Standard Operating Procedure for the genetic identification of crustacean species using cytochrome *c* oxidase I and *16S* rRNA gene segments

SEATRACES
CONSORTIUM
2020

CONTENTS

BACKGROUND

Illegal fisheries, fraud and mislabelling are issues representing a serious risk to the existence of this important economic activity for the Atlantic Area Regions. In order to address these issues, the international project SEATRACES ('Smart Traceability and Labelling Toolbox for a Sustainable Seafood production') of the Interreg Atlantic Area program aims to demonstrate to stakeholders and consumers that labelling and traceability are important for the protection and valorisation of Atlantic Area's fisheries and aquaculture. Within the project new, validated and standardized methods and tools are being developed to facilitate a harmonised control.

SCOPE

This document describes a procedure for the identification of single crustacean species in foodstuffs to the genus or species level based on Sanger sequencing.
The methodology has been tested to distinguish a variety of different commercially relevant and other crustacean species. The method is designed to work on fresh, frozen, and processed samples. It is not suitable for mixtures of samples containing DNA of different species. For some closely related species with relatively recent divergence, the method may only be able to determine the sample to genus level.

ABBREVIATIONS

DNA: Deoxyribonucleic acid
PCR: Polymerase chain reaction
SOP: Standard Operating Procedure
COI: Mitochondrial cytochrome *c* oxidase I gene

PRINCIPLE OF THE METHOD

DNA is extracted using a suitable method for crustaceans. Two mitochondrial gene fragments, a 658 bp cytochrome *c* oxidase I (COI) (Lobo et al., 2013) and an approximately 520 bp *16S* fragment (Palumbi et al., 1991) are amplified. The amplification of both fragments is necessary as in some cases, the simple use of only COI may not be sufficient as it may not work for all crustacean species and there may be a lack of sufficient COI data for some species. It has been noticed in sequence alignments that COI nucleotide sequences were ambiguous for some genera, e.g. *Metapenaeus* spp. and *Heterocarpus* spp., since either sequences were identical to other species or sequences of the same species differed substantially. From our laboratory experience, the COI primers may not work for some species such as *Liocarcinus holsatus*, *L. depurator*, *Trachypenaeus* spp., and *Portunus pelagicus*. Inconsistent amplification success was observed for *Cancer pagurus, Squilla mantis* and some Brachyura species.
The primers used for the amplification of these two sections are universal primers. In the further procedure, the PCR amplicons/products are sequenced in order to determine the nucleotide sequences. The evaluation of the determined sequences is carried out by homology search through comparing the query sequences with reference sequence entries in public databases, which enables a taxonomic determination with assignment of the sequence to a species or genus level on the basis of the degree of identity.

MATERIALS AND EQUIPMENT

Reagents and solutions

- Molecular grade water

DNA extraction

QIAamp DNA Mini Kit (Qiagen, Hilden), alternative DNA extraction methods may be used; PCR

- Thermostable Taq Polymerase (Hot-start)
- 10x PCR-Reaction buffer (including or separate $MgCl_2$ and deoxyribonucleoside triphosphate mix. *c*= 10 mmol/l)
- Primers

Primer name	DNA sequence of oligonucleotide	Reference
LoboF1	5'- KBTCHACAAAYCAYAARGAYATHGG-3'	Lobo *et al.*, 2013
LoboR1	5'- TAAACYTCWGGRTGWCCRAARAAYCA-3'	Lobo *et al.*, 2013
16Sar	5'- CGCCTGTTTATCAAAAACAT-3'	Palumbi *et al.*, 1991
16Sbr	5'- CCGGTCTGAACTCAGATCACGT-3'	Palumbi *et al.*, 1991

- Agarose
- Suitable DNA length standard for assessing the amplification product length

Commercial kits

- QIAamp DNA Mini Kit (Qiagen, Hilden) recommended or equivalent; CTAB has also been successfully tested and the Wizard DNA Clean up system (Promega) was successfully used by some laboratories in the pre-test.

Plastic

- Pipette tips (containing protective filters for PCR)
- 1.5 ml reaction tubes
- PCR tubes

Other materials

- Sterile dissection equipment
- Disposable plastic gloves

Equipment

- Precision pipettes 1-1000 µL
- Microcentrifuge
- Centrifuge, at least 13,000 rpm
- Vortex shaker
- Thermal shaker
- UV-spectrophotometer or fluorometer, to determine the concentration of DNA
- Thermocycler
- Gel electrophoresis device
- Gel documentation system
- DNA sequencer or Sequencing company

Electronic files/computer software

- A computer with a text editor e.g. notepad.
- Freely available sequence editing software e.g. Bioedit, MEGA, ProSeq.
- Internet access is required to utilise the Barcode of Life System: http://www.boldsystems.org/ and GenBank (blast.ncbi.nlm.nih.gov/Blast.cgi)

PROCEDURES

I. Sample preparation

All samples should be stored frozen at -20 °C until processed. Prior to processing, thaw samples at room temperature.
The test portion used for DNA extraction should be representative of the samples to be analysed. The external surfaces of samples submitted for analysis may have been affected through preservation treatments or bacterial breakdown. In order to minimize the risk of contamination or DNA degradation, samples shall not be taken from the surface of the tissue sample. Remove outer layers of tissue which are in contact with the environment before taking a subsample. Use sterile dissection equipment where appropriate.
For the QIAamp DNA Mini Kit (Qiagen, Hilden), 25 mg wet weight are cut into small pieces and placed into 1.5 mL reaction tubes. Using a commercial kit please refer to manufacturer's instructions on needed quantities and extraction procedures.

