JNCC Report

No. 699

Understanding and mitigating errors and biases in metabarcoding: an introduction for non-specialists

Mark Preston, Martin Fritzsche and Paul Woodcock

April 2022

ISSN 0963 8091

For further information please contact:
Joint Nature Conservation Committee
Monkstone House
City Road
Peterborough PE1 1JY
www.jncc.gov.uk

This report should be cited as:
Preston, M., Fritzsche, M. & Woodcock, P. 2022. Understanding and mitigating errors and biases in metabarcoding: an introduction for non-specialists. JNCC Report No. 699, JNCC, Peterborough, ISSN 0963-8091.

Author affiliation:
Mark Preston, Prismea Limited, Oxford, OX3 8SB
Martin Fritzsche, National Institute for Biological Standards and Control, South Mimms, EN6 3QG
Paul Woodcock, Joint Nature Conservation Committee, Peterborough, PE1 1JY

JNCC EQA Statement:
This report is compliant with the JNCC Evidence Quality Assurance Policy:
https://jncc.gov.uk/about-jncc/corporate-information/evidence-quality-assurance/

Acknowledgement:
This project was led and funded by the Defra DNA Centre of Excellence, which aims to transform Defra Group science to support policy and delivery by sharing and demonstrating the wider potential of existing methods; enabling the rapid development and trialling of new DNA based method applications; providing a focal point for leadership in Defra, UK and across other Government Departments; and building shared capability including skills and facilities.

Summary

DNA metabarcoding is a potentially powerful tool for increasing the breadth and efficiency of monitoring carried out by environmental public bodies. However, methods and data interpretation require skills across several fields, each with a diversity of complex and still-evolving methods. This has practical consequences for implementation by environmental public bodies, including **(i)** methods that would be suitable for routine monitoring may be adopted slowly or not at all, due to lack of confidence in results or difficulty in interpretation, and/or **(ii)** methods that are not suitable for the intended purpose may be used, due to lack of awareness of potential limitations.

To support non-experts in making more informed judgments on the development and interpretation of metabarcoding, we outline potential errors and biases that can occur, and how these can influence results. We also describe approaches to mitigation. The aim is not to critique metabarcoding or to evaluate it in relation to other options, and the report is not intended to provide non-specialists with the technical detail to design and carry out metabarcoding projects independently. Rather, we hope it will complement other resources in helping to give end-users a greater appreciation and awareness of potential problems and their mitigation.

Error and biases are discussed in terms of false positives (erroneous detection) and false negatives (erroneous non-detection), where and why these can arise in metabarcoding, and the potential impact on results. The following aspects of metabarcoding are covered: *Contamination* (excluding during initial collection in the field), *Degradation*, *PCR Amplification*, *Sequencing*, and *Bioinformatics*. We also stress that where differences in results between metabarcoding and more established methods occur, these do not necessarily mean an error in metabarcoding – discrepancies sometimes arise for other reasons (e.g. because the approaches sometimes sample communities in different ways, or because other methods also sometimes make mistakes).

We conclude by highlighting that many decisions in the implementation of metabarcoding can involve trade-offs. In particular, **(i)** actions to mitigate errors and biases tend to increase costs, and **(ii)** some mitigation strategies to reduce the risk of false positives can increase the risk of false negatives (and vice versa). The impact of false positives and false negatives for environmental public bodies will depend on the application, so it is not appropriate to provide prescriptive all-purpose recommendations. However, it is important to ensure clear and consistent communication between end-users of results and specialists so that the former appreciate potential limitations and the latter understand end-user uncertainties and requirements. This should occur throughout the process to guide study design, implementation, reporting and documenting results, and the final interpretation.

Contents

1 Introduction

Metabarcoding is a potentially powerful tool that can increase the breadth and efficiency of monitoring carried out by environmental public bodies. It involves using interspecific variation in particular regions of the genome (barcodes) to distinguish between species, and so can allow the species composition of a sample to be determined. This can include organisms collected directly as well as species that have shed DNA into the sample (environmental DNA, or eDNA). However, metabarcoding methods and data interpretation require skills across several fields, including ecology, microbiology, molecular technology, advanced statistics, and computer science. In each of these fields there are many possible approaches, and methods are often complex and still evolving. This has practical consequences for implementation by environmental public bodies, including **(i)** methods that would be suitable for routine monitoring may be adopted slowly or not at all, due to a lack of confidence in results or difficulty in interpretation, and/or **(ii)** methods that are not suitable for the intended purpose may be used, due to lack of awareness of potential limitations.

To support non-expert users in making more informed judgments over applying metabarcoding methods and interpreting results, this report outlines potential errors and biases that can occur in metabarcoding, and how these can influence results. Broadly speaking, we view **errors** as randomness in measurements around the true value and **bias** as a systematic error that means a measurement is consistently too big or small. Sources of potential error and bias are presented in Section 2 for each aspect of the metabarcoding process. Problems introduced during study site selection and sampling are not expressly considered (see e.g. Deiner *et al.* 2017; Liu *et al.* 2019 for more information on these topics).

We also describe some common approaches to mitigating potential errors and bias, though note that these are covered in more technical depth elsewhere (e.g. Alberdi *et al.* 2018; Furlan *et al.* 2020). In addition, we stress that mitigation decisions need to reflect practical considerations such as the resources available and the required accuracy of results, and that habitat-specific factors may mean certain errors are more problematic for some applications than others. As these considerations will differ for every study, the importance of some of the suggested mitigations will depend on the intended use.

The aim of this report is to give non-specialists with some existing understanding of metabarcoding a greater appreciation and awareness of potential problems and their mitigation. It is *not* to critique metabarcoding or to evaluate it as a monitoring tool by comparing with other options (and we stress that other methods such as morphological identification can also make mistakes, meaning that the true answer is not always clear – see Section 3). The report is also not intended to give the level of technical detail needed for non-specialists to design and carry out metabarcoding projects independently. Rather, we hope that it is accessible to end-users and provides a complement to related documents funded by the Defra DNA Centre of Excellence (Jones *et al.* 2020a, b) and to other resources introducing metabarcoding to non-specialists (e.g. Haenfling *et al.* 2017; Barsoum *et al.* 2018; Bruce *et al.* 2021). Collectively, these can help environmental public bodies build up the level of expertise needed to more confidently adopt metabarcoding methods where appropriate.

2 Overview of Types of Error and Bias

As with all methods for identifying species in a sample, metabarcoding is imperfect and can result in incorrect identification or uncertainty about the presence or absence of a taxon. The two most important error types are:

1. **False negative: metabarcoding does not identify the taxon even though it is present in the sample**. This is similar to conventional identification that overlooks a taxon in a sample.

2. **False positive: metabarcoding identifies a taxon as being present in the sample when it is not.** This is similar to conventional identification that misidentifies a taxon in a sample or erroneously records it as being present for other reasons.

The error types are related, **and there is often a trade-off in which minimising the number of false-negatives can lead to an increase in false-positives**. Decisions on what to prioritise depend on the importance of different types of error for meeting the study objectives. For example, for a given monitoring application, is it more important:

- To survey an area with maximum completeness (low false negatives)? This might be required, e.g. for whole community assessments and detecting rare species.

- To minimise the chance of falsely detecting species that are not present (low false positive)? This is important where false positives would lead to costly action that is unnecessary (e.g. erroneous detection of a notifiable pest or disease resulting in movement bans or destruction of material).

- To achieve an estimate of species composition that most closely reflects the true picture (optimal balance of false positives and false negatives)?

Table 1 summarises potential errors in metabarcoding, grouped according to the stage in the methods at which they apply. This table can be used as an aid to experimental design. It can also support interpretation of results, especially those that present unexpected findings. For example, it might be informative to look at the approach to minimising false positives if the final results contain unexpected species or surprisingly high species richness.

