MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 819–830 doi: 10.1111/1755-0998.12355

Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods

J. PIÑOL,*† G. MIR,‡§ P. GOMEZ-POLO¶ and N. AGUSTÍ¶
*Universitat Autònoma Barcelona, Cerdanyola del Vallès, 08193, Spain, †CREAF, Cerdanyola del Vallès, 08193, Spain, ‡CRAG CSIC IRTA UAB UB, Cerdanyola del Vallès 08193, Spain, §The Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia, ¶IRTA, Carretera de Cabrils, Km. 2, E-08348 Cabrils, Barcelona, Spain

Abstract

The quantification of the biological diversity in environmental samples using high-throughput DNA sequencing is hindered by the PCR bias caused by variable primer–template mismatches of the individual species. In some dietary studies, there is the added problem that samples are enriched with predator DNA, so often a predator-specific blocking oligonucleotide is used to alleviate the problem. However, specific blocking oligonucleotides could coblock nontarget species to some degree. Here, we accurately estimate the extent of the PCR biases induced by universal and blocking primers on a mock community prepared with DNA of twelve species of terrestrial arthropods. We also compare universal and blocking primer biases with those induced by variable annealing temperature and number of PCR cycles. The results show that reads of all species were recovered after PCR enrichment at our control conditions (no blocking oligonucleotide, 45 °C annealing temperature and 40 cycles) and high-throughput sequencing. They also show that the four factors considered biased the final proportions of the species to some degree. Among these factors, the number of primer–template mismatches of each species had a disproportionate effect (up to five orders of magnitude) on the amplification efficiency. In particular, the number of primer–template mismatches explained most of the variation (~3/4) in the amplification efficiency of the species. The effect of blocking oligonucleotide concentration on nontarget species relative abundance was also significant, but less important (below one order of magnitude). Considering the results reported here, the quantitative potential of the technique is limited, and only qualitative results (the species list) are reliable, at least when targeting the barcoding COI region.

Keywords: annealing temperature, insects, Ion Torrent, PCR bias, PCR cycles

Received 13 May 2014; revision received 17 November 2014; accepted 21 November 2014

Introduction

High-throughput DNA sequencing techniques are being increasingly applied to the analysis of biological diversity in environmental samples via the identification of DNA sequences from heterogeneous mixtures (Steele & Pires 2011; Taberlet *et al.* 2012; Carew *et al.* 2013). These techniques were originally applied to microbial communities (e.g. Sogin *et al.* 2006), but they are now common in eukaryote organisms as well: fungi (Fouts *et al.* 2012; Xu *et al.* 2012), invertebrates (Porazinska *et al.* 2009), plants (Hiiesalu *et al.* 2012) and in studies of the diet of herbivores (Kowalczyk *et al.* 2011; Raye *et al.* 2011) and predators (Deagle *et al.* 2009; Bohmann *et al.* 2011; Brown *et al.* 2012). These studies are based on the use of the PCR to amplify DNA of a certain genomic region, followed by high-throughput sequencing (Quail *et al.* 2012).

One acknowledged shortcoming of this methodology is that it only provides a qualitative list of species present in a sample, but not their relative abundances. Restricting the analysis to a certain genomic region for which there is ample information in the genetic databases requires amplifying the DNA of the region of interest, which is performed via PCR with universal primers for the taxonomic group of interest. Unfortunately, PCR generates a well-known bias (Wintzingerode *et al.* 1997; Polz & Cavanaugh 1998; Pinto & Raskin 2012), which is related to primer–template mismatches and to purely stochastic effects. As a result, some species increase their relative abundance, others decrease it, while others are not amplified and so remain absent in the final sequenced mixture. However, this well-known PCR bias has rarely been quantified using heterogeneous samples

Correspondence: Josep Piñol, Fax: +34-93-5814151; E-mail: Josep.Pinol@uab.es

of known composition followed by high-throughput sequencing of the PCR product. Only a few tests have been carried out so far in eukaryotes, based on mock communities in which the identity and the approximate relative abundance of the species are known (Porazinska *et al.* 2009; Amend *et al.* 2010; Deagle *et al.* 2013; Egge *et al.* 2013; Pochon *et al.* 2013).

In dietary studies, there is the added problem that samples (faeces or gut contents) are often enriched with predator DNA, so the result of the PCR amplification could be dominated by predator DNA rather than by prey DNA, especially if prey and predator are phylogenetically close (*e.g.* spiders eating spiders). Sequencing this PCR product produces a great deal of noninformative predator reads, reducing the potential of high-throughput sequencing. To overcome this problem, the so-called predator-specific blocking oligonucleotides have been used to reduce amplification of the predator DNA (Vestheim & Jarman 2008; Deagle *et al.* 2009, 2010, 2013; Shehzad *et al.* 2012). However, specific blocking oligonucleotides could block prey DNA to some degree, along with the targeted predator (Vestheim & Jarman 2008). Thus, the use of predator-specific blocking oligonucleotides may introduce an additional bias in the analysis of DNA from heterogeneous mixtures. To our knowledge, the effect of the blocking oligonucleotide concentration on the amplification efficiency of nontarget species in the mixture has never been quantified.

The main objective of this study was to measure with a structured experimental design the PCR bias induced by universal and blocking primers in a heterogeneous sample. For this purpose, we prepared a solution with the same concentration of DNA of several species of terrestrial arthropods. Using a pair of universal primers targeting the mitochondrial COI barcoding region, we PCR-amplified this mock community at different concentrations of a specific blocking oligonucleotide of one of the species in the mixture. We then sequenced all the PCR products with a high-throughput sequencing platform and counted the number of reads of each of the species in the mixture. Under the null hypothesis of no PCR bias, we would expect a proportion of $1/n$ reads for each of the n species in the mixture, although we anticipated otherwise. In particular, although we mixed equal amounts of total DNA, we did not account for differences in genome size and mtDNA copies across species.

As a secondary objective, using the same structured experimental design, we wanted to compare the effects of the annealing temperature and the number of PCR cycles on the amplification efficiency in relation to the biases induced by universal and blocking primer mismatches. We are aware that the effects of annealing temperature (the higher this is, the more specific the reaction) and the number of PCR cycles (the higher this is, the more divergent the final product) are well known (McPherson & Moller 2006). The key point is the quantitative comparison of the universal primer, blocking oligonucleotide, annealing temperature and PCR cycles effects in a single experiment to determine which factors most affect the amplification efficiency (Sipos *et al.* 2007).