II. DNA extraction

For DNA extraction, any suitable extraction method can be used. If using a commercial kit, DNA extraction success should be validated in advance. For example, the QIAamp DNA Mini Kit (Qiagen, Hilden) has been evaluated as suitable application. Using a commercial kit please refer to manufacturer's instructions.
In order to certify no contamination during extraction, it is necessary to add an extraction blank control.

It is recommended to measure DNA concentrations of the sample extractions and adjust the concentration to 10 ng/µl with elution buffer in order to ensure the PCR is successful.

III. PCR Amplification

The method was validated for a total volume of 20 µL per PCR. The reagents shown in Table a) and Table b) should be used for the COI and *16S* PCR, respectively.
Reagents are completely thawed at room temperature and should be centrifuged briefly before usage. A PCR reagent mixture is prepared containing all PCR components in the given concentrations except for the DNA extract. The amount of PCR mixture depends on the total volume per PCR and the total number of the reactions including a sufficient pipetting reserve**.**

Materials:
HotStarTaq DNA Polymerase Kit (Qiagen, Hilden), water (ddH_2O), $MgCl_2$, primer

Procedure:
1. Prepare a sample plan describing the DNA samples being analysed and the serial order in the rack or on plate.
2. Organise your DNA extractions (i.e. defrost, if necessary) according to the plan.
3. Alongside every set of reactions ensure a negative control (non-template DNA), the extraction blank control, and preferentially a positive control (this can be determined internally in each lab) must be included.
4. In the following, a recipe for a mastermix containing all components needed for the PCR, except the template DNA, is listed for one reaction. Each reaction has a total volume of 20 µL. Prepare a mastermix by multiplying each component for the number of samples to be analysed, with addition of an extra 10 % to the total volume of each of the reagent components.

a) COI fragment:
PCR Mastermix per reaction using a total volume of 20 µL:

Reagent components	**Volume per reaction**
Qiagen HotStarTaq-Master Mix *	10 µL
ddH_2O	7.4 µL
Primer-Forward LoboF1 (100 pmol/µL)	0.1 µL / 500 nM
Primer-Reverse LoboR1 (100 pmol/µL)	0.1 µL /500 nM
$MgCl_2$ (25 mM)	0.4 µL
Total volume	**18 µL**

Note*: If not using the Qiagen HotStarTaq-Master Mix, the total $MgCl_2$-concentration in the test must be considered: 2mM.

b) *16S* fragment:
PCR Mastermix per reaction using a total volume of 20 µL:

Reagent components	**Volume per reaction**
Qiagen HotStarTaq-Master Mix Kit	10 µL
ddH_2O	8.8 µL
Primer-Forward 16Sar (100 pmol/µL)	0.1 µL / 500 nM
Primer-Reverse 16 Sbr (100 pmol/µL)	0.1 µL /500 nM
Total volume	**19 µL**

5. Vortex the master mix thoroughly.

6. a) Place 18 µL of the COI master mix or b) 19 µL of the *16S* master mix into every tube/well (can use the same pipette tip during this step).

7. a) Add 2 µl for COI or b) 1 µL for *16S* of DNA template (10 ng/µL) to each tube/well following your sample plan.

8. a) Thermal conditions for the **COI PCR reaction** are described in the following:

- initial denaturation 95°C for 5min
- followed by 40 steps (in some cases 35 cycles are sufficient) of denaturation at 94°C for 30 sec, annealing at 54°C for 90 sec, elongation at 72°C for 60 sec
- final elongation 72°C for 5 min

8. b) Thermal conditions for the ***16S* PCR reaction** are described in the following:

- initial denaturation 95°C for 5min
- followed by 35 steps of denaturation at 94°C for 60 sec, annealing at 55°C for 60 sec, elongation at 72°C for 60 sec
- final elongation 72°C for 5 min

9. Place the tubes/plate in the PCR machine and run the PCR programme.

10. Once completed the PCR reactions can be stored in the fridge at 4°C. For long term storage freezing at -20°C is recommended.

IV. Evaluation of PCR products

The success of the PCR can be assessed by agarose gel electrophoresis. As this method is a standard in molecular biology, it is only briefly described.

Approximately 5 µL of each PCR reaction is separated in an agarose gel (e.g. 2% (w/v)) and evaluated with a gel documentation system. The size of the PCR products can be assessed by the addition of a standard DNA ladder showing bands of a defined length in one of the lanes.
The success of the PCR is evaluated as follows:

- No bands should be visible in the negative and extraction blank control.
- If the PCR product show single bands of the expected size of the respective gene fragment in the agarose gel electrophoresis, the sample may be sequenced.
- If the sample is negative for both targets, an inhibition control should be carried out. This can be done either by dilution of sample DNA or by using an internal inhibition control assay. If there is no inhibition, it may be possible that the primers do not match sufficiently to amplify the target sequence, the DNA extraction may not have been successful, or the DNA was too fragmented.

