PeerJ Preprints | NOT PEER-REVIEWED

1 **Development and validation of DNA metabarcoding COI primers for aquatic**
2 **invertebrates using the R package "PrimerMiner"**
3

4 Vasco Elbrecht [1, 2]*, Florian Leese [1,3]
5
6 Affiliations:
7 1) Aquatic Ecosystem Research, Faculty of Biology, University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen,
8 Germany
9 2) Department of Animal Ecology, Evolution and Biodiversity, Ruhr University Bochum, Universitätsstraße 150, 44801
10 Bochum, Germany
11 3) Centre for Water and Environmental Research (ZWU) Essen, University of Duisburg-Essen, Universitätsstraße 2,
12 45141 Essen, Germany
13 *Corresponding author: vasco.elbrecht@uni-due.de, phone: +49.201-1834053
14

15 **Abstract**

16 1) DNA metabarcoding is a powerful tool to assess biodiversity by amplifying and sequencing a standardized gene
17 marker region. However, typically used barcoding genes, such as the cytochrome c oxidase subunit I (COI) region for
18 animals, are highly variable. Thus, different taxa in communities under study are often not amplified equally well and
19 some might even remain undetected due to primer bias. To reduce these problems, optimized region- and/or ecosystem-
20 specific metabarcoding primers are necessary.
21 2) We developed the R package PrimerMiner, which batch downloads DNA barcode gene sequences from BOLD and
22 NCBI databases for specified target taxa and then applies sequence clustering to reduce biases introduced by differed
23 number of available sequences per species. To design primers targeted for freshwater invertebrates, we downloaded
24 COI data for the 15 most important invertebrate groups relevant for stream ecosystem assessment. Four primer sets with
25 high base degeneracy were developed and their performance tested by sequencing ten mock community samples
26 consisting each of 52 freshwater invertebrate taxa. Additionally, we evaluated the developed primers against other
27 metabarcoding primers *in silico* using PrimerMiner.
28 3) Amplification and sequencing was successful for all ten mock community samples with the four different primer
29 combinations. The developed primers varied in amplification efficiency and amount of taxa detected, but all primer sets
30 detected more taxa than standard Folmer barcoding primers. Additionally, the BF / BR primers amplified taxa very
31 consistently, the BF2+BR2 and BF2+BR1 primer combination even better than a previously tested ribosomal marker
32 (16S). Except for the BF1+BR1 primer combination, all BF / BR primers detected all 42 insect taxa present in the mock

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

PeerJ Preprints NOT PEER-REVIEWED

samples. *In silico* evaluation of the developed primers showed that they are also likely to work very well on other non aquatic invertebrate samples.

4) With PrimerMiner, we here provide a useful tool to obtain relevant sequence data for targeted primer development and evaluation. Our sequence datasets generated with the newly developed metabarcoding primers demonstrate that the design of optimized primers with high base degeneracy is superior to classical markers and enable us to detect almost 100% of animal taxa present in a sample using the standard COI barcoding gene. Therefore, the PrimerMiner package and primers developed using this tool are useful beyond assessment of biodiversity in aquatic ecosystems.

Key words: Primer development, DNA metabarcoding, primer bias, ecosystem assessment, in silico PCR, data mining

PeerJ Preprints NOT PEER-REVIEWED

42 **1) Introduction**

43 DNA barcoding allows for the reliable identification of collected specimens without prior knowledge of species
44 taxonomy. A prerequisite is the availability of reliable reference databases. For animals, the usefulness of the
45 cytochrome c oxidase I (COI) gene for species identification has been widely demonstrated and extensive reference
46 databases exist for many taxonomic groups (Larsen *et al.* 2011; Ratnasingham & Hebert 2013). However, identifying
47 single specimens using DNA barcoding is still quite expensive, as DNA has to be extracted individually, the barcoding
48 region amplified and then typically Sanger sequenced (Cameron *et al.* 2006; Stein *et al.* 2014). Recent advances in high
49 throughput sequencing (HTS) make it now possible to extract DNA and sequence the barcoding region from bulk
50 environmental samples often containing hundreds to thousands of specimens. This technique, coined DNA
51 metabarcoding, has been already widely used to generate comprehensive taxa list for a wide range of ecosystems and
52 environments (Taberlet *et al.* 2012). One main challenge of DNA metabarcoding is the often severe primer bias that
53 prevents detection of 100% of the taxa present in a sample and limits quantification of biomass from read abundances
54 (Piñol *et al.* 2014; Elbrecht & Leese 2015).

55 A universal barcoding primer pair, amplifying a fragment of suitable length for HTS is thus the most critical component
56 when assessing environmental samples with metabarcoding. As a coding gene, the COI barcoding region shows high
57 codon degeneracy throughout its sequence, making design of "truly" universal primers difficult (Deagle *et al.* 2014;
58 Sharma & Kobayashi 2014). Several universal COI barcoding primers have been developed of which many are now
59 used or could be used in metabarcoding studies (Figure 1, Folmer *et al.* 1994; Hebert *et al.* 2004; Meusnier *et al.* 2008;
60 Van Houdt *et al.* 2010; Zeale *et al.* 2010; Shokralla *et al.* 2011; Leray *et al.* 2013; Geller *et al.* 2013; Gibson *et al.* 2014;
61 Shokralla *et al.* 2015; Brandon-Mong *et al.* 2015). Despite being universal and often including several degenerate bases,
62 these metabarcoding primers typically recover only 80-90% or even less of the taxa present in a sample (Leray *et al.*
63 2013; Elbrecht & Leese 2015; Brandon-Mong *et al.* 2015). Furthermore, many primers used in metabarcoding have
64 never been thoroughly evaluated with respects to primer bias and the proportion of undetected taxa. Thus, the
65 development and critical evaluation of universal primers is still a pressing issue.

66 Details on primer design and/or used sequence data are often not described extensively (e.g. (Hajibabaei *et al.* 2011;
67 Shokralla *et al.* 2015)). Typically many reference barcode sequences for the taxonomic target groups are taken from
68 NCBI or BOLD and aligned (Zeale *et al.* 2010; Leray *et al.* 2013; Gibson *et al.* 2014) or alternatively only
69 mitochondrial genomes or a small subset of barcoding sequences are used (Geller *et al.* 2013; Deagle *et al.* 2014;
70 Brandon-Mong *et al.* 2015). A key problem when downloading complete datasets is that some taxa are typically
71 overrepresented with hundreds of sequences deposited (e.g. because many sequences are available from detailed
72 phylogeographic studies). This can in principle be circumvented when using only mitochondrial genomes. However,

PeerJ Preprints | NOT PEER-REVIEWED

73 typically such data sets are very limited as mitochondrial genomic sequences are still rare for many taxonomic groups.
74 However, obtaining good quality reference data is essential in manual and software based primer development. While
75 there are many programs available to aid primer development (e.g. Primer3, Untergasser *et al.* 2012, EcoPCR, Ficetola
76 *et al.* 2010), the challenge of batch downloading and systematically preparing obtained sequence data for primer
77 development has not been tackled until now. Therefore, we have developed the R package PrimerMiner. The software
78 allows the user to selectively batch download and cluster sequences into Operational Taxonomic Units (OTU).
79 Clustering is independent to reported taxonomy and reduces biases introduced by misidentified taxa, database
80 redundancies and overrepresented taxa. PrimerMiner includes visualisation tools to manually search for suitable
81 metabarcoding primers. Further new *in silico* primer evaluation tools are introduced with PrimerMiner, which take type
82 and position of mismatches between primer and template into account.
83 To test the PrimerMiner approach, we designed four DNA metabarcoding primer sets, targeting 15 freshwater
84 invertebrate taxa of central importance in bioassessment programs. All primer sets were evaluated using ten mock
85 communities with 52 taxa each, which have been used for primer evaluation in previous studies (Elbrecht & Leese 2015;
86 Elbrecht et al. 2016). Additionally, the developed primers were evaluated *in silico* against commonly used DNA
87 barcoding and metabarcoding primers.
88
89

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

PeerJ Preprints NOT PEER-REVIEWED

2) Material and Methods

The PrimerMiner R package

PrimerMiner is a fully automated R based sequence downloader and processor that condenses sequence data from NCBI and BOLD into Operational Taxonomic Units (OTUs) (Figure 2). It can download sequence data for a selected gene marker and specified taxonomic groups and also extract the respective target gene sequence from mitochondrial genomes if available. Thus, PrimerMiner takes full advantage of available partial sequences and mitochondrial genomes, laying a good data basis for primer development. All sequences are then clustered with a custom threshold (default 3%) and the OTUs for each taxonomic group exported as a fasta file for subsequent alignments and automated primer design with specific software or manually. The clustering strategy adopted in PrimerMiner has several key advantages: 1) Overrepresented taxa and duplicated sequences are merged into few OTUs. 2) Taxonomic variation within species is retained (wobble bases) while rare haplotypes can be ignored when generating OTU consensus sequences. 3) Highly diverse species are automatically represented by two or more OTUs. 4) Clustering is taxonomy-independent and thus can deal with misidentified species as long as their order / family was identified correctly.

