MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 620–633 doi: 10.1111/1755-0998.12103

Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count?

BRUCE E. DEAGLE,*[1] AUSTEN C. THOMAS,†[1] AMANDA K. SHAFFER,‡ ANDREW W. TRITES† and SIMON N. JARMAN*

*Australian Antarctic Division, Channel Highway, Kingston, Tas 7050, Australia, †Marine Mammal Research Unit, Fisheries Centre, University of British Columbia, 2202 Main Mall, Vancouver, BC V6T 1Z4, Canada, ‡Point Defiance Zoo and Aquarium, 5400 North Pearl Street, Tacoma, WA 98407, USA

Abstract

A goal of many environmental DNA barcoding studies is to infer quantitative information about relative abundances of different taxa based on sequence read proportions generated by high-throughput sequencing. However, potential biases associated with this approach are only beginning to be examined. We sequenced DNA amplified from faeces (scats) of captive harbour seals (*Phoca vitulina*) to investigate whether sequence counts could be used to quantify the seals' diet. Seals were fed fish in fixed proportions, a chordate-specific mitochondrial 16S marker was amplified from scat DNA and amplicons sequenced using an Ion Torrent PGM™. For a given set of bioinformatic parameters, there was generally low variability between scat samples in proportions of prey species sequences recovered. However, proportions varied substantially depending on sequencing direction, level of quality filtering (due to differences in sequence quality between species) and minimum read length considered. Short primer tags used to identify individual samples also influenced species proportions. In addition, there were complex interactions between factors; for example, the effect of quality filtering was influenced by the primer tag and sequencing direction. Resequencing of a subset of samples revealed some, but not all, biases were consistent between runs. Less stringent data filtering (based on quality scores or read length) generally produced more consistent proportional data, but overall proportions of sequences were very different than dietary mass proportions, indicating additional technical or biological biases are present. Our findings highlight that quantitative interpretations of sequence proportions generated via high-throughput sequencing will require careful experimental design and thoughtful data analysis.

Keywords: DNA barcoding, Ion Torrent, metabarcoding, next-generation sequencing, pinniped diet

Received 2 May 2012; revision received 26 February 2013; accepted 7 March 2013

Introduction

The advent of high-throughput sequencing methods allows genetic markers to be characterized at an unprecedented scale and has greatly enhanced the scope of studies using DNA-based identification methods (Valentini *et al.* 2009b). One area of particular interest is analysis of species diversity in environmental samples via recovery of many taxonomically informative sequences from DNA mixtures. High-throughput sequencing was initially applied in ecological studies to characterize microbial taxa (e.g. Sogin *et al.* 2006), but has been extended into the realm of eukaryotic organisms including studies focussed on microscopic eukaryotes (e.g. Porazinska *et al.* 2009; Bik *et al.* 2012), soil fungal communities (e.g. Buée *et al.* 2009), diversity of invertebrate or vertebrate populations (e.g. Hajibabaei *et al.* 2011; Andersen *et al.* 2012) and food species in diets of herbivores and carnivores (e.g. Deagle *et al.* 2009; Valentini *et al.* 2009a). These studies used PCR to amplify a variety of different markers and often employed molecular tagging techniques to distinguish between different strata or individual samples to take advantage of the large amount of data produced by each high-throughput sequencing run (e.g. Meyer *et al.* 2007). This enables the analysis of dozens of environmental samples in parallel, and hundreds or thousands of sequences can be recovered from each to provide a profusion of data about species diversity.

The goal of many environmental barcoding studies is to infer relative taxon abundance from proportions of

Correspondence: Bruce Deagle, Fax: +61 3 6232 3288; E-mail: bruce.deagle@aad.gov.au

[1]These authors contributed equally to the study.

different sequence reads recovered (Amend *et al.* 2010; Deagle *et al.* 2010). However, there are a myriad of potential biases associated with using sequence counts to quantify organisms. These include potential biases caused by biological attributes of the target taxa (e.g. taxon-specific variation in DNA copy number per cell, variation in tissue cell density or differences in environmental persistence). Technical biases can also be introduced at each laboratory and analytical step. Biases caused by target-specific differences in PCR amplification have been well scrutinized because a PCR amplification step is also crucial in traditional clone sequencing approaches (Polz & Cavanaugh 1998; Acinas *et al.* 2005; Sipos *et al.* 2007), but technical biases unique to high-throughput sequencing are just beginning to be evaluated. These include unavoidable sampling variance between template DNA molecules, but also systematic biases that cause final sequence counts to deviate from proportions present in template DNA molecules. For example, it has recently been reported that tagged PCR primers used for multiplex amplicon sequencing can impact bacterial community profiles obtained through pyrosequencing (Berry *et al.* 2011). Another study using pyrosequencing to look at fungal communities found that sequence count differences between species were due in part to biases introduced during bioinformatic filtering (Amend *et al.* 2010). Biases in sequences recovered based on GC content have also been documented from the Ion Torrent sequencer (Quail *et al.* 2012).

Several dietary DNA barcoding studies have used high-throughput sequencing to characterize food DNA amplicons recovered from faecal (scat) samples (reviewed in Pompanon *et al.* 2012), and in many cases, sequence counts have been reported as a semi-quantitative proxy for diet composition (Deagle *et al.* 2009; Soininen *et al.* 2009; Kowalczyk *et al.* 2011; Murray *et al.* 2011; Brown *et al.* 2012). One study using pyrosequencing found the proportions of four primary fish prey amplicon sequences recovered from little penguin scats were similar to those obtained with parallel qPCR analysis, suggesting that sequencing-related biases were not large (Murray *et al.* 2011). Another study of Australian fur seal diet showed that prey sequence proportions generated by pyrosequencing were consistent when two different-sized mtDNA barcoding amplicons were used (Deagle *et al.* 2009). The sequence counts from these studies are generally presented as fixed values, as in other related fields (e.g. Yergeau *et al.* 2012), despite the fact that counts are potentially influenced by many decisions made throughout the experimental procedure and bioinformatic pipeline (see Amend *et al.* 2010).

Here, we examine count data of fish DNA sequences recovered from scats of captive harbour seals fed a constant diet. The analysis was carried out using amplicon sequencing on the Life Technologies Ion Torrent Personal Genome Machine™ (Ion PGM) sequencer (Rothberg *et al.* 2011). Our initial objective was simply to see whether proportions of prey in diet were reflected in the proportion of prey sequences recovered; however, our analysis highlighted the fluidity of the count proportions and led us to examine the influence of experimental factors on the recovered prey sequence proportions. We specifically considered (i) sequences obtained from the forward and reverse read directions, (ii) samples marked with different identification tags (added before or after sample PCR amplification) and (iii) data filtered with various levels of quality control stringency and different minimum read length thresholds. The interactions between these factors were also considered and a subset of samples was re-examined on a second sequencing run to see whether results were congruent.

Materials and methods

Overview of genetic analysis

In the current study, a chordate-specific mitochondrial marker (~120 bp) was amplified from scats of captive seals (targeting the three fish species in their diet) and the amplicons were examined in two Ion Torrent sequencing runs. Amplicons were labelled with a unique combination of a 3-bp sequence incorporated onto PCR primers (tag sequences – Tag A, Tag B or Tag C) and one of 16 different 11-bp multiplex identifier sequences (MIDs) added after PCR amplification. In Run I, amplicon sequences from 48 scat samples were analysed, and sufficient data obtained from 39 of these. For this run, sequences over 100 bp were considered (Run I – 100 bp) and a parallel analysis included shorter sequences (Run I – 90 bp). In Run II, amplicons from eight scat samples were analysed in triplicate (with a different primer tag in each replicate). The second run was carried out with newer sequence chemistry, and most sequences were >100 bp, so one data set was considered (Run II – 100 bp). Details are outlined below.

