Title: Slippage of degenerate primers can cause variation in amplicon length

Running Title (45 char max): Primer degeneracy and varying amplicon length

Word count: _____

Authors: Vasco Elbrecht[1*], Paul D.N. Hebert[1,2], Dirk Steinke[1,2]

Affiliations:

1) Centre for Biodiversity Genomics, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

2) Department of Integrative Biology, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

*Corresponding author: Vasco Elbrecht (elbrecht@uoguelph.ca),

Abstract:

Metabarcoding studies often employ degenerate primers to reduce amplification bias and increase the number of detected taxa. However, degeneracy has the disadvantage of lowering binding specificity although the exact mechanisms and potential biases introduced by such off-target amplification are not fully understood.

We examined sequences recovered from the ten most abundant operational taxonomic units (OTUs) in two mock communities to investigate the specificity and binding behavior of five degenerate primer sets. Our results indicate that primers frequently bound 1-2 bp upstream in taxa where a homopolymer region was present in the amplification direction. As well, although less frequent, degeneracy occasionally led to primer binding 1 bp downstream. Some widely used primer sets were severely affected by this slippage effect, while others were not.

Our study shows that primer slippage can produce taxon-specific length variation in amplicons and subsequent length variation in recovered sequences. While this variation will only have small impacts on OTU designation by clustering algorithms that ignore terminal gaps, primer sets employed in metabarcoding projects should be evaluated for their sensitivity to slippage. Moreover, steps should be taken to reduce slippage by improving protocols for primer design. For example, the flanking region adjacent to the 3' end of the primer is not considered by current primer development software although GC clamps in this position could mitigate slippage. While degeneracy is important to ensure the universality of a primer, binding in homopolymer regions should be avoided.

Key words: Primer development, degeneracy, metabarcoding, primer slippage, bias, length variation

Introduction

Metabarcoding permits the rapid assessment of biodiversity (Baird & Hajibabaei 2012) using amplicon-based high-throughput sequencing (Taberlet *et al.* 2012). For metazoans, a segment of the cytochrome *c* oxidase subunit I (COI) gene is used (Hebert *et al.* 2003) as it offers species-level resolution coupled with access to extensive reference data (Ratnasingham & Hebert 2007). However, sequence variability in this gene region makes primer design difficult, especially when analyzing bulk samples that include a broad array of taxa (Sharma & Kobayashi 2014). Mismatches between primer and template DNA can lead to substantial primer bias, causing some taxa to remain undetected (Piñol *et al.* 2014; Elbrecht & Leese 2015). Although ribosomal markers provide more conserved primer binding sites (Deagle *et al.* 2014a), well designed COI primers can match or exceed the performance of ribosomal markers (Elbrecht *et al.* 2016; Clarke *et al.* 2017).

A key component of successful COI metabarcoding primers is primer degeneracy to allow matching at variable binding sites (Elbrecht & Leese 2017). Tools such as PrimerMiner support the automated download and processing of reference sequence data for the taxonomic group(s) targeted for analysis (Elbrecht & Leese 2016). Sequence alignments built from such datasets help to identify suitable primer binding sites. Matching to variable binding sites is optimized by inserting degenerate bases into the primer sequence. However, high degeneracy raises the chance that primers will also bind to non-target regions (Linhart & Shamir 2002). While sequences from non-target regions can be filtered out bioinformatically or by size selection of PCR products (assuming different amplicon lengths), such filtration can reduce the yield of target fragments. Thus, primer degeneracy is a tradeoff between maximizing taxon recovery (Piñol *et al.* 2014; Elbrecht & Leese 2017) and primer specificity. In a previous study, we observed length variation among sequences recovered from most primer combinations (Figure S6 in Elbrecht & Leese 2017), but did not investigate the mechanisms underlying this variation or the extent of variation in this effect at a species level.

In this study we analyze binding specificity for five primer sets in studies on mock assemblages of freshwater and marine macroinvertebrates (Leray & Knowlton 2017; Elbrecht & Leese 2017; Vamos *et al.* 2017). These datasets were chosen because haplotype sequences for most specimens were known, allowing precise determination of primer binding behavior on both the targeted binding regions and flanking areas.

