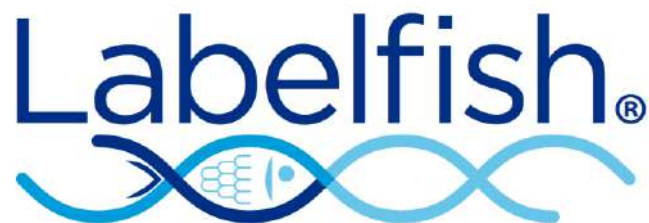


STANDARD OPERATING PROCEDURE FOR THE GENETIC IDENTIFICATION OF FISH SPECIES USING DNA BARCODING (MITOCHONDRIAL CYTOCHROME-C-OXIDASE I SEQUENCING)

Prepared by the Labelfish Consortium



December 2014



CONTENTS

1. BACKGROUND	3
2. PURPOSE	3
3. SCOPE	3
4. DEFINITIONS & ABBREVIATIONS	3
5. PRINCIPLE OF THE METHOD	3
6. MATERIALS & EQUIPMENT	3
6.1 <i>Water</i>	4
6.2 <i>Solutions, standards and reference materials</i>	4
6.3 <i>Commercial kits</i>	4
6.4 <i>Plastic-ware</i>	4
6.5 <i>Equipment</i>	4
6.6 <i>Other materials</i>	4
6.7 <i>Electronic files / computer software</i>	5
7. PROCEDURES	5
7.1 <i>Sample preparation</i>	5
7.2 <i>DNA Extraction</i>	5
7.3 <i>PCR Amplification</i>	6
7.4 <i>PCR Product Check</i>	8
7.5 <i>DNA Sequencing</i>	8
7.6 <i>Raw Data Processing</i>	9
7.7 <i>Generating a consensus sequence</i>	16
7.8 <i>Identifying the species on the Barcode of Life Database</i>	21
7.9 <i>Issues with Interpreting the Species Identification</i>	30
7.10 <i>Quality Assurance</i>	31
8. TROUBLESHOOTING	
8.1 <i>How to interpret the results of the phylogenetic tree</i>	31
8.2 <i>Does the sequences length influence the ability to identify the species?</i>	
8.3 <i>Identifying false identified species in the reference database</i>	33
8.4 <i>Identifying unknown species</i>	38
8.5 <i>Specific problems regarding the identification of Thunnus species</i>	44
9. REFERENCES	47

1. BACKGROUND

The Labelfish project is an EU InterReg funded network of laboratories in the “Atlantic Area” of Europe, aiming to develop harmonised & standardise methods for the authentication of seafood products (www.labelfish.eu).

2. PURPOSE

The purpose of this SOP is to provide a genetic method for the identification of fish species, in order to support the implementation of food labelling/authenticity testing.

3. SCOPE

This method is suitable for the qualitative identification of DNA (deoxyribonucleic acid) in fish products. It has been tested against a very broad taxonomic range of fish species (but has failed in a small minority of cases, <5% of species tested; Ivanova *et al.*, 2007). The assay is designed to work with fresh, smoked, salted and frozen samples. It is also successful with cooked products, but success is dependent on the intensity of cooking. It is not suitable for highly processed foods e.g. tins of tuna. It is also unsuitable for the identification of complex fish products containing DNA from multiple species. For some species of relatively recent evolutionary origin, this method may only be able to identify the sample down to the genus level (e.g. some tunas of the genus *Thunnus*, or redfish of the *Sebastes* genus). In these cases, additional tests might be required for species level identification.

4. DEFINITIONS & ABBREVIATIONS

DNA: Deoxyribonucleic acid

PCR: Polymerase Chain Reaction

SOP: Standard Operating Procedure

UV: Ultraviolet

CO1/COI: Mitochondrial cytochrome c oxidase 1 gene

5. PRINCIPLE OF THE METHOD

The following is taken from the international Barcode of Life Project (<http://www.barcodeoflife.org/>);

“Barcoding uses a very short genetic sequence from a standard part of the genome the way a supermarket scanner distinguishes products using the black stripes of the Universal Product Code (UPC). Two items may look very similar to the untrained eye, but in both cases the barcodes are distinct. The gene region that is being used as the standard barcode for almost all animal groups is a 648 base-pair region in the mitochondrial cytochrome c oxidase 1 gene (“CO1”). COI is proving highly effective in identifying many animal groups”.

6. MATERIALS & EQUIPMENT

The sections below report all the equipment and materials required to apply this protocol.

N.B. Batch numbers of kits used must be recorded.

6.1 Water

General use: Distilled or de-ionised water

PCR procedures: Sterile, DNase-, RNase- and Protease-free water e.g. Fisher Scientific DNA free water, product code: BPE2470-1

6.2 Solutions, standards and reference materials

The present SOP was validated using a ring trial based on 13 “blind” reference tissues (list of voucher specimens is held by the LABELFISH consortium). Details on the ring-trial procedure and results are available upon request to the LABELFISH consortium.

6.3 Commercial kits

DNA Extraction: The method has been validated using the ‘DNeasy Blood & Tissue Kit’ supplied by Qiagen (Product code 69504). DNA extraction kits from other suppliers must be shown to be appropriate before use.

6.4 Plastic ware

N.B. It is essential that all plastic-ware is sterile before use.

Item	Detail	Example Supplier	Product code
Pipette tips (filtered)	10, 20, 200 & 1000µl	Starlabs	S1120
PCR tubes	single, strip or 96-well	Starlabs	I1402
1.5ml tubes	1.5 ml	Starlabs	S1615

6.5 Equipment

The following items of equipment are required to undertake the analysis. Several alternative suppliers/models are available for each item. These must be shown to be appropriate before use.

Item	Detail	Example supplier	Product code
Precision pipettes	1-1000µl	Starlabs	G8900
Bench top vortex		Labnet	VX-100
Thermocycler	ABI Vereti 96 well		
Thermal mixer	to hold 1.5 ml tubes	Eppendorf	5355
DNA quantifier	Accurate to +/- 1 ng		ND1000
Microcentrifuge	to hold 1.5 ml tubes	Eppendorf	5452

Optional – laminar flow hood

6.6 Other materials

Disposable plastic gloves, sterile dissection equipment.

6.7 Electronic files / computer software

A computer with a text editor e.g. notepad.

Freely available sequence editing software e.g. Bioedit, FinchTV, ProSeq.

Internet access is required to utilise the Barcode of Life System: <http://www.boldsystems.org/>

7. PROCEDURES

It is essential to wear disposable plastic gloves during all laboratory procedures and to use pipette tips that are sterile and fitted with filters.

7.1 Sample preparation

All samples should be stored frozen at -20°C until processed. Samples can be stored frozen indefinitely.

N.B. In the ring trial ethanol-preserved samples were utilised.

The external surfaces of samples submitted for analysis may have been affected through preservation treatments or bacterial breakdown. Where possible, obtain subsamples for DNA extraction from the least degraded area of tissue in order to minimise contaminant DNA and DNA degradation. This will typically mean removing outer layers of tissue in contact with the environment before taking a subsample. Use sterile dissection equipment where appropriate.

7.2 DNA Extraction

Materials:

The extraction should be carried out with the Qiagen DNeasy Blood & Tissue Kit, following the manufacturer's protocol. It is recommended that the manufacturer's guidelines are checked each time kits are ordered to ensure any updates/changes made since development of this SOP are incorporated.

Procedure:

1. Cut up approx. 25 mg tissue into small pieces and place into a 1.5 ml tube.
2. Include an empty 1.5 ml tube as an extraction control. This is treated following the same procedure and carried through to the PCR stage (7.3).
3. Add 180 µl Buffer ATL (tissue lyser).
4. Add 20 µl proteinase K and vortex for 15 seconds.
5. Incubate in a thermal mixer at 56°C for 2 hours.
6. Vortex for 15 seconds.
7. Add 200 µl Buffer AL (cell lyser) to the sample and vortex for 15 seconds.
8. Add 200 µl 100% ethanol and vortex for 15 seconds.
9. Pipette the mixture into a DNeasy Mini spin column placed in a 2 ml collection tube.
10. Centrifuge at 8000 rpm for 1 minute
11. Discard the eluate and replace the collection tube.
12. Add 500 µl Buffer AW1 (wash 1).
13. Centrifuge at 8000 rpm for 1 minute
14. Discard the eluate and replace the collection tube.

15. Add 500 µl Buffer AW2 (wash 2).
16. Centrifuge at 13,000 rpm for 3 minutes
17. Discard the eluate and collection tube. Place the spin column in a 1.5 ml tube.
18. Pipette 100 µl Buffer AE (elution) directly onto the spin column membrane.
19. Incubate at room temperature for 1 minute
20. Centrifuge at 8000 rpm for 1 minute to elute DNA.
21. Discard the spin column, close the tube and store the eluate containing DNA at 4°C for up to one week or in a freezer (-20°C) long term.
22. DNA extract quantification. Extracted DNA must be quantified to assess the extraction process and enable normalisation of DNA concentration. One common method is to use a Nanodrop ND 1000 Spectrophotometer. DNA should be diluted to 10-50ng/µl using DNA-free water. Negative controls should read ~0 ng/µl.

Controls:

A negative extraction control (with no tissue) should be run in parallel with all batches of sample extraction and quantified alongside all tissue extractions.

7.3 PCR Amplification

Materials:

BIOTAQ DNA polymerase 500Units (Bioline Catalogue number BIO-21040, also contains reaction buffer & MgCl₂)

dNTP mix 10mM final concentration (Bioline Catalogue number BIO-39053, each dNTP at 2.5mM concentration)

Procedure:

1. Create a sample plan (ideally in Excel) describing the DNA being analysed and it's locations in the rack/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 (i.e. ultra pure water) and a positive control (*this can be determined internally in each lab, but the DNA must have originated from a fish for which the species has been accurately identified, or previously experimentally determined via COI sequencing; i.e. it needs to have successfully been PCR amplified previously*) are included.
4. Make up the primers to a 0.01 mM (i.e. 10 pM/µL) concentration.

Primers:

Primer Name	Primer sequence (5'-3')	References
VF2_t1	TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC	Ward <i>et al.</i> 2005
FishF2_t1	TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC	Ward <i>et al.</i> 2005
FishR2_t1	CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA	Ward <i>et al.</i> 2005

FR1d_t1	CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA	Ivanova <i>et al.</i> 2007
---------	---	----------------------------

5. Prepare the PCR reactions as follows (this following recipe is enough for 1 reaction and requires multiplication for the number of samples being analysed, in order to account for pipetting error it is also recommended to add 10% to the total volume of each of the reagents utilised);

PCR master mix, per reaction with a total volume 20 µl;

10 µL of 10% trehalose (e.g. Sigma-Aldrich, catalogue number T-5251)

2.7 µL of ultra pure water

2 µL 10×reaction buffer

1 µL MgCl₂ (50 mM)

0.2 µL of each primer (0.01 mM)

0.4 µL of the Bioline 10mM dNTP mix

0.1 µL of BIOTAQ *Taq* DNA Polymerase

6. Vortex master mix thoroughly.

7. Place 17 µL of the master mix into every tube/well (can use the same pipette tip during this step).

8. Aliquot 3 µl of DNA template to each tube/well following your sample plan.

Reagent	Per Reaction
10% trehalose	10
ddH ₂ O	2.7
10X buffer	2
50mM MgCl ₂	1
Primer VF2_t1	0.2
Primer FishF2_t1	0.2
Primer FishR2_t1	0.2
Primer FR1d_t1	0.2
dNTPs 10mM total mix	0.4
Taq	0.1
TOTAL	17

9. Thermal conditions for the PCR reaction are; 94°C for 2 min, 35 cycles of 94°C for 30 sec, 52°C for 40 sec, and 72°C for 1 min, with a final extension at 72°C for 10 min (the “hot lid” option should also be selected).

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

11. Once completed the PCR reactions can be stored in the fridge at 4°C. But for long term storage (i.e. greater than a week) freezing at -20°C is recommended.

7.4 PCR Product Check

Gel electrophoresis of DNA in an agarose gel is a standard technique in molecular biology, but equipment, reagents, staining and visualisation varies considerably between laboratories, and according to local health & safety controls. Therefore, this SOP suggests general conditions that need to be adapted to each laboratory.

1. Make a 1-2% agarose gel

2. Once set, load 4 µL of the PCR product into the well (the addition of loading buffer/dye may be necessary).

3. Include appropriate size standard in one lane (e.g. 5 µl Bioline hyperladder 100, catalogue number BIO-33056).

4. Run at 100V for approximately 1 hr (depending on size of gel), ensuring the DNA does not run off the gel.

5. Visualise your DNA fragments in UV light (with appropriate safety precautions); if the PCR reaction has been successful the positive control will have a single bright band of approximately 700 base pairs in length. Your negative controls should not contain bands. A band in the lanes corresponding to your samples indicates successful amplification.

6. Keep a permanent record of your gel (electronic and/or hard copy) as proof that the PCR amplification was successful and contaminant free.

7.5 DNA Sequencing

For this SOP it is assumed that the majority of laboratories do not have access to Sanger sequencing equipment in-house, therefore it is recommended that the PCR products are sent to an external company for PCR clean up and sequencing reaction. The requirements for the

sequencing services vary, especially in terms of the volume and concentration of PCR product and sequencing primers required. This needs to be checked specifically with your preferred service provider.

1. Estimate concentration of your PCR product. This can be done from the record you made of your PCR products when run on the agarose gel, by comparing the brightness of the bands to the size standard that was run (that has a standard concentration of DNA). This information is usually required by the sequencing service.

2. When placing an order for sequencing it is important to make clear that for each PCR product two sequencing reactions are required; one utilising the forward primer and a second utilising the reverse primer (so for each sample two complementary sequences will be obtained).

3. Ensure the PCR products are cleaned before the sequencing reaction is attempted. This can usually be completed by the external sequencing company (but there are protocols/kits to do this in-house e.g. ExoSAP-IT- USB Corporation; Cleveland, OH Cat. No. 78201).

