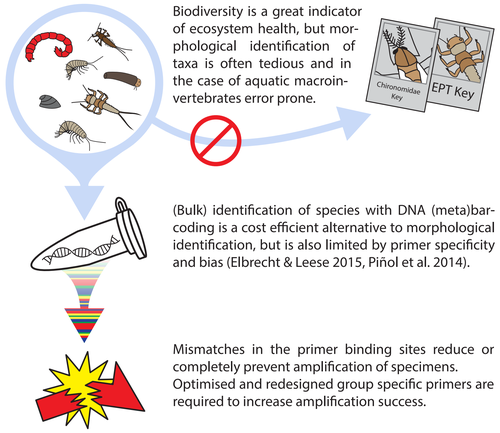
R package: PrimerMiner

Optimising primers and designing mini barcodes based on partial sequences

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

[1]Ruhr University Bochum, Germany - Department of Animal Ecology, Evolution and Biodiversity
[2]University of Duisburg and Essen, Germany - Aquatic Ecosystem Research - GeneStream.de

The PrimerMiner story!



Biodiversity is a great indicator of ecosystem health, but morphological identification of taxa is often tedious and in the case of aquatic macroinvertebrates error prone.

(Bulk) identification of species with DNA (meta)barcoding is a cost efficient alternative to morphological identification, but is also limited by primer specificity and bias (Elbrecht & Leese 2015, Piñol et al. 2014).

Mismatches in the primer binding sites reduce or completely prevent amplification of specimens. Optimised and redesigned group specific primers are required to increase amplification success.

The problem: Primer redesign requires extensive sequence data, but there is no comprehensive tool available to batch download sequence data.

The R script PrimerMiner helps you to batch download sequence information for specified groups, which then can be used to generate DNA barcoding primers or mini-barcodes for applications like DNA & eDNA metabarcoding.

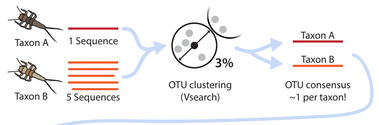
Here is how it works:

A PrimerMiner batch downloads sequences from NCBI and BOLD, for any specified marker (e.g. COI or 16S).

NCBI + BOLDSYSTEMS

B The target marker is also extracted from partial and complete mitochondrial sequences found on NCBI.

C All sequences are clustered (default 3% similarity), so species which are well (or over-) represented in the databases are not biasing the dataset.



geneious

The sequences are then exported as fasta files, and can be aligned (e.g. with Geneious and MAFFT). The resulting alignments can be visualized with PrimerMiner for custom Primer design (recommended) or other primer design software can be used.

The solution: PrimerMiner is a free tool to batch download and cluster sequences for specified groups, providing the ideal dataset to design primers.

What's happening under the hood and why?

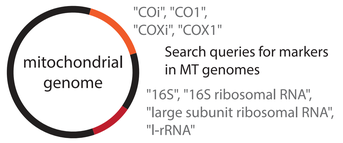
A Sequences are downloaded from BOLD and NCBI using their respective APIs.
- BOLD mostly contains standard barcoding markers like Cytochrome Oxidase I.
- You can add your own sequences to the PrimerMiner pipeline.

NCBI BOLDSYSTEMS

B When mitochondrial genomes are available on NCBI, the target marker is extracted if present in these sequences, even if they are only partial genomes (> 2000 bp).

Mitochondrial genomes are downloaded as GenBank files and markers found using the different spelling styles provided in the config.txt file.

By default 100 bp are extracted left and right from the target marker to account in mistakes in the annotations.



PRC **primers** are needed to copy the target barcoding marker for DNA sequencing.

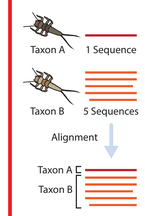
Universal primers are supposed to target a wide reange of organisms, but missmatches in the 3' primer end often lead to unequal amplification and even taxa being not detected (Elbrecht & Leese et al. 2015, Piñol et al. 2014).

With **OTU clustering** sequences of one marker are merged by similarity into **O**perational **T**axonomic **U**nits.

Each OTU stands for one putuative species, but diverse species or sequences with wrong taxonomy or sequencing errors can produce two or more OTUs.