Methods

One adult specimen of each of the following species was collected in two different localities of northeast Spain: La Selva del Camp (Tarragona) (*Theridion* sp. (Araneae: Theridiidae), *Philodromus cespitum* (Walckenaer) (Araneae: Philodromidae), *Phyllocnistis citrella* (Stainton) (Lepidoptera Gracillaridae), *Adalia decempunctata* (L.) (Coleoptera: Coccinellidae), *Aphis spiraecola* Patch (Hemiptera: Aphididae) and *Trichopsocus* sp. (Psocoptera: Trichopsocidae); and El Maresme (Barcelona) (*Orius majusculus* (Reuter) and *Orius laevigatus* (Fieber) (Hemiptera: Anthocoridae), *Episyrphus balteatus* (De Geer) (Diptera: Syrphidae), *Macrolophus melanotoma* (Costa) (Hemiptera: Miridae) and *Entomobrya* sp. (Collembola: Entomobryidae). In addition, the parasitoid *Aphidius colemani* (Dalman) (Hymenoptera: Braconidae) was provided by Biobest N.V. (Westerlo, Belgium).

DNA was extracted using the DNeasy Tissue Kit (QIAGEN; Hilden, Germany; protocol for animal tissues). Total DNA was eluted into 100 μL of elution buffer and stored at −20° C. A conventional PCR was run (using control PCR conditions shown in Table 1) with the ZBJ-ArtF1c and ZBJ-ArtR2c primers (Zeale *et al.* 2011) on DNA of each species and the products ran on agarose gels; all the selected species showed a band at the expected size. This primer set is known to amplify a wide range of insect orders (Ephemeroptera, Odonata, Plecoptera, Orthoptera, Dermaptera, Hemiptera, Neuroptera, Lepidoptera, Trichoptera, Diptera, Hymenoptera, Coleoptera, Blattodea), spiders and springtails (Zeale *et al.* 2011; Clarke *et al.* 2014; Piñol *et al.* 2014).

Sanger sequencing and COI sequences

To identify the sequence reads finally obtained with the universal primers used with the high-throughput DNA sequencing platform (see below), it was necessary to know the COI sequences (barcode) of all the species included in the mixture. To do so, each individual DNA extract was amplified using a 2720 thermal cycler (Applied Biosystems, CA, USA). PCR volumes (20 μL) contained 1.2 μL of template DNA, 18 μL of Platinum PCR Supermix High Fidelity (Invitrogen, CA, USA) and 0.4 μL (10 μM) of the primers LCO1490 (5-GGTCAA CAAATCATAAAGATATTGG-3′) and HCO2198 (5′-TA AACTTCAGGGTGACCAAAAAATCA-3′) (Folmer *et al.*

Table 1 Variable experimental conditions of the two ION TORRENT runs. The first chip was devoted to analysing the effect of the concentration of blocking oligonucleotide and the second to the effects of annealing temperature and number of cycles. The concentration of the blocking primer is indicated in proportion to the concentration of universal primer (i.e. a value of αx indicates that the concentration of blocking primer is α times the concentration of the universal primer). Please notice that there is one PCR condition (no blocking primer, 45 °C and 40 cycles; marked in bold) that appears twice, once in each chip. This condition was used as a reference or control condition

Chip	*Orius majusculus* blocking oligonucleotide concentration	Annealing temperature (°C)	Number of PCR cycles
1	**0x**	**45**	**40**
1	1x	45	40
1	2x	45	40
1	5x	45	40
1	10x	45	40
1	20x	45	40
1	40x	45	40
2	0x	40	40
2	**0x**	**45**	**40**
2	0x	50	40
2	0x	55	40
2	0x	45	25
2	0x	45	30
2	0x	45	35

1994). Samples were amplified for 35 cycles at 94° C for 30 s; 45° C for 45 s; and 68° C for 90 s. A denaturation cycle of 94° C for 10 s started the PCR, and the reaction was terminated with a final extension at 68° C for 10 min. Target DNA and water were included as positive and negative controls, respectively. PCR products were purified using the QIA-quick PCR Purification Kit (Qiagen) and sequenced in both directions on an ABI 3730 automated sequencer (Applied Biosystems) using the Big Dye version 3 sequencing kit. The chromatograms were inspected and edited using GENEIOUS version 6.1.7 created by Biomatters (available from http://www.geneious.com/). Only eleven of twelve samples (all except *Entomobrya* sp.) were successfully sequenced.

We counted the number of primer–template mismatches between the Zeale *et al.* (2011) universal primers and the Sanger sequences of the eleven species. The results are given separately for the forward and reverse primers and for the three 3′ terminal bases of each primer and for the entire primer. However, it has only been possible to analyse 21 bases of the forward primer because the first nine overlap the Folmer's primer (Folmer *et al.* 1994) used to amplify the barcoding region of these species.

Mock sample preparation and high-throughput DNA sequencing

The amount of DNA in each of the above twelve extracts described in the previous section was quantified using Qubit® 2.0 Fluorometer (dsDNA HS Assay, Life Technologies) and PicoGreen (Quant-iT™ PicoGeen® dsDNA Assay Kit, Life Technologies). After that, a mock sample was prepared for the high-throughput sequencing using 5 ng of each individual extraction.

We prepared the samples for high-throughput DNA sequencing with fusion primers, following the ION TORRENT recommendations for unidirectional sequencing. Briefly, one pair of primers was designed with the ION TORRENT primer A linked to the specific forward primer and the ION TORRENT primer trP1 linked to the specific reverse primer. Here, the specific primer set refers to the primers ZBJ-ArtF1c (5-AGATATTGGAACWTTATATTTTATTTTTGG-3′) and ZBJ-ArtR2c (5-WACTAATCAATTWCCAAATCCTCC-3′) (Zeale *et al.* 2011), which amplify a 157-bp COI fragment. To test the effect of different PCR conditions on our mock sample, we used seven different indexed versions of the forward fusion primer to allow multiplexing of up to seven samples in an individual sequencing run. All PCR products were purified using the QIA-quick PCR Purification Kit (Qiagen) and checked separately (DNA 1000 kit, Bioanalyzer 2100, Agilent Technologies). We selected fragments of the expected size (E-Gel® Size Select 2% Agarose Gel, Life Technologies), quantified the size-selected DNA (DNA High Sensitivity kit, Bioanalyzer 2100, Agilent Technologies) and prepared an equimolar pool of the seven PCR products. We then diluted, amplified (emulsion PCR, Ion OneTouch™ 200 Template Kit v2 DL, Ion Torrent, Life Technologies) and sequenced this pool on an ION TORRENT PGM platform, as described by the manufacturer (Ion Torrent, Life Technologies). We used two ION TORRENT 314 chips, with the sequencing chemistry for 300 bp read length and the version 2.2 of the TORRENT SUITE software for base calling.

