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

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^{o}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 & $MgCl_2$)

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 *et al*. 2005
FishF2_t1	TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC	Ward *et al*. 2005
FishR2_t1	CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA	Ward *et al*. 2005
FR1d_t1	CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA	Ivanova *et al*. 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 $MgCl_2$ (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
ddH2O	2.7
10X buffer	2
50mM MgCl2	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. great 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

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

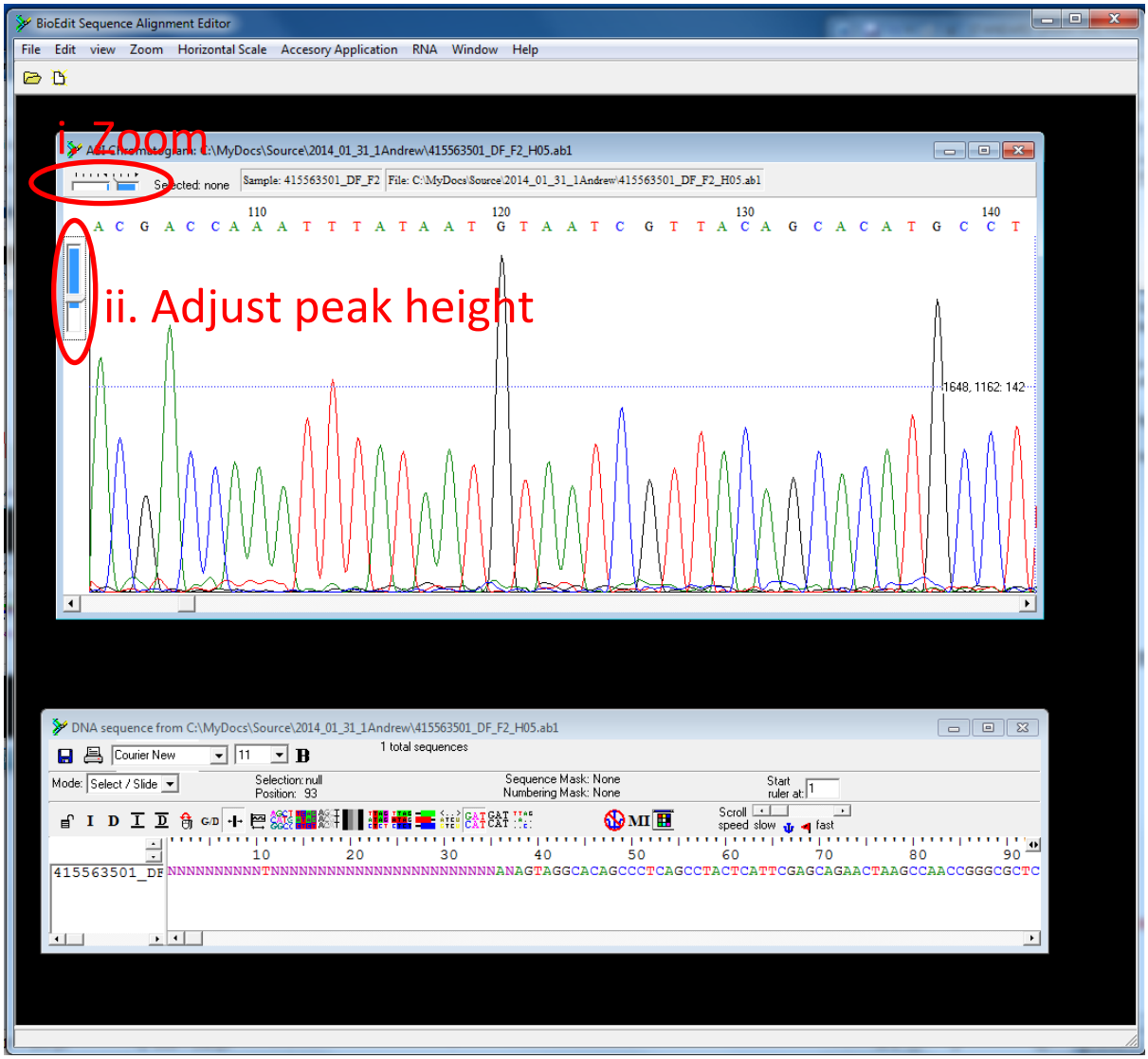
1. Open the BioEdit software by clicking on the BioEdit.exe icon

