Opportunities and challenges in metabarcoding approaches for helminth community identification in wild mammals

TUOMAS AIVELO[1]* *and* ALAN MEDLAR[2]

[1] *Department of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland*
[2] *Institute of Biotechnology, University of Helsinki, Viikinkaari 5, PO Box 56, Finland*

(*Received 26 January 2017; revised 16 March 2017; accepted 6 April 2017; first published online 23 May 2017*)

SUMMARY

Despite metabarcoding being widely used to analyse bacterial community composition, its application in parasitological research remains limited. What interest there has been has focused on previously intractable research settings where traditional methods are inappropriate, for example, in longitudinal studies and studies involving endangered species. In settings such as these, non-invasive sampling combined with metabarcoding can provide a fast and accurate assessment of component communities. In this paper we review the use of metabarcoding in the study of helminth communities in wild mammals, outlining the necessary procedures from sample collection to statistical analysis. We highlight the limitations of the metabarcoding approach and speculate on what type of parasitological study would benefit from such methods in the future.

Key words: mammals, genetic identification, high-throughtput sequencing, cestodes, nematodes, community ecology.

INTRODUCTION

In recent decades parasitology research has moved on from the one host – one parasite paradigm towards studying systems composed of multiple hosts and parasites. While traditional DNA sequencing is used to identify individual parasites (e.g. Jenkins *et al*. 2005; Asmundsson *et al*. 2008; Kutz *et al*. 2013; Budischak *et al*. 2015; papers in this issue), high-throughput sequencing, when combined with universal primers, can screen for multiple parasites concurrently (Tanaka *et al*. 2014; Aivelo *et al*. 2015; Avramenko *et al*. 2015; Lott *et al*. 2015). Such approaches collectively referred to as metabarcoding, can interrogate parasite communities faster, cheaper and more accurately than traditional methods (Bik *et al*. 2012).

Traditionally, the identification of parasite species has relied on adult individuals. In the case of helminth taxa, for example, it is difficult to distinguish between species from their egg, cyst or larval forms and in some cases, identification relies on sex-specific traits (Gasser, 2006). This poses a problem, as in many research settings adult individuals are unavailable and therefore assessments can only be made at higher taxonomical levels (Floyd *et al*. 2002). For example, non-invasive parasite assessment relies on egg and larvae samples. While it is sometimes possible to grow adult worms from eggs and larvae, coproculture is laborious and works with only a limited number of parasite species (Gasser, 2006). This problem is exacerbated by the ever-growing specialization necessary to distinguish between intestinal helminths. It is therefore practically impossible to perform exhaustive research projects on component communities of parasites based solely on morphology. In comparison, in a well-designed metabarcoding study, assigning amplicons to putative species is straightforward and does not need taxon-specific knowledge. Metabarcoding therefore opens up new avenues for parasitological research in, for example, longitudinal studies or studies in host animals, which cannot be killed due to endangerment or other ethical reasons (Table 1).

Intestinal helminths have different life cycles and modes of dispersal. Different protocols are therefore required to survey different taxa, with even sample collection requiring specialist knowledge. In comparison, metabarcoding is relatively straightforward irrespective of taxa studied and helminths at all life cycle stages could, in theory, be identified using the same workflow (Goldstein and DeSalle, 2011).

As with any scientific endeavour, metabarcoding is not without its challenges or critics; mostly related to the perceived disadvantages versus 'traditional taxonomy' (Ebach and Holdrege, 2005; Mitchell, 2011). We sidestep this debate, instead focusing exclusively on practical issues. The debate has been covered extensively elsewhere and we point the interested reader to Casiraghi *et al*. (2010); Taylor and Harris (2012) and Collins and Cruickshank (2013). Furthermore, much of the criticism against metabarcoding has been directed

* Corresponding author: Department of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland. E-mail: tuomas.aivelo@ieu.uzh.ch

Parasitology (2018), **145**, 608–621.
doi:10.1017/S0031182017000610

Table 1. A partial list of proposed benefits of the metabarcoding approach to the parasite identification, adapted from Aivelo (2015)

Benefit	Examples
All life stages can be identified	Leung *et al*. (2009); Locke *et al*. (2011)
Cryptic species recognized	Ferri *et al*. (2009); Ogedengbe *et al*. (2011)
Molecular methods make high-throughput processing of samples possible	Tanaka *et al*. (2014); Aivelo *et al*. (2015)
Identification of wide range of parasites	Tanaka *et al*. (2014)
Use of bulk environmental samples	Not yet done, but see Deagle *et al*. (2009)
Homology of genes is easier to predict than homology of morphological characters	Under debate, see Silva *et al*. (2010)
DNA sequences are digital and easy to communicate from laboratory to laboratory	Routine procedure

towards 'DNA taxonomy' – a method of building a new taxonomic framework. Delimitation of new species based on DNA is a highly contentious issue, whereas DNA identification of described species is more widely accepted (Lee, 2004; DeSalle *et al*. 2005).

In this review we discuss a generic workflow for metabarcoding studies (Fig. 1), identify challenges and discuss the options for sample processing, choice of barcode regions and bioinformatics analysis. We conclude with a critical analysis of metabarcoding, outlining the limitations and identifying future advancements necessary for better and more effective studies. While metabarcoding can involve both environmental samples and bulk processing of isolated samples (Bik *et al*. 2012), we will focus on identification of parasites from fecal samples.

METABARCODING IN PRACTICE

While metabarcoding has not been used in many parasitological studies so far, high-throughput sequencing has had a major impact on the study of helminth phyla, e.g. nematodes and cestodes. Indeed, DNA has been successfully isolated from a majority of parasitic groups (Caron, 2009). Moreover, identification of human and domestic animal pathogens is routinely performed using molecular methods (Ferri *et al*. 2009). Genetic parasite identification is generally based on polymerase chain reaction (PCR) amplification using either species- or group-specific primers, but other methods, e.g. restriction fragment length polymorphisms (RFLP), are also used (Lott *et al*. 2015). At the moment, DNA sequencing is the preferred method for molecular biodiversity studies (Gasser *et al*. 2008; Valentini *et al*. 2009; Creer *et al*. 2010; Bik *et al*. 2012) and it is therefore expected to become more common in parasitological studies.

Non-parasitological fecal analysis methods have also been developed, for example, there have been many more studies in diet analysis (Symondson, 2002; King *et al*. 2008; Deagle *et al*. 2009). For many vertebrate and invertebrate groups, diet analysis can be performed by amplification with universal primers and high-throughput sequencing. Some of the diet analyses have also detected parasites as non-target identifications, demonstrating that parasites can be readily amplified from fecal samples (Srivathsan *et al*. 2015, 2016).

As only a limited number of parasitological studies have used the metabarcoding approach so far, it implies that either there is limited interest in high-throughput community analysis of intestinal parasites or that there are impediments to its use. The former is probably untrue, as the paradigm of parasitological research is moving towards considering multiple species parasite communities (Archie and Ezenwa, 2011; Bordes and Morand, 2011; Viney and Graham, 2013). To better understand the problems associated with helminth metabarcoding, we are going to review the four studies performed to date: helminths in rats by Tanaka *et al*. (2014); strongylids in wallabies by Lott *et al*. (2015); nematodes in ruminants by Avramenko *et al*. (2015) and our study of nematodes in mouse lemurs, which is the only study performed on free-living animals (Aivelo *et al*. 2015).

Sample collection in the field

Genetic studies in the field are fraught with difficulties related to facilities and limited supplies. In spite of conditions, it is crucial that contamination be kept to a minimum and that samples are stored appropriately to simplify downstream processing.

While contamination is less abundant in helminth samples than bacterial samples, caution must still be observed. General rules for reducing cross-contamination also apply to parasite DNA studies: equipment should be sterilized between samples and negative controls (feces known to be uninfected) should be collected for validation purposes. Sterilization is not always possible in field conditions, but, for example, drying traps in direct sunlight provides sufficient ultraviolet radiation to kill many parasite taxa (Gaugler *et al*. 1992). Other equipment can be washed with ethanol to minimize the risk of contamination.