First chip – effect of the blocking oligonucleotide concentration

We used the blocking oligonucleotide BloOm2 (5-TATATTTTATTTTTGGGATATGAGCAGGAATAC-3′-C3) developed to block the amplification of *O. majusculus* with the Zeale primers (P. Gómez-Polo & N. Agustí, unpublished results). The C3 spacer at the 3′ end of the oligonucleotide inhibits the elongation by the Taq polymerase. This blocking oligonucleotide competes with the universal primer for the priming sites of the DNA template, so the final number of reads of *O. majusculus* depends on the ratio of blocking to universal primer concentrations. We

tested the effect of the blocking oligonucleotide at seven different concentrations: 0x (control; no blocking oligonucleotide), 1x, 2x, 5x, 10x, 20x and 40x (where αx indicates that the concentration of blocking oligonucleotide is α times the concentration of the universal primer; Table 1, chip one). In all cases, the DNA was amplified in 50 μL reaction volumes containing 3 μL of template DNA, 45 μL of Platinum PCR Supermix High Fidelity (Invitrogen Corporation) and 1 μL of each pair of 10 μM fusion primers. Each of the 40 PCR cycles consisted of 94 °C for 30 s, 45 °C for 45 s and 68 °C for 45 s following an initial denaturation step at 94 °C for 5 min and before a final extension step at 68 °C for 10 min.

Second chip – effects of the annealing temperature and the number of PCR cycles

In a second ION TORRENT 314 chip, we tested the effects of the annealing temperature and the number of PCR cycles on the amplification efficiency of the DNA of the species in the mixture. We tested four different annealing temperatures (40, 45, 50 and 55 °C) and four different number of PCR cycles (25, 30, 35 and 40 cycles). The remaining conditions were the same as in the previous experiment, with no blocking oligonucleotide (Table 1, chip two).

Processing and analysis of data

The output of the ION TORRENT was treated bioinformatically to ascertain how many reads of each species were obtained in each experimental condition. The ION TORRENT software itself separated into individual FASTQ files the reads corresponding to each experimental condition (Table 1), taking advantage of the index included in the forward fusion primers. All subsequent analyses were carried out in parallel with the 14 files corresponding to each experimental situation.

First of all, we eliminated the primer sequence from the 5′ end of each read using TagCleaner (Schmieder *et al.* 2011). We then trimmed (truncated to 150 bp at the 3′ end), discarded those with a mean quality score lower than 25 and downloaded the sequences in FASTA format (all using PRINSEQ, Schmieder & Edwards 2011). Each FASTA file was locally blasted (blastn; Zhang *et al.* 2000) against a database containing the sequences of the COI region of the eleven species for which we obtained their DNA barcode. We imported the output from the BLAST alignment into MEGAN version 4.70.4 (MEtaGenomics ANalyzer; Huson *et al.* 2011) to count the number of reads of each species analysed. The selected MEGAN parameters were as follows: Min Support = 1; Min Score = 100.0; Top Percentage = 1.0; Win Score = 0: Min Complexity = 0.44.

As we did not obtain a good Sanger sequence of the COI region for the collembolan (*Entomobrya* sp.), we tried to identify any sequences in the output files that could belong to this species using the following method. First, we downloaded as FASTA files the sequences with no hits from MEGAN. Then, we created a database containing all quality sequences of Entomobryidae from GenBank (188 sequences; search: Entomobryidae [Organism] AND COI [Gene Name] AND Barcode[Keyword]; accessed on 30 July 2014). Finally, we locally blasted the files containing the 'No hit' sequences against this Entomobryidae database and counted the number of sequences that produced a hit above 80% of similarity. As the blasted reads did not match any of the other eleven species, we assumed that all the hits at this low level of similarity belonged to the collembolan that we introduced in the mock sample.

Regression and correlation analyses were conducted with respect to the number of primer mismatches with the number of reads of each species using R version 2.15 (R Core Team 2012). The effect of blocking oligonucleotide concentration on the relative abundance of nontarget species was tested with the same software package with a general linear model with a binomial error structure and logit link.

Results

All species except *Entomobrya* sp. were amplified using LCO1490/HCO2198 primers. The obtained Sanger sequences were submitted to GenBank (Accession nos shown in Table 2).

High-throughput DNA sequencing

The first chip produced 108158 ± 3123 (mean ± SE) raw reads per experimental condition; 71.8 ± 1.1% of them passed the quality control and 88.9 ± 1.4% of them produced a hit to one of the species included in the mixture. In summary, we obtained 69082 ± 2144 reads per experimental condition, which allowed us to test the effect of the blocking oligonucleotide concentration on the amplification efficiency of the individual species.

The second chip produced 72775 ± 8179 raw reads per experimental condition; 58.2 ± 2.2% of them passed the quality control and 95.7 ± 1.0% of them produced a hit to one of the species of the mixture. In summary, we obtained 40512 ± 3690 reads per experimental condition to test the effect of the annealing temperature and of the number of PCR cycles on the amplification efficiency of the individual species (see Table S1, Supporting information for details).

Table 2 Number of template–primer mismatches between the Zeale *et al.* (2011) universal primers and the sequences of the eleven studied species. The results are given separately for the forward and reverse primers and for their sum and for the three 3′ terminal bases (3 bases) of the primer and for the entire primer length (entire; this value includes the number of mismatches in the 3 terminal bases). The alignments are given fully in Table S1 (Supporting information). The GenBank Accession no. corresponds to the Sanger sequence obtained in the present study

Order	Family	Species	GenBank Accession no.	Forward primer		Reverse primer		Both primers	
				3 bases	Entire	3 bases	Entire	3 bases	Entire
Araneae	Theridiidae	*Theridion* sp.	KJ467508	1	6	1	2	2	8
Araneae	Philodromidae	*Philodromus cespitum*	KJ467506	0	1	1	2	1	3
Lepidoptera	Gracillaridae	*Phyllocnistis citrella*	KJ467507	0	2	0	0	0	2
Coleoptera	Coccinellidae	*Adalia decempunctata*	KJ467504	0	2	0	0	0	2
Diptera	Syrphidae	*Episyrphus balteatus*	KJ467502	1	4	0	0	1	4
Hymenoptera	Braconidae	*Aphidius colemani*	KJ467503	0	1	0	1	0	2
Hemiptera	Anthocoridae	*Orius majusculus*	KJ467501	0	1	0	1	0	2
Hemiptera	Anthocoridae	*Orius laevigatus*	KJ467500	0	1	0	0	0	1
Hemiptera	Miridae	*Macrolophus melanotoma*	KJ467499	1	3	1	3	2	6
Hemiptera	Aphididae	*Aphis spiraecola*	KJ467505	0	2	1	2	1	4
Psocoptera	Trichopsocidae	*Trichopsocus clarus*	KJ467509	0	0	0	1	0	1