The sections below provide more detail, beginning with an introductory explanation and categorisation of the different sources of errors and biases and the impact on results. Additional information and approaches to mitigation are then given. The structure broadly follows the general eDNA metabarcoding workflow steps and best practices described by Furlan *et al.* (2020), though our target audience is non-specialists and so we do not provide the same level of technical detail.

Two of the main methods of controlling for errors are **(i)** sufficient biological replicates (collection of repeat samples in the field) and technical replicates (repeat analyses on the same sample), and **(ii)** positive controls (samples known to contain target DNA) and negative controls (sample known to lack target DNA). These approaches are needed at the appropriate points, although note that avoiding/mitigating some errors also requires more specific solutions.

Table 1: Summary of sources of potential error in metabarcoding. Sources of errors are classified into broad groups, and types of errors are categorised as False Positive (FP) or False Negative (FN). Errors resulting in false positives may lead to additional species being reported in a sample. Errors resulting in false negatives may lead to species being reported as absent from a sample when they are present. False positives and false negatives both cause the reported species composition of a sample to differ from the true picture. Some of the main approaches to mitigating errors are also shown. It will not necessarily be feasible or appropriate to implement every approach to mitigation in every circumstance – decisions should be informed by concern over the risk of false positives and false negatives given the intended application and the feasibility of the mitigation. Note that the table does not make a judgement on the magnitude or frequency of these errors or on the practicality of mitigation, because this depends on the application.

Error Group	Error Source	Explanation	Error	Mitigation	
				Experimental design	**Technical intervention**
Contamination	Cross-sample	Unintentional transfer of DNA between samples	Mainly FP	● Negative controls at relevant steps ● Use dual and rotating indexes	● Follow good lab practice ● Separate pre-and post-PCR ● Laboratory decontamination
	Environment	Incorporation of DNA from lab sources into samples	Mainly FP	As above	● Laboratory decontamination ● Protective laboratory gear
	Consumables	Incorporation of DNA from labware/reagents into samples	FP	● Replicates using consumables from different manufacturers	● Decontaminate consumables
Degradation	Sample collection and storage methods	Reduction in the amount of usable DNA in the sample	FN	● Tight time scheduling of DNA collection, transport, storage, and processing ● Positive controls ● Knowledge of degradation rate in sampling environment	● Optimised storage – e.g. low temperatures on arrival at lab (and ideally after collection in the field), use of high purity ethanol, liquid nitrogen or suitable buffers

Error Group	Error Source	Explanation	Error	Mitigation	
				Experimental design	Technical intervention
PCR amplification	Inhibition	Chemicals and enzymes that inhibit PCR, reducing the DNA available for sequencing	FN	• Positive controls	• Inhibitor neutralisation • Use inhibitor resistant PCR kits
	Unequal amplification	Production of more copies of the barcode region for some species than others, due to species-specific PCR biases	FN	• Use taxon-specific primers • Replicates using different primers • Positive controls	• Reduce the number of PCR cycles
	Biomass differences and non-target amplification	Reduced ability to detect organisms that contribute a small proportion of the DNA in a sample	FN	• Size sorting of organisms • Only use part of larger organisms (e.g. insect appendages) • Use taxon-specific primers	• Increase sequencing depth
	PCR errors	PCR does not always create exact copies of the barcode region	FP or FN	• Identify if barcode contains hard-to-amplify genomic areas • Only accept species if detected in a certain number of technical replicates	• Use high-fidelity PCR enzymes • Reduce PCR cycles and length of elongation step
Sequencing	Sequencing error	DNA bases are read incorrectly by the sequencer, giving an incorrect sequence	FP or FN	• Suitable thresholds for assigning sequences to species	• Monitor sequencing error profile provided by the sequencer

Error Group	Error Source	Explanation	Error	Mitigation	
				Experimental design	**Technical intervention**
Sequencing (contd.)	Sequencing failure	Technical problems cause the sequencing run to fail	FN	• Trained personnel • Monitoring of sequencer performance	• Ensure sufficient sequence complexity
	Index misassignment	Indexes associate sequences with samples. Misassignment can lead to mixing of sample data	FP or FN	• Negative controls • Quantify unknown barcode combinations	• Unique dual indexes • Use correct library storage • Remove free adapters
Bioinformatics	Unsuitable read length	Short read lengths may match to more species – longer reads give more unique and confident identification	FP		• Use overlapping paired reads • Use mechanical shearing with bioinformatic processing to produce longer reads
	Inadequate read depth	Low number of reads covering DNA section of interest reduces detection of species with low abundance	FN	• Saturation studies to determine optimum read depth	
	Quality control and trimming	Bioinformatic processing to remove low quality data	FP	• Appropriate use of quality control (QC) software • Full QC reports	• More stringent parameters for removing sequences with errors

Error Group	Error Source	Explanation	Error	Mitigation	
				Experimental design	**Technical intervention**
Bioinformatics (contd.)	Incorrect identification of sequences	Sequences are assigned to the wrong species (or genus)	FP or FN	● Use methods that require more accuracy (ASV over OTU – see Section 2.5.4) if samples allow	● Test software on data (or simulated data) with known outputs ● Test parameters using saturation and thresholding
	Unsuitable bioinformatics software, or changes in approach	Using the wrong software, or inappropriate parameters, can give incorrect results. Changing software or parameters can also affect comparability	FP or FN	● Use software designed for the specific experiment ● Include positive control data ● Ensure methods fully documented in report	● Test software on data (or simulated data) with known outputs
	Inaccurate or incomplete databases	Sequences can only be correctly assigned to a species when there is a correct match with the sequence in the barcode database.	FP or FN	● Identify database with sufficient data on species/taxa of interest	● Use larger and/or higher quality (curated) databases

2.1 Contamination

This section considers contamination in relation to '*the introduction of DNA into the sample after it has been collected*'. For discussions of contamination in sample collection see e.g. Goldberg *et al.* (2016); Liu *et al.* (2019); Furlan *et al.* (2020); Sepulveda *et al.* (2020). Contamination mainly results in false-positive errors, but in extreme cases can lead to false-negatives, e.g. when a real signal of a rare taxon is drowned out or suppressed by contamination from species that were not present in the sample when it was collected (also see Section 2.3.3). Sources of contaminating DNA can be other samples processed in the same laboratory (sample cross-contamination), the laboratory itself (environmental contamination) or the labware and reagents used to conduct the experiment (consumable-derived contamination). The risk of contamination is increased because modern metabarcoding protocols can detect very low DNA concentrations, making it more difficult to distinguish between real signals and noise.

An effective way to detect all types of contamination is to use *negative controls* that are free of input material but otherwise identical to real samples – i.e. containing the same buffers and reagents (ideally from the same batches or lots) and processed identically. Ideally, negative controls should be included at all laboratory steps because this allows contamination at each stage to be evaluated. Technical replicates are also important in understanding contamination. A technical replicate is defined as '*the repeated and independent measurement of the same biological sample*'. Comparing technical replicates can show how much random variation is introduced through laboratory protocols and analysis methods. In metabarcoding, the number of false positives arising from contamination can be reduced by only accepting species that are detected in multiple technical replicates. However, this must be balanced against the potential increased risk of false negatives for taxa with low abundance, and the greater cost of analysing more replicates.

Back to Table 1

2.1.1 Sample Cross-contamination

Cross-contamination is the unintentional transfer of DNA between samples and can occur in the field and in the laboratory. We focus on laboratory practices – see the references given in Section 2.1 for information on cross-contamination in sample collection. Cross-contamination between samples in the same study will tend to result in the same species being erroneously detected in additional samples, and so may inflate species richness within samples and/or lead to more similar species composition between samples. Contamination risks could therefore be an important consideration if interpreting results that unexpectedly show these patterns.

