level. Furthermore, to assess the relation between sample size and MOTU diversity, we constructed two sample-based rarefaction curves (for invertebrates and for all prey) using the observed-richness function implemented in *EstimateS* v9.1.0 (http://viceroy.eeb.uconn.edu/estimates/).

Two categories of prey were taken into consideration during our analysis: invertebrates (excluding Rotifera) and microorganisms (including diatoms, red and green algae, and Rotifera). In this work, we focused on invertebrates as we collected faeces from invertebrate-eating species. In contrast, microorganisms are less relevant since some uncertainty remains about their ingestion (secondary predation and/or nontargeted ingestion when predating invertebrates).

3 | RESULTS

3.1 | HTS data filtering

High-throughput sequencing of the six amplicon libraries (288 PCR products: 40 faecal samples, 2 T_{pos}, 2 T_{ext}, 2 T_{pai}, 1 T_{PCR}, 1 T_{tag}; 2 primer pairs; 3 replicates/sample/primer pair) generated a total of about 8.6 million (M) paired-end reads (per base read quality plots available in Fig. S2). The number of reads, variants, replicates and samples that were validated by each filtering step is reported in Table 1, for the whole data set and for T_{pos1} only (as a sample example). After the initial quality filtering and assignment steps (steps 1 and 2), the number of reads per replicate varied between 137 and 64,718 (median: 26,943) for faecal samples, 21,613 and 32,490 (median: 26,943) for positive controls, and 16 and 230 (median: 38) for negative controls in the MFZR data set. For ZFZR, read counts per replicate varied between 83 and 76,279 (median: 26,603) for faecal samples, 19,144 and 42,281 (median: 36,645) for positive controls, and 15 and 167 (median: 48) for negative controls. The total number of variants excluding singletons was 27,921 and 23,619 for MFZR and ZFZR, respectively (Table 1). During the filtering stage, the three LFN filters drastically reduced the number of variants: 99.2% and 99.5% of the variants were discarded. Nevertheless, the remaining variants still represented 75% (MFZR) and 78% (ZFZR) of the reads present before the LFN filtering (Step 4). All negative control replicates for both markers were eliminated at this step. Additionally, three *P. microps* samples (P1, P2 and P4) were retained for MFZR data set only. After eliminating the variants present in only one replicate per sample (Step 5), one sample for MFZR (14Ben09) and two samples for ZFZR (14Ben09, 14Deo04) were discarded. The *Obiclean* step (Step 6) reduced further the number of variants (by 32% for MFZR and 35% for ZFZR). When applying the Renkonen filter (Step 7) and eliminating variants that occur in only one replicate per sample (Step 8), one MFZR sample (14Ben05) and two ZFZR samples (14Ben02, 14Ben05) were further discarded. *UCHIME2* detected four MFZR and 17 ZFZR variants as chimeras (Step 10) and four MFZR and two ZFZR variants were potential pseudogenes (Step 11). The final 81 MFZR and 61 ZFZR variants represented <0.3% of the number of variants validated at Step 2, but they corresponded to over 70% of the reads initially assigned to samples.

After the contigation of the MFZR and ZFZR variants, 38 distinct contigs and 66 distinct variants (43 amplified by MFZR only and 23 amplified by ZFZR only) were retained. A total of 93 variants and contigs were found in the *Z. asper* faecal samples and six in the *P. microps* samples. The DNA of *Z. asper* was neither detected in the mock samples nor in the faecal samples. At the end of the HTS filtering process, all negative controls were eliminated, and 38 of the 40 initial faecal samples were represented by at least one variant or contig (samples 14Ben05 and 14Ben09 were not validated). Seven variants/contigs were identified in the mock samples (T_{pos1} and T_{pos2}): six that correspond to the six organisms used for the preparation of the mock samples (Figure 4a, Table S3), and one unexpected variant (MFZR_082136) assigned to an undetermined Eukaryota (Fig. S4). This variant is present in both mock community samples but absent from all other samples. Therefore, it is likely that it comes from an organism ingested by or attached to one of the prey individuals used for the construction of the mock sample.

When using the pipeline based on *UNOISE2* and *UNCROSS* for filtering our data set, a total of 18 variants were retained for each mock community sample (17 present in both). While this pipeline allowed the validation of the six variants expected in the mock samples, it retained 12 additional and unexpected variants (including MFZR_082136).

3.2 | Taxonomic identification and resolution

The 99 variants and contigs identified for *Z. asper* and *P. microps* faecal samples were analysed using the SAP, BOLD, LTG and FTA analyses (see Table S3). Regarding the variants and contigs assigned to invertebrates in *Z. asper* samples (n = 42), the taxonomic resolution of each analysis was quantified using the IR index (Figure 2a). The IR of FTA ($IR_{FTA} = 5.8 \pm 0.3$) was significantly higher than those calculated for SAP ($IR_{SAP} = 4.1 \pm 1.1$), BOLD ($IR_{BOLD} = 5.2 \pm 1.2$) and LTG ($IR_{LTG} = 5.7 \pm 0.3$). The FTA approach resulted in the assignment of most variants and contigs to the species level (Figure 2b). It should be noted that a very close performance (although significantly lower) was obtained by the fully automatic LTG approach.

During the cross-validation step, the taxonomic assignments of SAP, LTG and BOLD were compared: (i) in 6% of the cases, all three analyses suggested the same taxon; (ii) in 87% of cases, the analyses were consistent but assigned the variants/contigs to different taxonomic levels and the lowest taxonomic level was retained; (iii) in 7% of cases, the three approaches produced conflicting assignments. Variants or contigs with unsatisfactorily taxonomic levels or conflicting assignments were subjected to a phylogenetic analysis. All variants and contigs assigned to *Simulium* sp. were also included in the phylogenetic analyses, since at least one of them showed a conflicting assignment. Phylogenetic analyses were conducted for a total of 33 variants and contigs (sequence alignment deposited in Dryad; https://doi.org/10.5061/dryad.f40v5), which resulted in (i) the resolution of the assignment of nine variants and (ii) the refinement of the identification of a further 14 variants (Figs S3 and S4). In one case, biogeographical data were also considered for the FTA decision (Table S3; contig_0260 identified as *Antocha vitripennis* since it is recorded in sampling area, whereas *Austrophorocera Janzen04* sequences originate from Costa-Rica).