The latest version is available on GitHub with a quick video guide on YouTube (https://github.com/VascoElbrecht/PrimerMiner). An internet connection as well as Mac OSX or Linux operating system is required, as PrimerMiner relies on Vsearch for clustering (https://github.com/torognes/vsearch). The program is configured with a txt file, which is created by running "batch_config()". Target orders or families for which sequences should be obtained have to be specified in a csv file. Thus inclusion of a subset of taxa from a lower taxonomic level is possible. For example, for a certain order, a subset of families can be downloaded (e.g. only aquatic Coleoptera families) by specifying these in the second table column. Downloading data from higher than family taxonomy can cause the download to fail if group names are not unique and is thus not recommended. By running "batch_download()" matching barcode sequences are downloaded and processed.

By default, complete and partial COI sequences are download from the BOLD and NCBI databases. Additionally, the target marker is extracted from mitochondrial genomes if available on NCBI. Sequences are then dereplicated and clustered using Vsearch with 3% similarly threshold. Majority consensus sequences for each OTU are written into a fasta file for each group. All raw sequencing data as well as intermediate files and summary statistics are automatically saved. Subsequently, the generated OTU sequences for each group have to be aligned with e.g. Geneious (Kearse *et al.* 2012) and can then be used in other primer development tools or visualized for manual primer development using the "plot_alignments()" command.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

PeerJ Preprints | NOT PEER-REVIEWED

121 **Primer development using PrimerMiner**

122 The PrimerMiner package v0.2 was used to download COI and cluster sequences for the 15 most assessment relevant
123 freshwater invertebrate groups from NCBI and BOLD (Accessed February 2015, Taxa table S1). Sequences were
124 aligned with MAFFT v7.017 (Katoh *et al.* 2002) as implemented in Geneious 8.1.7 (Kearse *et al.* 2012) and the
125 alignment for each group was visualized with PrimerMiner. The alignment plot was used to identify suitable primer
126 binding sites manually. Two forward (BF1, BF2) and two reverse primers (BR1, BR2) were designed with high base
127 degeneracy. Fusion primers were generated by adding Illumina adapters and inline barcodes as described in (Elbrecht &
128 Leese 2015) to increase sequence diversity and allow for a one step PCR protocol.
129

130 **Testing of DNA metabarcoding primers on mock communities**

131 Amplification success of the BF / BR primers was evaluated using ten mock communities, each containing a set of 52
132 freshwater invertebrates used in previous studies (Elbrecht & Leese 2015). The identical DNA aliquot and one step
133 PCR protocol as in (Elbrecht & Leese 2015) was used for all four primer combinations. As in the previous studies, each
134 sample was uniquely tagged from both sides, but sometimes only 25 ng instead of 50 ng DNA was used in PCR (see
135 Figure S1). For each primer combination all ten samples were run in the PCR setup, using one PCR replicate per sample.
136 Ready-to-load products were magnet-bead purified (left sided, 0.8x SPRIselect, Beckman Coulter, Bread, CA, USA)
137 and quantified using the Qubit HS Kit (Thermofisher Scientific, Carlsbad, CA, USA). For each primer combination,
138 equimolar amounts of amplicons were pooled into one library (taking fragment length differences into account, Figure
139 S1). The library was sequenced on one lane of a HiSeq 2500 (rapid run, 2x250 bp) with 5% PhiX spike-in, carried out
140 by the DNA Sequencing Center of Brigham Young University, USA.
141 Bioinformatic processing of high throughput data was kept as similar as possible to previous studies (Elbrecht & Leese
142 2015; Elbrecht *et al.* 2016). In short, reads were demultiplexed (R script **- will be supplied in next version of preprint**)
143 and paired end reads merged using Usearch v8.1.1831 -fastq_mergepairs with -fastq_merge_maxee 1.0 (Edgar &
144 Flyvbjerg 2015). Where necessary, reads were converted into reverse complement. For each primer combination all ten
145 replicates were pooled, and sequences which were present only one single time in the dataset (singletons) removed prior
146 to clustering with Usearch (cluster_otus, 97% identity, strand plus, includes chimera removal) (Edgar 2013).
147 Dereplicated reads for each of the 40 samples (including singletons) were compared against the respective OTU dataset,
148 using usearch_global with a minimum match of 97% and strand plus. Like in previous studies, low abundance OTUs
149 without at least one sample above 0.003% sequences assigned, were considered unreliable and excluded from the
150 dataset. Taxonomy of the remaining OTUs was identified and manually verified using the BOLD and NCBI database.
151 To ensure that taxonomy was consistently assigned across primer combinations and in comparison to the reference COI

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

PeerJ Preprints NOT PEER-REVIEWED

152 study (Elbrecht & Leese 2015), the most abundant sequence for each OTU in each sample was extracted using an R
153 script, and the haplotype of all individual specimens assembled if possible.
154

155 ## *In silco* evaluation of primers

156 PrimerMiner has powerful *in silico* primer evaluation capabilities, allowing or evaluation of single primers and primer
157 pairs on any given sequence alignment. Unlike ecoPCR (Ficetola *et al.* 2010), PrimerMiner factors in the position and
158 type of each primer / sequence mismatch, which gives a more comprehensive picture, as amplification success is highly
159 dependent on a good matching 3 ' primer end (Piñol *et al.* 2014) (**add more refs**). Using the command
160 "primer_evaluation()" PrimerMiner calculates individual penalty score for each template to primer mismatch, factoring
161 in the position and type of mismatch and thus giving a more realistic evaluation of amplification efficiencies. Penalty
162 scores for position and mismatch type are fully customisable, by providing your own penalty tables. Mismatch
163 evaluations for each sequence are stored in a table, allowing full transparency and processing in other programs. With
164 the function "combine_2_primers()" two primer pairs can be evaluated against each other using the generated tables
165 with "primer_evaluation()", giving a maximum threshold under a primer pair is considered working for amplification.
166 All metabarcoding primers shown in Figure 1 were evaluated against 30 insect orders alignments following the
167 taxonomy by (Misof *et al.* 2014). Data was downloaded and clustered with PrimerMiner v0.3b on April 2016.
168
169
170