Cross-contamination in the laboratory can occur particularly if metabarcoding is being conducted continuously. This is because metabarcoding relies on the Polymerase Chain Reaction (PCR) to produce many copies of DNA barcode regions. The set of DNA barcodes used in this process is very limited and so without precautions, PCR products will accumulate in the laboratory environment and contaminate reagents and samples.

To reduce the risk of cross-contamination in the laboratory, it is crucial to physically separate laboratory areas where DNA of low concentration (pre-PCR) is handled from areas where high concentrations (post-PCR) are handled (Furlan *et al.* 2020). Equipment (e.g. centrifuges or pipettes), supplies (e.g. pens and notebooks), and protective gear (e.g. gloves and lab coats) must not travel between these areas without being decontaminated. All surfaces and equipment must be regularly decontaminated to keep environmental concentrations of potentially contaminating DNA to a minimum (Jones *et al.* 2020b). Wiping with diluted bleach

is highly effective for this purpose (Champlot *et al.* 2010). For sensitive equipment, non-corrosive commercial reagents such as DNA AWAY (Thermo Fisher) or DNA-Erase (VWR) can be used. Irradiation with UV light and γ-radiation can also be effective for decontaminating surfaces and reagents but may be harder to implement because it depends on the distance to the UV source and because it can damage plasticware and is a health hazard.

To reduce the risk of DNA-containing aerosols forming in the laboratory, filtered pipette tips are recommended for all metabarcoding protocol steps. Samples should also be handled in laminar flow cabinets equipped with HEPA filters and UV sterilisation lights until indexing[1] of the NGS library preparation protocol has been completed. Rotating the set of indexes between experiments is recommended (see Index Misassignment).

Cross-contamination can also occur due to experimenter mistakes such as accidentally pipetting into the wrong tube or well, potentially leading to both false-positive and false-negative errors. Good laboratory practice (e.g. clear labelling and protocols, experienced staff) and/or automated liquid handling platforms can help reduce this risk. In recent years, comparatively cheap (e.g. Opentrons OT-2) and small-footprint (e.g. Andrew Alliance Andrew+) systems have lowered the entry barrier for small laboratories to automate experimental workflows. More advanced systems (e.g. Hamilton Vantage) offer the additional benefits of complete audit trails for sample tracking and protocol execution (individual pipetting steps, heating, cooling, shaking, etc.) and thereby improve reproducibility.

Back to Table 1

2.1.2 Environmental Contamination

Environmental contamination is defined here as the unintentional incorporation of DNA into samples from sources present in the laboratory (e.g. microorganisms, arthropods, fungi, rodents, etc.). Sometimes contamination from these organisms may be obvious in results (e.g. species reported that could not feasibly have been part of the original sample) but sometimes less so (e.g. if laboratory organisms are also potentially present in the sampling location).

Environmental contamination can be avoided by using sterile and DNA-free protective equipment. The correct use of this equipment, as well as washing and replacement routines, should be specified in the relevant Standard Operating Procedures (SOPs). Many of the measures described above for avoiding sample cross-contamination also apply to environmental contamination – e.g. rigorous cleaning and decontamination, biosafety cabinets.

Back to Table 1

2.1.3 Consumable-derived Contamination

Contamination of labware or reagents used in metabarcoding can occur during the manufacturing process. This can be difficult to avoid – DNA from humans, bacteria, and domestic animals has been detected in reagents used in DNA extraction and PCR (Champlot *et al.* 2010) and natural and artificial plasmid DNA has been identified in reagents used for various molecular biology applications (Wally *et al.* 2019). Decontaminating reagents and consumables is challenging because it risks decreasing the performance of the metabarcoding protocol and thereby introducing new errors and biases. Treatment of

[1] Indexing involves adding a unique sequence of nucleotides to each sample. These allow samples to be pooled for sequencing and then separated back to the appropriate sample. Also see Index Misassignment

reagents with UV- and γ-irradiation, various chemicals, and autoclaving are options but can vary in decontamination effectiveness or impair metabarcoding performance (Champlot *et al.* 2010). However, using negative controls – and, if necessary, testing a range of different products – can help to identify contaminated consumables and reagents.

Back to Table 1

2.2 Degradation

Under ideal conditions such as permafrost, DNA can survive for millions of years (van der Valk *et al.* 2021). However, the stability of DNA is affected by abiotic (e.g. pH, UV radiation, temperature) and biological (DNA-degrading organisms) factors that can reduce DNA persistence to days-weeks (Strickler *et al.* 2015; Collins *et al.* 2018; Harrison *et al.* 2019). In the absence of preservatives, degradation affects both eDNA and DNA within organisms that are sampled directly. Degraded DNA is less likely to be detected by metabarcoding, and so DNA degradation can cause false negatives. Minimising degradation risk can be particularly important if detecting species at low abundance (e.g. rare species, invasive species at the early stages of establishing). Similarly, the approach to accounting for degradation risk should be investigated if interpreting samples with unexpectedly low species richness or species absences.

Consulting literature to understand DNA degradation rates in the sampled environment is strongly recommended during the experimental design stage. False negatives induced by degradation can also be distinguished from other explanations by using positive controls – these contain DNA from species known to be present at the survey site and so are expected to give a positive result. To reduce the risk of degradation, it is important to plan sample collection, transport, and processing to minimise exposure to adverse environmental conditions as far as possible.

To prevent further degradation after collection and extraction, ideally the DNA would be either processed immediately or stored in ethanol or optimised buffers and at -20°C or in liquid nitrogen. Where this is not realistic for logistical or cost reasons, less stringent methods can still give good results if applied in the right way, and with positive controls (Jones *et al.* 2020b; Bruce *et al.* 2021). For highly diluted DNA, specialised plastic storage tubes with reduced sample-to-surface binding (e.g. Eppendorf LoBind) are recommended. Focusing analyses on DNA that is present in multiple copies per cell (e.g. DNA from organelles such as mitochondria) can also help mitigate degradation and increase the chances of detecting species in degraded samples.

Back to Table 1

2.3 PCR Amplification

The amount of DNA extracted from organisms or from the environment is usually too low for direct detection by Next Generation Sequencing (NGS). Therefore, classical metabarcoding protocols rely on copying (or amplifying) the genomic regions (barcodes) by Polymerase Chain Reaction (PCR). The ability of PCR to generate billions of copies of a single molecule (also called amplicons) makes it a very sensitive technique but also a major potential source of several types of error and bias. These include *PCR inhibition* and *unequal amplification*, as well as non-target amplification and differences in initial biomass (each of which can result in false negatives) and *PCR errors* (which can cause false positives or false negatives) as described below.

Note that PCR-free protocols for metabarcoding have been developed, such as ultra-deep sequencing (Zhou *et al.* 2013) and physical enrichment of mitochondrial DNA (Macher *et al.* 2018). These eliminate PCR biases but may be more costly (e.g. Zaiko *et al.* 2018), and therefore amplicon-based protocols that use PCR are still most commonly used.

Back to Table 1

2.3.1 Inhibitors

Chemicals and enzymes that inhibit PCR can lead to false-negative errors. Environmental samples from different sources (e.g. soil, water, air, physical specimens) vary widely in the level and type of inhibitors present. It is therefore important to consider the risk of inhibition during experimental design. The approach to accounting for inhibition should also be investigated if interpreting samples with unexpectedly low species richness or species absences, or from environments that are likely to contain a high concentration of inhibitors.

The effects of inhibition can be assessed in experimental design using positive controls that contain well-defined amounts of DNA. These are either added to the sample ("spike-in control") or processed and analysed in parallel to the real samples ("run control"). Positive controls can be introduced at different points during the metabarcoding protocol and allow the performance of the whole process (or individual steps) to be assessed by comparing the DNA quantity in the control with the expectation given the amount of control DNA added. If performance is poorer than expected, possible causes such as inhibitors should be investigated.

