MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2016) **16**, 714–726 doi: 10.1111/1755-0998.12490

Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material

AUSTEN C. THOMAS,*,† BRUCE E. DEAGLE,‡ J. PAIGE EVESON,§ CORIE H. HARSCH* and ANDREW W. TRITES*

**Marine Mammal Research Unit, University of British Columbia Vancouver, British Columbia, Canada, †Science Department, Smith-Root Inc., Vancouver, Washington, USA, ‡Australian Antarctic Division, Channel Highway, Kingston, Tasmania, Australia, §CSIRO Oceans and Atmosphere, GPO Box 1538, Hobart, Tasmania 7001, Australia*

Abstract

DNA metabarcoding is a powerful new tool allowing characterization of species assemblages using high-throughput amplicon sequencing. The utility of DNA metabarcoding for quantifying relative species abundances is currently limited by both biological and technical biases which influence sequence read counts. We tested the idea of sequencing 50/50 mixtures of target species and a control species in order to generate relative correction factors (RCFs) that account for multiple sources of bias and are applicable to field studies. RCFs will be most effective if they are not affected by input mass ratio or co-occurring species. In a model experiment involving three target fish species and a fixed control, we found RCFs did vary with input ratio but in a consistent fashion, and that 50/50 RCFs applied to DNA sequence counts from various mixtures of the target species still greatly improved relative abundance estimates (e.g. average per species error of 19 ± 8% for uncorrected vs. 3 ± 1% for corrected estimates). To demonstrate the use of correction factors in a field setting, we calculated 50/50 RCFs for 18 harbour seal (*Phoca vitulina*) prey species (RCFs ranging from 0.68 to 3.68). Applying these corrections to field-collected seal scats affected species percentages from individual samples (Δ 6.7 ± 6.6%) more than population-level species estimates (Δ 1.7 ± 1.2%). Our results indicate that the 50/50 RCF approach is an effective tool for evaluating and correcting biases in DNA metabarcoding studies. The decision to apply correction factors will be influenced by the feasibility of creating tissue mixtures for the target species, and the level of accuracy needed to meet research objectives.

Keywords: diet analysis, DNA barcoding, environmental DNA, predator prey interactions

Received 23 July 2015; revision received 12 November 2015; accepted 17 November 2015

Introduction

High-throughput DNA sequencing is currently changing the way that biologists characterize assemblages of organisms, ranging from human intestinal microbes to whole eukaryotic communities (Eckburg *et al.* 2005; Bik *et al.* 2012; Taberlet *et al.* 2012a; Willerslev *et al.* 2014). Traditional methods for characterizing groups of organisms generally involved acquiring a representative sample of a community and then individually identifying each organism in the sample using a classification protocol such as a reference collection or taxonomic key. In the burgeoning field of DNA metabarcoding, genetic markers that can be recovered from broad groups of taxa are used to simultaneously characterize all species (or higher level taxonomic groups) contained in an environmental sample using high-throughput DNA amplicon sequencing (Taberlet *et al.* 2012b; Cristescu 2014). These new tools have allowed insight into systems that were largely unexplored due to methodological limitations, and have redefined the current level of understanding for several systems (e.g. Fonseca *et al.* 2010).

While DNA metabarcoding has many clear advantages, the process of characterizing groups of organism from amplified DNA sequences can be quite complex and requires careful study design and data analysis in order to avoid a biased interpretation (Creer *et al.* 2010; Pompanon *et al.* 2012). For example, chimeric sequences, contaminants and clustering algorithms can bias even the most basic outputs of DNA metabarcoding studies such as species richness (Coissac *et al.* 2012; Nguyen *et al.* 2014). Risk of biased interpretation is particularly apparent when researchers attempt to glean insight from

Correspondence: Austen C. Thomas, Fax: 360-573-2064; E-mail: athomas@smith-root.com

the proportions of species DNA sequences that result from amplicon sequencing (Zhou *et al.* 2011; Deagle *et al.* 2013). Differences in sequence read abundance between species are often used to infer the relative differences in mass or numerical abundance of species contained in a sample (Deagle *et al.* 2009; Soininen *et al.* 2009; Kowalczyk *et al.* 2011; Murray *et al.* 2011; Brown *et al.* 2012). For example, in a fascinating recent application of metabarcoding, DNA sequence reads were used to document changes in the proportional biomass of plant taxa over >50 thousand years based on eDNA in sediments and preserved megafauna diet samples (Willerslev *et al.* 2014). While such quantitative interpretation can vastly improve the value of DNA metabarcoding data to ecologists, numerous studies have documented biases that strongly impact sequence read abundance (Amend *et al.* 2010; Berry *et al.* 2011; Pinto & Raskin 2012; Deagle *et al.* 2013).

Previous attempts to control biasing factors in DNA metabarcoding studies have primarily focused on correcting for a single source of bias, or altering protocol steps that are known to introduce bias (Berry *et al.* 2011; Shokralla *et al.* 2012; Lundberg *et al.* 2013; Zarzoso-Lacoste *et al.* 2013). The objective of several recent bias correction efforts has been to account for species differences in template DNA copy number or DNA density (i.e. template copy number per gram of organism tissue) that cause certain species to be overrepresented and others underrepresented (Kembel *et al.* 2012; Angly *et al.* 2014). For example, Angly *et al.* (2014) documented variation in 16S rRNA gene copy number across microbial lineages, and used those data to correct amplicon counts in microbial community profiles. Copy number corrections and bias-mitigating alterations to laboratory protocols have proved useful for enhancing the quantitative capabilities of DNA metabarcoding; however, the presence of other technical factors often still prevents investigators from using DNA sequence proportions to infer relative organism mass or abundance.

An alternative approach to correcting for individual biases is to create control materials for target organisms, which when sequenced alongside environmental samples can be used to create correction factors that account for multiple sources of bias simultaneously (Huggett *et al.* 2013; Thomas *et al.* 2014). Using control materials, it is possible in a single correction step to account for biases due to copy number, DNA extraction, PCR amplification, DNA sequencing and bioinformatic filtering. However, the challenge in implementing control material correction factors comes in the transition from the laboratory to the field, where the goal is to characterize samples of unknown composition. For example, a recent metabarcoding diet study with seals demonstrated that by sequencing a fish tissue mixture that matched the diet of captive seals, food tissue correction factors could be calculated (Thomas *et al.* 2014). When these diet specific corrections were applied to prey DNA sequences from seal scats, the sequence percentages were much better aligned with seal diet composition. These results have limited applicability however, because they required a priori knowledge of the seal's exact diet in order to calculate correction factors.

A more generic approach was proposed which involves creating a prey library of tissue mix standards that could be used to correct sequence counts from samples of unknown composition. Such a prey library would consist of 50/50 mixtures of food tissues, wherein one species is held constant (i.e. present in all mixtures) and the other species is varied between mixes. Relative differences in the percentages of DNA sequences from mixtures would thus indicate the species-specific bias of the variable food species, and could be used to create relative correction factors (RCFs) useful for field studies. Such 50/50 RCFs would be most effective with samples of unknown composition if they proved to be consistent regardless of input proportion, and remained consistent regardless of species composition (i.e. no interactive effects between species).

