PrimerMiner: an R package for development and *in silico* validation of DNA metabarcoding primers

Vasco Elbrecht [1]*, Florian Leese [1,2]

Affiliations:

1) Aquatic Ecosystem Research, Faculty of Biology, University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen, Germany

2) Centre for Water and Environmental Research (ZWU) Essen, University of Duisburg-Essen, Universitätsstraße 2, 45141 Essen, Germany

*Corresponding author: vasco.elbrecht@uni-due.de, phone: +49.201-1834053

Abstract

1) DNA metabarcoding is a powerful tool to assess biodiversity by amplifying and sequencing a standardized gene marker region. Its success is often limited due to variable binding sites that introduce amplification biases. Thus the development of optimized primers for communities or taxa under study in a certain geographic region and/or ecosystems is of critical importance. However, no tool for obtaining and processing of reference sequence data in bulk that serve as a backbone for primer design is currently available.

2) We developed the R package PrimerMiner, which batch downloads DNA barcode gene sequences from BOLD and NCBI databases for specified target taxonomic groups and then applies sequence clustering into operational taxonomic units (OTUs) to reduce biases introduced by the different number of available sequences per species. Additionally, PrimerMiner offers functionalities to evaluate primers *in silico*, which are in our opinion more realistic then the strategy employed in another available software for that purpose, ecoPCR.

3) We used PrimerMiner to download cytochrome c oxidase subunit I (COI) sequences for 15 important freshwater invertebrate groups, relevant for ecosystem assessment. By processing COI markers from both databases, we were able to increase the amount of reference data 249-fold on average, compared to using complete mitochondrial genomes alone. Furthermore, we visualized the generated OTU sequence alignments and describe how to evaluate primers *in silico* using PrimerMiner.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

4) With PrimerMiner we provide a useful tool to obtain relevant sequence data for targeted primer development and evaluation. The OTU based reference alignments generated with PrimerMiner can be used for manual primer design, or processed with bioinformatic tools for primer development.

Key words: Primer development, primer evaluation, primer bias, ecosystem assessment, *in silico* PCR, data mining, DNA barcoding, high throughput sequencing, monitoring

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

Introduction

Metabarcoding allows rapid DNA-based species identification from environmental samples often containing hundreds or thousands of specimens. DNA is extracted in bulk, a barcoding marker amplified and sequenced with high throughput sequencing followed by bioinformatic processing (Taberlet *et al.* 2012). However, typically used marker regions like the cytochrome c oxidase subunit I (COI) fragment for animals, often show highly variable primer binding sites (Deagle *et al.* 2014; Sharma & Kobayashi 2014) leading sometimes to substantial primer bias and species remaining undetected (Piñol *et al.* 2014; Elbrecht & Leese 2015). Therefore, we argue that for optimal performance metabarcoding primers should be optimised to the specific taxonomic groups of interest in the geographic region / ecosystem under study. Many primers used in metabarcoding studies, however, are often not thoroughly evaluated or were developed on a limited reference dataset or mitochondrial genomes (Geller *et al.* 2013; Deagle *et al.* 2014; Brandon-Mong *et al.* 2015) or large but often not order / family specific datasets of partial marker sequences (Zeale *et al.* 2010; Leray *et al.* 2013; Gibson *et al.* 2014). However, downloading all relevant reference data from the respective databases can also introduce biases. A key problem is that certain taxa are often overrepresented with hundreds of sequences, e.g. from sequences available from phylogeographic studies. This artificial inflation for specific taxa can in principle be circumvented when using only mitochondrial genomes. Yet, such datasets are often very limited because mitochondrial genomic sequences are still rare for many taxonomic groups. However, obtaining good quality reference data is essential for primer development and evaluation. Another drawback when developing primers is that available *in silico* primer evaluation tools such as ecoPCR v1.01 (Ficetola *et al.* 2010) are based on in our opinion oversimplified models, which do not take into account the position and type of mismatches. However, as mismatches at or near the 3' end of the primer have a much greater effect than other mismatches (Stadhouders *et al.* 2010; Piñol *et al.* 2014) it is important to focus specifically on these. To address these concerns in DNA metabarcoding primer development we have developed the R package PrimerMiner, which allows for batch downloading and processing of sequences for specified groups. The software also has additional alignment visualisation capabilities and novel *in silico* primer evaluation capabilities.