While positive controls can detect the presence of inhibitors, they do not remove the effects. Where inhibitors are likely to be a problem, specific methods and inhibitor-resistant PCR kits can be used if the improvements justify the additional costs of these kits. Modern PCR formulations contain improved polymerases and additives such as betaine and Triton X-100 that increase the resistance to inhibitors, although high concentrations of chemicals (e.g. phenol) and enzymes (e.g. proteases) are still problematic. Specialised literature can help choose appropriate methods to remove or neutralise the specific inhibitors (Schrader *et al.* 2012; Majaneva *et al.* 2018). Note also that some preservatives commonly used as fixatives for specimens, such as formaldehyde, are potent inhibitors and should be avoided for metabarcoding studies.

Back to Table 1

2.3.2 Unequal Amplification

Unequal amplification arises because PCR can favour copying DNA for some species over others, and so can bias results. Amplification is the production of many copies of a barcode region, which occurs in PCR through a series of cycles. The efficiency with which PCR amplifies barcodes is species-specific (Elbrecht & Leese 2015). This means that for the same amount of input DNA, PCR may result in more copies of the barcode region for some species than others. Unequal amplification (or uniformly low amplification efficiency) can lead to false negatives, particularly for species at low abundance. It also impairs assessments of the abundance of different taxa in a sample, since the same starting biomass can result in different quantities of DNA following PCR.

The choice of primers is of great importance for amplification, and consequently for metabarcoding studies in general. Primers are short single strands of DNA bases designed to bind (anneal) to the target DNA during PCR and facilitate amplification of the target sequence (the barcode). Imperfect primer binding to DNA from a given species can result in low amplification efficiencies and therefore lead to false-negative errors.

Although the majority of metabarcoding studies use a standard set of genes such as cytochrome c oxidase subunit I (COI) or 16S rRNA as barcodes, the optimal design of primers to amplify these barcodes is still an active field of research. The risk of problems with unequal amplification can be reduced if the taxonomic focus can be clearly defined during experimental design and taxa-specific primers are then used (e.g. Gibson *et al.* 2014). If no clear guidance exists, replicated analysis of the same sample with different primer pairs allows the effect of the choice of primers to be quantified. The impact of unequal amplification will increase with the number of PCR cycles used to create copies of the barcode region: lowering the number of cycles can therefore limit the differences between species (but can increase the risk of false negatives – also see PCR errors).

Back to Table 1

2.3.3 Biomass Differences and Non-target Amplification

In addition to species-specific differences in amplification, the performance of PCR depends on the amount of input DNA. Because PCR is a stochastic process, low initial amounts of specific DNA molecules may mean that some of these will spontaneously drop out of the reaction while remaining molecules will have a higher chance of being amplified at each PCR cycle. This can increase the risk of false negatives for species that contribute a relatively small proportion of the starting DNA in a sample. There are two main situations in which this can pose a problem.

Firstly, for taxa in which the size of organisms in a sample can differ widely (e.g. invertebrates), results may be skewed towards species that contribute a larger proportion of the starting biomass at the expense of smaller or rarer species (Elbrecht *et al.* 2017). This problem can be mitigated prior to analysis by size-sorting or by only using parts of larger organisms, or by increasing the Read Depth (Elbrecht *et al.* 2021) although each of these approaches impose additional costs.

Secondly, where broad primers are used in order to survey a wide range of taxa, these may also amplify large amounts of non-target DNA (e.g. microbial DNA either in the sample or as contamination). This can make it more difficult to detect DNA from the target taxa (Collins *et al.* 2019), again potentially increasingly the risk of false negatives. If available, more taxon-specific primers can help mitigate this problem but also narrow the set of species that can be detected. Using multiple primers is also an option but increases costs (Bruce *et al.* 2021).

Back to Table 1

2.3.4 PCR Errors

The polymerase enzyme that generates DNA copies sometimes makes mistakes that incorporate the wrong bases into the DNA sequence. These errors lead to mismatches between the barcode sequences generated and the reference sequences to which they are compared, potentially resulting in species misidentification. This can cause a false-positive (the sequence is misidentified as belonging to a different species, potentially not present in the sample) and/or false-negative (the sequence is not classified as the taxon that is present in the sample).

The accuracy with which the polymerase enzyme makes copies of the DNA template is referred to as “fidelity”. High-fidelity PCR polymerases are recommended for metabarcoding, because they combine low error rates with the ability to excise (“proof-read”) bases that were incorrectly incorporated into the DNA copy. Using a lower number of thermal cycles in PCR can also help reduce errors, because errors can occur in each cycle and are then amplified in the remaining cycles. However, this risks false negatives because DNA from organisms with low initial abundance may remain under-amplified and so undetected. Lastly, because

amplification errors are semi-random (some DNA sequences are more error-prone than others), technical replicates can help to mitigate impacts with presences accepted only for species found in the majority of replicates.

In addition to occasionally incorporating the wrong DNA bases, PCR can produce chimeric sequences or 'chimaeras'. These are PCR products that are mixtures of two or more original sequences. Chimaera occurrence can be reduced by limiting the number of PCR cycles and by reducing the length of the DNA sequence elongation step that occurs during the PCR program. However, these measures increase the likelihood of not detecting rare taxa (false negative). The detection of sequence chimaeras should be part of the bioinformatic quality control pipeline.

Back to Table 1

2.4 Sequencing

DNA sequencing is the process of determining the order of bases in a DNA molecule. Errors in this process can result in differences with the true sequence, whilst in some situations sequencing may fail altogether. Sequences can also be matched back to the wrong sample if index misassignment has occurred.

A full comparison of sequencing approaches and mitigation of errors for different sequencing technologies is beyond the scope of this article (see e.g. Goodwin *et al.* 2016; Amarasinghe *et al.* 2020 and references therein). As such, we focus primarily on *Illumina* instruments that most researchers and public bodies are using/familiar with. However, many of the limitations and considerations transfer to other manufacturers, and where possible we have tried to ensure the text is broadly applicable.

Illumina implements a next-generation sequencing (*NGS* - also referred to as High Throughput Sequencing or HTS) technology called sequencing-by-synthesis. In this approach, DNA is amplified on a solid surface (on Illumina machines called "flow-cell"). The many identical copies of an individual DNA fragment ensure there is sufficient signal to read the sequence. Every time the polymerase enzyme incorporates a new DNA base into one of the copies, a light signal is emitted in a process called "fluorescence". Each of the four DNA bases emits a different colour, allowing the sequence of the DNA molecule to be determined. A single sequencer can simultaneously gather information from millions of individual DNA molecules, and so this technology is also referred to as "massively-parallel" sequencing. An excellent overview of various NGS technologies and novel developments is given by Goodwin *et al.* (2016).

Box 1: Library preparation (see Bohmann *et al.* 2021 for more information)

Before DNA can be read by the sequencer, it undergoes a process called *library preparation*, an integral part of metabarcoding. A *sequencing library* is a pool of DNA fragments, each with adapter DNA molecules attached. These adapters allow the DNA to interact with the sequencing platform. In the case of Illumina, the adapters attach to the *flow cell* surface and help to start the sequencing process. The adapters also contain unique combinations of short DNA sequences called indexes.

Libraries from different samples are usually pooled so they can be sequenced simultaneously on the same instrument ("*multiplexing*"). The indexes identify which sequencing read belongs to which sample. Before analysis of the sequencing data can begin, the reads are computationally assigned to their respective samples according to their indexes.

[Note that some metabarcoding protocols employ fusion primers that contain both the index and adapter sequences. This has the advantage of streamlining the protocol and reducing the number of PCR cycles. However, the approach might tie the library preparation protocol to a specific sequencer model and reduces the flexibility of combining samples from different experiments on a single sequencing run].