C Sequences are clustered to reduce the influence of well represented taxa on the dataset. This is done with Vsearch.
https://github.com/torognes/vsearch

Downloading data with out clustering.

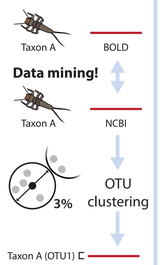


Sequences from taxon B are overrepresented!

However, for primer design each taxon should contribute the same amount of sequences.

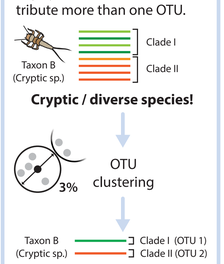
Clustering of sequences solves the problem of overrepresentated sequences and has additional advatages outlined below!

Removes duplicates (records with identical haplotypes).



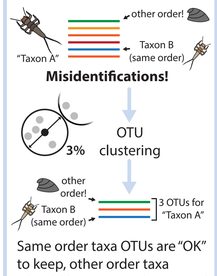
Identical sequences are deduplicated!

Each Taxon is represented the most common sequence. Highly diverse taxa can contribute more than one OTU.



Both clades are represented with one consensus OTU each!

Clustering taxon name independent, misidentification not (as big of) an issue.



Same order taxa OTUs are "OK" to keep, other order taxa (should be rare) can be manually detected in the alignments.

Quick start: it's easy!

You need two files:
- A csv table containing the groups which sequences should be downloaded for (e.g. "taxa.csv"). If you want only specific subgroups of e.g. Coleoptera, you can specify these in the second column.
- A configuration file, wich can be created with R and then modified (e.g. to select another marker).

Just run in R: `batch_config("config.txt")` in R

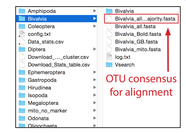
config.txt

```
# Configuration file for batch download

# General settings
Taxon_table = "path/to/csv/name_of_table.csv" # csv containing taxa to download
Taxon_sep = "," # table entries seperated by ,
Marker = c("COi", "CO1", "COXi", "COX1") # specify target gene
Download = T # if FALSE, no data will be downloaded
Merge_and_Cluster_data = T # if set to FALSE, sequences are not merged / clustered

# NCBI download, see ?Download_GB for details
download_GB = T
merge_GB = T
maxlength_GB = 2000
custom_query_GB = NULL
clipping_left_GB = 0
clipping_rigth_GB = 0

# Mitochondria download, see ?Download_mito and ?Mito_GB2fasta for details
download_mt = T
merge_mt = T
minlength_mt = 2001
maxlength_mt = 88888
custom_query_mt = NULL
clipping_left_mt = 0
clipping_rigth_mt = 0
add_mt = 100
rm_dup = T
no_marker = T

# BOLD download, see ?Download_BOLD for details
download_bold = T
merge_bold = T
clipping_left_bold = 0
clipping_rigth_bold = 0

# Clustering sequences, see ?Clustering for details
vsearchpath = "Vsearch"
id = 0.97
cmd = NULL
threshold = "Majority"

# Write summary statistics
summstats = T
```

Order	Family
Plecoptera	
Ephemeroptera	
Trichoptera	
Coleoptera	
	Dryopidae
	Dytiscidae
	Elmidae
	Georissidae
	Gyrinidae
	Haliplidae
	Helophoridae
	Hydraenidae
	Hydrochidae
	Hydrophilidae
	Hygrobiidae
	Noteridae
	Psephenidae
	Scirtidae
	Spercheidae
Odonata	
Megaloptera	
Turbellaria	
	Dendrocoelidae
	Dugesiidae
	Planariidae

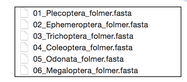
To start the batch sequence download in R run:
`batch_download("taxa.csv", "config.txt")`

PrimerMiner will take ~30-60 minutes depending on dataset size. Fasta files containing a consensus sequence for each taxon are created, as well as summary statistics for sequences downloaded and clustered.



The consensus sequences for each putative taxon can then be aligned and used for Primer development and validation of existing primers.
We recommend aligning mitochondiral marker sequences with MAFFT, and then map all marker sequence against the mitochondrial consensus sequence with **Map to reference** using Geneious.