4. Send your carefully labelled PCR products and sequencing reaction primers to the sequencing service, according to their instructions. The sequencing primers differ from those used in the PCR amplification and are detailed below;

Primer Name	Primer sequence (5'-3')
M13F (-21)	TGTAAAACGACGGCCAGT
M13R (-27)	CAGGAAACAGCTATGAC

Additional Resources;

A protocol developed by the consortium for the barcode of life is available below and deals with procedures 7.1 – 7.5 in greater detail, providing some useful background information and potential troubleshooting:

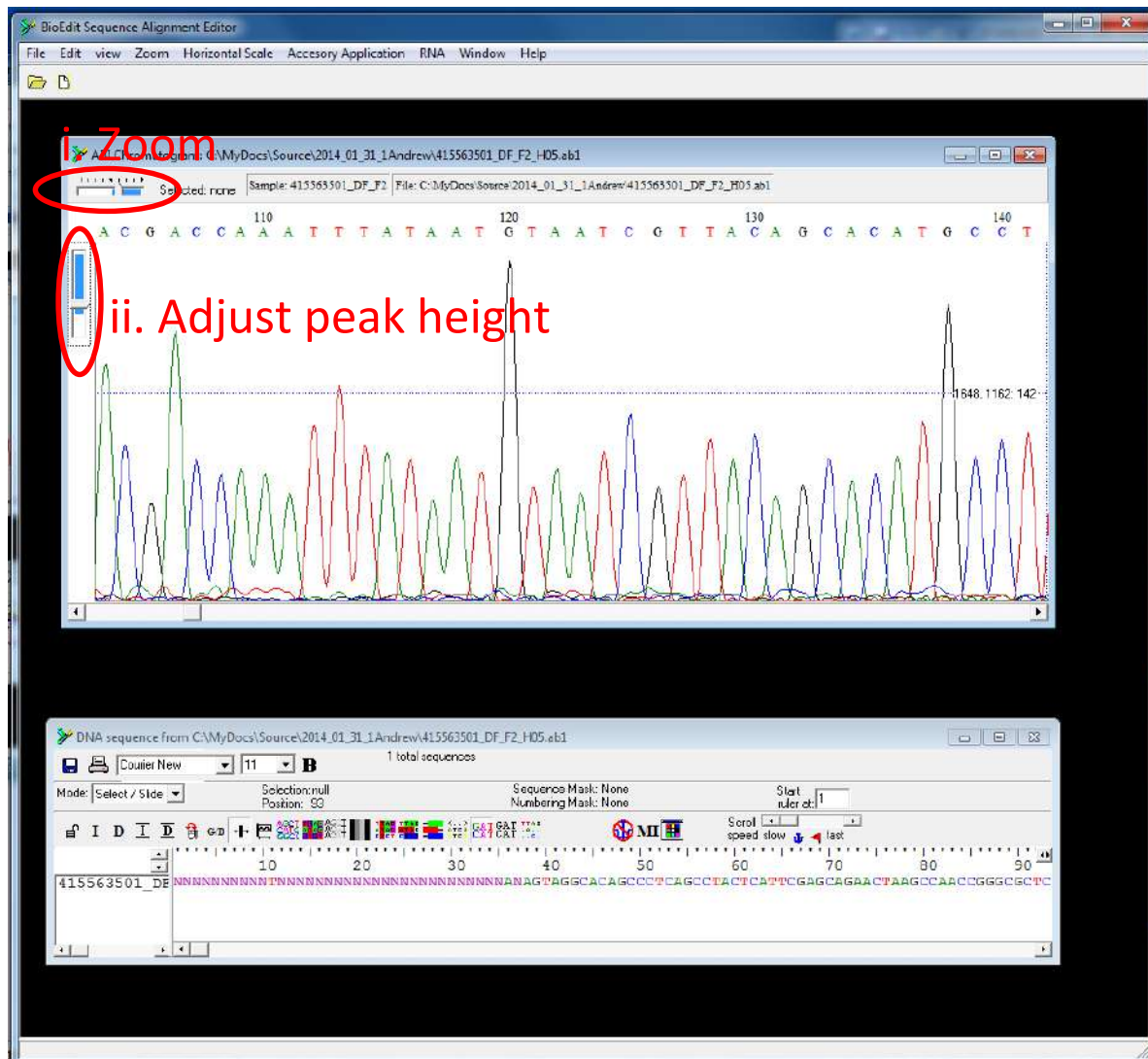
[http://www.barcodeoflife.org/sites/default/files/Protocols for High Volume DNA Barcode Analysis.pdf](http://www.barcodeoflife.org/sites/default/files/Protocols%20for%20High%20Volume%20DNA%20Barcode%20Analysis.pdf)

7.6 Raw Data Processing

Sequencing services usually supply the results in a range of files, but it is the ABI data file (.abi) required in the SOP (it is important to ensure the sequencing company will supply these before making an order, but it is usually standard). The raw data needs to be checked and edited before it can be used.

The ABI files can be viewed and edited with a number of freely available software packages (mentioned in section 6.7). This SOP has been tested using BioEdit, which can be downloaded from the following webpage: <http://www.mbio.ncsu.edu/bioedit/bioedit.html>

4. In order to optimise the view of the trace within the software, both (i.) zoom and (ii.) relative peak height function are present that can be adjusted to your preference;

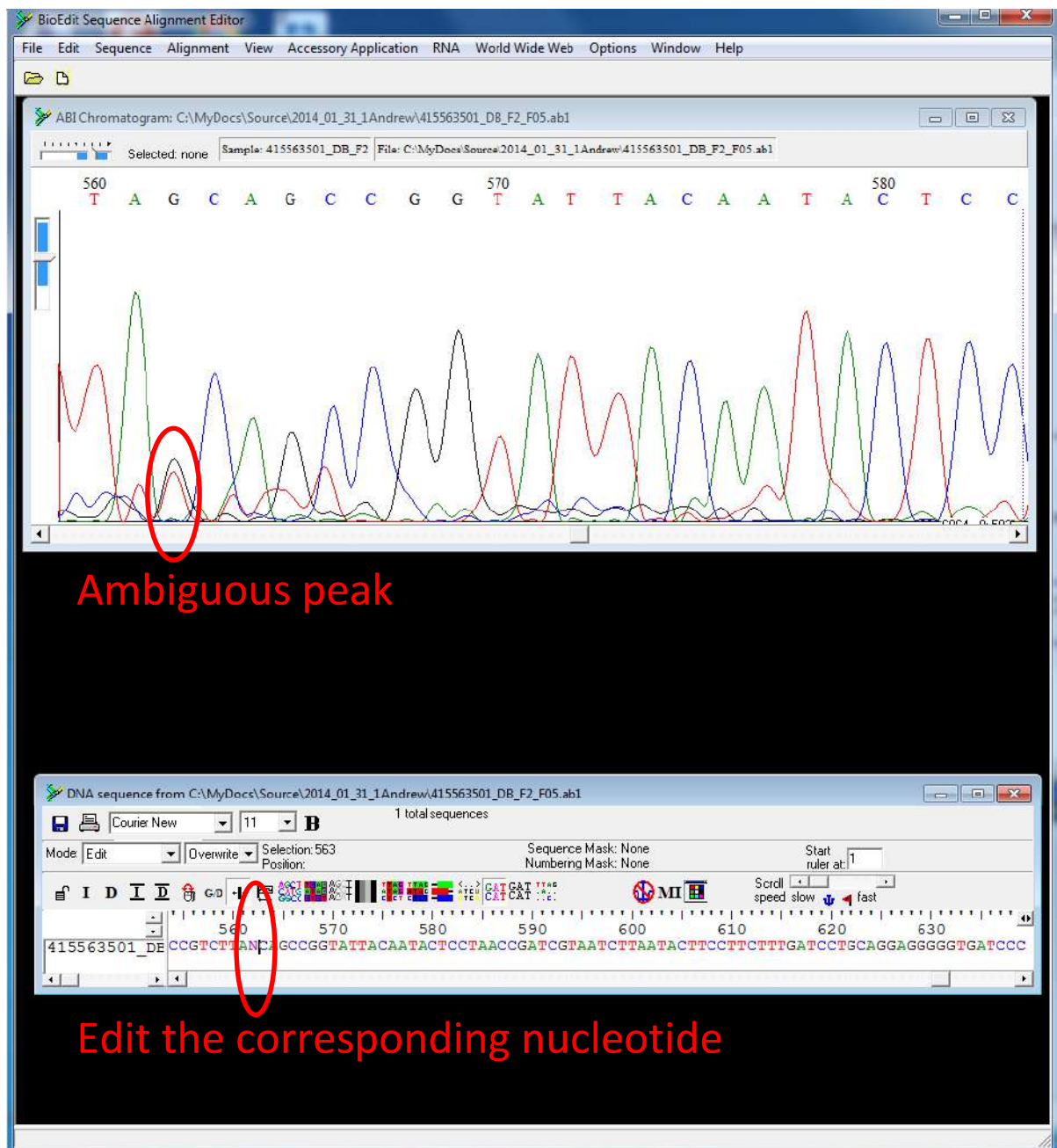


5. The sequence must be checked by eye to ensure the sequence reaction has worked successfully and the trace is of high quality. The majority of the trace should consist of a series of clear peaks (as in the figure above.) If the reaction has failed, or contamination is present, the peaks will look weak and/or it will be impossible to clearly resolve a single peak at each nucleotide position. If this is the case the results are not high quality enough for use.

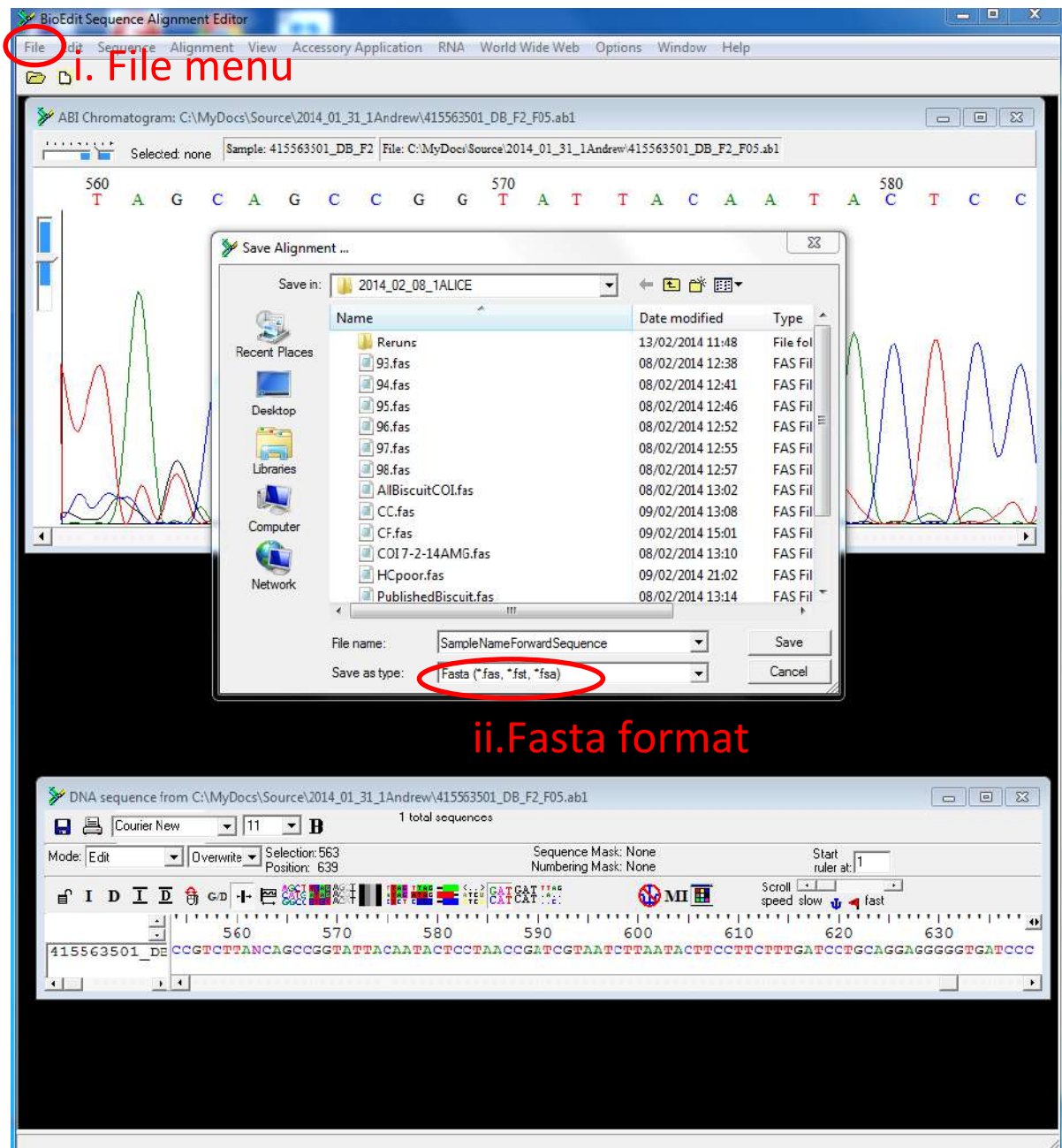
6. Often the quality of the sequencing reaction is poor at either ends of the trace (as below). In this case the ambiguous region at either end can simply be deleted, just leaving the high quality sequence (i.e. delete the flanking sequence at each end until you are confident that you can easily call each peak). *In the example below unambiguous peaks appear approximately after nucleotide position 42.*

The screenshot displays the BioEdit Sequence Alignment Editor interface. The main window shows an "ABI Chromatogram" titled "C:\MyDocs\Source\2014_01_31_Andrew\415563501_DF_F2_H05.ab1". Below the title bar, there's a menu bar with options like File, Edit, Sequence, Alignment, View, Accessory Application, RNA, World Wide Web, Options, Window, and Help. A toolbar contains icons for file operations and editing. The central area features a chromatogram plot with four colored traces (green, black, red, blue) representing DNA bases. Above the plot, the sequence is displayed as "NNN NNN NN NT NNNNNNNNNNNNNNNNNNNNNNNNA G TAG GCACA GCCCT CAG CCTACTCATTCGAGCAGA ACTAA GCCAAC".

8. Ambiguous nucleotides may also arise within the middle of the sequences (see below). Provided the sequence window is in the editing mode, these can simply be overwritten with the letter “N” (indicating uncertainty about the call). **Do not delete them.**



9. Once only high quality sequence remains, it is necessary to click on the nucleotide sequence window (so it is selected as the active window) and save the sequence. This is done in the file menu (i.) in the uppermost toolbar and selecting “Save As”. The file can be renamed (e.g. with the name of the original sample with indication as to whether the sequence was generated with the forward or reverse primer) and must be saved in fasta format (as ii. below):



Additional Resources;

The Barcode of life systems page has an excellent description for assembling and editing sequences and the common errors that can arise:

http://www.boldsystems.org/index.php/resources/handbook?chapter=7_validation.html

Although not part of the SOP, it is also possible to get a free BOLD Systems account and upload ABI trace files onto the workspace, where the system can make an automated check of the quality of your sequence – see trace submission in the BOLDsystems handbook;