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

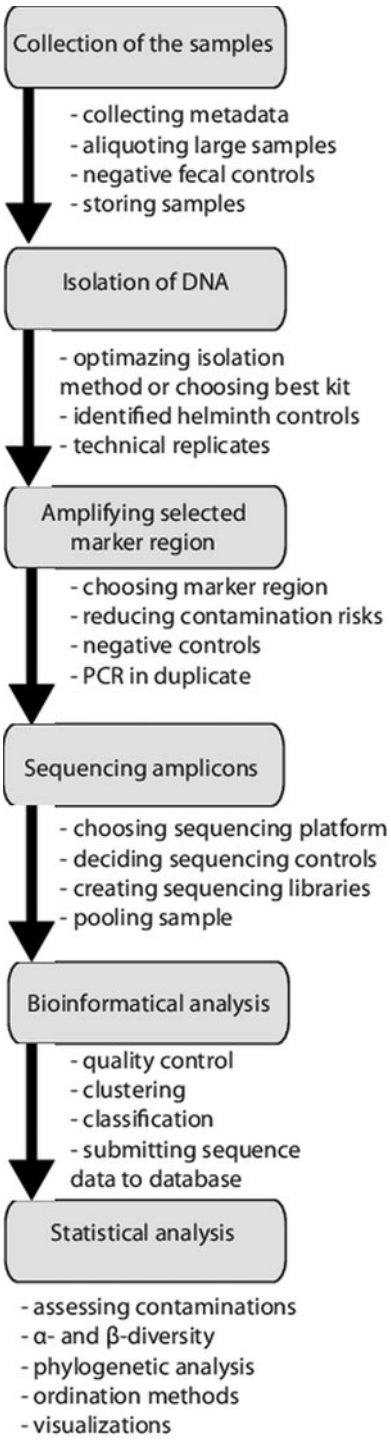

Fig. 1. Metabarcoding workflow. First DNA needs to be isolated from the collected samples, then the desired gene region is amplified with PCR and the amplicons sequenced. The resulting sequences need to be processed to deal with sequencing errors and assign sequences to reliable taxonomic identifications. This data can then be used for community analysis. Figure adapted from Aivelo (2015). PCR, polymerase chain reaction.

Aivelo *et al*. (2015) encountered several groups of nematodes, which were unlikely to be mouse lemur parasites. First, there were high levels of contamination from soil nematodes. As soil nematodes are ubiquitous, they can be transferred by the host, e.g. from their feet, to their feces. Second, the host could simply have consumed nematodes along with food or nematode-infected prey. Especially in the case of carnivorous hosts, this could prove to be an important source of contamination, though it has not been studied systematically. This was not the case, however, as only living nematodes were isolated with Baermann's method. Unfortunately, this resulted in contamination from flies that laid eggs on the samples, with some amplicons identified as intestinal parasites of arthropods. Luckily, this contamination always coincided with dipteran sequences which were also amplified by the universal primers. During the first year this contamination was abundant, whereas during the second year the laboratory had a door and contamination was much rarer (Aivelo, *in preperation*)

Presence-absence studies need to assess false negatives and false positives. While it is generally difficult to assess the probability of false negatives, there are several ways to do this. The most straightforward approach is to validate the presence of helminths with other methods, e.g. morphologically. This is easy to do in bulk samples, where helminths are isolated prior to DNA isolation. A similar approach can be used to assess false positives: negative controls from fecal samples known not to contain any helminths should be processed alongside other samples. Also, replicates of samples from different processing points (duplicate sampling, duplicate DNA isolation, duplicate PCR reactions) can be used to decrease of chances of false negatives.

While the best option is to process the samples when they are fresh, this is often not possible in field conditions and they need to be preserved. There is a trade-off between different end-uses of samples: parasitological samples are routinely fixed in formaldehyde for morphological examination, making DNA extraction very difficult. Likewise, 96% ethanol prevents sample degradation, but quickly fixates the DNA. Samples for metabarcoding should be stored in 70% ethanol, and if they are being stored for long periods of time, should be kept in a freezer at least at −18 °C. If samples need to be stored indefinitely or very high quality material is required, RNAlater (Ambion, Austin, USA) provides sufficient quality for downstream processing as it quickly permeates and stabilizes tissues, preventing DNA degradation.

For many wild mammals, the amount of available fecal matter can be substantially greater than required and, therefore, a sample must be extracted. The aggregation of parasite eggs is poorly studied, but it seems that in humans, helminth eggs are distributed without clear patterns either from front to back (Martin and Beaver, 1968; Yu *et al*., 1998; Krauth *et al*., 2012; or surface to centre of the fecal pellet (Woodstock *et al*., 1971; Ye *et al*., 1998; Krauth *et al*., 2012). Thus, we would recommend

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

sampling large fecal pellets from several locations or first homogenizing fecal pellets before sampling.

Sample processing in the laboratory

In comparison with bacteria, helminth DNA contamination in the laboratory is expected to be lower as it is less prevalent in general. If other nematodes are being handled in same laboratory as high-throughput sequencing samples, extra care should be taken to avoid contamination. Separate areas for pre-PCR work (DNA isolation, PCR set up) and post-PCR (preparing potential second PCR reaction or sequencing libraries) to reduce contamination substantially. Positive controls can be amplified preferably after all metabarcoding samples have been amplified.

DNA isolation and contaminations. Linking gene sequences to actual taxa is easier when known specimens are processed alongside collected samples. Preferably these are the same as in the studied host population, as positive controls or by partially matching fecal samples to examination of respective intestines. If actual species identifications can be attached to sequences, it also allows the resolution of the marker genes to be assessed. These validation steps should be performed in a pilot study, as in this way the DNA isolation and amplification strategies can be optimized for the studied communities (Cowart *et al*. 2015).

In comparison with other particles in fecal matter, intestinal parasites are generally abundant and in good condition. Eggs, cysts and larvae are living material and not degraded in comparison with other fecal matter like diet contents. Nevertheless, if genetic material is rare, fecal samples need to be treated carefully, as the first steps of PCR can easily lose rare material (Jarman *et al*. 2004). Conventional barcoding uses relatively long sequences (~300 bp) but environmental samples can contain degraded DNA where the average length could be closer to 150 bp (Valentini *et al*. 2009). Obviously, high quality DNA amplifies better than degraded DNA and in fecal matter, the host DNA tends to be higher quality than other material.

It has become commonplace to perform two technical PCR replicates on bacterial DNA, as this has led to more reliable estimation of the community composition. With helminths, the required number of replicates depends on the rarity of the species present (Ficetola *et al*. 2015). Sequencing depth is also an important factor for similar reasons (Smith and Peay, 2014). Replication levels should always be considered in relation to the study design. In fresh samples, the technical replicability is generally good (Ficetola *et al*. 2015), so resources should be directed to collection of multiple samples in the field. In cases where some phases of sample processing, e.g. DNA isolation or PCR amplification, have lower success rates and therefore higher chance of bias, technical replicates should also be performed.

In the four previous helminth metabarcoding studies, intestinal parasite DNA has been successfully isolated from fecal samples using different methods: by beadbeating with different kits and by using proteinase K with and without flash-freezing to rupture cell structure (Table 2). Using kits in general gives high-quality isolated DNA, but have lower yield than traditional methods. Different kits have differential isolation success rates in different groups. Therefore, depending on the kit, different kinds of bias can be introduced. Again, this should be a less significant problem when studying presence and absence of helminths. In quantitative analysis, as the number of species will be lower compared with bacterial metabarcoding , the bias should be easier to account for (see the section 'Quantitative *vs* qualitative'). Nevertheless, the success of DNA extraction should be assessed beforehand from all targeted parasite taxa. While half of the studies used the fecal pellets for DNA isolation, the other two separated helminths first from the fecal matter (Table 2). First separating the parasite larvae and using those for DNA isolation leads to comparatively clean samples with high quantities of target taxa and limited amounts of fecal bacteria DNA.

Fecal samples have a high concentration of inhibitors that make amplification prone to failure. Again, fecal isolation kits are adapted to offset the effect of these inhibitors, but there are differences in fecal composition from species to species and thus optimal isolation kits can only be found by trial and error. Also, in cases where special isolation methods are used, these DNA isolates can then be further cleaned with suitable kits.

Each helminth phyla has unique problems related to DNA isolation from the life stages present in feces (Hu *et al*. 2007). DNA isolation in nematodes is more difficult than in other helminth phyla as the cuticle is resistant to some isolation methods and the adult specimens have a low number of cells, resulting in small quantities of genetic material (Dawkins and Spencer, 1989; Gasser *et al*. 1993; Harmon *et al*. 2006). Nevertheless, nematodes are usually excreted in high numbers and some species hatch before defecation, meaning that there are live multicellular larvae available for sequencing (Stear *et al*., 1995, 2006; Mes, 2003). In comparison, other helminths are excreted as highly resistant eggs, which contain small amounts of DNA and from which DNA isolation is more challenging (Bretagne *et al*. 1993; Mathis and Deplazes, 2006).