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

PeerJ Preprints NOT PEER-REVIEWED

171 **3) Results**

172 **Developed primers using PrimerMiner**

173 Using the alignments of 15 freshwater assessment relevant groups (Figure S2) we designed four primer pairs (Figure
174 S3). Sequence coverage was increased 249 (SD=395) times on average by including COI barcode sequences to the
175 mitochondrial reads (Figure S2). The two forward and two reverse primers show high base degeneracy to amplify as
176 many insect taxa as possible. Amplified regions range from 217 bp for internal barcodes up to 421 bp for combinations
177 using a degenerated version of the HCO2198 primer. While samples in this study were tagged uniquely from both sides,
178 the inline barcodes allow for tagging of up to 72 samples for each primer combination (see Figure S4 for recommended
179 primer combinations).
180 All four primer combinations were tested on ten invertebrate mock samples on a Illumina HiSeq sequencer. PCR
181 efficiency varied across primer combinations, with PCRs involving the BF2 primer showing good amplification
182 whereas those with BF1 primer always showing decreased yields (Figure S5). Amplification efficiency with fusion
183 primers was always substantially lower than the positive control (standard COI Folmer primers, without Illumina tail).
184 Sequencing was successful for all samples, with obtaining very similar amounts of sequences for all replicates (on
185 average 1.55 million reads per sample, SD = 0.2, Figure S1A). Cluster density on the lane was low (402 k/mm^2)
186 yielding only 48.74% of the expected sequencing output, yet with good sequence quality (Q30 ≥ 92.17%, raw data
187 deposited on SRA: SRX1619153). The amplified read lengths had an influence on the amount of sequences retained in
188 bioinformatic processing. Longer amplicons have less overlap when PE merged and are thus excluded more often due
189 to expected errors > 1 (Figure S1B). Additionally primer combinations that used the P5_BF12 primer lost more
190 sequences than other combinations, as ~1/5 of the reads had poor Phred scores. Furthermore, there were issues with the
191 BF1 and BF2 primer which showed insertions or deletions on the 3' end affecting sequence length by 1-2 bp across all
192 replicates (Figure S6). Some primer combinations also amplified up to 1.35% shorter or longer fragments than expected
193 (Figure S7).
194

195 **Amount of taxa recovered**

196 All insect taxa present in the mock samples were detected with each primer combination (Table 1), with exception of
197 the BF1 + BR1 combination that failed to amplify the Scirtidae (Coleoptera). All primers failed for some of the other
198 metazoan taxa, with the BF1 + BR2 combination showing the least amount of undetected taxa. In comparison to the
199 traditional Folmer primers (Folmer et al. 1994), all BF / BR freshwater primers showed a more consistent and equal
200 read abundance across the mock samples (Figure 3). As in Elbrecht et al. (2016), the standard deviation from the
201 expected abundance and precision for the primer pairs was estimated, which summarizes the variance in amplification.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

PeerJ Preprints NOT PEER-REVIEWED

Primer combination BF1 + BR1 showed the highest inconsistencies in read abundance, while the BF2 + BR2 combination showed even higher precision than a previously tested 16S marker (Elbrecht et al 2016). The proportion of detected non-insect metazoan taxa varied between primer combinations, with the combination BF1+BR2 detecting all but one taxon.

Table 1: Number of species recovered with the newly developed primers and data on 16S and Folmer primers from previous tests (Elbrecht & Leese 2015; Elbrecht *et al.* 2016).

Taxonomic group	Number of specimens	Number of specimens recovered with specific primer combination											
		LCO1490+HCO2198		16S ins		BF2+BR2		BF2+BR1		BF1+BR2		BF1+BR1	
		7	(88%)	8	(100%)	8	(100%)	8	(100%)	8	(100%)	8	(100%)
Plecoptera	**4**	4	(100%)	4	(100%)	4	(100%)	4	(100%)	4	(100%)	4	(100%)
Trichoptera	**15**	13	(86%)	15	(100%)	15	(100%)	15	(100%)	15	(100%)	15	(100%)
Diptera	**8**	7	(88%)	7	(88%)	8	(100%)	8	(100%)	8	(100%)	8	(100%)
Other insects	**7**	7	(100%)	7	(100%)	7	(100%)	7	(100%)	7	(100%)	6	(86%)
Other metazoa	**10**	5	(50%)	2	(20%)	7	(70%)	6	(60%)	9	(90%)	6	(60%)
Σ All insects	**42**	38	(91%)	41	(98%)	42	(100%)	42	(100%)	42	(100%)	41	(98%)
SD*		1.01		0.62		0.84		0.65		0.71		0.54	
Precision**		0.72		0.37		0.58		0.35		0.49		0.28	
Σ All taxa	**52**		43 (83%)	43	(83%)	49	(94%)	48	(92%)	51	(98%)	47	(90%)

* Standard deviation (SD) of logarithmic sequence abundance for all samples that worked (specimens with < 0.003% read abundance discarded)

** Precision defined as the SD of the mean log distance to the expected abundance, calculated for each morphotaxon.

In silico evaluation of primers

Figure 4 gives an overview of 11 forward and 12 reverse primers evaluated against OTUs of all insect orders. Reference data for binding sites of the standard Folmer primers HCO and LCO was very limited, and six out of 29 orders had below 100 sequences in total. Primer efficiencies were very similar across orders but varied between primers. However, primers incorporating wobble bases (BF1, BF2, BR1, BR2, jgHCO2198) or inosin (Ill_B_F, ArF5, Il_C_R, ArR5) performed better than primers with no or just few wobble bases. Figure S8 shows an evaluation of primer pairs, giving results consisted to evaluations of individual primers. It should be noted that some primers from the literature are not only poorly matching because they lack wobble bases, but can be affected by additional problems (see Figure S2, "critical mismatches"). For instance, near the 3' ends, the EPT-long-univR has a completely unnecessary second inosine at a conserved position, while the Uni-MinibarF1 had a "T" at a position where more than half of the reference OTUs had an "A". Furthermore, the L499 primer targets a highly variable region. Finally, certain primers show mismatches to particular groups, e.g. the ZBJ-ArtF1c and BR1 primers do not match well to sequences of Bivalvia.

PeerJ Preprints NOT PEER-REVIEWED

229 ## 4) Discussion

230 We used PrimerMiner to develop four primer sets for freshwater invertebrates based on OTU sequence alignments
231 generated of mitochondrial and COI barcodes from NCBI and BOLD. By not only using only mitochondrial genomes
232 but also including COI barcode data from BOLD and NCBI primer design was built upon a solid and balanced data
233 basis. Clustering helped to avoid overrepresentation of taxa with many sequences available in data bases, making sure
234 that each species is represented by only a few majority consensus OTU sequences. Due to the high variability
235 throughout the COI gene alignments (Sharma & Kobayashi 2014) and complexity of the task, we here decided to search
236 for primers manually, instead of using available software solutions. We deliberately decided to not factor in nucleotide
237 variability present in only few groups (mostly non-insect Metazoa), to limit the degeneracy of the primers to a
238 reasonable level.
239 We further decided (and recommend) to develop COI metabarcoding primers internal of the Folmer region, as sequence
240 coverage is still quite limited on the Folmer primer binding sites (Figure 4). We consider 100 OTUs for a given order as
241 a minimum coverage for a primer binding site to capture its variability and select necessary wobble bases. Due to the
242 codon degeneracy, larger alignments do not necessarily give much additional information. Thus, for the HCO binding
243 region it is often possible to obtain reliable information while the sequence depth of the LCO primers is often limited to
244 mitochondrial genomes (<100 OTUs available). In conclusion, PrimerMiner is an efficient and valuable tool to obtain
245 and visualize meaningful sequence data to design and evaluate universal metabarcoding primers, tailored to the
246 taxonomic groups present in the studied ecosystem.
247

248 ### Amplification success of mock communities

249 All primer sets amplified the ten mock communities successfully. By factoring in the different amplicon lengths in
250 library pooling we obtained similar amount of reads for each sample. All degenerated COI primers showed superior
251 detection rates (up to 100% of insects and 98% of all morphotaxa) and more consistent read abundances compared to
252 the standard Folmer barcoding primers that lacked any base degeneracy (Folmer *et al.* 1994; Elbrecht & Leese 2015).
253 The primer sets BF2 in combination with BR1/BR2 even showed better detection rates and higher precision than a
254 previously used 16S primer, which was tested on the same communities (Elbrecht *et al.* 2016). Also *in silico* analysis of
255 the BF / BR primers against all insect taxa on NCBI and BOLD confirmed their excellent detection rates, with mean
256 success rates near 100%. (Deagle *et al.* 2014) strongly argued against the use of degenerated primers to be used in DNA
257 metabarcoding, proposing the use of ribosomal markers with more conserved binding regions instead. However, we
258 here clearly show that highly degenerated COI primers are not only feasible but also superior to ribosomal
259 metabarcoding of animals when it comes to primer performance and available reference databases.