2. Open the ABI file from your sample by selecting the file menu and the open option. Select the ABI sample from your sample

3. This will open up two windows within the software; (i.) The chromatogram, i.e. the sequence trace or peaks corresponding to the signal from each of the nucleotides in the DNA sequence; (ii.) A long string of letters, predominantly made up of A, T, C, & G, which correspond to the software's interpretation of the peaks and conversion into a representative nucleotide sequence;

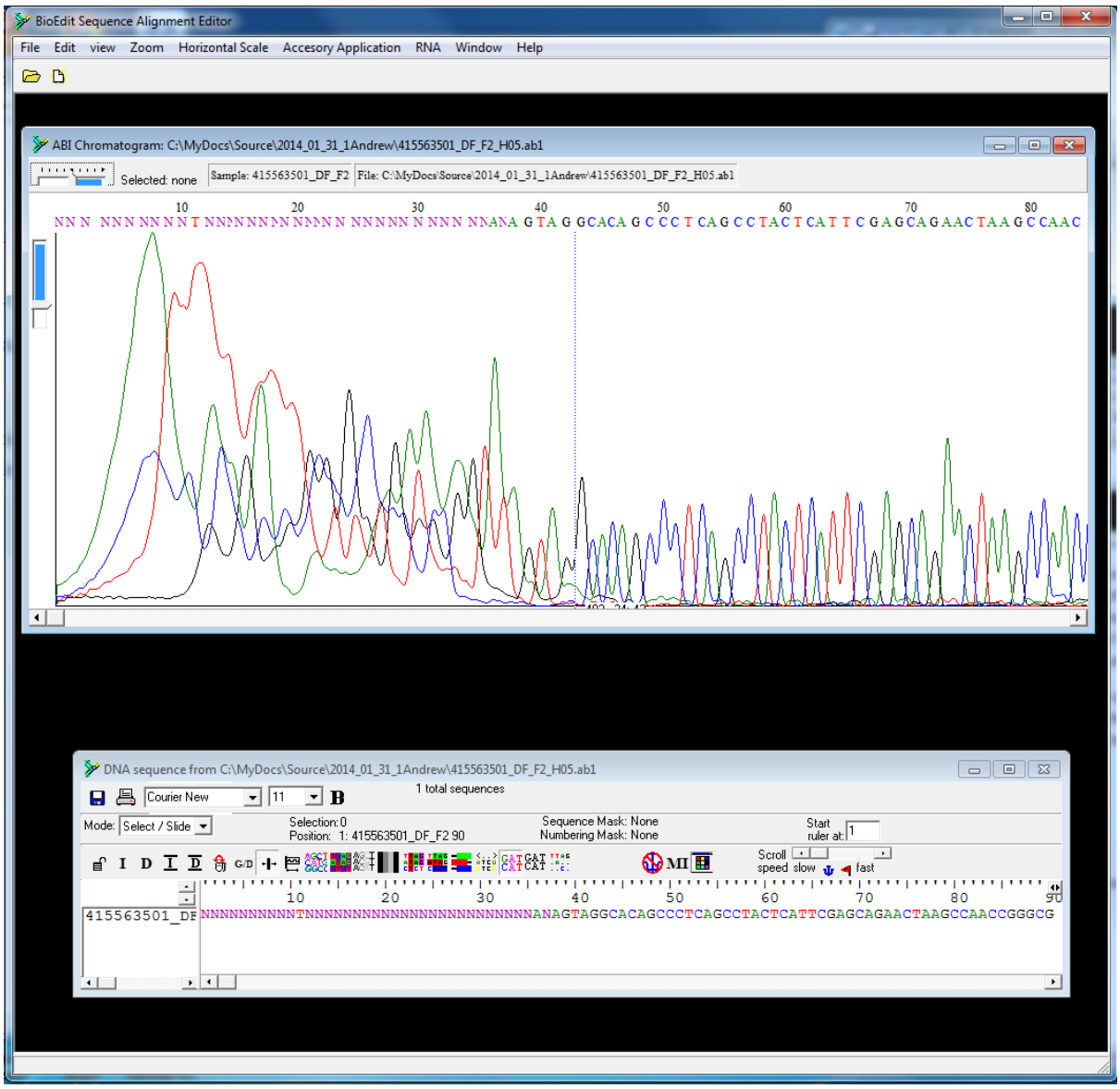

BioEdit Sequence Alignment Editor
File Edit view Zoom Horizontal Scale Accesory Application RNA Window Help
ABI Chromatogram: C:\MyDocs\Source\2014_01_31_1Andrew\415563501_DF_F2_H05.ab1
Sample: 415563501_DF_F2
File: C:\MyDocs\Source\2014_01_31_1Andrew\415563501_DF_F2_H05.ab1
110 120 130 140
i. Trace
1648, 1162: 142
DNA sequence from C:\MyDocs\Source\2014_01_31_1Andrew\415563501_DF_F2_H05.ab1
Courier New
11
1 total sequences
Mode: Select / Slide
Selection: null
Position: 93
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
Scroll speed slow fast
10 20 30 40 50 60 70 80 90
415563501_DF
NNNNNNNNNNTNNNNNNNNNNNNNNNNNNNNNNNNNANAGTAGGCACAGCCCTCAGCCTACTCATTCGAGCAGAACTAAGCCAACCGGGCGCTC
ii. Nucleotide sequence

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;