Our objective was therefore to test the feasibility of using 50/50 RCFs derived from a prey library of tissue mixes to improve the relationship between mass percentages and DNA sequence percentages. Here, we create a model system using four fish species tissues, treating one species as the control and calculating RCFs from 50/50 mixtures of the control fish and the other three species. We then demonstrate how 50/50 RCFs can be used to correct sequence percentages from other mixtures of variable mass composition. We also generate a small prey library for Pacific harbour seals (*Phoca vitulina*) to evaluate the range of potential correction factors that would be produced using the 50/50 RCF method. Finally, we apply the prey library-derived correction factors to a subset of wild seal scat samples to determine the impact of 50/50 RCF correction in a real-world scenario. Although this study is focused on biases involved in seal diet analysis, the general framework for implementing 50/50 RCFs is widely applicable to any metabarcoding study that can feasibly create control mixtures of the target organisms (e.g. mixture of bacterial cultures, target insect species).

Materials and methods

Evaluation of tissue correction factors

Our first goal was to evaluate the feasibility of using 50/50 RCFs to improve the relationship between mass percentages and DNA sequence percentages. This involved testing whether the RCF for a given species remained consistent regardless of (i) input proportions (i.e. test whether the RCF calculated for species x using species y as a control remained the same regardless of the relative proportion of x to y by mass in the tissue mixture), and (ii) species composition (i.e. test whether the RCFs calculated using a given control species remained effective at correcting the sequence proportions in a sample mixture, regardless of the species composition of the mixture). If the RCFs are dependent on sample species composition, this would likely render any attempt at species correction factors unfeasible due to the sheer number of potential species combinations that could occur in a sample.

An experiment was set up involving four species: herring (*Clupea pallasii*), capelin (*Mallotus villosus*), atka (*Pleurogrammus monopterygius*) and mackerel (*Scomber japonicas*), where mackerel was used as the control. Pairwise tissue mixtures were created including one of the test species (herring, capelin or atka) and the control species (mackerel), where the mass percentage of the test species in each paired mixture progressively increased from 20% to 80% (e.g. the pairwise mass ratio combinations of herring and mackerel were 20:80, 40:60, 50:50, 60:40 and 80:20).

Tissue mixtures were created in three homogenization steps. First, representative samples of each fish species were chopped into pieces and individually ground using a standard meat grinder. Second, the coarse ground fish tissue was further homogenized with a bladed food processor. At this stage, a 4 g mixture of species tissue was created by combining the variable 'test fish' homogenate with the 'control fish' homogenate in a 20-mL vial. Lastly, 95% ethanol was added to the samples for preservation and they were processed with a Fisher Scientific PowerGen homogenizer, creating a finely ground ethanol/fish slurry. DNA was extracted, amplified and sequenced from the homogenized mixture, and the sequence proportions of the test and control species were calculated (see 'Genetic analysis' below).

Species-specific RCFs were calculated for each tissue mixture similarly to those in Thomas *et al.* (2014), but adapted for use with a control species:

$$\mathrm{RCF}_{p,\mathrm{t}} = \frac{M_\mathrm{t}}{M_\mathrm{c}} \times \frac{S_\mathrm{c}}{S_\mathrm{t}} \qquad (1)$$

where t is the test species, c is the control species, and M_t and M_c are the mass percentages (or grams) in the tissue mix of the test and control fish, respectively. S_t and S_c are the DNA sequence percentages (or counts) from the tissue extraction of the test and control fish, respectively, and p is the percentage of the test species in the mixture (i.e. $p = M_\mathrm{t}/(M_\mathrm{t} + M_\mathrm{c}) \times 100$). Using this equation, a correction factor can be calculated for any paired ratio of test fish and control fish after sequencing. RCFs >1 indicate that a species is underestimated relative to the control, and RCFs <1 indicate a species is overestimated. Note that $RCF_{50,\mathrm{t}}$ denotes what we have termed the 50/50 RCF for species t.

The overall process of estimating 50/50 RCFs for a set of test species and a given control species is illustrated in Fig. 1.

After calculating RCFs for all mixtures of the test and control species, we evaluated whether, for a given test species t, $RCF_{p,\mathrm{t}}$ was approximately the same for all values of p. In other words, are correction factors the same regardless of the input proportion used, or do they vary at values greater or lesser than 50%?

Next, to investigate whether the RCFs remained effective regardless of the species composition in the mixture being corrected, 50/50 RCFs were used to correct the DNA sequences resulting from the following tissue mix-

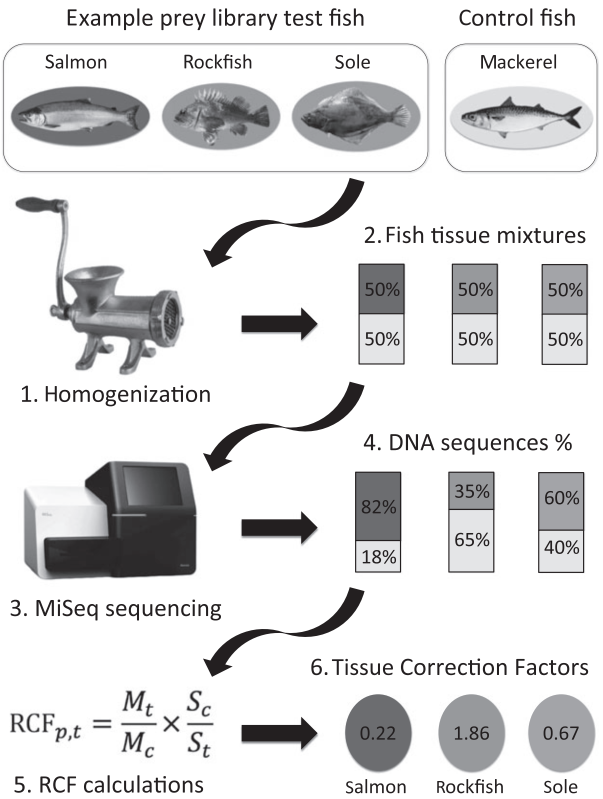

Fig. 1 Six steps involved in calculating relative correction factors (RCFs) from a prey tissue library: (1) homogenization of the control fish and test fish species, (2) creation of 50%/50% mixtures by mass of the control fish and various test fish homogenates, (3) Illumina amplicon sequencing of tissue mix DNA, (4) bioinformatic calculation of species DNA sequence proportions, (5) calculation of 50/50 RCFs and (6) numerical RCFs resulting from the prey tissue library. Colours indicate different fishes: salmon (red), rock fish (blue), sole (green), mackerel (yellow).

tures: (i) all pairwise mixtures of the three test species, where the mass percentage of one species progressively increased from 20% to 80%, similar to mixtures with the control (e.g. the mass ratio combinations of herring and capelin were 20:80, 40:60, 50:50, 60:40 and 80:20), and (ii) three-way mixtures of herring, capelin and atka in the ratios of 33:33:33 and 60:20:20. Two replicates of each mass ratio and species combination were made to evaluate technical variability.

To correct the sequence counts from a given sample using 50/50 RCFs, the count for each species is simply multiplied by the appropriate species-specific RCF:

$$\hat{N}_{\mathrm{t}} = N_{\mathrm{t}} \times RCF_{50,\mathrm{t}}$$

where N_{t} and $\hat{N}_{\mathrm{t}}$ are the observed and corrected sequence counts, respectively, from the sample for species t. The corrected sequence counts can then be expressed as percentages for comparison with the input mass percentages (i.e. $\hat{p}_{\mathrm{t}} = \hat{N}_{\mathrm{t}} \sum_{s \in S} \hat{N}_s$. where S denotes the set of all species in the sample).