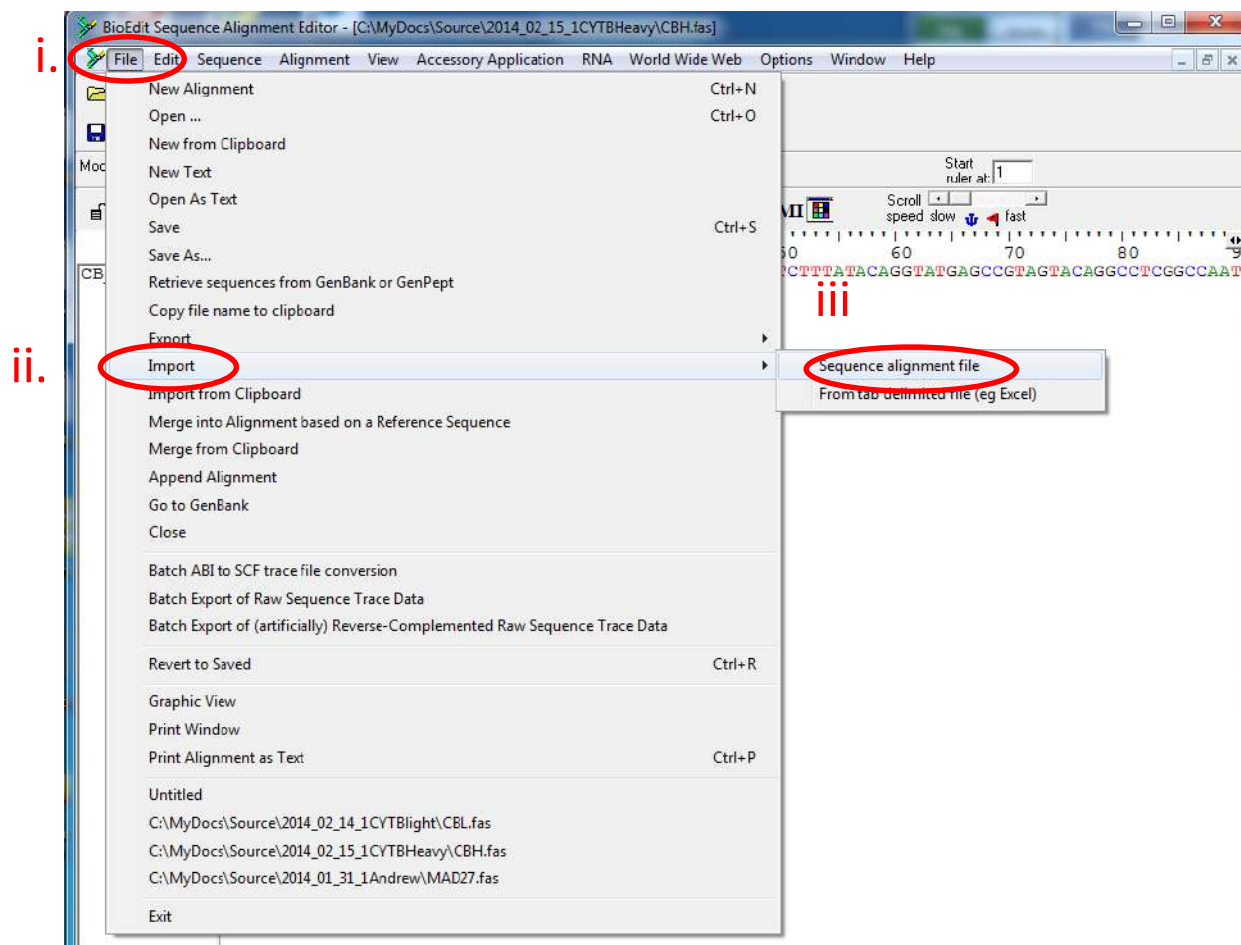
http://www.boldsystems.org/index.php/resources/handbook?chapter=3_submissions.html§ion=trace_submissions

7.7 Generating a consensus sequence

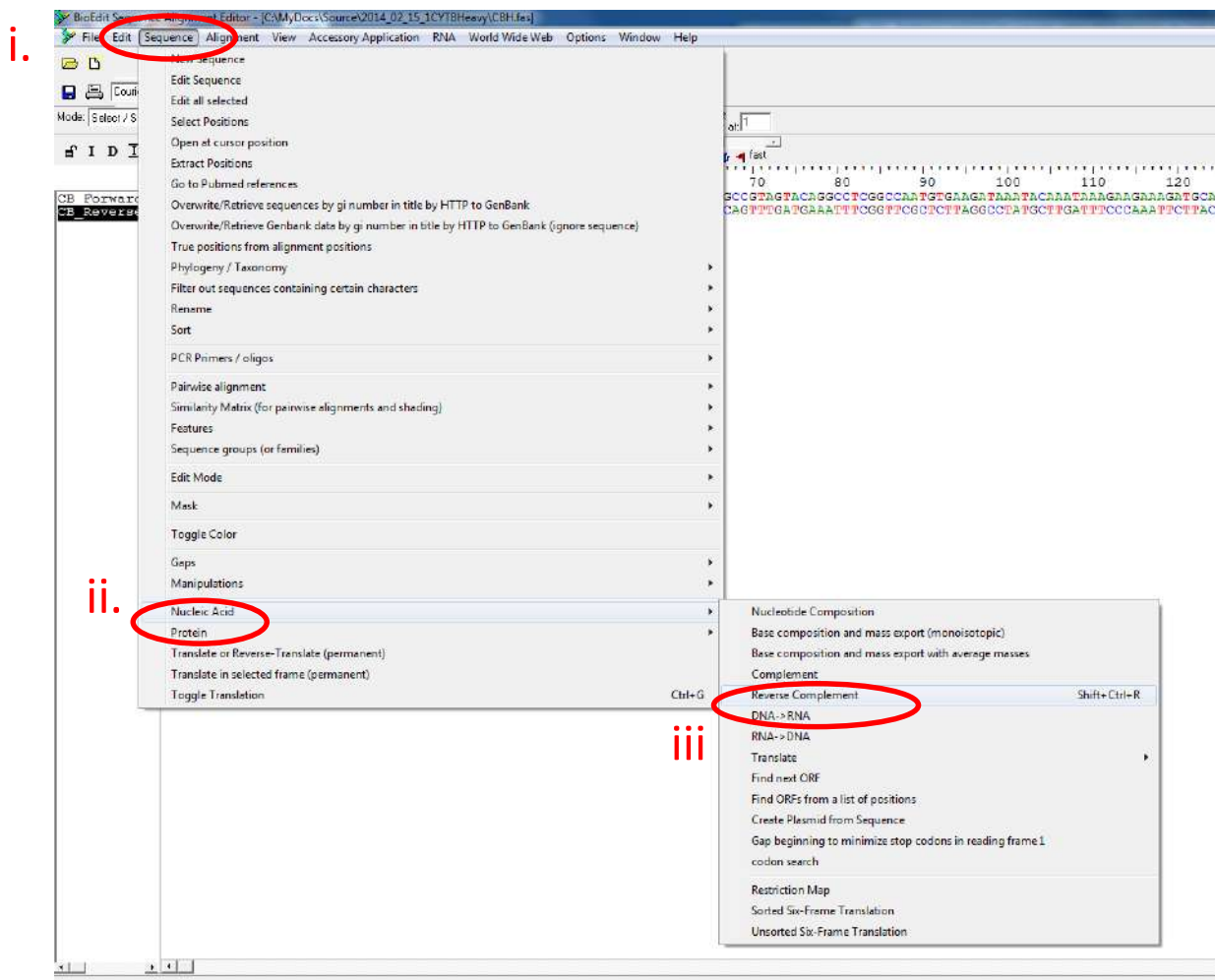
Each of the samples should have been sequenced in both the forward and reverse directions, meaning these complementary/overlapping sequences can be combined into a consensus. This serves as an important way of checking the accuracy of the sequence, and can help remove any ambiguous bases and generate a longer total sequence.

1. Start the BioEdit software by clicking on the BioEdit.exe icon and open the edited forward fasta files generated from the sample (as in 7.6). This will only open a nucleotide sequence window (there will be no trace window).

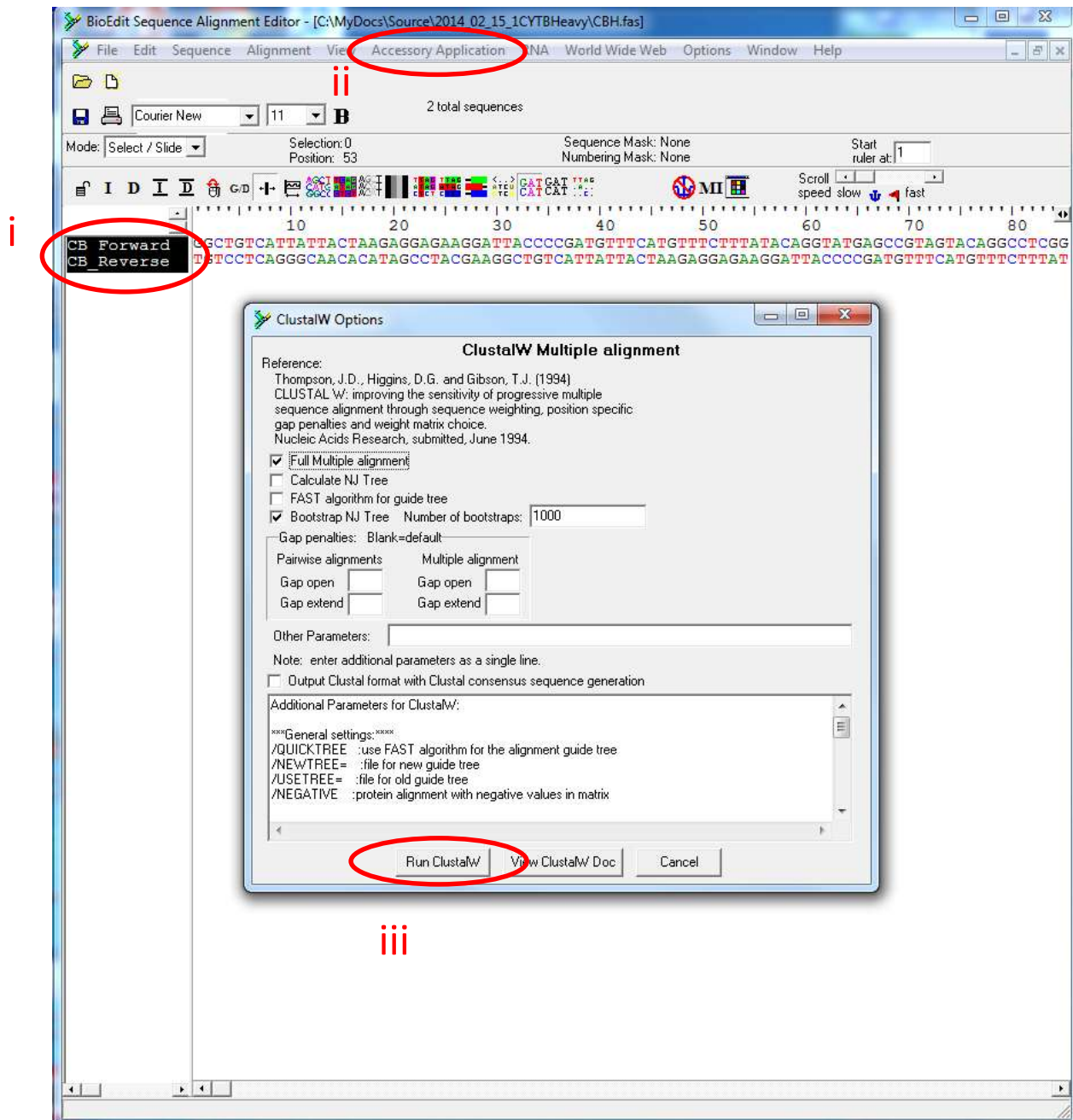
2. It is then necessary to import the reverse fasta file into the software. This is done in the file menu in the uppermost toolbar and selecting import, then sequence alignment file and locating your complementary reverse sequence fasta file.



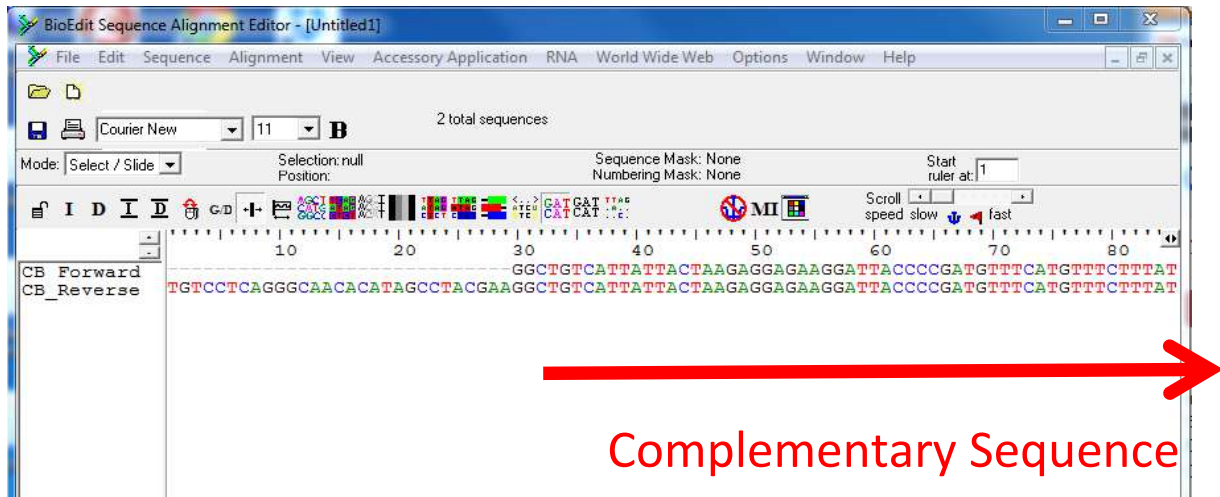
3. Select the reverse sequence within the nucleotide sequence window, just by clicking on its name on the far left. Then, in the sequence menu in the uppermost toolbar, select Nucleic Acid, followed by Reverse Complement.



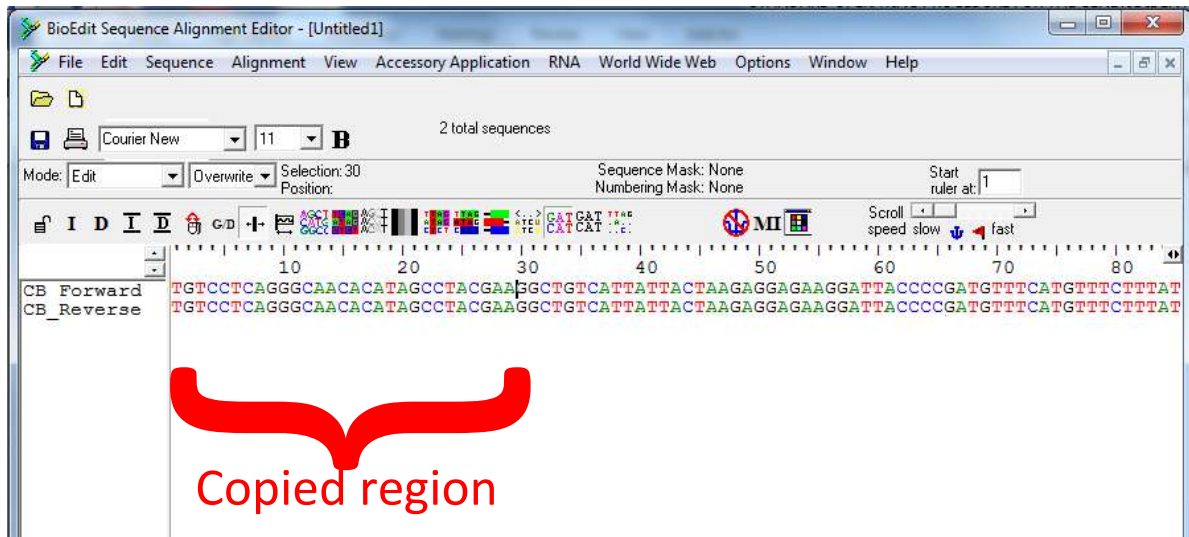
4. Select both the forward and reverse sequence within the nucleotide sequence window, using shift and select (i). Then, in the sequence menu in the uppermost toolbar, select Accessory Application (ii), followed by ClustalW Multiple Alignment. Leave the settings as defaults and click on the Run ClustalW tab (iii).