TABLE 1 Read, variant, sample and replicate counts after each filtering steps. Between parentheses, data for T_{pos1} when applicable

		ZFZR				MFZR			
Step	**Filter**	**Read count**	**Variants**	**Replicates**	**Samples**	**Read count**	**Variants**	**Replicates**	**Samples**
0	Total number of read pairs	4 422 025	NA	NA	NA	4 186 396	NA	NA	NA
1	Merge read pairs (*PEAR*)	4 161 519	NA	NA	NA	3 903 413	NA	NA	NA
2	Assign to replicates, discard singletons	3 061 455	23 619	144	48	3 046 833	27 921	144	48
3	Filter out non-COI	3 052 270 (81 264)	23 217 (1 384)	144 (3)	48 (1)	3 022 982 (70 797)	26 954 (1 773)	144 (3)	48 (1)
4a	LFN_{pos}	2 425 759 (62 796)	945 (5)	144 (3)	48 (1)	2 307 139 (53 468)	549 (13)	144 (3)	48 (1)
4b	LFN_{tag}	3 002 234 (81 135)	23 217 (1 355)	144 (3)	48 (1)	2 984 499 (69 671)	26 954 (1 712)	144 (3)	48 (1)
4c	LFN_{neg}	2 489 306 (63 232)	530 (9)	119 (3)	41 (1)	2 550 218 (55 514)	1 248 (52)	125 (3)	42 (1)
4	All LFN	2 383 667 (62 796)	171 (5)	112 (3)	39 (1)	2 280 261 (52 603)	214 (12)	122 (3)	42 (1)
5	Variants present in at least two replicates per sample	2 370 145 (62 493)	135 (3)	109 (3)	37 (1)	2 261 388 (52 481)	135 (10)	121 (3)	41 (1)
6	*Obiclean*	2 287 734 (62 493)	88 (3)	109 (3)	37 (1)	2 188 485 (50 938)	92 (7)	121 (3)	41 (1)
7	Renkonen	2 247 766 (62 493)	82 (3)	101 (3)	35 (1)	2 177 917 (50 938)	91 (7)	115 (3)	40 (1)
8	Variants present in at least two replicates per sample	2 247 421 (62 493)	80 (3)	101 (3)	35 (1)	2 177 736 (50 938)	89 (7)	115 (3)	40 (1)
9	Consensus over replicates	2 247 421 (62 493)	80 (3)	NA	35 (1)	2 177 736 (50 938)	89 (7)	NA	40 (1)
10	Chimeras discarded (*UCHIME2*)	2 234 992 (62 493)	63 (3)	NA	35 (1)	2 175 729 (50 938)	85 (7)	NA	40 (1)
11	Pseudogenes discarded	2 220 391 (62 493)	61 (3)	NA	35 (1)	2 171 780 (50 938)	81 (7)	NA	40 (1)

NA, not applicable.

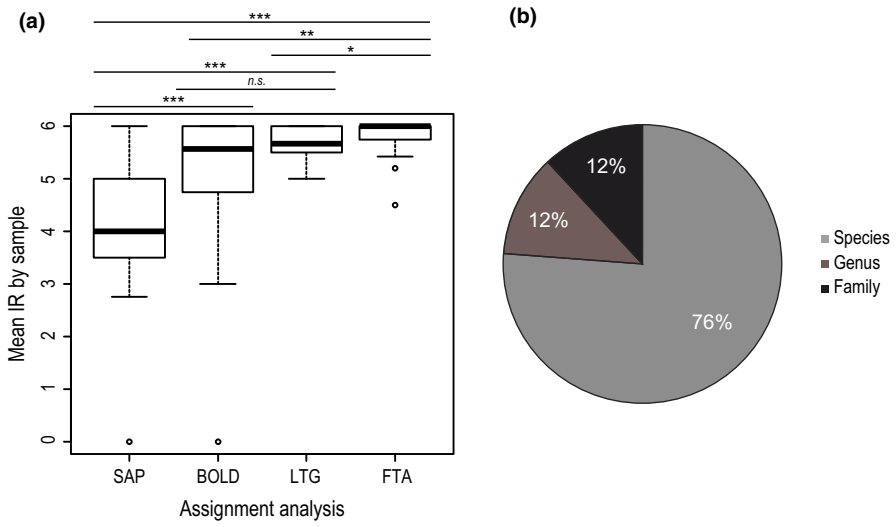

FIGURE 2 Taxonomic assignment analysis. (a) The distribution of index resolution (IR; using invertebrates only) among *Zingel asper* samples was represented by box plots for each taxonomic assignment procedure. Significant differences between taxonomic analyses were indicated at the top; n.s., nonsignificant; $*p < .05$; $**p < .005$; $***p < .0005$). (b) The taxonomic levels of the *Z. asper* invertebrate prey identified using the FTA procedure

3.3 | Taxonomic coverage of the minibarcode primer set

In vitro tests showed that 87% and 85% of the taxa were amplified by ZFZR and MFZR, respectively. The combination of both primer pairs leads to a mean amplification success of 94% (Table 2): 100% for Arthropoda, 80% for Mollusca, 90% for Annelida and 65% for Teleostei. The PCR success was generally similar among samples of the same species, with a few exceptions (Table S1).