Feeding trials and scat sampling

The feeding trial was carried out with five adult female harbour seals at Point Defiance Zoo and Aquarium (Tacoma, WA, USA) between 1 July and 17 August 2011. The seals occupied a single pool and were fed a constant diet of four species in fixed proportions: capelin (*Mallotus villosus*) (40%), Pacific herring (*Clupea pallasii*) (30%), chub mackerel (*Scomber japonicus*) (15%) and market squid (*Loligo opalescens*) (15%). Individual species within daily rations were weighed to the nearest 0.1 kg and distributed evenly across three meals in which seals consumed every fish. Daily food intake varied based on seal body mass and their interest in food, but diet proportions were maintained within measurement precision (2.0% SD per species; see

Table S1, Supporting information for a complete record of each animal's diet).

During the trial, seal scat samples were collected from pool and haul-out areas (generally within 2–4 h of deposition), put into Ziploc bags and stored at −20 °C. We wanted to completely homogenize samples because prey DNA is not evenly distributed in pinniped scats (Deagle *et al.* 2005). We also wanted to remove all prey hard parts, so they did not influence the genetic data, and to make the protocol useful for studies incorporating parallel hard-part analysis (e.g. Tollit *et al.* 2009). To accomplish this, our sampling procedure involved transferring thawed individual scats into a 500-mL plastic container lined with a 124-μm nylon mesh strainer. We poured 200 mL of 90% ethanol over the scat which was then manually homogenized to form an ethanol-scat slurry. The strainer was removed along with prey hard parts, and the ethanol preserved scat sediment was stored at −20 °C for up to 3 months. DNA extraction was performed on approximately 20 mg of material using QIAamp DNA Stool Kit (QIAGEN) following Deagle *et al.* (2005) with elution in 100 μL AE buffer.

Amplicon library preparation

The barcoding marker we used was a mitochondrial 16S fragment which is approximately 120 bp in length and has been used previously for differentiating fish species (see Deagle *et al.* 2009). We amplified this marker with primers Chord_16S_F (CGAGAAGACCCTRTGGAGCT) and Chord_16S_R_Short (CCTNGGTCGCCCCAAC) which bind to sites that are almost completely conserved in chordates. Amplicons from the three fish species are within a few base pairs in length but differ by more than 20% sequence divergence (see Table S2, Supporting information for sequence alignments including primer binding region). Initially, we also ran PCRs with a second primer set which would amplify squid DNA in addition to fish (see Deagle *et al.* 2009); however, the amplicon length was >250 bp and initial Ion Torrent library preparations failed (new library preparation procedures now allow sequencing of fragments >400 bp). Therefore, this marker was abandoned and the squid diet portion excluded from subsequent analyses. To limit amplification of seal DNA, a 32-bp blocking oligonucleotide (see Vestheim & Jarman 2008) matching harbour seal sequence was used in PCR (with a modified C3 spacer at the 3′-end to prevent extension; details in Table S2, Supporting information). All PCR amplifications were performed in 20 μL volumes using a Multiplex PCR Kit (QIAGEN). Reaction mixture contained 10 μL master mix, 0.25 μM of each primer, 2.5 μM blocking oligonucleotide and 2 μL template DNA. Thermal cycling conditions were 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. Products were checked on 1.8% agarose gels.

We prepared amplicon libraries for two Ion Torrent sequencing runs. The first (Run I) contained equal volumes of DNA amplified from 48 individual scats with each sample being uniquely labelled (see below). The second (Run II) was a reanalysis of three new PCR amplifications of DNA from each of eight scats characterized in the initial run. The purpose of this Run II was to see whether the results (and technical biases) were consistent between runs. Ion Torrent protocols existing at the time only allowed differentiation of 16 samples, so a two-step sample tagging process was used to differentiate between amplicons from the 48 individual scat samples in Run I and the 24 samples in Run II (Fig. 1). Both tagging approaches are routinely used to differentiate samples in studies employing high-throughput sequencing platforms. In step 1, short tags added to the 5′ end of the primer were incorporated into amplicons during PCR. In our case, we amplified DNA extracted from each scat sample using primers containing one of three different 3 bp primer tags (Tag A = CAT, Tag B = GCA, Tag C = TAC; for a given sample both forward and reverse primers had identical tags).The forward primer contained an additional 3-bp spacer (ATG) after the primer tag. These tags allowed us to identify three

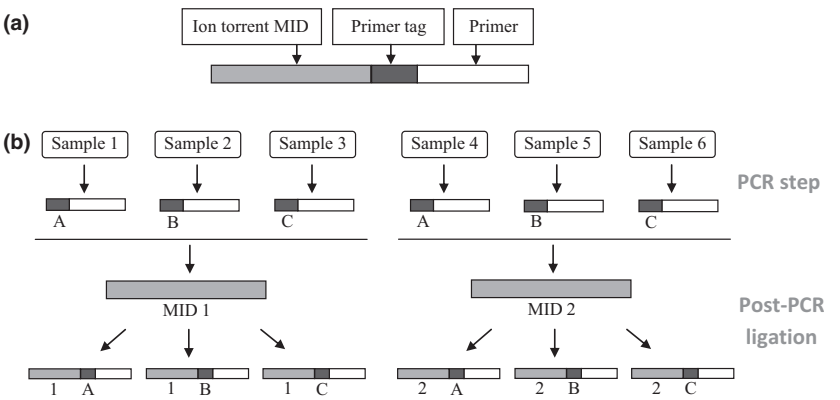

Fig. 1 Schematic showing (a) the combination of multiplex identifier sequence (MID) and primer tag used to identify amplicons from individual samples. (b) The sample labelling procedure. First, this involved PCR amplification of scat template DNA using one of three tagged primer sets (A, B, C). Second, an Ion Torrent MID was ligated to the amplicons (16 different MIDs), such that all samples received a unique combination of primer tag and MID.

groups of PCR amplicons. In step 2, we used the Ion Barcoding 1–16 kit (Life Technologies; part no. 4468654 Rev. B) which ligates up to 16 unique 11-bp multiplex identifier sequences (MIDs) onto amplicons post-PCR. PCR amplicons containing unique tagged primers were assigned to one of 16 Ion Torrent MIDs, thus creating 48 unique combinations of primer tags and MIDs for individual samples in Run I. This tagging scheme was used in part to evaluate tag-specific bases. Individual tagging of samples could also have been achieved using many uniquely tagged primer sets; however, that approach would not allow for replication of primer tags sufficient to evaluate tag biases. Sequencing Run II was intended in part to decouple the potential effects of individual sample variability and MID sequence from the effects of primer tags. In this sequencing run, each of eight samples was amplified with all three tagged primer sets, and the MID sequence was kept constant for each sample.