5. The software will take a few seconds to align these complementary sequences and the result is a large region of overlapping sequence. As these two sequences come from the same sample they should match perfectly with no mismatching nucleotides. However, any ambiguous nucleotides (i.e. "N") can now be resolved from the complementary sequence. Any mismatches also need to be resolved by consulting the original trace files and deciding which nucleotide call is correct (if this is not possible an "N" can be used at the position where the sequences mismatch, as section 7.6).



6. The region where the complementary sequences do not overlap on the reverse sequence needs to be copied and pasted (using the ctrl+c and ctrl+v keyboard shortcuts) onto the end forward sequence, creating a full length barcode.



7. The reverse sequence can now be deleted and this full length barcode sequence can be saved (as in 7.6), by renaming it "*sampleNameComplementary*" and saving it in fasta format.

8. The final step in generating a DNA barcode is to remove the primers. This can be done by referring to the primer sequences 7.3 & 7.5 and removing them from both ends of your sequence. It can perhaps more easily be done by aligning the consensus sequence with a full length barcode downloaded from BOLD. The standard barcode length for most animal species

is 648bp, so your edited sequence should be approximately this long. Below is a full length barcode for Atlantic cod (*Gadus morhua*), in text format, obtained through the application of the steps illustrated above.

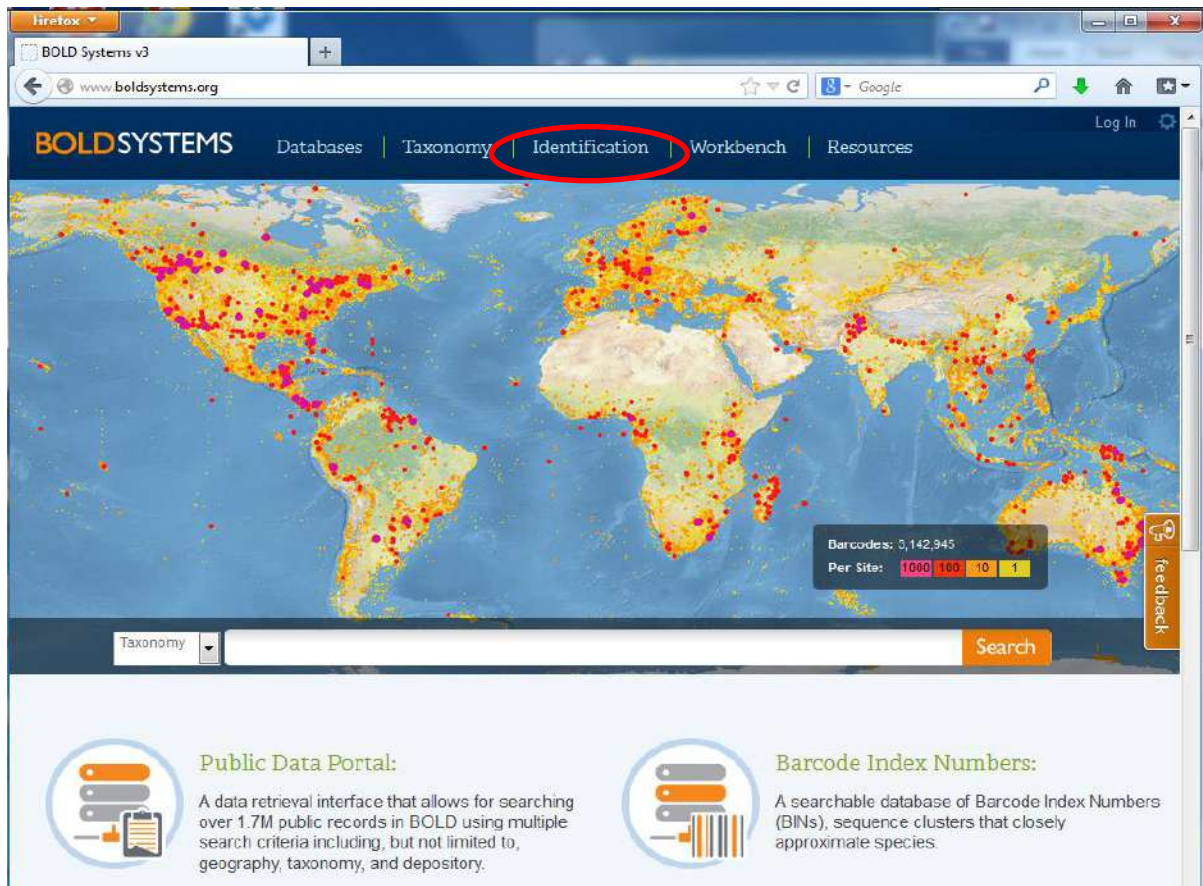
>GadusMorhuaSCFAC839-06

```
CCTTTATCTCGTATTTGGTGCCTGAGCCGGCATAGTCGGAACAGCCCTAAGCCTGCTCA
TTCGAGCAGAGCTAAGTCAACCTGGTGCACCTTCTTGGTGATGATCAAATTTATAATGTG
ATCGTTACAGCGCACGCTTTCGTAATAATTTTCTTTATAGTAATACCACTAATAATTGGA
GGCTTTGGGAACTGACTCATTCTCTAATGATCGGTGCACCAGATATAGCTTTCCCTCG
AATAAATAACATAAGCTTCTGACTTCTTCCTCCATCTTTCTGCTCCTTTTAGCATCCTCT
GGTGTAGAAGCTGGGGCTGGAACAGGCTGAACTGTCTATCCACCTTTAGCCGGAAACC
TCGCTCATGCTGGGGCATCTGTTGATCTCACTATTTTTCTCTTCATCTAGCAGGGATTT
CATCAATTCTTGGGGCAATTAATTTTATTACCACAATTATTAATATGAAACCTCCGGCAAT
TTCACAGTACCAAACACCCCTATTTGTTTGAGCAGTACTAATTACAGCTGTGCTTCTACT
ATTATCTCTCCCCGTCTTAGCAGCTGGTATCACATACTTCTAACTGACCGTAATCTTAA
TACTTCTTTCTTTGACCCTGCTGGAGGAGGTGATCCCATTTTATACCAACA
```

7.8 Identifying the species on the Barcode of Life Database

In order to identify what species your consensus, full-length COI sequence originates from it is necessary to utilise freely available data that has been submitted to the Barcode of Life (BOLD) project. This includes a comprehensive database of COI sequence data that has been collected by individuals and organisation across the globe and is constantly being updated with new data.

1. Start by navigating to the **BOLD Systems** webpage (<http://www.boldsystems.org/>) and select the “Identification” tab at the top of the webpage.



2. This page acts as a portal allowing the consensus sequence generated in the laboratory to be referenced against the entire BOLD database of reference data, i.e. from known species. **Various search options** are possible that relate to different collections of reference data, but the **default settings** provide an excellent initial step at identifying the species. However, it is important to ensure that the “Animal Identification (COI)” tab (i) and the “Species Level Barcode Records” database (ii) are both selected. The consensus sequence obtained from the sample can then be cut & pasted into the empty box at the bottom of the page (iii); in this example the published sequence from *Gadus morhua* included in the previous section has been utilised. The easiest way to copy the consensus sequence in your fasta file is to force windows to open the .fas file in Notepad. Alternatively, make a copy of the .fas file and edit the file extension to .txt allowing it to be opened in Notepad. Once the sequence has been entered, hit the submit button at the bottom of the page.

Firefox
BOLD Systems: Management & Analysis...
www.boldsystems.org/index.php/IDS_OpenIdEngine
User Public

BOLDSYSTEMS Databases | Taxonomy | Identification | Workbench | Resources

Identification Request

Print

i. **Animal Identification [COI]** Fungal Identification [ITS] Plant Identification [rbcL & matK]

The BOLD Identification System (IDS) for COI accepts sequences from the 5' region of the mitochondrial Cytochrome c oxidase subunit I gene and returns a species-level identification when one is possible. Further validation with independent genetic markers will be desirable in some forensic applications.

Historical Databases: [Jul-2013](#) [Jul-2012](#) [Jul-2011](#) [Jul-2010](#) [Jul-2009](#)

Search Databases:

ii. ☒ **All Barcode Records on BOLD (2,742,418 Sequences)**
Every COI barcode record on BOLD with a minimum sequence length of 500bp (warning: unvalidated library and includes records without species level identification). This includes many species represented by only one or two specimens as well as all species with interim taxonomy. This search only returns a list of the nearest matches and does not provide a probability of placement to a taxon.

☐ **Species Level Barcode Records (1,779,732 Sequences/146,084 Species/58,357 Interim Species)**
Every COI barcode record with a species level identification and a minimum sequence length of 500bp. This includes many species represented by only one or two specimens as well as all species with interim taxonomy.

☐ **Public Record Barcode Database (555,693 Sequences/62,894 Species/14,985 Interim Species)**
All published COI records from BOLD and GenBank with a minimum sequence length of 500bp. This library is a collection of records from the published projects section of BOLD.

☐ **Full Length Record Barcode Database (1,318,574 Sequences/132,900 Species/50,814 Interim Species)**
Subset of the Species library with a minimum sequence length of 640bp and containing both public and private records. This library is intended for short sequence identification as it provides maximum overlap with short reads from the barcode region of COI.

iii. Enter sequences in fasta format:

```
>GadusMorhuaSCFAC839-06
CCCTTTATCTCGTATTGGTGCCTGAGCCGGCATACTCGGAACAGCCCTAAGCCCTCCTCATTGAGCAGAGCTAAG
TCAACCTCGTCCACTTCTGCTCATCAAAATTTATAATCTGATCGTTACAGCCGACGCTTCTGTAATAATTTT
CTTTATAGTAATACCACTAATAATTGGAGGCTTTGGGAAGTCACTCATTCTCTAATGATCGGTGCACAGATAT
AGCTTTTCCTCGGAATAAATAAGCTTCTGACTTCTTCTCCATCTTCTCTGCTCCTTTAGCATCCTCTGG
TGTAGAACGCTCGGCTCGAACAGGCTCAACTGCTATCCACCTTTAGCCGGAAACCTCGCTCATGCTCGGCGATC
TGTTGATCTCACTATTTTCTCTTCTATCTAGCAGGATTTCAATCTTGGGGCAATTAAATTTATTACCAC
AATTATTAATGAAACCTCGGCAATTTACAGTACCAAAACCCCTATTTGTTTGAGCAGTACTAATTACAGC
TCTGCTTCTACTATTATCTCTCCCGCTTTAGCAGCTGGTATACAACTCTTAACCTACCGTAATCTTAATAC
TTCTTTCTTTGACCTCGCTGGAGGAGGTATCCCATTTTATACCAACA
```

Email me the results ☐ **Submit**

Feedback

3. In a few seconds the browser will update and give you the results of the search, revealing the records contained in the database that **yields the closest match** in terms of sequence similarity. First, it is important to save a screen grab of the results as proof of the result, something similar to the picture below (this can be done using the print screen option, pasting directly into Paint or a Microsoft Office software and saving as a picture).

www.boldsystems.org/index.php/IDS_IdentificationRequest

Gadus chalcogrammus

User Public

BOLDSYSTEMS Databases | Taxonomy | Identification | Workbench | Resources

Specimen Identification Request

Query: GadusMorhuaSCFAC839-06 Top Hit: Chordata - Gadiformes - Gadus morhua (100%)

Search Result:

A species level match could not be made, the queried specimen is likely to be one of the following:

[Gadus morhua](#)
[Gadus chalcogrammus](#)

For a heirarchical placement - a neighbor-joining tree is provided: [Tree Based Identification](#)

Identification Summary:

Taxonomic Level	Taxon Assignment	Probability of Placement (%)
Phylum	Chordata	100
Class	Actinopterygii	100
Order	Gadiformes	100
Family	Gadidae	100
Genus	Gadus	100

Similarity Scores of Top 99 Matches:

TOP 20 Matches:

Display option: Top 20

Phylum	Class	Order	Family	Genus	Species	Similarity (%)	Status
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Early-Release
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Early-Release
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Early-Release
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	morhua	100	Published

4. This screen also contains a lot of information that will allow a **confident identification** to be made from your sequence. At the top the search result is returned; in BOLD this 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. **However**, in the example below, two species have been returned (highlighted in red), prompting BOLD to display the message “**A species match could not be made**, the queried specimen is likely to be one of the following”. It is possible to interrogate the results further and still make a robust identification.

www.boldsystems.org/index.php/IDS_IdentificationRequest

BOLDSYSTEMS Databases | Taxonomy | Identification | Workbench | Resources

User Public

Specimen Identification Request

Query: *GadusMorhua*SCFAC839-06 Top Hit: Chordata - Gadiformes - *Gadus morhua* (100%)

Search Result:

A species level match could not be made. the queried specimen is likely to be one of the following:

Gadus morhua
Gadus chalcogrammus

For a heirarchical placement - a neighbor-joining tree is provided: Tree Based Identification

Identification Summary:

Taxonomic Level	Taxon Assignment	Probability of Placement (%)
Phylum	Chordata	100
Class	Actinopterygii	100
Order	Gadiformes	100
Family	Gadidae	100
Genus	Gadus	100

Similarity Scores of Top 99 Matches:

TOP 20 Matches:

Phylum	Class	Order	Family	Genus	Species	Similarity (%)	Status
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Early-Release
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Early-Release
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Early-Release
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private
Chordata	Actinopterygii	Gadiformes	Gadidae	Gadus	<i>morhua</i>	100	Private

Display option: Top 20

In the example below, it is clearly illustrated that there is **100% sequence similarity** between our example consensus sequence and the *Gadus morhua* records. It is also clear that there is a sudden reduction in the level of similarity observed between the consensus sequence and the records originating from *Gadus morhua* (which are 100-99.35% similar) and those from *Gadus chalcogrammus* (whose highest similarity is 98.53%), as indicated by the red arrow (iii.). The 100% sequence match criterion alongside the reduced similarity between our consensus sequences and any other matching species record, are both strong indicators that the sequence originated from *Gadus morhua*.

26

6. Besides referencing your sequence against the BOLD reference database, it is **also important to produce a simple tree** to graphically display the results of the homology search (although this is not a highly robust phylogenetic reconstruction). First click on the “Tree Based Identification” tab (i.), then a new window will pop up and the tree can be saved as a pdf by selecting the “Download Tree” option. This can then be saved as a permanent record of the results, to be kept alongside the previous screen-grab.

The image consists of two screenshots of the BOLD Systems website. The top screenshot shows the 'Specimen Identification Request' page. The query is 'GadusMorhuaSCEAC839-06'. The top hit is 'Chordata - Gadiformes - Gadus morhua (100%)'. The search result indicates that a species level match could not be made. A red circle highlights the 'Tree Based Identification' link, with a red 'i.' next to it. The bottom screenshot shows the 'Tree Result' page. It displays various options for viewing and downloading the tree. A red circle highlights the 'Download Tree' button, with a red 'ii.' next to it. The page also includes a 'Feedback' button on the right and a note at the bottom regarding sequence accuracy.

