Strategies for sample labelling and library preparation in DNA metabarcoding studies

Kristine Bohmann[1], Vasco Elbrecht[2], Christian Carøe[1], Iliana Bista[3], Florian Leese[4], Michael Bunce[5], Douglas W. Yu[6], Mathew Seymour[7], Alex Dumbrell[8], and Simon Creer[9]

[1]University of Copenhagen Faculty of Health and Medical Sciences
[2]ETH Zurich
[3]Affiliation not available
[4]University of Duisburg-Essen
[5]Curtin University
[6]University of East Anglia
[7]Swedish University of Agricultural Sciences
[8]University of Essex
[9]Bangor University

May 19, 2021

Abstract

Metabarcoding of DNA extracted from environmental or bulk specimen samples is increasingly used to detect plant and animal taxa in basic and applied biodiversity research because of its targeted nature that allows sequencing of genetic markers from many samples in parallel. To achieve this, PCR amplification is carried out with primers designed to target a taxonomically informative marker within a taxonomic group, and sample-specific nucleotide identifiers are added to the amplicons prior to sequencing. This enables assignment of the sequences back to the samples they originated from. Nucleotide identifiers can be added during the metabarcoding PCR and/or during 'library preparation', i.e. when amplicons are prepared for sequencing. Different strategies to achieve this labelling exist. All have advantages, challenges and limitations, some of which can lead to misleading results, and in the worst case compromise the fidelity of the metabarcoding data. Given the range of questions addressed using metabarcoding, the importance of ensuring that data generation is robust and fit for purpose should be at the forefront of practitioners seeking to employ metabarcoding for biodiversity assessments. Here, we present an overview of the three main workflows for sample-specific labelling and library preparation in metabarcoding studies on Illumina sequencing platforms. Further, we distil the key considerations for researchers seeking to select an appropriate metabarcoding strategy for their specific study. Ultimately, by gaining insights into the consequences of different metabarcoding workflows, we hope to further consolidate the power of metabarcoding as a tool to assess biodiversity across a range of applications.

Strategies for sample labelling and library preparation in DNA metabarcoding studies

Kristine Bohmann[1], Vasco Elbrecht[2], Christian Carøe[1], Iliana Bista[3,4], Florian Leese[5], Michael Bunce[6], Douglas W. Yu[7, 8, ,9], Mathew Seymour[10], Alex J. Dumbrell[11], Simon Creer[12]

Affiliations

[1]Section for Evolutionary Genomics, Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, 1353 Copenhagen, Denmark

[2]Department of Environmental Systems Science, ETH Zurich, Zürich, Switzerland

[3] Department of Genetics, University of Cambridge, Downing Street, CB2 3EH, United Kingdom

[4] Wellcome Sanger Institute, Hinxton, CB10 1SA, United Kingdom

[5]Aquatic Ecosystem Research, Faculty of Biology, University of Duisburg-Essen, 45141 Essen, Germany

[6]Trace and Environmental DNA (TrEnD) Laboratory, School of Molecular and Life Sciences, Curtin University, WA 6162, Perth, Australia

[7]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China

[8]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR4 7TJ, UK

[9]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan, 650223 China

[10] Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[11]School of Life Sciences, University of Essex, Colchester, CO4 3SQ, United Kingdom

[12]Molecular Ecology and Fisheries Genetics Laboratory, School of Natural Sciences, Bangor University, Gwynedd, LL57 2UW, United Kingdom

Running title: Sample-labelling strategies in metabarcoding

Abstract

Metabarcoding of DNA extracted from environmental or bulk specimen samples is increasingly used to detect plant and animal taxa in basic and applied biodiversity research because of its targeted nature that allows sequencing of genetic markers from many samples in parallel. To achieve this, PCR amplification is carried out with primers designed to target a taxonomically informative marker within a taxonomic group, and sample-specific nucleotide identifiers are added to the amplicons prior to sequencing. This enables assignment of the sequences back to the samples they originated from. Nucleotide identifiers can be added during the metabarcoding PCR and/or during ‘library preparation’, i.e. when amplicons are prepared for sequencing. Different strategies to achieve this labelling exist. All have advantages, challenges and limitations, some of which can lead to misleading results, and in the worst case compromise the fidelity of the metabarcoding data. Given the range of questions addressed using metabarcoding, the importance of ensuring that data generation is robust and fit for purpose should be at the forefront of practitioners seeking to employ metabarcoding for biodiversity assessments. Here, we present an overview of the three main workflows for sample-specific labelling and library preparation in metabarcoding studies on Illumina sequencing platforms. Further, we distil the key considerations for researchers seeking to select an appropriate metabarcoding strategy for their specific study. Ultimately, by gaining insights into the consequences of different metabarcoding workflows, we hope to further consolidate the power of metabarcoding as a tool to assess biodiversity across a range of applications.

Keywords

Amplicon sequencing, Biodiversity assessment, Environmental DNA, High-throughput sequencing, Illumina sequencing, Library preparation

Introduction

In recent years, the analysis of environmental DNA (eDNA) and DNA extracted from bulk specimen samples has experienced an enormous surge in popularity in basic and applied biodiversity studies seeking to detect plants and animal taxa (Taberlet*et al.* 2012a; Creer *et al.* 2016; Jarman *et al.*2018). Within the field of genetic biodiversity assessment, DNA metabarcoding is currently the most widely used approach, as it allows targeted, parallel, and as such relatively cost-effective, identification of multiple taxa from DNA extracted from e.g. soil, water, faeces as well as from bulk samples of organisms (Taberlet *et al.*2012b). Here, the application of metabarcoding ranges widely; e.g., detection of invasive species in water samples (e.g. Pochon*et al.* 2013); assessment of water quality via identification of freshwater invertebrates in bulk specimen samples (e.g. Elbrecht*et al.* 2017) and environmental samples (e.g. Seymour*et al.* 2020); identification of plant-pollinator interactions via pollen trapped on the bodies of modern (e.g. Lucas*et al.* 2018) and historical (e.g. Gous*et al.* 2019) pollinator specimens; detection of vertebrate wildlife via invertebrate 'samplers' of vertebrate blood or feces (e.g. Calvignac-Spencer *et al.* 2013), assessment of e.g. niche partitioning (e.g. Razgour*et al.* 2011) and ecosystem services (e.g. Aizpurua*et al.* 2017) through detection of diet items in gut and faecal samples. Furthermore, metabarcoding is explored for implementation in routine biomonitoring around the world (Pont*et al.* 2018, 2021; Li *et al.* 2018, 2019; Aylagas *et al.* 2018; Zizka *et al.* 2020) (www.danubesurvey.org; www.syke.fi), and is an integral component of the proposals for the Next Generation of Biomonitoring programmes (Bohan *et al.* 2017).