Package description

PrimerMiner a batch sequence downloader

PrimerMiner is an R based package that batch downloads sequences data from NCBI and BOLD and clusters them into operational taxonomic units (OTUs) using Vsearch (Figure 1, https://github.com/torognes/vsearch). Sequence data for a chosen genetic marker is downloaded for each specified taxonomic group. Target sequences are also extracted from

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

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 into OTUs using a 3% sequence similarity by default. OTU consensus sequences are saved in a fasta file for each taxonomic group and can then be aligned and used for manual or software-based primer design. This clustering strategy utilized in PrimerMiner has several key advantages:

1) Overrepresented taxa and duplicated sequences are merged into few OTUs.
2) Taxonomic variation within a species can be retained (wobble bases), while rare haplotypes are ignored in the OTU consensus sequences.
3) Highly diverse (potentially cryptic) 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 or family was identified correctly.

For batch downloading sequence data PrimerMiner needs two files, one configuration file that can be generated by running `batch_config("config.txt")` in R and a taxonomy table specifying the groups for which data should be downloaded (csv table, see Table S1). The default options in the configuration file are optimised for the COI barcoding marker, but other marker genes can be downloaded as well (e.g. `Marker = "16S"`). At this point only COI markers are downloaded from BOLD and marker extraction is limited to mitochondrial genomes but support for additional markers and chloroplast genomes is planned. By default the majority consensus sequence for each OTU is saved for each group, but OTU diversity can be captured by including wobble bases in the consensus sequences (e.g. `threshold = 0.1`). PrimerMiner is optimised to download data on order level (fist table column, Table S1) with the option to specify a subset of taxa on family level in the second column if, e.g. only aquatic Coleoptera (beetles) families should be targeted. Downloading data for taxonomic groups above family level rank (genus or species level) can cause problems, if taxonomic names are used across several groups. By running `batch_download("taxa_table.csv", "config.txt")` matching barcode sequences are downloaded and processed. This can take a few hours depending on dataset size, but PrimerMiner will keep you updated and report once it has finished in the R console.

To visualize the downloaded sequence data, OTUs from each group have to be aligned and trimmed to the target marker region. We deliberately did not automate this process as part of the PrimerMiner strategy, as we highly recommend to visually inspect the alignments and remove misaligned sequences and gaps. Different markers or groups typically need specific attention because databases can contain errors or alignment parameters need to be adjusted, thus we refrain from automating this step at this moment. OTUs can be aligned with tools like MAFFT (Katoh *et al.* 2002) as implemented in

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

Geneious (Kearse *et al.* 2012) and gaps in the alignment can be removed using the "Strip Alignment Columns" feature in Geneious. We generally consider ~100 OTUs for each order as a minimum coverage to capture its variability of primer binding sites and select necessary wobble bases. Once the alignments are generated and exported as fasta files, they can be visualized using the `plot_alignments(c("file1.fasta", "file2.fasta"))` function in PrimerMiner to manually search for new and evaluate existing primers (Figure 2). These alignments can then also be used to develop primers with automated tools such as Primer3 (Untergasser *et al.* 2012) or ecoPCR (Ficetola *et al.* 2010).

To test the PrimerMiner package we have downloaded and processed COI sequence data for 15 important freshwater macroinvertebrate groups. The alignments are available as sample data as part of the PrimerMiner package and are visualised in Figure 2. Compared to using mitochondrial genomes alone sequence coverage of the 15 groups was increased 249-fold (SD = 395) on average by including COI OTU sequences generated with PrimerMiner (number of OTUs for each group divided by available mitochondrial genomes). Additionally, in Figure 2 we have added a few commonly used barcoding and metabarcoding primers binding in the Folmer LCO primer binding site (Folmer *et al.* 1994). Some of these primers show a high number of mismatches due to the lack of degeneracy (LCO1490, Folmer *et al.* 1994, ZBJ-ArtF1c, Zeale *et al.* 2010), or in the case of the Uni-MinibarF1 primer even strong mismatches on the 3' end of the primer (Meusnier *et al.* 2008). These problematic primers likely introduce biases in metabarcoding studies (Piñol *et al.* 2014; Elbrecht & Leese 2015) illustrating how critical careful primer evaluation is not only when designing novel primers but also when choosing primers from the literature.

In silico primer evaluation

PrimerMiner also introduces powerful *in silico* primer evaluation capabilities, allowing the evaluation of single primers and primer pairs on any given sequence alignment. Unlike ecoPCR (Ficetola *et al.* 2010) PrimerMiner factors in the adjacency, position and type of each mismatch between primer and template sequence. This is important because amplification success depends highly on good matches at the 3' end of the primer (Stadhouders *et al.* 2010; Piñol *et al.* 2014). Using the command `evaluate_primer("Alignment.fasta", "Primersequence", bind_start, bind_stop, save="filename.csv", mm_position ="Position_v1.csv", adjacent=2, mm_type="Type_v1.csv")` PrimerMiner calculates an individual penalty score for each template to primer mismatch in a table format, which can then be further analysed.