PeerJ Preprints NOT PEER-REVIEWED

260 While our developed primers show very reliable amplification results, there are also problems associated with the
261 primers itself and well as the applied metabarcoding protocol. First, while the use of fusion primers potentially
262 decreases the chance of index switching and reduces needed laboratory work, it also reduces PCR efficiency
263 substantially (Schnell *et al.* 2015). While primer combinations involving BF2 primers were less affected by this issue, it
264 was more pronounced with the BF1 primer especially in combination with BR1. Further, concerns have been raised by
265 biases associated with use of indexed primers (O'Donnell *et al.* 2016). While we could not observe any obvious effects
266 in our current data set (most taxa are detected to equal proportions regardless of primer index), there was a drop in
267 sequence quality when using the BF12 primer. Whether this is a systematic effect associated with the primer index or a
268 problem in e.g. primer synthesis / quality cannot be determined from this data set. However, independent of this
269 possible bias, it did not have any effects on the number of detected taxa. Additionally, 17% of reads from the BF2+BR2
270 primer combinations were discarded due to low expected error values, as the overlap was limited with 250 PE
271 sequencing of a 421 bp region on the HiSeq system. Further, with highly degenerated primers, the specificity of the
272 primers decreases (Deagle *et al.* 2014) potentially amplifying non target regions or unexpected lengths. This effect was
273 often minimally, with few sequences deviation from the expected length (below <0.5 % for most primers sets), with
274 these numbers being potentially inflated by PCR / Sequencing errors and pseudo genes (e.g. Nemuridae). However,
275 more problematically the BF1 and BF2 primers were affected by shifting effects making up to 40% of the sequences 1-2
276 bp shorter or longer at the primer binding side. It is not particularly clear what causes this effect, which can be also to
277 observed lesser degrees in in datasets from previous studies (Elbrecht & Leese 2015; Elbrecht *et al.* 2016). Potentially
278 the high degeneracy of the forward primers in combination with low diversity nucleotides at the primer's 3 ' end (e.g.
279 C[cta]TT[tc]CC in BF2) makes this effect particularly pronounced. Thus we recommend designing primers with two
280 unique nucleotides on the 3 ' end. The effect of this minimal shifting shortens the read length by 1-2 bp which has no
281 effect on detection on taxa (OTUs will still match the same reference taxon, regardless of 1-2 bp being clipped form the
282 sequence). However, when calculation OTU based biodiversity indices, the small shift might lead to a bias in these
283 metrics due to inflated OTU numbers. This might be countered by increasing the OTU clustering threshold to e.g. 4%,
284 but we advice to take OTU based diversity measures with caution using the BF / BR primer set.
285

286 **Primer success is determined by base degeneracy**

287 *In silico* analysis of 23 potentially suitable primers for COI DNA metabarcoding sowed that high primer degeneracy
288 leads to the best amplification of freshwater and insect taxa. This was also confirmed experimentally, with the tested
289 macroinvertebrate mock communities showing high primer bias with standard Folmer primers (Elbrecht & Leese 2015),
290 and very consisted amplification with higher detection rates with the primers developed in this study. While other

PeerJ Preprints NOT PEER-REVIEWED

primers from (Gibson *et al.* 2014) and (Shokralla *et al.* 2015) probably lead to equally good amplification rates as the BF/BR primers, a lack of degeneracy can lead to substantial primer bias. While these biases might not affect PCR on single organisms for DNA barcoding strongly, they will substantially skew detection rates of complex multispecies bulk samples, in the worst case leading to taxa remaining undetected (Piñol *et al.* 2014; Elbrecht & Leese 2015). This might already be the case, when primers have to little degeneracy like the mlCOIint primers by (Leray *et al.* 2013), which have a maximum degeneracy of two nucleotides at each position. The mlCOIint primers were tested with two mock communities, containing DNA from previously barcoded taxa (Leray & Knowlton 2015). Leray et al. 2015 reported that up to 35% of taxa remained undetected, which is consistent with the *in silico* primer evaluations in this study. Probably even more problematic are primers, which show no base degeneracy at all. While the primer bias associated with the high variation of the COI gene have been well known (Clarke *et al.* 2014; Deagle *et al.* 2014; Sharma & Kobayashi 2014; Piñol *et al.* 2014; Elbrecht & Leese 2015), primers without base degeneracy like ZBJ-Art by (Zeale *et al.* 2010) are widely used and recommended e.g. for gut content analysis (Pompanon 2012???). This can be really problematic, as large proportions of biodiversity might be missed or underrepresented in studies using these primers. Even when primers have good success rates for barcoding of single specimens (Meusnier *et al.* 2008), they are likely to introduce huge primer bias in metabarcoding studies. Thus careful evaluation of primers to the specific groups of interest in the planned metabarcoding study is curtail. PrimerMiner provides helpful tools to obtain and evaluate group specific sequence data needed for theses evaluations. Further, the efficiency of popular primer sets should additionally tested using mock communities, to detect specific biases introduced by the primers or the specific metabarcoding protocol.

Recommended approaches for assessment of insects and freshwater taxa

The success of every DNA metabarcoding project depends on well designed primers, which amplify the target communities as consistent as possible. Amplification bias depends on primer binding regions, which can be more conserved in ribosomal genes than in COI. Thus 18 and 16S markers have been proposed as suitable alternatives, despite lacking comprehensive reference databases (Clarke *et al.* 2014; Deagle *et al.* 2014; Elbrecht *et al.* 2016). Given the *in silico* evaluations and better performance of the BF2 + BR1 / BR2 primer sets, it can be settled that ribosomal markers are not necessary for reliable DNA metabarcoding on animals. Thus, to only remaining challenge is to find the ideal COI metabarcoding marker, suitable for your groups of interest. PrimerMiner can be a helpful tool to evaluate existing markers and if needed build new ones. Also we encourage to try and evaluate combining primers from different primer sets and test them *in silico*.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

PeerJ Preprints | NOT PEER-REVIEWED

321 When using DNA metabarcoding approaches for ecosystem assessment, protocols from the literature should be
322 critically evaluated. We recommend using the illumina HiSeq sequencer in rapid run mode (250 bp PE reads) and
323 include replications to reduce changes fro tag switching and exclude false OTUs from the dataset. While we have
324 previously encouraged the use of fusion primer due to their ease of use (single step PCR, (Elbrecht & Leese 2015)), we
325 have to acknowledge that they decrease PCR efficiency, and thus two step PCRs might be better suited for
326 environmental samples which often contain PCR inhibitors.
327 Additionally, metagenomic approaches using enrichment for mitochondrial genomes could be suitable for assessment of
328 ecosystems, which potentially less bias as the PCR amplification step can be omitted (Liu *et al.* 2015). However, as
329 briefly discussed in (Elbrecht *et al.* 2016), metagenomics methods have to be further validated and mitochondrial
330 reference genome libraries completed (Dowle *et al.* 2015). Thus, the selection of a specific metabarcoding or
331 metagonomics approach depends on future developments, available resources and expertise in the laboratories.
332 However, it is clear that if one decides to apply DNA metabacoding, primers have to be carefully evaluated or even
333 newly developed to optimally amplify the targeted groups of the specific project.
334

335 Conclusions

336 With PrimerMiner, we have developed a useful R package for primer development and evaluation, which we here used
337 to design new DNA metabarcoding primers targeting freshwater invertebrates. Our *in silico* evaluations as well as mock
338 communities metabarcoding experiments clearly indicated that with highly degenerated COI primers almost 100% of
339 the taxa were not only detected, but also amplified with highly similar read numbers. Thus, we argue that COI is the
340 marker of choice to use in animal metabarcoding, dismissing other markers such as ribosomal markers as a suitable
341 alternative due to the poor reference data for these. We additionally encourage a more thorough *in silico* and *in vivo*
342 evaluation of existing primers, as many are not suitable for DNA metabarcoding due to low base degeneracy,
343 potentially high primer bias or critical design flaws.
344
345
346
347
348
349
350