Metabarcoding relies on PCR amplification of extracted DNA with primers designed to target a taxonomically informative marker for a selected taxonomic group (Taberlet*et al.* 2012b) (Fig. 1). The backbone of metabarcoding analyses is the addition of sample-specific nucleotide identifiers to amplicons and the use of these to assign metabarcoding sequences back to the samples they originated from ('demultiplexing'). This allows pooling of hundreds to thousands of samples for sequencing and thereby full utilisation of the capacity of high-throughput sequencing platforms (Fig. 1). Amplicon labelling can be achieved at two stages during a metabarcoding workflow: prior to library build as 5' nucleotide 'tags' on amplicons and/or during library build as library indices. The strategies to achieve this labelling can be categorised into three main approaches (Fig. 2). All three approaches have advantages, challenges and limitations, which - if not considered - can result in misleading data interpretation, and in the very worst case can lead to unusable data and considerable wasted time and money, as for instance in the case of the so-called 'tag-jumps' (Schnell *et al.* 2015; Esling *et al.* 2015; Carøe & Bohmann 2020). Despite this, in contrast to discussions on metabarcoding substrate selection, DNA extraction and data processing, the strategies for amplicon labelling and library preparation workflows have received little systematic attention in the metabarcoding literature (although see Murray *et al.* 2015).

Here, we present an overview of the three most commonly used workflows with which to achieve sample-specific labelling and library preparation in metabarcoding studies and how they can potentially influence the resulting data. For the sake of simplicity, we focus on metabarcoding of plants and animals in basic and applied biodiversity studies with sequencing on arguably the most used high-throughput sequencing platform series today, the Illumina sequencing platforms. Doing so, we provide critical considerations for researchers to choose the optimal metabarcoding strategy for generating reliable data tailored to their individual study;for example, regarding sample type and number, research question, speed of laboratory processing, contamination risk, budget and whether similar studies are to be carried out in the laboratory in the future. Ultimately, by gaining detailed and critical insights into the consequences of choosing different metabarcoding workflows, we hope to further increase the potential of metabarcoding as a reliable tool for use across a wide range of applications.

Tagging and indexing approaches in metabarcoding studies

Today, the most commonly used high-throughput sequencing platform for metabarcoding studies is the Illumina series, where for example the MiSeq, iSeq, HiSeq, NextSeq, and NovaSeq have been employed (Jarman *et al.*2018). These platforms offer high-throughput, relatively low error rates, and \soutlong paired-end reads, typically up to 150bp of each paired read on the NextSeq550/1000/2000, HiSeq 3000/4000 and NovaSeq (up to 250 bp on SP flow cell), and 300bp of each paired read on the MiSeq platform (www.illumina.com, applied in e.g. Shehzad *et al.* 2012b; Quéméré *et al.* 2013; Hope *et al.* 2014; Elbrecht *et al.* 2017; Stoeck *et al.* 2018; Singer *et al.* 2019).

The sequencing depth required per sample is commonly much lower in metabarcoding studies than in shotgun sequencing studies (e.g. Srivathsan *et al.* 2015; Stat *et al.* 2017), and in metabarcoding studies it is (economically) feasible to sequence tens, hundreds, or even thousands of samples per sequencing run. To allow pooling and parallel sequencing of this magnitude, different molecular labelling systems have been developed. For metabarcoding studies, the addition of sample-specific identifiers to PCR amplicons can be achieved either as nucleotide tags during the metabarcoding PCR, or as library indices when converting amplicons into sequencing libraries.

Metabarcoding approaches can be divided into three overall strategies for adding nucleotide tags and library indices (Taberlet *et al.*2018) (Fig. 2):

1. The 'one-step PCR' approach in which sample DNA extracts are amplified and built into sequence libraries in one reaction. Here, metabarcoding primers carry sequencing adapters and library indices, referred to as 'fusion primers' (Fig. 2B). This approach is used in e.g. Kozich et al. (2013), Elbrecht and Leese (2015), Sickel et al. (2015), Grealy et al. (2016), Berry et al. (2017), Elbrecht et al. (2017), Hardy et al. (2017), Seersholm et al. (2018) and Bessey et al. (2020). In the one-step PCR approach, each PCR replicate or sample is a sequencing library and as such is returned as a separate fastq file following sequencing. It should be noted that a few studies modify this approach by adding nucleotide tags to the fusion primers instead of library indices (e.g. Elbrecht & Steinke 2018). When doing that, each PCR replicate is not an individual sequencing library.
2. The 'two-step PCR' approach in which sample DNA extracts are PCR-amplified with two primer sets. In the primary reaction metabarcoding primers carry 5' sequence overhangs of ca. 33-34 nucleotides in length and no nucleotide tags. The sequence overhangs allow the resulting amplicons to be targeted by the second round of primers, which carry sequencing adapters and indices (Fig. 2C). Most commonly, two consecutive PCRs are carried out, such as in Miya et al. (2015), de Vere et al, (2017), Galan et al. (2017), Kaunisto et al. (2017), Swift et al. (2018) and Vesterinen et al. (2018). However, a few studies carry out only one reaction with the two primer sets, such as Clarke et al. (2014a). The two-step PCR approach is based on Illumina's 16S rRNA system originally developed for microbiome studies (www.illumina.com). In the two-step approach, each PCR replicate is an individual sequencing library and as such is returned as a separate fastq file following sequencing. It should be noted that a few studies modify the two-step PCR approach to include nucleotide labelling in the first PCR, see Kitson et al. (2018).
3. The 'tagged PCR' approach, in which sample DNA extracts are PCR amplified with metabarcoding primers that carry 5' nucleotide tags. The individually tagged PCR products are pooled, and ligation-based library preparation is carried out on pools of 5' tagged amplicons. The ligated adapters can themselves contain indices, which eliminates the need for a second PCR step (e.g. Thomsen *et al.* 2016; Carøe & Bohmann 2020), or the adapter ligation can be followed by a PCR step with indexed primers (e.g. Hope *et al.* 2014; Bohmann *et al.* 2018). This approach was first demonstrated by Binladen et al. (2007) on the 454 FLX platform and has been since been used in e.g. Shehzad et al. (2012a), Hibert et al. (2013), Hope et al. (2014), Thomsen et al. (2016), Apothéloz-Perret-Gentil et al. (2017), Sigsgaard et al. (2017), Bakker et al. (2017), Kocher et al. (2017), Thomsen and Sigsgaard (2019) and Lynggaard et al. (2020) (Fig. 2D). In this approach, each library pool of PCR replicates is a sequencing library and is returned as a separate fastq file, each of which can contain data from a large number of PCR replicates.

All three main strategies offer the option to add extra nucleotides to shift PCR amplicons in relation to each other and thereby to increase sequence complexity on the flow cell ('heterogeneity spacers', see e.g. De Barba et al. 2014; Elbrecht & Leese 2015; Bohmann et al. 2018). Note that given the inconsistent use of terminology in the metabarcoding literature, for clarity, we use the original term for nucleotide tags in amplicon sequencing as used by Binladen *et al* . (2007) and Illumina's terminology to describe the nucleotide reads that are used to demultiplex sequencing libraries, the i5 and i7 index reads. That is, 5' nucleotide tags are sequenced with the metabarcoding marker and primers in the Illumina sequencing read 1 (and read 2 for paired-end sequencing), while library indices are sequenced as separate index reads, i.e. if dual-indexing is performed as i5 and i7 reads (Fig. 2A) (https://support.illumina.com).

