DR. BRUCE E DEAGLE (Orcid ID : 0000-0001-7651-3687)

MR. EERO JUHANI VESTERINEN (Orcid ID : 0000-0003-3665-5802)

DR. ELIZABETH LLOYD CLARE (Orcid ID : 0000-0002-6563-3365)

Article type : Special Issue

Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data?

Bruce E. Deagle* (corresponding author: Bruce.Deagle@aad.gov.au)

Austen C. Thomas†

Julie C. McInnes*

Laurence J. Clarke‡*

Eero J. Vesterinen[1]¶

Elizabeth L. Clare[2]

Tyler R. Kartzinel[3]

J. Paige Eveson§

*Australian Antarctic Division, Channel Highway, Kingston, Tasmania, Australia

†Science Department, Smith-Root Inc., Vancouver, Washington, USA

‡ Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Tasmania, Australia

[1] Biodiversity Unit and Department of Biology, University of Turku, Turku, Finland

¶ Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland

[2] School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

[3] Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island, USA

§CSIRO Oceans and Atmosphere, GPO Box 1538, Hobart, Tasmania, Australia

Abstract

Advances in DNA sequencing technology have revolutionised the field of molecular analysis of trophic interactions and it is now possible to recover counts of food DNA sequences from a wide range of dietary samples. But what do these counts mean? To obtain an accurate estimate of a consumer's diet should we work strictly with datasets summarising frequency of occurrence of different food taxa, or is it possible to use relative number of sequences? Both approaches are applied to obtain semi-quantitative diet summaries, but occurrence data is often promoted as a more conservative and reliable option due to taxa-specific biases in recovery of sequences. We explore representative dietary metabarcoding datasets and point out that diet summaries based on occurrence data often overestimate the importance of food consumed in small quantities (potentially including low-level contaminants) and are sensitive to the count threshold used to define an occurrence. Our simulations indicate that using relative read abundance (RRA) information often provide a more accurate view of population-level diet even with moderate recovery biases incorporated; however, RRA summaries are sensitive to recovery biases impacting common diet taxa. Both approaches are more accurate when the mean number of food taxa in samples is small. The ideas presented here highlight the need to consider all sources of bias and to justify the methods used to interpret count data in dietary metabarcoding studies. We encourage researchers to continue addressing methodological challenges, and

acknowledge unanswered questions to help spur future investigations in this rapidly developing area of research.

1. Introduction

Many recent studies documenting trophic interactions make use of metabarcoding, an approach which combines high-throughput sequencing (HTS) with DNA barcoding to identify the food remains present in faecal samples or stomach contents (Nielsen *et al.* 2017). When HTS first became available the potential applications in diet studies were clear and the methods were quickly embraced by the community (Deagle *et al.* 2009; Valentini *et al.* 2009). In a comprehensive review of DNA-based diet analysis by King *et al*. (2008) the possibility of using HTS was only briefly mentioned as a 'Future Direction', and just four years later another review paper focussed entirely on this approach (Pompanon *et al.* 2012). While many underlying technical and biological details vary between dietary metabarcoding studies, the general workflow is now well defined. It involves extraction of total DNA from the dietary sample, PCR amplification of DNA barcode markers from food taxa of interest, and then DNA sequencing for taxonomic classification of the recovered sequences. The workflow has been applied to determine diet in a range of animals, from invertebrates to large mammalian herbivores and carnivores (representative studies summarised in Table 1).

The rapid adoption of HTS to characterise complex mixtures of DNA is not unique to dietary studies; over the last ten years the technology has produced a wealth of new genetic data providing insight into almost all areas of biology (Goodwin *et al.* 2016). One feature of HTS is that it provides counts of DNA sequences in each sample and therefore it has the potential not only to provide a qualitative list, but also to quantify what DNA is present. The interpretation of sequence read counts as a numerical representation of sample composition is common in many HTS applications. For example, studies sequencing transcripts to determine differences in gene expression (Finotello & Di Camillo 2015), profiling microbe communities (Vandeputte *et al.* 2017) or measuring epigenetic variation (Schield *et al.* 2016) all rely on sequence read counts. However, decisions about how to interpret read counts is certainly not routine and the validity of interpretations is sometimes questioned even in fields where the practice is well established (e.g. Edgar 2017; Olova *et al.* 2017). These debates are constructive, and should motivate researchers to test underlying assumptions and justify their interpretations, but can give rise to the impression that count data are always misleading.

The reality is that all metabarcoding studies use sequence counts to some extent. In dietary investigations, count data are used either to record the occurrence of food species within samples based on a threshold number of sequences (i.e. presence/absence of taxa), or to calculate the percentage of DNA belonging to each food species as a proxy for relative biomass consumed (i.e. relative abundance of taxa; Figure 1). The conversion of sequence counts to occurrence data is often considered a more conservative approach than using proportional data. In their introduction to the Molecular Ecology Special Issue on 'Molecular Detection of Trophic Interactions', Symondson & Harwood (2014) pointed out that authors of many metabarcoding papers "*now simply record numbers of predators testing positive for a target prey or plant species, providing a pragmatic and useful surrogate for truly quantitative information*". This sentiment, that focusing only on occurrence data is a reliable and safe option, is now common in the literature and this step in the analysis pipeline is often uncritically applied as the default option. Using counts as an indication of biomass in samples is more controversial. Indeed, the difficulties of obtaining an accurate biomass signature from sequence counts include both technical and biological biases that affect barcode marker recovery rates from different taxa (Amend *et al.* 2010; Deagle *et al.* 2009; Pompanon *et al.* 2012). Therefore in the best-case scenario sequence read counts can only provide a rough estimate of proportional abundance. Still, to accept the notion that relative sequence counts provide no meaningful information would mean that, within one sample, a few DNA sequences from one food taxon is equivalent to 10,000 sequences from another. Most molecular ecologists would interpret these disparate counts to mean that there are differences in template DNA abundance (as long as methods used to collect the data are reasonable) and that there is some biological basis for that difference. Ignoring this difference may inhibit ecological understanding.

Here, we review the approaches taken to interpret sequence count data in dietary metabarcoding studies and consider their implications. Throughout the paper we will refer to the two general approaches as 'occurrence' (i.e. presence/absence of taxa) and 'relative read abundance' (RRA; i.e. proportional summaries of counts). The end product of both methods is the same, a semi-quantitative surrogate for the true diet, and our goal is to critically evaluate these different interpretations. We point out that converting sequence read counts to occurrence information can introduce strong biases and thus we suggest it is not always a "conservative" approach. We also assess the scale of biases in recovery of sequences from different food taxa in study systems where it has been examined. Using simulations, we explore the impact of these biases on data summaries (both based on occurrence and read counts). In this light, we evaluate factors that impact dietary metabarcoding data summaries and consider when using sequence count data as an indication of relative biomass within samples might be justified to provide a more nuanced picture of animal diet.

The issues we consider on how best to summarise dietary data have implications for all metabarcoding studies (Taberlet *et al.* 2018) and similar issues have been considered extensively in traditional diet studies (e.g. Barrett *et al.* 2007; Laake *et al.* 2002). In HTS-based diet studies the ideas are most relevant when the underlying objective is to estimate the true diet of a particular consumer (i.e. the relative biomass contributions of alternative diet species). This may not be a clearly stated goal, but is often implicit in outcomes of dietary metabarcoding studies. Approaches for summarising sequence counts may be of less concern in studies aiming to provide a list of taxa consumed by a particular species (niche breadth), a qualitative description of trophic interactions in a food web, or an indicator of dietary differences between sites. We focus mainly on dietary studies using DNA extracted from faecal material. The use of HTS to identify food in stomach contents is common in invertebrates, and also fish, but the material recovered is in various states of digestion and the sequence counts are less likely to contain a meaningful quantitative signal based on RRA compared to the more consistent signal seen in faecal material (Deagle *et al.* 2013; Nakahara *et al.* 2015).

2. Current Practice

Non-dietary metabarcoding studies use a range of approaches to interpret sequence count data, and these vary depending on the targeted organisms. Recent papers published in Molecular Ecology on bacterial/archaeal communities all make use of RRA, although half of these studies also presented summaries based on taxon occurrences (Table S1). There is widespread acknowledgement of taxon-specific biases in recovery of the bacterial/archaeal barcode markers, but RRA is accepted as a flawed, but useful, measure of these diverse communities that cannot be easily characterized by other means (Forney *et al.* 2004; Ibarbalz *et al.* 2014). There is no clear consensus in metabarcoding of eukaryotic communities: RRA is sometimes used exclusively (often the case in studies of fungi), whereas metazoan studies use either occurrence data only or both metrics in tandem (recent examples listed in Table S2).