Material and Methods

To investigate the specificity of primer binding, we analyzed five different primer sets used for metabarcoding of mock communities with known composition (Leray & Knowlton 2017; Elbrecht & Leese 2017; Vamos *et al.* 2017). COI sequences spanning the Folmer region (Folmer *et al.* 1994; Hebert *et al.* 2003) were available for most taxa which allowed

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

the analysis of potential length variation in amplicons generated by each primer set and specimen at a haplotype level. Table 1 describes the primer combinations analyzed.

The datasets were retrieved from the NCBI Sequence Read Archive and demultiplexed using the JAMP v0.34 pipeline (github.com/VascoElbrecht/JAMP, Elbrecht & Leese 2017; Vamos *et al.* 2017). Only sequences of the randomly selected mock community "B" were analyzed (Elbrecht & Leese 2015). Sequence data from a study that examined marine macroinvertebrates (Leray & Knowlton 2017) were downloaded from figshare. The results from sequencing run 1 were used without demultiplexing to ensure sufficient sequencing depth. Raw sequences were paired end merged using Usearch v10.0.240 (fastq_mergepairs -fastq_pctid 90 -fastq_maxdiffs 999 -fastq_trunctail 0, Edgar 2010) and imported into Geneious 11.0.4 (Kearse *et al.* 2012). Based on OTU tables from the original studies, the ten most abundant OTUs for each primer combination were selected for analysis to ensure sufficient sequencing depth and to reduce stochastic effects (Leray & Knowlton 2017). Sequences from sample B and from the marine mock sample were mapped against the known haplotype sequence for each selected taxon (lowest sensitivity, a 100% match, and zero gaps in the sequence, haplotypes from Script S2 in Elbrecht & Leese 2017). Flanking regions in the sequence alignment were extracted for each taxon, and the length distribution of each primer sequence was determined. A few sequences (no more than three per taxon) were much longer than expected, likely due to sequencing artifacts, and were therefore excluded from further analysis, which still included several thousand sequences per taxon (Table S1). A t-test was used for each primer to differentiate between OTUs where 10% or more reads were affected by length variation and those that were unaffected. All R scripts used are available as supporting information (Scripts S1).

Results

Two of the three reverse primers (BR1, fwhR2) were not associated with length variation (>99% sequences had the expected length), but the other reverse and all four forward primers showed length variation (Table 1). A 1 bp insertion was present at the 3'end of some (<10%) amplicons generated by the fwhR1 and the BF2 primers, (Figure 1B, Figure S1C). Importantly, the 3' end of the fwhR1 primer binds to a homopolymer region with up to six cytosines in some species while the BF2 primer targets a low complexity region of cytosine and thymine. In those cases where taxa amplified with the BF2 primer were unaffected by deletions, some sequences were affected by 1 bp insertions (Figure 1B). Many of the sequences retrieved with the four forward primers (BF1, BF2, fwhF2 and mlCOIintF) were 1 - 2 bp shorter than expected (Figure 1, Figure S1C & D). The incidence of these truncated sequences varied among primers and with the nature of templates, with their frequency rising when a low diversity cytosine primer binding region extended in the direction of elongation. This

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

effect was particularly dramatic for some taxa amplified with the mlCOIintF primer. For example, 80% of the sequences were shorter than expected for OTU_92 where the primer bound to a homopolymer region spanning seven cytosines (Figure S1F). Interestingly, in taxa where this low diversity region was directly followed by a set of different nucleotides (e.g. a poly C region followed by A, T or G), <2% of the sequences were affected by deletions (Figure 1 and Figure S1D & F). There was significantly more length variation between OTUs where binding sites were followed by low diversity regions than those binding sites that were flanked by high diversity variation for all tested primers ($p = 0.003$, t-test, Table 1). Some primers, such as BF2, were associated with both insertions and deletions (Figure 1A). In a few cases, larger changes in sequence length were detected, apparently linked to compositional variation in the primer binding site. For example, OTU_3 possessed a tandem repeat (ACCC) within the primer binding region and when it was amplified with BF1, about 6% of the sequences possessed a 4 bp deletion in the amplicon as primer sequences were only 16 bp long instead of 20 bp.