In this article, we discuss the three main metabarcoding strategies. One approach not mentioned here is library preparation on individual unlabelled PCR products through a ligation-based library preparation protocol with or without an index PCR step. However, such ligation based protocol would entail several steps on each PCR product, such as end-repair and ligation of adapters (e.g. carrying indices such as in Illumina's TruSeq Nano DNA Library Prep kit, see Zizka et al (2019). The reason that we do not consider this approach a main metabarcoding strategy is due to low reported use of this method, its high cost and workload and thereby limited throughput (Zizka *et al.* 2019).

Pros and cons of metabarcoding approaches

The ability to tag and index amplicons to fully harvest the power of high-throughput sequencing comes at a price; the labelling and pooling of hundreds of PCR replicates is highly complex and entails costs associated with preventing, detecting, and eliminating errors and biases. None of the metabarcoding approaches presented here is perfect; rather each of them has pros and cons. Below, we outline the advantages and disadvantages, specifically addressing issues related to cross-contamination risk, PCR amplification efficiency, chimera formation, tag-jumping, index-misassignment, cost, and workload. The issues associated with each metabarcoding strategy are important to keep in mind for choosing a metabarcoding strategy and for designing laboratory workflows and interpreting results.

Cross-contamination risk

During the metabarcoding PCR (here specified as the PCR in which the metabarcoding marker is targeted), relatively short DNA sequences (typically <300 bp) are enriched through amplification. Especially when targeting trace amounts of DNA, PCR amplification can be highly susceptible to contamination and thereby to false positives. The risk of contamination when preparing metabarcoding PCRs is the same no matter which of the three overall metabarcoding approaches is used. Moreover, regardless of the metabarcoding strategy employed, cross-contamination can happen between nucleotide tagged and indexed primer stocks (which are delivered at very high molarity). The risk of this happening will be similar between the strategies and will depend on the number of samples and the chosen setup within the employed strategy. In the following, rather than discussing primer contamination, we will focus on how the three main metabarcoding approaches differ in risk of cross-contamination between PCR products after the metabarcoding PCR.

In the one-step PCR approach (Fig. 2B) and the tagged PCR approach (Fig. 2D), PCR products are labelled during the metabarcoding PCR amplification. In the one-step PCR approach, the metabarcoding PCR is carried out with primers that target the selected marker and carry both sequencing adapters and indices. This way, the indexed PCR products can be immediately sequenced following this one PCR step (Fig. 2B). If the indexed ready-to-sequence libraries are to be pooled into one pool before sequencing, then cross-contamination between indexed amplicon libraries is obviously not of concern. However, if more sequencing pools are made in which the same index combinations occur across multiple samples, then cross-contamination between the sequencing pools can be an issue. A solution is to process them in separate sequencing run batches to avoid cross-contamination. In the tagged PCR approach, amplicons will be 5' nucleotide tagged following the metabarcoding PCR, which means that cross-contamination between tagged PCR products is not of concern. However, until the amplicon pools are indexed during library preparation there is a risk of cross-contamination between amplicon pools if the same tag combinations are used in different amplicon pools (Schnell *et al.*2015). Some laboratories do not reuse tag-primer combinations to further reduce contamination risk (see Murray*et al.* 2015).

In the two-step approach, sample-specific labelling is not carried out during the metabarcoding PCR. This creates a risk of cross-contamination between unlabelled PCR products when handling them prior to the second PCR (Zizka *et al.*2019). Therefore, this metabarcoding approach has the greatest theoretical risk of cross-contamination between PCR amplicons (Fig. 2C). It is worth mentioning that a few studies adopt modifications of the two-step approach that eliminates this kind of cross-contamination. One is to include nucleotide labelling in the first PCR, see Kitson et al. (2018), and the other is to carry out both of the two PCRs, i.e. to include both two primer sets, in the same reaction, see for example Clarke et al. (2014a).

Irrespective of the chosen approach, cross-contamination can be detected and filtered out by including sample replicates, PCR replicates, and positive and negative controls. Thus, these should be included in the laboratory workflow and sequencing (e.g. Bista*et al.* 2017). An important measure that enables one to filter out potential contamination during data processing is to use different nucleotide tag and/or library index combinations on each sample's individual PCR replicates as this will allow for restrictive sequence processing across each sample's PCR replicates (Alberdi *et al.*2018).

PCR amplification

PCR amplification introduces biases, such as primer biases and errors, such as nucleotide substitutions and chimeras (e.g. Polz & Cavanaugh 1998; Haas *et al.* 2011; Murray *et al.*2015; Piñol *et al.* 2015). Two of the three main metabarcoding strategies allow practitioners to carry out only a single PCR step before sequencing, namely the one-step PCR with fusion primers approach and the tagged metabarcoding PCR approach in which PCR-free library building is carried out (Fig. 2B and D). Because an extra PCR step adds an additional risk of introducing errors, these two approaches offer an advantage over the two-step PCR method and the tagged PCR approach in which the workflow includes an index PCR step (Fig. 2C and D).

Apart from minimizing the number of PCR steps, the 5' nucleotide additions to metabarcoding primers should be considered. Bulk sample and eDNA extracts consist of complex mixtures of DNA from a large number of organisms, which in the case of eDNA can be degraded (Taberlet *et al.*2012a). With such DNA extracts, the primers are faced with the task of amplifying (trace) copy number target DNA from different taxa (Taberlet *et al.*2012b) potentially distorted by primer biases, inhibitors and potentially abundant predator or host DNA (e.g. Deagle *et al.* 2014; Clarke *et al.* 2014b; Murray *et al.* 2015). To add to this, nucleotide additions to primers can decrease PCR efficiency (Schnell *et al.*2015; Murray *et al.* 2015).

The three main metabarcoding strategies have different lengths of nucleotide additions on the 5'-end of metabarcoding primers. The longest 5'-nucleotide additions are found in the one-step PCR approach where up to 60 nucleotides (sequence adapters and indices) are added to one or both of the primers, making the complete primer often over 80 bp long (e.g. Elbrecht & Leese 2015). In the two-step PCR approach (Fig. 2C), the sequence overhangs on the metabarcoding primers used in the first PCR are approximately half the length of the fusion primers, e.g. 33-34 nucleotides, if using Illumina® Nextera Indices. The tagged PCR approach has the shortest nucleotide additions to the metabarcoding primers (Fig. 2D) with tags of typically 5-10 nucleotides in length (Coissac 2012; De Barba *et al.* 2014; e.g. Alberdi *et al.* 2018). The long additions to the metabarcoding primers cause a decrease in PCR efficiency (Murray *et al.* 2015) and in line with this, the two-step PCR approach has been shown to have a marginal increase in detection of taxa as compared to the one-step fusion primer approach (Zizka *et al.* 2019). Even the short nucleotide additions in the tagged PCR approach have been shown to decrease PCR efficiency (Schnell *et al.*2015). Thus, no method is free of decreased PCR efficiency caused by the nucleotide additions to 5'-end of metabarcoding primers. However, it has to our knowledge not been formally tested whether - and to what extent - the shorter nucleotide tag additions in the tagged PCR approach offers greater PCR efficiency and taxonomic detection than the two other approaches, and thereby it can only be speculated that it is the most sensitive when it comes to detection of taxa in low abundance amongst the three main approaches. Regardless of metabarcoding strategy, we stress the importance of optimising PCR amplifications (usually by qPCR) to detect PCR inhibition, identify samples with low template quantity and track PCR efficiency issues (Murray *et al.*2015; Yang *et al.* 2021).