Amplification of barcode genes. The barcode region of choice for metazoans is cytochrome oxidase I (COI or coxI) (Hebert *et al*. 2003*a*, *b*). COI has

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

Table 2. Comparison of the different metabarcoding studies performed on helminths of mammals.

	Tanaka *et al*. (2014)	Aivelo *et al*. (2015)	Lott *et al*. (2015)	Avramenko *et al*. (2015)
Host	Brown and black rat (*Rattus norvegicus* and *R. rattus*)	Rufous mouse lemur (*Microcebus rufus*)	Red kangaroo (*Macropus rufus*)	Cattle (*Bos taurus*)
Host environment	Free-living in urban environment	Free-living in rainforest	Zoo and wildlife park animals	Domestic animals
Parasite taxa	nematode, cestode, protist	nematode	nematode	nematode
Marker gene	18S	18S	ITS-2	ITS-2
Fecal processing	No	Baermann's method	No	Modified Wisconsin method
DNA isolation kit	Mobio PowerSoil	No	Biolane ISOLATE Fecal	No
Cell disruption	Beadbeating	No	Beadbeating	Deep-freezing
Cell lysis		Proteinase K		Proteinase K
Amplification	Two-step PCR	Two-step PCR	Two-step PCR	Two-step PCR
Amplicon length	~150 bp	~350 bp	~400 bp	~320 bp
Sequencing	MiSeq	454	MiSeq	MiSeq
Sequence analysis pipeline	QIIME	Séance	SeqMan NG and mothur	FLASH
OTU picking	UCLUST – Closed reference against SILVA with all eukaryotes	Séance – *de novo*	Mothur – *de novo*	no OTU picking
Identity threshold	95%	99%	97%	97%
OTU identification	QIIME and BLAST against in-house database	MegaBLAST against NCBI NR database	BLAST against NCBI database	BLASTN of individual sequences against in-house database and Genbank
OTU contamination threshold	not used	5 amplicons in at least two samples	100 amplicons over all samples	not used

Table 3. A selection of studies, which have either used Sanger sequencing barcoding for parasite identification or have used metabarcoding approaches to identify parasites or closely related taxa

Helminth group	Marker gene	Targeted/observed species	Metabarcoding or barcoding	Reference
Cestodes	18S	Helminths in mammal fecal sample	MBC	Tanaka *et al*. (2014)
	COI	Mammal helminths	BC	Galimberti *et al*. (2012)
Cestodes and trematodes	COI, 18S	Previously collected specimens	BC	Van Steenkiste *et al*. (2015)
		Fish helminths	BC	Moszczynska *et al*. (2009)
Trematodes	COI	Fish helminths	BC	Leung *et al*. (2009)
	18S	Snail parasites	BC	Routtu *et al*. (2014)
	18S, 28S, ITS, COI	Rhabdocoels and monogeneans from fish	BC	Vanhove *et al*. (2013)
Nematodes	18S	Soil nematodes	MBC	Porazinska *et al*. (2009)
		Helminths in mammal fecal sample	MBC	Tanaka *et al*. (2014)
		Helminths in mammal fecal sample	MBC	Aivelo *et al*. (2015)
	COI	Previously collected specimens	BC	Derycke *et al*. (2010)
		Canned fish	BC	Siddall *et al*. (2012)
		Previously collected vertebrate parasites	BC	Prosser *et al*. (2013)
	ITS-2	Helminths in mammal fecal sample	MBC	Lott *et al*. (2015)
		Helminths in mammal fecal sample	MBC	Avramenko *et al*. (2015)

been widely successful, although it does not work for every phylum. Standardization is progressing slowly in the difficult phyla (Frézal and Leblois, 2008) and there are also difficulties within otherwise well-functioning phyla (Santos *et al*. 2011). The length of the amplified region is also affected by sequencing method (see the section 'Sequencing'), but as a general rule, the targeted regions should be shorter if the amount of DNA is low or otherwise degraded. To illustrate the diversity of marker genes used in the literature, we have collected a non-exhaustive list of studies that barcoded cestodes, trematodes and nematodes in Table 3 along with the different marker genes used. In comparison with other helminth groups, there is currently a lack of barcoding studies targeting the helminths in phylum Acantocephala. As metabarcoding is still rarely applied in this domain, we included both regular barcoding of helminths and metabarcoding performed on closely related non-parasite groups. 18S rRNA gene databases have the highest coverage of described species in helminth groups, though COI is becoming more common due to BOLD (Barcode of Life Database) efforts. 18S can be used when there is interest in higher taxonomical groups, whereas species-level identification requires use of COI or ITS (Blasco-Costa *et al*. 2016).

Sequencing. We will not extensively review different sequencing platforms in this paper. Shokralla *et al*. (2012) have a good, though already quite old, overview of the different sequencing platforms. The most widely used sequencing platforms for metabarcoding are the Roche 454, which has already been phased-out of production, and the Illumina MiSeq. Roche was one of the early adopters of next-generation sequencing technology and provided quite long (150–1000 bp) reads, which could be utilized for species identification. Lately, the 300 bp read pairs provided by MiSeq are not only faster and cheaper than the 454, but has become the method for choice for metabarcode studies. While MiSeq sequencing has a high error rate compared with traditional Sanger sequencing, it provides paired-end sequences, which can be merged into a consensus sequence. All sequencing platforms exhibit different kinds of bias. For example, 454 suffers from homopolymer errors (Quince *et al*. 2011) and Ion PGM is biased by high GC content (Quail *et al*. 2012). These differences make it more difficult to compare community analyses performed on different sequencing platforms.

Bioinformatics

There is a wide range of software available for processing sequence data and performing statistical analysis on the resulting observations. Many programs used for metabarcoding have been integrated into pipelines; the best supported being Mothur (Schloss *et al*. 2009) and QIIME (Caporaso *et al*. 2011). While both integrate software together, they have differing philosophies: Mothur includes optimized re-implementations of existing methods, whereas QIIME does not

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

implement these functions itself, but provides many 'wrapper' scripts designed to interoperate with one another. Both are frequently updated.

In general, bioinformatics for metabarcoding follows three basic steps: quality control, clustering and classification. While these broad steps are not always clearly demarcated, it will serve as a conceptual framework to contrast each method.

Quality control. Quality control is highly dependent on sequencing platform, for example, Roche 454 data must be denoised to remove homopolymer errors (Quince *et al*. 2011) and, analogously, paired-end data from Illumina MiSeq should be merged into longer contigs if they overlap. Irrespective of platform, sequencing involves many artificial oligonucleotide sequences, including PCR primers, adapters and multiplexing barcodes; all of which must be removed. As different library preparation kits use different adapters, it is crucial to know precisely what was used as their inclusion can lead to the identification of spurious species (Schloss *et al*. 2011). From the perspective of quality control, however, the inclusion of adapter sequences in the raw data is beneficial as sequences known *a priori* are quality controls that are independent of the quality metrics output by the sequencer. As a result, high numbers of mismatches in the primer or barcode sequences are assumed to indicate low quality reads, which are discarded. After removing adapter sequences, the resulting amplicons must be filtered for quality: trimming low quality bases from the 3' end and then rejecting sequences that are either too short or noisy. Sequences that are exact duplicates are removed and a count kept in their place. Finally, chimeric sequences formed during PCR amplification need to be identified and removed as their inclusion leads to an inflation of species diversity. Chimera detection operates on the principle that putative chimeras must have sufficient similarity to two parent sequences, both of which are found in higher abundance than the child sequence (Haas *et al*. 2011).

Clustering. Estimation of species diversity requires de-duplicating data to summarise intra-species variation, reduce the computational effort needed to perform downstream analysis and respect operational definitions (for example, 3% dissimilarity delimiting bacterial species). Unique sequences are clustered, with the clusters termed *operational taxonomic units* (OTUs) (Blaxter *et al*. 2005; Bik *et al*. 2012). The resulting OTUs depend on the similarity threshold provided by the user and which clustering algorithm is used, i.e. how the similarity threshold is interpreted. Commonly used algorithms include: hierarchical clustering (including single-, average- and complete-linkage clustering), centroid-based and closed-reference clustering that we will cover in turn.