In dietary metabarcoding studies, it is common to only interpret sequence data after conversion to taxon occurrences (representative studies summarised in Table 1). This conversion is done in various ways. During initial processing of sequence reads, most researchers discard rare sequences to avoid incorporation of background sequencing errors (e.g. Quéméré *et al.* 2013). After this a summary table of remaining sequence reads in each sample is produced (often with similar sequences being clustered) and sequences are assigned taxonomy. Then, when converting these read counts to occurrence data, a threshold number of reads is often required for each taxon to be tallied as an occurrence. Sequencing depth can vary considerably between samples, so in practice a threshold percentage of reads is often used (e.g. 1% of food sequences McInnes *et al.* 2017b), or sequencing depth can be rarefied to a common level (O'Rorke *et al.* 2016). These approaches normalize detection across samples, so that more sequences are required for an occurrence to be recorded in samples with higher read depths.

Once occurrences are recorded in individual samples, several metrics can be used to summarise the diet across samples. Those considered here are percent frequency of occurrence (%FOO), percent of occurrence (POO) and weighted percent of occurrence (wPOO) (Figure 1; see Box 1 for details).

Box 1: Some metrics used to summarise sequence data in dietary metabarcoding studies

Occurrence Data

Frequency of occurrence (FOO) is the number of samples that contain a given food item, most often expressed as a percent ($\%FOO$). Percent of occurrence (POO) is simply $\%FOO$ rescaled so that the sum across all food items is 100%. Weighted percent of occurrence ($wPOO$) is similar to POO, but rather than giving equal weight to all occurrences, this metric weights each occurrence according to the number of food items in the sample (e.g., if a sample contains 5 food items, each will be given weight 1/5). Intuitive graphical representations are shown in Figure 1, mathematical expressions are as follows:

$$\%FOO_i = \frac{1}{S}\sum_{k=1}^{S} I_{i,k} \times 100\%$$

$$POO_i = \frac{\sum_{k=1}^{S} I_{i,k}}{\sum_{i=1}^{T}\sum_{k=1}^{S} I_{i,k}}$$

$$wPOO_i = \frac{1}{S}\sum_{k=1}^{S} \frac{I_{i,k}}{\sum_{i=1}^{T} I_{i,k}}$$

where T is the number of food items (taxa), S is the number of samples, and I is an indicator function such that $I_{i,k}$ = 1 if food item i is present in sample k, and 0 if not.

Many metabarcoding diet studies make use of both %FOO and POO (e.g. Xiong *et al.* 2017). POO provides a convenient view since each food taxon contributes a percentage of total diet (unlike %FOO which does not sum to 100%). In POO summaries samples with a high number of food taxa have a stronger influence, whereas in wPOO each sample is weighted equally (i.e. lower weighting to food taxa in a mixed meal) and this may be more biologically realistic (wPOO is the same as split-sample frequency of occurence; see Tollit *et al.* 2017 and references within).

Read Abundance Data

Using the sequence counts, relative read abundance (RRA_i) for food item i is calculated as:

$$RRA_i = \frac{1}{S}\sum_{k=1}^{S}\frac{n_{i,k}}{\sum_{i=1}^{T} n_{i,k}} \times 100\%$$

where $n_{i,k}$ is the number of sequences of food item i in sample k.

Some dietary metabarcoding studies present RRA data along with occurrence summaries, although relatively few have relied solely on information obtained from RRA (Table 1). In almost all of these studies, the number of sequences obtained per sample are converted to percentages (Figure 1a), because the absolute counts (i.e. sample sequencing depth) are dependent on several factors unrelated to the overall importance of the sample (amount of starting material used, DNA extraction efficiency, standardization of samples before HTS, etc.). To provide an overall diet summary, sample-specific RRA values can be averaged across samples; when doing so, each sample is given equal weight (Box 1; Figure 1). The RRA of taxa in each sample will vary depending on genetic marker, laboratory protocol, and bioinformatic filtering strategy (Alberdi *et al.* 2017; Deagle *et al.* 2013). Ensuring laboratory methods are robust (i.e. focussing on samples with sufficient target DNA and checking replicates) and using a standardised bioinformatics pipeline without excessive filtering can help ensure RRA data are reproducible and precise (Alberdi *et al.* 2017; Deagle *et al.* 2013; Emmanuel *et al.* 2017; McInnes *et al.* 2017a; Murray *et al.* 2015).

3. Does converting read counts to occurrence data solve our problems?

It is often assumed that because conversion to occurrence data moderates the impact of taxa-specific bias in marker signal, it provides a trustworthy, or at least conservative, view of diet. While it is true that occurrence-based summaries of diet are less affected by recovery bias, it is not necessarily the case that they provide a more accurate representation of overall diet. Our simulations suggest POO summaries are highly consistent but generally less accurate representation of overall diet compared to RRA summaries even when moderate taxa-specific recovery biases are present (see Box 2 for details).

Box 2: Simulations evaluating approaches for summarising population-level diet composition

To compare how effectively occurrence and RRA methods reconstruct population-level diet we simulated HTS read counts for samples derived from a population with a fixed diet and investigated the impact of taxa-specific sequence recovery biases (Figure 2). Basically, we show how population-level diet estimates vary given a range of biases that can impact any food taxa in the diet. Our simulation results are for a population with 40 food taxa in its diet, occurring in exponentially declining abundance. Sequencing was simulated for 100 scat samples assuming a mean of either 3 or 20 food taxa per sample, and assuming different sequence recovery bias scenarios: no bias, low bias or high bias. The biases introduce positive or negative biases of up to 4x and 20x (low and high biases respectively) relative to a standard. Diet summaries were made using: (1) RRA; (2) POO with a 1% minimum sequence threshold. For further details see Supporting Information and R scripts in the Dryad Digital Repository (doi:10.5061/dryad.jt07145).

Overall results show that with these parameters RRA summaries were on average more accurate but had higher variance than POO summaries. POO produced more consistent estimates less impacted by recovery biases, but only outperformed RRA when the largest recovery biases corresponded to the most common food items. Both methods were more accurate when the number of food taxa per sample was small: with a small number of food taxa per sample POO estimates provide more realistic enumeration of rare items and RRA estimates are less impacted by sequence recovery biases (since biases are only expressed in the context of other taxa in a sample).

-

The primary drawback of occurrence datasets is that the importance of rare food taxa are often artificially inflated at the expense of food taxa eaten in large amounts, effectively flattening the rank-abundance species curves typically seen in dietary datasets (Figure 1; Box 2). This effect can be illustrated in metabarcoding data from a population-level diet study of killer whales (Figure 3). This study concluded that the whale population's diet consisted primarily of Chinook salmon (~80%) based on high RRA of this species in most samples (Ford *et al.* 2016). If we consider the killer whales' diet as occurrence (POO; each food species occurrence given equal value), the view changes considerably because other salmon species and halibut frequently detected at low levels become important prey. The threshold level used to count an occurrence also impacts the relative importance of these fish prey; a lower threshold increases the importance of rare diet items (Figure 3). A similar pattern is seen in seal population-level diet estimates calculate with RRA and POO (Figure 4a). These different outcomes have substantial implications when diet percentages are combined with bioenergetics estimates and consumer population size to derive estimates of prey consumption (Chasco *et al.* 2017). Another implication of rare-item inflation occurs in studies of niche partitioning. Here, the conclusion that species feed on separate resources may be inaccurate because separation may be driven primarily by partitioning of rare diet items, which are given similar weight as shared important food. In contrast, the conclusion that species overlap in their dietary niche is potentially less likely (i.e. requiring overlap in both primary and rare food items), but may therefore be more biologically meaningful when found (Clare 2014).

How much influence rare diet taxa have in overall diet estimates depends to some extent on the foraging strategy of the focal species and food distribution. In cases where small amounts of rare diet items are consumed in most feeding bouts, the importance of these items could be strongly over-estimated in occurrence-based summaries (as seen in the simulations with a high number of taxa per scat sample; Box 2). This may be the situation for some large grazing herbivores that forage continuously across a grassland, often eating relatively rare plant taxa in proportion to their availability (i.e., non-selective feeding). In contrast, when rare diet items are eaten sporadically, their DNA would be detected only occasionally and diet estimates would be more realistic. For instance, some carnivores feed sporadically, individualistically, and in discrete foraging events such that prey occurrences may provide a more meaningful indication of how often each taxon is predated (Codron *et al.* 2016). The feeding ecology of a species is reflected to some extent in the number of food taxa in individual faecal samples and this varies widely between studies (Table 1). This value provides insight into the potential impact of rare-item inflation bias. For example, in Figure 1, the zebra faecal samples have many food taxa per sample and when summarised as occurrences, these have a predictably flat rank-abundance curve; this curve would be generated regardless of the true amount of each plant consumed in each meal (Box 2).