Development of a harbour seal prey library

The next experiment consisted of calculating 50/50 RCFs for 18 harbour seal prey species in order to build up a library of correction factors. The prey library was not intended to create a complete set of RCFs for harbour seal prey. Rather, it was designed to assess the range of potential correction factor values across representative prey, and to see whether there are similarities in bias between closely related prey species.

Fresh whole samples of fish species that are known to occur in the diets of harbour seals in British Columbia, Canada, were collected opportunistically from one of two sources: (i) as bycatch in annual trawl surveys conducted by the Department of Fisheries and Oceans Canada, or (ii) purchased directly from fishermen shortly after landing at their port of call. To prevent water loss that could affect mass ratios, all samples were sealed in zip-type freezer bags and immediately frozen after collection in a nondefrosting freezer at −20 °C.

For each of the prey species in the sample collection (n = 18), tissue mixtures were made up comprising 50% of the prey species and 50% of the control species, where mackerel was again used as the control. The process described in the previous section and illustrated in Fig. 1 was used to calculate 50/50 RCFs for each tissue mixture. When possible, four replicate samples were made for each prey species in the library. Two replicates were made from homogenized tissue of multiple individual fish of the test species, and the other two contained tissues only from one individual fish each. The purpose of this design was to evaluate variability in the resulting sequence percentages that is due to (i) technical variation in sample processing, and (ii) biological variation between individual fish such as mtDNA density in tissue.

Wild harbour seal scat samples

The harbour seal scats we collected were part of a larger study directed towards assessing the impacts of harbour seals on salmon populations in British Columbia. At known harbour seal haulout sites, individual seal scats were collected into a 500-mL plastic jar lined with a 126-μm nylon mesh paint strainer. Samples were either preserved immediately in the field by adding 300 mL 95% ethanol to the collection jar, or they were taken to the laboratory and frozen at −20 °C within 6 h of collection. Samples were manually homogenized inside the paint strainer to separate the scat matrix material from hard prey remains (e.g. bones, cephalopod beaks), and the strainer was removed from the jar leaving behind the ethanol preserved scat matrix for genetic analysis.

The harbour seal prey library we generated did not contain all known diet species for harbour seals in British Columbia. Therefore, to assess the impacts of 50/50 RCFs on seal diet estimates, we selected a subset of 10 scat samples that contained only prey species that were included in our library, thereby allowing for 50/50 RCF correction of all species represented.

Genetic analysis

Tissue mixes and scat samples were subsampled, centrifuged and dried to remove ethanol prior to DNA extraction. Tissue extractions were carried out using the QIAGEN DNeasy Blood & Tissue Kit, and scat extractions were carried out with QIAGEN QIAamp DNA Stool Mini Kit according to the manufacturer's protocols. For additional details on the extraction process, see Deagle *et al.* (2005) and Thomas *et al.* (2014).

The metabarcoding marker we used to quantify fish proportions was a 16S mtDNA fragment (~260 bp) previously described in Deagle *et al.* (2009) for pinniped scat analysis. We used the combined Chord/Ceph primer sets: Chord_16S_F (GATCGAGAAGACCCTRTGGAGCT), Chord_16S_R (GGATTGCGCTGTTATCCCT), Ceph_16S_F (GACGAGAAGACCCTAWTGAGCT) and Ceph_16S_R (AAATTACGCTGTTATCCCT). This multiplex PCR is designed to amplify both chordate and cephalopod prey species DNA.

To take full advantage of sequencing throughput, we used a two stage labelling scheme to identify individual samples. This involved both uniquely tagged PCR primers (n = 96) and labelled MiSeq adapter sequences. The open-source software package EDITTAG was used to create 96 primer sets each with a unique 10-bp F primer tag and an edit distance of 5. This indicates that five insertions, substitutions or deletions would have to occur in order to cause one sample's sequences to be mistaken for another (Faircloth & Glenn 2012).

To ensure that all PCR conditions were identical to those used to amplify seal scat DNA in a related study, a blocking oligonucleotide was included in all PCRs to limit amplification of seal DNA (Vestheim & Jarman 2008). The oligonucleotide (32 bp: ATGGAGCTTTAAT-TAACTAACTCAACAGAGCA-C3) matches harbour seal sequence (GenBank Accession no. AM181032) and was modified with a C3 spacer, so it is nonextendable during PCR (Vestheim & Jarman 2008).

All PCR amplifications were performed in 20 μL volumes using the Multiplex PCR Kit (Qiagen). Reactions contained 10 μL (0.5×) master mix, 0.25 μM of each primer, 2.5 μM blocking oligonucleotide and 2 μL template DNA. Thermal cycling conditions were as follows: 95 °C for 15 min followed by 34 cycles of 94 °C for 30 s, 57 °C for 90 s and 72 °C for 60 s.

Sequencing libraries were prepared from pools of 96 samples using an Illumina TruSeq™ DNA sample prep kit which ligated uniquely labelled adapter sequences to each pool. Libraries were then pooled and DNA sequencing was carried out on Illumina MiSeq using the MiSeq Reagent Kit v2 (300 cycle) for SE 300-bp reads. Samples for this study were sequenced on multiple different runs as part of the larger study; however, typically between 4 and 6 libraries (each a pool of 96 individually identifiable samples) were sequenced on a single MiSeq run.

Sequences were automatically sorted (MiSeq post processing) by amplicon pool using the indexed TruSeq™ adapter sequences. FASTQ sequence files for each library were imported into QIIME for demultiplexing and sequence assignment to species (Caporaso *et al.* 2010). For a sequence to be assigned to sample, it had to match the full forward and reverse primer sequences, and match the 10-bp primer tag for that sample (allowing for up to two mismatches in either primers or tag sequence).

To assign DNA sequences to a fish species, we created a custom BLAST reference database of 16S sequences using an iterative process. First, using a list of the fish species of Puget Sound, we searched GenBank for the 16S sequence fragment of all fishes known to occur in the region (71 fish families 230 species) (DeVaney & Pietsch 2006; Benson *et al.* 2012). Reference sequences for each prey species were included in the database if the entire fragment was available, and preference was given to sequences of voucher specimens. GenBank contained 16S sequences for 192 of the 230 fish species in the region, and the remaining 38 species were mostly uncommon species unlikely to occur in seal diets.

Next, the DNA sequences that were assigned to scat or tissue samples were clustered with USEARCH (similarity threshold = 0.99; minimum cluster size = 3; de novo chimera detection), and a representative sequence from each cluster was entered in a GenBank nucleotide BLAST search (Altschul *et al.* 1990; Edgar 2010). If the top matching species for any cluster was not included in the existing database (or the sequence differed indicating allelic variation), the top matching entry was put in the reference database. The procedure was repeated with every new batch of sequence data to minimize the potential for incorrect species assignment or prey species exclusion.

For all DNA sequences successfully assigned to a sample, a BLAST search was carried out against our custom 16S reference database. A species was assigned to a sequence based on the best match in the database (threshold BLAST N e-value <1e^{-20}), and the proportions of each species' sequences were quantified by sample after excluding harbour seal sequences or any identified contaminants (Caporaso *et al.* 2010).