V. Sequencing of the PCR products

Sequencing of PCR products is carried out according to the method available for the testing laboratory. Usually, Sanger sequencing using fluorescence-labelled dideoxynucleoside triphosphates is being performed to determine the nucleotide sequence. If no suitable equipment is available in the laboratory, sequencing can also be performed by an external service provider. The requirements for the respective sequencing companies may vary. Thus, the specific requirements especially in terms of volume and concentrations of PCR products and primers should be checked in advance. If necessary, PCR products can be purified using commercially available kits.
For sequencing of the COI and *16S* PCR products, the primers used for the generation of the amplicons serve as sequencing primers. For each PCR product, sequencing should be performed with the forward and reverse primers, so two complementary sequences for each gene fragment and sample will be obtained.

VI. Raw data processing

General

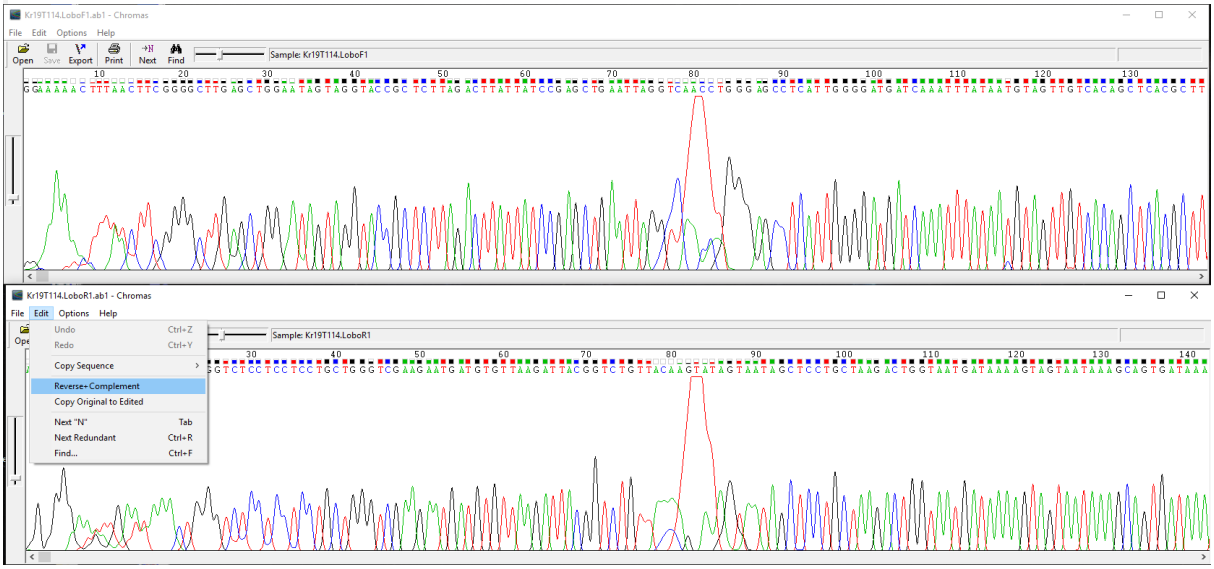
- Chromatograms need to be checked to ensure the sequencing has been successful and quality of the sequencing is satisfying, and the base calling is correct. Approximately 80% of the expected read length should have been determined.
- Misassigned nucleotides to chromatogram peaks should be corrected in the sequence using appropriate software and evaluating the fluorescent peak data.
- The sequence analysis should be performed of both strands. The overlapping or complementary sequences need to be combined to a consensus sequence. This needs to be done to ensure the quality of the sequence and to remove ambiguous bases.
- Primer sequences need to be removed from the sequences before the comparison with gene databases.

Example for raw data processing

In the following section, an example for raw data processing is presented. A further detailed and comprehensive example (for fish) is provided by the Labelfish SOP “STANDARD OPERATING

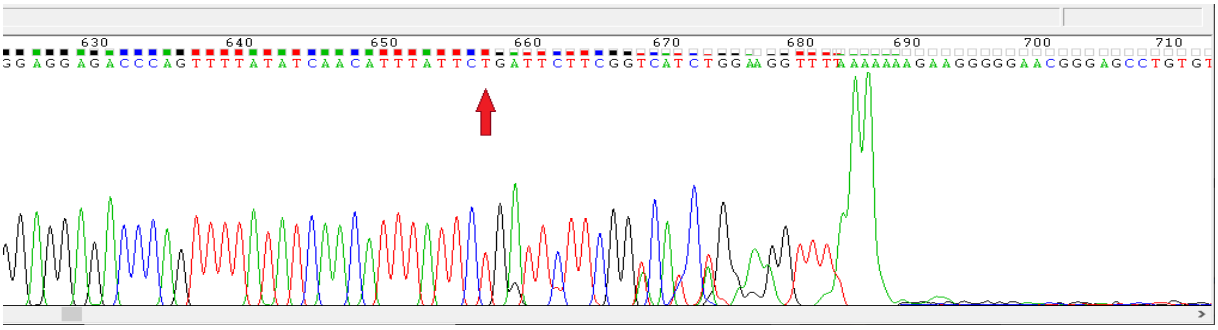
PROCEDURE FOR THE GENETIC IDENTIFICATION OF FISH SPECIES USING DNA BARCODING (MITOCHONDRIAL CYTOCHROME I SEQUENCING)" (http://labelfish.eu/noticia/labelfish-standard-operating-procedure/)