Sequencing

We used the Ion OneTouch™ System (Life Technologies) to prepare amplicons (already containing MIDs and associated capture and sequencing primers) for sequencing following the appropriate user's guide protocol. In the single year that we have been working with the Ion Torrent system, at least four different sequencing kit upgrades have been released. Therefore, the two sequencing runs we report here were carried out with different kits. The first run was performed using the Ion OneTouch Template kit (p/n 4468660) and the second with the Ion OneTouch™ 200 Template Kit v2 (p/n 4478316). The resultant enriched Ion Sphere™ particles were loaded onto 314 Ion semiconductor sequencing chips, and sequencing was carried out on the Ion PGM sequencer. Bidirectional sequencing was performed (i.e. sequence reads started from forward and reverse PCR primers), but reads were not paired. Each run was expected to produce approximately 100 000 reads. For Run I, expected read length was 100 bp (~75 bp being target-specific sequence, as this estimate includes the PCR primer and primer tag), so the full 16S fragment was not covered in a single read. In the second run, due to improved chemistry, reads were expected to be 200 bp in length which covers the full amplicon.

Bioinformatics

The Ion Torrent platform automatically sorted sequences based on the 16 MIDs, removed the MID sequence and output a single FASTQ file for each MID. Quality metrics were based on reanalysis of raw data carried out at the end of the study with TORRENT SUITE software version 2.0.1. All postsequencing analyses (except for taxonomic assignment; see below) were carried out using the R language (R Development Core Team 2010) making use of the Bioconductor packages ShortRead (Morgan *et al.* 2009) and Biostrings (all relevant FASTQ files and R code are available in Dryad). Our approach was slightly unconventional in that we kept all of sequences above the cut-off sequence length in the final database. This included sequences that were low quality, taxonomically unassigned and those that did not match a primer. Briefly, the procedure involved importing FASTQ output files into R, and sequences along with quality information were extracted. Sequences and quality information were trimmed to 100 bp, and data from shorter sequence reads were discarded. Sequences were exported in FASTA format, and prey species assignment was performed using the software package QIIME (Caporaso *et al.* 2010). In QIIME, a BLAST search for each sequence (removing tag and start of primer sequence) was performed against a local reference database containing 16S sequences for the three fish species and harbour seal. The match of each Ion Torrent sequence to reference sequences was assessed based on having a BLASTN *e*-value less than a relatively strict threshold value of $E < 1e\text{-}20$ and a minimum identity of 0.9. The minimum identity score and our predefined reference sequences prevented assignment of chimeric sequences. Resultant species assignments (including a category for sequences with no blast hit) were imported back into the R workspace. Sequence quality scores for all base calls were incorporated into the data set and mean quality scores were calculated. For each sequence, read direction was determined and sequences were matched to their individual sample of origin where possible (based on primer sequence, MID number and primer tag). For a sequence to be linked to a specific sample and read direction, it had to match the 3-bp primer tag and the first 11 bases of the primer (11 bp chosen to avoid a homopolymer run in the reverse primer). This included the ATG spacer sequence in the forward primer, and we allowed for mismatches at two variable sites in the reverse primer. The resultant data set, containing all sequences in the original 100-bp FASTQ files and related information, could then be queried based on quality score, read direction, tag identity, MID identity, etc., and sequences tallied based on taxonomic assignments. In Run I, many of the sequences were <100 bp in length; thus, for comparison, a parallel data set was created using a 90-bp size cut-off.

Results

Overview of sequence data (Run I + II)

The sequencing of Run I (amplicons from 48 individually identifiable scats) produced 330 594 reads with a

mean length of 102 bp (33.70 Mbp of data; 23.72 Mbp of Q20 Bases). The total number of Ion Torrent sequences generated varied considerably between the 16 MIDs (mean = 18 687, range = 1–45 972) with 22 338 sequences unassigned to a MID. The low sequence counts from some MIDs are likely due to errors made in the course of a complex MID labelling protocol (pooling of PCR products with different tags within a MID show very even recoveries, so this step is unlikely to cause these differences). Three of the 16 MIDs were excluded from further analyses due to low overall sequence counts (<400 sequences/scat sample). For the remaining 13 MIDs, representing 39 scat samples, a total of 297 049 sequences were exported into FASTQ files (mean read number per MID = 22 850; range = 2206–45 972). Of these sequences, 63% (n = 188 534) were over 100 bp in length (93% were more than 90 bp in length) and these were assigned to local reference sequences using BLAST.

For the Run I, 100-bp data set, 70% (n = 131 571) of sequences could be linked with a specific scat sample based on their match with a PCR primer and associated tag. The mean quality score for sequences matching a primer was 25.0 vs. 20.5 for those without a match. Of the primer-matched sequences, 84% (n = 110 270) were assigned to species in our reference library based on local BLAST assignation. The vast majority of assigned sequences matched the three fish species in the seals' diet, with only 2.3% (n = 2522) identified as harbour seal. While only sequences with an identified primer and taxonomic assignment were considered in the final analysis, we also examined the discarded sequences. More than half of the sequences that were excluded because they did not match a primer could be assigned to a prey species (n = 30 614) using our stringent local BLAST. A subset of sequences without taxonomic assignment (including those without a primer match, or a primer match but no local BLAST match) were characterized against the full NCBI nucleotide database. The top hit for the majority of these sequences were feeding trial prey species, but were below the minimum identity (potentially including chimeric sequences). Most others had no strong matches in the database; however, a small percentage of sequences matched those from the preceding Ion Torrent sequencing run (humpback whale nuclear gene amplicons; 0.5% of 100-bp data set; n = 971). These contaminating sequences likely resulted from carry over in the OneTouch instrument used for emulsion PCR (a new cleaning procedure for maintaining the instrument between runs has since been implemented by Life Technologies). Overall, results from the Run I, 90-bp data set, were similar to those reported for the 100-bp data set and are reported in supplementary material (Fig. S1, Supporting information).

A subset of samples were resequenced in Run II; these amplicons were from new PCR amplifications of DNA from eight scats characterized in the initial sequencing run. DNA from each scat was amplified in triplicate (once with each set of tagged primers), and amplicons from each of the eight samples were labelled with a separate MID sequence. This run with new sequencing chemistry produced 405 211 reads with a mean length of 151 bp (61.30 Mbp of data; 31.84 Mbp of Q20 Bases). The total number of Ion Torrent sequences was more consistent between the 8 MIDs used in Run II (mean = 37 010, range = 24 334–48 391) with 104 140 sequences unassigned to a MID. Despite only 8 MIDs being employed in Run II, some sequences were allotted to each of the 16 potential MIDs. The sequences from eight unused MIDs represented only 0.6% of sequences (n = 2328; range 6–1415 per MID) and were generally low-quality sequences (100 bp mean = 17.6). These sequences primarily matched prey species of this study and likely represent rare mis-assignment of sequences between MIDs rather than contamination because these amplicons had not been sequenced in the previous 10 runs. Low levels of contaminating sequences from the previous sequencing run were present (despite using the new OneTouch cleaning protocol). The contaminants were sheared long-range PCR amplicons (human DNA) and were apparent because many of the recovered sequences exceeded the maximum size of target mtDNA amplicons.

From the eight correctly classified MIDs, a total of 296 079 sequences were exported into FASTQ files and 96% of these were over 100 bp in length. For the Run II, 100-bp data set, 56% (n = 159 952; mean quality 26.8) could be linked to a specific PCR sample based on the primer; this percentage was low compared with Run I due to more nontarget sequences (without close blast matches) being recovered. Of sequences which contained primers, 87% (n = 139 630) were assigned to species in our reference library. Harbour seal sequences made up 5.8% (n = 8087) of these assigned sequences.