PrimerMiner evaluates these parameters in the following order:

1) Scoring of mismatches based on position (based on the provided table in `mm_position`).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

2) Adjustment of mismatch types (based on the provided table in `mm_type`).

3) Increasing of penalties for two adjacent mismatches (by default penalty scores are duplicated `adjacent=2`).

4) Reduction of penalty scores if wobble bases are present in the target sequence, which partially match the primer.

Penalty scores for position and mismatch type are fully customisable by providing your own penalty tables. Example mismatch type and mismatch position scoring tables are included in the example data of the PrimerMiner package. Two penalty score tables can be evaluated with `primer_threshold(table1.csv, table2.csv, threshold=100)` using a defined threshold under which a primer pair is considered not amplifying. As detailed penalty scores are provided for each sequence and primer combination more elaborate statistical analysis is easily possible.

Package availability, documentation and sample data

The PrimerMiner package and example sequence alignments (figure 2) are available on github (https://github.com/VascoElbrecht/PrimerMiner), alongside extensive documentation in the GitHub wiki including YouTube video tutorials. After downloading the package can be installed with `install.packages("path_to_file", repos = NULL, type="source", dependencies=T)` in R. As the PrimerMiner is bundled with the Vsearch v1.10.2 binary, it cannot be made available on CRAN. At this time a windows version is not available, because for windows only Usearch v8.1.1861 is available, which in contrast to Vsearch produces poor quality alignments of OTU sequences (Edgar 2013).

Conclusions

We here described the first release of the package PrimerMiner, a useful tool to obtain reference sequence data for targeted primer development and evaluation using its alignment visualisation capabilities as well as refined *in silico* evaluation functions. A thorough *in silico* and *in vivo* evaluation of existing primers is needed, as many primers might not be suitable for DNA metabarcoding due to low base degeneracy, potentially high primer bias or critical design flaws. Additionally, developing primers specific to ecosystems, geographic areas and taxonomic groups of interest is encouraged to maximize performance.

Acknowledgements: FL and VE are supported by a grant of the Kurt Eberhard Bode foundation to FL. We thank Jan Mache and Edith Vamos for proofreading and two anonymous reviewers for their comments comments which improved

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

this manuscript.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

Figures

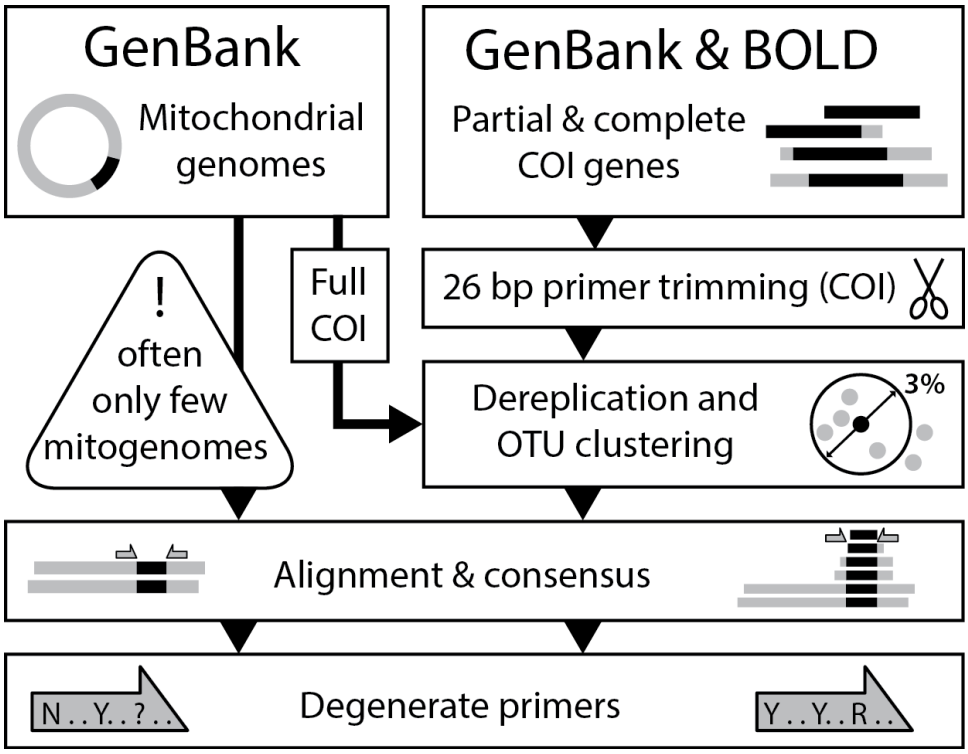

Figure 1: Overview of the principle behind the PrimerMiner package for sequence downloading and clustering. Both mitochondrial genomes as well as partial gene sequences are downloaded and clustered to make maximum use of the available sequence information while minimising biases introduced by overrepresented taxa in the sequence data. Primer trimming is necessary if primers have not been removed from all sequences in the database.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