In silico tests showed that the taxonomic coverage at the species level for Arthropoda was 63% and 66% for MFZR and ZFZR, respectively (Table S4). When combining the coverage of the two primer pairs, the taxonomic coverage increased to 77% for Arthropoda, varying from 62% (Hymenoptera) to 100% (Lepidoptera). The discriminating power of the minibarcoding amplicons was also estimated in silico: 79%–100% of the analysed species could be differentiated by at least one mutation for the analysed taxa (Table S4).

TABLE 2 Taxonomic coverage of the metabarcoding primers (in vitro tests)

			ZFZR		MFZR		Combined success rate	
Taxa	Number of taxa	Number of samples	Species rate (%)	Sample rate (%)	Species rate (%)	Sample rate (%)	by species (%)	by sample (%)
Insecta								
Trichoptera	12	25	100	96	92	84	100	96
Ephemeroptera	27	60	100	90	96	85	100	90
Plecoptera	9	18	100	94	100	78	100	94
Diptera	54	76	98	97	92	87	100	100
Coleoptera	17	23	100	100	76	83	100	100
Odonata	7	12	86	92	86	92	100	100
Hymenoptera	2	2	100	100	100	100	100	100
Hemiptera	4	4	80	80	100	100	100	100
Lepidoptera	3	4	100	100	100	100	100	100
Blattodea	8	10	100	100	100	100	100	100
Arachnida	8	10	100	100	100	100	100	100
Crustacea	9	16	89	75	89	75	100	75
Myriapoda	3	3	100	100	100	100	100	100
Mollusca	15	25	47	52	73	76	80	80
Annelida	10	15	80	67	90	80	90	80
Platyhelminthes	1	1	100	100	100	100	100	100
Teleostei	20	61	35	20	35	34	65	51
Total	209	365	87	78	85	77	94	86

To further evaluate the complementarity of the two primer pairs (ZFZR and MFZR), we looked at their performance in detecting the 81 MOTUs identified in the *Z. asper* and *P. microps* faecal samples (Table S5): ~30% MOTUs were detected with both primer pairs, 23% were detected with ZFZR only, and 47% were detected with MFZR only (Figure 3a). When focusing on invertebrates, 53% of MOTUs were detected with one of the two primer sets only (Figure 3b). The ZFZR pair was more efficient in detecting Diptera (both for Chironomidae and Simuliidae), whereas MFZR detected more Baetidae (Ephemeroptera) and Crustacea.

3.4 | Sample composition

The number of variants and contigs in the *Z. asper* faecal samples varied from 1 to 24 (6.6 ± 5.5). At least one variant of invertebrates was present in all *Z. asper* samples, with a mean value of 3.4 (± 1.7) (Table S3; Figure 4a). These samples also varied in terms of prey composition: whereas *Baetis* was detected in 79% of the samples, the frequency of other prey varied widely. Ephemeroptera, Diptera and Trichoptera appeared to be the most abundant prey at the population scale (Figure 4b). Four Ephemeroptera families were identified with a predominance of Baetidae (five species), which also displayed the highest MNI (22% of the total MNI in the *Z. asper* samples; Figure 4b). The main groups of Diptera that were preyed upon were Simuliidae and Orthocladiinae (subfamily of Chironomidae). Noninvertebrate taxa were mainly composed of metazoan microorganisms: Bacillariophyta, Rotifera, Rhodophyta and Oomycetes. Furthermore, none of the dietary MOTU accumulation curves approached an asymptote (Figure 4c).

Regarding the *P. microps* samples, prey DNA was detected in samples P3 and P5 only (one and four variants/contigs, respectively), and the DNA of *P. microps* was detected in all five faecal samples (Table S3).

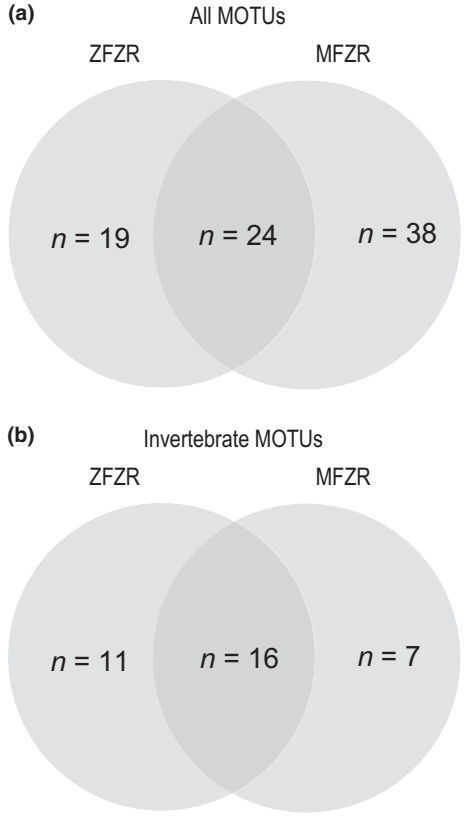

FIGURE 3 Complementarity of minibarcoding primer sets based on *Zingel asper* and *Pomatoschistus microps* samples