1. Open chromatograms of both strands using appropriate software (e.g. ChromasLite).
2. Reverse complement sequence of reverse strand

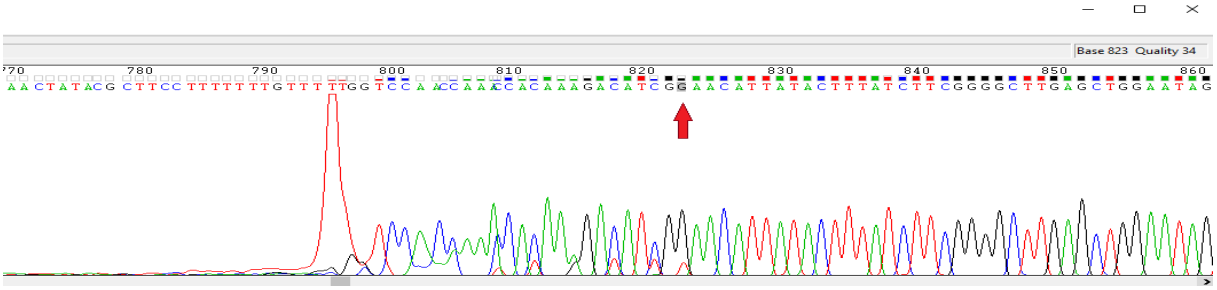


3. Search for primers and delete them

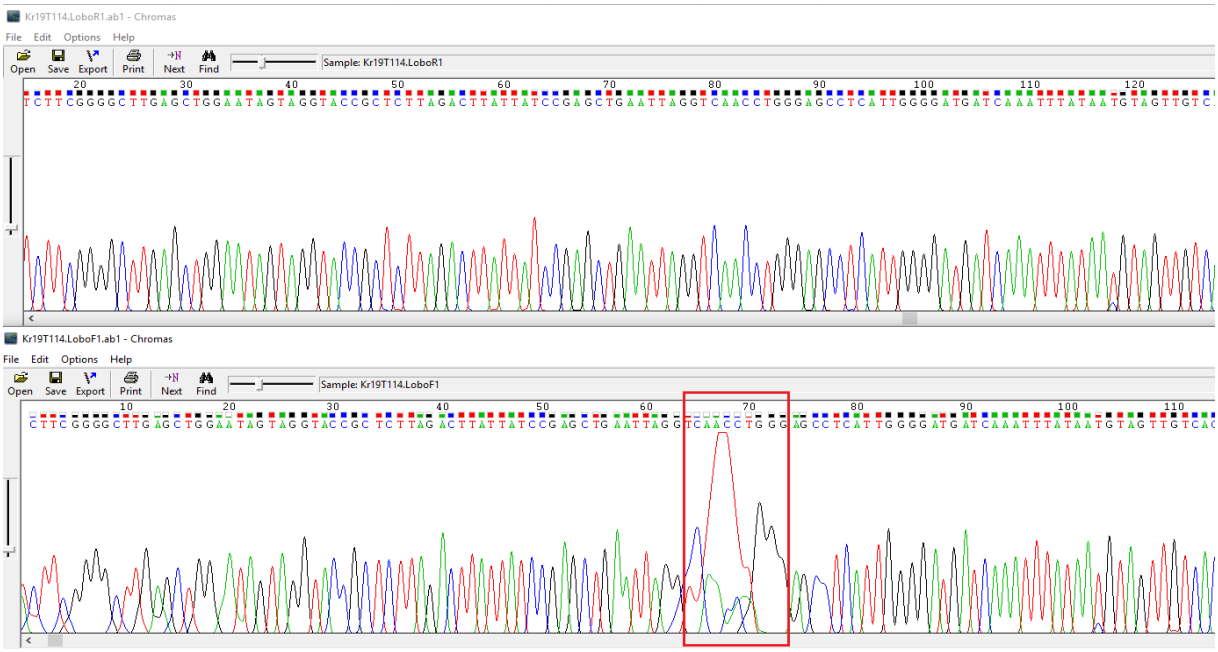
Forward strand



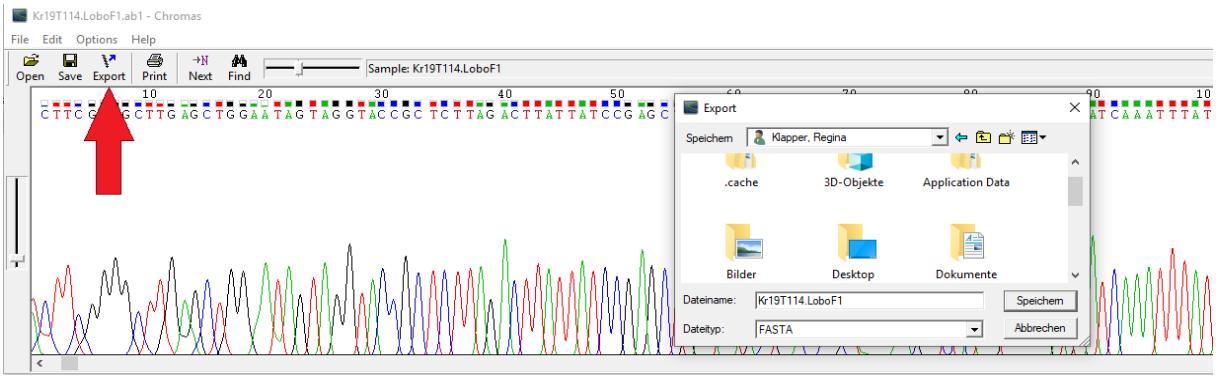
Reverse strand



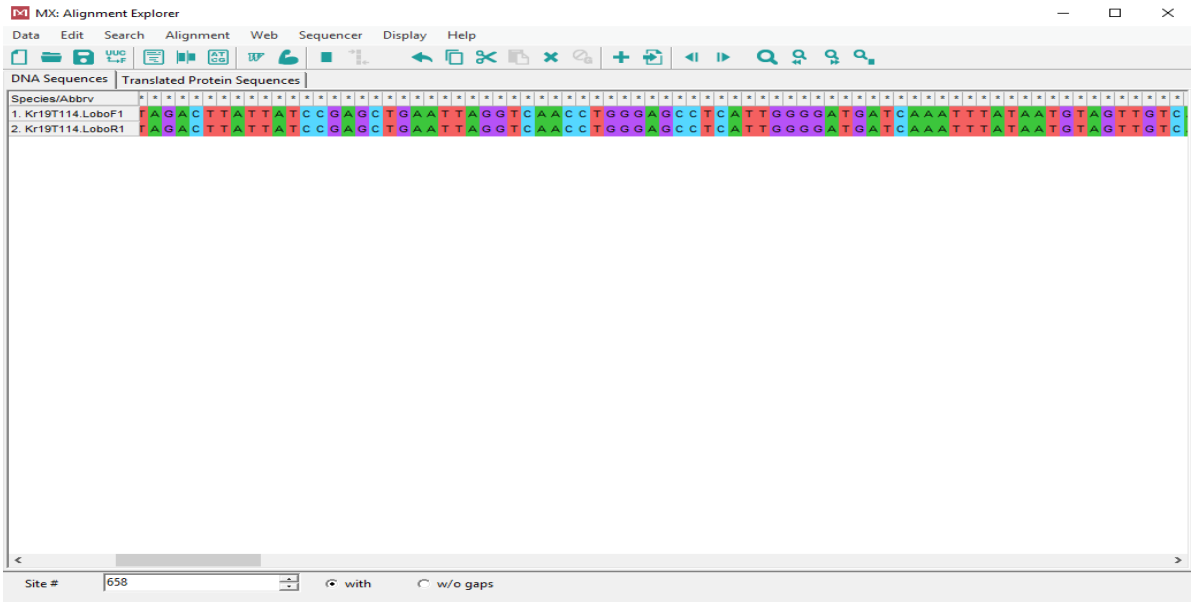
4. Check base calling and peaks, correct misassigned nucleotides



5. Export .fasta-file from chromatogram



6. Align sequences by the use of an appropriate programme (e.g. MEGA, Bioedit)



- Open MEGA> Align> Edit/Build Alignment> Select an Option: Create a new alignment> OK> Are you building a DNA or protein sequence alignment: DNA.
- Insert sequences in .fasta-format either by copy/paste or by Edit> Insert Sequence from File> Choose your .fasta-file
- Click on Alignment> Align by ClustalW> OK (Keep default settings)
- In case not all complementary bases are in concordance as indicated by an asterix, check the chromatogram again