Differential amplification of the species

If the template concentration of the species in the mixture (n = 12) was the same and they were amplified with the same efficiency, the expected proportion of reads should be approximately 1/12 (~0.08). Using the two control runs (no blocking oligonucleotide, annealing temperature of 45 °C and 40 PCR cycles; Table 1), results showed that only six species had a frequency close to the expected value (within one order of magnitude): *P. citrella*, *A. decempunctata*, *A. colemani*, *O. majusculus*, *O. laevigatus* and *T. clarus* (Fig. 1). Five species amplified with an efficiency 2–3 orders of magnitude lower: *Philodromus* sp., *E. balteatus*, *M. melanotoma*, *A. spiraecola* and *Entomobrya* sp. Finally, the spider *Theridion* sp. amplified with an even lower efficiency.

The number of template–primer mismatches explained quite well the differential amplification of the species (the comparison was limited to eleven species, as we did not obtain the Sanger sequence of *Entomobrya* sp.). The six species with stronger amplification had no mismatches in any of the three bases at the 3′ end of both the forward and reverse primers (Fig. 2b). Three of the four species with an intermediate amplification efficiency had one mismatch, and two species (*M. melanotoma* and *Theridion*) had two mismatches in these terminal regions (one in the forward primer and one in the reverse). If the entire sequence of the primers is considered (Fig. 2a), the species with the higher amplification efficiency had one or two mismatches; those with intermediate amplification efficiency had three to six mismatches (*M. melanotoma* had four mismatches), and *Theridion* had eight mismatches (Table S2, Supporting information). In summary, the total number of mismatches between both primers (forward + reverse; Table 2) and the sequence of each species explained 73% of the variation of the proportion of reads (Fig. 2a). The number of mismatches between the 3′ end (three bases) of both primers (forward + reverse; Table 2) and the DNA sequence of each species performed even better, as it explained 81% of the variation of the proportion of reads (Fig. 2b).

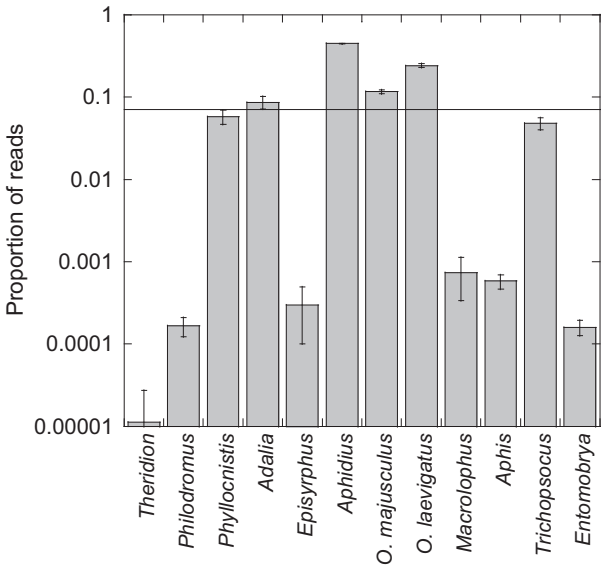

Fig. 1 Proportion of reads of the twelve species present in the mixture in two control ION TORRENT runs. Vertical bars indicate the standard error. PCR conditions: no blocking primer, annealing temperature 45 °C, 40 cycles. The horizontal line indicates the expected proportion of reads (=1/12) if all species are amplified with the same efficiency.

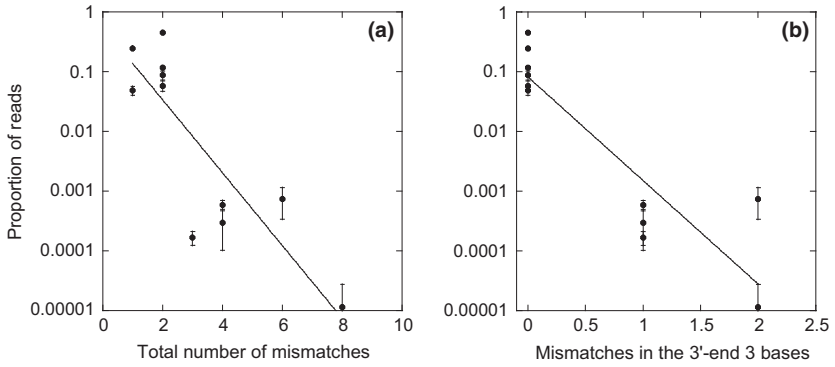

Fig. 2 (a) Relationship between the proportion of reads in an ION TORRENT run and the total number of mismatches between the two primers and the corresponding region of each species (equation: $\log_{10}(y) = -0.25\ -0.61\ x$, $r^2 = 0.73$, $F_{1,9} = 24.35$, $P = 0.0008$) (b) The same relationship but only counting the number of mismatches between the last 3 bases of the 3′ end of the primers (considering both the forward and reverse primers) and the priming site of the species (equation: $\log_{10}(y) = -1.09\ -1.73\ x$, $r^2 = 0.81$, $F_{1,9} = 38.24$, $P = 0.0002$). The priming sites of each species were obtained by Sanger sequencing of the COI region. Each proportion is the mean value of two independent PCRs and sequencing runs; vertical bars indicate the SE.

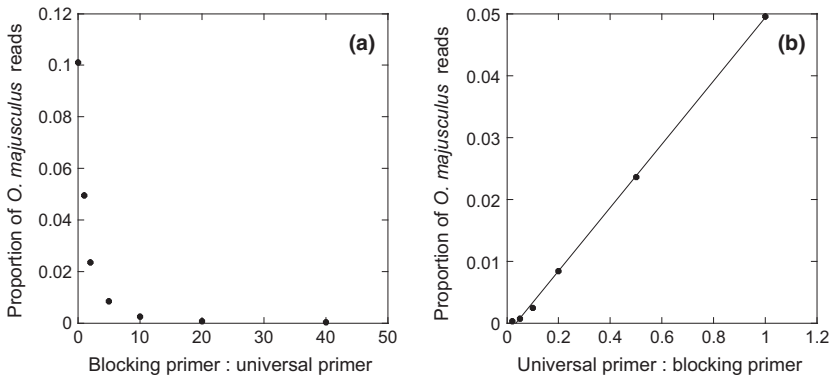

Fig. 3 (a) Effect of the concentration of the blocking primer on the proportion of the *Orius majusculus* reads in an ION TORRENT run. The concentration of blocking primer is reported as the proportion of the concentration of *O. majusculus* blocking primer over the concentration of the universal primer (i.e. 10x means that the concentration of blocking primers was ten times higher than the concentration of the universal primer). (b) The same plot, but using the inverse of the value of the A plot in the *x*-axis. In this diagram, the value with no blocking primer (0x) is not represented because it is not possible to compute the inverse of zero. The values follow very closely a linear trend ($r^2 = 0.999$, $P < 0.001$), suggesting that the blocking primer competes for the active site of the polymerase (the priming site), a characteristic of the enzymatic allosteric inhibition. Other PCR conditions are described in Table 1.