In all experiments, the errors reported are based on the absolute deviation from the uncorrected or previously corrected percentage value (i.e. $|$uncorrected % − corrected %$|$), after averaging replicate samples. When calculating the effects of correction factors for many samples combined (e.g. all pairwise species mixes), summary values reported are the average and SD of the calculated absolute deviations, combining all samples in that group.

Results

Evaluation of tissue correction factors

The experiment to evaluate the feasibility of using 50/50 RCFs with mackerel as the control species revealed several interesting trends (Fig. 2). First, there was a positive relationship for all test species between the mass percentage of the species in a tissue mix and the DNA sequence percentage of that species (Fig. 2a) However, the corresponding RCFs differed depending on the tissue mass percentage of the test species (Fig. 2b). In particular, when a species was present in high proportion (i.e. >50% by mass), it was generally underestimated by DNA sequence percentages relative to when it was present at 50% (i.e. the correction factor required was larger than the 50/50 RCF); and, conversely, when a species was present in low proportion (i.e. <50% by mass), it was generally overestimated by DNA sequence percentages relative to when it was

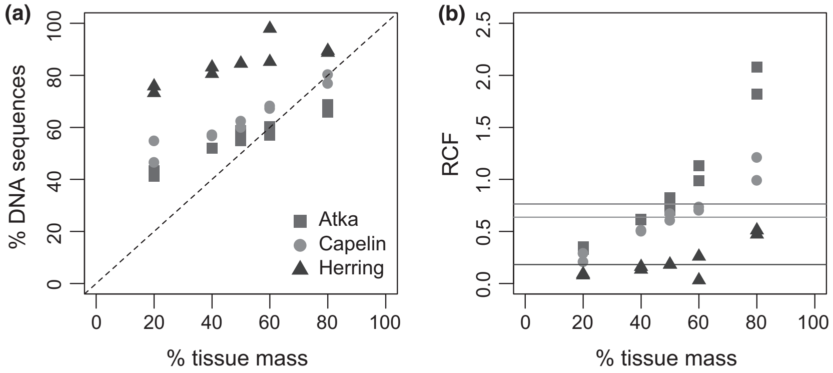

Fig. 2 Testing whether species correction factors remain consistent regardless of input mass percentage. (a) Percentage of DNA sequences recovered from tissue mixes of three test species (atka, capelin and herring) mixed individually with mackerel (the control species) in ratios of 20:80, 40:60, 50:50, 60:40 and 80:20 by mass. Two replicate tissue mixes were analysed for each test species and input ratio. (b) The relative correction factors (RCFs) calculated for each test species and input ratio. Horizontal lines indicate the 50/50 RCF value for each species. In both plots, the *x*-axis displays the percentage of the test species by mass in the tissue mix.

present at 50% (i.e. the correction factor required was smaller than the 50/50 RCF).

Although the RCFs for a given test species were proportion dependent, they were reasonably consistent for input percentages between 40% and 60% (Fig. 2). Moreover, in all mixes, the ranked species bias was consistent, that is herring was always the most overestimated, followed by capelin, then atka (the least abundant based on sequence percentages). These two factors suggest that using 50/50 RCFs to correct sequence proportions from unknown sample mixtures may still be reasonable.

Using mackerel as the control species, the 50/50 RCFs (mean and SD of the two estimates) for the three test species were herring (RCF = 0.18 ± 0.00), capelin (RCF = 0.64 ± 0.03) and atka (RCF = 0.76 ± 0.06). Applying these correction factors to DNA sequence counts from the pairwise tissue mixtures of these three test species reduced the average estimate error from 21 ± 15% (uncorrected) to 9 ± 6% (50/50 RCF corrected), averaging the errors of all pairwise combinations and replicates (Fig. 3). For the two tissue mixtures that combined all three test species, the RCFs improved estimates even more than in pairwise mixtures: average estimate error 19 ± 8% (uncorrected) and 3 ± 1% (50/50 RCF corrected) (Fig. 4).

We also explored the possibility of further correcting the estimates using proportion-dependent RCFs. To do so, we used the 50/50 RCF corrected sequence counts in place of the original sequence counts in eqn (1) to calculate new proportion-dependent RCFs (PRCFs) for each test species. We found that the relationship between the logarithm of the PRCFs and the input mass percentages for a given test species could be well approximated by a linear model (Fig. 5). Furthermore, the lines did not differ significantly between the three test species (*F*-statistic = 0.37, d.f. = 4, 24; *P*-value = 0.83), with the common line estimated to be:

$$\begin{aligned} \log_{10}(\mathrm{PRCF}_p) &= -0.59 + 0.012p \\ \Rightarrow \mathrm{PRCF}_p &= 10^{-0.59+0.012p} \end{aligned}$$

One replicate mixture of herring and mackerel resulted in a clear outlier relative to all other mixtures; this point was excluded from the consensus line calculation. We applied the appropriate PRCF as estimated from this linear equation to the 50/50 corrected sequence percentages. Specifically, if the 50/50 RCF corrected sequence count and percentage for test species *t* were $\hat{N}_\mathrm{t}$ and $\hat{p}_\mathrm{t}$, respectively, then we calculated the proportion-dependent corrected count ($\tilde{N}_\mathrm{t}$) as:

$$\begin{aligned} \tilde{N}_\mathrm{t} &= \hat{N}_\mathrm{t} \times \mathrm{PRCF}_{\hat{p}_\mathrm{t}} \\ &= \hat{N}_\mathrm{t} \times 10^{-0.59+0.012\hat{p}_\mathrm{t}} \end{aligned}$$

PRCFs mildly improved estimates for the pairwise test fish mixtures, but increased relative variability: average estimate error = 9 ± 6% (50/50 RCF corrected), changed to 5 ± 5% (proportion corrected) (Fig. 3). However, proportion-dependent correction substantially reduced the accuracy of estimates for mixtures that included all three species: average estimate error = 3 ± 1% (50/50 RCF corrected), changed to 8 ± 5% (proportion corrected) (Fig. 4).

Seal prey library

All fish species in the prey library tissue mix experiment were successfully identified in the bioinformatic sequence assignment pipeline. Within a prey species,