Quality control measures and prevention of equipment failure are specific to the sequencing platform, and so best-practices defined by the manufacturers should be followed, ideally with experienced operators. Overall, the sequencing step introduces little random variation into the data and is highly reproducible across instruments of the same type, experimenters, and sequencing runs. Therefore, sequencing replicates (repeated sequencing of the same library) are not considered necessary for Illumina platforms. However, understanding of the errors and biases introduced during sequencing is still required to decide on appropriate bioinformatics strategies and to interpret the results.

Back to Table 1

2.4.1 Sequencing Error

Sequencing error occurs when an individual base of the DNA sequence is read incorrectly by the sequencer. The probability of this differs between technologies (Haenfling *et al.* 2017) and also depends on the type of error (e.g. insertion or deletion of a DNA base vs substitution of a base). Sequencing errors do not necessarily affect metabarcoding results, because with a small number of errors the sequence may still closely resemble the correct barcode. However, errors can lead to false positives (if the erroneous sequence matches another species) and false negatives (if the error means that a species is incorrectly identified as absent from the sample). Errors sometimes also generate sequences that cannot be assigned to a species – this might be a particular concern where barcode libraries are very incomplete so many correct sequences could also be unassignable. Since modern sequencers provide estimates of the accuracy of the sequencing process, sequencing error can be monitored and to a certain degree controlled bioinformatically (see Section 2.5). As such, the bioinformatics approach (e.g. thresholds for species assignment, treatment of reads with low abundance) influences whether sequencing error tends to lead to false positives, false negatives, unassignable reads, or have relatively little effect on the final results.

Back to Table 1

2.4.2 Sequencing Failure

Although the reproducibility of sequencing runs is generally very high, individual runs can encounter technical problems that negatively affect the results and can cause the failure of the run. Manufacturing problems and the introduction of air bubbles during sequencing can lead to non-random errors that lower the accuracy of metabarcoding if undetected. Therefore, quality control using manufacturer-specific (e.g. Illumina's Sequence Analysis Viewer) or generalised bioinformatic software (e.g. FastQC and MultiQC) is recommended.

Another potential cause of sequencing failure is low sequence complexity, which is the case when sequencing a limited set of amplicons, as in metabarcoding. Illumina sequencers identify which nucleotide (A, C, G or T) is at each position in all of the DNA fragments simultaneously (i.e. all the first nucleotide in every fragment at the same time, then all the second nucleotides together etc). Each iteration is called a cycle. For example, a 150bp read will be the result of 150 cycles. If all of the reads have the same nucleotide at the same cycle this can lead to the sequencing run being aborted due to lack of sequence complexity. This problem can be avoided by running more complex samples alongside the metabarcoding amplicons on the same sequencing run, by including a PhiX spike, by simultaneously using different sets of metabarcoding primers in the same study, or by using frame-shifted PCR primers (Lundberg *et al.* 2013).

Continuous monitoring of performance metrics for the sequencing instrument is highly recommended to detect any unusual trends or deviations that might affect the metabarcoding analysis, and to pre-emptively detect and resolve technical problems that would lead to run failures. For Illumina systems, activation of the manufacturer's "Proactive" remote monitoring service and use of analysis tools such as MegaQC is recommended.

Back to Table 1

2.4.3 Index Misassignment

Indexes are unique sequences of DNA added to each sample to allow sequences to be assigned back to the correct sample. The indexing process is imperfect and so there is a risk of misassignment, meaning that the sequence would be assigned to the wrong sample (Bohmann *et al.* 2021). Certain combinations of sequencing instruments (e.g. systems using patterned flow cell technology), library types (e.g. libraries that contain large concentrations of free adapters), and library handling can increase the frequency of index misassignment and lead to false-positives and false-negatives. However, best practices as recommended by the sequencing system manufacturers can mitigate this problem. For Illumina, these recommendations include (Illumina 2018):

- Use of unique dual indexes (one at each end of the DNA fragment) during library preparation
- Removal of free adapters in the library using cleanup beads, gels or spin columns
- Treatment of libraries with adapter-blocking reagents
- Storage of libraries at -20°C for a maximum of one week after pooling

While it is difficult to remove misassigned sequencing reads from the data bioinformatically, negative controls and quantifying unknown index combinations allows problematic levels of misassignment to be identified and distinguished from sample cross-contamination.

Back to Table 1

2.5 Bioinformatics

Bioinformatics is defined by Haenfling *et al.* (2017) as: '*the field of biology that uses computer science, statistics, mathematics and engineering to study and process large biological data sets, especially sequence reads generated from High Throughput Sequencing*'. The bioinformatics applied to sequencing data in metabarcoding follows a common path:

1. The sequencing data are processed for quality and formatted into barcodes;
2. The barcodes are (optionally) *clustered* based on the similarity of the sequences;
3. The barcodes/clusters are compared to databases to identify taxa;
4. Results are presented.

The sequencing data used in the bioinformatics step of metabarcoding are supplied from the sequencer as strings of letters (called a *read*), composed of A's, C's, G's and T's for each DNA molecule sequenced. The letters symbolise the four bases that constitute DNA (adenine, cytosine, guanine, and thymine). Illumina sequencers return read-pairs – one read from each end of the DNA fragment being sequenced. These reads are commonly 150, 250 or 300 base pairs (bp) long. (While single-end sequencing is possible, it is not typically used for metabarcoding due to its lower information value).

The quality of bioinformatics results depends partly on the bioinformatics methods employed but also on the quality of the sequencing data and the database. Poor quality data (e.g. caused by contamination or by overamplification that adds or changes the input DNA for sequencing) reduces the accuracy of results. Unfortunately, it is hard to distinguish between these different sources of error in the final results and reports.

Good experimental design and *a priori* testing of the bioinformatics components with existing data/known results is valuable to ensure that methods are appropriate and repeatable. Testing also provides important information on the sampling and replication requirements, and on the volume of sequencing data needed for statistically robust answers.

Back to Table 1

2.5.1 Unsuitable Read Length

The read length is the number of DNA base pairs in a sequence. Short read lengths may match to more species in DNA barcode databases, increasing the uncertainty in assigning species names and increasing the risk of false positives. This risk can be mitigated using longer reads, which give more confidence of an accurate match with the database and reduce false positives. However, longer sequencing reads can cost more, and so this has to be balanced against other cost considerations such as the depth of sequencing needed (see Read Depth).

Longer reads can be obtained by combining the paired (forward and reverse ends of the DNA strand) sequencing reads together into a longer sequence - this can give better matches and so reduce false positives. For short PCR fragments, sequencers read from both ends of the DNA, so the nucleotides in the middle occur in both the forward and the reverse reads. This enables bioinformatic software to reconstruct the full fragment from the overlap. For example, if the barcode region is 500bp or below, then using 300bp overlapping paired reads will give higher sensitivity (e.g. 300bp paired reads give 100bp overlap). For longer reads only reading nucleotides at each end will not give any information on the middle

(e.g. 150bp forward and reverse reads for a 500bp fragment will miss the middle). In this case shorter reads could be used, or DNA could be mechanically sheared into smaller fragments and then bioinformatically processed to gather information about the nucleotides along the entire fragment length. Longer barcode regions influence database applicability and data re-use (see Databases).

Back to Table 1

2.5.2 Inadequate Read Depth (Volume of Sequencing)

Read depth (or sequencing depth) is defined as '*the number of reads in a sample that cover a section of interest of a genomic region*'. In relation to a single position in a DNA sequence, read depth indicates how often the base at this position has been read by the sequencer. Sequencing depth varies across different regions of a barcode and so an average value is often stated for the region of interest. If there is insufficient read depth, some DNA molecules in a sample may not be sequenced. This can mean that a sample is not fully surveyed and so some taxa will be missed (false negatives), affecting assessments of species occurrence, diversity, and composition. In general, the read depth to obtain a full species list will be greater for samples that have high species richness or that have many low abundance species (or low volume of input DNA).