BOLD SYSTEMS Databases | Taxonomy | Identification | Workbench | Resources Log In

Specimen Identification Request

Query: *GadusMorhuaSCEAC839-06* Top Hit: Chordata - Gadiformes - Gadus morhua (100%)

Search Result:

A species level match could not be made, the queried specimen is likely to be one of the following:

[Gadus morhua](#)
[Gadus chalcogrammus](#)

For a heirarchical placement - a neighbor-joining tree is provided: [Tree Based Identification](#)

BOLD SYSTEMS

Tree Result -

Tree Parameters :

Datatype: Model: Labels:

PDF tree : [View Tree](#) [Download Tree](#)

Export Tree to Newick Format : [Download File](#)

Taxonomy Report : [View Report](#)

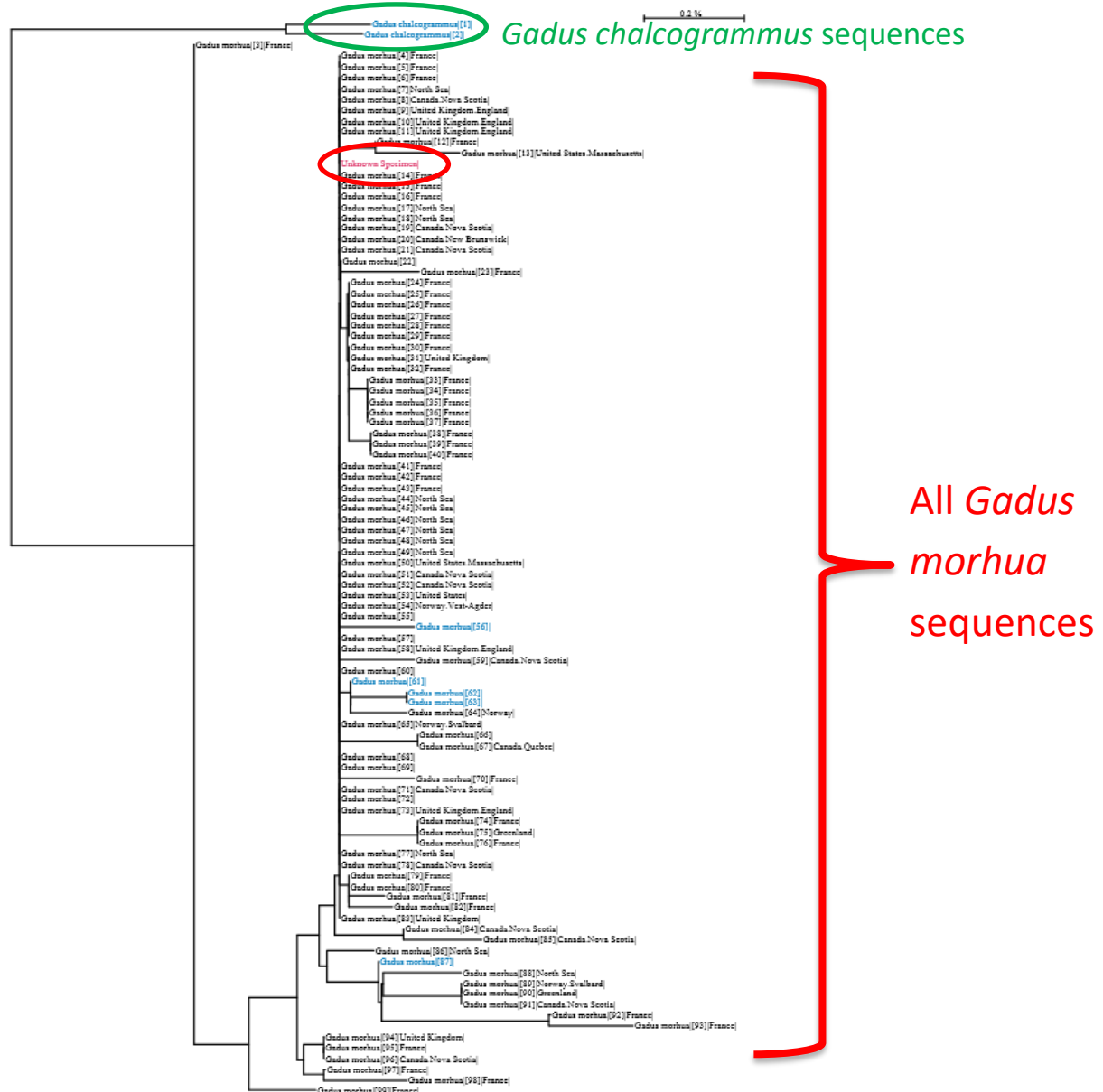
Image List : [View Image List](#)

Spreadsheet : [View Spreadsheet](#)

NOTE : Query sequence will be marked red on the tree with BOLD sequences in black. GenBank sequences that are not represented in BOLD may be less up to date in terms of sequence accuracy and taxonomic identification and will be marked in blue.

Feedback

7. In the tree diagram the uploaded sequence is highlighted in red. In order to make a clear identification, this “unknown specimen” should **only cluster with sequences originating from a single species** (i.e. from a *monophyletic* group). The tree generated from our example sequence is below; our uploaded sequence is clearly shown (highlighted in red) nested within sequences exclusively originating from *Gadus morhua*, with *Gadus chalcogrammus* (highlighted in green) forming a separate branch some distance from our unknown specimen. This is further evidence that this sequence originated from *Gadus morhua*.



8. In cases where **BOLD** returns more than one species and displays the message: "A species match could not be made, the queried specimen is likely to be one of the following", an additional search can also be made, **utilising a different set of reference data**. Return to the identification request portal and upload the sequence, but select the "Public Record Barcode Database" (this restricts the search to sequences that have been published). In some instances this may help provide an unambiguous identification and the results can be generated and saved as above (with a screen-grab and tree, relating to this search).

The screenshot shows the BOLD Systems website interface. The navigation bar includes links for Databases, Taxonomy, Identification, Workbench, and Resources. The 'Identification Request' section has three tabs: Animal Identification [COI], Fungal Identification [ITS], and Plant Identification [rbcL & matK]. The 'Animal Identification [COI]' tab is active. Below the tabs, there is a description of the BOLD Identification System (IDS) and a list of search databases. The 'Public Record Barcode Database' is highlighted with a red circle. Below the database list, there is a text box for entering sequences in FASTA format, which contains a sample sequence for *Gadus chalcogrammus*. A 'Submit' button is located at the bottom right of the text box.

BOLD SYSTEMS Databases | Taxonomy | Identification | Workbench | Resources

Identification Request [Print](#)

Animal Identification [COI] Fungal Identification [ITS] Plant Identification [rbcL & matK]

The BOLD Identification System (IDS) for COI accepts sequences from the 5' region of the mitochondrial Cytochrome c oxidase subunit I gene and returns a species-level identification when one is possible. Further validation with independent genetic markers will be desirable in some forensic applications.

Historical Databases: [Jul-2013](#) [Jul-2012](#) [Jul-2011](#) [Jul-2010](#) [Jul-2009](#)

Search Databases:

- ☐ All Barcode Records on BOLD (2,742,418 Sequences)
Every COI barcode record on BOLD with a minimum sequence length of 500bp (warning: unvalidated library and includes records without species level identification). This includes many species represented by only one or two specimens as well as all species with interim taxonomy. This search only returns a list of the nearest matches and does not provide a probability of placement to a taxon.
- ☐ Species Level Barcode Records (1,739,732 Sequences/146,084 Species/58,357 Interim Species)
Every COI barcode record with a species level identification and a minimum sequence length of 500bp. This includes many species represented by only one or two specimens as well as all species with interim taxonomy.
- ☒ **Public Record Barcode Database (555,894 Sequences/62,894 Species/14,985 Interim Species)**
All published COI records from BOLD and GenBank with a minimum sequence length of 500bp. This library is a collection of records from the published projects section of BOLD.
- ☐ Full Length Record Barcode Database (1,318,574 Sequences/132,900 Species/50,814 Interim Species)
Subset of the Species library with a minimum sequence length of 640bp and containing both public and private records. This library is intended for short sequence identification as it provides maximum overlap with short reads from the barcode region of COI.

Enter sequences in fasta format:

```
>GadusMorhuaSCFAC839-06
CCITTATCTCGTATTGGTGGCTGAGCCGGCATAGTCGGAACAGCCCTAAGCCTGCTCATTGGAGCAGAGCTAAG
TCAACCTGGTGCACTTCTTGGTGATCAATTTTAAATGTGATCGTTACAGCGCAGCTTTCGTAATAATTTT
CTTTATAGTAATACCACTAATAATTGGAGCCTTTGGGAAGTCACTTCTCTAATGATCGGTGCACAGATAT
AGCTTTCCCTCGAATAAATAACATAAGCTTCTGACTTCTCTCCATCTTCTCTGCTCTTTAGCATCCTCTGG
TGTAGAAGCTGGGGCTCGAACAGGCTGAAGTCTCTATCCACCTTTAGCCGGAACCTCGCTATGCTGGGGCATC
TGTTGATCTCACTATTTTCTCTTCTATCAGCAGGCAATTCATCAATCTTGGGGCAATTAATTTTATTACAC
AATTATTAATATGAACCTCCGGCAATTCACAGTACCAACACCCCTATTTGTTGACCACTACTAATTACAGC
TGTGCTTCTACTATTATCTCTCCCGCTTAGCAGCTGGTATCACAATCTCTAAGTACCGCTAATCTTAATAC
TTCTTTCTTGACCTGCTGGAGGAGTATCCCATTTTATACCAACA
```

[Submit](#)

9. Alternatively, if the sample for example comes from a rare or exotic fish, there may be **no matching records** in the “Species Level Barcode Records” database that demonstrate high levels of sequence similarity. The screen grab and tree are still essential records, especially as BOLD may still be able to assign the sequence to a **genus** or **family**, which still provides potentially useful information (and is often enough to help check for mislabelling). An **additional search is also possible**, in this case, by selecting the “All Public Records on BOLD” (this is the broadest database). This may yield a stronger match and the results can be saved as above (with a screen-grab and tree, relating to this search). If *a-priori* information about the species that sample supposedly originates from is available (i.e. the label), it is also possible to check if a species is represented in the database within the taxonomy tab at the top of the window. For further information on troubleshooting see 7.10.

BOLD SYSTEMS Databases | Taxonomy | Identification | Workbench | Resources

Log In

Identification Request

[Print](#)

Animal Identification [COI]

Fungal Identification [ITS]

Plant Identification [rbcL & matK]

The BOLD Identification System (IDS) for COI accepts sequences from the 5' region of the mitochondrial Cytochrome c oxidase subunit I gene and returns a species-level identification when one is possible. Further validation with independent genetic markers will be desirable in some forensic applications.

Historical Databases: [Jul-2013](#) [Jul-2012](#) [Jul-2011](#) [Jul-2010](#) [Jul-2009](#)

Search Databases:

- ☒ **All Barcode Records on BOLD (2,742,410 Sequences)**
Every COI barcode record on BOLD with a minimum sequence length of 500bp (warning: unvalidated library and includes records without species level identification). This includes many species represented by only one or two specimens as well as all species with interim taxonomy. This search only returns a list of the nearest matches and does not provide a probability of placement to a taxon.
- ☐ **Species Level Barcode Records (1,739,732 Sequences/146,084 Species/58,357 Interim Species)**
Every COI barcode record with a species level identification and a minimum sequence length of 500bp. This includes many species represented by only one or two specimens as well as all species with interim taxonomy.
- ☐ **Public Record Barcode Database (555,693 Sequences/62,894 Species/14,985 Interim Species)**
All published COI records from BOLD and GenBank with a minimum sequence length of 500bp. This library is a collection of records from the published projects section of BOLD.
- ☐ **Full Length Record Barcode Database (1,318,574 Sequences/132,900 Species/50,814 Interim Species)**
Subset of the Species library with a minimum sequence length of 640bp and containing both public and private records. This library is intended for short sequence identification as it provides maximum overlap with short reads from the barcode region of COI.

Enter sequences in fasta format:

```
>GadusMorhuaSCFAC839-06
CCTTTATCTCGTATTTGGTGCCTGAGCCGGCATAGTCGGAACACCCCTAACCCCTGCTCATTGCGAGAGACCTAAG
TCAACCTGCTGCACCTTCTTGGTCATGATCAAATTTATAATGTGATCGTTACACGGCAGCGTTTCGTAATAATTTT
CTTTATAGTAATACCACTAATAATTGGAGGCTTTGGGAACCTGACTCATTCTCTAATGATCGGTGCACAGATAT
AGCTTTCCTCGGAATAAATAACATAAGCTTCTGACTTCTTCTCCATCTTCTGCTCTTTTAGCATCTCTGG
TGTAGAAGCTGGGGCTGGAACAGGCTGAACCTGCTATCCACCTTTAGCCGGAACCTCGCTCATGCTGGGGCATC
TGTGATCTCTCAATTTTCTCTCATCTAGCAGGATTTTCATCAATCTTGGGGCAATTAATTTATACAC
AATTATTAATATGAACCTCCGGCAATTTACAGTACCAACACCCCTATTCTTTGACCACTAATTTATACAGC
TCTGCTCTACTATTATCTCTCCCGCTTACGACCTGCTATCACAATCTTCAACTGACCGTAATCTTAATAC
TTCTTTCTTTGACCTGCTGGAGGATGATCCCATTTATACCAACA
```

[feedback](#)

[Submit](#)

7.9 Quality Assurance

Extraction Control – negative control

This is included to check for extraction kit contamination. Only negligible DNA should be detected during quantification (<2ng/μl). If significant levels of DNA are detected, sterilize all equipment and repeat DNA extractions.

PCR Amplification – negative control

This is included to check for background laboratory contamination. No PCR product/band should be produced during procedure 7.4 PCR Product Check.

PCR Amplification – positive control

This should yield a strong PCR product/band (7.4 PCR Product Check) to ensure there are no issues with amplification.

7.10 Issues with Interpreting the Species Identification

1. **Every sample** should have **results** from searches in one (and occasionally two databases) with corresponding **screen-grabs and tree summarising** the results. BOLD may yield a completely unambiguous identification, but further interpretation of the results may be required to try and find the clearest species identification (as in 7.8). However, there will be **cases where** a species is lacking from the database, making a **species level identification impossible**. Despite this, the results may still yield other broader taxonomic information e.g. the genus or family the sample is likely to have originated from. It is important to note that the database is continually being updated and is becoming more comprehensive over time.

2. Another possible outcome is that despite examining the highest matching records and the tree, the **identification remains ambiguous** e.g. two species have 100% similarity to the uploaded sequence, so the end result is an ambiguous match to both. Some commercial groups of fishes, e.g. some *Thunnus* species of tunas and *Sebastes* species of “redfish”, are very closely related/difficult to distinguish and further testing may be required to successfully identify them. In such circumstances **laboratories should** indicate on the official reporting that the **sample was identified to the genus level**, (e.g. *Thunnus* spp), and/or indicate the only two species creating ambiguity (e.g. *Thunnus albacares* or *Thunnus obesus*). However, this SOP is designed to be as universal as possible (i.e. applicable to the broadest range of fish products) and generates positive information to distinguish species (even if this may not always yield a match down to the species level).

3. In case the BOLD database is unable to identify your sequence, other publically available reference databases could be queried, e.g. GenBank (www.ncbi.nlm.nih.gov/). However, the correct use of these databases falls out of the scope of this SOP.