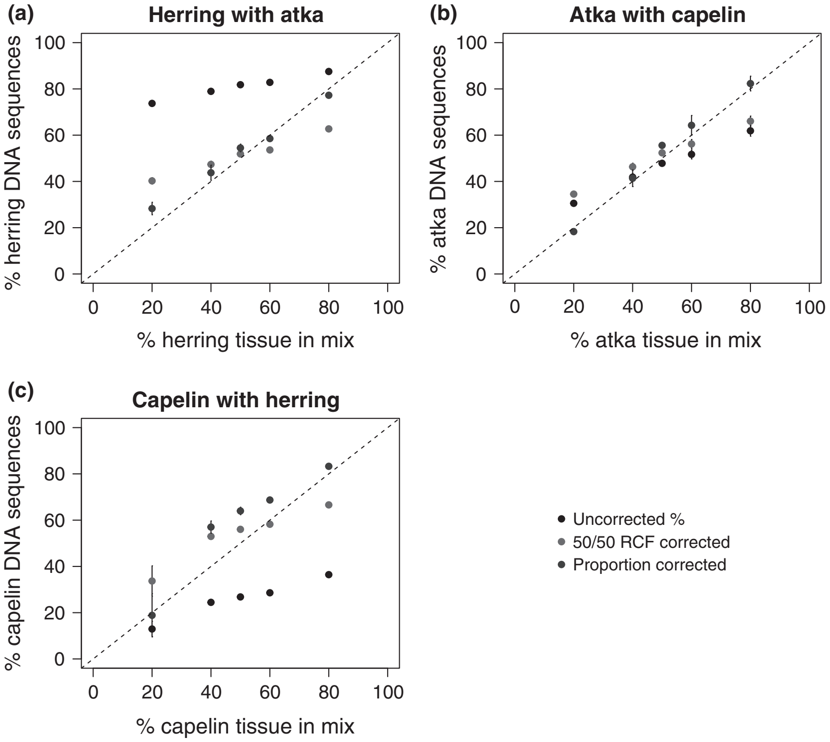

Fig. 3 Evaluating the effectiveness of correction factors based on mixtures with a control species (mackerel), by applying corrections to other pairwise species combinations. Displaying proportion of DNA recovered from tissue mixes which did not include the control species (a) herring mixed with atka, (b) atka mixed with capelin and (c) capelin mixed with herring in pairwise ratios of 20/80, 40/60, 50/50, 60/40 and 80/20. Black dots indicate the uncorrected sequence percentages; red dots indicate DNA percentages after the 50/50 relative correction factors (RCFs) from mackerel mixtures have been applied to both test species; and blue dots indicate percentages after both 50/50 RCFs and the proportion-dependent RCFs (Fig. 5) have been applied to both species.

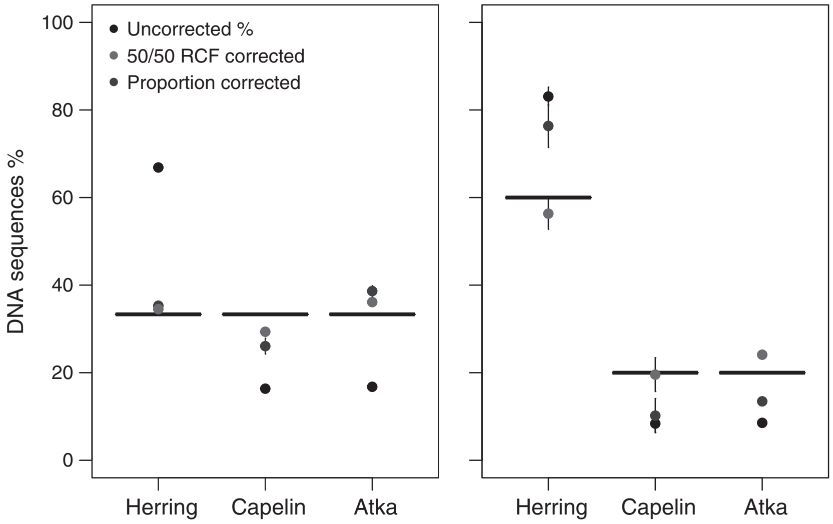

Fig. 4 Relative correction factors applied to DNA sequence percentages obtained from mixtures of three test species (herring, capelin and atka), as opposed to pairwise mixtures. Black bars indicate the species mass percentage in the tissue mixture sequenced: left (herring/capelin/atka: 33/33/33%), right (herring/capelin/atka: 60/20/20%). Black dots indicate average uncorrected DNA sequence percentage of two replicates, and error bars indicate standard deviation. Red dots show sequence percentages after 50/50 RCFs (from mixtures with the control species) have been applied to all three test fish species. Blue dots indicate average values after both 50/50 RCF correction, and the proportion-dependent correction factors have been applied.

there was generally little variability in the DNA sequence percentages between biological and technical replicate samples (Fig. 6). For example, the average deviation between two replicate samples containing multiple individuals was 2.6%, and the average deviation between samples of individual fishes of the same species was 3.9% (Fig. 6). By comparison, the average amount that a species' DNA sequences percentage deviated from

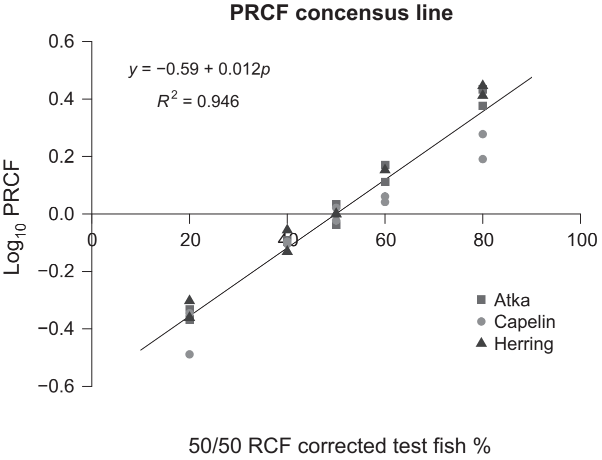

Fig. 5 Linear equation for the log-transformed proportion-dependent relative correction factors (PRCFs), plotted against the 50/50 RCF corrected sequence percentages of all pairwise mixtures combining test fishes (herring, capelin, atka) with the control fish (mackerel). This equation can be used to derive the proportion-dependent correction factor for any species, using the 50/50 RCF corrected sequence % for that species in a mixture. PRCFs can then be applied in a final correction step to account for proportion-dependent biases (see Results section for details).

the tissue mix mass percentage (i.e. 50%) was 9.5% across all species in the prey library (ranging from 0.1% for whitebait smelt to 28.6% for juvenile walleye pollock).

The 50/50 RCFs calculated for each species in the library using mackerel as the control ranged from 0.68 to 3.68. Grouping taxonomically, herrings, smelts, lingcod and dogfish generally required minor correction relative to mackerel: Pacific sardine (RCF = 0.87 ± 0.03), American shad (RCF = 0.98 ± 0.05), juvenile Pacific herring (RCF = 1.32 ± 0.11), northern anchovy (RCF = 1.13 ± 0.03), whitebait smelt (RCF = 1.00 ± 0.03), eulachon (RCF = 1.13 ± 0.04), lingcod (RCF = 1.17 ± 0.07) and spiny dogfish (RCF = 0.92 ± 0.01). Rockfishes, English sole and cods (including hake) were generally underestimated relative to mackerel, with juvenile pollock being the most underestimated species: copper rockfish (RCF = 2.08 ± 0.17), quillback rockfish (RCF = 1.90 ± 0.04), canary rockfish (RCF = 1.26 ± 0.07), English sole (RCF = 3.09 ± 0.11), Pacific hake (RCF = 1.56 ± 0.23) and juvenile walleye pollock (RCF = 3.68 ± 0.06). The salmonids were variable, with coho salmon being the most overestimated species relative to mackerel: chum salmon (RCF = 0.96 ± 0.05), coho salmon (RCF = 0.68) and pink salmon (RCF = 1.54 ± 0.05). Only one cephalopod species was tested, which was underestimated and exhibited relatively high variability between replicates: market squid (RCF = 2.90 ± 0.58).