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

PeerJ Preprints NOT PEER-REVIEWED

351 **Acknowledgements:** FL and VE are supported by a grant of the Kurt Eberhard Bode foundation to FL. We thank
352 Bianca Peinert for her help with primer development, and Edward Wilcox and ScienceExchange for HiSeq sequencing.
353
354 **Declaration of interest:** The authors report no conflicts of interest. The authors alone are responsible for the content
355 and writing of the paper.
356
357
358 **Figures**
359
360 **Figure 1**: Selection of potential COI primer sets for DNA metabarcoding of insects, targeting the Folmer region. Primer
361 pairs are shown based on typically used / suggested combinations used in the literature, but also other combinations are
362 possible. Table S1 gives an overview of the exact primer sequences and sources.
363
364 **Figure 2**: Overview of the principle behind the PrimerMiner package for sequence downloading and clustering. Both
365 mitochondrial genomes as well as partial gene sequences are downloaded and clustered, to make utilise the maximum
366 of available sequence information while minimising biases introduced by overrepresented taxa in the sequence data.
367 Primer trimming is necessary if database sequences have not been properly trimmed.
368
369 **Figure 3**: Comparison of the COI Folmer primer performance and the four tested primer combinations newly developed.
370 All primer combinations were tested with the same ten bulk samples each containing 52 morphologically distinct
371 macroinvertebrate taxa. The 52 taxa are shown on the x-axis with the relative number of reads obtained for each
372 morphotaxon by black dots on the logarithmic y-axis (mean read abundance indicated by red circles), for each
373 respective primer combination. Sequence abundance was normalized across the ten replicates and the amount of tissue
374 used in each DNA extraction. Only OTUs with a minimum read abundance of 0.003% in at least one of the ten samples
375 were included in analyses. Number of samples for which a morphotaxon was not detected is indicated by orange and
376 red numbers in each plot. A thick vertical line in light red indicates if a morphotaxon was not detected.
377
378 **Figure 4**: **Preliminary data, error penalties subject to changes / kind of mismatch not jet implemented!** Overview
379 of *in silico* evaluation of primer performance using PrimerMiner with OTU data from 29 insect orders. Primer
380 performance is shown for each group in pie charts (red = failure, green = working, grey = missing data / gaps). Every
381 primer sequence match with a mismatch penalty score of above 50 is considered a failure. Every order with at least 100
382 OTUs is used for calculation of the average and the box plot showing the mean penalty scores for each group. Good

383 primers have a yellow background.

384 **References**

385 Brandon-Mong, G.J., Gan, H.M., Sing, K.W., Lee, P.S., Lim, P.E. & Wilson, J.J. (2015). DNA metabarcoding of
386 insects and allies: an evaluation of primers and pipelines. *Bulletin of Entomological Research*, **105**, 717–727.

387 Cameron, S., Rubinoff, D. & Will, K. (2006). Who Will Actually Use DNA Barcoding and What Will It Cost?
388 *Systematic Biology*, **55**, 844–847.

389 Clarke, L.J., Soubrier, J., Weyrich, L.S. & Cooper, A. (2014). Environmental metabarcodes for insects: in silicoPCR
390 reveals potential for taxonomic bias. *Molecular ecology resources*, **14**, 1160–1170.

391 Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014). DNA metabarcoding and the cytochrome
392 c oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**, 20140562–20140562.

393 Dowle, E.J., Pochon, X. & Banks, J.C. (2015). Targeted gene enrichment and high-throughput sequencing for
394 environmental biomonitoring: a case study using freshwater macroinvertebrates. *Molecular Ecology*.

395 Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, **10**,
396 996–998.

397 Edgar, R.C. & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing
398 reads. *Bioinformatics*, **31**, 3476–3482.

399 Elbrecht, V. & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing
400 Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol (M. Hajibabaei,
401 Ed.). *PloS one*, **10**, e0130324–16.

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

404 Ficetola, G.F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P. & Pompanon, F. (2010). An in
405 silico approach for the evaluation of DNA barcodes. *BMC genomics*, **11**, 434.

406 Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial
407 cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology*,
408 **3**, 294–299.

409 Geller, J., Meyer, C., Parker, M. & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome coxidase
410 subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular ecology resources*, **13**, 851–
411 861.

412 Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W. & Hajibabaei, M.
413 (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods
414 through DNA metasystematics. *Proceedings of the National Academy of Sciences*, **111**, 8007–8012.

415 Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. & Baird, D.J. (2011). Environmental Barcoding: A Next-Generation
416 Sequencing Approach for Biomonitoring Applications Using River Benthos. *PloS one*.

417 Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H. & Hallwachs, W. (2004). Ten species in one: DNA barcoding
418 reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National*
419 *Academy of Sciences of the United States of America*, **101**, 14812–14817.

420 Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence
421 alignment based on fast Fourier transform. *Nucleic acids research*, **30**, 3059–3066.

422 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S.,
423 Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. (2012). Geneious Basic: An integrated and
424 extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**,
425 1647–1649.

426 Larsen, S., Pace, G. & Ormerod, S. (2011). Experimental effects of sediment deposition on the structure and function of

PeerJ Preprints | NOT PEER-REVIEWED

427 macroinvertebrate assemblages in temperate streams. *River Research and Applications*, **27**, 257–267.

428 Leray, M. & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of
429 marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*,
430 201424997–6.

431 Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. & Machida, R.J. (2013). A new
432 versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan
433 diversity: application for characterizing coral reef fish gut contents. *Frontiers in zoology*, **10**, 1–1.

434 Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., Zhang, H., Misof, B., Kjer, K.M., Tang, M., Niehuis, O., Jiang, H. &
435 Zhou, X. (2015). Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics
436 biodiversity analysis. *Molecular ecology resources*.

437 Meusnier, I., Singer, G.A., Landry, J.-F., Hickey, D.A., Hebert, P.D. & Hajibabaei, M. (2008). A universal DNA mini-
438 barcode for biodiversity analysis. *BMC genomics*, **9**, 214.

439 Misof, B., Liu, S., Meusemann, K. & Peters, R.S. (2014). Phylogenomics resolves the timing and pattern of insect
440 evolution. ….

441 O'Donnell, J.L., Kelly, R.P., Lowell, N.C. & Port, J.A. (2016). Indexed PCR Primers Induce Template-Specific Bias in
442 Large-Scale DNA Sequencing Studies (A.R. Mahon, Ed.). *PloS one*, **11**, e0148698–11.

443 Piñol, J., Mir, G., Gomez-Polo, P. & Agustí, N. (2014). Universal and blocking primer mismatches limit the use of
444 high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular ecology resources*,
445 1–12.

446 Ratnasingham, S. & Hebert, P.D.N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index
447 Number (BIN) System (D. Fontaneto, Ed.). *PloS one*, **8**, e66213–16.

448 Schnell, I.B., Bohmann, K. & Gilbert, M.T.P. (2015). Tag jumps illuminated - reducing sequence-to-sample
449 misidentifications in metabarcoding studies. *Molecular ecology resources*, **15**, 1289–1303.

450 Sharma, P. & Kobayashi, T. (2014). Are 'universal' DNA primers really universal? *Journal of applied genetics*, **55**,
451 485–496.

452 Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., Golding, G.B. & Hajibabaei, M.
453 (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq
454 platform. *Scientific Reports*, **5**, 9687.

455 Shokralla, S., Zhou, X., Janzen, D.H., Hallwachs, W., Landry, J.-F., Jacobus, L.M. & Hajibabaei, M. (2011).
456 Pyrosequencing for Mini-Barcoding of Fresh and Old Museum Specimens (S.-O. Kolokotronis, Ed.). *PloS one*, **6**,
457 e21252.