Summaries based on occurrences become less accurate when samples are pooled (i.e. when sequence reads from individual scats are not identifiable; Clare *et al.* 2014; Deagle *et al.* 2009; Ford *et al.* 2016) because rare diet taxa present in any one of the pooled samples are weighted equally to taxa found in all of the pooled samples. The time period over which food consumption is integrated in a faecal DNA sample (influenced by gut passage time) can affect these data in a similar way, since longer integration will mean rare taxa have a greater likelihood of being present in each sample.

The inflated importance of rare sequences in occurrence summaries could also magnify some problems encountered in diet metabarcoding. There are occasions when exogenous DNA can contaminate a sample of interest. This includes field-based contamination from non-food eDNA (McInnes *et al.* 2017a), laboratory contamination (De Barba *et al.* 2014), and misassignment of sequence-to-sample during HTS (i.e. tag-jumping; Schnell *et al.* 2015). These problems will generally have less influence in RRA summaries since the real food items should dominate unless samples are very poor quality. A similar issue is the detection of secondary predation (i.e. DNA from gut contents of ingested prey). Depending on the study system and research question, secondary predation may or may not be a serious problem. However, occurrence-based datasets are expected to over-emphasise these detections and ruling out secondary predation in occurrence summaries may require information of RRA, examination of prey co-occurrence, or expert knowledge (Bowser *et al.* 2013; Hardy *et al.* 2017; McInnes *et al.* 2017b).

4. Does RRA actually reflect food biomass?

The relationship between proportions of biological material in a sample and sequence reads recovered by HTS has been studied in many experiments by sequencing artificial mixtures with known composition. These 'mock communities' are most relevant to dietary metabarcoding studies when made from food tissues similar to what is being consumed. Both mitochondrial and chloroplast DNA markers are present in multiple copies in each cell and copy number varies between tissue types (e.g. leaves versus roots; Ma & Li 2015) and physiological state (e.g. juvenile vs. gravid adult; Veltri *et al.* 1990). Getting a thoroughly homogeneous mix of tissues in a small volume suitable for DNA extractions is challenging; therefore, mixtures made from DNA extracted separately for each taxa are sometimes used (e.g. Ford *et al.* 2016; Krehenwinkel *et al.* 2017b; Piñol *et al.* 2015). However, results from purified genomic DNA mixture may have little biological meaning because differences in cell density and genome size will confound results (i.e. low recovery from a species could be a bias, or the species may have a large genome and therefore fewer markers are in the fixed amount of DNA added to the mixture) (Piñol *et al.* 2015). Mixtures of PCR products can identify technical biases (i.e. good for assessing PCR primers), but miss underlying biological differences.

Conclusions from analyses of mock communities vary from no relationship to good correlations between the composition of the mixture and sequence reads (Edgar 2017; Kimmerling *et al.* 2018; Pornon *et al.* 2016). One reason for these different conclusions is that the range of concentrations analysed varies considerably across studies, from nearly equal mixtures of a few taxa, to mixtures containing many taxa in very different abundances. For example, consider two mixtures: (A) three species in the ratio 20:30:50 and

(B) eight species in the ratio 1:1:4:4:10:10:20:50. In the first mixture even modest relative deviations in recovery would result in a poor correlation, whereas the second would be less impacted by the same relative level of bias. High variability between studies is also due to biotic differences in target organisms and technical differences (e.g. different barcode markers, PCR primers, sequencing platforms, etc.). This variation makes it difficult to generalise, and considerable work is required to understand the reliability of RRA in any system. Two taxonomic prey groups that have been the focus of several dietary metabarcoding studies, and for which mock communities have been examined, are fish and insects. These groups provide some insight into the expected scale of biases.

In metabarcoding of fish mixtures, conserved PCR primers are generally employed and documented recovery biases are moderate. In their killer whale study, Ford *et al.* (2016) analysed known percentages of DNA extracted from four fish species and the RRA of each fish corresponded well to input (generally within 5% of expected values) providing confidence in their conclusions. Using prey species of harbour seals Thomas *et al.* (2016) carried out a detailed study on sequence recovery from blended tissue mixtures. Various taxa (primarily fish; n=18) were sequenced in 50:50 tissue mixes with a control fish, and the extent of deviations from the control fish measured. The recovered sequences varied from 20% to 60%, a 3-fold variation in marker recovery relative to the control. A recent study looking at recovery of barcode markers from bulk samples of larval fish avoided marker amplification by directly sequencing all DNA, then bioinformatically recovering relevant marker sequences (Kimmerling *et al.* 2018). They found strong correspondence between biomass in known mixtures and sequence counts, suggesting that without PCR amplification biases, biological differences in mtDNA density between these fish are small. Even studies looking at fish environmental DNA samples have found a relationship between fish density and recovered sequence counts (Lacoursière-Roussel *et al.* 2016; Port *et al.* 2015; Thomsen *et al.* 2016).

Many studies have sequenced DNA from insect mock communities; however, rather than considering if read counts are proxies for input biomass, the focus of these studies has generally been to test if taxa can be detected (Alberdi *et al.* 2017; Clarke *et al.* 2014; Elbrecht & Leese 2015; Yu *et al.* 2012). The reason for this focus is that insect communities tend to be complex, with many rare taxa, and the recovery biases large. In studies by Yu *et al.* (2012) and Clarke *et al.* (2014), a paltry 43-76% of species known to be present in mock communities were recovered. A study that included a mixture containing equal amounts of purified DNA from 12 arthropod species (10 insects, 2 spiders), reported RRA values for half of the species that were more than 100 times lower than expected (i.e. expected 8% and recovered at <0.08% (Piñol *et al.* 2015)). Another arthropod study found consistent relationships between percentages of DNA and RRA; however, the slope of the correlation deviated from the expected value of 1 in different insect orders and with different DNA markers, which was attributed to copy number variation (Krehenwinkel *et al.* 2017b). Even a change in PCR primers used to amplify a marker from the same gene can produce very different results (Alberdi *et al.* 2017). Most diet studies looking at insectivorous predators focus on occurrence data because of the generally poor correlation between biomass and read counts (Table 1), but methodological improvements may change this (Jusino *et al.* 2017).

Diet studies incorporate more complexity than analysis of mock communities due to potential differential digestion of food taxa. Relatively few captive feeding experiments have examined how well dietary DNA counts reflect known diet, but studies have been carried out on sheep (Willerslev *et al.* 2014), deer (Nakahara *et al.* 2015), penguins (Deagle *et al.* 2010) and seals (Thomas *et al.* 2014). These have focussed on simple diets (~2-6 diet items) and results generally show that comparisons between major and minor diet components are reflected in RRA. For example, the diet of sheep fed two plants in ratios of 0:100, 25:75, 50:50, 75:25, 100:0 had a good correlation with the percentages of DNA marker sequences amplified from rumen content (Willerslev *et al.* 2014). In a study on captive deer, >90% of the diet was made up of three plant species with two other species fed in low amounts. In this case >90% of sequences came from the three dominant taxa, but considering just these taxa, the correlation between what went in and what came out was poor (Nakahara *et al.* 2015). Similarly, in faecal samples from captive penguins fed pilchards as the majority of their diet, sequence reads from pilchards were most common in the data; however, the three other fish species fed in mass ratios 45:35:20 produced sequences counts of 60:6:34 (Deagle *et al.* 2010).

Detailed captive feeding studies examining quantitative prey DNA recovery have been carried out on captive seals and sea lions (Bowles *et al.* 2011; Deagle & Tollit 2007; Thomas *et al.* 2014). Early studies used quantitative PCR rather than HTS and found the amount of marker DNA recovered provided a reasonable indication of biomass ingested (Bowles *et al.* 2011; Deagle & Tollit 2007). A trial with harbour seals by Thomas *et al.* (2014) compared HTS data from food tissue (affected by biological and technical biases) with faecal DNA (affected by digestion as well). The scale of bias introduced by digestion was generally smaller than biases observed in undigested fish tissue mix. Since digestion bias may be in the same or opposite direction to tissue biases, the overall effect is expected to increase variance in prey-specific recovery biases compared to tissue mixes. These seal studies all excluded prey hard parts from DNA extractions, but in other systems where this may not be feasible, digestion biases could be larger. For example, faeces from insectivorous animals often contain relatively undigested hard body parts (i.e. exoskeleton). The impact on DNA recovery is difficult to assess: hard fragments will contain undigested DNA, but the DNA may not be extracted as efficiently as DNA present from soft bodied prey (Clare 2014).