Theoretically, the reduced PCR efficiency in the one-step and two-step PCR approaches caused by the long overhangs on primers might be counteracted by spiking the PCRs with metabarcoding primers without any 5' attachments (e.g. Murray*et al.* 2015). However, this has been shown to have modest PCR efficiency improvements for the one-step approach (e.g. Murray*et al.* 2015). Alternatively, a pre-enrichment before the metabarcoding PCR can be carried out, i.e. running a PCR with metabarcoding primers (with no nucleotide additions) prior to the metabarcoding PCR as done in Zizka et al. (2019) and Elbrecht & Steinke (2018) for the one-step PCR approach. However, this not only introduces another PCR amplification step, but can increase the risk of cross-contamination between PCR products due to the initial unlabelled PCR amplification step (e.g. Murray *et al.* 2015).

Apart from the length of the nucleotide additions, it has been investigated whether differences in nucleotide tag sequences can result in biases in the tagged PCR approach. Although one study shows that such tag bias is an issue (O'Donnell *et al.*2016), other studies show that it is not (Leray & Knowlton 2017; Yang *et al.* 2021). If tag bias does exist, it should theoretically be minimised if different tags are used on each sample's PCR replicates.

*Chimeras & tag-jumps*Chimeras can be formed during all PCR steps in any metabarcoding workflow (Fig. 2B-D). Chimeras are sequences that consist of two or more different template sequences, and the majority are thought to result from incomplete primer extension during the elongation phase of the PCR cycle (Meyerhans*et al.* 1990; Wang & Wang 1997; Judo *et al.* 1998; Shin*et al.* 2014). The probability of chimera formation increases when similar template sequences are amplified in the same PCR reaction (e.g. Judo *et al.* 1998; Smyth *et al.* 2010, but see also Fonseca*et al.* 2012), such as during the metabarcoding PCR (Fig. 2B-D) or during the index PCR-amplification of pools of tagged amplicons (Fig. 2D). There are different consequences of chimeric sequences depending on where they arise. If they are created during a PCR-amplification of a single sample's DNA extract, the chimeras will be intra-sample chimeras, which can be falsely interpreted as novel taxa and erroneously inflate measures of diversity. If, on the other hand, chimeras are created during a PCR-amplification of pooled tagged amplicons, such as in the tagged PCR approach (Fig. 2D), the chimeras may be inter-sample chimeras, which can result in tag-jumps and false attribution of amplicon sequences to samples (Schnell *et al.*2015). This can also lead to false positives and inflation of diversity.

All metabarcoding approaches are prone to intra-sample chimeras. However, as chimera formation increases when similar sequences are amplified in the same PCR reaction (e.g. Judo *et al.* 1998; Smyth *et al.* 2010), the use of metabarcoding primers with long 5' overhangs, as in the one-step and two-step approaches, might be more prone to chimera formation since they carry long and similar sequences at the 5' end of the primers. However, this hypothesis requires testing. Intra-sample chimeras can be reduced by limiting the number of PCR cycles (Haas *et al.* 2011). Also, if samples are subjected to multiple, independent PCRs, chimeras can be filtered out by keeping only sequences that occur in multiple PCR replicates, the ‘restrictive approach-described in Alberdi et al, (2018). Chimera detection programmes such as UCHIME (Edgar *et al.* 2011) can be used for further clean-up.

Inter-sample chimeras can cause havoc in metabarcoding studies. They can only occur in the tagged PCR approach where library build is carried out on pooled tagged amplicons from different samples (Fig. 2D). Here, tag-jumps can create sequences with new combinations of the nucleotide tags used in the amplicon pool (Schnell *et al.*2015). If the new combinations of tags are already used in the amplicon pool, it will cause false assignment of sequences to samples, which should be avoided at all costs (Schnell *et al.*2015; Esling *et al.* 2015). Such tag-jumps can also have the consequence that negative controls are seemingly not negative following bioinformatic sorting of sequences to samples. It should be noted that tag-jumps can also occur due to T4 DNA Polymerase activity in the blunt-ending step during library preparation, as demonstrated in library building for the Roche/454 sequencing platform (van Orsouw *et al.* 2007; Palkopoulou *et al.* 2016) and for the Illumina sequencing platform (Carøe & Bohmann 2020). The rate of tag-jumping has been estimated from ca. 2% to up to 49% of total sequences (Schnell *et al.* 2015; Esling *et al.* 2015; Carøe & Bohmann 2020). This broad range can be caused by factors affecting inter-sample chimera formation during the index PCR. For example, DNA template and primer concentration, PCR cycle number, and sequence similarity (e.g. Judo *et al.* 1998; Smyth *et al.* 2010; Carøe & Bohmann 2020). The range of tag-jump proportions highlights the unreliability of including an index PCR step in the tagged PCR approach.

To avoid tag-jumps in the tagged PCR approach, and thereby prevent false assignment of sequences to samples, it is important to refine index PCR parameters to decrease the likelihood of chimera formation - or better yet, to omit the index PCR step (Fig. 2D). Further, blunt-ending using T4 DNA Polymerase should be circumvented during library preparation (Schnell *et al.* 2015; Palkopoulou *et al.* 2016; Carøe & Bohmann 2020). If both T4 DNA Polymerase blunt-ending and index PCR are eliminated during library preparation of pools of tagged amplicons, tag-jumps can practically be eliminated (Carøe & Bohmann 2020).

If the library preparation protocol contains a T4 DNA blunt-ending step and/or an index PCR step, and thereby can be assumed to generate tag-jumps, they can be detected and removed by using 'twin-tags' during the original PCRs (e.g. F1-R1, F2-R2,...), because tag-jumped sequences would then produce non-twinned tag combinations not used in the set-up (e.g. F1-R2, F2-R3,...) (e.g. Schnell *et al.* 2015; Yang *et al.* 2021). However, using twin tags comes at the price of buying many more versions of tagged primers and building more libraries (Schnell *et al.*2015). If twin tags are not used, chimera removal software can remove some chimeric sequences carrying false combinations of used tags (Schnell *et al.*2015).

The extent of tag-jumping and spillover of taxa between samples can be detected through inclusion of positive controls consisting of synthetic oligos or taxa not expected to occur in the dataset. However, note that such controls do not enable confident elimination of false positives caused by tag-jumps. The extent of tag-jumping can also be assessed by comparing all observed combinations of used tags to all originally used tag combinations (Schnell *et al.*2015; Zepeda Mendoza *et al.* 2016).