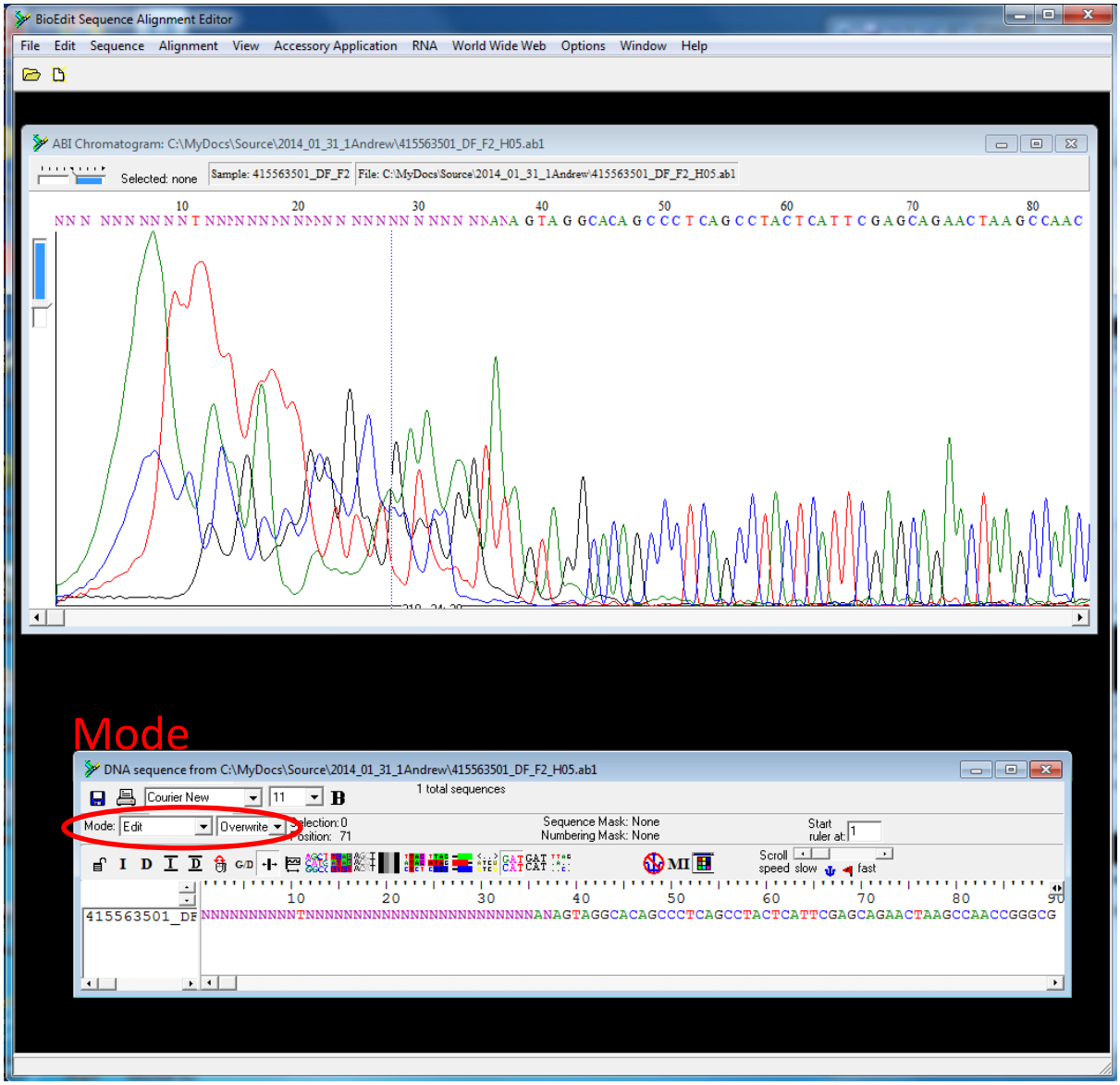


5. The sequence must be checked by eye to ensure the sequence reaction has worked successfully and the trace is of high quality. The majority of the trace should consist of a series of clear peaks (as in the figure above.) If the reaction has failed, or contamination is present, the peaks will look weak 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.