Another approach to understanding how much of a signal is present in counts from DNA sequences is to compare results with other methods of diet analysis. In a study of large mammalian herbivores, Kartzinel *et al.* (2015) found a nearly one-to-one correlation between estimates of C_4 grass (family Poaceae) consumption based on stable isotopes analyses and RRA based on metabarcoding of the chloroplast marker (trnL-P6). The use of alternative proxies for diet composition can also reveal complexities. Craine *et al.* (2015) used similar protocols to Kartzinel *et al.* (2015) but found C_4 grass RRA to be under-represented compared to measures based on stable isotopes. They suggested that chloroplast density scales with foliar nitrogen concentrations so that RRA values could reflect dietary sources of protein, and thus may deviate from dietary sources of biomass as represented by carbon stable isotopes. If RRA values based on this marker occasionally reflect an animal’s source of protein more closely than its source of carbon (i.e., biomass), this knowledge can enable count data to still be interpreted appropriately.

Several studies have used traditional morphological analysis of food remains to help cross-validate RRA data (Soininen *et al.* 2009; Thomas *et al.* 2017). Thomas *et al*. (2017) analysed DNA and prey hard parts in a large studies of seal populations diet over several seasons, while there were minor differences between methods in prey recovery and taxonomic resolution, RRA and hard part occurrences provided a very similar picture (Figure 4b). Cross-validation has the problem that all methods of diet determination are biased, so if there is disagreement the correct answer may be unclear (Soininen *et al.* 2009). However, congruence between datasets is reassuring and known biases can be taken into account when making conclusions (e.g. jellyfish are digested quickly, so occurrence in faecal DNA but not stomach contents is credible; Jarman *et al.* 2013; McInnes *et al.* 2017b). Large differences in results between methods warrant further investigation; multiple lines of independent evidence provide the strongest support for any conclusion.

Overall, assessing recovery bias between food taxa is complex, specific to a study system, and can require significant effort. In some cases, broad correlations are likely, but this cannot be taken for granted and very large biases may occur (e.g. Pawluczyk *et al.* 2015).

5. A view of the way forward in interpreting sequence counts

What should be considered best practice given the potential biases we have outlined in diet metabarcoding studies? First of all, we should take a step back and remember that getting estimates of the true diet of any species using any method is a challenging proposition – all methods of diet analysis have biases. A well-designed metabarcoding diet study may provide as accurate an estimate as any other approach, and has the benefits of providing high taxonomic resolution, detecting rare foods and can potentially solve otherwise intractable problems (e.g. liquid feeding). We should also remember that other classic experimental design issues, such as collecting appropriate sample sizes and getting representative samples, will potentially have a bigger impact on study outcomes than the diet estimation method. Furthermore, dietary metabarcoding has a huge variety of applications, many of which do not require highly accurate dietary proportions.

Still, we will inevitably come to a point in dietary metabarcoding studies where we need to decide how to interpret sequence counts. It is often the case that the overarching views of population-level diet are consistent regardless of how sequence counts are summarised (i.e. when commonly occurring food items are also represented by the highest number of sequences). This is most likely to be the case when faecal samples contain a limited number of food taxa (in the extreme case where there is only one taxon per sample, occurrence and RRA estimates are identical and recovery biases have no impact). However, some outcomes will depend on how we consider counts. Occurrence summaries are less affected by differential recovery of markers from food taxa, but tend to put much more weight on food consumed in small quantities and potential contaminants. RRA can potentially provide a weighting of food present in a sample based on biomass, but differential recovery of markers (especially from dominant food taxa) can impact data summaries. Our strongest recommendation is that if one approach is relied on heavily, some justification should be given for its use, and potential biases inherent in the method should be acknowledged and taken into account when drawing conclusions.

5.1 Using occurrence data

Many future diet studies will have almost no information on the scale of biases in the recovery of sequences from specific food taxa. The use of occurrence data may be a sensible approach, but careful consideration of the impact of this choice is still required and the bioinformatics steps taken to produce this dataset should be documented. We recommend converting counts to percentages (excluding non-food sequences from total count) and then defining a minimum sequence percentage threshold to determine occurrences. This will limit the impact of variation in read depth. The threshold is a trade-off between maximizing inclusion of real diet sequences and excluding low-level background noise (secondary predation, contamination, sequencing errors). A 1% threshold may be suitable for many situations, but when diets are extremely diverse with potentially large recovery biases (e.g. some bats species), then a much lower threshold may be justified (e.g. 0.01% in Alberdi *et al.* 2017). In these cases, ensuring contaminant sequences do not influence results requires additional vigilance (De Barba *et al.* 2014; Nguyen *et al.* 2015). Given that many of the issues we have raised regarding the use of occurrence data stem from the disproportionate influence of rarer sequences, it may seem advantageous to use a higher minimum sequence threshold (e.g. >5% constitutes occurrence). While this type of summary can provide insight, rare taxa that make up a small percentage of sequences in many samples would be missed completely (Alberdi *et al.* 2017) and taxa-specific biases in recovery also have a larger impact on these high threshold occurrence summaries (see simulations in Figure S1 comparing different threshold levels). Since the purported benefit of occurrence-based approaches is to record food taxa even when there is strong bias against them, thresholds higher than 1% cannot be generally recommended.

The sequencing depth required per sample is directly related to the minimum threshold; in diverse and/or potentially highly biased situations warranting a very low threshold (e.g. 0.05%), high numbers of reads per sample would be needed (e.g. >10000). Lower read depth is sufficient with a 1% threshold and increasing replication (biological or technical) would be preferable to having redundant sequences within samples. Even when sequence counts are not used directly, it is important that these data are available (and ideally the sequence reads archived too) with appropriate explanatory files outlining potential biases. This allows others to revisit the data and will allow insight in future comparative meta-analyses.

Summaries of data based only on occurrence information will remain appropriate in many situations, simply because these are more predictably inaccurate and less impacted by recovery bias. This includes dietary metabarcoding studies that use DNA from food remains in gut contents since differences in time since ingestion will have a major impact on the relative number of reads recovered per taxon (Egeter *et al.* 2015; Greenstone *et al.* 2014). In studies using faecal samples, occurrence summaries will often be appropriate when food is clearly differentially digested, the sequence recovery bias is known to be high (e.g. many animals with an insectivorous diet), or this bias is unknown and results cannot be cross-validated. Note, that this appropriateness may differ between dietary analyses of relatively similar consumers. For example, most bat diet studies only analyse occurrence data, but the bat population shown in Figure 1 has relatively low diet richness compared to other bats and RRA may be suitable (Vesterinen *et al.* 2016).

5.2 Using RRA

Incorporation of RRA into analyses will have the most benefit when individual faecal samples contain several food taxa and the same food taxa occur across many samples. In these cases, occurrence summaries may provide very inaccurate summaries (Box 2). Unfortunately RRA-based summaries from these types of samples can be most affected by recovery biases (Box 2) and careful decisions about how to interpret data are required. When there is uncertainty surrounding which method will be more accurate, presentation of results summarised with both methods is recommended. Conclusions relying heavily on RRA should include justification as to why the counts are expected to contain a roughly accurate signature. One way to justify interpretations based on RRA is through cross-validation of the diet data with alternative methods, and this is recommended whenever possible (see Figure 4). Alternatively, mock community experiments and/or feeding trials can be carried out, but this is feasible in a limited number of situations. In study systems where diet is relatively well known, examining biases in a small number of dominant food taxa can ensure they are not drastically over or underestimated and will lend support to using RRA information. The dominant diet items have by far the strongest impact on RRA diet summaries as significant shifts in percentages of these species will adjust percentages of all food taxa (i.e. unit sum constrained data must sum to 100%). One question that inevitably arises is, at what point does “semi-quantitative” RRA information stop being useful? Our simulations indicate that even in scenarios with 20x overestimation of some food and 20x underestimation of others (i.e. in 50:50 mixtures this could lead to 400 fold recovery bias) the population-level RRA summaries often still provides a more accurate view of diet compared to POO (Figure 2). But the limits of usefulness will depend on the application. It is probable that comparisons between closely related food taxa will provide more reliable RRA data, because biological differences should be smaller and technical biases less pronounced (e.g. animal COI primer binding sites will be more conserved, or length differences in the plant trnL-P6 marker will be low). However, it is risky to make generalizations and to transfer specific methodological findings between study systems.