Fish species proportions in 39 scats (Run I)

The proportion of three fish species consumed in the diet was known, so our initial objective was to simply see whether these proportions were reflected in the sequence counts. Based on previous experiments, we expected relatively low variation in the proportions of prey sequences amplified from scats of animals fed a consistent diet. (Deagle & Tollit 2007; Bowles *et al.* 2011). If we consider the average composition of 39 scat samples based on all assigned sequences >100 bp, there was little variability in the proportions of sequences assigned to the fish species (Fig. 2). These sequence proportions do

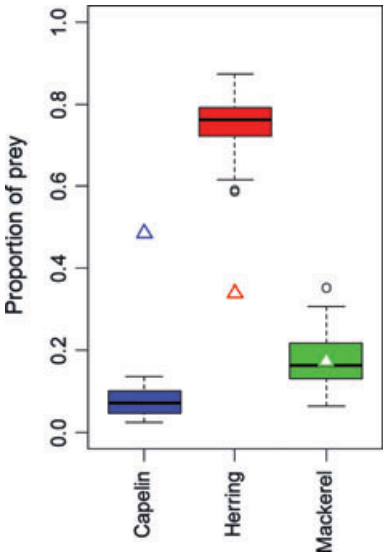

Fig. 2 Comparison between mass proportion of three fish species fed to seals (triangles) vs. overall proportions of sequence reads recovered (box plots). Box plots were generated from the sequence read proportions from 39 individual scat samples (Run I – 100 bp) using combined forward and reverse reads.

not match proportions of the three species consumed. Capelin was considerably underrepresented (7.3 ± 3.0% SD vs. 48.5% of fish diet), herring was considerably overrepresented by sequence proportions (74.8 ± 7.0% SD vs. 34% of fish diet), and mackerel matched the diet (17.9 ± 6.7% SD vs. 17.5% of fish diet). The discrepancy could be caused by many factors (such as PCR bias or biological differences between prey). However, here our focus is specifically on how choices made throughout the experimental procedure and during bioinformatic sorting impact proportions of various species in the sequence counts.

Influence of read direction and size cut-off (Run I)

Despite forward and reverse DNA strands being present in equimolar amounts after PCR, sequencing read direction substantially influenced the proportions of sequences assigned to each fish species (Fig. 3; Table 1). In the forward read direction, by far the largest percentage of sequences were herring (85.5 ± 9.5% SD) with very few sequences from mackerel (10.0 ± 7.2% SD) or capelin (4.5 ± 3.5% SD). In the reverse read direction, proportions of sequences were substantially different: herring (47.4 ± 17.7 SD), followed by mackerel (39.5 ± 17.1% SD), then capelin (13.1 ± 5.6% SD). Sequence counts indicate the differences between forward and reverse reads were primarily driven by a bias favouring herring fragment reads in the forward direction (Fig. 4; Table 1). The proportions of various sequences recovered were also influence by the arbitrary sequence length cut-off point used to define the final data set. When all sequences >90 bp (Run I – 90 bp) were considered (rather than only those >100 bp), the differ-

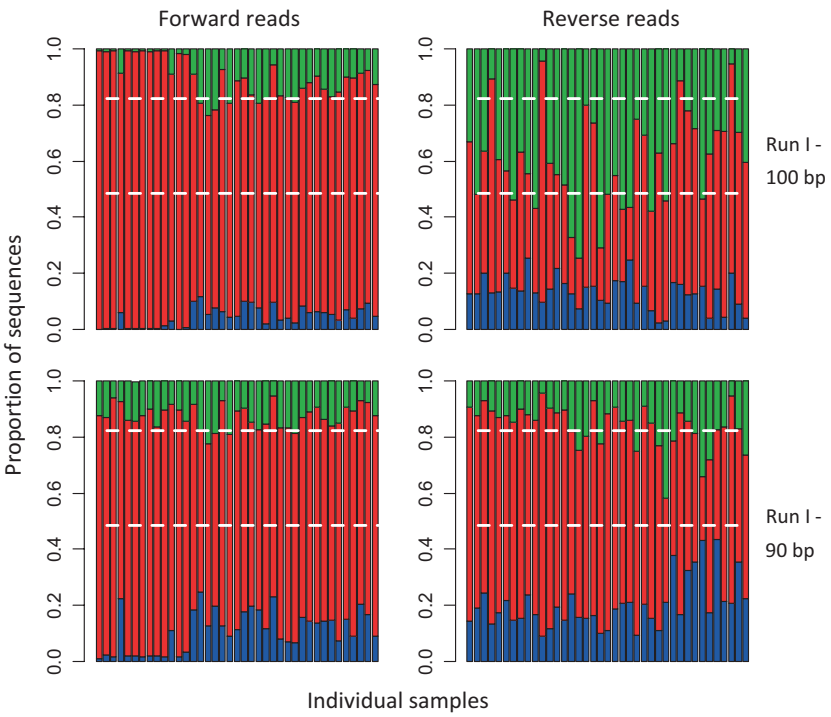

Fig. 3 Bar plots showing proportions of fish sequences recovered from 39 individual seal scats in sequenced in Run I (blue = capelin, red = herring, green = mackerel). Each bar represents an individual sample, and proportions of forward and reverse reads are shown separately. Data were filtered to retain either sequences >100 bp (top) or >90 bp (bottom). Proportions of three fish species by mass in the diet are shown as dotted lines on plots.

Table 1 Sequence counts and percentages of three fish species recovered from seal scats in two Ion Torrent sequencing runs. Run I data are from 39 scat samples, and Run II data are from a subset of these scats (n = 8) each rerun in triplicate with different primer tags (A, B or C). Data are shown for various subsets of recovered sequences (both without quality filtering and when only sequences with high quality scores are considered)

			No quality filter			Mean sequence quality >28		
Data/Subset	Primer		Capelin	Herring	Mackerel	Capelin	Herring	Mackerel
Diet/Fish*		%	48.5	34	17.5			
Run I – 100 bp†	F	%	4.5 ± 3.5	85.5 ± 9.5	10.0 ± 7.2	2.8 ± 2.9	91.5 ± 6.5	5.7 ± 4.7
		Mean count¶	90	1586	209	25	822	52
	R	%	13.1 ± 5.6	47.4 ± 17.7	39.5 ± 17.1	15.1 ± 7.8	22.6 ± 22.6	62.3 ± 22.6
		Mean count¶	121	456	301	47	95	167
Run I – 90 bp‡	F	%	10.8 ± 7.2	76.7 ± 7.9	12.5 ± 4.2	6.5 ± 6.1	83.4 ± 6.3	10.0 ± 3.8
		Mean count¶	233	1588	276	61	860	107
	R	%	20.2 ± 8.8	64.0 ± 13.3	15.7 ± 7.9	29.7 ± 17	45.1 ± 26.4	25.1 ± 15.3
		Mean count¶	440	1348	313	239	394	176
Run II/TagA§	F	%	16.7 ± 4.2	68.4 ± 2.6	14.9 ± 4.3	15.8 ± 5	71.9 ± 3.7	12.3 ± 3.6
		Mean count¶	700	2940	651	425	2034	348
	R	%	19.2 ± 2.9	64.3 ± 3.7	16.6 ± 6.1	20.9 ± 3.1	64.8 ± 3.7	14.3 ± 5.6
		Mean count¶	707	2395	628	516	1608	359
Run II/TagB§	F	%	13.8 ± 3.6	74.2 ± 3.1	12 ± 3.5	12.5 ± 3.6	79.9 ± 3.5	7.6 ± 2.4
		Mean count¶	271	1469	238	170	1114	104
	R	%	14.9 ± 2.9	72 ± 4.4	13.2 ± 4.8	15.8 ± 3.3	72.6 ± 4.9	11.6 ± 4.5
		Mean count¶	229	1102	200	175	795	126
Run II/TagC§	F	%	14.4 ± 3.4	72.4 ± 2.7	13.2 ± 4.1	12.9 ± 3.4	79.5 ± 4.4	7.6 ± 2.9
		Mean count¶	376	1926	351	231	1476	137
	R	%	20.4 ± 4	61.1 ± 4.6	18.4 ± 6.5	35.0 ± 11.9	41.0 ± 12.1	24.0 ± 8.7
		Mean count¶	454	1382	424	333	468	248