Applying 50/50 RCFs to seal scats

50/50 RCFs derived from the prey library were applied to 10 wild harbour seal scat samples comprised of only those prey species represented in the prey library (Fig. 7). For individual samples, the average change in diet % per species was 6.7 ± 6.6% after applying 50/50 RCFs to all prey. The maximum amount that any prey species diet percentage changed was 23.9% for walleye pollock, which required significant positive correction (Fig. 7: sample 5). By contrast, population-level diet percentages calculated by averaging each species' DNA % across all samples were less affected by 50/50 RCFs, with the average change per species being 1.7 ± 1.2%, and a maximum change of 3.8% for walleye pollock (Fig. 7: population average).

Discussion

DNA metabarcoding is a powerful tool for the simultaneous characterization of multiple species in an environmental sample, with a seemingly endless range of potential applications. However, to fully take advantage of the data produced by next-generation sequencing platforms in metabarcoding studies, a practical method is needed to control the biasing factors that are known to affect DNA sequence read abundance. Our testing of species-specific correction factors from tissue mixtures of the target organisms (fish tissue homogenates) produced several results that will likely be of interest to researchers using sequence read abundance to quantify relative proportions of species. First, we found that increasing a species mass proportion results in a consistently greater proportion of DNA sequences, supporting the idea that sequence read abundance can be used as a measure of relative mass composition. There was, however, a strong proportion-dependent effect on sequence read abundance, such that when a species was present as a low mass proportion its relative counts tended to be inflated compared to those in the 50/50 mix, and the opposite when the species was at a high mass proportion. A similar finding was reported by Kembel *et al.* (2012) while applying gene copy number corrections to empirical environmental data sets. They noted that gene abundances (microbial 16S sequence reads) were generally higher for the rarest taxa, and lower for the most abundant taxa relative to estimated organism abundances. Our combined results suggest that the observed phenomenon may be inherent to the data produced by next-generation amplicon sequencing, and should be considered in future metabarcoding studies.

One potential explanation for the observed proportion-dependent bias is that template DNA available in high copy number during PCR may be more likely to

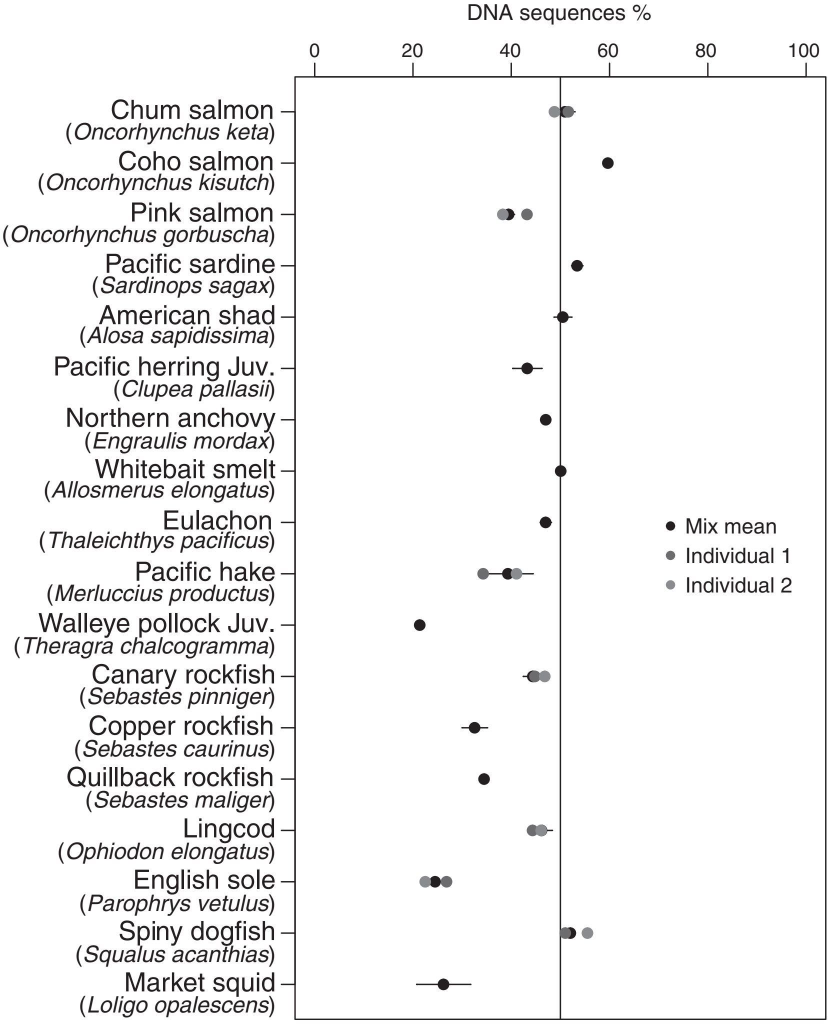

Fig. 6 Proportions of DNA sequences counted after Illumina amplicon sequencing of tissue samples that contained 50% of each test species by mass and 50% chub mackerel (the control species). Purple and blue dots indicate replicate samples of those species for which individual fishes were sequenced (indicating biological replicate variation). Black dots and error bars (SD) are samples from multiple combined individual fish of the test species (indicating technical replicate variation).

self-anneal rather than binding to PCR primers, which would partially inhibit amplification. By contrast, template DNA available in low concentration has a much lower probability of self-annealing because the single stranded fragments are more likely to encounter primers instead of the complimentary DNA strand. Thus, in our simple pairwise mixtures of two fish species DNA, it is conceivable that the PCR for the high abundance species is less efficient than the same reaction for the low abundance species. Our observation that this bias was less apparent in more complex mixtures (>2 species) is also consistent with this explanation, as we would expect the problem of self-annealing to be limited to instances where there is an overwhelming difference between species template DNA concentration during PCR.

A proportion-specific relationship between DNA sequence % and biomass % could pose a significant challenge to bias correction efforts; however, we did find that the magnitude of the bias was highly predictable and consistent between species mixtures. The linear equation from the log-transformed PRCFs enabled us to calculate the level of correction would be required for any resulting DNA sequence % after the initial 50/50 RCFs had been applied to sequence counts. The second-stage correction factor based on that relationship mildly improved estimates for the pairwise species mixes, but it increased variability and worsened proportional estimates for mixtures of more than two species. Given that most mixtures contain more than two species and proportional improvement with PRCFs was minimal, our findings suggest that proportion-based correction is likely not worth pursuing for application to field-collected samples.

Although correction factors were influenced by species input proportion, we did not find strong evidence of an interactive effect between co-occurring species in mixtures. For example, when both species in the pairwise mixtures were corrected with 50/50 RCFs (generated using a common control fish species), the resulting proportional estimates were highly consistent (Fig. 3: 50/50