Effect of the blocking oligonucleotide concentration

As expected, the proportion of *O. majusculus* reads decreased with the concentration of the blocking oligonucleotide, which was designed for this target species (Fig. 3a). At the highest concentration of blocking oligonucleotide (40x), some reads of *O. majusculus* were still found (0.03%, i.e. a reduction of ca. 300 times compared with nonblocking oligonucleotide). The relationship between blocking oligonucleotide concentration and the proportion of reads became linear after the inverse transformation of the concentration (Fig. 3b).

With the exception of *M. melanotoma*, absent at the highest blocking oligonucleotide concentration (40x), all the species were detected irrespectively of the blocking oligonucleotide concentration. However, the presence of the blocking oligonucleotide modified the proportion of reads of most nontarget species (Fig. 4). *Aphidius colemani*, *M. melatonoma* and *O. laevigatus* were significantly coblocked by the *O. majusculus* blocking oligonucleotide, so they decreased their abundance at high concentrations of the blocking oligonucleotide (Table 3). On the contrary, *P. citrella*, *A. decempunctata*, *E. balteatus*, *A. spiraecola* and *T. clarus* significantly increased their relative abundance with increasing concentrations of blocking oligonucleotide. The effect was particularly remarkable for *P. citrella* (3.4 times increase from 0x to 40x), *A. decempunctata* (4.0 times increase),

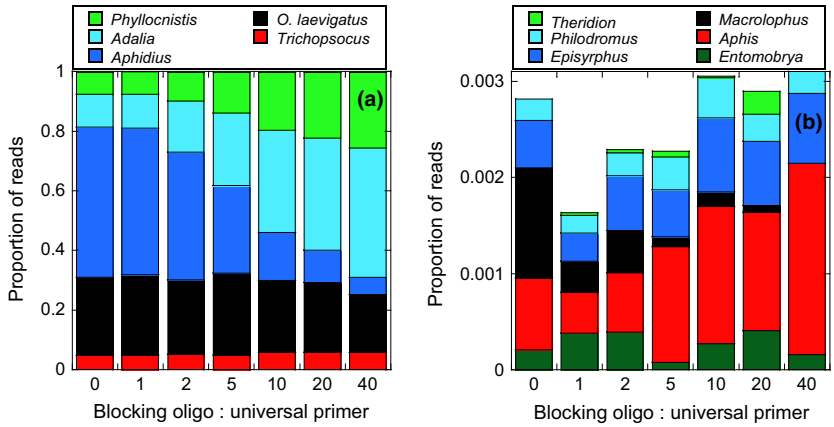

Fig. 4 The effect of the blocking primer concentration on the proportion of sequences obtained in an ION TORRENT run with all nontarget species (all except *Orius majusculus*). (a) Well-amplified species and (b) poorly amplified species.

Table 3 Summary of the general linear model for the effect of the blocking oligonucleotide concentration on the proportion of reads of nontarget species (expressed as the proportion blocking oligonucleotide: universal primer). The table contains the estimate of the coefficient of the model, its standard error, the statistic *z* and its statistical significance (*P*-value). If a conservative table-wide Bonferroni correction is applied, *P*-values below (0.05/11 = 0.0045) indicate a statistical significant effect of the blocking oligonucleotide concentration on the proportion of reads of that species (in **bold**)

Species	Coefficient	St error	*z*-value	*P*-value
Macrolophus melanotoma	**−0.173**	**0.026**	**−6.63**	**$3.4 \cdot 10^{-11}$**
Aphidius colemani	**−0.090**	**0.0005**	**−193**	**$<2 \cdot 10^{-16}$**
Entomobrya sp.	−0.007	0.007	−1.03	0.30
Orius laevigatus	**−0.001**	**0.0003**	**−36.9**	**$<2 \cdot 10^{-16}$**
Philodromus cespitum	0.003	0.006	0.46	0.65
Trichopsocus clarus	**0.005**	**0.0005**	**10.4**	**$<2 \cdot 10^{-16}$**
Episyrphus balteatus	**0.013**	**0.004**	**2.99**	**0.003**
Aphis spiraecola	**0.025**	**0.003**	**8.87**	**$<2 \cdot 10^{-16}$**
Theridion sp.	0.028	0.012	2.39	0.02
Phyllocnistis citrella	**0.031**	**0.0003**	**114**	**$<2 \cdot 10^{-16}$**
Adalia decempunctata	**0.039**	**0.0002**	**162**	**$<2 \cdot 10^{-16}$**

A. colemani (8.7 times decrease) and *M. melatonoma* (0 reads at 40x). Finally, *Theridion* sp., *Philodromus* sp. and *Entomobrya* sp. relative abundance were not significantly related to the blocking oligonucleotide concentration.

The explanatory power of the mismatches between the blocking oligonucleotide and the priming site of the ten nontarget species was weak. The two species more similar to the blocking primer of *O. majusculus* (*A. colemani* and *M. melanotoma*; four mismatches each; Table 4) were also those more negatively affected by the presence of the blocking oligonucleotide (Fig. 4). The third most similar species (*O. laevigatus*; five mismatches) was also significantly coblocked, albeit with a very low effect (Table 3). It is worth noticing that all ten nontarget species had a mismatch at the residue just before the C-3 spacer, but only in the three species significantly coblocked, that residue was a 'T' (Table 4). However, the overall regression between the proportion of reads at 40x to reads at 0x (a measure of the effect of the blocking oligonucleotide) against the number of mismatches between the blocking oligonucleotide and the priming site of each species was nonsignificant ($r^2 = 0.24$; $F_{1,8} = 2.50$, $P = 0.15$).