7. Build consensus sequence.

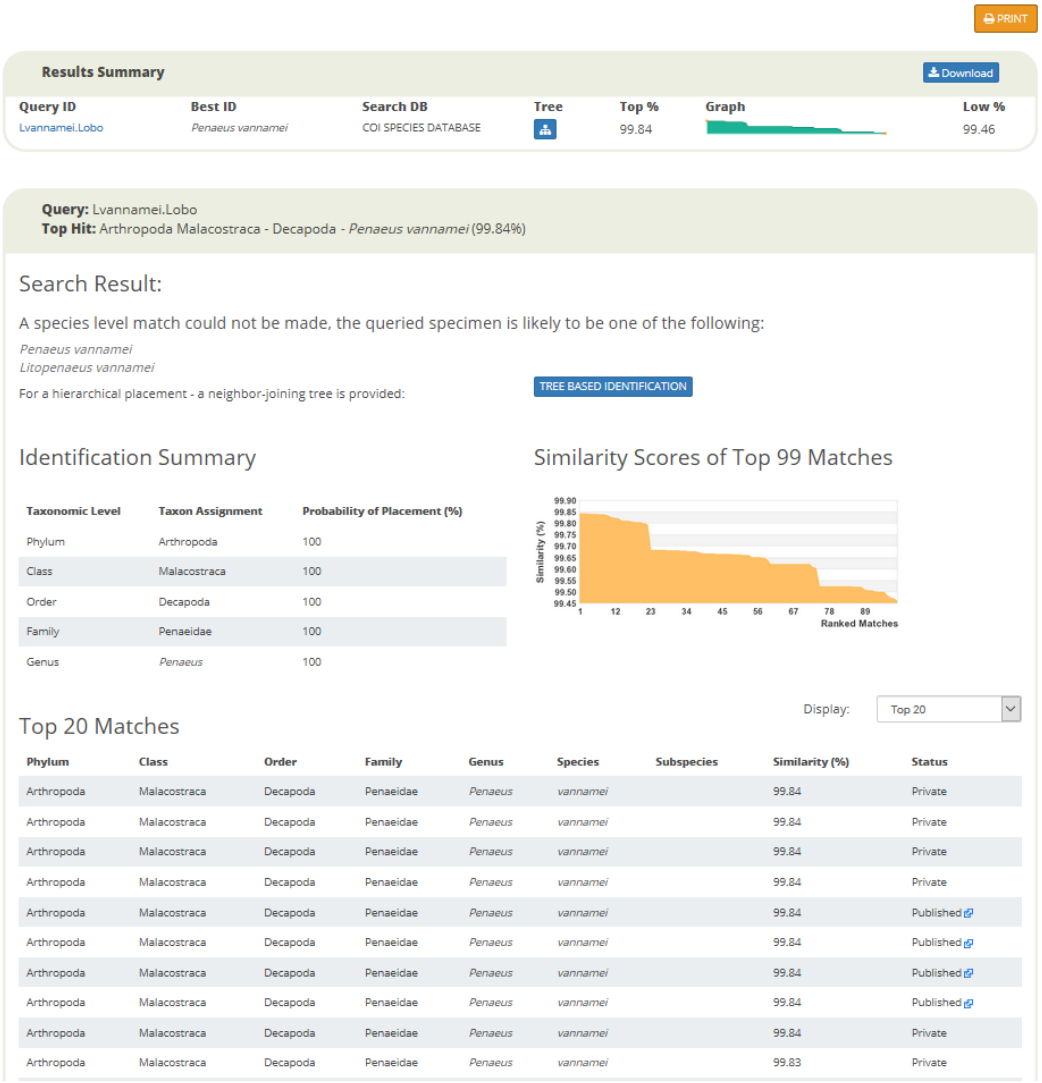
This can be done in Bioedit or alternatively manually. As a result, the final full length barcode is shown in text format (An example with shortened sequence is shown).

>Lvannamei.Lobo
AACATTATACTTTATCTTCGGGGCTTGAGCTGGAATAGTAGGTACCGCTCTTAGACTTAT
TATCCGAGCTGAATTAGGTCAACCTGGGAGCCTCATTGGGGATGATCAAATTTATAATGT
AGTTGTCACAGCTCACGCTTTTGTAATAATTTTTTTTATAGTTATACCAATTATAATTGG
AGGATTTGGTAATTGACTAGTACCTTTAATGTTAGGTGCCCCAGATATAGCCTTCCCTCG
AATGAATAATATAAGCTTCTGGTTATTACCTCCTTCTCTTACATTGCTTTTATCAAGAGG
AATGGTTGAAAGAGGTGTCGGAACCGGATGAACGGTATACCCTCCTTTATCTGCCAGTAT
TGCTCACGCTGGAGCTTCAGTAGATCTTGGAATTTTCTCTCTTCACTTAGCTGGAGTATC
...