Table 1: The specificity of binding to different template strands for three forward and four reverse primers. The performance of each primer was examined for the ten most abundant taxa in each PCR reaction (for which the template sequences were known). The exact primer length distribution and number of sequences used for this analysis are also provided in Table S1. For primers where no length variation was observed or for primers where all taxa showed length variation, no t-test could be applied (NA) due to the lack of groups (slippage vs no slippage).

Primer combination	Primer tested	Length variation	Proportion with expected length (± SD)	t-test (p value)	Figure	Data set
P5_BF1_0 + P7_BR1_4	BF1	- 1 to 2 bp for some taxa	80.42 (±18.94)	0.003	Fig 1A	(Elbrecht & Leese 2017)
	BR1	No variation	99.44 (±0.04)	NA	Fig S1A	
P5_BF2_0 + P7_BR1_4	BF2	- 1 to 2 bp for some taxa, + 1 for all taxa	62.14 (±14.07)	NA	Fig 1B	(Elbrecht & Leese 2017)
	BR1	No variation	99.45 (±0.04)	NA	Fig S1B	
P5_fwhF1_3 + P7_fwhR1_1	fwhR1	+ 1 for all taxa	96.24 (±1.22)	NA	Fig S1C	(Vamos et al. 2017)
P5_fwhR2_2 + P7_fwhF2_3	fwhR2	No variation	99.33 (±0.05)	NA	Fig S1E	(Vamos et al. 2017)
	fwhF2	- 1 to 2 bp for some taxa	82.23 (±22.05)	0.003	Fig S1D	(Vamos et al. 2017)
mlCOIintF + jgHCO2198, complete run 1	mlCOIintF	- 1 to 2 bp for some taxa	70.08 (±29.93)	0.003	Fig S1F	(Leray & Knowlton 2017)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

Discussion

This study describes length variation created when degenerate primers bind to low diversity regions of their target template. This length variation does not reflect the presence of an indel in the primer or the template, but rather results from the primer binding 1 - 2 bp downstream or upstream from its expected site. In such situations, amplicons are 1 - 2 bp longer or shorter than expected once primers are trimmed during bioinformatics processing. The fact that primer sequences were successfully trimmed from >99% of the reads in each sample (Elbrecht & Leese 2017) indicates that this variation reflects primer slippage rather than indels in the primer itself. Additionally, we detected a taxon-specific slippage effect in datasets from several independent studies that used different primer sets, making it unlikely this effect is caused by flaws in oligo synthesis. While we previously described primer-dependent length variation resulting from metabarcoding samples amplified with BF1 / BF2 (Elbrecht & Leese 2017), the present study demonstrates this phenomenon for a wider range of primer sets using individual COI barcoded specimens. The overall results indicate that when the 3' end of a primer binds to a low diversity region, the primer often also binds 1-2 bp away from its target binding region. We argue this process is influenced by primer degeneracy, by the composition of the template DNA, and by the length of the low diversity region in the template DNA, being most prominent when it exceeds the length in the primer binding region.

If primer slippage occurs, it usually involves a homopolymer region (e.g. CCCC) at the 3' end and leads to the deletion of 1 - 2 bases. Insertions were less common and were limited to single base inserts in the primer sets examined in this study. Figure 2 depicts how these indels are likely caused through off-target primer binding. Analysis of variation in the incidence of these events among taxa indicated that forward primer slippage only occurred when a homopolymer region extended in the extension direction of the primer. This constraint means that primer slippage is highly template dependent with marked differences among species or even between haplotypes of a species. The explanation for this pattern is clear - the primer is prevented from binding upstream if the homopolymer region is interrupted by different nucleotides, preventing forward slippage. This also means that primer slippage can be prevented by targeting regions with higher diversity, or by providing two different base pairs at the (usually conserved) 3’ end (e.g. a GC clamp). For example, the BR1 primer binds in regions with up to a 4 bp homopolymer of cytosine, but it does not show signs of primer slippage because of the GC at its end and the absence of another cytosine flanking the primer binding region. In cases where the DNA template shows similar repetitive patterns, slippage of more than 2 bp is possible, e.g. OTU_3 amplified with the BF1 primer (Figure 1A). In cases where the homopolymer region does not extend beyond the primer binding site, slippage can still occur in the opposite direction, leading to single base insertions as evidenced by both the fwhR1 and BF2 primers. The BF2 primer is particularly affected by insertions as it can bind to a poly-thymine /cytosine region, linked by a degenerate mixed base (Y = T or C).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

In general, when primer slippage occurs, it leads to deletions rather than insertions, likely reflecting the irreversible nature of primer shifts. For example, if a primer is successfully incorporated and amplified one bp upstream of the usual binding site, it will shorten the homopolymer, making all successively amplified fragments shorter as well, or even leading to further forward shifts (Figure 2C).