Effect of the annealing temperature

As expected, the main effect of increasing the annealing temperature in the PCR was to increase the specificity of the universal primer. Consequently, the number of species amplified at 40 and 45 °C was higher (twelve species) than at 50 and 55 °C (seven species) (Fig. 5a,b). At 50 and 55 °C, there was not a single read of *Theridion* sp., *Philodromus* sp., *E. balteatus*, *M. melanotoma* or *Entomobrya* sp. *Aphis spiraecola* also reduced its amplification efficiency (Fig. 5b). It is worth remembering that these species had the most mismatches with the universal primer used here (Table S2, Supporting information). In the well-amplified species (Fig. 5a), *O. laevigatus* and

	1 10 20 30
Orius majusculus	TATATTTTATTTTTGGGATATGAGCAGGAATAC
Blocking oligo	TATATTTTATTTTTGGGATATGAGCAGGAATAC
Macrolophus melanotoma	TATATTTTATATTAGGGATATGGGCAGGAATAT
Aphidius colemani	TATATTTTATTTTTGGTATATGATCAGGTATAT
Orius laevigatus	TATATTTCATTTTTGGAATATGGGCAGGTATAT
Philodromus cespitum	TATATTTAATTTTTGGTGCTTGAGCAGCAATAG
Trichopsocus clarus	TATATTTTATTTTTGGGTTTTGAGCGGGGACTG
Episyrphus balteatus	TATATTTTCTATTCGGAGCTTGAGCTGGAATAG
Aphis spiraecola	TATATTTTTTATTTGGTATTTGATCAGGAATAA
Theridion sp.	TGTATTTAGTATTCGGTGCTTGAGCTGCTATAG
Phyllocnistis citrella	TATATTTTTTATTTGGAATTTGATCAGGAATAG
Adalia decempunctata	TATATTTTTTATTTGGCCTATGGGCAGGTATAG

Table 4 Alignment of the blocking primer BloOm2 of *Orius majusculus* with the ten other nontarget species of arthropods considered in this study. The mismatches between the blocking primer and the priming regions of the sequences of each species are highlighted

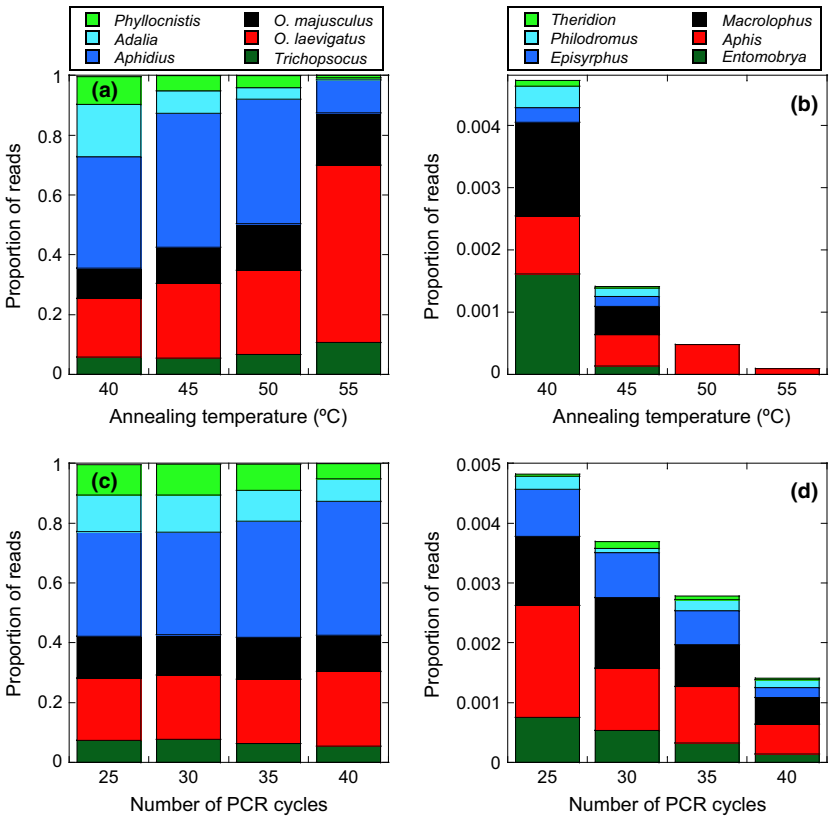

Fig. 5 The effect of the annealing temperature (Up) and number of PCR cycles (Down) on the proportion of reads in an ION TORRENT run. All 12 species are included: well-amplified species (a, c) and poorly amplified species (b, d). PCR conditions: no blocking primer, 40 cycles, annealing temperature as indicated (a, b); no blocking primer, PCR cycles as indicated, annealing temperature 45 °C (c, d).

T. clarus increased their proportion of reads at 55 °C. These two species matched the primer sequence most closely (only one mismatch).

Effect of the number of PCR cycles

In comparison with the effect of the annealing temperature, the effect of the number of cycles was not so drastic, as all the species were present under all the experimental conditions. There was a slight increase in the standard deviation of the percentage of reads of each species, from 11% at 25 and 30 cycles to 13% at 35 cycles and 14% at 40 cycles. The main consequence of increasing the number of PCR cycles was a reduction in the proportion of the less well-amplified species in the final mixture as the number of cycles increased (Fig. 5d).

Discussion

The four factors analysed in the present study (universal primer–template mismatches, blocking oligonucleotide concentration, annealing temperature and number of PCR cycles) had the expected effect on the final proportion of individual species in the mixture. However, the number of mismatches between the universal primers and the priming region of each species had a disproportionate effect (up to 5 orders of magnitude) on the amplification efficiency and, consequently, on the number of reads finally obtained.

Amplification efficiency and primer mismatches

All twelve species of arthropods in the mock community were detected with the control PCR conditions (Table 1), so the method provides excellent results from a qualitative point of view (Yu *et al.* 2012). However, the relative proportions of each species (all at the same initial concentration in the mock sample) were not reproduced at all: in the control conditions, the number of reads of six species comprised more than 99.5% of the total (thousands to tens of thousands reads for each species), five species were much less represented in the final mixture (tens of reads), and one species appeared only as a single read (Fig. 1). Considering these results, it is clear that the method provided a very poor quantitative description of the mock community. Similar results have been reported by other authors: Porazinska *et al.* (2009, 2010) using an artificial mixture of nematodes; Egge *et al.* (2013) using a mock community of eleven species of haptophytes; and Deagle *et al.* (2013) using captive seals fed with a diet of three fish species. The only two studies that reported a significant relationship between biomass and number of reads recovered are Hiiesalu *et al.* (2012) and Kelly *et al.* (2014), but in both experiments, the number of species was very low (two to five plant and four fish species, respectively). The most parsimonious explanation of the differential amplification of the different species in the mixture comes from the number of mismatches between the universal primer used and the priming site of each species. This single factor explains ~3/4 of the amplification efficiency of the individual species. The rest of the variation has to be caused by other factors not considered in the present study.