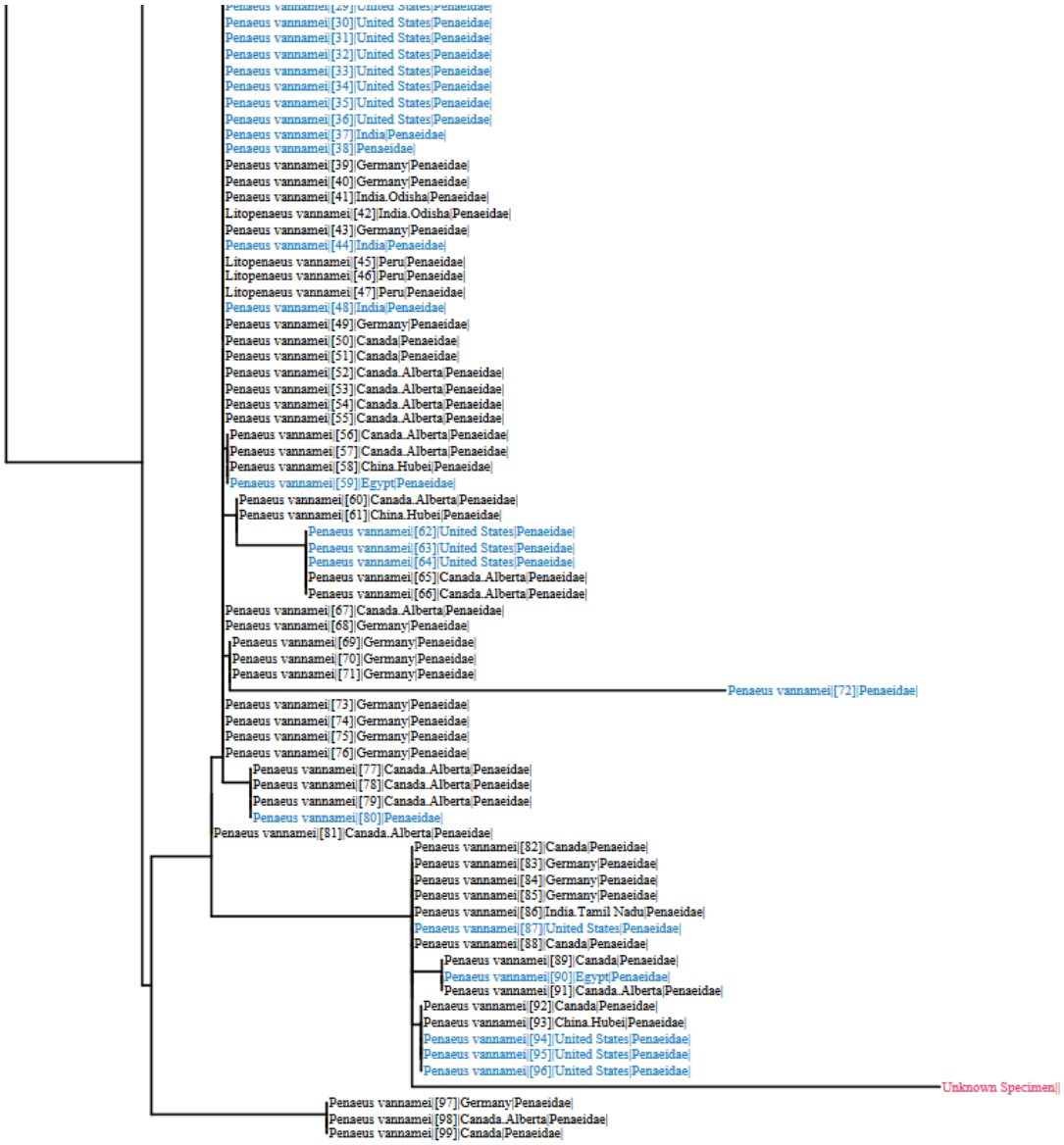
VII. Sequence comparison for sequence identification in public databases

Sequence comparison for sequence identification on the Barcode of Life Database
The Barcode of Life Data System (BOLD) is an online workbench and database that supports the assembly and use of DNA barcode data (COI sequences). The database is freely available and was built by members of the barcoding community (Ratnasingham and Hebert, 2007).

- Go to http://www.boldsystems.org/
- Click on identification>Species Level Barcode Records
- Enter .fasta-formatted sequences in the forward orientation through copy/paste from the Editor to the respective data entry field> Submit
- After a few seconds, the browser will update to the results page showing the closest match including the percentage identity. For documentation, safe the page by choosing the option “print” and safe as .pdf-file.



- The results page provides further information which allows a confident identification from the sequence. In this example, the search result explains that a species level match could not be made, which in BOLD generally means any species that has a sequence record that is 98% similar (or more) will be returned. Often this will just be a single species allowing an unambiguous identification to be made for the sample. In this example, the species name has been changed, so actually a species level match could be obtained.
- The graph “Similarity Scores of Top 99 Matches” shows the percent similarity for each of 99 top matching records in the database against your entered sequence. In the display options, you can also choose to display the Top 99 Matches. In this example the percent similarity of the top 99 matches ranged from 99.84-99.46 %, including solely *Litopenaeus vannamei*.
- Another option is the display of a simple tree to graphically display the results of the homology search. The tree may also be downloaded as .pdf-file. For the clear identification to species level, the entered sequence named as “unknown specimen” clusters only with sequences originating from a single species, so from a monophyletic group.