One of the main methods to ensure optimum read depth is to perform a saturation study. Saturation studies involve sequencing a sample to a very high depth and running the bioinformatics analyses on subsets of these sequencing data. Plotting the number of sequences in each subset against the primary experimental statistical measure (e.g. species richness) for that subset will produce a curve that shows the sequencing depth beyond which there is little gain from more reads (i.e. false negatives are reduced to an acceptable level). Examples of optimal read depth might be the power to detect a single species at 95% power or 1% limit of detection; or when a diversity metric no longer increases with more reads. Saturation studies can help understand trade-offs between the number of samples that need to be sequenced, how many can be sequenced in one run, how many runs may be needed, and even how many field samples may be needed.

Note that the number of reads matching the barcode of a particular species do not allow a precise estimate of the absolute number of DNA molecules from that species in the collected sample. This is because there are several steps in metabarcoding protocols that amplify DNA in a partly stochastic and species-specific way. The methods paper by Kivioja *et al.* (2012) provides a good introduction to this topic, describing the use of unique molecular identifiers (UMIs) – also see, e.g. Fonseca (2018). Non-NGS technologies such as droplet digital PCR (ddPCR) offer solutions and can be used in combination with NGS to improve the accuracy of the analysis (Wood *et al.* 2019). Note that for eDNA, even exact knowledge about the absolute abundance of DNA at a sampling site only allows limited conclusions on the absolute abundance of the species present, since many small organisms might deposit the same amount as fewer bigger organisms (but see Di Muri *et al.* 2020).

Back to Table 1

2.5.3 Quality Control and Trimming

Base quality is the rate of errors in a DNA sequence. The likelihood of error for every base sequenced can be estimated by sequencers. If too many incorrect bases are included in the processing, then the chance of false positives increases. Bioinformatic pre-processing uses error estimates to shorten or remove reads from processing, and so the approach to this quality control can influence what sequences are retained for analysis.

Base quality is measured on a logarithmic scale, known as the PHRED or Q-scale: Q10 equates to 1:10 chance of base error; Q20 to 1:100 (1%); and Q30 to 1:1000 (0.1%). Quality control tools, such as Trimmomatic or Cutadapt, will commonly remove data below Q30 as well as any known sequences added to the barcode during the sequencing library preparation.

Back to Table 1

2.5.4 Incorrect Identification of Sequences

DNA sequences may be incorrectly identified, with misidentification (false positives) and no identification (false negatives) potentially occurring and so giving an inaccurate picture of species composition.

Identification involves comparing each pair of reads against a database and then assigning the reads to a taxon. There are many methods to achieve this, with the two most common using operational taxonomic units (OTUs) or amplicon sequence variants (ASVs). OTU methods cluster similar sequences together and match these against the database.

Sequences within a cluster are not necessarily identical to the database, and so the threshold for deciding whether an OTU is sufficiently similar to a species in the database will influence the risk of false positives and false negatives. If OTUs are required to match the database very closely, this will reduce the risk of false positives but increase the risk of false negatives. Conversely, if a lower threshold of similarity between the OTU and the database is used, false positives become more likely and false negatives become less likely. ASV require exact matches with the database so may reject more reads as not matching a database entry and therefore result in less data being used, and potentially less statistical power.

The choice of technique depends on the volume of data and the type of errors introduced in the field, laboratory, and sequencing. In high error situations, e.g. from amplification error due to a large number of cycles, the clustering in OTU methods may give better results. In low error situations, ASV methods may provide better results, including down to a lower taxonomic level. See articles by Glassman and Martiny (2018) and Porter and Hajibabaei (2020) for more in-depth discussion.

[Note that non-barcoding methods including genome skimming, whole genome and metagenomic analyses use different tools.]

Back to Table 1

2.5.5 Unsuitable or Changing Bioinformatics Analyses

Using the wrong software or the correct software with inappropriate parameters can give incorrect results. Different versions of the same software, different parameters with the same software, or different software that uses similar algorithms, is likely to produce different but comparable results. However, a given piece of bioinformatics software will produce consistent results with the same input data and parameters, and so to maintain comparability, bioinformatics analyses (software and parameters) should be held as constant as possible (Almeida *et al.* 2018).

The appropriateness of comparing or combining results that use different bioinformatics analyses depends on the experimental aims, and also on the statistical analyses used. However, provided the raw data and metadata are available, re-analysis using different bioinformatics methods is possible in future (though this would impose some additional cost).

Back to Table 1

2.5.6 Inaccurate or Incomplete Databases

Sequences can only be assigned to a species when there is a match with a sequence in the barcode database. The ability to do this depends on the size, quality, and specificity of the database. Poor coverage and gaps in databases mean that some taxa may be under-represented, with species missing from databases – this leads to non-identifications and false-negative results. Lower quality databases may have errors, either because a sequence in the database has been assigned to the wrong species (e.g. due to incorrect morphological identification) or because errors in sequencing mean that some DNA bases for a species included in the database are incorrect. These errors can lead to misidentifications and false-positive results when using the database. Conversely, a high-quality database with high specificity may contain many distinct species entries with similar sequences and so give finer taxonomic resolution (e.g. identification to species level).

Errors and biases associated with incomplete or lower quality databases can be mitigated to an extent by using multiple, larger, or higher quality databases with sufficient data on the taxa of interest. However, such databases do not always exist, and this should be reflected in the presentation and interpretation of results. For example, if an endangered or invasive species is not in a database then it will never be reported in the final species list and so (unless there is other evidence) cannot be viewed as absent. Similarly, if a database has low quality, this limits the confidence that can be placed on species identifications.

The database considerations are influenced by the primer choice - only the database entries that overlap with the region amplified for sequencing are important when choosing the database. All other barcode regions in the database are not relevant. As such, using a common region and trying to maximise the read length is highly recommended.

Back to Table 1

3 Different Methods, Different Results

Metabarcoding is potentially valuable in providing *complementary* information alongside conventional methods – e.g. data on additional taxa. However, in other circumstances there might be an expectation that metabarcoding will replicate what has been found using conventional methods such as morphological identification of specimens. This can be important where DNA sequencing is considered an alternative to the current approach. Although errors and bias in metabarcoding can lead to incorrect results as described in Section 2, it is important to be aware that results from metabarcoding and conventional methods can differ for other reasons.

Firstly, conventional methods can also have false positives, false negatives, and specific biases. For example, a given field survey method will be more effective at detecting some species over others for a range of reasons, such as the conspicuousness and activity patterns of the species. The impact of this error on species composition has parallels with PCR bias in metabarcoding, which could also mean some species are more likely to be detected than others. Similarly, just as sequencing or database errors in metabarcoding can cause misidentification and false positives and false negatives, conventional identification can make mistakes – particularly for assemblages that are species rich and/or contain many morphologically similar species. It is therefore important to stress that differences in results between metabarcoding and conventional identification do not necessarily mean that the error has occurred in metabarcoding.

Secondly, in some cases conventional methods and metabarcoding survey the community differently. For example, conventional surveys of waxcap fungi in grasslands rely on the detection of aboveground parts, whereas metabarcoding of soil samples can also detect fungi that do not have aboveground fruiting bodies during the survey period. Similarly, if using conventional methods, adult insect stages can be much easier to detect and identify than juvenile stages whereas metabarcoding can potentially identify both stages equally effectively. If used correctly metabarcoding and conventional methods are both potentially valid, but the possibility of different results creates problems for comparing methods to understand whether metabarcoding is a suitable alternative. Differences between conventional approaches and metabarcoding that arise because the community is surveyed in different ways relate primarily to eDNA, because if physical specimens are collected then conventional identification and metabarcoding approaches begin with the same input material. However, note that metabarcoding can also identify undigested prey species in specimens, which may not have been detected using conventional methods.