4 | DISCUSSION

4.1 | Producing robust HTS data sets

The choice of the primer sets is crucial in diet metabarcoding as it may lead to false negatives due to insufficient taxonomic coverage (Pompanon et al., 2012). In this study, we used two primer sets to amplify a short but informative COI region (MFZR and ZFZR). In vitro tests on our lcDNA samples showed that the primer sets were complementary and enabled coverage of a large taxonomic spectrum, especially for invertebrates (100% of items), whereas in silico tests suggested lower coverage. Our *Z. asper* faecal samples displayed a taxonomic diversity comparable to a previous morphological-based diet study (Cavalli, Pech, & Chappaz, 2003; summary in Fig. S5). Interestingly, the complementarity of primer sets was higher in the *Z. asper* faecal samples (53% of invertebrates MOTUs detected by one of the two primer sets only) compared to in vitro tests (11%). Increased complementarity in multiplexed samples is likely to be the result of preferential primer binding (e.g. Thomas, Deagle, Eveson, Harsch, & Trites, 2015), which justifies the use of more than one primer pair. In our case, the selection of two primer pairs led to the detection of all prey in the positive samples and to a high prey diversity within and among *Z. asper* samples. Furthermore, the successful detection of prey in *P. microps* faecal samples suggested that these primer sets are applicable for species that live in nonfreshwater environments. However, in vitro tests on lcDNA samples revealed that the combined use of ZFZR and MFZR displayed insufficient taxonomic coverage in some groups (e.g. 80% of Mollusca). Nevertheless, our workflow could easily be adapted to other predators and environments, thanks to the availability of several "universal" minibarcodes located in the COI region that we have targeted (e.g., Brandon-Mong et al., 2015; Hajibabaei, Shokralla, Zhou, Singer, & Baird, 2011; Leray et al., 2013; Shokralla et al., 2015).

The most commonly used filtering pipelines when dealing with HTS metabarcoding data, such as *mothur* (Schloss et al., 2009) and *QIIME* (Caporaso et al., 2010), rely on the clustering of reads into MOTUs. In these clustering-based approaches however, almost each step and each parameter in the bioinformatics pipeline influence the outcome, and the number of taxa is often overestimated (e.g. Brown et al., 2015; Clare et al., 2016; Majaneva, Hyytiäinen, Varvio, Nagai, & Blomster, 2015). Therefore, each data set will need a specific analysis method and parameters should be tailored to the purpose of the study (Flynn et al., 2015). We therefore favoured and developed a clustering-free pipeline that relies on a series of stringent filtering

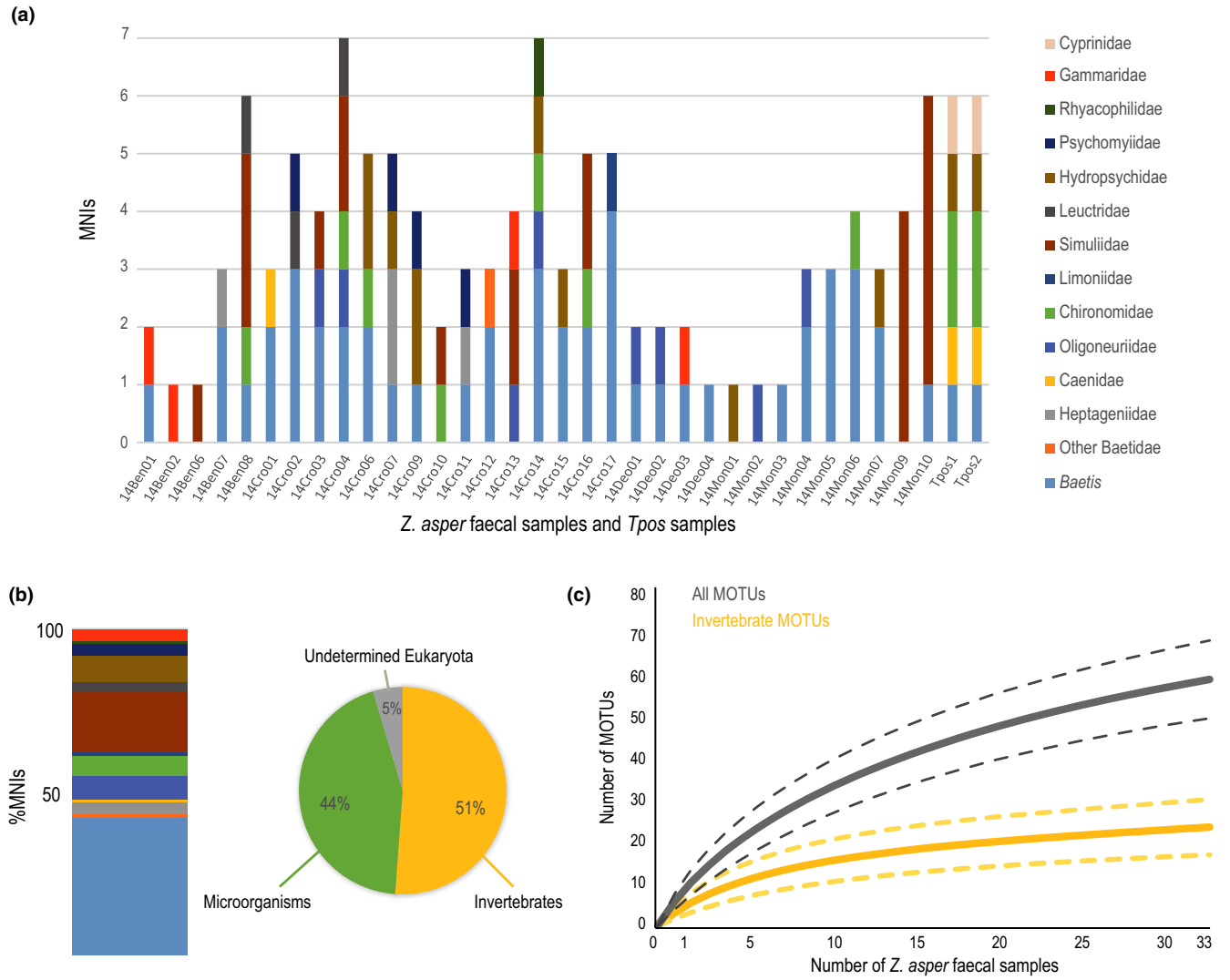

FIGURE 4 The diet composition of *Zingel asper*. (a) Invertebrate composition by sample based on absolute minimal number of individuals (MNIs). (b) Invertebrate composition at the population level, and proportions of microorganisms and invertebrates based on cumulative MNIs. (c) Dietary molecular operational taxonomic unit accumulation curves for *Z. asper* samples (±95% confidence intervals)