458 Stein, E.D., Martinez, M.C., Stiles, S., Miller, P.E. & Zakharov, E.V. (2014). Is DNA Barcoding Actually Cheaper and
459 Faster than Traditional Morphological Methods: Results from a Survey of Freshwater Bioassessment Efforts in the
460 United States? (M. Casiraghi, Ed.). *PloS one*, **9**, e95525.

461 Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012). Environmental DNA. *Molecular Ecology*, **21**, 1789–
462 1793.

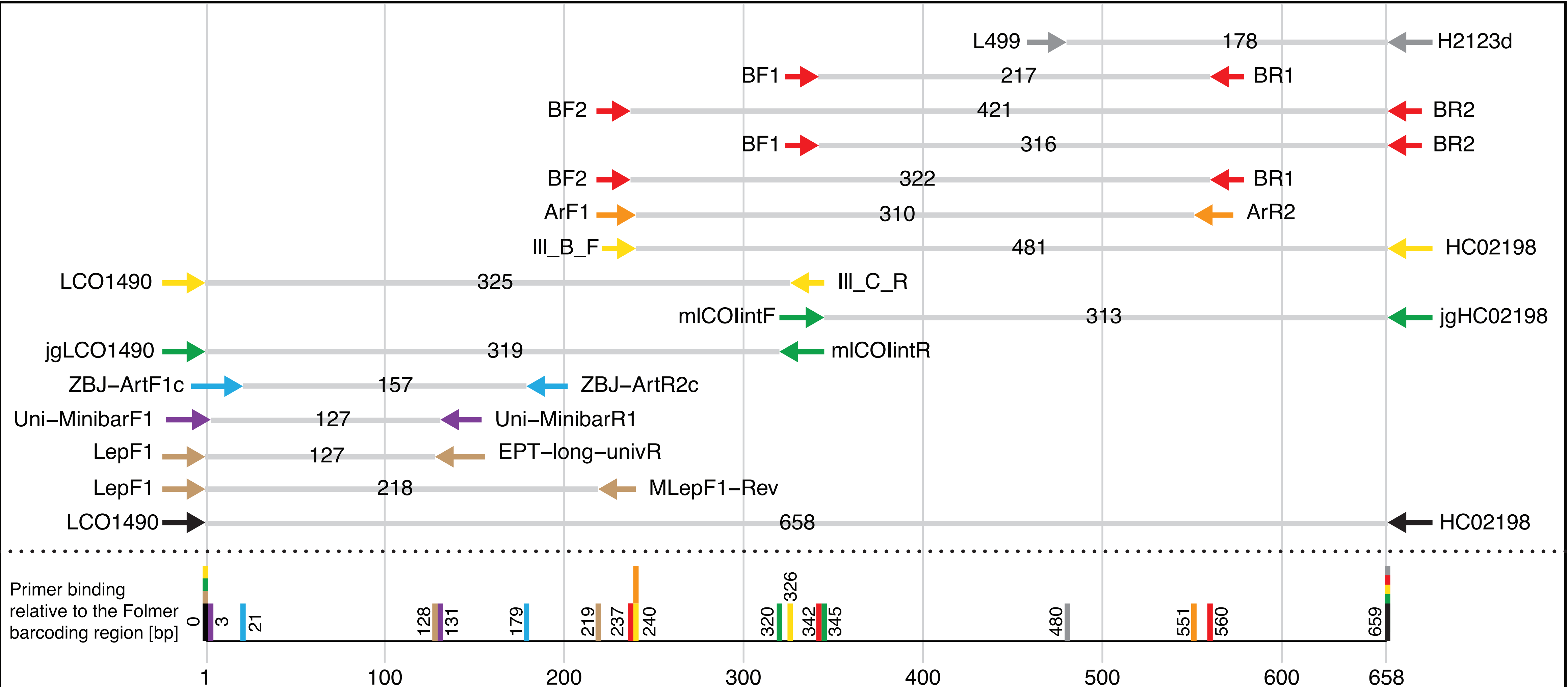
463 Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. & Rozen, S.G. (2012). Primer3--new
464 capabilities and interfaces. *Nucleic acids research*, **40**, e115–e115.

465 Van Houdt, J.K.J., Breman, F.C., Virgilio, M. & De Meyer, M. (2010). Recovering full DNA barcodes from natural
466 history collections of Tephritid fruitflies (Tephritidae, Diptera) using mini barcodes. *Molecular ecology resources*,
467 **10**, 459–465.

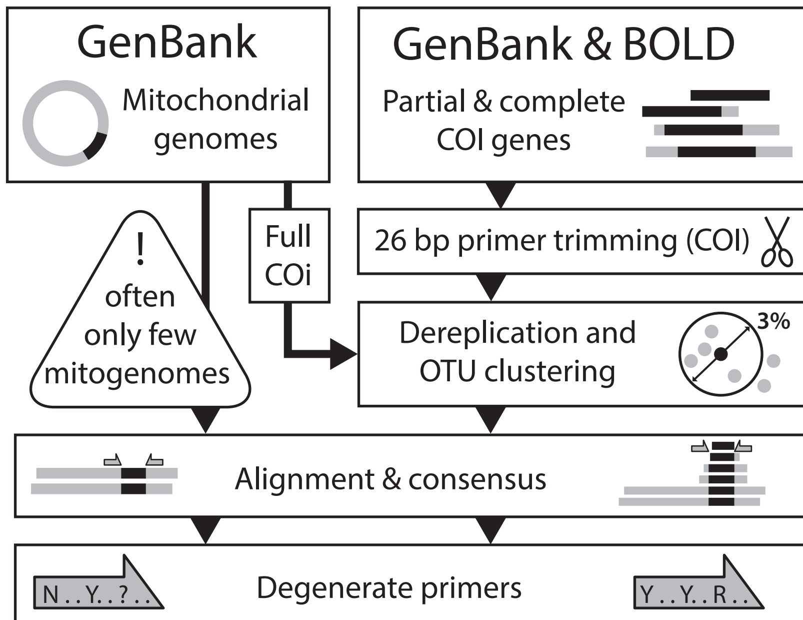
468 Zeale, M.R.K., Butlin, R.K., Barker, G.L.A., Lees, D.C. & Jones, G. (2010). Taxon-specific PCR for DNA barcoding
469 arthropod prey in bat faeces. *Molecular ecology resources*, **11**, 236–244.