Hierarchical clustering requires a distance matrix between all sequence pairs as input; building up clusters by agglomeration. In this way all data points start out as independent clusters, but are combined if the distance between clusters is less than the similarity threshold. Where approaches differ is how the distance between clusters is calculated. In single-linkage clustering the distance between two clusters is the minimum distance between all pairs of data points, each data point coming from a different cluster. In average-linkage clustering, the distance is defined as the average distance between all pairs of data points, one from each cluster. Highly abundant sequences are more likely to be error-free, so a weighted average can be used instead. Unfortunately, hierarchical clustering can be slow for large datasets as the number of calculations scales quadratically with the number of unique input sequences. Hierarchical clustering is available in the Mothur pipeline (Schloss *et al*. 2009).

Centroid clustering avoids pre-calculating the full distance matrix and takes advantage of abundance information from exact duplicates. The method considers all sequences in descending order of abundance, comparing each data point with all existing clusters. Each cluster is defined by a representative sequence, the centroid, against which new data points are compared using the similarity threshold. If a sequence is not sufficiently close to an existing cluster, a new one is created with that sequence as the centroid. While the total amount of computation is lower than would be required for a full distance matrix, the procedure can result in a 'long tail' of low abundance clusters due to the suboptimal selection of cluster centroids. This approach is used by UCLUST (available in QIIME) (Edgar, 2010) and Séance (Medlar *et al*. 2014), among others.

One issue with the clustering methods described so far is the underlying assumption that all sequences are homologous for the same loci, which might not always be the case. In study designs that use multiple libraries, hierarchical and centroid-based clustering would erroneously produce multiple clusters per species. Instead, closed-reference OTU picking finds the closest match for each input sequence from a reference database. Each reference sequence is considered to be a cluster in downstream analysis. Closed-reference OTU picking is only applicable for the most commonly sequenced marker genes and can result in bias in the case of an incomplete database and, as such, can leave sequences that are not sufficiently close to any reference sequence unclassified. Despite these limitations, closed-reference OTU picking is common in barcoding studies and is implemented in QIIME (Caporaso *et al*. 2011).

For helminths, as with other organisms, the most appropriate method is dependent on the study design. As stated previously, the number of helminth species present in a single host is usually quite low,

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

making centroid clustering far more efficient than hierarchical clustering. For similar reasons clustering can be performed using a high similarity threshold, e.g. 99%, however, the optimal value will depend on the marker gene region selected. Closed-reference clustering warrants some consideration too, however, as the specific scientific question may revolve around specific species, the diversity of which are well represented in sequence databases.

Removing potential contaminations. As low level amplicon contamination from non-target sequences is inevitable, some OTUs might be composed entirely of contamination. A majority of contamination can be avoided by filtering out sequences with a low copy number, requiring OTUs to be present in multiple samples (in longitudinal settings) and requiring a minimum number of amplicons before a sample can be said to contain a given OTU (see Table 2 for examples). Any thresholds are dependent not only on the level of contamination, but also sequencing depth, amplification and sequencing biases and the number of amplicons generated by the less abundant helminths.

Classification. Clusters alone are not especially useful and need to be taxonomically classified so results can be contrasted with existing knowledge, to allow for contamination to be removed (for example, clusters derived from host DNA) and to enable comparisons between different samples. Classification is inherently reference-based, so it is important to inspect results critically if there is a possibility of sampling previously undescribed or under represented species.

The RDP classifier uses a naïve Bayes approach to exploit correlations between k-mer frequencies to perform classification (Wang *et al*. 2007). The RDP classifier provides pre-built models for bacterial 16S and fungal ITS genes, but only to the genus level. Uncertainty is quantified with bootstrapping. Unfortunately, it is difficult to recommend for helminth identification: firstly, and most practically to a majority of users, there is no prebuilt model for helminths. To use the RDP classifier, in this context, requires the collection of suitable sequence data and the retraining of the model. This is further complicated as, while there are many sequences in databases, there are few sequences per genus (for example, the median number of nematode sequences per genus in the SILVA database is 7, and per species is 1). This presents a problem because statistical methods work better with more examples per taxon. Secondly, the efficacy of such methods is evaluated using cross-validation and without many sequences per taxon, even our evaluation of the method in artificial settings would be compromised. This is less of an issue for bacteria, where the number of reference sequences for 16S is far greater.

Sequence similarity based methods, e.g. BLAST (Camacho *et al*. 2009), can give more specific classifications than statistical methods, but the results need to be post-processed to summarise ambiguities. For example, significant BLAST hits may be found for multiple species, so a reasonable summary is the lowest common ancestor (LCA) in the taxonomy (Huson *et al*. 2007; Medlar *et al*. 2014). BLAST-based classification suffers from the same issues as the RDP classifier, in that the lack of reference data can produce spuriously classifications. One final issue with BLAST is that, unless you have a curated database, you need to avoid significant hits from metagenomics and environmental samples found in public databases such as NCBI's NR database as these provide no information, but can crowd out well annotated sequences in the list of results.

We can try to overcome the lack of reference data using phylogenetic methods. Instead of classifying sequences directly based on their similarity to reference sequences, we can instead build a phylogenetic tree using full length gene sequences and extend that tree with short amplicons. The end result shows the evolutionary context of each sequence given an evolutionary model. These methods are implemented in pplacer (available in QIIME), which uses maximum likelihood (Matsen *et al*. 2010) and Séance, which performs evolutionary alignment extension (Medlar *et al*. 2014).

All methods for classification come with serious caveats, so we recommend using multiple methods. Aivelo *et al*. (2015) used both BLAST and phylogenetic placement, the concordance of which provided confidence or prompted further inspection. If in-house databases are used, parasite species identification needs to be based on morphological data and thus voucher samples are imperative, so identifications can be later verified (Janzen *et al*. 2009; Astrin *et al*. 2013). The same is true for any new sequence data added to the curated databases.

Statistical analysis

In comparison with many other applications of metabarcoding, helminth species richness is generally low: whereas the number of bacterial OTUs in mammals ranges in the hundreds, there is rarely more than 10 helminth species present. This allows for traditional statistical tools, including alpha and beta diversity comparisons, to be used in the analysis (e.g. Lott *et al*. 2015 and Aivelo *et al*. 2015). While in the early stages, metabarcoding techniques are mostly used to explore the presence or absence of certain taxa. As such, the data are vulnerable to false negatives: species that are present in the host, but due to egg shedding or unsuccessful sampling, went undetected. Previously mentioned replication can be used to reduce false negatives and the limits of detection can be modelled (Ficetola *et al*. 2015;

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

Furlan *et al*. 2016). Furthermore, while limits of detection can pose a significant risk with environmental DNA samples (Hunter *et al*. 2016), we would expect that the amount of parasite eggs or larvae is always proportionally high. Thus, the threshold for the sensitivity of the assay can be higher than for many environmental DNA applications.

Metabarcoding data can also be well-suited for community ecological analysis. Tools include dissimilarity matrices, which have been used extensively with bacterial 16S studies (Mills *et al*. 2006), or model-based ordination methods used for comparing sampling events (Warton *et al*. 2015). The collected metadata can be incorporated into models to see how much of the species associations are accounted for similar habitat requirements or phylogenetic relatedness and how much are accounted by other factors, including species interactions (Aivelo and Norberg, 2016).

Presence-absence analysis (see also 'Quantitative versus qualitative' later) also makes helminth metabarcoding studies more comparable, as presence-absence data are expected to vary less than community composition data. Nevertheless, the lack of definite species identification poses a clear problem: the choice of marker region and the parameters used for OTU clustering affect the number of OTUs. Thus, comparisons between metabarcoding studies should be done only when these variables are similar.

LIMITATIONS OF METABARCODING

Metabarcoding is by no means the silver bullet for in-situ species identification as some researchers may have hoped. The most significant limitation is that only a minority of species are described, a subset of which have been reliably sequenced and are present in public sequence databases (Wilson *et al*. 2011). Different methods of identifying OTUs, like incorporating evolutionary information, can resolve this problem partially; species can be differentiated even though they are not identified as a specific species. Despite these issues, things will improve over time as the coverage of barcode sequences and primers increases.