steps based on read counts of each variant in each replicate (for a similar approach, see De Barba et al., 2014). For this purpose, the thresholds used in our filtering approach were inferred from mock community samples (Step 4a) and negative controls (Step 4c). It has to be noted that this conservative approach will discard all variants for which frequencies are below the LFN thresholds. In the context of complex community samples, such as samples collected for biodiversity assessment and monitoring (e.g., Elbrecht & Leese, 2017; Lanzén, Lekang, Jonassen, Thompson, & Troedsson, 2016; Leray & Knowlton, 2015), some low-abundance taxa may not be validated. In fact, their read count may fall below one or more LFN thresholds, making the corresponding variant undistinguishable from noise. In this case, the analysis of several biological replicates (i.e., distinct DNA extractions corresponding to different fractions of a given sample) should help with the detection and the validation of low-abundance taxa (Lanzén, Lekang, Jonassen, Thompson, & Troedsson, 2017; Zhan et al., 2014). In addition, we recommend that the mock samples approximate the complexity of the communities sampled, by approximating their expected taxonomic composition and even the differential abundance of taxa.

Ignoring the mistagging bias (Schnell et al., 2015) may lead (i) to overestimate the number of taxa in a sample and (ii) to blur the differentiation between samples with respect to their taxonomic composition. To assess and overcome mistagging bias, two strategies were recently suggested: (i) increasing the number of unused tag combinations using Latin square design (Esling et al., 2015); and (ii) using fusion primers (Herbold et al., 2015) to avoid the creation of intersample chimeras when pooling samples during the sequencing library preparation. In the last case however, PCR efficiency is reduced substantially (Schnell et al., 2015) and template-specific amplification bias may be inflated (O'Donnell, Kelly, Lowell, & Port, 2016). Following Esling et al. (2015), we chose a variant frequency-

dependent approach (LFN_{tag}; Step 4b) to control mistagging. In our case, the threshold of the LFN_{tag} filter is based on mock community samples and on the co-analysis of samples from different habitats (brackish-water and freshwater) in the same HTS run.

In our workflow, LFN thresholds appear as the most critical parameters for the filtering and the validation of HTS data: most variants that are expected to be false positives were discarded at Step 4 (see Table 1). Nevertheless, subsequent filtering steps (steps 5–11) are also important for the validation of the data. Expectedly, omitting the filtering steps 5–11 will decrease the robustness of the data and inflate the taxonomic and/or genetic diversity within samples, biasing biodiversity estimations. In our case, for example, omitting filtering steps 5–11 is expected to inflate the MNI estimator (see below). As a matter of fact, further false positives were controlled, such as variants with PCR/sequencing errors (Step 6) and chimeras (Step 10). Moreover, our workflow included three PCR replicates for each primer pair to avoid false-positive detections (see Ficetola, Taberlet, & Coissac, 2016). These PCR replicates were sequenced separately and used at filtering steps 5, 7 and 8 for validating reproducible variants only. Several authors, however, used PCR replicates differently in the case of complex community samples (e.g., Lanzén et al., 2016; Leray & Knowlton, 2017): PCR replicates were pooled and then sequenced jointly to limit the impact of random sampling amplification biases and hence minimizing false negatives due to low-abundance taxa. Nevertheless, this approach does not allow minimizing the false-positive detections. In our case, the size of faecal samples allowed one single DNA extraction reaction. However, in the case of larger or more complex samples, several biological replicates (i.e., distinct fractions of the same sample) may be analysed for controlling false negatives (see Lanzén et al., 2017; Zhan et al., 2014). The combined use of biological replicates and of PCR replicates (sequenced separately) should ensure both avoiding false-positive detections and maximizing the coverage of taxa diversity.

Alternatively, for complex community samples, more relaxed filtering thresholds (at steps 4a and 4c) and more relaxed reproducibility criteria (at steps 5, 7 and 8) might be considered if one wants to maximize the detection low-abundance taxa. In this case however, the presence of these taxa in the final data set will remain uncertain since confounded with the low-frequency noise and the false positives. These low-frequency taxa will therefore require to be validated by some complementary field data (for estimating the probability of false negatives, as suggested by: Leray & Knowlton, 2017) and/or by the use of site occupancy-detection models (that can account for the presence of false positives; e.g., Lahoz-Monfort, Guillera-Arroita, & Tingley, 2016).

Recently, new clustering-based methods have been developed for denoising HTS reads (e.g., *UNOISE2*: Edgar, 2016b; *Swarm*: Mahé, Rognes, Quince, De Vargas, & Dunthorn, 2015). These methods avoid clustering the reads based on fixed similarity level and their outcome depends on a very few parameters. We therefore used a modified version of our pipeline, where our denoising steps (LFN and *Obiclean*, steps 4 and 6) were replaced by *UNOISE2* and *UNCROSS*. This filtering approach retained 12 unexpected variants in the mock community samples, showing that our pipeline is more accurate for restituting the composition of our mock samples. In fact, the bioinformatics pipeline we developed retained all the expected variants and filtered out all but one nonexpected variants in mock controls (this latter variant was likely introduced during construction of the mock samples; see "Results"), as well as all variants in negative controls. The final data set contained only 0.3% of the original variants, but these variants represented over 70% of the reads, reinforcing that the eliminated variants could be considered to be noise.