These issues have important implications for both primer design and for the bioinformatics analysis of sequence data. While it is generally recommended to avoid homopolymer regions in binding sites (Abd-Elsalam 2003), the variability of the COI barcoding fragment (Deagle *et al.* 2014b; Sharma & Kobayashi 2014), and the high degeneracy needed to reduce primer bias (Piñol *et al.* 2014; Elbrecht & Leese 2015), places strong constraints on primer design. Nevertheless, metabarcoding primers should be designed which bind to two different nucleotides at the 3' end to reduce the chance of primer slippage. Further, because primer slippage events are highly template specific, the sequence attributes of both the primer binding region and its flanking regions should be considered. To our knowledge, software currently employed for primer design only considers the nucleotide composition of the targeted binding site, and ignores the flanking region in the extension direction. Therefore, we recommend evaluating that all primers used for metabarcoding analysis be tested for their susceptibility to slippage. Our study clearly shows that commonly used or recommended metabarcoding primers such as e.g. mlCOIintF (Leray *et al.* 2013) or BF2 (Elbrecht & Leese 2017) are susceptible to substantial primer slippage.

Primer slippage can lead to a large proportion of sequences being a few bp longer or, more likely, shorter than expected. As this effect is highly template specific and differs between taxa, it can introduce substantial biases during bioinformatic processing. It can skew the representation of certain species or haplotypes, especially if a metabarcoding dataset is filtered to an exact amplicon length. If, on the other hand, sequences of slightly different length are included in the analysis, they could introduce a substantial bias by generating false OTUs if terminal gaps are counted as differences (Flynn *et al.* 2015). Thus, when analyzing metabarcoding data, it is essential to know if a primer set is sensitive to slippage, and if the results generated by the clustering algorithm are impacted by such variation. It is fairly easy to test for primer slippage by examining patterns of length variability in the amplicons and their location. If more than 10% of the sequences are 1-2 bp shorter than expected after primer trimming and the length variation is concentrated near the ends of the sequence, primer slippage is a likely cause.

Conclusions

This study shows that high primer degeneracy, when combined with low sequence diversity in the primer binding sites and flanking regions, can lead to slippage, producing sequences that are a few bp shorter or longer than the expected amplicon

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

length. As this effect is template specific, its extent can vary substantially, even among closely related species in a particular sample. This variation can create analytical complexity, especially when clustering algorithms consider flanking regions. Importantly, primer slippage can be mitigated by repositioning primers to more heterogeneous binding sites and by considering their flanking regions when designing primer sets.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