8. TROUBLE SHOOTING

Section 7.8 of the Standard Operating Procedure (SOP) explains step-by-step the methodology to correctly identify a sample using the Barcode of Life Database. The effectiveness of the SOP was tested by conducting a collaborative blind ring trial among 13 different laboratories. Yet, some inconsistencies were recorded in the correct identification of the samples. Here, guidelines are presented to address these specific issues in utilising BOLD for species identification.

8.1 How to interpret the results of the phylogenetic tree

BOLD uses neighbour-joining trees, which group sequences together hierarchically, based on the number of amino acid or nucleotide differences. The arrangement of the specimens in the tree is based on sequence similarities, with the sequences that are most similar placed closer together on the tree, and with the branch length proportional to the degree of similarity. The percentage of similarity between sequences can be measured against the legend (usually 2%). The longer the branch the more disparity between the sequences, as specimens of the same species have more similar sequences and cluster closer together than specimens from different species.

Robust species identification is straightforward when the unknown sequence is clustered within a monophyletic group containing reference sequences of just one species. In phylogenetics a clade or 'monophylum' is a group of species/records consisting of an ancestor and all its descendants. The ancestor may be an individual, a population or even a species. In Figure 1, a 'cladogram' or 'phylogenetic tree' of a biological group is depicted showing the last common ancestor at the bottom of the composite tree. The blue and red subgroups on the left and right hand side of the picture represent clades, or monophyletic taxonomic groups. Each shows the last common ancestor and all descendant branches. The green, central, subgroup is not a clade; it is not 'monophyletic'; rather it represents a 'paraphyletic' group, which is incomplete because the blue clade, although descended from it, is excluded.

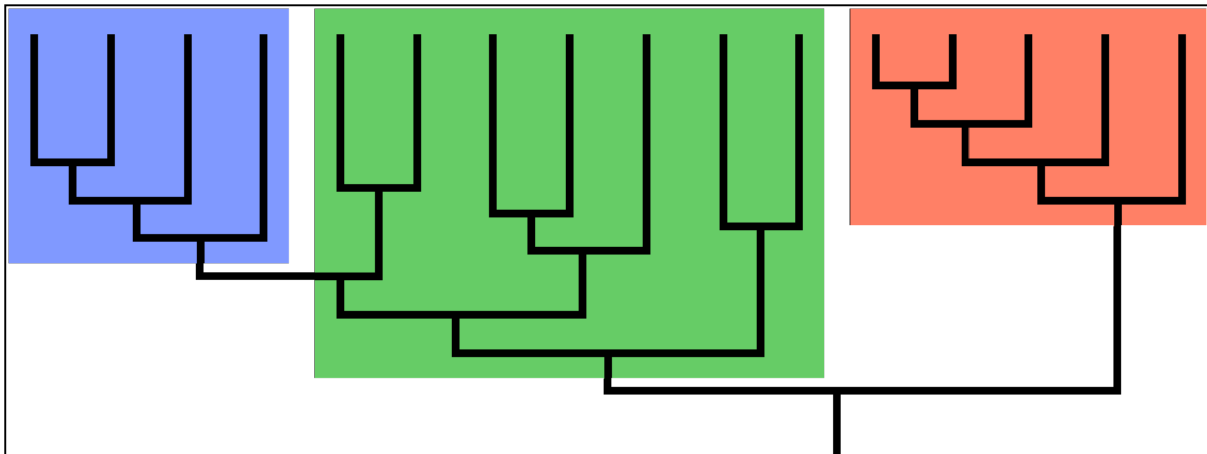


Figure 1 A tree or cladogram showing the last common ancestor at the bottom of the composite tree

Samples that fail to produce monophyletic trees, with species failing to form monophyletic groups can complicate the interpretation of results. Unexpected or ambiguous identification outcomes can reveal interesting findings, which could be associated with biologically relevant patterns, or they can reveal errors such as misidentification in the data base record or contamination of a sample (See below for a worked example; 8.2). For more information on how to build a Taxon ID Tree, and the parameters you can select to tailor your tree, please refer to the [BOLD Handbook](#).

8.2 Does the sequences length influence the ability to identify the species?

The Taxon tree functionality allows generating dendrograms from sequences using the Neighbour joining algorithm. This option is based on the distance matrix which is generated by aligning the sequences. BOLD provides an integrated alignment browser that allows users to analyse and edit sequences without needing to import them into a 3rd party software. The available BOLD alignment options are either based on amino-acid or nucleotide sequences. In turn, the distance matrix is calculated using the Kimura2 Parameter (default), Jukes Cantor or pairwise Distance models. More specifically, the tool identifies consensus bases from each group, compares them to those from the remaining sequences in other groups and then characterizes how unique each consensus base is.

Short sequences may influence the shape of the ID tree as they are less likely to align correctly with longer sequences and results should be interpreted with caution.

8.3 Identifying false identified species in the reference database

All submitted sequences on BOLD are assigned to a project, with a minimum of information requirements like the taxonomy, specimen detail and collection data. The person who creates a project is automatically assigned as the project manager. The process of creating a unique

sequence and project ID is useful to identify the sequences used throughout the barcoding process, but also as quality assurance.

Some records in the BOLD database may be flagged. This could indicate a contaminated sequence or a misidentified species. Flags serve two purposes: they act as an alert to inform project managers that an issue has been detected in their records, and they prevent a record from being included in the BOLD ID Engine and Taxonomy Browser.

Project managers can change the taxonomy of the sample or re-edit the sequence to resolve the flag. On the other hand, users that detect an issue with one of the records on BOLD Public Data Portal can log into BOLD to add a comment or a tag to that particular record. If you do not have a BOLD account, you can contact the BOLD support staff by emailing support@boldsystems.org. We strongly advise our SOP users to help the scientific community by alerting the BOLD support team in case of doubt. Through this collaborative action we all build on the quality assurance of the reference database.

Example:

Specimen sequences are identified by looking up the closest match of sequence similarity with the available records in the various BOLD search reference databases.

As part of the LABELFISH ring-trial, one of the test samples returned the following sequence:

```
TTNNAACGACGCGCCAGTCGACTAATCATAAAGATATCGGCACCCCTTTATCTCGTATTTGGTGCTTGAGCC
GGAATAGTAGGGACTGCCTTAAGTCTGCTCATTGAGCGGAACCTAAGCCAGCCTGGCGCCCTTTAGGGGA
CGACCAATCTATAATGTCATTGTTACAGCACACGCATTTGTAATAATTTTTTCATAGTAATACCAATTATAAT
CGGAGGTTTCGGAACTGACTTATTCCACTCATGATCGGTGCCCCGACATAGCATTCCCCCGTATGAATAA
TATGAGCTTCTGACTCCTCCCCCTTCATTCTTCTACTCCTTGCCCTCCTCTGGTGTTGAAGCCGGGGCCGG
TACTGGGTGAACAGTCTACCCACCACTAGCAGGGAACCTTGCCACGCAGGTGCATCAGTTGACTTAACTAT
CTTTCCCTCCACCTAGCCGGAATTTTCATCCATTCTTGGGGCCATTAATTTTCACTACTACCATTATTAATATGA
AACCCCGAGCCATTTTACAATACCAACGCCACTATTTGTGTGAGCCGTCTTAATTACAGCTGTCCTTCTTCT
TCTGTCCCTCCCGAGTCTTGCTGCGGGAATTACTATGCTCCTCACAGACCGAAACCTAAACACCACCTTCTTT
GACCCAGCCGGAGGAGGGGACCCAATTCTTTACCAACATCTTTCTGATTCTTCGGACACCCTGAAGTGTCAT
TAGCTGTTTCNG
```

The BOLD search generated the outcome in Figure 2. At the top, the search result is returned, with no unambiguous match, but two potential species returned (*Boops boops* and *Oblada melanura*, top left), prompting BOLD to display the message

“A species match could not be made”, the queried specimen is likely to be one of the following”.

Only two of the top >99% matches are recorded as *Oblada melanura* (highlighted in red, Figure 2)

In order to make a robust identification several steps have to be taken:

- The graph entitled “Similarity Scores of Top 99 Matches” shows the percent similarity for each of 99 top matching records in the database against your consensus sequences. In this example, it is clearly illustrated only *Boops boops* are included in the top matches between 100% and 99.38% similarity.
- The “Top 99 matches” show that the lowest level of similarity (around 90%) appears with a species belonging to the same family, but different genus *Sarpa salpa*.

However, within the “Top 99 matches”, 2 records are reported, among *Boops boops*, as belonging to *Oblada melanura* (Figure 2), reducing the 100% certainty for correct species identification.

Apart from investigating the similarity levels of your sequences with the public database it is important to produce a simple Neighbour Joining tree (Figure 3).

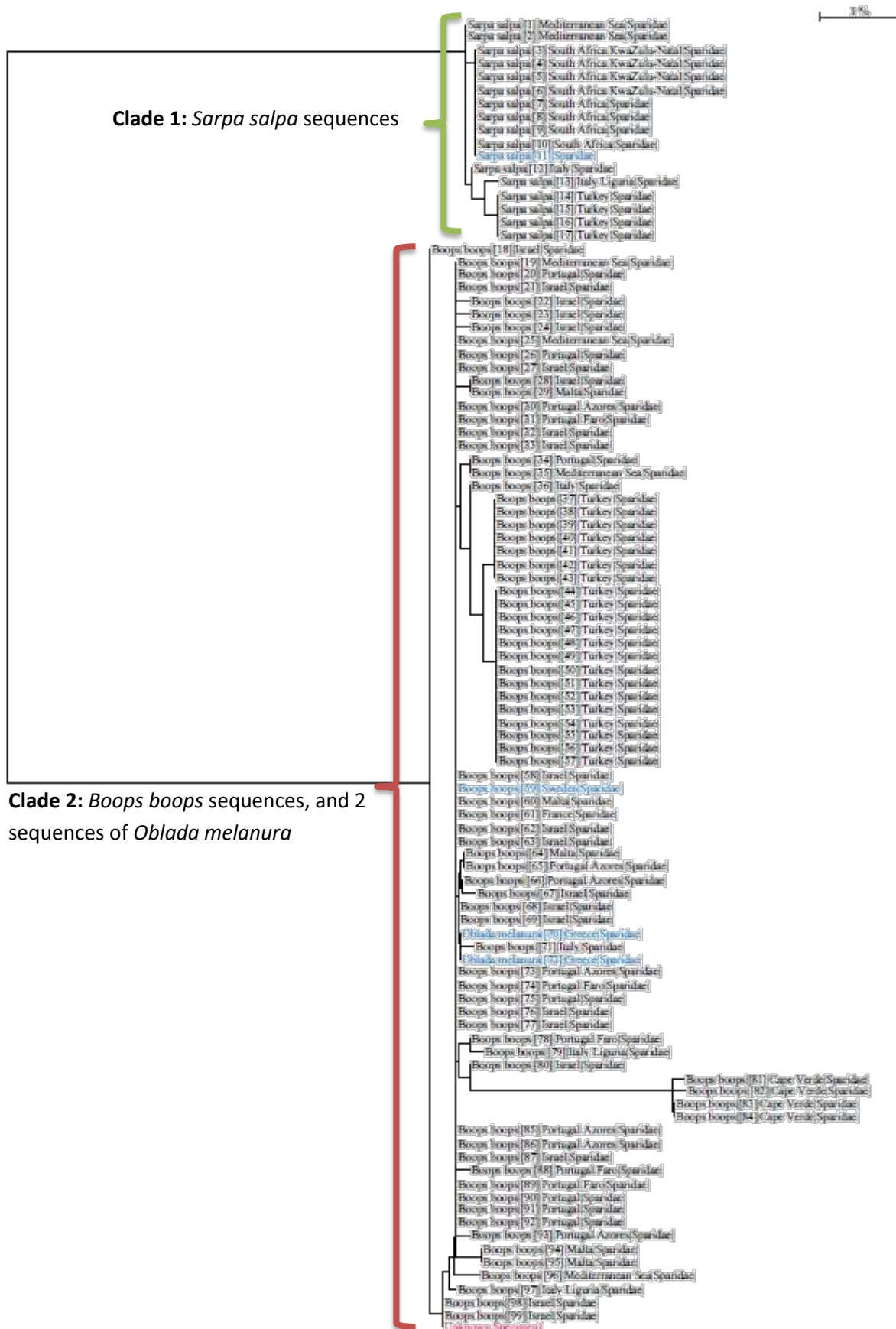


Figure 3 A Neighbouring tree can be consulted when identifying a sequence on BOLD. The example shown here illustrates that the unknown sequence (in red) is clustered within a clade of *Boops boops* records, a second clade is shown composing only of *Sarpa salpa* sequences. This monophyletic clade is clearly well separated from our unknown sequence. Within the *Boops boops* clade however, two sequences are found originating from *Oblada melanura*.

In the above tree (Figure 3) the queried sequence is highlighted in red and is nested in the tree dominated by *Boops boops* sequences. Only two sequences in this clade are recorded as *Oblada melanura*, while the BOLD database contains many more sequences of this particular species, which in fact show a similarity score far lower than 1%. The threshold in nucleotide difference between species is commonly taken at 2% (Avisé 2000). This raises suspicion whether or not these sequences are to be trusted. Further investigation would be recommended.

Consulting the BOLD Taxonomy browser allows users to examine the number of records available in BOLD of any species (Figure 4). For this specific case the Taxonomy Browser shows there are in total 43 sequence records of *Oblada melanura* and 103 *Boops boops*.

Specimen Records : 5,411,504 Specimens with Barcodes : 4,013,098 Species with Barcodes : 238,748

Animals:

- [Acanthocephala \[552\]](#)
- [Annelida \[40476\]](#)
- [Arthropoda \[4137733\]](#)
- [Brachopoda \[147\]](#)
- [Bryozoa \[1885\]](#)
- [Chaetognatha \[421\]](#)
- [Chordata \[450143\]](#)
- [Cnidaria \[15313\]](#)
- [Cyclophora \[326\]](#)
- [Echinodermata \[33658\]](#)
- [Echiura \[79\]](#)
- [Gnathostomulida \[15\]](#)
- [Hemichordata \[45\]](#)
- [Mollusca \[116434\]](#)
- [Nematoda \[9351\]](#)
- [Nemertea \[1486\]](#)
- [Onychophora \[835\]](#)
- [Platyhelminthes \[13015\]](#)
- [Porifera \[2818\]](#)
- [Priapulida \[23\]](#)
- [Rotifera \[9135\]](#)
- [Sinuicula \[446\]](#)
- [Tardigrada \[1770\]](#)
- [Xenoturbellida \[4\]](#)

Plants:

- [Bryophyta \[7659\]](#)
- [Chlorophyta \[11539\]](#)
- [Lycopodiophyta \[816\]](#)
- [Magnoliophyta \[302555\]](#)
- [Pinophyta \[6689\]](#)
- [Pteridophyta \[9575\]](#)
- [Rhodophyta \[42609\]](#)

Fungi:

- [Ascomycota \[71118\]](#)
- [Basidiomycota \[41630\]](#)
- [Chytridiomycota \[233\]](#)
- [Glomeromycota \[3072\]](#)
- [Mycetozoa \[222\]](#)
- [Zygomycota \[2693\]](#)

Protists:

- [Chlorarachniophyta \[67\]](#)
- [Ciliophora \[561\]](#)
- [Heterokontophyta \[17364\]](#)
- [Pyroplasmophyta \[2222\]](#)

BOLD SYSTEMS Databases Taxonomy Identification Workbench Resources

Oblada melanura {species} - [Chordata](#); [Actinopterygii](#); [Perciformes](#); [Sparidae](#); [Oblada](#);

Sub-taxa

BOLD Stats

Specimen Records:	46	Public Records:	39
Specimens with Sequences:	43	Public Species:	1
Specimens with Barcodes:	37	Public BINs:	6
Species:	1		
Species With Barcodes:	1		

Figure 4 The top panel illustrates the BOLD Taxonomy search engine which is accessible through the home page by clicking on the Taxonomy tab. Whenever needed, taxa can be found by using the search option. The lower panel depicts the outcome for *Oblada melanura*, illustrating that the BOLD reference database contains 43 sequences.