4.2 | Towards a quantitative approach: the MNI

Due to PCR biases, the number of reads is a very poor estimator of abundance in metabarcoding studies (Elbrecht & Leese, 2015), and only presence/absence of MOTUs can be obtained with clustering-based approaches, since different alleles coming from the same taxon are confounded in the same MOTU. Alternatively, several studies highlighted the pertinence of determining the number of distinct sequences belonging to the same taxon when assessing genetic diversity (González-Tortuero et al., 2015; Shokralla et al., 2015) or prey biomass (Jo et al., 2016). In this study, thanks to rigorous filtering procedures we produced a robust data set of variants and contigs that is reliable and relatively free from false positives and artefacts. Moreover, the variants corresponding to pseudogenes were filtered out. Therefore, we are confident that the number of COI variants/contigs can be used to estimate MNIs. However, in some cases, that are prey taxa prone to heteroplasmy or that present tissue-specific mitochondrial variants, an overestimation bias should be considered when using MNIs. The MNI statistic may therefore deserve a further evaluation by using appropriate mock samples as controls.

In the case of *Z. asper* prey, the biological analyses suggested that the final variants can be used as a reliable estimation of their DNA diversity in faeces and therefore that the MNI statistic adequately reflects relative prey abundance. In fact, congruence between the MNI and previously observed prey diversity based on morphological gut-content analysis in Durance River (Cavalli et al., 2003) was observed. More specifically, the family Baetidae exhibited the highest MNIs in *Z. asper* faecal samples, which was also the most abundant prey found in morphological gut-content analysis (Fig. S5).

4.3 | Taxonomic assignment of prey

Taxonomic assignment procedures are critical in metabarcoding studies, and several innovative approaches were recently developed (e.g., Porter et al., 2014; Somervuo, Koskela, Pennanen, Henrik Nilsson, & Ovaskainen, 2016; Somervuo et al., 2017). However, most metabarcoding studies are based on one single taxonomic assignment approach, and yet using different procedures as well as different reference databases should improve the reliability and the accuracy of MOTU identification. For this study, we combined different assignment procedures to combine their respective advantages. Although our strategy can be time-consuming—some of the steps cannot be fully automated—the IR statistic showed that it did significantly improve the precision of the assignments. This resulted in a very low

proportion of high-level taxonomic assignments (e.g., Eukaryota), and a high proportion of low-level taxonomic assignments for variants of invertebrates (species level for most). In contrast, most invertebrates identified by morphological-based gut-content analysis were identified at the Family level in Cavalli et al. (2003). Moreover, our detailed assignment of *Z. asper* prey supports the use of COI as the favoured target gene for invertebrates (see also Brandon-Mong et al., 2015). The COI appears as the one of the most appropriate gene for invertebrate metabarcoding because (i) it reveals species-level variation (Elbrecht et al., 2016), and (ii) a huge amount of annotated sequences is available in public databases (Deagle et al., 2014).

The taxonomic assignment of invertebrates indicated that the Baetidae found in *Z. asper* faeces belonged mostly to the genus *Baetis* although other genera were also present in the sampling area (e.g., *Alainites*, *Acentrella*, *Centroptilum* and *Procloeon*). According to Tachet et al. (2010), *Baetis* has a higher affinity towards coarse substrates (e.g., stones, pebbles) with higher water velocity than other Baetidae, which in turn prefer epiphyte lifestyle on macrophyte or algae substrate. Additionally, the genera *Hydropsyche* (Hydropsychiidae) and *Simulium* (Simuliidae), all epibenthic and rheophilic taxa, occurred at non-negligible frequencies in *Z. asper* faecal samples. Consequently, the cumulated frequency based on MNI of *Baetis*, *Hydropsyche* and Simuliidae was ~70% (see Figure 4). This suggests that *Z. asper* actively selected rheophilic and epibenthic macroinvertebrates, which constitutes accurate information related to habitat use of this critically endangered species.

5 | CONCLUSION AND PERSPECTIVES

Diet analyses are critical to gain a better understanding of prey/habitat relationships and feeding habitats (e.g., Corse et al., 2010; Sanchez-Hernandez, 2014), especially when a fine and accurate taxonomic identification of prey can be achieved (e.g., Adamczuk & Mieczan, 2015). In this study, stringent wet-laboratory conditions, carefully selected primer sets, PCR replicates and nonarbitrary filtering thresholds based on control samples led to the validation of a robust data set dedicated to the study of the diet of a critically endangered fish species. The robustness of our data set allowed us to take into account the prey genetic variability and to obtain a semiquantitative estimate of diet through the use of MNI. Furthermore, a species-level identification for most of the invertebrate prey was obtained through a complementary taxonomic assignment approach. On the whole, our approach produced a robust data set for ecological analyses and opens perspective for more precision on feeding habitats of invertebrate-eaters. This new information would in turn improve conservation strategies such as habitat restoration or the selection of optimal re-introduction sites. Finally, our results reinforced previous findings that suggested diet metabarcoding can be a powerful tool in trophic ecology as it allows the determination of large scale and highly resolved networks (Evans, Kitson, Lunt, Straw, & Pocock, 2016). It may also produce a level of precision that reveals unexpected food web structures (Roslin & Majaneva, 2016). This article proposes a from-benchtop-to-desktop workflow that provides an efficient tool for the study of invertebrate-eater diets and will, we hope, stimulate and inspire future trophic works using metabarcoding approaches.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for their constructive comments, which improved our manuscript significantly. We warmly thank Caroline Costedoat, André Gilles, Georges Olivari and the technicians of the French Office National de l'Eau et des Milieux Aquatiques (ONEMA) for their assistance in the field, especially Guillaume Verdier. This work is part of the French Plan National d'Action en faveur de l'apron du Rhône (2012–2016), supervised by the Direction Régionale pour l'Environnement, l'Aménagement et le Logement de Rhône-Alpes and coordinated by the Conservatoire d'Espaces Naturels Rhône-Alpes. E.C. was supported by a postdoctoral grant from Électricité de France (EDF) and ONEMA. This study was funded by the Syndicat Mixte d'Aménagement du Val Durance (SMAVD), the Agence de l'Eau Rhône-Méditerranée-Corse (AERMC) and the Conseil Régional de Provence-Alpes-Côte d'Azur. This work was conducted in accordance with permits from the French Direction Départementale des Territoires des Hautes-Alpes (DDT 05). Data used in this work were partly produced through the molecular facilities of LabEx CeMEB (Montpellier) and CIRAD (Montpellier).