*Percentage composition of fish species in seals' diet.
†Data from sequencing Run I (amplicons from 39 scats) including only sequences >100 bp in length.
‡Data from sequencing Run I including all sequences >90 bp in length.
§Data from sequencing Run II, amplicons from eight scats run in triplicate with different primer tags (A,B or C).
¶Mean number of sequences recovered per sample within data subset.

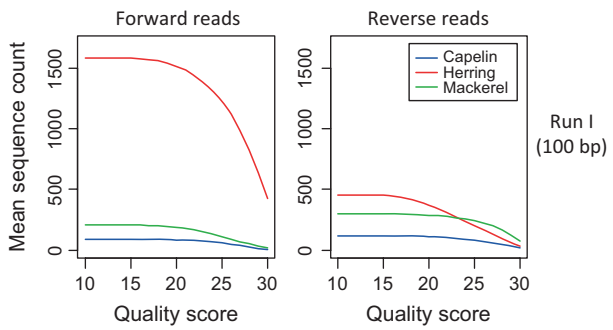

Fig. 4 Mean sequence counts for fish in 39 individual seal scats for various levels of quality filtering (Run I – 100 bp; forward and reverse reads are shown separately).

ences between forward and reverse reads were less dramatic (Fig. 3; Table 1).

Tag and MID biases (Run I)

In addition to being influenced by read direction, sequence proportions were also influenced by primer tags added during PCR to trace sequences back to their sample of origin (Fig. 5). In the forward read direction, a higher proportion of herring DNA fragments were amplified and sequenced from primers containing Tag A (97.1 ± 4.6% SD) than from either Tag B (77.8 ± 5.8% SD) or Tag C (81.6 ± 2.7% SD). In the reverse read direction, there was more variation in prey proportions within each tag, but substantial differences between tags were also apparent (e.g. Herring Tag A: 46.2 ± 17.6% SD; Tag B 37.9 ± 17.8% SD; Tag C 58.1 ± 11.8% SD). The differences in species composition between tags were not consistent between read directions, suggesting values do not represent the true differences between samples (Fig. 5). In Run II, we processed individual samples with different tags to examine the tag effect further (see below). Only three samples were sequenced with each Ion Torrent MID, so we had little power to evaluate variability in sequence proportions between MIDs. However, there were some differences between MIDs that warrant further examination. For example, the length of sequence reads between MIDs varied slightly; in our analysis, only 61% of the sequences from MID#4 were longer than

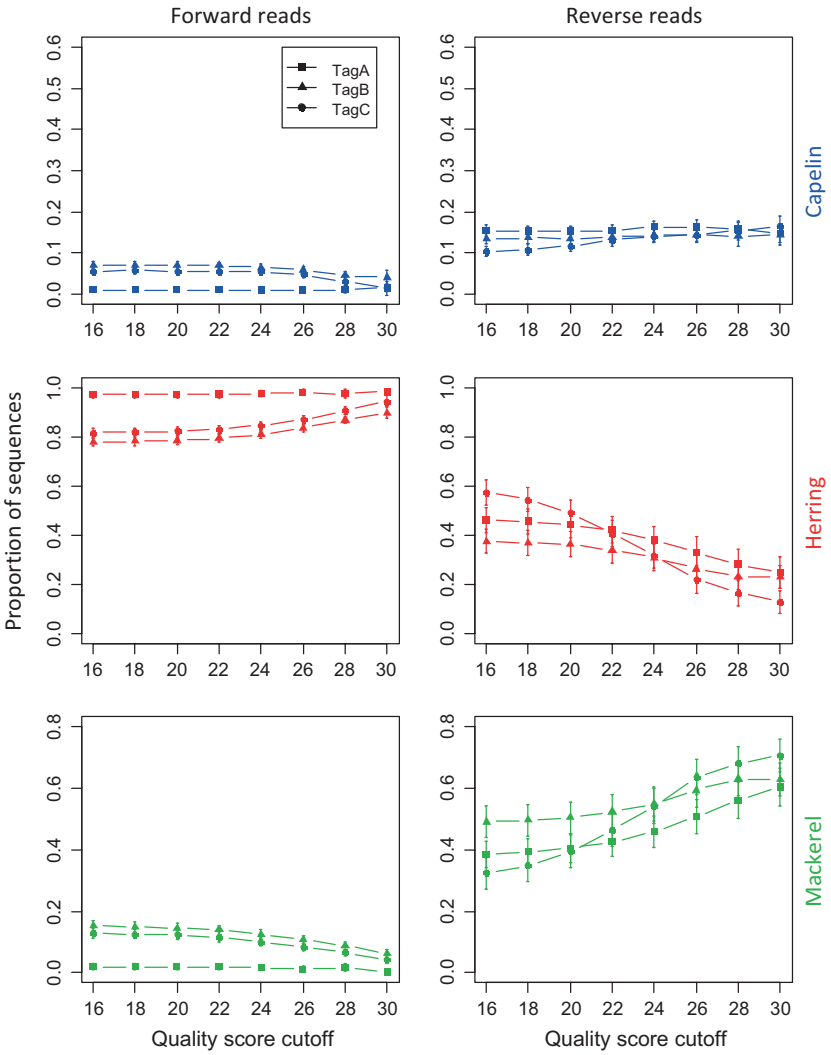

Fig. 5 Plots depicting the interacting effects of three different primer tags (A, B, C) and eight different quality filter cut-off values on proportions of fish sequences detected in 39 scats (Run I – 100 bp). Sequence proportions for each tag (represented by different shapes) at a given quality score cut-off (varies along the *x*-axis) and add up to 1. Results for forward and reverse read directions are displayed separately. Error bars represent standard error.

100 bp vs. 87% of sequences from MID#5. The quality scores also varied slightly between MIDs. For example, 28% of herring sequences labelled with MID#4 had mean quality scores over 30, vs. only 10% labelled with MID#5 (calculated over 100 bp for both).