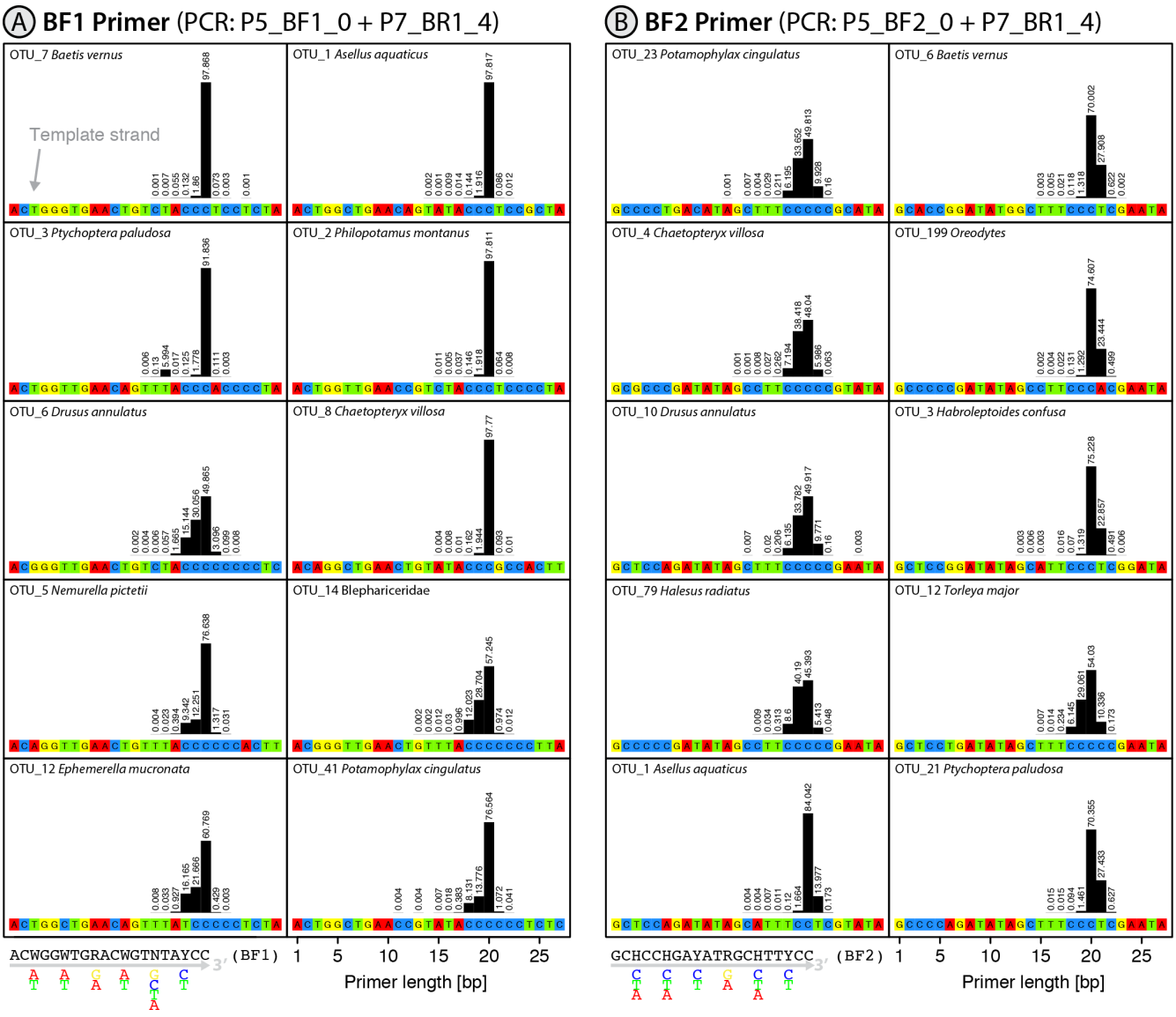

Figure 1: Plot showing the primer binding sites and bar plots depicting the length distribution of binding primers for the 10 most abundant OTUs in the mock sample B (sequence data from Elbrecht & Leese 2017). The sample was amplified with the N5_BF1_0 + N7_BR1_4 and N5_BF2_0 + N7_BR1_4 primer set, and the length distribution of the incorporated primers is shown for both the BF1 and BF2 primer (A and B). Relative abundance is given as a percentage above each bar.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