*Misassignment of library indices*Incorrect assignment of indices between pooled libraries can cause sequence reads to be incorrectly assigned to libraries. Misassigned indices have been attributed to the formation of mixed clusters on the sequencing flow cell, i.e. clusters originating from two different template molecules or clusters growing into each other, to low levels of free index primers present in the sequence library and to bulk amplification of pooled libraries (Nelson*et al.* 2014; Sinha *et al.* 2017; Vodak *et al.* 2018; Costello *et al.* 2018; Valk *et al.* 2019). Regardless of how index misassignment occurs, if it occurs in metabarcoding studies it can cause incorrect assignment of amplicon sequences to libraries, which can cause incorrect assignment of sequences to samples and false positives. This phenomenon can affect all three metabarcoding approaches (Fig. 2). To avoid index misassignment it is recommended to dual-index libraries with unique library index combinations (Kircher *et al.*2012; Sinha *et al.* 2017),*www.illumina.com*). Further, stringent bead purification (or size selection) can remove free adapters/primers from the libraries (Owens *et al.* 2018). The labelling in the different metabarcoding approaches further allows for accounting for potential incorrect assignment of sequences to libraries. In the tagged PCR approach, unique tagging of PCR replicates across all pooled libraries can be used to account for (and detect) index misassignment. However, this can be costly. In the one-step PCR with fusion primers approach, a tweaked protocol where nucleotide tags are used instead of i7 and i5 of library indices (e.g. Elbrecht & Steinke 2018) creates one single library that is thereby free of index misassignment. As with tag-jumping, the extent of incorrect assignment of indices and spillover of taxa between samples can be detected through inclusion of positive controls consisting of taxa not expected to occur in the data set and by comparing all observed to all used combinations of used indices when demultiplexing libraries.

It is important not to mistake tag-jumping, index misassignment or cross-contamination between PCR products with cross-contamination of the primers themselves. Due to the high concentration of primers upon synthesis, cross-contamination (e.g. by aerosols) can manifest itself as low numbers of sequence reads and could be misinterpreted as tag-jumps or index-bleeding. Due to the risk of primer cross-contamination, some laboratories avoid ordering primers in 96-well plates. There are anecdotal reports that primer contamination can also occur at primer synthesis (or purification). As mentioned, the risk of cross-contamination between nucleotide tagged primer stocks and indexed primer stocks, which could e.g. occur during resuspension of primers, will generally be the same no matter which of the three overall metabarcoding approaches is used. In the first PCR step in the two-step PCR approach, the primers are unlabelled and any cross-contamination that might occur will not have consequences.

Cost Metabarcoding primers in the tagged and one-step PCR approaches have to be labelled with either nucleotide tags or indices, whereas the metabarcoding primers in the two-step approach are generally not individually labelled. Due to the different labelling systems in the three primary metabarcoding approaches, there are different costs associated with them.

The fusion primers for the one-step PCR approach are the most expensive metabarcoding primers amongst the three approaches. This is (i) because differently indexed versions are purchased for each metabarcoding primer set and (ii) because the increased oligo length results in lower yield of the full length product. If unique matching indices are used to account for index misassignment, one-step PCR can become increasingly expensive for larger scale studies. However, this needs to be factored against the potential cost of repeating runs due to artifacts and contamination, and the fact that only a single PCR step is needed to go from sample extract to library.

In the tagged PCR approach (Fig. 2D), the metabarcoding primers are relatively inexpensive compared to the one-step PCR fusion primers as they only add 5' tags of 5-10 nucleotides in length. However, as with the one-step PCR approach, these need to be purchased in many tagged editions for each metabarcoding primer set. Furthermore, if tag-jumping is to be taken into account by only using each tag once in a library amplicon pool, e.g. by only amplifying with twin forward and reverse tags, then metabarcoding primer sets have to be ordered in many differently labelled editions (Schnell *et al.* 2015). To keep costs down, this needs to be balanced by pooling fewer PCR products into each library and thereby creating more sequence libraries (Fig. 2D). However, if a library preparation protocol is used that does not create tag-jumps, tags can be freely combined, which lowers the number of tagged primers that must be purchased (Schnell *et al.*2015; Carøe & Bohmann 2020). In contrast to the other two metabarcoding approaches, the tagged PCR approach includes library preparation on pools of amplicons, and the cost of this therefore has to be taken into account. This can however be kept low if a protocol that does not generate tag-jumps is used and only a few libraries have to be made.

If a large number of metabarcoding primer sets are used, the two-step approach offers a relatively inexpensive solution. In the two-step PCR approach, the metabarcoding primers are generally synthesized with 5' tails containing no tags or indices. This means that the same primer set can be used across multiple samples and projects. This has the benefit that trying out new metabarcoding primer sets does not entail buying many labelled versions of the metabarcoding primer sets, as it does in the other metabarcoding approaches (Fig. 2B-D). However, the second primer set in the two-step PCR approach is costly as it has to include both the sequence complementary to the sequence overhang, the sequence adapters and the library indices (Fig. 2C). It is worth noting that, just as with the one-step PCR approach, many labelled index primers will have to be purchased if twin dual-indices are used to account for incorrect assignment of indices to libraries. This second primer set is, however, applicable across different metabarcoding primer sets and can thereby be used across many metabarcoding studies.

Laboratory workload The one-step PCR approach is without doubt the quickest method for generating sequence-ready libraries, as it only requires a single PCR-step to achieve both amplification and library preparation of the metabarcoding amplicons (Fig. 2B), and it has been used in the field to rapidly turn-around sequence data. The workload for the two-step PCR approach and the tagged PCR approach depends, to some extent, on how many sample extracts and PCR replicates are to be processed. If it is a relatively high number, the tagged PCR approach is the quickest due to the library build being performed on pooled amplicons rather than through a PCR step on individual PCR products. However, as with all molecular biological workflows, carefully organised liquid handling and automation provide solutions to high-throughput studies.

Choosing a metabarcoding approach

It is clear that there is no such thing as a perfect metabarcoding sample-labelling approach, and that choosing which one is right for a given study or lab should be an informed trade-off of pros and cons balanced to the needs. Within metabarcoding studies, those needs can range widely.

Metabarcoding studies range from those that look for one or a few taxa within sample units (e.g. Bohmann*et al.* 2018) to studies that look for many taxa within sample units (e.g. Seersholm *et al.* 2018), and sample numbers can range from tens (e.g. Elbrecht*et al.* 2017), to hundreds (Rodgers*et al.* 2017; e.g. Galan *et al.* 2017) or even thousands (e.g. Schnell *et al.* 2018; Ji *et al.* 2020). The research question and experimental set-up can require taxonomic identifications to be made within individual samples (e.g. Coghlan*et al.* 2012), while in other studies, taxonomic identifications from pools of individual samples or from a number of samples within e.g. a geographic location is the goal (e.g. Grealy *et al.* 2016; Schnell *et al.* 2018). Sample types can range from bulk specimen samples consisting of high quality DNA from pools of entire organisms (e.g. Tang*et al.* 2015) to environmental samples in which DNA from target organisms can be fragmented and scarce (e.g. Stat*et al.* 2017). Furthermore, studies differ in how many metabarcoding primer sets are used - from only one (e.g. Bohmann *et al.* 2011; Drinkwater *et al.* 2018) to several (e.g. De Barba *et al.* 2014; Drummond *et al.* 2015; Zhang *et al.* 2018). Furthermore, the budget for a metabarcoding project will differ between studies, and lastly so will whether the metabarcoding primers are to be used in future studies. Lastly, some applications of metabarcoding, such as biosecurity or forensics, will necessitate a 'high bar' for data fidelity and controls.