The second problem is far more difficult to solve: there are no truly universal primers, but a large number of taxon-specific primers (Deagle *et al*. 2014). Optimizing metabarcoding regions is difficult as the failure of some species to amplify is masked by successful amplifications of other species (Deagle *et al*. 2014). Furthermore, there is a tradeoff between coverage of species and resolution. For example, while nematode-specific primers can successfully amplify many nematode species, the resolution of the 18S gene region is often too low for differentiation at even the genus level (Aivelo, 2015). Also, as Deagle *et al*. (2014) points out, the variation quickly becomes saturated, meaning it is difficult to design primers, which are specific for only one phylum. The obvious solution for these problems would be to use longer amplicons or several primers, so called primer cocktails, to sequence several diagnostic gene regions (Prosser *et al*. 2013). For example, sequencing multiple gene targets might be necessary to successfully barcode all the intestinal nematodes present within one study. This makes the methods much more difficult to design and validate than the original conception of barcoding. Amplification-free approaches may solve the problem of amplification bias (Zhou *et al*. 2013), however, these approaches are limited by the quality of DNA required and the general problem of low ratio of DNA from taxa of interest in fecal matter compared with host or diet DNA. Nevertheless, diet analysis using shotgun sequencing seems to identify at least some helminths as a side-catch (Srivathsan *et al*. 2015, 2016).

The third problem is that sequencing error rates can be high and these errors can inflate diversity estimates (Meyer and Paulay, 2005; Quince *et al*. 2011). Of course, sequencing error profiles are only understood because they are easy to control for, but all stages of the metabarcoding pipeline interact with one another and can negatively impact results. For example, with DNA extraction, low quality DNA can induce downstream errors in amplification. Sequencing is affected by proper amplification, where the choice of region and length of amplicons dictates the chances of success. During sequencing, different platforms have different error profiles and finally in bioinformatics improper data handling will lead to inaccurate statistical inferences.

The fourth problem is that the metabarcoding is still costly and requires expert knowledge from a number of fields. The falling cost of high-throughput sequencing has quickly expanded its use to all areas of biology. Nevertheless, the costs are not trivial as the cost of sequencing 96 samples using the Illumina MiSeq platform, for example, costs approximately 2000 euros at the time of writing, and the costs of building sequencing libraries are much higher. Sequencing is not a trivial process either and the support necessary for planning and executing a metabarcoding study differs between research institutes. Metabarcoding sequence analysis also requires bioinformatics skills, so the researcher must be comfortable with and well-versed in the use of tools to analyse amplicon data, or else be able to collaborate or hire a specialist.

Fecal sampling

Many of the problems encountered during a metabarcoding study are not related to the barcoding approach per se, but to non-invasive surveys of parasites in general. Fecal analysis identifies parasites indirectly, as it can only detect parasites that are

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

laying eggs at the moment or, rather, at a certain time prior to sampling (Stear *et al*. 1995, 2006; Gillespie, 2006). Fecal analysis is also known to be less sensitive than terminal sampling or surveying intestines for helminth identification (Jorge *et al*. 2013). Nevertheless, in many situations, invasive sampling is not possible, for example, in longitudinal studies, when studying endangered species or if the value of acquired data is low compared with killing an animal. In these cases, metabarcoding can be an especially valuable method.

Not every amplicon in feces that matches a helminth sequence is actually an intestinal parasite. While other helminth groups are exclusively parasitic, identifying parasitic nematodes can be difficult: potentially contaminating soil nematodes and actual parasitic nematodes can be closely related and they might not be reliably identified if the reference databases are limited. As previously described, contaminations can come from many different sources: from the soil, from the prey, during the sample processing or during the DNA isolation and amplification. Especially in the field, contaminations may be unavoidable, despite researchers' best efforts.

Identification of species

Metabarcoding does not work directly with species, but with OTUs. As OTUs are defined as clusters of amplicons with high similarity, they can contain multiple species that happen to have sufficient sequence similarity by chance (Creer *et al*. 2010; Powers *et al*. 2011; Clare *et al*. 2016). OTU composition is therefore affected by the choice of marker gene region (Powers, 2004; Tang *et al*. 2012). The difficult relationship between OTUs and species could limit the research questions for which metabarcoding can provide answers. For most of community ecology, though, the species identity is not an essential question. In comparison, when carrying out epidemiological studies, it is important to link OTUs with pathogenic outcomes. Also, it should be noted that sharing the same OTU in two different host species does not necessarily mean that both host species share the same parasite species. Metabarcoding is therefore inappropriate for studies interested in parasite sharing between different host species, unless additional taxon-specific genomic regions with higher resolution for species identification are used.

While metabarcoding approaches have quickly become more common, it has not been widely discussed what community ecology concepts, like diversity or richness means in relation to OTUs. Within microbiota research, these concepts, along with different beta diversity metrics, have been widely used, but critical consideration of the importance of different metrics has been rare (Dethlefsen *et al*. 2007; Robinson *et al*. 2010). Nevertheless, community ecology is becoming more central as there is more and more data on different host-associated systems (Belden and Harris, 2007; Costello *et al*. 2012; De Schryver and Vadstein, 2014; Trosvik and de Muinck, 2015)

Quantitative vs qualitative

If contamination and limits of detection are taken into account, metabarcoding can provide reliable presence-absence data on helminth communities. In contrast, using metabarcoding for quantitative assessment of helminth communities within their hosts is more challenging (Gasser, 2006; Deagle *et al*. 2013). For gastrointestinal parasites, parasite load is defined as the number of parasites within a host individual and is therefore measured with invasive techniques such as the dissection of the gastrointestinal tract to count the number of adult parasites (Poulin and Morand, 2000). Measurement of parasite load non-invasively is bound to encounter the same problems as assessing parasite presence: parasite load can be assessed only indirectly (Jorge *et al*. 2013). The most commonly used method for non-invasive quantification of gastrointestinal helminth load is to quantify the parasite eggs in fecal matter. Fecal egg count (FEC, or eggs per gram, EPG) is often used as a proxy for parasite load but this is obviously problematic as FEC does not necessarily correlate with the number of parasite individuals within host (Stear *et al*. 1995, 2006; Gillespie, 2006).

The number of eggs found in feces varies daily and can only be used as an indicator for prolific egg layers (Prichard and Tait, 2001). Even if we assume that the ratio of egg biomass is equal to adult parasite biomass, different species have different quantities of DNA within their eggs. Amplification success rates differ due to variation in the number of ribosomal and mitochondrial genes and the choice of DNA isolation methods, primers and sequencing affects the resulting community composition (Fouhy *et al*. 2016). Depending on the amplification or sequencing steps, there can be different error rates within different amplicons leading to quality control discarding proportionally more amplicons from certain species.

Nevertheless, there are similar frustrations in 16S microbial sequencing and the bacterial microbiota research community has applied quantitative methods within these limitations (Brooks *et al*. 2015; de la Cuesta-Zuluaga and Escobar, 2016). We believe that helminth amplicon frequency can be used as a crude measure of abundance and is comparable across samples given appropriate design; ensuring that sample processing is uniform. The veracity of results from different quantitative analysis will wary in different host-parasite systems and depends on both the biology of the system and the amount of methodological validation carried out. There has been some

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

progress in quantitative diet analysis, with well-validated correction factors (Thomas *et al*. 2016), but further investigation is necessary to understand whether a similar approach is feasible in helminth metabarcoding.

FUTURE RESEARCH DIRECTIONS

As we have outlined above, metabarcoding provides both opportunities for faster and more extensive community composition analysis, but – like any method – has limitations that need to be taken into consideration. While not all scientific questions about helminth communities can be answered with metabarcoding, some cannot be addressed in any other manner. Indeed, the promise of metabarcoding lies in situations where traditional approaches cannot be used effectively. To summarize the possibilities for parasitology, we outline future research directions where metabarcoding can be employed and identify methodological questions that need to be resolved.

Parasite community analysis in animals that are difficult to capture

There are numerous reasons why some animals are not feasible to capture: they might be large or the act of capture might be unnecessarily disturbing. In these cases researchers might wait patiently nearby while the animal defecates and then collect the sample. Alternately, parasite analysis could be performed using feces found opportunistically on the ground. Metabarcoding could provide more power to fecal analysis.

Nevertheless, there is little data on how different sampling methods compare with each other and how retrievable parasite community structures are. We do not yet know, how fresh feces needs to be and do not know all the circumstances that lead to contamination, for example, when collecting feces from the ground. More methodological studies would hopefully lead to consensus as to what are the optimal methods in relation to work needed, price, yield and quality of DNA, as high-throughput methods rely on the usability of fecal DNA.