Quality filtering bias (Run I)

Data reported up to this point were not quality-filtered beyond initial processing by the Torrent Suite software; however, postsequencing quality control in amplicon sequencing studies is generally carried out based on quality scores assigned to sequences. For amplicons in the current study, sequence quality generally diminished along the length of the read; quality scores were initially similar between species and then diverged, becoming species-specific as sequences became different (Fig. 6). Particular sequencing positions in both forward and reverse read directions had notably low quality scores. This was particularly apparent at the start of reads where species share the same sequence. For example, in the reverse read direction, nucleotide quality score dropped dramatically to its lowest point at sequencing position 15 (mean quality score = 18.2). That position corresponds with the third C in the CCCC homopolymer of the reverse primer. The majority of these primer sequences were incorrectly called as CCC (even when considering only higher-quality reads in Run I that matched the first 11 bp of the primer and were taxonomically assigned). Overall, mean sequence quality scores varied between species, and to some degree with sequencing direction (Fig. 6). In the forward direction, when quality scores were averaged over 100 bp, the highest-quality sequences overall were herring (mean = 27.4) followed by capelin (mean = 25.9) and then mackerel (mean = 25.2). For the reverse direction, the opposite trend was observed (Reverse: mackerel = 27.3; capelin = 26.2; herring = 25.0). These species differences in sequence quality resulted in predictable biases in sequence counts that were introduced

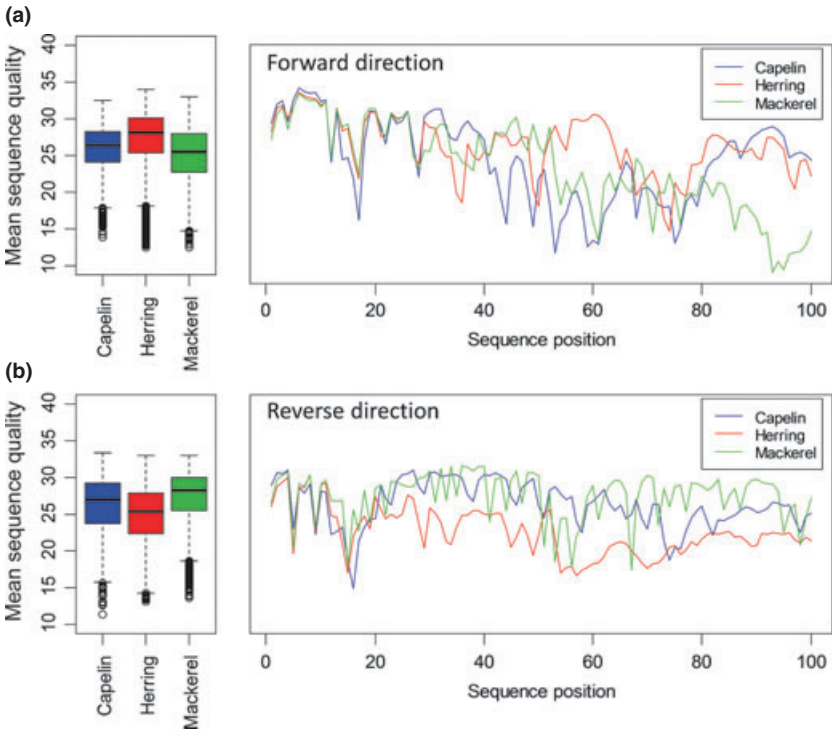

Fig. 6 Sequence quality scores vary between species and between (a) forward and (b) reverse reads. Box plots show summary of mean quality scores (median, range and upper/lower quartiles across 100 bp of sequence from Run I; n = 110 270 sequences). Line plots show variation in mean quality at positions along the sequence for each of target species in the same data set.

during quality filtering. For example, as quality score cut-off stringency increased for the reverse reads, more of the relatively higher-quality mackerel sequences were present and fewer of the lower-quality herring sequences were retained (Fig. 5).

Interactions (Run I)

The proportions of sequences assigned to the three species were also affected by interactions between the factors evaluated in this study (sequencing direction, size cut-off, primer tag and quality cut-off value; see Fig. 5). As mentioned above, sequence proportions in the forward direction responded differently to quality filtering than they did in the reverse direction. For example, the proportion of mackerel sequences decreased with additional quality filtering in the forward direction, whereas it increased with stricter quality filtering in the reverse direction. This effect was smaller in the 90-bp amplicon data set compared with the 100-bp data set (Fig. S1, Supporting information). Sequence proportions also responded differently to the primer tags depending on the level of quality filtering and read direction. In the forward direction, Tags B and C tended to converge with Tag A when the level of quality filtering was increased, but Tag A sequence proportions were virtually unchanged (Fig. 5). In contrast, in the reverse direction, Tag C responded more strongly to quality filtering that the other two. Again these effects were somewhat dependent on sequence length cut-off used (Fig. S1, Supporting information).

Rerun of subset of samples (Run II)

In the absence of quality filtering, the eight rerun samples produced reasonably consistent results between samples, between the three amplifications with different primer tags and between sequencing directions (Table 1; Fig. S1, Supporting information). These results were in general agreement with overall results obtained from the 39 scats sequenced in Run I (all assigned sequences >100 bp in Run II: capelin = 16.4 ± 3.3%; herring = 69.0 ± 4.4%; mackerel = 14.6 ± 4.7% vs. data in Fig. 2). However, mean values in Runs I and II were produced by quite different underlying values. For example, in Run I, the bias towards herring sequences in the forward read direction was much stronger compared with Run II (although observed to some extent in both runs; Table 1). Direct comparison of eight individual samples between runs was hampered because in Run I they were all amplified using primers labelled with Tag A, and these samples had a very large proportion of herring in the forward direction (Fig. 5). This Tag A effect was not observed in Run II, in fact Tag A replicates in Run II had a lower proportion of herring compared with other tags (individual samples had always had less herring when labelled with Tag A compared with Tag B, the mean difference was 5.8%; Table 1). While the tag effect

was minor in the second run, this bias and differences between sequencing directions became much more substantial with increased filtering based on quality scores (Table 1; Fig. S1, Supporting information). Quality filtering had the strongest impact on sequences labelled with Tag C in the reverse direction, similar to the effect seen in Run I. This three base pair primer tag (TAC) produces a homopolymer of three C's when combined with the reverse primer (only two C's with the other tags). This may explain why samples labelled with Tag C were more influenced by quality filtering. While the quality of sequences assigned to each species was generally higher in Run II compared with Run I, this was not the case for mackerel reverse read sequences. The lower relative quality of mackerel sequences in Run II compared with other species resulted in mackerel sequences being less common when high levels of quality filtering are applied (the opposite of the effect seen in Run I: Fig. 5; Fig. S1, Supporting information).

Discussion

There is considerable optimism about the use of high-throughput sequencing methods in DNA-based surveys of biodiversity, but biases associated with the approach are only beginning to be examined. Environmental barcoding studies generally characterize short PCR products, and these amplicon sequencing experiments are more strongly influenced by biases than more common applications of high-throughput sequencing such as resequencing of genes, genome or transcriptomes. In the latter experiments, biases can be overcome to a large extent by having multiple overlapping reads of the same regions. Here, we focus on sequence count proportion biases in the context of a DNA-based diet analysis of seals. Captive seals received a constant diet containing three fish species, and mtDNA barcode amplicons were recovered from their scat using an Ion PGM sequencer. To our knowledge, this is the first study examining biases obtained using Ion Torrent technology amplicon sequencing, although some biases have been evaluated in the context of bacterial genome resequencing (Quail *et al.* 2012). Overall, proportions of fish sequences recovered from the seal scats were not directly related to diet proportions; furthermore, the sequence proportions we recovered depended on many technical factors (e.g. influence of read direction, sequence identifier tags, quality filtering).