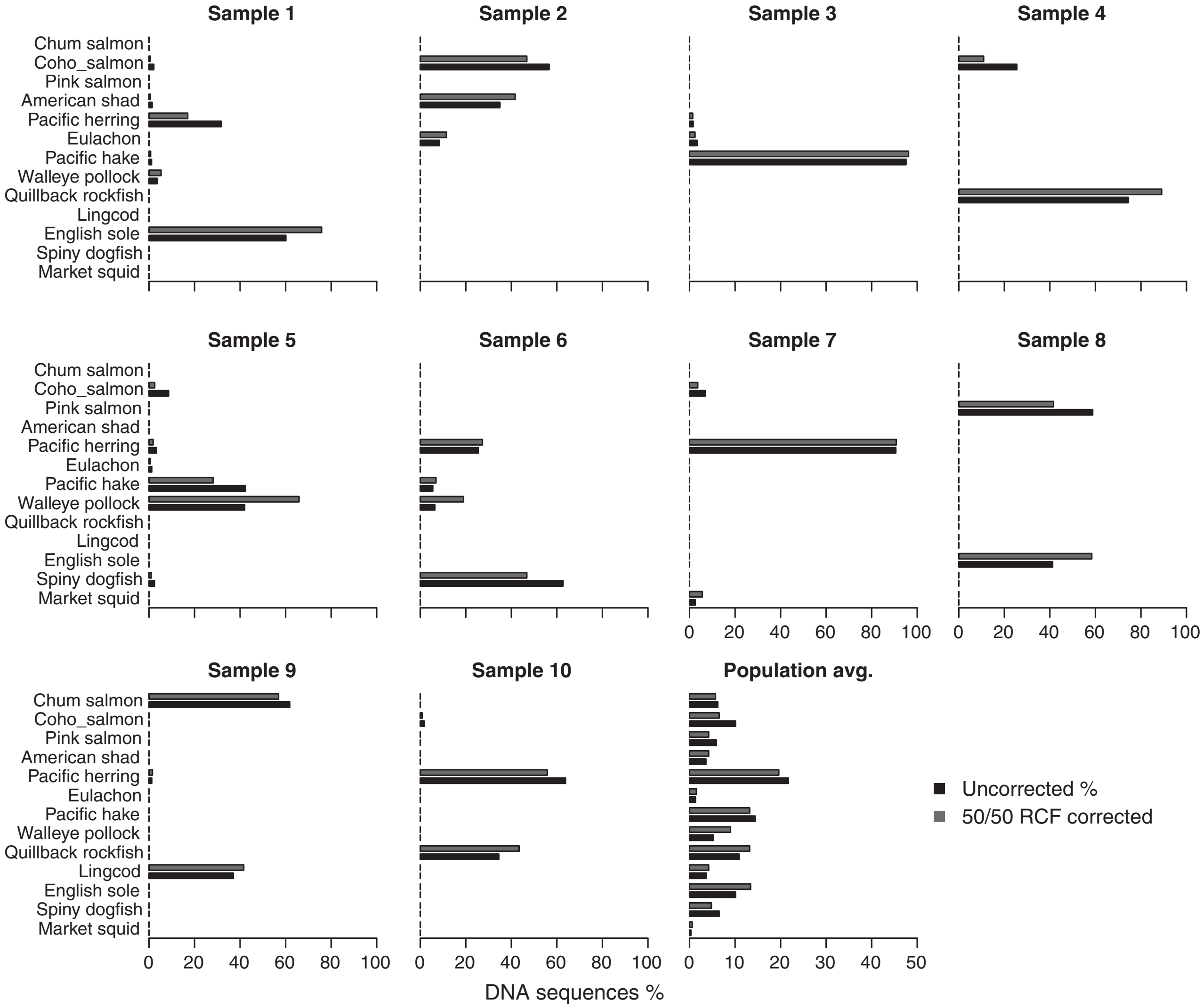

Fig. 7 Sequence percentages recovered from harbour seal scats collected in British Columbia, Canada. These samples only contained prey species included in our 50/50 tissue library. Black bars indicate the uncorrected percentages of species DNA sequences per sample. Red bars indicate sequence percentages after 50/50 RCFs from the prey library were applied to sequence counts for each sample. The bottom right panel displays the population average estimate for all ten samples combined.

RCF corrected). This implies that the magnitude of bias for a particular species does not change depending on the other species present in the mixture, and indicates that 50/50 RCFs created by combining test and control species are a viable means of correcting for the majority of species-specific bias.

In all instances, 50/50 RCFs improved the relationship between DNA sequences % and tissue mass % when applied to DNA sequence counts. After applying 50/50 corrections to DNA sequences of the pairwise mixtures, the average estimate error was reduced from 21% (uncorrected) to 9% (50/50 RCF corrected). Most of the remaining error after 50/50 RCF correction was due to deviation in the high and low mass proportions (Fig. 3). The effectiveness of the 50/50 RCFs was much more pronounced in the mixtures of all three test species, reducing the average estimate error from 19% to 3% (a sixfold reduction in average error) (Fig. 4). This consistent accuracy improvement from 50/50 RCFs, and the apparent lack of an interactive effect between species, suggests that 50/50 RCFs could be a useful approach for increasing the accuracy proportional biomass estimates in field-based DNA metabarcoding studies.

In order to apply 50/50 RCFs in a metabarcoding study with field-collected samples, a tissue library of potential target organisms would need to be generated such as the seal prey library created in our study. As anticipated, sequencing of the 50/50 prey library resulted in substantial variation in the percentages of sequences recovered between different fish species, indicating a range of species-specific biases (Fig. 6). The fact that there was very little variability between replicate samples suggests that the biases detected (i.e. deviation from 50%) are indicative of a true species-specific biases, and not due to individual variation or experimental error. Species of a common family tended to have similar correction factor values, supporting the notion that there is some phylogenetic structure to the biases detected (Angly *et al.* 2014).

We demonstrated how the 50/50 RCF approach can be used in a field study by applying our prey library-derived RCFs to sequence data from wild harbour seal scat samples. Our results indicated that the average magnitude of improvement from 50/50 RCFs for any individual species in a sample was approximately 7% per diet species—although this will depend largely on the number of species in the sample and the species proportional differences. For example, the magnitude of improvement from 50/50 RCF correction would be larger if a sample contained equal proportions of two species, compared to a sample containing equal proportions of three species. The degree of change to sample percentages can be substantial when co-occurring species present in large proportions require opposing correction factors.

The impact of 50/50 RCF correction was far less pronounced when samples were aggregated to create a population-level diet estimate (Fig. 7). The average change due to 50/50 RCF correction to any individual species in the population diet estimate was <2%, indicating that there is a strong bias-mitigating effect of averaging samples when generating population diet estimates. These results imply that the choice of whether or not to apply 50/50 RCFs in metabarcoding studies will likely be driven by the level at which proportion information is needed (i.e. individual samples vs. aggregate estimates), and the degree of accuracy required to effectively answer the research questions.

While 50/50 RCFs may provide a solution to multiple sources of bias in a single correction, there are other sources of bias that are not accounted for using this approach that require consideration. Most notably are biases introduced by differential degradation of species DNA due to either digestion (in the case of diet studies), or other degenerative processes responsible for degrading environmental DNA. A metabarcoding diet study with penguins suggested that differential DNA degradation due to digestion was the most significant cause of bias in the study system (Deagle *et al.* 2010). In those cases, additional bias correction efforts (e.g. lipid correction; Thomas *et al.* 2014) may be needed in order to achieve a highly accurate representation of mass proportion from DNA sequence counts of environmental samples.

When to use 50/50 RCFs

In many DNA metabarcoding studies, the primary challenge is simply to detect all species present in an environmental sample, such as when samples consist of many phylogenetically dissimilar taxa that require multiple degenerate primers to achieve amplification of most species. In those circumstances, it is likely unrealistic to expect accurate estimates of species proportion based on DNA sequence read abundances (e.g. Clarke *et al.* 2014; Elbrecht & Leese 2015), and correction factors are likely not worth pursuing in that stage of methodological development. However, in study systems focused on a limited number of species which have conserved barcode priming regions, 50/50 RCFs offer potential to improve proportional estimates by accounting for multiple sources of bias. The 50/50 RCF approach will be particularly useful when biases to sequence read abundance are substantial and the resulting species correction factor magnitudes are large. Even when it is not possible to generate a complete tissue library, a 50/50 RCF library consisting of a subset of key species could be used to screen for large species-specific biases and aid in the interpretation of sequencing results.