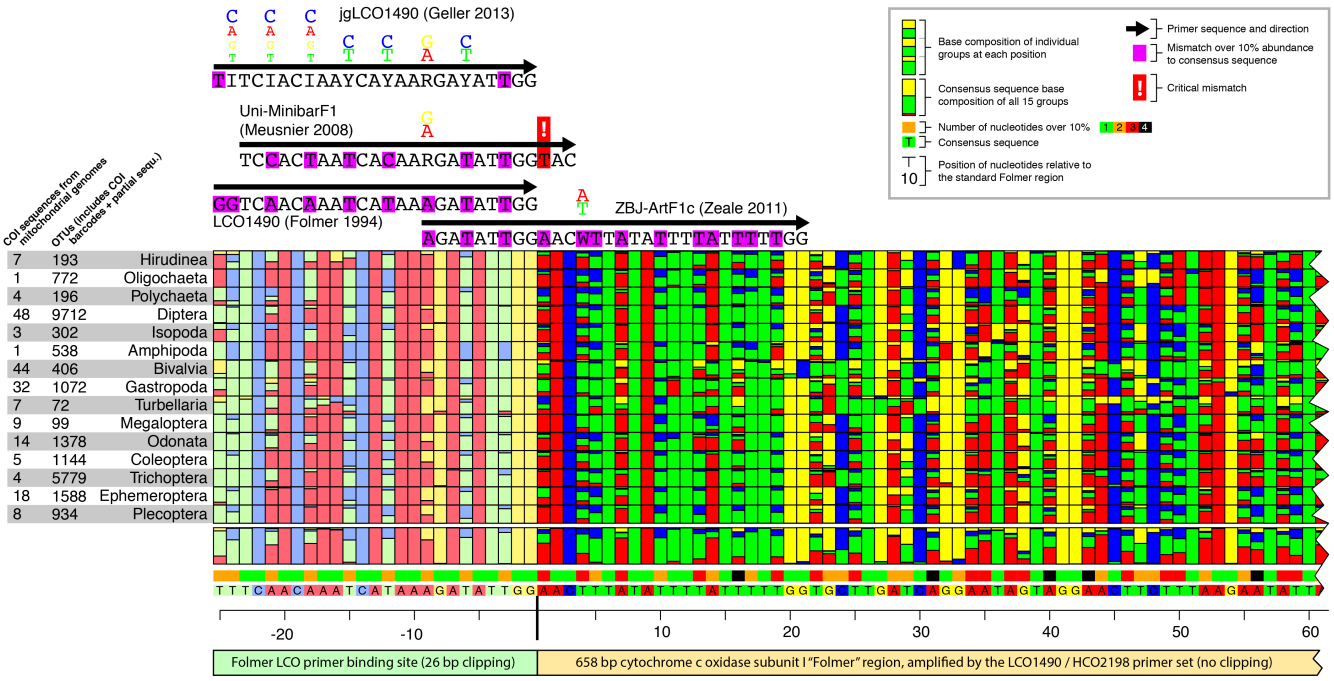

Figure 2: Example base composition plot generated with PrimerMiner for the COI Folmer region of 15 important freshwater invertebrate groups. The sequences for the Folmer binding regions (opaque colours) have been downloaded in February 2015 and had 26 bp clipping applied, as many clusters were affected by sequences which still contained the primer sequences. Sequences from the actual Folmer region were downloaded April 2015 not trimmed, as only the region amplified by the primers was used (fasta OTU sequences of the Folmer region are available as sample data in the PrimerMiner package). Additionally, common primers from the literature were added to the plot and mismatches above 10% indicated with pink colour (Folmer *et al.* 1994; Meusnier *et al.* 2008; Zeale *et al.* 2010; Geller *et al.* 2013).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

References

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

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**, 20140562–20140562.

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

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 (M. Hajibabaei, Ed.). *PloS one*, **10**, e0130324–16.

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

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.

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

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

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

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., 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.

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

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*, 1–12.

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

Stadhouders, R., Pas, S.D., Anber, J., Voermans, J., Mes, T.H.M. & Schutten, M. (2010). The Effect of Primer-Template Mismatches on the Detection and Quantification of Nucleic Acids Using the 5′ Nuclease Assay. *The Journal of Molecular Diagnostics*, **12**, 109–117.

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

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. & Rozen, S.G. (2012). Primer3--new

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016

capabilities and interfaces. *Nucleic acids research*, **40**, e115–e115.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2352v1 | CC BY 4.0 Open Access | rec: 15 Aug 2016, publ: 15 Aug 2016