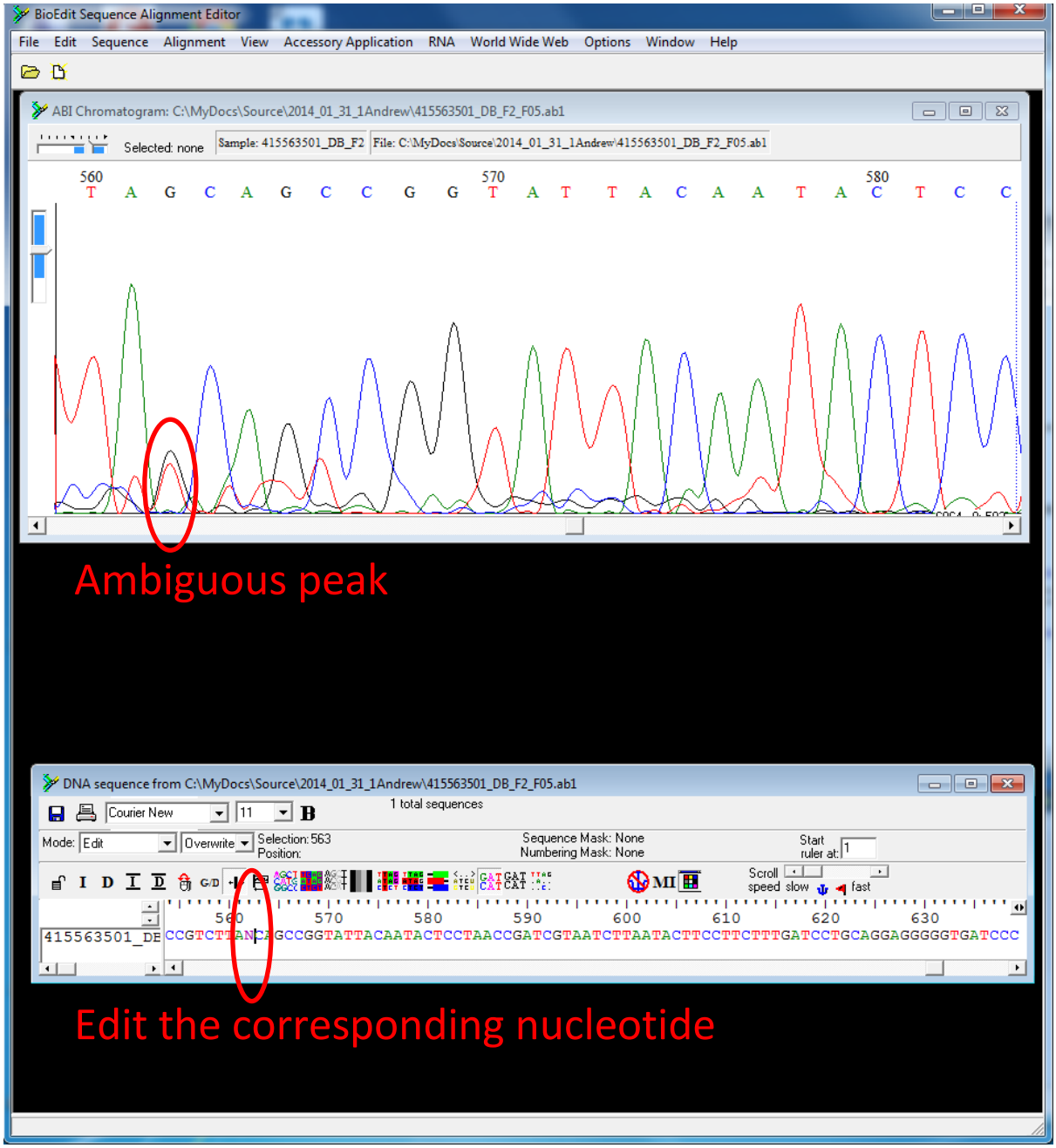
6. Often the quality of the sequencing reaction is poor at either ends of the trace (as below). In this case the ambiguous region at either end can simply be deleted, just leaving the high quality sequence (i.e. delete the flanking sequence at each end until you are confident that you can easily call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


BioEdit Sequence Alignment Editor
File Edit view Zoom Horizontal Scale Accesory Application RNA Window Help
ABI Chromatogram: C:\MyDocs\Source\2014_01_31_1Andrew\415563501_DF_F2_H05.ab1
Selected: none
Sample: 415563501_DF_F2
File: C:\MyDocs\Source\2014_01_31_1Andrew\415563501_DF_F2_H05.ab1
10 20 30 40 50 60 70 80
NN N NNN NN N N T NNNNNNNNNNNNNNNNNNNNNNNNNNNNNANA G TA G GCACA G C C C T CA G C C TAC T CA T T C G A G CA G AA C TAA G C CAA C
DNA sequence from C:\MyDocs\Source\2014_01_31_1Andrew\415563501_DF_F2_H05.ab1
Courier New
11
1 total sequences
Mode: Select / Slide
Selection: 0
Position: 1: 415563501_DF_F2 90
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
Scroll speed slow fast
10 20 30 40 50 60 70 80 90
415563501_DF
NNNNNNNNNNTNNNNNNNNNNNNNNNNNNNNNNNNNNANAGTAGGCACAGCCCTCAGCCTACTCATTCGAGCAGAACTAAGCCAACCGGGCG