Further refinements to increase confidence in RRA dietary metabarcoding data are possible. Because conversion to occurrence datasets has been seen as a necessary remedy for biases in sequence recovery, there has been less incentive for researchers to test new protocols and evaluate markers on their ability to obtain accurate RRA data. While it is sensible to use standard DNA barcode markers, by ignoring information in RRA during marker development we might have inadvertently imposed limitations on the field. Fortunately, we are starting to move towards a point where markers used in different applications are better understood and alternative less-biased approaches are being explored. This includes the use of multiple target markers (Stat *et al.* 2017) and PCR-free approaches (Srivathsan *et al.* 2016) that can be combined with prey DNA enrichment (Krehenwinkel *et al.* 2017a). Inclusion of control materials in sequencing runs can also ensure consistency between experiments (Hardwick *et al.* 2017). For the most accurate diet estimates, correction factors can be developed to take into account known biological differences between taxa in mixtures (e.g. gene copy number differences; Angly *et al.* 2014; Vasselon *et al.* 2018). Such species-specific correction factors have been developed for fish, with the intent of applying them in field-collected seal diet samples (Thomas *et al.* 2016).

While the effort needed to justify the RRA approach may be challenging, the possibility of obtaining more accurate diet estimates will make it worthwhile in many situations. We have seen such effort undertaken in papers addressing broad ecological questions (Kartzinel *et al.* 2015; Willerslev *et al.* 2014), and in diet studies of marine predators, where population consumption has significant fisheries management implications (Ford *et al.* 2016; Thomas *et al.* 2017). This approach should also be possible in monitoring programs, such as those carried out on seabird diet (Jarman *et al.* 2013; Sydeman *et al.* 2017), where the long-term investment warrants the development of robust DNA-based methods that provide the best possible data.

5.3 Outstanding issues

There are a number of issues in the diet metabarcoding literature that have an impact on both occurrence and RRA summaries that have yet to be clearly addressed. Appropriate statistical analysis of metabarcoding data is one area that needs more development, in particular how to deal with unit sum constrained data that is biased (i.e. POO and RRA summaries add to 100% therefore any biases will impact the magnitudes of the other diet components). This becomes particularly confusing when comparisons are made between populations eating some different food taxa since relative comparisons are difficult, and biases may only impact one population (Aizpurua *et al.* 2018).

Another issue is the impact of collecting data with markers that have low taxonomic resolution (McInnes *et al.* 2017b) or collating data at higher taxonomic levels to increase certainty in taxonomic assignment (Biffi *et al.* 2017). Depending on how broad the grouping are, occurrence summaries may not be very informative as many occurrences are potentially pooled. For RRA it is unclear whether pooling counts from multiple taxa will nullify fine-scale stochasticity in recovery biases, or magnify lineage-specific biases. A related issue is how to summarise data from diet metabarcoding studies using multiple markers. When markers are targeting the same food taxa, either additive (i.e. include detections by any marker) or restrictive strategies (only include food detected by all markers) could be logically applied in occurrence and RRA summaries (Alberdi *et al.* 2017). The situation is even more complex when a "universal" primer set is used to define the broad diet and group-specific primers subsequently improve taxonomic resolution for particular groups (e.g. a marker targeting all plants together with several that offer greater resolution for specific plant families). Errors based on the universal marker will be propagated when attempting to incorporate data from the other primer sets (i.e. if the grass family is estimated to be 20% of a diet instead of the true 40%, then the perceived importance of each grass species is reduced).This problem can be avoided to some extent by reporting each component separately, but this provides an unsatisfactory synthesis for omnivorous and other species with a very diverse diet that can only be characterised with several markers (De Barba *et al.* 2014). Studies that use a marker capturing only one component of the diet need to be very clear that the results comprise an unknown amount of the total diet.

Simulations such as the ones outlined in this paper can help establish which scenarios are most sensitive to biases from alternate summaries (either occurrence or RRA). When informed by experimental work to assign an error range to each parameter, and combined with sensitivity analysis, this can identify which sources of bias have the largest impact on conclusions. There are many downstream application and we have not considered impacts in specific situations. For example, it would be very interesting to see how switching between occurrence and RRA datasets affects outputs in the context of quantitative food web studies (Banašek-Richter *et al.* 2009; Roslin & Majaneva 2016).

The ultimate test for how to deal with sequence counts in HTS diet analyses will remain in empirical studies. We hope this opinion piece will be a starting point to highlight the need to consider all sources of bias and to justify the methods used when confronting count data in metabarcoding studies. We also hope that this critique is not discouraging to researchers approaching this new and rapidly developing area of research, as no single study should be rightly expected to address all issues arising from DNA-based diet analyses. Instead, our aim is to encourage researchers to continue addressing methodological challenges, and acknowledge unanswered questions to help spur future investigations. As the field matures, we envisage publication standards will emerge to provide the most robust diet data and provide an accurate indication of the uncertainty associated with dietary assessments.

Acknowledgements

We thank our many colleagues for discussions on this topic. We also thank Vasco Elbrecht and two other reviewers for providing comments that helped improve the manuscript.

Data Accessibility

Data in figures have been deposited in the Dryad Digital Repository (doi:10.5061/dryad.jt07145) along with R scripts to produce the plots and R scripts used in our simulations.

Author Contributions

All Authors contributed ideas and to the writing of the paper

Figure Captions

Figure 1: Information in faecal samples from dietary metabarcoding datasets of an albatross (McInnes *et al.* 2017c), an insectivorous bat (Vesterinen *et al.* 2016) and Grevy's zebra (Kartzinel *et al.* 2015). (a) Individual-level data in 10 faecal samples viewed using different metrics. Colours represent different food taxa. (b) Population-level summaries of these datasets showing the top 15 food taxa (%FOO ranking); 1% threshold used for occurrence in POO and wPOO calculations. In the lower plots, sum contribution of remaining food taxa are plotted at end. In each example population data include only collections from one site and samples with >50 food taxa reads; the albatross data only considers the fish component of the diet (i.e. fish specific PCR primers).

Figure 2: Simulation results: (a) difference between estimated population diet and true diet proportions (compared using Bray-Curtis dissimilarity metric) for RRA and POO summary methods under different bias scenarios. The first plot shows an example bias vector (for both low and high bias) used in one simulation with differential recovery values for each food taxa. The boxplots summarise results from 1000 simulations for each bias scenario where the average number of taxa per sample was 3 or 20, with 100 samples per simulation. (b) In these simulations the most common taxa (T1) was forced to have the greatest positive bias or the greatest negative bias (low bias scenario = Low T1+ or Low T1-; high bias scenario = High T1+ and High T1-). Plots show the bias vectors and the corresponding population diet summaries are illustrated as bar plots. Numbers on top of bars are Bray-Curtis dissimilarity compared to true diet. Again, the average number of taxa per sample was 3 or 20. See Box 2 text for details.

Figure 3: Killer whale diet in the Salish Sea illustrated with bipartite graphs constructed from data in Ford *et al*. (2016) using either (a) RRA (b) POO with a 0.1% threshold or (c) POO with a 1% threshold. Samples (DNA from faecal material) are shown on left of each plot and were pooled according to collection dates (Early, Middle, Late) in different years. The overall diet calculated by the different methods is shown on the right of each plot (includes the seven prey taxa with >1% of sequences in at least one sample). Line thickness shows contribution of taxa in each sample to the overall diet.

Figure 4: Comparison between population-level diet percentages for harbour seals calculated with DNA metabarcoding and also parallel analysis of prey hard part remains (data from Thomas *et al*. 2017). Each point is a prey taxa, colours show different collection sites and symbols differentiate sampling times (e.g. black triangles represent all prey at the Fraser site in autumn 2012). Data are from over 1000 faecal samples and 14 comparisons are plotted (stratified by site, year and season) (a) Relationship between POO and RRA summaries based on a DNA metabarcoding dataset; POO puts more weight on food consumed in small quantities. (b) Relationship between wPOO and RRA; wPOO reduces the importance of samples with many prey (c) Relationship between RRA and prey remain (split sample frequency of occurrence model) summaries, a strong agreement despite biases in both approaches. To allow comparison of hard parts and DNA data higher taxonomic groupings of prey were used in many cases (e.g. salmon bones were rarely distinguished to species so DNA detections of salmon species were merged).