PrimerMiner can visualize the alignments to aid finding good universal primers (see picture below). Alignments in a folder can be plotted with the following function:
`plot_alignments("path/to/folder")`



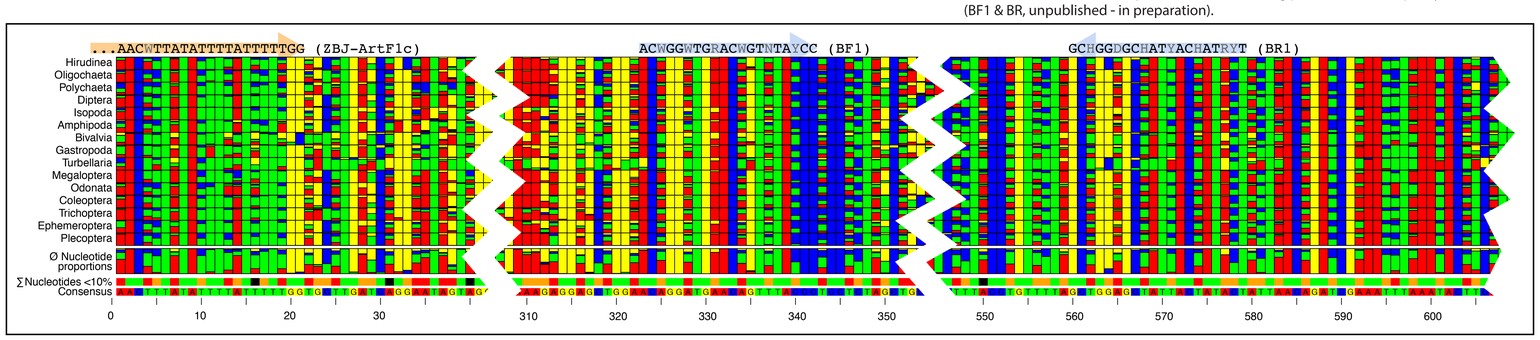
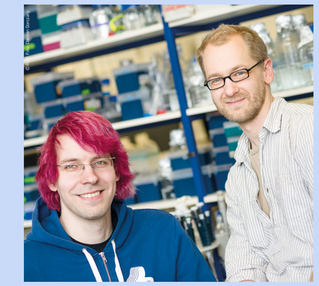


Figure 1: PrimerMiner can be used to validate existing primers and design new ones. The example below shows the base proportions for the standard DNA Barcoding gene (Cytochrome oxidase 1 Folmer regoin, Folmer et al. 1994) for 15 assessment relevant freshwater invertebrate orders. For orders like Coleoptera only sequences from families which can be found in freshwater were downloaded (see Table above).
We used the diagram for validation of existing primers, like the mini barcoding primers from Zeale et al. 2011 (here only ZBJ-ArtF1c is shown). This primer set would potentially have problems to detect Amphipoda and Bivalvia in gut contents due to common mismatches at the 3'end. We also developed new metabarcoding primers for water quality assessment (BF1 & BR, unpublished - in preparation).



Vasco Elbrecht (PhD student) and **Florian Leese** are developing laboratory protocols & bioinformatics for assessing stream health with DNA based methods.

Tweet **@luckylionde @leese_lab**
E-Mail **Vasco.Elbrecht@rub.de**

https://goo.gl/ATu51x

Learn more about stream assessment with metabarcoding and problems with primer bias in this 4 minute **video**, or visit our **talk on Thursday** in room Rozanski 104 session 5 (4:15 pm).

GitHub

The PrimerMiner R package is open source and available on GitHub
http://www.github.com/primerMiner

System requirements
- A Mac or Linux system (clustering is done with an UNIX based script)
- A good internet connection

References:

- **Elbrecht & Leese** (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. PLoS ONE
- **Folmer, Black, Hoeh, Lutz & Vrijenhoek** (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates.
- **Piñol, Mir, Gomez-Polo & Agustí** (2014). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods.
- **Zeale, Butlin, Barker, Lees & Jones** (2011). Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces.

Feedback and question corner!

Take a post-it and write a comment or question. If you include your e-mail address I will come back to you.

Feedback?
Questions?