Longitudinal analysis of parasite community composition

Invasive sampling is clearly impossible in longitudinal analysis of parasite communities within host individuals. While the modus operandi of parasitological research has been cross-sectional studies by killing and invasively sampling host individuals, the succession of parasite communities within host (component communities) has rarely been studied. This has been very difficult, as the required taxonomic accuracy has not been attainable by morphological methods, whereas genetic methods can be used to link successive parasite samples to the same OTUs. While the repeatability of parasite measures from individual hosts has been low (Stear *et al*. 1995), longitudinal sampling provides a more reliable picture of the parasite community. In turn, this longitudinal data on variation in within-host parasite communities can be linked to other measurements of individual hosts, including fitness, body condition, behaviour or genetic data.

Integration with phylogenetic information

One of the benefits of metabarcoding is that it also provides phylogenetic information. If the OTUs are placed in a phylogenetic tree, this could be used to provide a more concise picture on the phylogenetic relationships between metabarcoded parasites. Furthermore, this could be used in statistical analysis by taking phylogenetic distances into account. This is already common in microbiota research, but thus far rarely used in parasitological research (Bordes and Morand, 2009; Poulin, 2010; Xu *et al*. 2013).

Previously unknown parasite communities

Metabarcoding can also be useful in studying host populations, of which we do not know the parasite assemblages. Even if the parasite species have not been described, OTUs can be defined much more accurately than morphospecies (depending on the marker region). This potentially provides far more opportunities in most host species, which are rarely studied and whose parasites are unknown. This could also enable broadening many classical ecological studies, which have not yet looked at the host-related symbionts, to take into account parasite communities.

CONCLUDING REMARKS

We have outlined several reasons for adopting metabarcoding practices to study helminths in wild mammalian hosts, including faster and more accurate identification of parasites, especially from fecal samples, which have been previously difficult to analyze. Based on our own experiences and the few other helminth metabarcoding studies, we outline challenges and considerations in sampling in the field, laboratory work and data analysis. There are certain limitations with the metabarcoding approach, like the lack of truly universal primers, problems with contamination and the challenges to quantify relative abundances. We also suggest potential future research directions, which could benefit from metabarcoding, especially longitudinally surveying of host individuals or studies on endangered species, which could be done with non-invasive fecal sampling.

https://doi.org/10.1017/S0031182017000610 Published online by Cambridge University Press

ACKNOWLEDGEMENTS

We thank two anonymous reviewers on their comments on the manuscript.

FINANCIAL SUPPORT

This work was supported by the Finnish Cultural Foundation (T.A.) and Oskar Öflunds Stiftelse (T.A.).

REFERENCES

Aivelo, T. (2015). Longitudinal Monitoring of Parasites in Individual Wild Primates. PhD thesis, University of Helsinki, Helsinki, Finland.

Aivelo, T. and Norberg, A. (2016). Parasite-microbiota interactions potentially affect intestinal communities in wild mammals. bioRxiv, New York, NY.

Aivelo, T., Medlar, A., Löytynoja, A., Laakkonen, J. and Jernvall, J. (2015). Tracking year-to-year changes in intestinal nematode communities of rufous mouse lemurs (Microcebus rufus). *Parasitology* **142**, 1095–1107.

Archie, E. A. and Ezenwa, V. O. (2011). Population genetic structure and history of a generalist parasite infecting multiple sympatric host species. *International Journal for Parasitology* **41**, 89–98.

Asmundsson, I. M., Mortenson, J. A. and Hoberg, E. P. (2008). Muscleworms, *Parelaphostrongylus andersoni* (Nematoda: Protostrongylidae), discovered in Columbia white-tailed deer from Oregon and Washington: implications for biogeography and host associations. *Journal of Wildlife Diseases* **44**, 16–17.

Astrin, J. J., Zhou, X. and Misof, B. (2013). The importance of biobanking in molecular taxonomy, with proposed definitions for vouchers in a molecular context. *ZooKeys* **365**, 67–70.

Avramenko, R. W., Redman, E. M., Lewis, R., Yazwinski, T. A., Wasmuth, J. D. and Gilleard, J. S. (2015). Exploring the gastrointestinal "nemabiome": Deep amplicon sequencing to quantify the species composition of parasitic nematode communities. *PLoS ONE* **10**, 1–18.

Belden, L. K. and Harris, R. N. (2007). Infectious the diseases in wildlife: ecology context community. *Frontiers in Ecology and the Environment* **5**, 533–539.

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R. and Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution* **27**, 233–243.

Blasco-Costa, I., Cutmore, S. C., Miller, T. L. and Nolan, M. J. (2016). Molecular approaches to trematode systematics: "best practice" and implications for future study. *Systematic Parasitology* **93**, 295–306.

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

Bordes, F. and Morand, S. (2009). Coevolution between multiple helminth infestations and basal immune investment in mammals: cumulative effects of polyparasitism? *Parasitology Research* **106**, 33–37.

Bordes, F. and Morand, S. (2011). The impact of multiple infections on wild animal hosts: a review. *Infection Ecology & Epidemiology* **1**, 7346.

Bretagne, S., Guillou, J. P., Morand, M. and Houin, R. (1993). Detection of *Echinococcus multilocularis* DNA in fox faeces using DNA amplification. *Parasitology* **106**, 193–199.

Brooks, J. P., Edwards, D. J., Harwick, M. D., Jr., Rivera, M. C., Fettweis, J. M., Serrano, M. G., Reris, R. A., Sheth, N. U., Huang, B., Girerd, P., Vaginal Microbiome Consortium, Strauss, J. F., III, Jefferson, K. K. and Buck, G. A. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studes. BMC Microbiology 15, 66.

Budischak, S. A, Hoberg, E. P., Abrams, A., Jolles, A. E. and Ezenwa, V. O. (2015). A combined parasitological molecular approach for noninvasive characterization of parasitic nematode communities in wild hosts. *Molecular Ecology Resources* **15**, 1112–1119.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Gonzalez Peña, A., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. (2011). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336.

Caron, D. A. (2009). New accomplishments and approaches for assessing protistan diversity and ecology in natural ecosystems. *BioScience* **59**, 287–299.

Casiraghi, M., Labra, M., Ferri, E., Galimberti, A. and De Mattia, F. (2010). DNA barcoding: a six-question tour to improve users' awareness about the method. *Briefings in Bioinformatics* **11**, 440–453.

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

Collins, R. A. and Cruickshank, R. H. (2013). The seven deadly sins of DNA barcoding. *Molecular Ecology Resources* **13**, 969–975.

Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M. and Relman, D. A. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science* **336**, 1255–1263.

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J. and Arnaud-Haond, S. (2015). Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS ONE* **10**, 1–26.

Creer, S., Fonseca, V. G., Porazinska, D. L., Giblin-Davis, R. M., Sung, W., Power, D. M., Packer, M., Carvalho, G. R., Blaxter, M. L., Lambshead, P. J. D. and Thomas, W. K. (2010). Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology* **19** (Suppl. 1), 4–20.

Dawkins, H. J. S. and Spencer, T. L. (1989). The isolation of nucleic acid from nematodes requires an understanding of the parasite and its cuticular structure. *Parasitology Today* **5**, 73–76.

de la Cuesta-Zuluaga, J. & Escobar, J. (2016). Considerations for optimizing microbiome analysis using a marker gene. *Frontiers in Nutrition* **3**, 26.

De Schryver, P. and Vadstein, O. (2014). Ecological theory as a foundation to control pathogenic invasion in aquaculture. *ISME Journal* **8**, 2360–2368.

Deagle, B. E., Kirkwood, R. and Jarman, S. N. (2009). Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology* **18**, 2022–2038.

Deagle, B. E., Thomas, A. C., Shaffer, A. K., Trites, A. W. and Jarman, S. N. (2013). Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources* **13**, 620–633.

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

Derycke, S., Vanaverbeke, J., Rigaux, A., Backeljau, T. and Moens, T. (2010). Exploring the use of cytochrome oxidase c subunit 1 (COI) for DNA barcoding of free-living marine nematodes. *PLoS ONE* **5**, e13716.

DeSalle, R., Egan, M. G. and Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society of London B: Biological sciences* **360**, 1905–1916.

Dethlefsen, L., McFall-Ngai, M. and Relman, D. A. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449**, 811–818.

Ebach, M. C. and Holdrege, C. (2005). More taxonomy, not DNA barcoding. *BioScience* **55**, 822–823.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461.