470

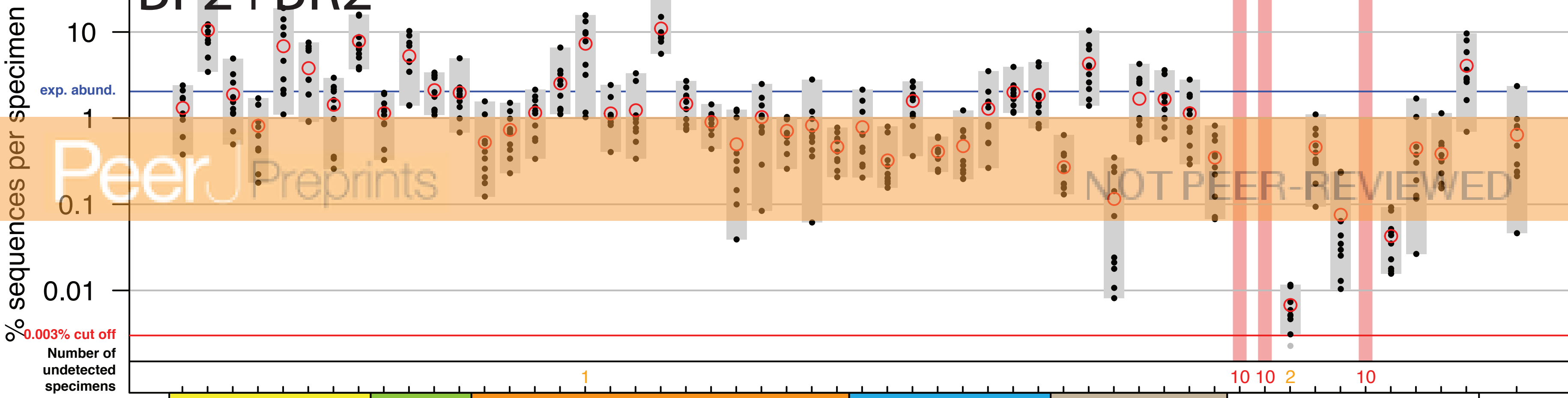
PeerJ Preprints NOT PEER-REVIEWED

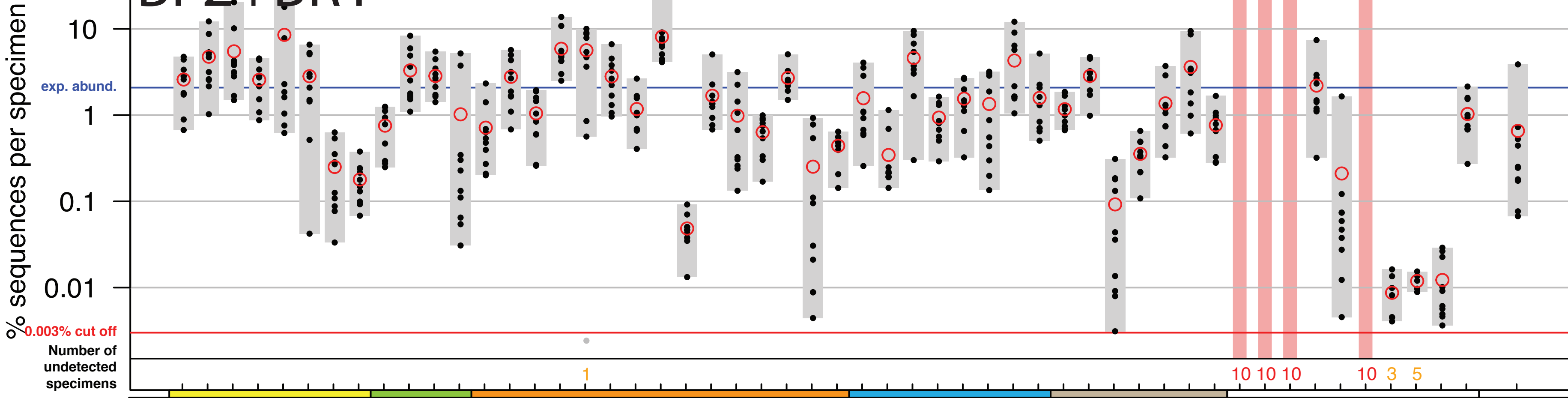


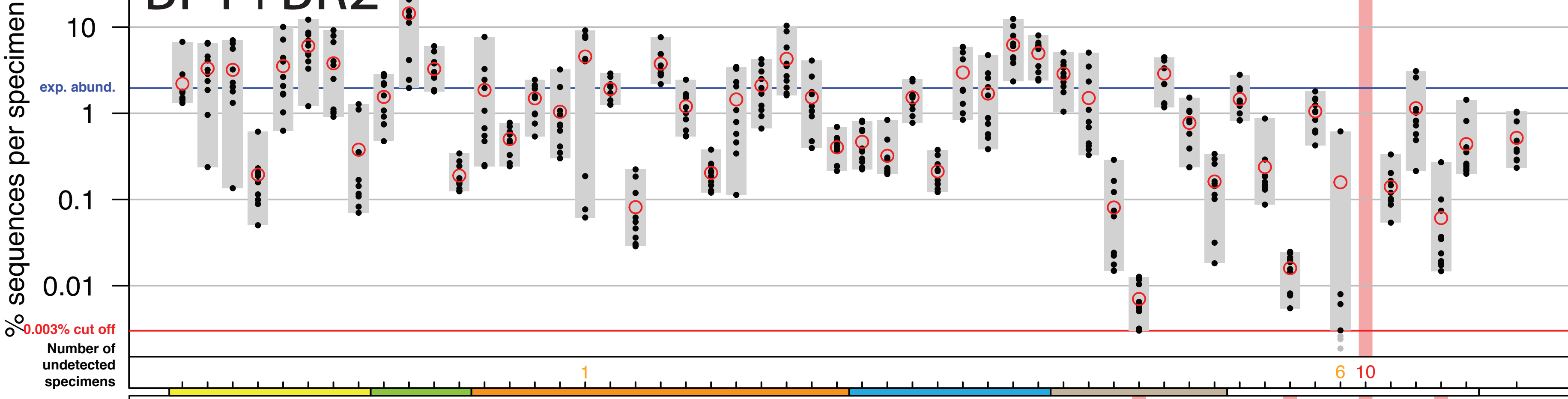
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

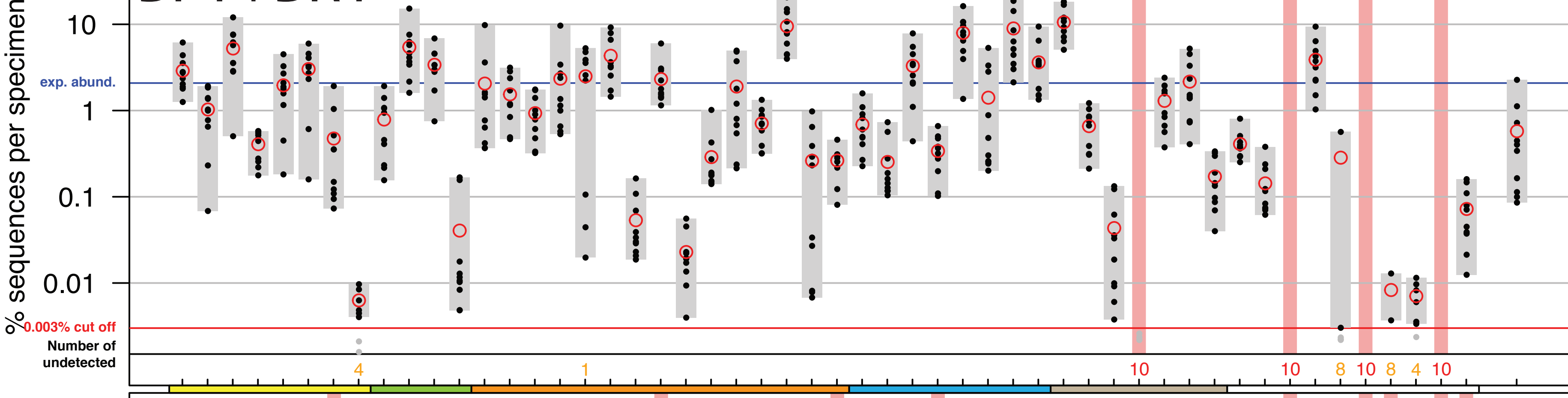

GenBank
Mitochondrial genomes
GenBank & BOLD
Partial & complete COI genes
! often only few mitogenomes
Full COi
26 bp primer trimming (COI)
Dereplication and OTU clustering
3%
Alignment & consensus
Degenerate primers
N . . Y. . ? . .
Y . . Y. . R . .