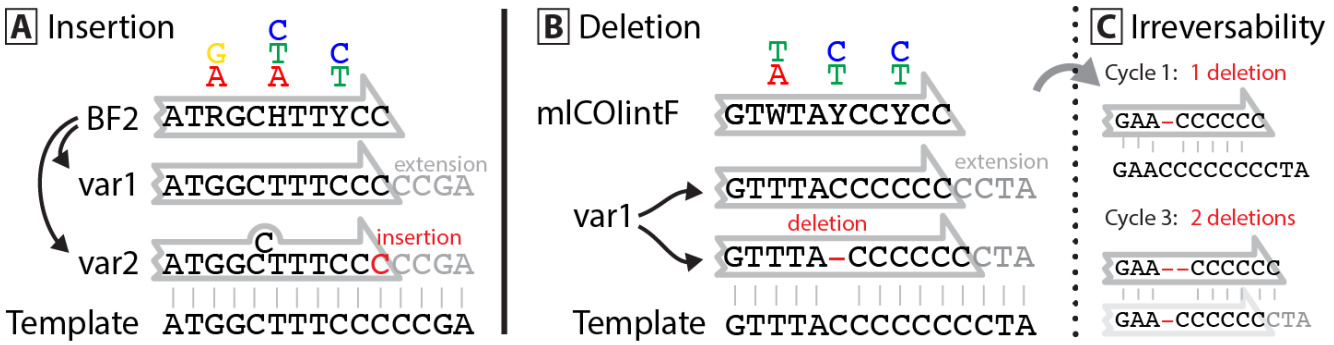

Figure 2: Proposed mechanism of primer slippage in binding regions with low diversity. Primer and template DNA are depicted with black letters, while nucleotides added during PCR are indicated by grey letters with insertions and deletions highlighted in red. **A**: Highly degenerate primers include many different primer versions (e.g. var1 & var2). These can slip "backwards" in low diversity binding regions, as the 3' primer tip can also match 1 bp upstream, leading to the incorporation of an additional base (shown here at the end of the BF2 primer). (**B**) Slippage in the forward direction is more common and follows a similar mechanism. The primer binds one position upstream which leads to the deletion of one nucleotide. When homopolymer regions are present at a primer binding site, forward slippage is much commoner than backward slippage. This effect is likely caused by the incorporation of primers throughout the PCR cycles (**C**), which can easily slip forward, but then shorten the homopolymer region providing less room for primers to bind and slip backward. If so, the extent of primer slippage should be PCR cycle dependent.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

Acknowledgements

This work was supported by the Canada First Research Excellence Fund. It represents a contribution to the 'Food From Thought' research program and to the European Cooperation in Science and Technology (COST) Action DNAqua-Net (CA15219).

Author contributions

V.E. developed the concept, analysed the data, and wrote the paper. D.S. and P.H. revised the paper.

Supporting information

Figure S1: Plots of length variation for six additional primers.

Table S1: Raw length distribution data and number of sequences used for each taxon and primer.

Scripts S1: R scripts used to analyze primer length distribution.

Abd-Elsalam, K.A. (2003). Bioinformatic tools and guideline for PCR primer design. *African Journal of Biotechnology*, **2**, 91–95.

Baird, D.J. & Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology* **21**, 2039–2044.

Clarke, L.J., Beard, J.M., Swadling, K.M. & Deagle, B.E. (2017). Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution*, **7**, 873–883.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**, 1–4.

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

Elbrecht, V. & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—Sequence relationships with an innovative metabarcoding protocol. *PloS one*, **10**, e0130324–16.

Elbrecht, V. & Leese, F. (2016). PrimerMiner: an R package for development and in silico validation of DNA metabarcoding primers. *Methods in Ecology and Evolution*, **8,** 622–626.

Elbrecht, V. & Leese, F. (2017). Validation and development of freshwater invertebrate metabarcoding COI primers for environmental impact assessment. *Frontiers in Environmental Science*. 5, 5–11.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., Coissac, E., Boyer, F. & Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, **4**, e1966–12.

Flynn, J.M., Brown, E.A., Chain, F.J.J., MacIsaac, H.J. & Cristescu, M.E. (2015). Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, **5**, 2252–2266.

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.

Hebert, P.D.N., Ratnasingham, S. & de Waard, J.R. (2003). Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, **270**, 96–99.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Leray, M. & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, **5**, e3006.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. & Machida, R.J. (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**, 1–1.

Linhart, C. & Shamir, R. (2002). The degenerate primer design problem. *Bioinformatics*, **18**, S172–S181.

Piñol, J., Mir, G., Gomez-Polo, P. & Agustí, N. (2014). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, **15**, 1–12.

Ratnasingham, S. & Hebert, P.D.N. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes*, **7**, 355–364.

Sharma, P. & Kobayashi, T. (2014). Are 'universal' DNA primers really universal? *Journal of Applied Genetics*, **55**, 485–496.

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

Vamos, E.E., Elbrecht, V. & Leese, F. (2017). Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*. **1**, e14625.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26763v1 | CC BY 4.0 Open Access | rec: 23 Mar 2018, publ: 23 Mar 2018