DATA ACCESSIBILITY

The lcDNA sequences were deposited in GenBank (GenBank ID: MF458551-MF458851). Supplementary data deposited in Dryad (https://doi.org/10.5061/dryad.f40v5) included: (i) custom COI database used during HTS filtering named COI-filtering-DB; (ii) Taxassign-DB; (iii) Perl script with the .csv files indicating the samples/tag combination correspondence; (iv) unfiltered HTS data; and (v) two alignments of COI sequences used during phylogenetic analysis.

AUTHOR CONTRIBUTIONS

V.D. and E.C. conceived and designed the study. E.C., V.D., G.A., R.C. and C.T. collected the samples. G.A. identified invertebrates morphologically. E.C., V.D. and E.M. developed the method. M.A. and J.F.M. contributed some key methodological points. E.C. and C.T. performed the experiments. E.M. and E.C. analysed the data. E.C., E.M. and V.D. wrote the manuscript. All authors drafted the article.

ORCID

Vincent Dubut http://orcid.org/0000-0002-5619-6909

REFERENCES

Adamczuk, M., & Mieczan, T. (2015). Different levels of precision in studies on the alimentary tract content of omnivorous fish affect predictions of their food niche and competitive interactions. *Comptes Rendus Biologies*, *338*, 678–687.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate–a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289–300.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B Biological Sciences*, *360*, 1935–1943.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). Obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*, 176–182.

Brandon-Mong, G.-J., Gan, H.-M., Sing, K.-W., Lee, P. S., Lim, P. E., & Wilson, J. J. (2015). DNA metabarcoding of insects and allies: An evaluation of primers and pipelines. *Bulletin of Entomological Research*, *105*, 717–727.

Brown, E. A., Chain, F. J. J., Crease, T. J., MacIsaac, H. J., & Cristescu, M. E. (2015). Divergence thresholds and divergent biodiversity estimates: Can metabarcoding reliably describe zooplankton communities? *Ecology and Evolution*, *5*, 2234–2251.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335–336.

Cavalli, L., Pech, N., & Chappaz, R. (2003). Diet and growth of the endangered *Zingel asper* in the Durance River. *Journal of Fish Biology*, *63*, 460–471.

Clare, E. L., Chain, F. J. J., Littlefair, J. E., & Cristescu, M. E. (2016). The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome*, *59*, 981–990.

Clare, E. L., Symondson, W. O. C., Broders, H., Fabianek, F., Fraser, E. E., MacKenzie, A., … Reimer, J. P. (2014). The diet of *Myotis lucifugus* across Canada: Assessing foraging quality and diet variability. *Molecular Ecology*, *23*, 3618–3632.

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: *In silico* PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, *14*, 1160–1170.

Corse, E., Costedoat, C., Chappaz, R., Pech, N., Martin, J.-F., & Gilles, A. (2010). A PCR-based method for diet analysis in freshwater organisms using 18S rDNA barcoding on faeces. *Molecular Ecology Resources*, *10*, 96–108.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources*, *14*, 306–323.

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: Not a perfect match. *Biology Letters*, *10*, 20140562.

Edgar, R. C. (2016a). UCHIME2: Improved chimera prediction for amplicon sequencing. *bioRxiv*, 074252. https://doi.org/10.1101/074252

Edgar, R. C. (2016b). UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, 081257. https://doi.org/10.1101/081257

Edgar, R. C. (2016c). UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads. *bioRxiv*, 088666. https://doi.org/10.1101/088666

Elbrech, V., & Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Frontiers in Environmental Science*, *5*, 11.

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—Sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, *10*, e0130324.

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J. N., … Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, *4*, e1966.

Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, *43*, 2513–2524.

Evans, D. M., Kitson, J. J. N., Lunt, D. H., Straw, N. A., & Pocock, M. J. O. (2016). Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Functional Ecology*, *30*, 1904–1916.

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., … Pompanon, F. (2010). An *In silico* approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*, 434.

Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, *16*, 604–607.

Flynn, J. M., Brown, E. A., Chain, F. J. J., MacIsaac, H. J., & Cristescu, M. E. (2015). Toward accurate molecular identification of species in complex environmental samples: Testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, *5*, 2252–2266.

Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., … Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the USA*, *111*, 8007–8012.

González-Tortuero, E., Rusek, J., Petrusek, A., Gießler, S., Lyras, D., Grath, S., … Wolinska, J. (2015). The quantification of representative sequences pipeline for amplicon sequencing: Case study on within-population ITS1 sequence variation in a microparasite infecting *Daphnia*. *Molecular Ecology Resources*, *15*, 1385–1395.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, *6*, e17497.

Herbold, C. W., Pelikan, C., Kuzyk, O., Hausmann, B., Angel, R., Berry, D., & Loy, A. (2015). A flexible and economical barcoding approach for highly multiplexed amplicon sequencing of diverse target genes. *Frontiers in Microbiology*, *6*, 731.

Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., & Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, *21*, 1552–1560.

Jo, H., Ventura, M., Vidal, N., Gim, J. S., Buchaca, T., Barmuta, L. A., … Joo, G. J. (2016). Discovering hidden biodiversity: The use of complementary monitoring of fish diet based on DNA barcoding in freshwater ecosystems. *Ecology and Evolution*, *6*, 219–232.

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources*, *16*, 673–685.

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology*, *25*, 4392–4406.

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2017). DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *PLoS ONE*, *12*, e0179443.

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the USA*, *112*, 2076–2081.

Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, *5*, e3006.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*, 34.

Mahé, F., Rognes, T., Quince, C., De Vargas, C., & Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420.