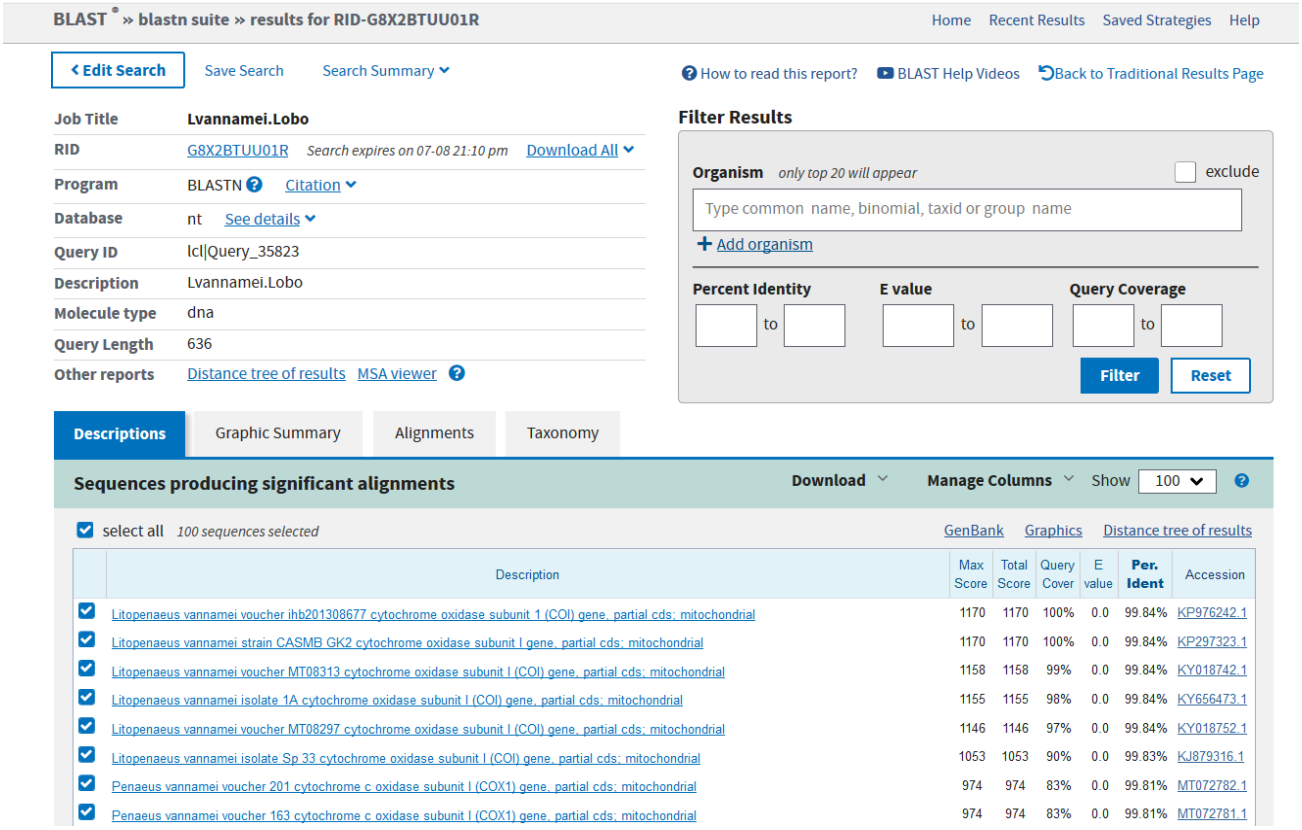
In case of issue interpreting the results, please read the Labelfish SOP. Further general information on DNA barcoding is also provided on the BOLD website http://www.barcodeoflife.org/ or in a handbook about the BOLD database http://www.boldsystems.org/index.php/Resources.

VIII. Sequence comparison for sequence identification on the NCBI GenBank

The NCBI database GenBank is one of the most extensive databases for molecular DNA and protein data. Unlike BOLD, GenBank is not restricted to COI barcoding sequences. Therefore, sequences obtained from other DNA fragments, such as *16S*, need to be examined in GenBank.
For comparison of own sequence data with database entries, the Basic Local Alignment Search Tool (BLAST) is used. It is based on an algorithm which compares entered nucleotide sequences to sequence entries and calculates the statistical significance.

- Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi >Nucleotide BLAST
- Enter your query sequence in the respective data entry field by copy/paste or alternatively by "upload file"

- Keep standard settings: Standard databases, Nucleotide collection, optimize for: Highly similar sequences (megablast). Then click on the BLAST button. In case there are many entries with high similarity, such as tuna sequences, change the number of Max target sequences to a higher number in the menu below “Algorithm parameters” to make sure, all sequences which have a similarity >98% to the query sequence are recognised
- The results of best hits are sorted by maximum score. Re-sort the list by maximum identity.



- In case additional relevant species shall be analysed, an alternative query can be performed, in which main hits or certain taxa are excluded. Go back to the data entry page and change the settings
- For documentation, the results can be saved by clicking on the browser options > print> print as .pdf
- Graphic examination of results can be performed by going to “Distance tree of results”. In the “Taxonomy” section, organisms which are listed in the top matches including the number of hits and the score are listed

Interpretation of database query results

- The reliability and specificity of the results in gene databases depend on the number of sequences in the database and the quality of the sequence. Therefore, it is important to gather information about the taxon you are investigating for additional species in the same genus, presence of declared and related species in GenBank or BOLD, and the number of sequences in the database before entering the sequence query. The specificity may vary between organism groups. The quality and quantity of data entries from crustaceans varies a lot.
- For a secure assignment of crustacean species, sequence identity should be at least 98 % with a sufficient coverage of >90%. However, if a species identification is possible must be decided from case to case. This is dependent on the variation within and between species and the number of sequences.

- An issue may be incorrectly assigned sequences in the databases. If a particular sequence is only entered by the same researchers, the results are considered as unconfirmed. In addition, if the same sequence is assigned to different species and if this is only the case for single entries in comparison to the number of top hits of another species, it has to be suspected that the sample/sequence has not been correctly identified.
- As both presented markers, COI and *16S*, are mitochondrial markers which are maternally inherited, hybrids cannot be identified.

References

Labelfish SOP "STANDARD OPERATING PROCEDURE FOR THE GENETIC IDENTIFICATION OF FISH SPECIES USING DNA BARCODING (MITOCHONDRIAL CYTOCHROME I SEQUENCING)" (http://labelfish.eu/noticia/labelfish-standard-operating-procedure/)

Lobo, J., Costa, P. M., Teixeira, M. A., Ferreira, M. S., Costa, M. H., & Costa, F. O. (2013). Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans. *BMC ecology*, *13*(1), 34.

Palumbi, S. (1991). Simple fool's guide to PCR.

Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular ecology notes*, *7*(3), 355-364.