Table 1 Use of sequence counts in 20 metabarcoding diet studies carried out using faecal DNA collected from a range of different species. Representative studies across a range of focal taxa carried out by different research groups are shown rather than trying to summarise all dietary metabarcoding studies.

Focal Taxa	**Reference**	**FOO†**	**RRA‡**	**Sample number**	**Number food taxa §**	**Taxa per sample¶**	**Marker**	**Target group**	**Sequences per sample$^{\Phi}$**	**Sequencer**	**Count data Available**
Snail	O'Rorke *et al.* (2016)	N	Y	35	>50	NR	ITS	fungus	3500 (rarefied)	MiSeq	Yes
Snail	Waterhouse *et al.* (2014)	Y	N	60	26	4.7	16S	earthworms	1047	454	No
Pigeon	Ando *et al.* (2013)	Y	Y	48	44	6.7	trnL	plants	743	454	No
Albatross	McInnes *et al.* (2017b)	Y	Y	447	~20	NR	18S	metazoan	>100 prey	MiSeq	Yes
Puffin	Bowser *et al.* (2013)	Y	Y	129	~40	NA	CO1, 16S	metazoan	>50 prey	454	No
Sandpiper	Gerwing *et al.* (2016)	Y	N	164	132	NA	CO1, 16S	metazoan, fish/cephalopod/ crustacea	721^	454	No
Desman (Rodent)	Biffi *et al.* (2017)	Y	N	383	156	5.8	CO1	arthropods	6910^	Ion Torrent	No
Bat	Clare *et al.* (2014)	Y	N	25 (pooled)	>158	NA	CO1	arthropods	>10000*	Ion Torrent	No
Bats	Burgar *et al.* (2014)	Y	N	64	>120	15	CO1	arthropods	230	454	No
Bat	Vesterinen *et al.* (2016)	Y	Y	82	59	NR	CO1	arthropods	995	Ion Torrent	Yes
Bat	Aizpurua *et al.* (2018)	Y	Y	79	>276	8.4	CO1, 16S	arthropods	>10000*	MiSeq	Yes (raw sequences)
Seal	Thomas *et al.* (2017)	N	Y	1166	71	3.2	CO1, 16S	salmon, fish and cephalopods	1227	MiSeq	No (Available on request)
Seal	Hardy *et al.* (2017)	Y	N	112	115	3 to 6	16S, 12S	vertebrates, invertebrates	>10000*	MiSeq	Yes
Killer Whale	Ford *et al.* (2016)	N	Y	13 (pooled)	16	NA	16S	fish	>10000*	MiSeq	Yes (raw sequences)
Bear	De Barba *et al.* (2014)	Y	N	91	>84	NA	trnL, 12S, 16S, ITS	plants, vertebrates, invertebrates	>500	HiSeq	Yes
Cats	Xiong *et al.* (2017)	Y	N	103	40	3.6-4.1	16S	vertebrates	>10000*	HiSeq	No
Monke	Quémér	Y	N	96	>13	13.9	trnL	plants	>10000*	Illumi	No

y	é *et al.* (2013)				0					na GA IIx	
Deer	Erickson *et al.* (2017)	N	Y	12	>91	71	rbcL	plants	>10000*	MiSeq	No
Large herbivores	Kartzinel *et al.* (2015)	Y	Y	292	>110	NA	trnL, ITS	plants	>10000*	HiSeq	Yes
Ibex and Goat	Gebremedhin *et al.* (2016)	Y	Y	39	>50	NR	trnL	plants	>8000	454	Yes

† For this table Frequency Of Occurrence (FOO) refers to any use of presence/absence data

‡ For this table Relative Read Abundance (RRA) refers to the use of sequence counts to weight taxa present in samples. This includes distance methods such as Bray-Curtis dissimilarity applied to sequence counts.

§ Taxonomic level of assignments varies between studies, therefore the number of taxa is not directly comparable.

¶ In some cases multiple markers were used, or multiple samples were pooled, making this value Not Applicable (NA). NR indicate the number of food taxa per sample was Not Reported.

ϕ Most studies report mean number of food taxa sequences recovered per sample, but variance is not usually provided. The minimum number was reported in some cases.

^ Unclear if these sequence counts include non-target DNA such as consumer DNA.

* The maximum value reported here was 10000 reads per sample.

References

Aizpurua O, Budinski I, Georgiakakis P*, et al.* (2018) Agriculture shapes the trophic niche of a bat preying on multiple pest arthropods across Europe: evidence from DNA metabarcoding. *Molecular Ecology* **27**, 815–825.

Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2017) Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*.

Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology* **19**, 5555-5565.

Ando H, Setsuko S, Horikoshi K*, et al.* (2013) Diet analysis by next-generation sequencing indicates the frequent consumption of introduced plants by the critically endangered red-headed wood pigeon (*Columba janthina nitens*) in oceanic island habitats. *Ecology and Evolution* **3**, 4057-4069.

Angly FE, Dennis PG, Skarshewski A*, et al.* (2014) CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* **2**, 11.

Banašek-Richter C, Bersier L-F, Cattin M-F*, et al.* (2009) Complexity in quantitative food webs. *Ecology* **90**, 1470-1477.

Barrett RT, Camphuysen K, Anker-Nilssen T*, et al.* (2007) Diet studies of seabirds: a review and recommendations. *ICES Journal of Marine Science* **64**, 1675-1691.

Biffi M, Gillet F, Laffaille P*, et al.* (2017) Novel insights into the diet of the Pyrenean desman (*Galemys pyrenaicus*) using next-generation sequencing molecular analyses. *Journal of Mammalogy*.

Bowles E, Schulte PM, Tollit DJ, Deagle BE, Trites AW (2011) Proportion of prey consumed can be determined from faecal DNA using real-time PCR. *Molecular Ecology Resources* **11**, 530-540.

Bowser AK, Diamond AW, Addison JA (2013) From puffins to plankton: a DNA-based analysis of a seabird food chain in the northern Gulf of Maine. *PLoS One* **8**, e83152.

Burgar JM, Murray DC, Craig MD*, et al.* (2014) Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. *Molecular Ecology* **23**, 3605-3617.

Chasco BE, Kaplan IC, Thomas AC*, et al.* (2017) Competing tradeoffs between increasing marine mammal predation and fisheries harvest of Chinook salmon. *Scientific Reports* **7**, 15439.

Clare EL (2014) Molecular detection of trophic interactions: emerging trends, distinct advantages, significant considerations and conservation applications. *Evolutionary Applications*, 1144–1157.

Clare EL, Symondson WOC, Fenton MB (2014) An inordinate fondness for beetles? Variation in seasonal dietary preferences of night-roosting big brown bats (*Eptesicus fuscus*). *Molecular Ecology* **23**, 3633-3647.

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

Codron D, Codron J, Sponheimer M, Clauss M (2016) Within-population isotopic niche variability in savanna mammals: disparity between carnivores and herbivores. *Frontiers in Ecology and Evolution* **4**, 15.

Craine JM, Towne EG, Miller M, Fierer N (2015) Climatic warming and the future of bison as grazers. *Scientific Reports* **5**, 16738.

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

Deagle BE, Chiaradia A, McInnes J, Jarman SN (2010) Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics* **11**, 2039-2048.

Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology* **18**, 2022-2038.

Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN (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 BE, Tollit DJ (2007) Quantitative analysis of prey DNA in pinniped faeces: potential to estimate diet composition? *Conservation Genetics* **8**, 743-747.

Edgar RC (2017) UNBIAS: An attempt to correct abundance bias in 16S sequencing, with limited success. *bioRxiv*, 124149.

Egeter B, Bishop PJ, Robertson BC (2015) Detecting frogs as prey in the diets of introduced mammals: a comparison between morphological and DNA-based diet analyses. *Molecular Ecology Resources* **15**, 306-316.

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

Emmanuel C, Emese M, Gaït A*, et al.* (2017) A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies. *Molecular Ecology Resources* **17**, e146-e159.

Erickson DL, Reed E, Ramachandran P*, et al.* (2017) Reconstructing a herbivore's diet using a novel rbcL DNA mini-barcode for plants. *AoB PLANTS* **9**.

Finotello F, Di Camillo B (2015) Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in functional genomics* **14**, 130-142.

Ford MJ, Hempelmann J, Hanson MB*, et al.* (2016) Estimation of a killer whale (*Orcinus orca*) population's diet using sequencing analysis of DNA from feces. *PLoS One* **11**, e0144956.

Forney LJ, Zhou X, Brown CJ (2004) Molecular microbial ecology: land of the one-eyed king. *Current opinion in microbiology* **7**, 210-220.