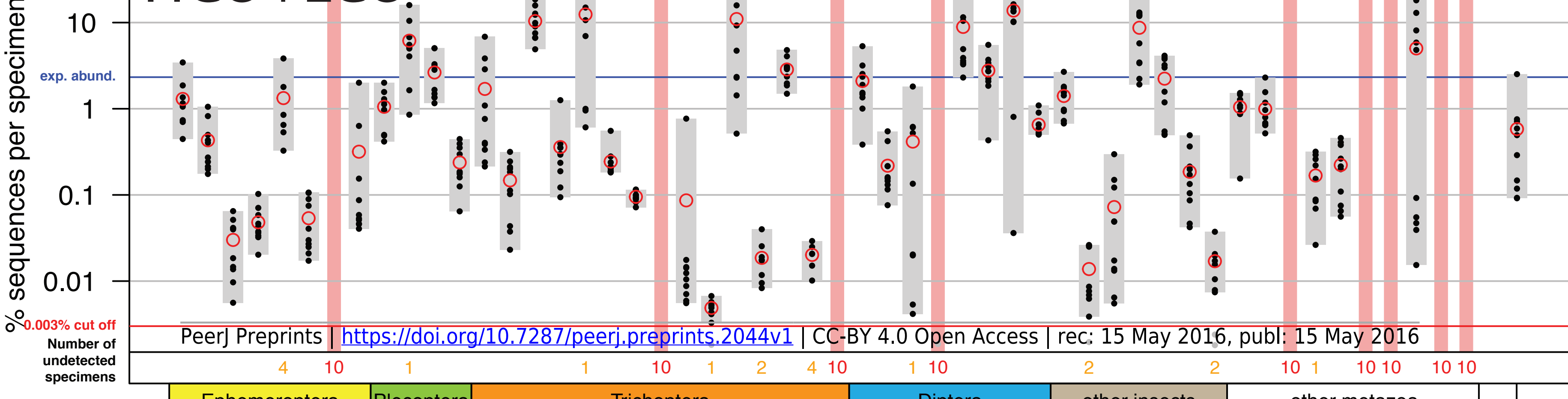
PeerJ Preprints NOT PEER-REVIEWED











PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016

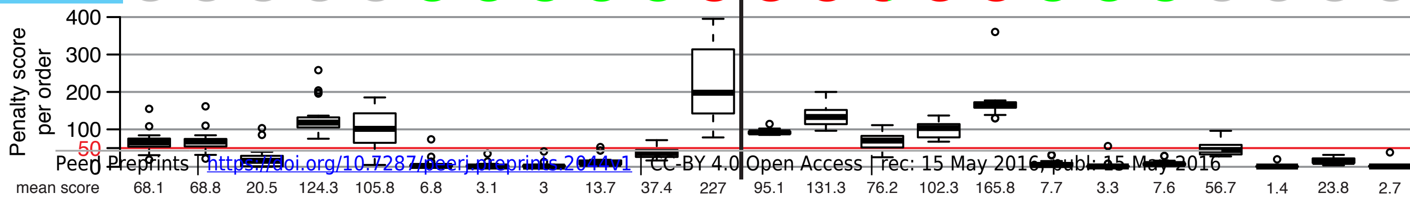
PeerJ Preprints
NOT PEER-REVIEWED

Forward Primers:

Number of OTUS	Order:	LCO1490	LepF1	jgLCO1490	Uni-MinibarF1	ZBJ-ArtF1c	BF2	ArF5	III_B_F	BF1	mlCOIintF	L499
137	Archaeognatha	70	70	100	30	30	100	100	100	98	51	3
5	Zygentoma	33	33	100	67	0	100	100	100	100	100	25
1555	Odonata	60	60	80	7	11	99	99	99	99	78	0
1621	Ephemeroptera	60	45	100	15	56	100	100	100	100	74	4
5	Zoraptera	33	33	100	67	0	100	100	100	100	100	25
60	Dermaptera	100	100	100	0	100	100	100	100	98	81	0
999	Plecoptera	33	17	75	8	39	100	100	100	100	79	13
2993	Orthoptera	50	35	88	11	22	98	96	96	98	68	1
1	Mantophasmatodea	0	0	0	0	0	100	100	100	100	0	0
2	Grylloblattodea	100	100	100	0	0	100	100	100	100	100	0
81	Embioptera	50	50	100	0	0	100	100	100	96	71	0
130	Phasmatodea	56	50	72	11	5	100	100	100	91	46	1
645	Mantodea	25	25	100	0	100	100	100	100	100	77	0
1280	Blattodea	60	80	100	0	15	100	100	100	100	61	0
275	Isoptera	21	20	99	0	4	100	98	98	100	69	0
1070	Thysanoptera_edited	44	39	83	0	6	59	60	60	96	63	0
15592	Hemiptera	53	53	73	10	29	96	98	98	99	61	4
1033	Psocodea	100	100	100	0	6	99	99	99	99	82	0
63450	Hymenoptera	63	64	85	2	20	78	99	98	97	73	0
130	Raphidioptera	56	50	72	11	5	100	100	100	91	46	1
108	Megaloptera	46	46	85	0	59	100	100	100	100	59	9
929	Neuroptera	67	74	93	14	56	100	100	100	100	71	1
117	Strepsiptera	33	33	33	0	0	99	100	100	92	59	0
35058	Coleoptera	48	46	82	13	40	100	100	100	100	75	12
5544	Trichoptera	33	33	50	0	53	100	100	100	100	81	1
105880	Lepidoptera	92	92	94	13	87	100	100	100	100	88	0
158	Siphonaptera	100	100	100	0	33	98	98	98	100	86	0
148	Mecoptera	100	100	100	25	80	98	98	98	100	82	43
78027	Diptera	90	92	96	23	57	100	100	100	100	79	20
	Average all insects	47	44	86	8	23	97	98	98	98	70	5
	Freshwater groups	48	45	85	3	53	100	100	100	100	75	9

Reverse Primers:

Number of OTUS	Order:	EPT-long-univR	Uni-MinibarR1	ZBJ-ArtR2c	MLepF1-Rev	mlCOIintR	III_C_R	ArR5	BR1	HC02198	jgHCO2198	H2123d	BR2
137	Archaeognatha	19	19	34	35	2	98	96	91	62	99	91	99
5	Zygentoma	0	25	0	0	0	100	100	100	33	100	33	100
1555	Odonata	13	2	40	42	2	99	100	97	53	100	92	100
1621	Ephemeroptera	22	11	25	11	7	100	100	89	47	100	84	100
5	Zoraptera	0	25	0	0	0	100	100	100	33	100	33	100
60	Dermaptera	7	10	36	5	7	98	100	90	100	100	100	100
999	Plecoptera	6	5	33	13	14	100	99	93	36	100	93	100
2993	Orthoptera	8	4	46	50	2	98	97	95	53	100	96	99
1	Mantophasmatodea	0	100	100	0	0	100	100	100	0	100	100	100
2	Grylloblattodea	0	0	0	50	0	100	100	100	0	100	0	100
81	Embioptera	0	9	35	16	0	100	96	100	41	100	98	100
130	Phasmatodea	10	2	41	33	0	95	100	91	83	100	100	100
645	Mantodea	20	14	48	30	0	100	98	98	70	100	94	99
1280	Blattodea	12	9	35	42	1	100	99	97	18	100	94	100
275	Isoptera	7	7	21	55	1	99	100	97	8	100	95	100
1070	Thysanoptera_edited	10	3	56	18	0	98	99	92	85	95	92	91
15592	Hemiptera	21	4	39	28	3	99	98	94	73	100	95	99
1033	Psocodea	11	11	43	14	3	99	100	89	59	100	75	100
63450	Hymenoptera	12	2	52	16	1	98	98	92	86	99	98	99
130	Raphidioptera	10	2	41	33	0	95	100	91	83	100	100	100
108	Megaloptera	4	7	69	13	5	100	100	96	100	100	100	100
929	Neuroptera	12	8	61	24	2	100	100	96	92	100	100	100
117	Strepsiptera	2	0	64	28	0	97	28	90	84	100	100	98
35058	Coleoptera	7	11	46	16	2	100	100	96	73	100	96	100
5544	Trichoptera	21	8	66	12	1	100	100	93	82	100	96	100
105880	Lepidoptera	5	9	81	13	0	100	100	94	90	99	98	99
158	Siphonaptera	0	5	75	2	0	100	99	94	81	100	97	100
148	Mecoptera	4	2	72	10	11	100	98	99	95	100	100	100
78027	Diptera	9	6	67	17	1	100	99	95	79	100	98	100
	Average all insects	11	7	49	25	3	99	96	94	64	100	96	99
	Freshwater groups	12	7	50	17	5	100	100	94	71	100	96	100



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2044v1 | CC-BY 4.0 Open Access | rec: 15 May 2016, publ: 15 May 2016