Majaneva, M., Hyytiäinen, K., Varvio, S. L., Nagai, S., & Blomster, J. (2015). Bioinformatic amplicon read processing strategies strongly affect eukaryotic diversity and the taxonomic composition of communities. *PLoS ONE*, *10*, e0130035.

Meusnier, I., Singer, G. A., Landry, J. F., Hickey, D. A., Hebert, P. D., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, *9*, 214.

Monti, F., Duriez, O., Arnal, V., Dominici, J.-M., Sforzi, A., Fusani, L., … Montgelard, C. (2015). Being cosmopolitan: Evolutionary history and phylogeography of a specialized raptor, the Osprey *Pandion haliaetus*. *BMC Evolutionary Biology*, *15*, 255.

Munch, K., Boomsma, W., Huelsenbeck, J. P., Willerslev, E., & Nielsen, R. (2008). Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, *57*, 750–757.

Murray, D. C., Coghlan, M. L., & Bunce, M. (2015). From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS ONE*, *10*, e0124671.

O'Donnell, J. L., Kelly, R. P., Lowell, N. C., & Port, J. A. (2016). Indexed PCR primers induce template-specific bias in large-scale DNA sequencing studies. *PLoS ONE*, *11*, e0148698.

Pompanon, F., Deagle, B. E., Symondson, W. O., Brown, D. S., Jarman, S. N., & Taberlet, P. (2012). Who is eating what: Diet assessment using next generation sequencing. *Molecular Ecology*, *21*, 1931–1950.

Porter, T. M., Gibson, J. F., Shokralla, S., Baird, D. J., Golding, G. B., & Hajibabaei, M. (2014). Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome *c* oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. *Molecular Ecology Resources*, *14*, 929–942.

R Development Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The barcode of life data system (www.barcodinglife.org). *Molecular Ecology Notes*, *7*, 355–364.

Razgour, O., Clare, E. L., Zeale, M. R. K., Hanmer, J., Schnell, I. B., Rasmussen, M., … Jones, G. (2011). High-throughput sequencing offers insight into mechanisms of resource partitioning in cryptic bat species. *Ecology and Evolution*, *1*, 556–570.

Richardson, R. T., Bengtsson-Palme, J., & Johnson, R. M. (2017). Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. *Molecular Ecology Resources*, *17*, 760–769.

Roslin, T., & Majaneva, S. (2016). The use of DNA barcodes in food web construction—terrestrial and aquatic ecologists unite! *Genome*, *59*, 603–628.

Sanchez-Hernandez, J. (2014). Age-related differences in prey-handling efficiency and feeding habitat utilization of *Squalius carolitertii* (Cyprinidae) according to prey trait analysis. *Biologia*, *69*, 696–704.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*, 7537–7541.

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*, 1289–1303.

Shehzad, W., McCarthy, T. M., Pompanon, F., Purevjav, L., Coissac, E., Riaz, T., & Taberlet, P. (2012). Prey preference of snow leopard (*Panthera uncia*) in South Gobi, Mongolia. PloS ONE, *7*, e32104.

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., … Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, *5*, 9687.

Sint, D., Raso, L., Kaufmann, R., & Traugott, M. (2011). Optimizing methods for PCR-based analysis of predation. *Molecular Ecology Resources*, *11*, 795–801.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. San Francisco, CA: W.H. Freeman and Co.

Soininen, E. M., Gauthier, G., Bilodeau, F., Berteaux, D., Gielly, L., Taberlet, P., … Yoccoz, N. G. (2015). Highly overlapping winter diet in two sympatric lemming species revealed by DNA metabarcoding. *PLoS ONE*, *10*, e0115335.

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., & Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics*, *32*, 2920–2927.

Somervuo, P., Yu, D. W., Xu, C. C., Ji, Y., Hultman, J., Wirta, H., & Ovaskainen, O. (2017). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods in Ecology and Evolution*, *8*, 398–407.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050.

Tachet, H., Richoux, P., Bournaud, M., & Usseglio-Polatera, P. (2010). *Invertébrés d'eau douce: Systématique, biologie, écologie*. Paris: CNRS Editions.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*, *30*, 2725–2729.

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2015). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, *16*, 714–726.

Vesterinen, E. J., Ruokolainen, L., Wahlberg, N., Peña, C., Roslin, T., Laine, V. N., … Lilley, T. M. (2016). What you need is what you eat? Prey selection by the bat *Myotis daubentonii*. *Molecular Ecology*, *25*, 1581–1594.

White, T. E. (1953). A method of calculating the dietary percentage of various food animals utilized by Aboriginal peoples. *American Antiquity*, *18*, 396–398.

Zarzoso-Lacoste, D., Bonnaud, E., Corse, E., Gilles, A., Meglécz, E., Costedoat, C., … Vidal, E. (2016). Improving morphological diet studies with molecular ecology: An application for invasive mammal predation on island birds. *Biological Conservation*, *193*, 134–142.

Zarzoso-Lacoste, D., Corse, E., & Vidal, E. (2013). Improving PCR detection of prey in molecular diet studies: Importance of group-specific primer set selection and extraction protocol performances. *Molecular Ecology Resources*, *13*, 117–127.

Zeale, M. R. K., Butlin, R. K., Barker, G. L. A., Lees, D. C., & Jones, G. (2011). Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources*, *11*, 236–244.

Zhan, A., He, S., Brown, E. A., Chain, F. J. J., Therriault, T. W., Abbott, C. L., … MacIsaac, H. J. (2014). Reproducibility of pyrosequencing data for biodiversity assessment in complex communities. *Methods in Ecology and Evolution*, *5*, 881–890.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*, 614–620.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Corse E, Meglécz E, Archambaud G, et al. A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies. *Mol Ecol Resour*. 2017;17:e146–e159. https://doi.org/10.1111/1755-0998.12703