Among the unaccounted factors of variation, a variable mtDNA copy number and genome size across species deserve special attention. These two factors must affect the final results as we prepared the mock sample by standardizing by the total DNA concentration of the twelve species (which include both genomic and mitochondrial DNA). The Animal Genome Database (Gregory 2014) provides the genome size of 817 species corresponding to the arthropoda orders considered in this study (Table 1); the 90% confidence interval of the genome size for these 817 species was 0.16–2.75 (measured as the c-value). This is a ~17-fold variation on genome size across species, much lower than the amplification bias we describe here that spans five orders of magnitude (Fig. 1). The mtDNA copy number must also play a role, but we are not aware of any reliable data of this variable across arthropod species. Nevertheless, it is worth keeping in mind that ~3/4 of the variance was already explained by template–primer mismatches and that, in consequence, only ~1/4 of the variance was not accounted for (mtDNA copy number, genome size, etc.). Although no further considered here, it is worth mentioning that part of the unexplained variance might be caused by the type of nucleotide mismatch (Kwok *et al.* 1990), by the short primer tags used to identify individual samples (Deagle *et al.* 2013), and even by the quality filtering in the bioinformatic pipeline (Deagle *et al.* 2013).

The reported results could be the consequence of a poor performance of the selected universal primer (Zeale *et al.* 2011). Considering all eleven species and both the forward and reverse primers, there were 3.2 ± 0.7 mismatches (mean ± SE), with a range from 1 to 8 total mismatches (Table 1). Leray *et al.* (2013) recently proposed a new general primer set targeting a 313-bp region within the same COI region. This primer set had 1.8 ± 0.1 mismatches with the template of 99 species of arthropods (range 0 to 5 mismatches). In this respect, this primer set (or the ones proposed by Clarke *et al.* 2014) could be a better choice than the one used here, but, even so, there is a wide range of variation in the number of mismatches and, presumably, in the differential amplification of species in a heterogeneous sample. In contrast, another proposed general primer for the COI region in animals, the mini-barcode set (Meusnier *et al.* 2008), has many more mismatches with the priming region of the eleven species studied here (5.3 ± 0.5 mismatches, and this only in the reverse primer, as the forward one lies almost totally outside the standard barcoding region and could not be checked; Table S3, Supporting information). Clarke *et al.* (2014) reported a rather poor performance of the mini-barcode primer set in comparison with the one used here.

Effect of the blocking oligonucleotide concentration

The strong bias induced by the universal primer mismatches discussed above may be extended conceptually to blocking oligonucleotides as well (Piñol *et al.* 2014). Species with a more similar sequence to the targeted species (= to the designed blocking oligonucleotide) could eventually be also blocked during their PCR

amplification. Here, we found a significant coblocking effect on three of the eleven nontarget species tested, and probably as a consequence, five species increased their relative abundance with the blocking oligonucleotide concentration (Fig. 4; Table 3). When there was either a decrease or an increase in relative abundance, the effect was of a much lower magnitude than the effect observed with the universal primers: only one of the eleven nontarget species was totally coblocked at the highest blocking oligonucleotide concentration, and the general effect was within one order of magnitude range. Two of the coblocked species belonged to the same superfamily than the targeted species (Hemiptera: Cimicoidea), but the third Hemipteran tested was not coblocked at all (*A. spiraecola*) and the third coblocked species was an Hymenopteran (*A. colemani*).

The relationship between the number and type of mismatches of the blocking oligonucleotide and the priming sites of the species was not as straightforward to understand as in the case of the universal primer. The three coblocked species shared two common characteristics: the lowest number of mismatches with the blocking oligonucleotide and a 'T' residue at the 3′ position (Kwok *et al.* 1990). These two observations could be useful for the design of specific blocking oligonucleotides in the future.

Effects of annealing temperature and of the number of PCR cycles

Our results as regards the effects of the annealing temperature and the number of cycles on the relative proportion of the species in the final mixture confirmed our expectations (Sipos *et al.* 2007). A low annealing temperature reduced the specificity of the primer and increased the range of the species amplified in the PCR. To amplify as many species as possible, it is advisable to reduce the annealing temperature as much as possible without inducing artefacts by amplifying other genomic regions. The lowest temperature tested in the present study was 40 °C, which did not produce amplification artefacts. The effect of the number of cycles was less important compared to the effect of the annealing temperature. Increasing the number of PCR cycles decreased the number of reads of the worst amplified species in the mixture. So, provided enough DNA can be obtained to fulfil the requirements of the high-throughput sequencing laboratory, it is advisable to reduce the number of PCR cycles as much as possible (Polz & Cavanaugh 1998). This recommendation is especially important in environmental, more diverse samples with distinct species at very different concentrations, in contrast to our mock sample where all the species were at the same concentration.

Is it possible to provide quantitative results?

In our mock sample, the DNA of all the species was initially present at the same concentration (before PCR amplification), but the final results showed that the reads of some species were five orders of magnitude more abundant than others. Most of this bias was a direct consequence of the mismatch between the universal primer and the priming sites within the sequences of each species. It could be argued that this problem could be alleviated using 'truly' universal primers. For instance, Leray *et al.* (2013) and Clarke *et al.* (2014) have recently proposed new universal primers sets (targeting the mitochondrial COI and 16S, respectively) that could eventually work better than the ones used here. However, even these promising new primer sets have a variable number of mismatches across species: Leray *et al.* (2013) primer set has 0–5 mismatches in 99 species of arthropods, and some species could not be amplified at all; and Clarke *et al.* (2014) allowed 0–2 mismatches per primer in their design (i.e. 0–4 mismatches per primer pair), and also some species could not be amplified. Some authors maintain that the quest for the 'truly' universal primer set is futile and that no such thing exists for diversified groups of organisms (Forney *et al.* 2004). Even if a primer pair was able to amplify all arthropod species, that would not solve the problem of differential amplification unless the number of mismatches was the same for all the species in the mixture, and this is certainly unrealistic.