Gebremedhin B, Flagstad Ø, Bekele A*, et al.* (2016) DNA Metabarcoding Reveals Diet Overlap between the Endangered Walia Ibex and Domestic Goats - Implications for Conservation. *PLoS One* **11**, e0159133.

Gerwing TG, Kim J-H, Hamilton DJ, Barbeau MA, Addison JA (2016) Diet reconstruction using next-generation sequencing increases the known ecosystem usage by a shorebird. *The Auk* **133**, 168-177.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333.

Greenstone MH, Payton ME, Weber DC, Simmons AM (2014) The detectability half-life in arthropod predator–prey research: what it is, why we need it, how to measure it, and how to use it. *Molecular Ecology* **23**, 3799–3813.

Hardwick S, Deveson I, Mercer T (2017) Reference standards for next-generation sequencing. *Nature Reviews Genetics* **18**, 473-484.

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

Ibarbalz FM, Pérez MV, Figuerola EL, Erijman L (2014) The bias associated with amplicon sequencing does not affect the quantitative assessment of bacterial community dynamics. *PLoS One* **9**, e99722.

Jarman SN, McInnes JC, Faux C*, et al.* (2013) Adélie penguin population diet monitoring by analysis of food DNA in scats. *PLoS One* **8**, e82227.

Jusino MA, Banik MT, Palmer JM*, et al.* (2017) An improved method for utilizing high-throughput amplicon sequencing to determine the diets of insectivorous animals. *PeerJ PrePrints*.

Kartzinel TR, Chen PA, Coverdale TC*, et al.* (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences* **112**, 8019-8024.

Kimmerling N, Zuqert O, Amitai G*, et al.* (2018) Quantitative species-level ecology of reef fish larvae via metabarcoding. *Nature ecology & evolution* **2**, 306.

King RA, Read DS, Traugott M, Symondson WOC (2008) Molecular analysis of predation: a review of best practice for DNA-based approaches. *Molecular Ecology* **17**, 947-963.

Krehenwinkel H, Kennedy S, Pekár S, Gillespie RG (2017a) A cost-efficient and simple protocol to enrich prey DNA from extractions of predatory arthropods for large-scale gut content analysis by Illumina sequencing. *Methods in Ecology and Evolution* **8**, 126-134.

Krehenwinkel H, Wolf M, Lim JY*, et al.* (2017b) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports* **7**, 17668.

Laake J, Browne P, DeLong R, Huber H (2002) Pinniped diet composition: a comparison of estimation models. *Fishery Bulletin* **100**, 434-447.

Lacoursière-Roussel A, Côté G, Leclerc V, Bernatchez L (2016) Quantifying relative fish abundance with eDNA: a promising tool for fisheries management. *Journal of Applied Ecology* **53**, 1148-1157.

Ma J, Li X-Q (2015) Organellar genome copy number variation and integrity during moderate maturation of roots and leaves of maize seedlings. *Current genetics* **61**, 591-600.

McInnes JC, Alderman R, Deagle BE*, et al.* (2017a) Optimised scat collection protocols for dietary DNA metabarcoding in vertebrates. *Methods in Ecology and Evolution* **8**, 192-202.

McInnes JC, Alderman R, Lea M-A*, et al.* (2017b) High occurrence of jellyfish predation by black-browed and Campbell albatross identified by DNA metabarcoding. *Molecular Ecology* **26**, 4831-4845.

McInnes JC, Jarman SN, Lea M-A*, et al.* (2017c) DNA metabarcoding as a marine conservation and management tool: a circumpolar examination of fishery discards in the diet of threatened albatrosses. *Frontiers in Marine Science* **4**, 277.

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

Nakahara F, Ando H, Ito H*, et al.* (2015) The applicability of DNA barcoding for dietary analysis of sika deer. *DNA Barcodes* **3**, 200-206.

Nguyen NH, Smith D, Peay K, Kennedy P (2015) Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist* **205**, 1389-1393.

Nielsen JM, Clare EL, Hayden B, Brett MT, Kratina P (2017) Diet tracing in ecology: method comparison and selection. *Methods in Ecology and Evolution*.

O'Rorke R, Holland BS, Cobian GM, Gaughen K, Amend AS (2016) Dietary preferences of Hawaiian tree snails to inform culture for conservation. *Biological Conservation* **198**, 177-182.

Olova N, Krueger F, Andrews S*, et al.* (2017) Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *bioRxiv*, 165449.

Pawluczyk M, Weiss J, Links MG*, et al.* (2015) Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and bioanalytical chemistry* **407**, 1841-1848.

Piñol J, Mir G, Gomez-Polo P, Agustí N (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources* **15**, 819-830.

Pompanon F, Deagle BE, Symondson WO*, et al.* (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology* **21**, 1931-1950.

Pornon A, Escaravage N, Burrus M*, et al.* (2016) Using metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific Reports* **6**, 27282.

Port J, O'Donnell J, Lowell N, Romero-Maraccini O, Kelly R (2015) Assessing the vertebrate community of a kelp forest ecosystem using environmental DNA. *Molecular Ecology*.

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

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

Schield DR, Walsh MR, Card DC*, et al.* (2016) EpiRADseq: scalable analysis of genomewide patterns of methylation using next-generation sequencing. *Methods in Ecology and Evolution* **7**, 60-69.

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

Soininen EM, Valentini A, Coissac E*, et al.* (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology* **6**, 16.

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

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

Sydeman WJ, Piatt JF, Thompson SA*, et al.* (2017) Puffins reveal contrasting relationships between forage fish and ocean climate in the North Pacific. *Fisheries Oceanography* **26**, 379-395.

Symondson WO, Harwood JD (2014) Special issue on molecular detection of trophic interactions: Unpicking the tangled bank. *Molecular Ecology* **23**, 3601-3604.

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

Thomas AC, Deagle BE, Eveson JP, Harsch CH, Trites AW (2016) Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources* **16**, 714-726.

Thomas AC, Jarman SN, Haman KH, Trites AW, Deagle BE (2014) Improving accuracy of DNA diet estimates using food tissue control materials and an evaluation of proxies for digestion bias. *Molecular Ecology* **23**, 3706-3718.

Thomas AC, Nelson BW, Lance MM, Deagle BE, Trites AW (2017) Harbour seals target juvenile salmon of conservation concern. *Canadian Journal of Fisheries and Aquatic Sciences* **74**, 907-921.

Thomsen PF, Møller PR, Sigsgaard EE*, et al.* (2016) Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS One* **11**, e0165252.

Tollit D, Fritz L, Joy R*, et al.* (2017) Diet of endangered Steller sea lions in the Aleutian Islands: New insights from DNA detections and bio-energetic reconstructions. *Canadian Journal of Zoology*.

Valentini A, Miquel C, Nawaz MA*, et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources* **9**, 51-60.

Vandeputte D, Kathagen G, D'hoe K*, et al.* (2017) Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**.

Vasselon V, Bouchez A, Rimet F*, et al.* (2018) Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods in Ecology and Evolution*.

Veltri KL, Espiritu M, Singh G (1990) Distinct genomic copy number in mitochondria of different mammalian organs. *Journal of cellular physiology* **143**, 160-164.

Vesterinen EJ, Ruokolainen L, Wahlberg N*, et al.* (2016) What you need is what you eat? Prey selection by the bat *Myotis daubentonii*. *Molecular Ecology* **25**, 1581-1594.

Waterhouse BR, Boyer S, Wratten SD (2014) Pyrosequencing of prey DNA in faeces of carnivorous land snails to facilitate ecological restoration and relocation programmes. *Oecologia* **175**, 737-746.

Willerslev E, Davison J, Moora M*, et al.* (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* **506**, 47.

Xiong M, Wang D, Bu H*, et al.* (2017) Molecular dietary analysis of two sympatric felids in the Mountains of Southwest China biodiversity hotspot and conservation implications. *Scientific Reports* **7**.

Yu DW, Ji Y, Emerson BC*, et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution* **3**, 613-623.