Second, in the “Top 99 matches” table (see Figure 2), one can check the details of all publically available records by clicking the blue arrow symbol: . The specimen record of both these *Oblada melanura* records appear to have been flagged and have received a tag illustrating the sequence could be misidentified (see red, Figure 5).

Consequently, the identification of the sequences as *Boops boops* is substantially strengthened as we have some grounds to discount the *Oblada melanura* records as being unreliable

BOLD SYSTEMS Databases | Taxonomy | Identification | Workbench | Resources

Public Data Portal - Specimen Record [Print](#)

Search Public Data [XML](#) [TSV](#) [FASTA](#) [TRACE](#) [XML](#) [TSV](#) [Show Help](#)

Specimen Data Sequences Combined

Record Details For ANGBF7288-12 [Back to Search: Records](#)

IDENTIFIERS:

Sample ID:	HM590730	Museum ID:	
Field ID:	HM590730	Collection Code:	
Deposited In:	Mined from GenBank, NCBI		

Tags & Comments [Add Tag](#) [Comments](#) [Associated Tags](#) [Mis-identified](#)

TAXONOMY:

Phylum:	Chordata	Subfamily:	
Class:	Actinopterygii	Genus:	<i>Oblada</i>
Order:	Perciformes	Species:	<i>Oblada melanura</i>
Family:	Scorpaenidae	BIN (Cluster ID):	BOLD:AAE7806

* Barcode Index Numbers(BIN): cluster barcode sequences to create OTUs that closely reflect species groupings

SPECIMEN DETAILS:

Voucher Status:		Reproduction:	
Tissue Descriptor:		Sex:	
Brief Note:		Life Stage:	
Detailed Notes:	<i>Oblada melanura</i>		

Specimen Images: N/A
Collection Site: N/A

Attribution:

Specimen Depository: Mined from GenBank, NCBI
Sequencing Center:
Photography:
Collectors:
Specimen Identification:
Project Manager: Dirk Steinke
Sequencing Support:

Figure 5. Specimen record of one of the *Oblada melanura* sequences. Whenever a sequence has been flagged, or a tag has been added to indicate issues with the relevant sequences, this would be observed here.

8.4 Identifying unknown species

If a sample comes from a rare or exotic fish, there may be no matching records in the “Species Level Barcode Records” database that demonstrate high levels of sequence similarity. An additional search is also recommended by selecting the “All Public Records on BOLD”, this is the broadest database as it includes both sequences available on BOLD as well as on GenBank (Figure 6). This may allow the user to assign the sequence to a genus or family and still provide potentially useful information (this is one of the big advantages of utilising a big public database). The search is conducted and interpreted in the same fashion as the SOP but the selection of this expanded database must be recorded in the results.

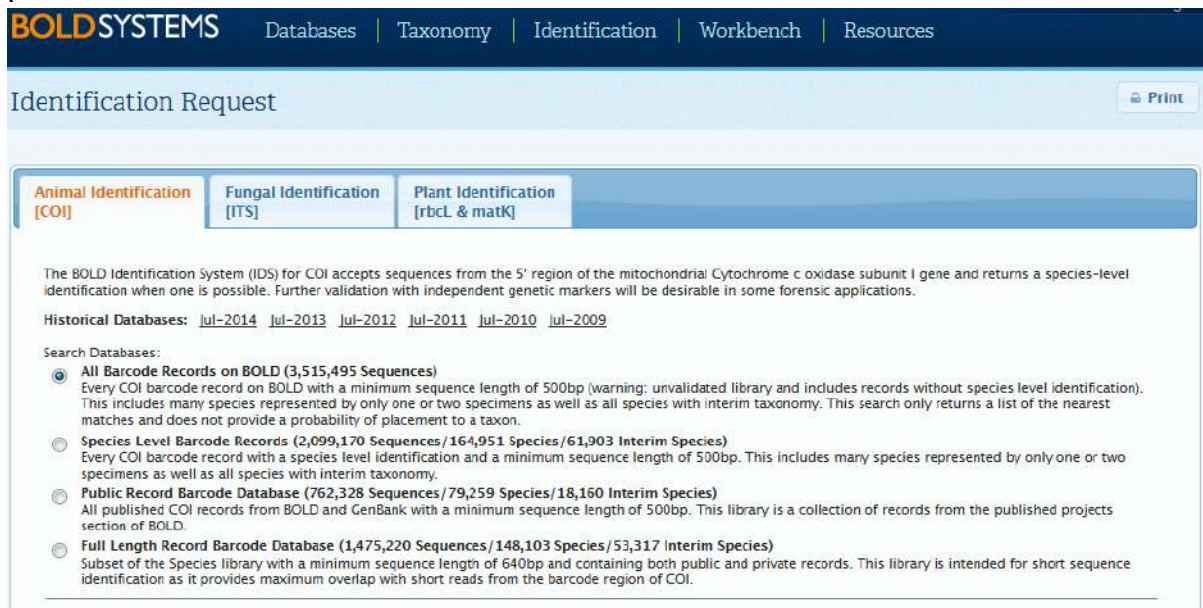


Figure 6 screenshot of the type of reference database option one can pick from on the BOLD database. The selected one illustrated the “All Public Records” option.

If *a-priori* information about the supposed species is available, i.e. from the label, it would be recommended to check if the species and the commercially significant counterparts belonging to the same genus are represented in the Taxonomy Browser. This Browser is a synthetic database that allows users to examine the progress of DNA barcoding by browsing through the different levels of the taxonomic hierarchy available on BOLD. Within this browser users are able to select between animal, plant, fungal and protist kingdoms and navigate from phylum to species level. To look up a specific taxon directly, use the search function by entering a taxonomic name into the search bar at the top of the Taxonomic Browser or on the BOLD main page.). Comparing this species list with database on marine organisms like Fishbase (www.fishbase.org) or World Register of Marine Species (WoRMS, www.marinespecies.org) allows users to determine whether all species belonging to a certain genus or family are available in the BOLD database:

Example:

Within the collaborative ring trial, two of the unknown samples belonged to the genus *Merluccius*. The Taxonomy Browser of BOLD provides an overview of all the species recorded belonging to this genus (Figure 7). Comparing this species list with database on marine organisms like Fishbase (www.fishbase.org) or World Register of Marine Species (WoRMS, www.marinespecies.org) shows that some species are not yet in the BOLD database:

- *Merluccius gayi peruanus*
- *Merluccius hernandezi*
- *Merluccius tasmanicus*
- *Merluccius patagonicus*

When a sequence from one of these species is queried in BOLD, high confidence identification will only be obtained for the genus level. Selecting the “All Public Records on BOLD”, will not increase the level of identification as neither of these species are included on GenBank. This should however not reduce the level of confidence in identifying the other *Merluccius* species.

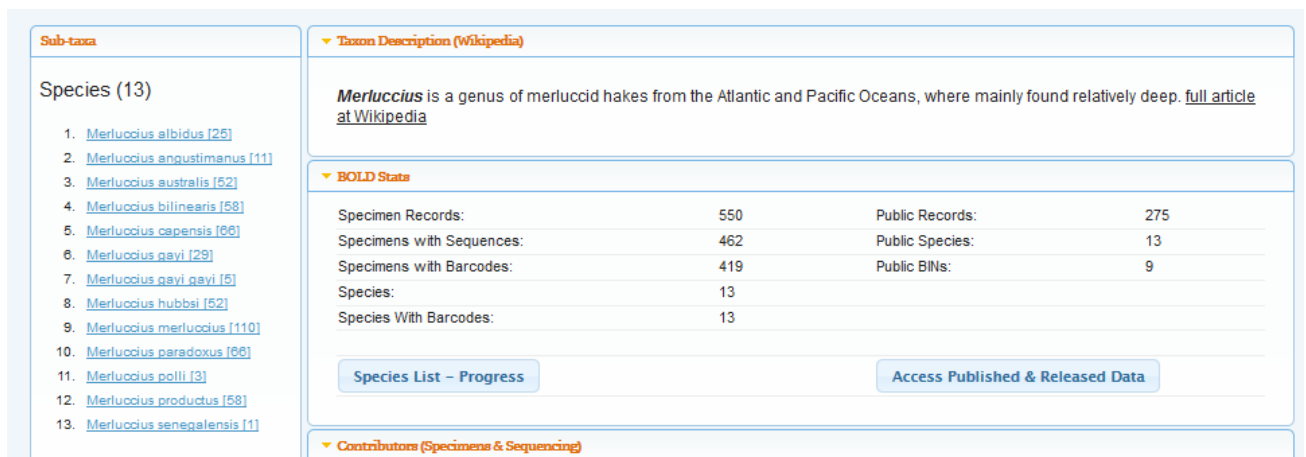


Figure 7 Taxonomy query in BOLD on the genus *Merluccius*.

An example is depicted in Figure 8 where the following sequence belonging to the *Merluccius* genus:

```
TTTGGTGCTTGAGCCGGCATAGTCGGAACAGCCCTAAGCCTGCTCATCCGGGCAGAACTAAGTCAACCCGG
CGCACTCCTGGGCGACGATCAAATTTATAACGTAATCGTCACGGCACACGCCTTCGTAATAATTTCTTTATA
GTAATACCGTTAATAATTGGAGGCTTTGGAACTGACTCGTTCCCCTAATGATCGGAGCCCCGACATGGCC
TTCCCCGAATAAATAACATAAGCTTCTGACTTCTTCCCTCCGTCTTTCTGCTCCTCCTAGCATCCTCCGGAG
TAGAAGCCGGAGCCGGGACAGGTTGAACAGTATACCCCCCTCTTGCAAGCAATCTTGCCCACGCTGGCGCC
AGCGTGGACCTCACTATTTTTTCACTTCACCTAGCAGGCGTTTCTCAATTCTAGGAGCAATTAATTTCAATTAC
TACTATTATTAATATGAAACCCCTGCAATCTCACAGTACCAGACACCCCTCTTTGTTTGATCCGTCCTTATTA
CAGCTGTCCTCCTCCTACTCTCCCTGCCCCGTCTTAGCCGCCGGCATCACAACTACTAACTGACCGAAACC
TCAACACCTCCTTCTTTGACCCCGCCGGTGGAGGGGACCCCATCCTATACCAGCATTT
```

is identified using BOLD. The similarity score illustrate the sequences belongs to *Merluccius hubbsi* (% similarity 100-99.82, Figure 8). The next encountered species belongs to *Merluccius productus* (98.58%), after that the level of similarity drops steeply for *Merluccius albidus* (96.23%, see red arrow Figure 8).

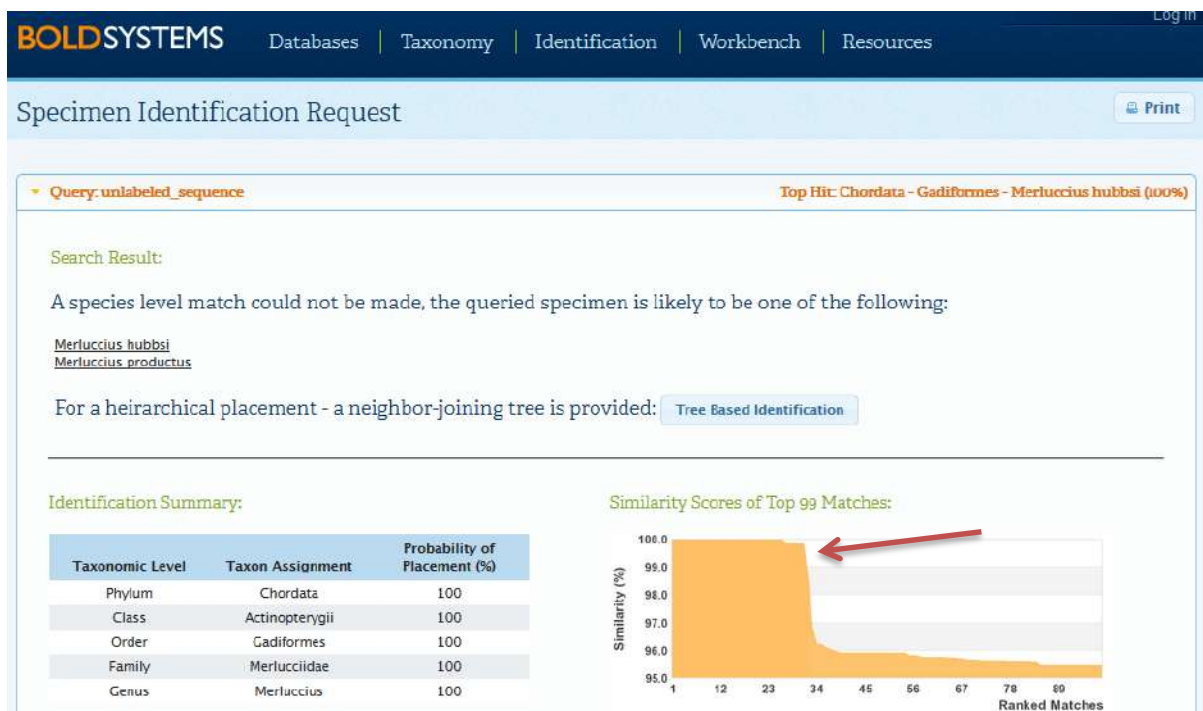


Figure 8 Specimen Identification output on BOLD. Top 99 Matches show the highest similarity scores are obtained for the species *Merluccius hubbsi* (100-98.58). Moving further down in the species list, the similarity score drops sharply to 96.82%.

Constructing a phylogenetic tree will help to further specify which species match has the highest confidence (Figure 9). The unknown sequence is clearly nested within sequences originating from *Merluccius hubbsi*. All other *Merluccius* species form a separate branch some distance away from our unknown specimen. Except for one sample recorded as *Merluccius productus* (Figure 9, blue arrow). Although this specific record does not have a tag added, the appearance of all other *Merluccius productus* records in another clade may indicate this particular sequence was indeed misidentified or should at least be treated with less confidence. As users you may want to raise your concern by contacting the BOLD support staff by emailing support@boldsystems.org or if you have a BOLD account by adding a comment or a tag to that particular record.