A multitude of combinations of the above metabarcoding study parameters exist, and as witnessed by this article, the significance of the pros and cons of the metabarcoding approaches will differ with them. For example, while the tagged PCR approach (Fig. 2D) might be more sensitive to low-abundance templates, the one-step PCR offers a quick turnaround (Fig. 2B). However, this comes at the cost of buying long fusion primers and is only worthwhile if the metabarcoding primers are to be used again.

When choosing a metabarcoding approach, the need for future multiplexing of the metabarcoding primers should be considered. That is, to use several metabarcoding primer sets that target different markers and taxonomic groups in individual PCR reactions to simultaneously screen for many taxonomic groups within the same reaction, and thereby keep costs and work load at a minimum (e.g. De Barba*et al.* 2014). For this, the nucleotide tagged primers in the tagged PCR approach should theoretically be the most applicable, whereas the long additions to the metabarcoding primers in the one-step and two-step PCR approaches will be far less conducive to multiplexing due to the extensive sequence homology.

Lastly, it should be noted that whatever metabarcoding strategy is chosen, it should be clear from the present article that one should not change workflows within an experiment. Moreover, there is some justified concern within the metabarcoding community that the nuances in metabarcoding workflows makes inter-lab comparison difficult (e.g. Murray *et al.* 2015; Zizka *et al.* 2019; Blackman *et al.* 2019).

PerspectivesAll metabarcoding strategies can generate robust data. However, like all laboratory workflows if they are not executed well or are inappropriate for the application, they may lead to flawed data. We advocate that just because PCR is a relatively simple method it does not mean that metabarcoding is simple, and there are many traps in metabarcoding workflows that can trip-up new users. Here, we have presented an overview of the three main metabarcoding strategies for assessment of biodiversity on Illumina sequencing platforms, and the downstream consequences for the resulting data with regards to cross-contamination risk, PCR amplification efficiency, chimera formation, tag-jumping, index-misassignment, cost, and workload. In doing so we wish to enable researchers and practitioners to make an informed choice of which metabarcoding strategy is best suited for their specific study. Ultimately, this is to avoid the worst case scenario, generation of unusable data and wasting a considerable amount of time and money, or even worse making wrong conclusions due to flawed data.

Metabarcoding of environmental DNA has some commonalities with the field of ancient DNA in which low quality and quantity of target DNA is also targeted amongst non-target (and potentially more abundant) templates. In the early days of ancient DNA studies, PCR-based techniques (including amplifying already amplified DNA to enhance signals) were used, which caused authentication issues, as amplification of modern templates was mistaken for true ancient signals. This was followed by urgent calls for precautions to ensure reliability and authenticity of ancient DNA sequences (e.g. Cooper & Poinar 2000; Pääbo *et al.* 2004). Also similarly to the field of ancient DNA, the take-home message should be that metabarcoding is becoming a self-critical and self-correcting field in which technical reliability is promoted and rewarded, with the long-term benefit of uptake by stakeholders who will employ metabarcoding for environmental management. Reputational setbacks as the result of practitioners not executing their metabarcoding workflows well will likely resonante across a variety of biomonitoring, forensic and bioseurity applications.

We thus stress the importance of being informed about the pros and cons of the chosen metabarcoding approach with regards to cross-contamination risk, PCR amplification efficiency, chimera formation, tag-jumping, index-misassignment, cost, and workload and to include appropriate quality assurance and quality control measures. This will help ensure that the generated data will facilitate informed data analysis and interpretation. Therefore, we advocate that metabarcoding publications should include detailed information about the metabarcoding strategy and how its challenges have been taken into account in the laboratory, data processing, and interpretation of results. Furthermore, it may be appropriate to eventually develop a set of metabarcoding guidelines similar to the MIQE guidelines for qPCR (Bustin *et al.*2009), ultimately further increasing the power and reliability of metabarcoding.

Acknowledgements

KB was supported by Independent Research Fund Denmark, DFF grant 5051-00140 and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856488). AJD was supported by the U.K. Natural Environment Research Council (NERC) grants NE/S006958/1, NE/M02086X/1, NE/S005560/1, and NE/S000291/1. SC benefitted from NERC Highlight Topic funding grant NE/N006216/1 and SC acknowledges support from NERC grant NE/N003756/1 and NE/M02086X/1. CC was supported by Carlsberg Foundation (CF18-1110, 'Archives'). FL, VE and SC are members of EU COST Action DNAqua-Net (CA15219). Moreover, the authors would like to thank their collective research groups whose laboratory and bioinformatic experiences over many years contributed to the ideas, initiation and completion of this manuscript.

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contributions

KB and SC conceived the idea for the manuscript, KB drafted the manuscript and figures and all authors contributed with edits and comments.

References

Aizpurua O, Budinski I, Georgiakakis P *et al.* (2017) Agriculture shapes the trophic niche of a bat preying on multiple pest arthropods across Europe: evidence from DNA metabarcoding. *Molecular ecology.*

Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in ecology and evolution / British Ecological Society.*

Apothéloz-Perret-Gentil L, Cordonier A, Straub F *et al.* (2017) Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular ecology resources*, **17**, 1231–1242.

Aylagas E, Borja Á, Muxika I, Rodríguez-Ezpeleta N (2018) Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks.*Ecological indicators*, **95**, 194–202.

Bakker J, Wangensteen OS, Chapman DD *et al.* (2017) Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact.*Scientific reports*, **7**, 16886.

Berry TE, Osterrieder SK, Murray DC *et al.* (2017) DNA metabarcoding for diet analysis and biodiversity: A case study using the endangered Australian sea lion (Neophoca cinerea). *Ecology and evolution*, **7**, 5435–5453.

Bessey C, Jarman SN, Berry O*et al.* (2020) Maximizing fish detection with eDNA metabarcoding.*Environmental DNA*.

Binladen J, Gilbert MTP, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PloS one*, **2**, e197.

Bista I, Carvalho GR, Walsh K*et al.* (2017) Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity.*Nature communications*, **8**, 14087.

Blackman RC, Mächler E, Altermatt F *et al.* (2019) Advancing the use of molecular methods for routine freshwater macroinvertebrate biomonitoring – the need for calibration experiments. *Metabarcoding Metagenom.*, **3**, 49–57.

Bohan DA, Vacher C, Tamaddoni-Nezhad A *et al.* (2017) Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. *Trends in ecology & evolution*, **32**, 477–487.