Sources of bias in amplicon sequence proportions

In our sequence counts from 39 scat samples, we observed large differences in sequence proportions from the three fish species between forward and reverse reads. For a given sample, forward and reverse sequences come from opposing strands of the same set of amplicons; so although the sequences differ (i.e. they are reverse complements), they should be present in equal numbers after PCR. During Ion Torrent sequencing,there are two additional amplification steps (one during library prep and then an emulsion PCR step) which could preferentially amplify certain DNA molecules (Quail *et al.* 2012). Alternatively, the sequencing process itself might be more efficient for certain sequences, resulting in deviation in proportions. A similar sequencing direction effect was noted in a previous pyrosequencing study, so this type of bias is not platform specific (Amend *et al.* 2010). Regardless of underlying cause, this type of bias could also affect representation of species in a mixture if there are large interspecific sequence differences.

Primer tags added to amplicons during initial template PCR, and identifier sequences ligated to products after PCR, have proven to be useful tools for differentiating between sequences with different origins within a high-throughput sequencing run. Recent evaluations of potential bias introduced by primer tags suggest that some tags are favoured in PCR and sequencing reactions, which leads to biased sequence proportions (Berry *et al.* 2011). Our results corroborate that conclusion. In our first sequencing run, 39 samples were split between three primer tags and proportions of sequences assigned to the three test species differed between tags. We explicitly analysed differences between PCR amplifications performed with different tags in a second sequencing run by examining eight samples using each of the three primer tags. Here, there were also differences between proportional estimates from different tags. For example, in Run II, the forward direction samples amplified with Tag A labelled primers always had less herring than when labelled with Tag B. In addition to PCR added tags, we also used different identification sequences added to amplicons post-PCR (MIDs). The post-PCR amplification steps mentioned above could differentially amplify MIDs; however, with only three samples per MID, we had very little statistical power to detect potential biases. The primer tag bias was generally not as large as the other biases we encountered, but the impact of biases caused by primer tag or MIDs could be particularly insidious as these identifiers are often used to discriminate between different groups of samples, or different experimental treatments. It would be prudent to design studies so that particular identifiers are used across treatments in different sequencing runs. With this type of design, it may be possible to evaluate tag introduced bias and if necessary correct for tag effects (or eliminate those tags producing outlier data). A two stage PCR (in which template DNA is first amplified using untagged primers and tagged primers are added during

the last few PCR cycles) has been suggested to reduce this bias (e.g. Berry *et al.* 2011; Hajibabaei *et al.* 2011). However, the increased risk of cross-contamination needs to considered, especially when amplifying from low-quality samples with small amounts of starting DNA template.

Some species produce higher-quality reads than others presumably due to their sequences differences; therefore, bioinformatic sorting based on quality scores introduces species-specific biases. While the number of sequences retained decreases as quality threshold goes up, there are abrupt decreases in sequences retained above a certain quality threshold for species with lower quality scores. The result is differing proportions of sequences from component species in data sets produced with different levels of quality filtering. We also observed that the distribution of quality scores for a particular species was occasionally bimodal, so changes in species composition based on quality were not always predictable based simply on species mean quality scores. One approach to deal with this bias may be to use less-quality filtering to avoid penalizing those sequences that tend to have a low quality score. However, retaining potential sequencing errors in data sets may result in difficulties with sequence assignment, so a trade-off will need to be made. As with pyrosequencing, the Ion Torrent sequence quality was particularly affected by homopolymer runs (see also Quail *et al.* 2012). During sequencing, these repeat sequences are called simultaneously, as signified by hydrogen ions being released during a single flow of nucleotide, and distinguishing multiple releases is problematic. Differences in frequency of homopolymers between species may lead to particularly strong divergences in quality score.

Interactions between the technical factors we evaluated were unexpected and highlight the difficulty in predicting sequence count biases likely to be present in a high-throughput sequencing data set. We found that differences between primer tags changed depending on stringency of quality filtering. This implies that both total number of sequences generated and sequence quality are somewhat dependant on the primer tag used in PCR. Also, primer tag biases were different between forward and reverse read directions, indicating that an interaction between template sequence and tag sequence is important, rather than a simply the tag sequence. The proportion of reverse reads from one primer tags that we used (Tag C) was particularly affected by quality filtering. Postsequencing examination of the primer sequence revealed that this particular tag created a 3-bp homopolymer when combined with the reverse primer (vs. two base pairs with other tags). This homopolymer lowers the quality scores of all sequences labelled with this tag resulting in more stringent filtering of these sequences relative to the other tags for a given quality cut-off level. Incorporating a small consistent spacer sequence between the identifier sequence and the primer could reduce this type of bias.

Our reanalysis of a subset of samples, to look at repeatability of sequence proportions and the repeatability of factors influencing those proportions, was somewhat confounded due to changes in sequence chemistry between runs. Despite this, overall sequence proportions were quite similar between runs. While this consistency is reassuring, the new results differ from the original data set in many aspects. In the second run, sequence proportions were considerably more similar between sequencing directions and between different primer tags (without stringent quality filtering). Some of the biases we observed in analysis of the original run were seen again (e.g. the Tag C effect mentioned in the previous paragraph), but other biases changed between runs (e.g. the quality of sequences obtained from different species changed slightly; thus, quality filtering had a different impact on sequence proportions). The extreme bias in Run I for recovery of herring sequences from forward reads labelled with Tag A (97% of these prey sequences) was not seen in the second run (68%), indicating an experiment-specific effect. This observation highlights the potential benefit of data averaging across multiple sequencing runs to minimize the influence of such outliers (although systemic biases will remain). The increased read length in the second run meant that very few sequences were filtered out due to short read length, an improvement since excluding sequences <100 bp in the first run magnified observed biases. In both runs, stringent quality filtering resulted in the largest deviations between proportions in forward and reverse reads. These results reaffirm that moderate levels of data filtering likely produce a more representative data set. This is likely to be especially important when there are large differences in sequence or quality score between amplicons.

Given that high-throughput sequencing technologies are currently in a period of rapid transition, it may be unrealistic to expect that one can define and correct for many of the platform-specific biases. For example, a recent Ion Torrent platform's software upgrade significantly changed sequence qualities derived from our first sequencing run (presumably due to ongoing improvements in algorithms used to process raw data); these types of changes make detailed analyses of sequence quality-related biases redundant soon after they are completed. This type of problem may become less of an issue as platforms stabilize; however, a new generation of single-molecule sequencing technologies is emerging and thus stabilization is unlikely to occur in the near future (Schadt *et al.* 2010). Spike-in standards (i.e., exoge-

nous DNA sequences) similar to those being promoted for reproducibility in RNA sequencing (e.g. Jiang *et al.* 2011) and ChIP-sequencing (e.g. Cheung *et al.* 2011) might be a useful approach to help control for complex biases and changing technologies.