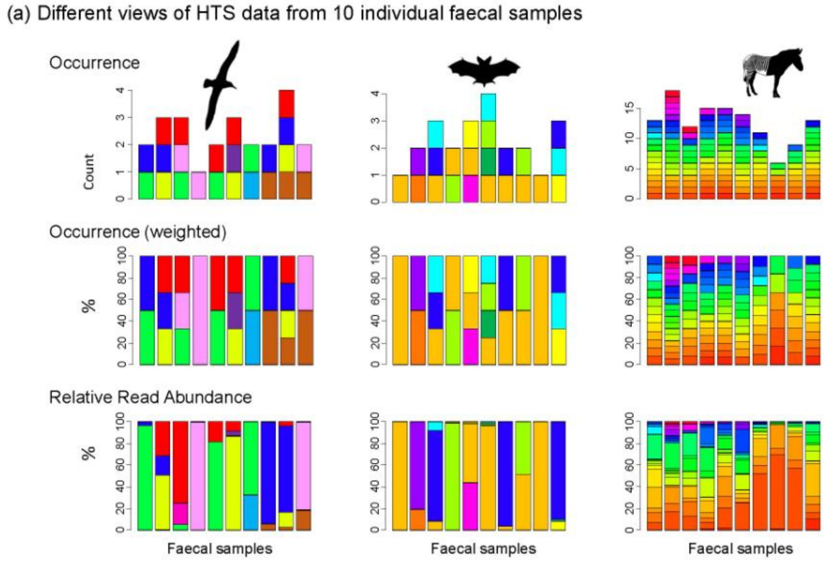

Figure 1a

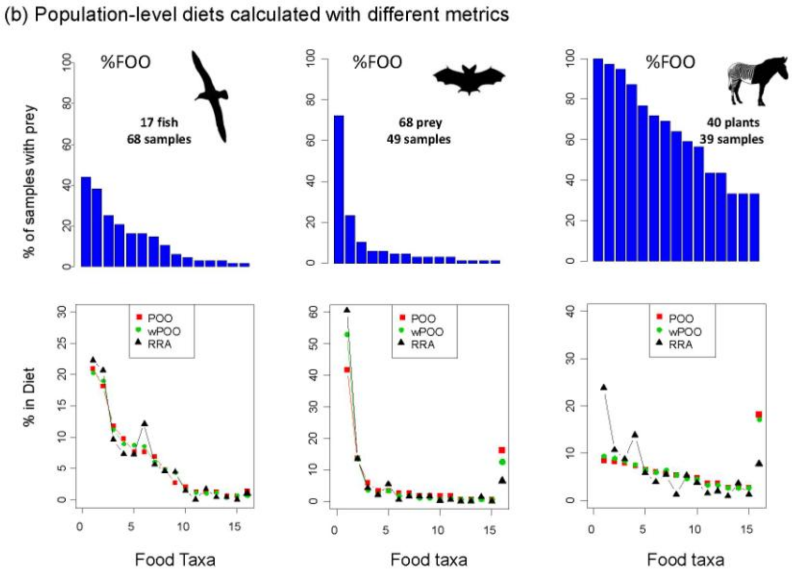

Figure 1b

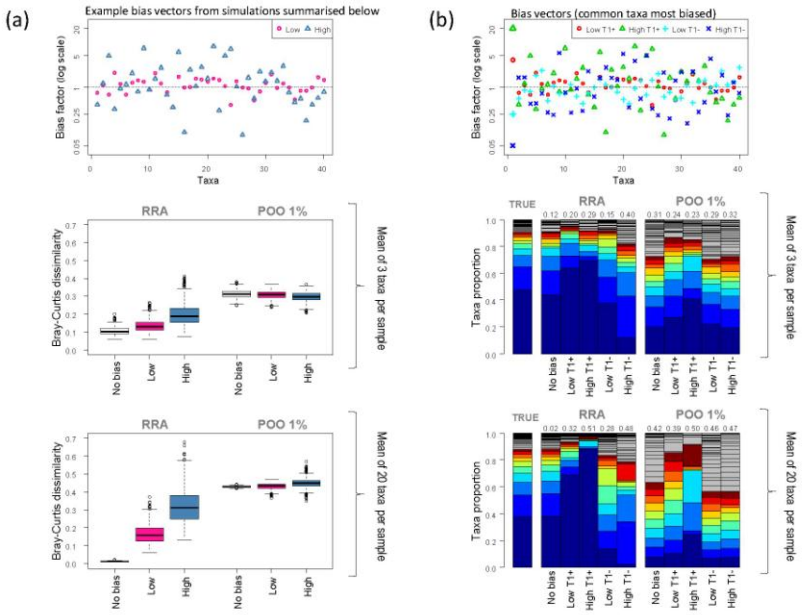

Figure 2

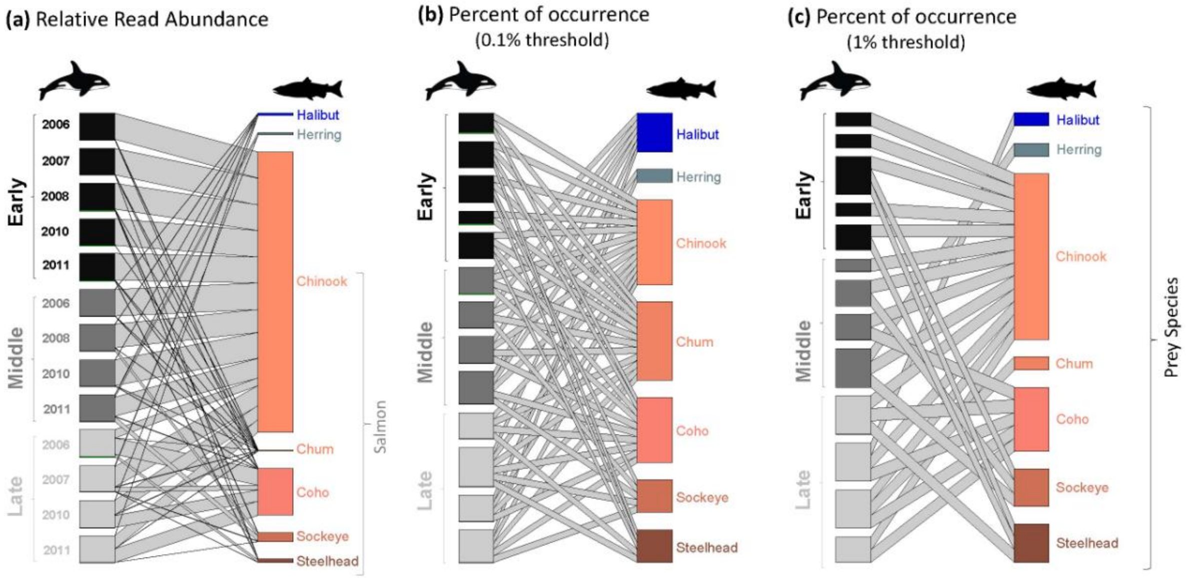

Figure 3

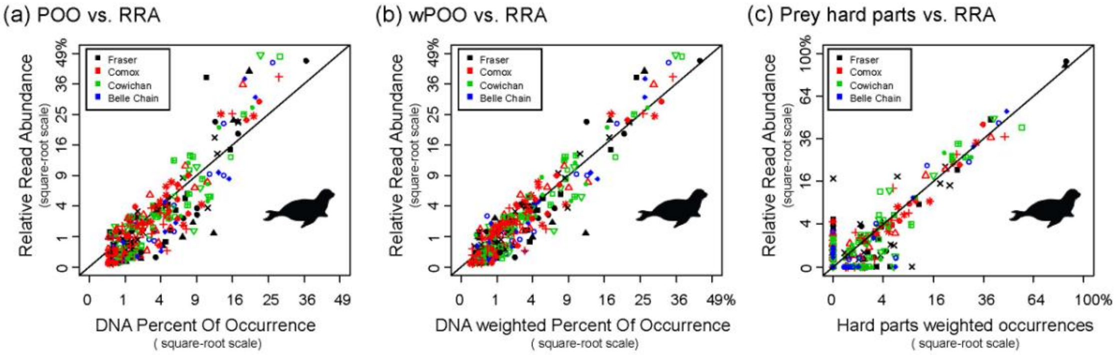
(a) POO vs. RRA
Relative Read Abundance
(square-root scale)
0
1
4
9
16
25
36
49%
Fraser
Comox
Cowichan
Belle Chain
0
1
4
9
16
25
36
49
DNA Percent Of Occurrence
(square-root scale)
(b) wPOO vs. RRA
Relative Read Abundance
(square-root scale)
0
1
4
9
16
25
36
49%
Fraser
Comox
Cowichan
Belle Chain
0
1
4
9
16
25
36
49%
DNA weighted Percent Of Occurrence
(square-root scale)
(c) Prey hard parts vs. RRA
Relative Read Abundance
(square-root scale)
0
4
16
36
64
100%
Fraser
Comox
Cowichan
Belle Chain
0
4
16
36
64
100%
Hard parts weighted occurrences
(square-root scale)

Figure 4