Bohmann K, Gopalakrishnan S, Nielsen M *et al.* (2018) Using DNA metabarcoding for simultaneous inference of common vampire bat diet and population structure.*Molecular ecology resources.*

Bohmann K, Monadjem A, Lehmkuhl Noer C *et al.* (2011) Molecular diet analysis of two african free-tailed bats (molossidae) using high throughput sequencing.*PloS one*, **6**, e21441.

Bustin SA, Benes V, Garson JA*et al.* (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical chemistry*, **55**, 611–622.

Calvignac-Spencer S, Merkel K, Kutzner N *et al.* (2013) Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity.*Molecular ecology*, **22**, 915–924.

Carøe C, Bohmann K (2020) Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples. *Molecular ecology resources*,**20**, 1620–1631.

Clarke LJ, Czechowski P, Soubrier J (2014a) Modular tagging of amplicons using a single PCR for high-throughput sequencing. *Molecular ecology.*

Clarke LJ, Soubrier J, Weyrich LS, Cooper A (2014b) Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular ecology resources*, **14**, 1160–1170.

Coghlan ML, Haile J, Houston J *et al.* (2012) Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. *PLoS genetics*, **8**, e1002657.

Coissac E (2012) OligoTag: a program for designing sets of tags for next-generation sequencing of multiplexed samples. *Methods in molecular biology* , **888**, 13–31.

Cooper A, Poinar HN (2000) Ancient DNA: Do It Right or Not at All. *Science*, **289**, 1139–1139.

Costello M, Fleharty M, Abreu J *et al.* (2018) Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC genomics*, **19**, 332.

Creer S, Deiner K, Frey S*et al.* (2016) The ecologist's field guide to sequence-based identification of biodiversity. *Methods in ecology and evolution / British Ecological Society*, **7**, 1008–1018.

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology letters*,**10**.

De Barba M, Miquel C, Boyer F*et al.* (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet.*Molecular ecology resources*, **14**, 306–323.

Drinkwater R, Schnell IB, Bohmann K *et al.* (2018) Using metabarcoding to compare the suitability of two blood-feeding leech species for sampling mammalian diversity in North Borneo. *Molecular ecology resources.*

Drummond AJ, Newcomb RD, Buckley TR *et al.* (2015) Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, **4**, 46.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* , **27**, 2194–2200.

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, Steinke D (2018) Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring.*Freshwater biology*, **5**, 1.

Elbrecht V, Vamos EE, Meissner K, Aroviita J, Leese F (2017) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in ecology and evolution / British Ecological Society*, **8**, 1265–1275.

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

Fonseca VG, Nichols B, Lallias D *et al.* (2012) Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic acids research*, **40**, e66.

Galan M, Pons J-B, Tournayre O *et al.* (2017) Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Molecular ecology resources.*

Gous A, Swanevelder DZH, Eardley CD, Willows-Munro S (2019) Plant–pollinator interactions over time: Pollen metabarcoding from bees in a historic collection.*Evolutionary applications*, **12**, 187–197.

Grealy A, Douglass K, Haile J*et al.* (2016) Tropical ancient DNA from bulk archaeological fish bone reveals the subsistence practices of a historic coastal community in southwest Madagascar. *Journal of archaeological science*,**75**, 82–88.

Haas BJ, Gevers D, Earl AM*et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*,**21**, 494–504.

Hardy N, Berry T, Kelaher BP*et al.* (2017) Assessing the trophic ecology of top predators across a recolonisation frontier using DNA metabarcoding of diets.*Marine ecology progress series*, **573**, 237–254.

Hibert F, Taberlet P, Chave J*et al.* (2013) Unveiling the diet of elusive rainforest herbivores in next generation sequencing era? The tapir as a case study. *PloS one*, **8**, e60799.

Hope PR, Bohmann K, Gilbert MTP *et al.* (2014) Second generation sequencing and morphological faecal analysis reveal unexpected foraging behaviour by Myotis nattereri (Chiroptera, Vespertilionidae) in winter. *Frontiers in zoology*,**11**, 39.

Jarman SN, Berry O, Bunce M (2018) The value of environmental DNA biobanking for long-term biomonitoring. *Nature Ecology & Evolution*, **2**, 1192–1193.

Ji Y, Baker CCM, Popescu VD*et al.* (2020) Measuring protected-area outcomes with leech iDNA: large-scale quantification of vertebrate biodiversity in Ailaoshan reserve. *BioRXiv*, 2020.02.10.941336.

Judo MS, Wedel AB, Wilson C (1998) Stimulation and suppression of PCR-mediated recombination.*Nucleic acids research*, **26**, 1819–1825.

Kaunisto KM, Roslin T, Saaksjarvi IE, Vesterinen EJ (2017) Pellets of proof: First glimpse of the dietary composition of adult odonates as revealed by metabarcoding of feces. *Ecology and evolution*, **7**, 8588–8598.

Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research*, **40**, e3.

Kitson JJN, Hahn C, Sands RJ*et al.* (2018) Detecting host-parasitoid interactions in an invasive Lepidopteran using nested tagging DNA metabarcoding.*Molecular ecology.*

Kocher A, de Thoisy B, Catzeflis F *et al.* (2017) Evaluation of short mitochondrial metabarcodes for the identification of Amazonian mammals (O Gaggiotti, Ed,). *Methods in ecology and evolution / British Ecological Society*, **8**, 1276–1283.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and environmental microbiology*, **79**, 5112–5120.

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

Li J, Hatton-Ellis TW, Lawson Handley L *et al.* (2019) Ground-truthing of a fish-based environmental DNA metabarcoding method for assessing the quality of lakes. *The Journal of applied ecology*, **56**, 1232–1244.

Li F, Peng Y, Fang W *et al.* (2018) Application of Environmental DNA Metabarcoding for Predicting Anthropogenic Pollution in Rivers. *Environmental science & technology*, **52**, 11708–11719.

Lucas A, Bodger O, Brosi BJ*et al.* (2018) Floral resource partitioning by individuals within generalised hoverfly pollination networks revealed by DNA metabarcoding.*Scientific reports*, **8**, 5133.

Lynggaard C, Yu DW, Oliveira G *et al.* (2020) DNA-based arthropod diversity assessment in amazonian iron mine lands show ecological succession towards undisturbed reference sites. *Frontiers in ecology and evolution*, **8**.

Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic acids research*, **18**, 1687–1691.

Miya M, Sato Y, Fukunaga T*et al.* (2015) MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society open science*,**2**, 150088.

Murray DC, Coghlan ML, Bunce M (2015) From benchtop to desktop: important considerations when designing amplicon sequencing workflows. *PloS one*, **10**, e0124671.

Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J (2014) Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys.*PloS one*, **9**, e94249.

O'Donnell JL, Kelly RP, Lowell NC, Port JA (2016) Indexed PCR Primers Induce Template-Specific Bias in Large-Scale DNA Sequencing Studies. *PloS one*,**11**, e0148698.

van Orsouw NJ, Hogers RCJ, Janssen A *et al.* (2007) Complexity Reduction of Polymorphic Sequences (CRoPS$^{\mathrm{TM}}$): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. *PloS one*, **2**, e1172.