4 Conclusions

Metabarcoding is a complex combination of several evolving scientific disciplines. Like all methods, metabarcoding has limitations that can result in errors and biases in results. However, it also has many strengths and can be applied to a range of ecological and conservation questions. As shown by the increasing volume of high-quality ecological research using metabarcoding, it is certainly possible to overcome problems provided study design is informed by a good understanding of the techniques.

Importantly, methodological choices and errors at one stage will influence later stages. Many decisions in the implementation of metabarcoding also involve trade-offs. For example, actions to reduce the risk of errors and biases and to improve reliability tend to increase costs and/or impose logistical constraints, though this is not always the case. The degree to which metabarcoding errors and biases must be reduced depends on the application – some mitigation actions may be required for all applications (e.g. suitable use of positive and negative controls) but the benefits of others might only outweigh the costs if there are important economic or conservation implications of results, or if metabarcoding is the only data source. Methodological decisions also need to consider the interdependence of strategies for mitigating errors – in several cases, mitigation to reduce the risk of false positives can increase the risk of false negatives and *vice versa* (Furlan *et al.* 2020). The relative importance of these two types of error will again depend on the ecological application.

For the above reasons, it is not appropriate to provide all-purpose recommendations for applying metabarcoding. However, some general sets of recommendations are given below (also see Haenfling *et al.* 2017; Liu *et al.* 2019; Bruce *et al.* 2021).

Firstly, it is vital to get the design of the statistical analysis correct pre-experiment. Consult a statistician or bioinformatician for study-design decisions that can be modelled and investigated to maximise the chance of success:

- **Design biological replication** (e.g. number of samples taken from the field) **and technical replication** (e.g. repeat measurements on the same sample) appropriately. This is important for reducing the impact of technical error and noise.

- **Include appropriate positive/negative controls** to help identify potentially spurious results (i.e. false positives or false negatives).

- **Choose the correct primers** for the focal taxa. These must have appropriate specificity and the chosen database must have suitable/sufficient content.

Once a study is underway:

- **Collect and store DNA appropriately** to avoid degradation (with positive controls to check this)
- **Minimise contamination from the environment** and other samples (with negative controls to check this)
- **Only PCR-amplify DNA as much as necessary** and rely on well-tested primers where possible
- **Use appropriate software for controlling sequencing errors**

In reporting and interpreting data:

- **Document all stages of the process fully** (e.g. Goldberg *et al.* 2016; Jones *et al.* 2020a; Bruce *et al.* 2021)
- **Ensure limitations are explained, particularly where reports are intended for non-specialists** (e.g. risks that results include false positives or false negatives).

Lastly, it is important to ensure clear and consistent communication between bioinformaticians and end-users of results. This will help bioinformaticians understand end-user concerns and needs, and end-users appreciate potential limitations and sources of error. This should occur throughout the process to guide study design, implementation, reporting and documenting results, and the final interpretation.

Acknowledgements

We are very grateful to the following people for valuable discussions and comments: Debbie Leatherland, Katie Clark, Joan Cottrell, Kat Bruce, David Bass and Kerry Walsh. Very helpful contributions were also made through Prismea by Amy Niven, Annabel Locke and Camilla Rossi. Additional constructive comments and review was provided by Maddie Harris and Lisa Hecker.

This project was led and funded by the Defra DNA Centre of Excellence, which aims to transform Defra Group science to support policy and delivery by sharing and demonstrating the wider potential of existing methods; enabling the rapid development and trialling of new DNA based method applications; providing a focal point for leadership in Defra, UK and across other Government Departments; and building shared capability including skills and facilities.

Glossary

Definitions in the Glossary published as part of a related project (Jones *et al.* 2020a) are used directly here (*italicised*) supplemented with additional terminology. Several other glossaries have also been produced recently and may contain additional useful information (e.g. Deiner *et al.* 2017; Haenfling *et al.* 2017; Liu *et al.* 2019; Bruce *et al.* 2021)

Amplicon: Copy of a section of DNA. In metabarcoding, amplicons are produced by PCR

Amplicon Sequence Variant (ASV): A single, exact DNA sequence. ASVs generated in metabarcoding are error-corrected and are one approach to describing the taxonomic composition of a sample (contrasted with **Operational Taxonomic Units** below).

Barcode: *Specific region of DNA that has been selected as a target for sequencing, because for a given taxon (e.g. vertebrates) it is consistently and sufficiently dissimilar between species (or genera) to identify them correctly. Commonly used DNA barcode regions vary for different taxa, and more than one barcode may be necessary to identify some groups to species (Jones* et al. *2020a).*

Clustering: A bioinformatic technique to treat reads and/or database entries that differ slightly in DNA sequences as a single entity for the purposes of taxon assignment. Clustering generates **Operational Taxonomic Units (OTUs)**. The similarity threshold for determining which sequences to include in a cluster is decided by the experimenter.

eDNA: *True environmental DNA is DNA shed by an organism into its environment, rather than a sample composed of the organism itself. For example, fish DNA captured from a water sample is eDNA, while bacteria captured from a water sample will primarily be composed of the bacteria (not eDNA shed by the bacteria). This definition is not always clear cut (Jones* et al. *2020a)*

Flow cell: Small, liquid-filled glass chamber in which the sequencing reaction occurs on Illumina sequencing machines

High Throughput Sequencing (HTS): Sometimes used interchangeably with Next Generation Sequencing (NGS). *Covers a range of methods and platforms that are capable of sequencing multiple DNA molecules in parallel, enabling hundreds of thousands or millions of DNA molecules to be sequenced at a time, from the same sample (Jones* et al. *2020a)*. Metabarcoding uses HTS methods.

Index: Unique DNA sequence added to sample DNA during library preparation that is used to identify which sequencing read belongs to which sample

Library: Pool of DNA fragments that have been processed to be run on a sequencer

Metagenomics: The sequencing of genomes present in a sample (rather than specific barcode regions). See e.g. Creer *et al.* (2016) for more information

Operational Taxonomic Unit: Grouping of similar DNA sequences that is generated by **Clustering**. OTUs are viewed as approximately equivalent to species, and so are one approach to describing the taxonomic composition of a sample. (Contrasted with **Amplicon Sequence Variants**).

PCR (Polymerase Chain Reaction): Laboratory method that makes copies of specific regions of DNA. *A PCR amplifies a target region of DNA exponentially, so that a small number of copies of target DNA at the start of the reaction can be amplified up to millions of*

*copies of the target (called PCR products or **amplicons**) by the end of multiple cycles (typically 25+). The target DNA region to be amplified is determined by **primers.** (Jones* et al. *2020a)*.

Primers: *Short synthetic stretches of DNA, that bind to conserved regions of DNA flanking the target sequence (Jones* et al. *2020a).* Primers facilitate amplification of the target sequence (the **barcode**)

Read: DNA sequence generated by a sequencing machine that corresponds to all or part of a single DNA fragment

Sequencer: Machine that reads the base sequence of DNA (in this text always referring to next-generation sequencing)

Sequencing depth: Number of reads in a sample that cover a section of interest in a genomic region

References

Alberdi, A., Aizpurua, O., Gilbert, M.T.P. & Bohmann, K. (2018) Scrutinising key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, **9**, 134-147.