Relevance to quantitative DNA diet studies

High-throughput sequencing has only been applied in a small number of DNA-based diet studies (reviewed in Pompanon *et al.* 2012), but these have generated considerable interest. In studies carried out to date, it is common for data to be generated from a single sequencing run and analysed with a static set of bioinformatic parameters (e.g. Deagle *et al.* 2009). These types of data overviews provide a misleading view of the precision of sequence proportions. While quantitative interpretations of sequence counts are often not discussed in detail, presentation of counts, or sequence proportions in graphs, implies some quantitative signature (e.g. Deagle *et al.* 2009; Soininen *et al.* 2009; Kowalczyk *et al.* 2011; Brown *et al.* 2012). In practical terms, the effect of any potential sequence recovery biases on overall diet estimates (based on many samples) will be dependent on the composition of wild-collected samples. If animals feed sequentially on different food items and most scats contain only a single dominant diet item, then biases will not be critical. However, if a mixture of food species is found in each scat (as in the current artificial feeding regime), then biases will be directly reflected in the final data set (Deagle & Tollit 2007). An alternative to quantifying sequence proportions that has been used by some high-throughput sequencing diet studies is to focus on frequency of occurrence data summaries to obtain an overall quantitative picture (e.g. Valentini *et al.* 2009a; Razgour *et al.* 2011; Shehzad *et al.* 2012). It is clear that inferring quantitative information from presence/absence data can have a number of problems (e.g. minor food items eaten frequently will appear to be an important part of the diet; see Laake *et al.* 2002). In addition, these presence/absence measures of animal diet are also likely to be influenced by stringency of quality filtering and other bioinformatic parameters affecting read number retained in the final data set. The low-level contamination between runs and mis-assignment of sequences between samples, observed in the current data sets, would drastically affect presence/absence data summaries. Given the already demanding requirement to avoid contamination during PCR in amplicon sequencing studies (see Pompanon *et al.* 2012), this additional source of potential contamination is particularly unwelcome.

Despite many technical factors influencing relative proportions of amplicon sequences recovered in the current study, the fact that for a given set of parameters, we observed consistent sequence proportions from scats of animals fed a constant diet is encouraging. The replicate PCR amplifications analysed in the second sequencing run produced very consistent results when there was no quality filtering. These results were also quite similar to the least-filtered data set from our first run (90-bp amplicons and no quality filtering). The consensus view across the two runs and both sequencing directions is that, in read count data, capelin was underrepresented (10–20% vs. 48.5% in diet), herring was overrepresented (65–75% vs. 34% in diet), and mackerel was quite close (10–20% vs. 17.5% in diet). The reason for the discrepancy between the diet and the proportion of recovered sequences is not clear based on data sets in the current study. It is possible that the observed bias is caused by differential PCR amplification, differences in DNA density of the fish species (i.e. herring may have more copies of mtDNA per gram of tissue than capelin), or there could be differential survival of the fish's DNA during digestion. If the biases are caused by either of the first two factors, it is possible that parallel analysis of fish tissue mixtures could allow species-specific correction factors to be developed for relatively simple systems – this is a possibility we are investigating further.

Conclusions

Due to the enormous amounts of data that can be generated by high-throughput sequencing of PCR amplicons, it is clear that this approach will be widely adopted to characterize mixed-species DNA samples. Our detailed analysis of three target species in a simple DNA mixture highlights that parameters in bioinformatic pipelines used to produce summaries of a data set can drastically affect proportions of sequences that are recovered. In our case, less stringent data filtering (based on quality scores or read length) produced more consistent results; however, other data sets may show a different trend, and retention of low quality sequences could have other consequences for field-based studies (e.g. species misclassification, or diversity overestimation). Therefore, it would be prudent for researchers to examine the impact on their own data rather than simply limiting filtering. Potential biases introduced by primer tags used to identify samples should also be considered in experimental design, both to allow for their detection and to reduce impacts. Finally, it would be useful to employ taxon-specific standards of known proportions in sequencing runs to begin systematically monitoring and accounting for taxon-specific biases. The issues that we have highlighted may be smaller than other well-documented forms of bias, such as impact of variation in PCR primer binding sites. This is particularly true for more complex

environmental samples where hundreds of diverse taxa may be simultaneously targeted. In these types of samples, further biases may also be introduced in extra bioinformatic processing steps that may be required (e.g. during more complex taxonomic assignment methods or during removal of chimeric sequences). With the high level of interest in environmental DNA barcoding shown by the molecular ecology community, we expect that high-throughput amplicon sequence data sets will be under increasing scrutiny, and as technologies stabilize, more accurate quantitative studies will be possible.

Acknowledgements

We thank the Point Defiance Zoo and Aquarium and their staff (Lisa Triggs, Chris Harris, Cindy Roberts and Amanda Chomos) and volunteers for conducting feeding trials. We also thank Andrea Polanowski and Cassy Faux for helpful laboratory guidance. Funding was provided by the Pacific Salmon Foundation and Australian Antarctic Science Program (Project 4014). All animal-related activities were undertaken in accordance with UBC Animal Care Committee guidelines. Thanks to four incredibly thorough reviewers and the handling editor for thoughtful comments.

References

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, **71**, 8966–8969.

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

Andersen K, Bird KL, Rasmussen M *et al.* (2012) Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Berry D, Mahfoudh KB, 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 and Evolution*, **27**, 233–243.

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.

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.

Buée M, Reich M, Murat C *et al.* (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist*, **184**, 449–456.

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

Cheung MS, Down TA, Latorre I, Ahringer J (2011) Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research* **39**, e72.

Deagle BE, Tollit DJ (2007) Quantitative analysis of prey DNA in pinniped faeces: potential to estimate diet composition? *Conservation Genetics*, **8**, 743–747.

Deagle B, Tollit D, Jarman S *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.

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* **6**, e17497.

Jiang LC, Schlesinger F, Davis CA *et al.* (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, **21**, 1543–1551.

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

Laake JL, Browne P, DeLong RL, Huber HR (2002) Pinniped diet composition: a comparison of estimation models. *Fishery Bulletin*, **100**, 434–447.

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research*, **35**, e97.

Morgan M, Anders S, Lawrence M *et al.* (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, **25**, 2607–2608.

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.

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

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

Porazinska DL, Giblin-Davis RM, Faller L *et al.* (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Molecular Ecology Resources*, **9**, 1439–1450.

Quail MA, Smith M, Coupland P *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

R Development Core Team (2010) *R: A Language and Environment For Statistical Computing*. R foundation for statistical computing, Vienna, Austria.

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

Rothberg JM, Hinz W, Rearick TM *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. *Human Molecular Genetics*, **19**, R227–R240.

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

Sipos R, Székely AJ, Palatinszky M *et al.* (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiology Ecology*, **60**, 341–350.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences, USA*, **103**, 12115–12120.

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.

Tollit DJ, Schulze AD, Trites AW *et al.* (2009) Development and application of DNA techniques for validating and improving pinniped diet estimates. *Ecological Applications*, **19**, 889–905.

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

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

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*, e12.

Yergeau E, Lawrence JR, Sanschagrin S *et al.* (2012) Next-generation sequencing of microbial communities in the Athabasca River and its tributaries in relation to oil sands mining activities. *Applied and Environmental Microbiology*, **78**, 7626–7637.

The focus of the study changed throughout its course, and all authors contributed to aspects of study conception and design. A.C.T. and A.K.S. ran the feeding trial. B.E.D. and A.C.T. did the laboratory work, analysed the data and wrote the article. S.N.J. and A.W.T. contributed to the manuscript.

Data Accessibility

R code and sequence data (FASTQ): Dryad doi:10.5061/dryad.953hv.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Summary plots of data from the Run I – 90 bp and Run II data sets.

Table S1 Captive seal feeding records used to calculate a combined mean diet.

Table S2 Amplicon sequences showing binding sites of PCR primers and the blocking oligonucleotide.