Ferri, E., Barbuto, M., Bain, O., Galimberti, A., Uni, S., Guerrero, R., Ferté, H., Bandi, C., Martin, C. and Casiraghi, M. (2009). Integrated taxonomy: traditional approach and DNA barcoding for the identification of filarioid worms and related parasites (Nematoda). *Frontiers in Zoology* **6**, 1.

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G. and Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources* **15**, 543–556.

Floyd, R., Abebe, E., Papert, A. and Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology* **11**, 839–850.

Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J. and Cotter, P. D. (2016). 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology* **16**, 123.

Frézal, L. and Leblois, R. (2008). Four years of DNA barcoding: current advances and prospects. *Infection, Genetics and Evolution* **8**, 727–736.

Furlan, E. M., Gleeson, D., Hardy, C. M. and Duncan, R. P. (2016). A framework for estimating the sensitivity of eDNA surveys. *Molecular Ecology Resources* **16**, 641–654.

Galimberti, A., Romano, D. F., Genchi, M., Paoloni, D., Vercillo, F., Bizzarri, L., Sassera, D., Bandi, C., Genchi, C., Ragni, B. and Casiraghi, M. (2012). Integrative taxonomy at work: DNA barcoding of taeniids harboured by wild and domestic cats. *Molecular Ecology Resources* **12**, 403–413.

Gasser, R. B. (2006). Molecular tools – advances, opportunities and prospects. *Veterinary Parasitology* **136**, 69–89.

Gasser, R. B., Chilton, N. B., Hoste, H. and Beveridge, I. (1993). Rapid sequencing of rDNA from single worms and eggs of parasitic helminths. *Nucleic Acids Research* **21**, 2525–2526.

Gasser, R. B., Bott, N. J., Chilton, N. B., Hunt, P. and Beveridge, I. (2008). Toward practical, DNA-based diagnostic methods for parasitic nematodes of livestock--bionomic and biotechnological implications. *Biotechnology Advances* **26**, 325–334.

Gaugler, R., Bednarek, A. and Campbell, J. F. (1992). Ultraviolet inactivation of heterorhabditid and steinernematid nematodes. *Journal of Invertebrate Pathology* **59**, 155–160.

Gillespie, T. R. (2006). Noninvasive assessment of gastrointestinal parasite infections in free-ranging primates. *International Journal of Primatology* **27**, 1129–1143.

Goldstein, P. Z. and DeSalle, R. (2011). Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *BioEssays* **33**, 135–147.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methé, B., DeSantis, T. Z., Petrosino, J. F., Knight, R. and Birren, B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* **21**, 494–504.

Harmon, A. F., Zarlenga, D. S. and Hildreth, M. B. (2006). Improved methods for isolating DNA from *Ostertagia ostertagi* eggs in cattle feces. *Veterinary Parasitology* **135**, 297–302.

Hebert, P. D. N., Cywinska, A., Ball, S. L. and DeWaard, J. R. (2003*a*). Biological identifications through DNA barcodes. *Proceedings. Biological sciences/Royal Society* **270**, 313–321.

Hebert, P. D. N., Ratnasingham, S. and DeWaard, J. R. (2003*b*). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences* **270**, S96–S99.

Hu, M., Jex, A. R., Campbell, B. E. and Gasser, R. B. (2007). Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. *Nature Protocols* **2**, 2339–2344.

Hunter, M. E., Dorazio, R. M., Butterfield, J. S. S., Meigs-Friend, G., Nico, L. G. and Ferrante, J. A. (2016). Detection limits of quantitative and digital PCR assays and their influence in presence-absence surveys of environmental DNA. *Molecular Ecology Resources* **17**, 221–229.

Huson, D. H., Auch, A. F., Qi, J., Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research* **17**, 377–386.

Janzen, D. H., Hallwachs, W., Blandin, P., Burns, J. M., Cadiou, J. M., Chacon, I., Dapkey, T., Deans, A. R., Epstein, M. E., Espinoza, B., Franclemont, J. G., Haber, W. A., Hajibabaei, M., Hall, J. P. W., Hebert, P. D. N., Gauld, I. D., Harvey, D. J., Hausmann, A., Kitching, I. J., Lafontaine, D., Landry, J. F., Lemaire, C., Miller, J. Y., Miller, J. S., Miller, L., Miller, S. E., Montero, J., Munroe, E., Green, S. R., Ratnasingham, S. *et al.* (2009). Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources* **9**, 1–26.

Jarman, S. N., Deagle, B. E. and Gales, N. J. (2004). Group-specific polymerase chain reaction for DNA-based analysis of species diversity and identity in dietary samples. *Molecular Ecology* **13**, 1313–1322.

Jenkins, E. J., Appleyard, G. D., Hoberg, E. P., Rosenthal, B. M., Kutz, S. J., Veitch, A. M., Schwantje, H. M., Elkin, B. T. and Polley, L. (2005). Geographic distribution of the muscle-dwelling nematode *Parelaphostrongylus odocoilei* in North America, using molecular identification of first-stage larvae. *Journal of Parasitology* **91**, 574–584.

Jorge, F., Carretero, M. A., Roca, V., Poulin, R. and Perera, A. (2013). What you get is what they have? Detectability of intestinal parasites in reptiles using faeces. *Parasitology Research* **112**, 4001–4007.

King, R. A., Read, D. S., Traugott, M. and Symondson, W. O. C. (2008). Molecular analysis of predation: a review of best practice for DNA-based approaches. *Molecular Ecology* **17**, 947–963.

Krauth, S. J., Coulibaly, J. T., Knopp, S., Traoré, M., N'Goran, E. K. and Utzinger, J. (2012). An in-depth analysis of a piece of shit: distribution of *Schistosoma mansoni* and hookworm eggs in human stool. *PLoS Neglected Tropical Diseases* **6**, e1969.

Kutz, S. J., Checkley, S., Verocai, G. G., Dumond, M., Hoberg, E. P., Peacock, R., Wu, J. P., Orsel, K., Seegers, K., Warren, A. L. and Abrams, A. (2013). Invasion, establishment, and range expansion of two parasitic nematodes in the Canadian Arctic. *Global Change Biology* **19**, 3254–3262.

Lee, M. S. Y. (2004). The molecularisation of taxonomy. *Invertebrate Systematics* **18**, 1–6.

Leung, T. L. F., Donald, K. M., Keeney, D. B., Koehler, A. V., Peoples, R. C. and Poulin, R. (2009). Trematode parasites of Otago Harbour (New Zealand) soft-sediment intertidal ecosystems: life cycles, ecological roles and DNA barcodes. *New Zealand Journal of Marine and Freshwater Research* **43**, 857–865.

Locke, S. A., McLaughlin, J. D., Lapierre, A. R., Johnson, P. T. J. and Marcogliese, D. J. (2011). Linking larvae and adults of *Apharyngostrigea cornu*, *Hysteromorpha triloba*, and *Alaria mustelae* (Diplostomoidea: Digenea) using molecular data. *Journal of Parasitology* **97**, 846–851.

Lott, M. J., Eldridge, M. D. B, Hose, G. C. and Power, M. L. (2012). Nematode community structure in the brush-tailed rock-wallaby, Petrogale penicillata: Implications of captive breeding and the translocation of wildlife. *Experimental Parasitology* **132**, 185–192.

Lott, M. J., Hose, G. C. and Power, M. L. (2015). Parasitic nematode communities of the red kangaroo, *Macropus rufus*: richness and structuring in captive systems. *Parasitology Research* **114**, 2925–2932.

Martin, L. K. and Beaver, P. C. (1968). Evaluation of Kato thick-smear technique for quantitative diagnosis of helminth infections. *American Journal of Tropical Medicine and Hygiene* **17**, 382–391.

Mathis, A. and Deplazes, P. (2006). Copro-DNA tests for diagnosis of animal taeniid cestodes. *Parasitology International* **55**, S87–S90.

Matsen, F. A., Kodner, R. B. and Armbrust, E. V. (2010). pplacer : linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree pplacer : linear time maximum-likelihood and Bayesian phy- logenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538.

Medlar, A., Aivelo, T. and Löytynoja, A. (2014). Séance : reference-based phylogenetic analysis for 18S rRNA studies. *BMC Evolutionary Biology* **14**, 235.

Mes, T. H. M. (2003). Technical variability and required sample size of helminth egg isolation procedures. *Veterinary Parasitology* **115**, 311–320.

Meyer, C. P. and Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology* **3**, e422.