Owens GL, Todesco M, Drummond EBM, Yeaman S, Rieseberg LH (2018) A novel post hoc method for detecting index switching finds no evidence for increased switching on the Illumina HiSeq X. *Molecular ecology resources*, **18**, 169–175.

Paabo S, Poinar H, Serre D*et al.* (2004) Genetic analyses from ancient DNA. *Annual review of genetics*, **38**, 645–679.

Palkopoulou E, Baca M, Abramson NI *et al.* (2016) Synchronous genetic turnovers across Western Eurasia in Late Pleistocene collared lemmings. *Global change biology*, **22**, 1710–1721.

Pinol J, Mir G, Gomez-Polo P, Agusti N (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular ecology resources*, **15**, 819–830.

Pochon X, Bott NJ, Smith KF, Wood SA (2013) Evaluating detection limits of next-generation sequencing for the surveillance and monitoring of international marine pests.*PloS one*, **8**, e73935.

Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental microbiology*, **64**, 3724–3730.

Pont D, Rocle M, Valentini A*et al.* (2018) Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation.*Scientific reports*, **8**, 10361.

Pont D, Valentini A, Rocle M*et al.* (2021) The future of fish-based ecological assessment of European rivers: from traditional EU Water Framework Directive compliant methods to eDNA metabarcoding-based approaches. *Journal of fish biology*, **98**, 354–366.

Quemere E, Hibert F, Miquel C*et al.* (2013) A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. *PloS one*, **8**, e58971.

Razgour O, Clare EL, Zeale MRK *et al.* (2011) High-throughput sequencing offers insight into mechanisms of resource partitioning in cryptic bat species.*Ecology and evolution*, **1**, 556–570.

Rodgers TW, Xu CCY, Giacalone J (2017) Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: Evidence from a known tropical mammal community.*Molecular ecology.*

Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular ecology resources*, **15**, 1289–1303.

Schnell IB, Bohmann K, Schultze SE *et al.* (2018) Debugging diversity-a pan-continental exploration of the potential of terrestrial blood-feeding leeches as a vertebrate monitoring tool. *Molecular ecology resources.*

Seersholm FV, Cole TL, Grealy A *et al.* (2018) Subsistence practices, past biodiversity, and anthropogenic impacts revealed by New Zealand-wide ancient DNA survey.*Proceedings of the National Academy of Sciences of the United States of America.*

Seymour M, Edwards FK, Cosby BJ *et al.* (2020) Executing multi-taxa eDNA ecological assessment via traditional metrics and interactive networks. *The Science of the total environment*, **729**, 138801.

Shehzad W, McCarthy TM, Pompanon F *et al.* (2012a) Prey preference of snow leopard (Panthera uncia) in South Gobi, Mongolia. *PloS one*, **7**, e32104.

Shehzad W, Riaz T, Nawaz MA*et al.* (2012b) Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (Prionailurus bengalensis) in Pakistan. *Molecular ecology*, **21**, 1951–1965.

Shin S, Lee TK, Han JM, Park J (2014) Regional effects on chimera formation in 454 pyrosequenced amplicons from a mock community. *Journal of microbiology* ,**52**, 566–573.

Sickel W, Ankenbrand MJ, Grimmer G *et al.* (2015) Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach.*BMC ecology*, **15**, 20.

Sigsgaard EE, Nielsen IB, Carl H *et al.* (2017) Seawater environmental DNA reflects seasonality of a coastal fish community. *Marine biology*,**164**, 128.

Singer G, Fahner NA, Barnes J, McCarthy A, Hajibabaei M (2019) Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *bioRxiv*, 515890.

Sinha R, Stanley G, Gulati GS*et al.* (2017) Index Switching Causes “Spreading-Of-Signal” Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing.*bioRxiv*, 125724.

Smyth RP, Schlub TE, Grimm A*et al.* (2010) Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, **469**, 45–51.

Srivathsan A, Sha J, Vogler AP, Meier R (2015) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (Pygathrix nemaeus). *Molecular ecology resources*, **15**, 250–261.

Stat M, Huggett MJ, Bernasconi R *et al.* (2017) Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment.*Scientific reports*, **7**, 12240.

Stoeck T, Pan H, Dully V, Forster D, Jung T (2018) Towards an eDNA metabarcode-based performance indicator for full-scale municipal wastewater treatment plants.*Water research*, **144**, 322–331.

Swift JF, Lance RF, Guan X*et al.* (2018) Multifaceted DNA metabarcoding: Validation of a noninvasive, next-generation approach to studying bat populations.*Evolutionary applications*, **12**, 2175.

Taberlet P, Bonin A, Zinger L, Coissac E (2018) *Environmental DNA: For Biodiversity Research and Monitoring.* Oxford University Press.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012a) Environmental DNA. *Molecular ecology*, **21**, 1789–1793.

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

Tang M, Hardman CJ, Ji Y*et al.* (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in ecology and evolution / British Ecological Society*, **6**, 1034–1043.

Thomsen PF, Moller PR, Sigsgaard EE *et al.* (2016) Environmental DNA from Seawater Samples Correlate with Trawl Catches of Subarctic, Deepwater Fishes.*PloS one*, **11**, e0165252.

Thomsen PF, Sigsgaard EE (2019) Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. *Ecology and evolution.*

Valk T van der, van der Valk T, Vezzi F *et al.* (2019) Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies.

de Vere N, Jones LE, Gilmore T *et al.* (2017) Using DNA metabarcoding to investigate honey bee foraging reveals limited flower use despite high floral availability.*Scientific reports*, **7**, 42838.

Vesterinen EJ, Puisto AIE, Blomberg AS, Lilley TM (2018) Table for five, please: Dietary partitioning in boreal bats. *Ecology and evolution*, **8**, 10914–10937.

Vodak D, Lorenz S, Nakken S*et al.* (2018) Sample-Index Misassignment Impacts Tumor Exome Sequencing.

Wang GC, Wang Y (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes.*Applied and environmental microbiology*, **63**, 4645–4650.

Yang C, Bohmann K, Wang X*et al.* (2021) Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution.*

Zepeda Mendoza ML, Bohmann K, Carmona Baez A, Gilbert MTP (2016) DAMe: a toolkit for the initial processing of datasets with PCR replicates of double-tagged amplicons for DNA metabarcoding analyses. *BMC research notes*, **9**, 255.

Zhang GK, Chain FJJ, Abbott CL, Cristescu ME (2018) Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities.*Evolutionary applications*, **11**, 1901–1914.

Zizka VMA, Elbrecht V, Macher J-N, Leese F (2019) Assessing the influence of sample tagging and library preparation on DNA metabarcoding. *Molecular ecology resources.*

Zizka VMA, Geiger MF, Leese F (2020) DNA metabarcoding of stream invertebrates reveals spatio-temporal variation but consistent status class assessments in a natural and urban river. *Ecological indicators*, **115**, 106383.

figures/Fig1/Fig1-eps-converted-to.pdf

figures/Fig2/Fig2-eps-converted-to.pdf