To be fair, explicit quantitative proportions of species using high-throughput DNA sequencing after PCR amplification with universal primers are rarely reported (Soininen *et al.* 2009). The discussion of many studies contains a paragraph advising against that 'temptation' (i.e. Valentini *et al.* 2009; Shehzad *et al.* 2012; Yu *et al.* 2012). However, many researchers, including ourselves (Piñol *et al.* 2014), have reported the numbers of reads of each individual species in the final mixture (Sogin *et al.* 2006; Hiiesalu *et al.* 2012), implicitly implying that these figures have some quantitative value. However, the results presented here and elsewhere (i.e. Porazinska *et al.* 2009; Egge *et al.* 2013) showed otherwise, and we discourage such a practice and suggest limiting the presentation of results to a list of species or OTUs (Bohmann *et al.* 2011; De Barba *et al.* 2014), at least when working with the barcoding COI region. It remains to be seen if this limitation also applies to more conserved genomic regions like the mitochondrial 16S (Clarke *et al.* 2104).

The quantitative potential of the reported technique is limited because of the PCR step conducted on a heterogeneous mixture. We see no other way to use high-throughput sequencing platforms for quantitative biodiversity assessment than to circumvent the PCR step, that

is, to directly sequence the DNA extracted from the sample (Taberlet *et al*. 2012; Yu *et al*. 2012). The present contribution joins several metabarcoding studies considering similar issues and already heading in this direction (Shehzad *et al*. 2012; Yu *et al*. 2012; Zhou *et al*. 2013; Clarke *et al*. 2014). In line with them, we would like to finish by highlighting what we think are our main contributions. First, we quantified for the first time the effect of blocking oligonucleotide concentration on non-target species relative abundances and showed that it affected most species but to a limited degree (i.e. less than one order of magnitude). Second, albeit being 'public knowledge' that universal primers bias the amplification efficiency of different species in heterogeneous DNA mixtures, we precisely showed that it spans 5 orders of magnitude and that most of the variance (~3/4) can be explained by the number of primer–template mismatches.

Acknowledgements

We thank several anonymous reviewers and Bruce Deagle for useful suggestions on preliminary versions of the manuscript. Financial help was provided by the MINECO grants CGL2010-18182 and AGL2011-24349. Carla Romeu-Dalmau and Laia Mestre provided some of the DNA extracts used here. We thank Vanesa Vega, Mercè Miquel, Oriol Casagran and Jordi Martinez-Vilalta for technical support.

References

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

Bohmann K, Monadjem A, Noer CL *et al*. (2011) Molecular dietary analysis of two African free-tailed bats (Molossidae) using high throughput sequencing. *PLoS ONE*, **6**, e21441.

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.

Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA (2013) Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology*, **10**, 45.

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

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

Deagle BE, 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.

Egge E, Bittner L, Andersen T, Audic S, de Vargas C, Edvardsen B (2013) 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine haptophytes. *PLoS ONE*, **8**, e74371.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

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

Fouts DE, Szpakowski S, Purushe J *et al*. (2012) Next generation sequencing to define prokaryotic and fungal diversity in the bovine rumen. *PLoS ONE*, **7**, e48289.

Gregory TR (2014). Animal Genome Size Database. http://www.genomesize.com.

Hiiesalu I, Pik MO, Metsis M *et al*. (2012) Plant species richness belowground: higher richness and new patterns revealed by next-generation sequencing. *Molecular Ecology*, **21**, 2004–2016.

Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN 4. *Genome Research*, **21**, 1552–1560.

Kelly RP, Port JA, Yamahara KM, Crowder LB (2014) Using environmental DNA to census marine fishes in a large mesocosm. *PLoS ONE*, **9**, e86175.

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.

Kwok S, Kellogg DE, McKinney N *et al*. (1990) Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic Acids Research*, **18**, 999–1005.

Leray M, Yang JY, Meyer CP *et al*. (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity; application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, **10**, 34.

McPherson M, Moller S (2006) *PCR*, 2nd edn. Taylor & Francis, London.

Meusnier I, Singer GAC, Landry JF, Hickey DA, Hebert PDN, Hajibabaei M (2008) A universal mini-barcode for biodiversity analysis. *BMC Genomics*, **9**, 214.

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.

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

Pochon X, Bott NJ, Smith KF, Wood SA (2013) Evaluating detection limits of next-generation sequencing for the surveillance and monitoring of international marine pests. *PLoS ONE*, **8**, e73935.

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

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.

Porazinska DL, Sung W, Giblin-Davis RM, Thomas WK (2010) Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. *Molecular Ecology Resources*, **10**, 666–676.

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 Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raye G, Miquel M, Coissac E, Redjadj C, Loison A, Taberlet P (2011) New insights on diet variability revealed by DNA barcoding and high-throughput pyrosequencing: chamois diet in autumn as a case study. *Ecological Research*, **26**, 265–276.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Schmieder R, Lim YW, Rohwer F, Edwards R (2011) TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics*, **11**, 341.

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*, **8**, 1951–1965.

Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on16S 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'. *PNAS*, **103**, 12115–12120.

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

Steele RE, Pires JC (2011) Biodiversity assessment: state-of-the-art techniques in phylogenomics and species identification. *American Journal of Botany*, **98**, 415–425.

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

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

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.

Wintzingerode FV, Göbel UV, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, **21**, 213–229.

Xu L, Ravnskov S, Larsen J, Nilsson RH, Nicolaisen M (2012) Soil fungal community structure along a soil health gradient in pea fields examined using deep amplicon sequencing. *Soil Biology and Biochemistry*, **46**, 26e32.

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

Zeale MRK, Butlin RK, Barker GLA, Lees DC, Jones G (2011) Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources*, **11**, 236–244.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.

Zhou X, Li Y, Liu S *et al.* (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4.

J.P. designed the study, conducted the bioinformatics and drafted the manuscript. G.M. quantified the individual DNA extracts, prepared the mock sample and was responsible for the ION TORRENT sequencing. P.G.-P. conducted the PCR laboratory work and analysed the Sanger sequences. N.A. codesigned the study, supervised the laboratory work and codrafted the manuscript. All authors played a role in editing the final version of the article.

Data Accessibility

High-throughput sequence data (FASTQ files): doi:10.5061/dryad.2p51n.

Database with the COI barcodes of the eleven species obtained by Sanger sequencing (FASTA file): doi:10.5061/dryad.2p51n.

Database with the COI barcodes of 188 species of Entomobryidae (Collembola) obtained from GenBank and used here (FASTA file): doi:10.5061/dryad.2p51n.

Supporting Information

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

Table S1 Number of reads obtained by high-throughput DNA sequencing from each experimental condition considered in this study

Table S2 Alignment of the Zeale *et al.* (2011) primers with the eleven species of arthropods considered in this study

Table S3 Alignment of the Meusnier *et al.* (2008) primers with the eleven species of arthropods considered in this study.