Almeida, A., Mitchell, A.L., Tarkowska, A. & Finn, R.D. (2018) Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience*, **7** (5). https://doi.org/10.1093/gigascience/giy054

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. & Gouil, Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, **21**, 30. https://doi.org/10.1186/s13059-020-1935-5

Barsoum, N., A'Hara, S., Cottrell, J. & Green, S. (2018) Using DNA barcoding and metabarcoding to detect species and improve forest biodiversity monitoring. Forestry Commission Research Note 032

Bohmann, K., Elbrecht, V., Caroe, C., Bista, I., Leese, F., Bunce, M. *et al.* (2021) Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13512

Bruce, K., Blackman, R.C., Bourlat, S.J., Hellstrom, M., Bakker, J., Bista, I. *et al.* (2021) A practical guide to DNA-based methods for biodiversity assessment. Advanced Books. https://doi.org/10.3897/ab.e68634

Champlot, S., Berthelot, C., Pruvost, M., Bennett, E.A., Grange, T. & Geigl, E.-M. (2010) DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS ONE*, **5**, e13042.

Collins, R.A., Wangensteen, O.S., O'Gorman, E.J., Mariani, S., Sims, D.W. & Genner, M.J. (2018) Persistence of environmental DNA in marine systems. *Communications Biology*, **1**, 185. https://doi.org/10.1038/s42003-018-0192-6

Collins, R.A., Bakker, J., Wangensteen, O.S., Soto, A.Z., Corrigan, L., Sims, D.W. *et al.* (2019) Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, **10**, 1985-2001.

Deiner, K., Bik. H.M., Machler, E., Seymour, M., Lacoursiere-Rousel, A. & Altermatt, F. (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology*, **26**, 5872-5895.

Creer, S., Deiner, K., Frey, S. Porazinska, D., Taberlet, P., Thomas, W.K., *et al.* (2016) The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, **7**, 1008-1018.

Di Muri, C., Lawson Handley, L. Bean, C.W., Li, J., Peirson, G. & Sellers, G.S. (2020) Read counts from environmental DNA (eDNA) metabarcoding reflect fish abundance and biomass in drained ponds. *Metabarcoding and Metagenomics*, **4**, e56959. https://doi.org/10.3897/mbmg.4.56959

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 (7)**, e0130324. https://doi.org/10.1371/journal.pone.0130324

Elbrecht, V., Peinert, B. & Leese, F. (2017) Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, **7**, 6918-6926

Elbrecht, V., Bourlat, S.J., Horren, T., Lindner, A., Mordent, A., Noll, N.W., *et al*. (2021) Pooling size sorted Malaise trap fractions to maximise taxon recovery with metabarcoding. *Peer J*, **9**, e12177. https://doi.org/10.7717/peerj.12177

Fonseca, V.G. (2018) Pitfalls in relative abundance estimation using eDNA metabarcoding. *Molecular Ecology Resources*, **18**, 923-926.

Furlan, E.M., Davis, J. & Duncan, R.P. (2020) Identifying error and accurately interpreting environmental DNA metabarcoding results: a case study to detect vertebrates at arid zone waterbodies. *Molecular Ecology Resources*, **20**, 1259-1276.

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

Glassman, S.I. & Martiny, J.B.H. (2018) Broadscale ecological patterns are robust to use of Exact Sequence Variants versus Operational Taxonomic Units. *mSphere*, **3** (4). https://doi.org/10.1128/mSphere.00148-18

Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F. & Murphy, M.A. (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, **7**, 1299-1307

Goodwin, S., McPherson, J.D. & McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**, 333-351.

Hänfling, B., Lawson-Handley, L., Lunt, D., Shum, P., Winfield, I. & Read, D. (2017) A review of recent advances in genetic methods to identify improvements in CAMERAS partners monitoring activities (Reference – CR/2015/09).

Harrison, J.B., Sunday, J.M., Rogers, S.M. (2019) Predicting the fate of eDNA in the environment and implications for studying biodiversity. *Proceedings of the Royal Society: Biological Sciences*, **286**, 20191409. http://dx.doi.org/10.1098/rspb.2019.1409

Illumina. (2018) Effects of index misassignment on multiplexing and downstream analysis. Illumina Inc. Pub. No. 770-2017-004-D QB 5746. Available at https://emea.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/index-hopping.html. Retrieved November 2021

Jones, E.P., Adams, I.P., Walshaw, J., Benucci, M., MacArthur, R., Boonham, N. & Bryce S. (2020a) Guidance for end users on DNA methods development and project assessment. JNCC Report No. 669a. JNCC, Peterborough, ISSN 0963-8091.

Jones, E.P., Adams, I.P., Walshaw, J., Benucci, M., MacArthur, R., Boonham, N. & Bryce S. (2020b) End-user Frequently Asked Questions on DNA-based methods for environmental monitoring. JNCC Report No. 669b. JNCC, Peterborough, ISSN 0963-8091.

Kivioja, T., Vaharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, **9**, 72-74.

Liu, M., Clarke, L.J., Baker, S.C., Jordan, G.J. & Burridge, C.P. (2020) A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, **45**, 373-385.

Lundberg, D.S., Yourstone, S., Mieczkowski, P., Jones, C.D. & Dangl, J.L. Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, **10**, 999-1002.

Macher, J.-N., Zizka, V.M.A., Weigand, A.M. & Leese, F.A. (2018) A simple centrifugation protocol for metagenomic studies increases mitochondrial DNA yield by two orders of magnitude. *Methods in Ecology and Evolution*, **9**, 1070-1074

Majaneva, M., Diserud, O.H., Eagle, S.H.C., Hajibabaei, M. & Ekrem, T. (2018) Choice of DNA extraction method affects DNA metabarcoding of unsorted invertebrate bulk samples. *Metabarcoding and Metagenomics*, **2**, e26664. https://doi.org/10.3897/mbmg.2.26664

Porter, T.M. & Hajibabaei, M. (2020) Putting COI metabarcoding in context: the utility of Exact Sequence Variants (ESVs) in biodiversity analysis. *Frontiers in Ecology and Evolution*, **8**, 248. doi: 10.3389/fevo.2020.00248

Schrader, C., Schielke, A., Ellerbroek, L. & Johne, R. (2012) PCR inhibitors – occurrence, properties and removal. *Journal of Applied Microbiology*, **113**, 1014-1026.

Sepulveda, A.J., Hutchins, P.R., Forstchen, M., Mckeefry, M.N. & Swigris, A.M. (2020) The elephant in the lab (and field): contamination in aquatic environmental DNA studies. *Frontiers in Ecology and Evolution*, **8**, 609973. doi:10.3389/fevo.2020.609973

Strickler, K.M., Fremier, A.K. & Goldberg, C.S. (2015) Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biological Conservation*, **183**, 85-92.

Van der Valk, T., Pecnerova, P., Diez-del-Molino, D., Bergstrom, A., Oppenheimer, J., Hartmann, S., *et al*. (2021) Million-year-old DNA sheds light on the genomic history of mammoths. *Nature,* **591**, 265-269.

Wally, N., Schneider, M., Thannesberger, J., Kastner, M.T., Bakonyi, T., Indik, S., *et al.* (2019) Plasmid DNA contaminant in molecular reagents. *Scientific Reports*, **9**, 1652. https://doi.org/10.1038/s41598-019-38733-1

Wood, S.A., Pochon, X., Laroche, O., von Ammon, U., Adamson, J. & Zaiko, A. (2019) A comparison of droplet digital polymerase chain reaction (PCR), quantitative PCR and metabarcoding for species-specific detection in environmental DNA. *Molecular Ecology Resources*, **19**, 1407-1419.

Zaiko, A., Pochon, X., Garcia-Vazquez, E., Olenin, S. & Wood, S.A. (2018) Advantages and limitations of environmental DNA/RNA tools for marine biosecurity: management and surveillance of non-indigenous species. *Frontiers in Marine Science*, **5**, 322. doi:10.3389/fmars.2018.00322

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X. Zhou, L., *et al.* (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2:4**. https://doi.org/10.1186/2047-217X-2-4