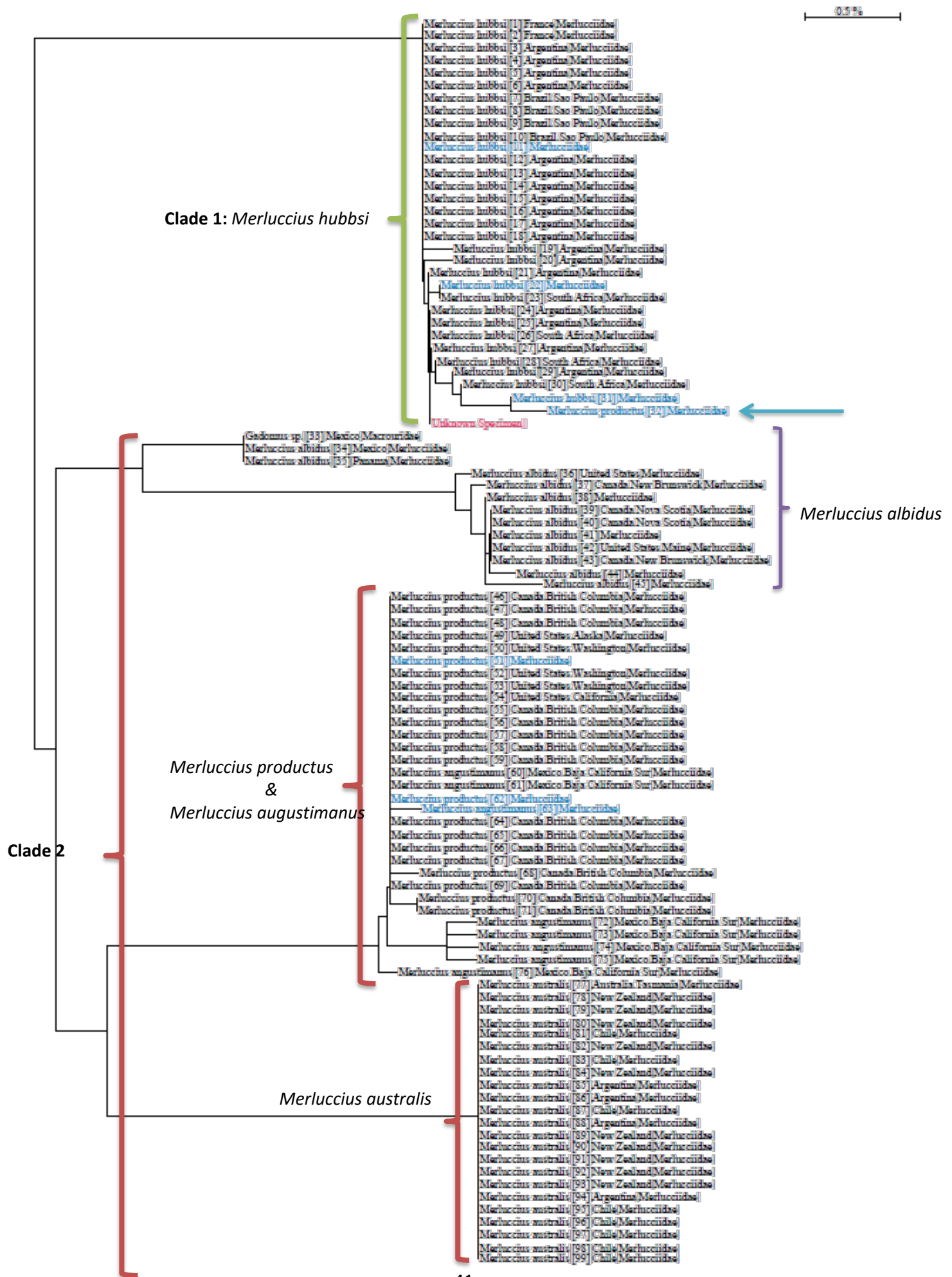


Figure 9 Phylogenetic tree of *Merluccius* species

8.5 Specific problems regarding the identification of *Thunnus* species

Due to recent divergence or introgression events, there are limitations associated with distinguishing between *Thunnus* species with the COI barcoding gene (these limitations are common amongst many methods). Especially problematic is the differentiation between i) *Thunnus albacares* and *Thunnus obesus* (Vinas & Tudela 2009, Hanner et al. 2011, Pedrosa-Gerasmio et al. 2012, Santini et al. 2013) and ii) *Thunnus thynnus*, *Thunnus orientalis* and *Thunnus alalunga* (Dawnay et al. 2007).

Identification of either *Thunnus albacares* or *Thunnus obesus* with high confidence without using additional methods (using additional markers like *cytb* or D-loop) is therefore difficult. The sequence similarity of the COI gene between these two species is high as they share some haplotypes and consequently constructing the phylogenetic tree in BOLD will not provide more detailed insight (Figure 10). Nevertheless, a BOLD search will still provide reliable identification to the genus level e.g. *Thunnus*, *Sarda* and *Katsuwonus*, which may still provide valuable information. Below we illustrate the difficulty to identify a sample of *Thunnus alalunga*.

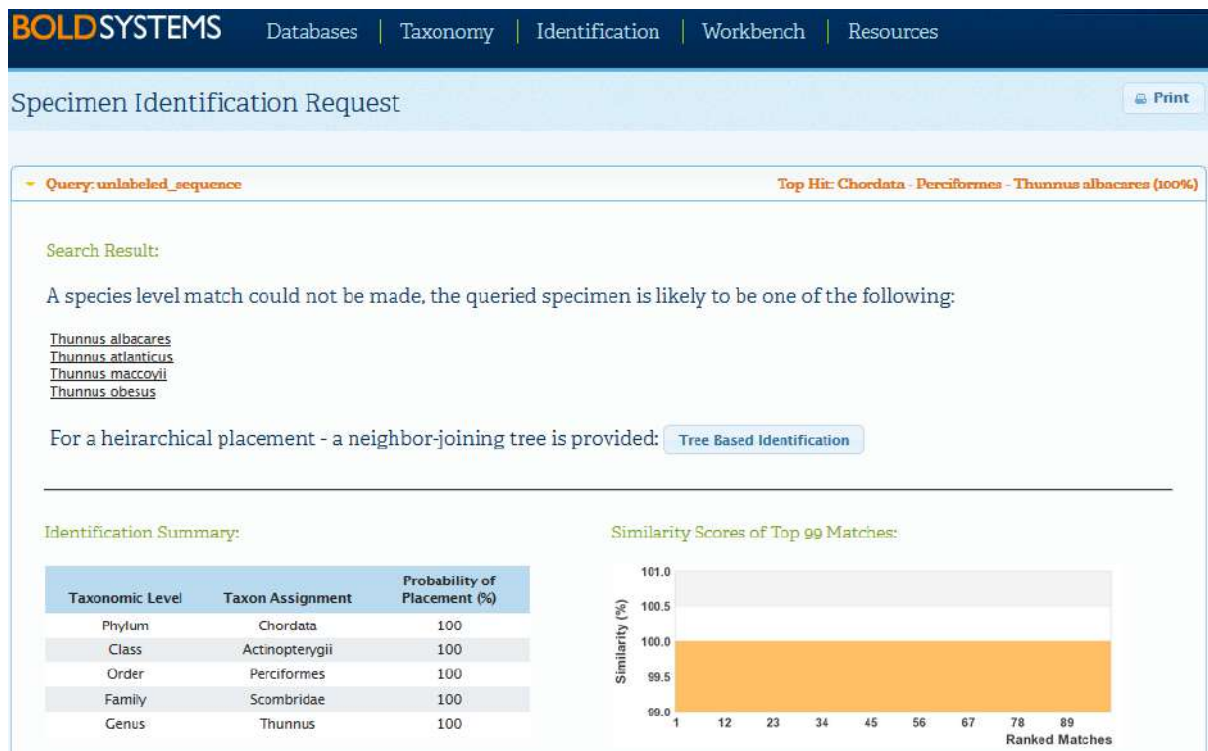


Figure 10 Specimen Identification output on BOLD for a *Thunnus albacares* sample. Clearly the BOLD database cannot distinguish between 4 different *Thunnus* species as some sequences exist under different names, which show a level of similarity of 100%.

First, the enquiry will reveal the sample belongs to the *Thunnus* genus, which should alert the user on possible difficulties for correct species level identification. Subsequently, one could focus only on the 100% similarity score matches as these are most robust. Depending on the chosen reference database, published and private or only published records, the 100% similarity scores will contain only *Thunnus alalunga* or *Thunnus alalunga* and *Thunnus obesus* respectively. The phylogenetic tree may further help determine to which species the unknown sample belongs.

The phylogenetic tree obtained from a *Thunnus alalunga* sample is depicted in Figure 11. The unknown sample is highlighted in red. Although the tree is unresolved (the different species in the tree are not represented in a single branch) which makes interpretation of the results complicated, we like to explain some insights into the species identification.

Three clusters can be identified in the tree:

- Cluster 1 contains only *Thunnus alalunga* sequences
- Cluster 2 consist of *Thunnus alalunga* sequences, and 2 *Thunnus obesus* records (see red arrow)
- Cluster 3 is nested within Cluster 2, but contains no *Thunnus alalunga* sequences

The longer branch of cluster 3 indicate lower similarity with the other two clusters (see 8.1), reducing the chance of our unknown sequences to be identified as either *Thunnus orientalis* or *Thunnus thynnus*.

Cluster 2 contains our unknown sample, 2 records of *Thunnus obesus* both not flagged, and sequences of *Thunnus alalunga*. One of the *Thunnus obesus* records showed up in the 100% similarity match. Elaborating on this, one could identify the sample as *Thunnus alalunga* based on the knowledge that out of the 87 sequences (composing clusters 1 & 2) only 2 records were from to *Thunnus obesus*, representing 2.3% chance of incorrect species identification. Additionally, the BOLD database contains 264 species records of *Thunnus obesus*, of which 190 are with barcode. In case the sample would be a *Thunnus obesus* more BOLD records of this species would be expected to appear. Additionally, when using the Public Record Barcode Database, instead of the Species Level Barcode Records (default reference database option in BOLD) 100% similarity matches will in most cases only return as either one of the two species, here 100% similarity returned *Thunnus alalunga*.

In such circumstances a laboratory could identify the sample as i) *Thunnus* species (very robust), ii) *Thunnus alalunga* (although can *Thunnus obesus* cannot be excluded 100%), or iii) the laboratory could suggest they need to perform an additional test if a 100% confident species identification is required.

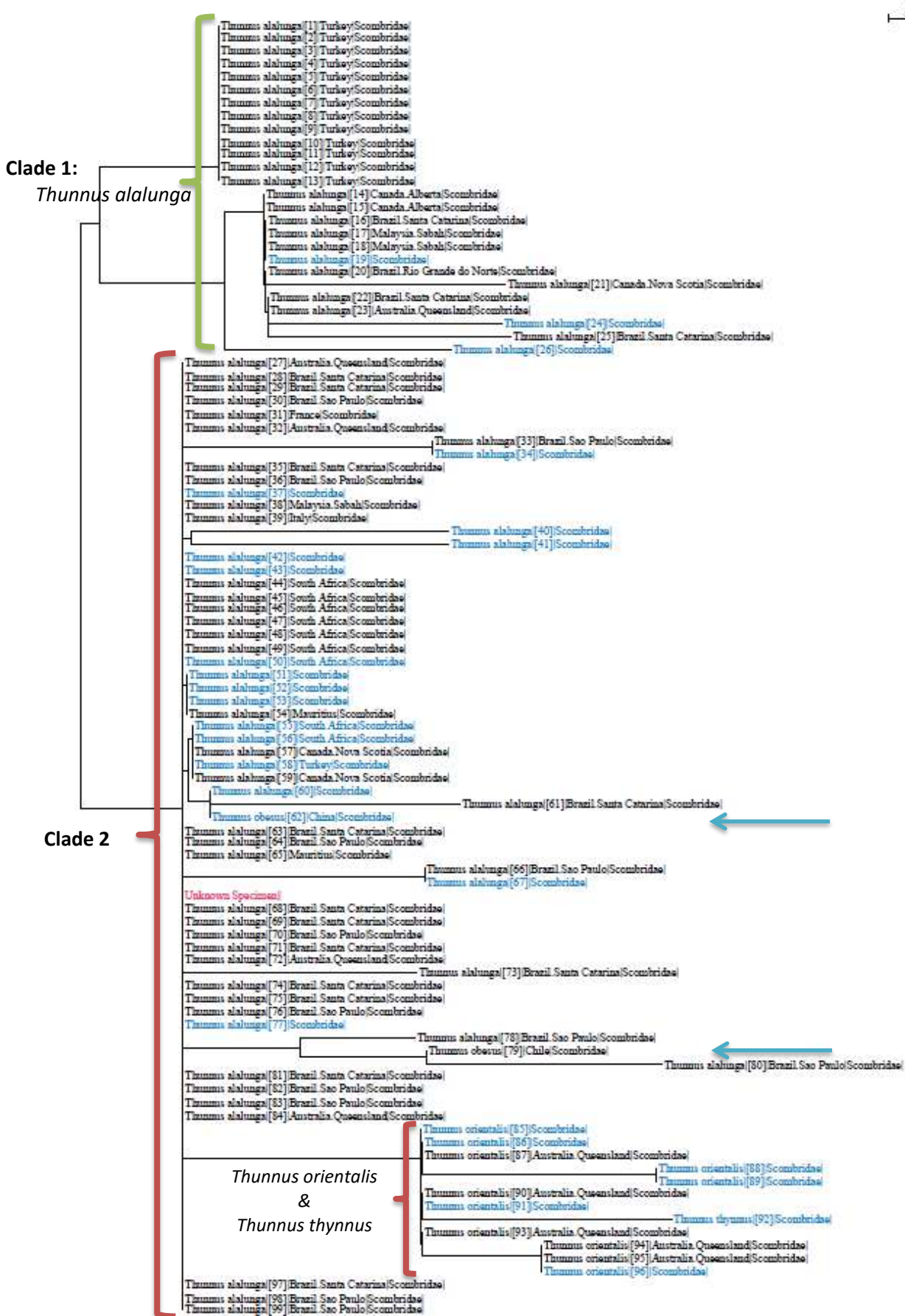


Figure 11 Phylogenetic tree constructed in BOLD for a *Thunnus alalunga* sample.

8.6 References

Avise 2000 Phylogeography. The history and formation of species. Harvard University Press, Cambridge, MA.

Dawnay et al. 2007 Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Science International*, **173**: 1-6.

Hanner et al. 2011 FISH-BOL and seafood identification: Geographically dispersed case studies reveal systemic market substitution across Canada. *Mitochondrial DNA*, **22**:106-122.

Ivanova et al. 2007 Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes*, **7**: 544–548.

Pedrosa-Gerasmio et al. 2012 Discrimination of Juvenile Yellowfin (*Thunnus albacares*) and Bigeye (*T. obesus*) Tunas using Mitochondrial DNA Control Region and Liver Morphology. *PlosOne*, 7:e35604.

Santini et al. 2013 First molecular scombrid timetree (Percomorpha: Scombridae) shows recent radiation of tunas following invasion of pelagic habitat. *Italian Journal of Zoology*, **80**: 210-221.

Vinas & Tudela 2009 A validated methodology for genetic identification of tuna species (Genus *Thunnus*). *PlosOne*, 4: e7606.

Ward et al. 2005 DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B*, **360**: 1847–57.

Additional Resources:

General information and a solid background to DNA barcoding are available below, including access to the barcode of life online community (including a forum that can potentially provide troubleshooting advice);

<http://www.barcodeoflife.org/>

A comprehensive hand book for utilising the BOLD database is available;

<http://www.boldsystems.org/index.php/Resources>