For metabarcoding diet studies, the goal is often to generate a population diet estimate from multiple individual diet samples, and the diet proportions of any individual sample are not especially important. Based on our results, the accuracy improvement to population diet estimates from 50/50 RCFs is subtle, and prey library-derived 50/50 RCFs may not be worth the effort unless high diet accuracy is needed. Small differences in population diet estimates can however lead to drastically different ecological conclusions. For example, over a 3-month period a difference of 2% Chinook salmon in the diets of 40 000 harbour seals in British Columbia could equate to a difference of ~15 million juvenile Chinook salmon being consumed by the seal population (Olesiuk 1993, 2010). This implies that accurate population-level diet information may be very important in this study system.

The benefits of 50/50 RCFs will be most apparent when it is important to provide accurate proportional information for a single environmental sample, or when multiple replicate samples from a single location are used to characterize species composition. Here, it is worthwhile to distinguish between aggregates of replicate samples such as those often employed in eDNA studies vs. population averages of many individual samples taken from separate animals or different sampling locations. Characterization of a single sampling site using DNA metabarcoding will be more vulnerable to bias because estimates are less affected by the bias-mitigating effects of averaging. Thus, 50/50 RCFs will likely prove beneficial for researchers using DNA sequence read abundances to characterize species composition from a single environmental sample or collection location.

It should be emphasized that the species-specific correction factors calculated using this approach are also specific to the experimental conditions of the methodological protocol. For example, one could not expect to produce the same RCF values we calculated if the blocking oligo was excluded from the PCRs (Piñol *et al.* 2014, 2015). Therefore, the control materials used to produce RCFs should be resequenced each time an alteration is made to the methodological protocol, such as a change in PCR conditions or the transfer of protocols between laboratories.

Conclusion

Quantitative inference based on DNA sequence counts is commonplace in the microbial ecology literature; although recent studies recognize the need to account for species differences in gene copy number that can largely impact estimates of relative abundance. Factors biasing sequence read proportions in most metabarcoding studies have until now limited analyses to descriptions of biodiversity, or at best, semi-quantitative estimates of the relative proportions of species. In this study, we outline a method by which researchers can control for many of the biasing factors involved in DNA metabarcoding using 50/50 mixtures of the target species and a control species. Although this method does not account for all biases, the correction factors generated from the 50/50 tissue library greatly improved the relationship between DNA sequence read abundance and mass percentage, and could facilitate quantitative inquiry in future studies. The usefulness of 50/50 RCFs as a tool in DNA metabarcoding studies will ultimately be dictated by the feasibility of creating tissue mixtures for the target species, and the level of accuracy needed to answer the research questions of interest.

Acknowledgements

The authors thank Dr. Carl Walters, Dr. Anthony Sinclair and Monique Lance for providing reviews on early drafts of this manuscript. We are also thankful for manuscript feedback from three anonymous reviewers and MER Editor Dr. Eric Coissac. The harbour seal scats were collected under an approved Canadian Department of Fisheries and Oceans Marine Mammal Research License (MML 2011-10), and a University of British Columbia Animal Care Permit (A11-0072). Financial support the project was provided by the Pacific Salmon Foundation.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

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

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**, 1–13.

Benson DA, Cavanaugh M, Clark K *et al.* (2012) GenBank. *Nucleic Acids Research*, **28**, 15–18.

Berry D, Ben Mahfoudh K, Wagner M, Loy A (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology*, **77**, 7846–7849.

Bik HM, Porazinska DL, Creer S *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233–243.

Brown DS, Jarman SN, Symondson WOC (2012) Pyrosequencing of prey DNA in reptile faeces: analysis of earthworm consumption by slow worms. *Molecular Ecology Resources*, **12**, 259–266.

Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

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

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.

Creer S, Fonseca V, Porazinska D *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, **19**, 4–20.

Cristescu ME (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, **29**, 566–571.

Deagle BE, Tollit DJ, Jarman SN *et al.* (2005) Molecular scatology as a tool to study diet: analysis of prey DNA in scats from captive Steller sea lions. *Molecular Ecology*, **14**, 1831–1842.

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, 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, 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.

DeVaney S, Pietsch TW (2006) Key to the fishes of puget sound. Available from: http://www.burkemuseum.org/static/FishKey/ (accessed 28 November 2012).

Eckburg PB, Bik EM, Bernstein CN *et al.* (2005) Diversity of the human intestinal microbial flora. *Science*, **308**, 1635–1638.

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

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. *PeerJ PrePrints*, **3**, e1258.

Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One*, **7**, e42543.

Fonseca VG, Carvalho GR, Sung W, et al. (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.

Huggett JF, Laver T, Tamisak S *et al.* (2013) Considerations for the development and application of control materials to improve metagenomic microbial community profiling. *Accreditation and Quality Assurance*, **18**, 77–83.

Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Computational Biology*, **8**, e1002743.

Kowalczyk R, Taberlet P, Coissac E *et al.* (2011) Influence of management practices on large herbivore diet—Case of European bison in Białowieża Primeval Forest (Poland). *Forest Ecology and Management*, **261**, 821–828.

Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL (2013) Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, **10**, 999–1002.

Murray DC, Bunce M, Cannell BL *et al.* (2011) DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches. *PLoS One*, **6**, e25776.

Nguyen NH, Smith D, Peay K, Kennedy P (2014) Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist*, **205**, 1389–1393.

Olesiuk PF (1993) Annual prey consumption by harbour seals (*Phoca vitulina*) in the Strait of Georgia, British Columbia. *Fishery Bulletin*, **91**, 491–515.

Olesiuk P (2009) An assessment of population trends and abundance of harbour seals (*Phoca vitulina*) in British Columbia. *DFO Canadian Science Advisory Secretariat Research Document* 2009/105, vi + 157 pp.

Piñol J, San Andrés V, Clare EL, Mir G, Symondson WOC (2014) A pragmatic approach to the analysis of diets of generalist predators: the use of next-generation sequencing with no blocking probes. *Molecular Ecology Resources*, **14**, 18–26.

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.

Pinto AJ, Raskin L (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One*, **7**, e43093.

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

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

Soininen E, 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**, e16.

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

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

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.

Vestheim H, Jarman SN (2008) Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. *Frontiers in Zoology*, **5**, 12.

Willerslev E, Davison J, Moora M *et al.* (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, **506**, 47–51.

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

Zhou J, Wu L, Deng Y *et al.* (2011) Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME Journal*, **5**, 1303–1313.

A.C.T. contributed to study design, sample processing, data analysis, manuscript writing and editing; B.E.D., to study design, data analysis, manuscript writing and editing; J.P.E., to data analysis, manuscript writing and editing; C.H.H., to sample processing and editing; A.W.T., to securing funding, study design and editing.

Data accessibility

All Illumina sequence data in FASTQ format and QIIME mapping files for samples have been archived in Dryad Digital Repository. Likewise, the BLAST database and processed sequence data from the tissue mix experiments have been archived to facilitate alternative analyses (doi: 10.5061/dryad.7dv96).