Mills, D. K., Entry, J. A., Voss, J. D., Gillevet, P. M. and Mathee, K. (2006). An assessment of the hypervariable domains of the 16S rRNA genes for their value in determining microbial community diversity: the paradox of traditional ecological indices. *FEMS Microbiology Ecology* **57**, 496–503.

Mitchell, A. (2011). DNA barcoding is useful for taxonomy: a reply to Ebach. *Zootaxa* **2772**, 67–68.

Moszczynska, A., Locke, S. A., McLaughlin, J. D., Marcogliese, D. J. and Crease, T. J. (2009). Development of primers for the mitochondrial cytochrome c oxidase I gene in digenetic trematodes (Platyhelminthes) illustrates the challenge of barcoding parasitic helminths. *Molecular Ecology Resources* **9**, 75–82.

Ogedengbe, J. D., Hanner, R. H. and Barta, J. R. (2011). DNA barcoding identifies Eimeria species and contributes to the phylogenetics of coccidian parasites (Eimeriorina, Apicomplexa, Alveolata). *International Journal for Parasitology* **41**, 843–850.

Porazinska, D. L., Giblin-Davis, R. M., Faller, L., Farmerie, W., Kanzaki, N., Morris, K., Powers, T. O., Tucker, A. E., Sung, W. and Thomas, W. K. (2009). Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Molecular Ecology Resources* **9**, 1439–1450.

Poulin, R. (2010). Decay of similarity with host phylogenetic distance in parasite faunas. *Parasitology* **137**, 733–741.

Poulin, R. and Morand, S. (2000). The diversity of parasites. *Quarterly Review of Biology* **75**, 277–293.

Powers, T. (2004). Nematode molecular diagnostics: from bands to barcodes. *Annual Review of Phytopathology* **42**, 367–383.

Powers, T., Harris, T., Higgins, R., Mullin, P., Sutton, L. and Powers, K. (2011). MOTUs, morphology, and biodiversity estimation: a case study using nematodes of the suborder Criconematina and a conserved 18S DNA barcode. *Journal of Nematology* **43**, 35–48.

Prichard, R. and Tait, A. (2001). The role of molecular biology in veterinary parasitology. *Veterinary Parasitology* **98**, 169–194.

Prosser, S. W. J., Velarde-Aguilar, M. G., León-Règagnon, V. and Hebert, P. D. N. (2013). Advancing nematode barcoding: a primer cocktail for the cytochrome c oxidase subunit I gene from vertebrate parasitic nematodes. *Molecular Ecology Resources* **13**, 1108–1115.

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., Gu, Y., Rothberg, J., Hinz, W., Rearick, T., Schultz, J., Mileski, W., Davey, M., Leamon, J., Johnson, K., Milgrew, M., Edwards, M., Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P.,

Rank, D., Baybayan, P., Bettman, B., Bentley, D. *et al*. (2012). A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13**, 341.

Quince, C., Lanzen, A., Davenport, R. J. and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**, 38.

Robinson, C. J., Bohannan, B. J. M. and Young, V. B. (2010). From structure to function: the ecology of host-associated microbial communities. *Microbiology and Molecular Biology Reviews* **74**, 453–476.

Routtu, J., Grunberg, D., Izhar, R., Dagan, Y., Guttel, Y., Ucko, M. and Ben-Ami, F. (2014). Selective and universal primers for trematode barcoding in freshwater snails. *Parasitology Research* **113**, 2535–2540.

Santos, A. M. C., Besnard, G. and Quicke, D. L. J. (2011). Applying DNA barcoding for the study of geographical variation in host-parasitoid interactions. *Molecular Ecology Resources* **11**, 46–59.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537–7541.

Schloss, P. D., Gevers, D. and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing Artifacts on 16s rRNA-based studies. *PLoS ONE* **6**, e27310.

Shokralla, S., Spall, J. L., Gibson, J. F. and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology* **21**, 1794–1805.

Siddall, M. E., Kvist, S., Phillips, A. and Oceguera-Figueroa, A. (2012). DNA barcoding of parasitic nematodes: is it kosher? *Journal of Parasitology* **98**, 692–694.

Silva, N. R. R., Da Silva, M. C., Genevois, V. F., Esteves, A. M., Ley, P., De Decraemer, W., Rieger, T. T. and Correia, M. T. D. S. (2010). Marine nematode taxonomy in the age of DNA: the present and future of molecular tools to assess their biodiversity. *Nematology* **12**, 661–672.

Smith, D. P. and Peay, K. G. (2014). Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS ONE* **9**, e90234.

Srivathsan, A., Sha, J. C. M., Vogler, A. P. and 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.

Srivathsan, A., Ang, A., Vogler, A. P. and Meier, R. (2016). Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. *Frontiers in Zoology* **13**, 17.

Stear, M. J., Bishop, S. C., Duncan, J. L., Mckellar, Q. A. and Murray, M. (1995). The repeatability of faecal egg counts, peripheral eosinophil counts, and plasma pepsinogen concentrations during deliberate infection with *Ostertagia circumcincta*. *International Journal for Parasitology* **25**, 375–380.

Stear, M. J., Abuagob, O., Benothman, M., Bishop, S. C., Innocent, G., Kerr, A. and Mitchell, S. (2006). Variation among faecal egg counts following natural nematode infection in Scottish Blackface lambs. *Parasitology* **132**, 275–280.

Symondson, W. O. C. (2002). Molecular identification of prey in predator diets. *Molecular Ecology* **11**, 627–641.

Tanaka, R., Hino, A., Tsai, I. J., Palomares-Rius, J. E., Yoshida, A., Ogura, Y., Hayashi, T., Maruyama, H. and Kikuchi, T. (2014). Assessment of helminth biodiversity in wild rats using 18S rDNA based metagenomics. *PLoS ONE* **9**, e110769.

Tang, C. Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T. G. and Fontaneto, D. (2012). The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 16208–16212.

Taylor, H. R. and Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources* **12**, 377–388.

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H. and Trites, A. W. (2016). Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources* **16**, 714–726.

Trosvik, P. and de Muinck, E. J. (2015). Ecology of bacteria in the human gastrointestinal tract--identification of keystone and foundation taxa. *Microbiome* **3**, 44.

Valentini, A., Pompanon, F. and Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution* **24**, 110–117.

Vanhove, M. P. M., Tessens, B., Schoelinck, C., Jondelius, U., Littlewood, D. T. J., Artois, T. and Huyse, T. (2013). Problematic barcoding in flatworms: a case-study on monogeneans and rhabdocoels (Platyhelminthes). *ZooKeys* **365**, 355–379.

Van Steenkiste, N., Locke, S. A., Castelin, M., Marcogliese, D. J. and Abbott, C. L. (2015). New primers for DNA barcoding of digeneans and cestodes (Platyhelminthes). *Molecular Ecology Resources* **15**, 945–952.

Viney, M. E. and Graham, A. L. (2013). Patterns and processes in parasite co-infection. *Advances in Parasitology* **82**, 321–369.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**, 5261–5267.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C. and Hui, F. K. C. (2015). So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution* **30**, 766–779.

Wilson, J. J., Rougerie, R., Schonfeld, J., Janzen, D. H., Hallwachs, W., Hajibabaei, M., Kitching, I. J., Haxaire, J. and Hebert, P. D. N. (2011). When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology* **11**, 18.

Woodstock, L., Cook, J. A., Peters, P. A. and Warren, K. S. (1971). Random distribution of schistosome eggs in the feces of patients with *Schistosomiasis mansoni*. *Journal of Infectious Diseases* **124**, 613–614.

Xu, B., Xu, W., Yang, F., Li, J., Yang, Y., Tang, X., Mu, Y., Zhou, J. and Huang, Z. (2013). Metagenomic analysis of the pygmy loris fecal microbiome reveals unique functional capacity related to metabolism of aromatic compounds. *PLoS ONE* **8**, e56565.

Ye, X. P., Donnelly, C. A., Anderson, R. M., Fu, Y. L. and Agnew, A. (1998). The distribution of *Schistosoma japonicum* eggs in faeces and the effect of stirring faecal specimens. *Annals of Tropical Medicine and Parasitology* **92**, 181–185.

Yu, J. M., de Vlas, S. J., Yuan, H. C., Gryseels, B. (1998). Variations in fecal *Schistosoma japonicum* egg counts. *American Journal of Tropical Medicine and Hygiene* **59**, 370–375.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J. and Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Giga Science* **2**, 4.