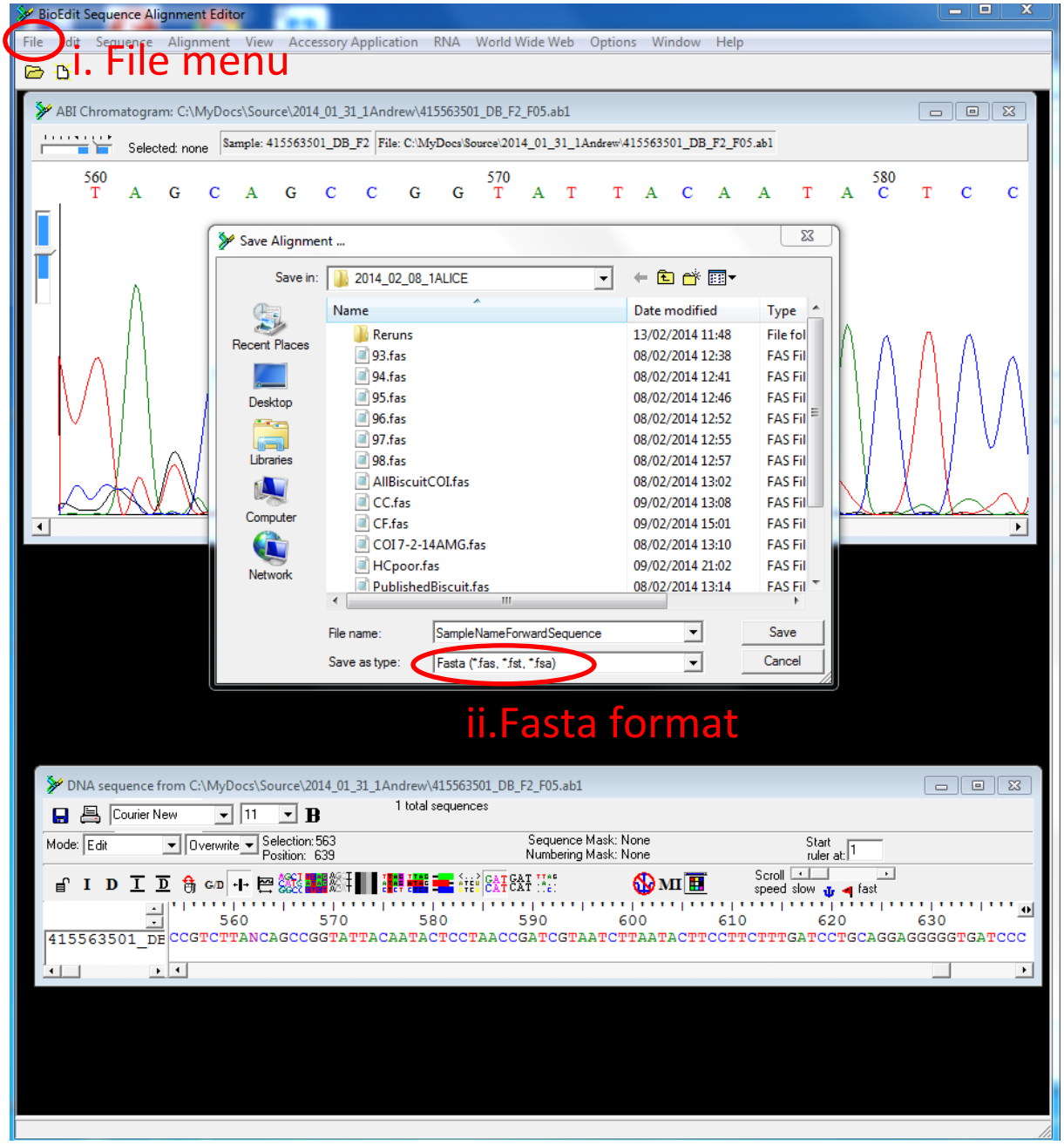
7. These low quality portions of the sequence at either end can be removed in the nucleotide sequence window. First it is necessary to switch the mode to edit, as indicated below, and then the sequence can be edited like any other text file. However, it is important to remember that the trace and sequence windows within BioEdit operate independently, so alterations to the sequence are not reflected in the trace window (meaning any edits to the sequence text will mean the nucleotide positions in the two windows will no longer match).



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). **<u>Do not delete them</u>**.



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;

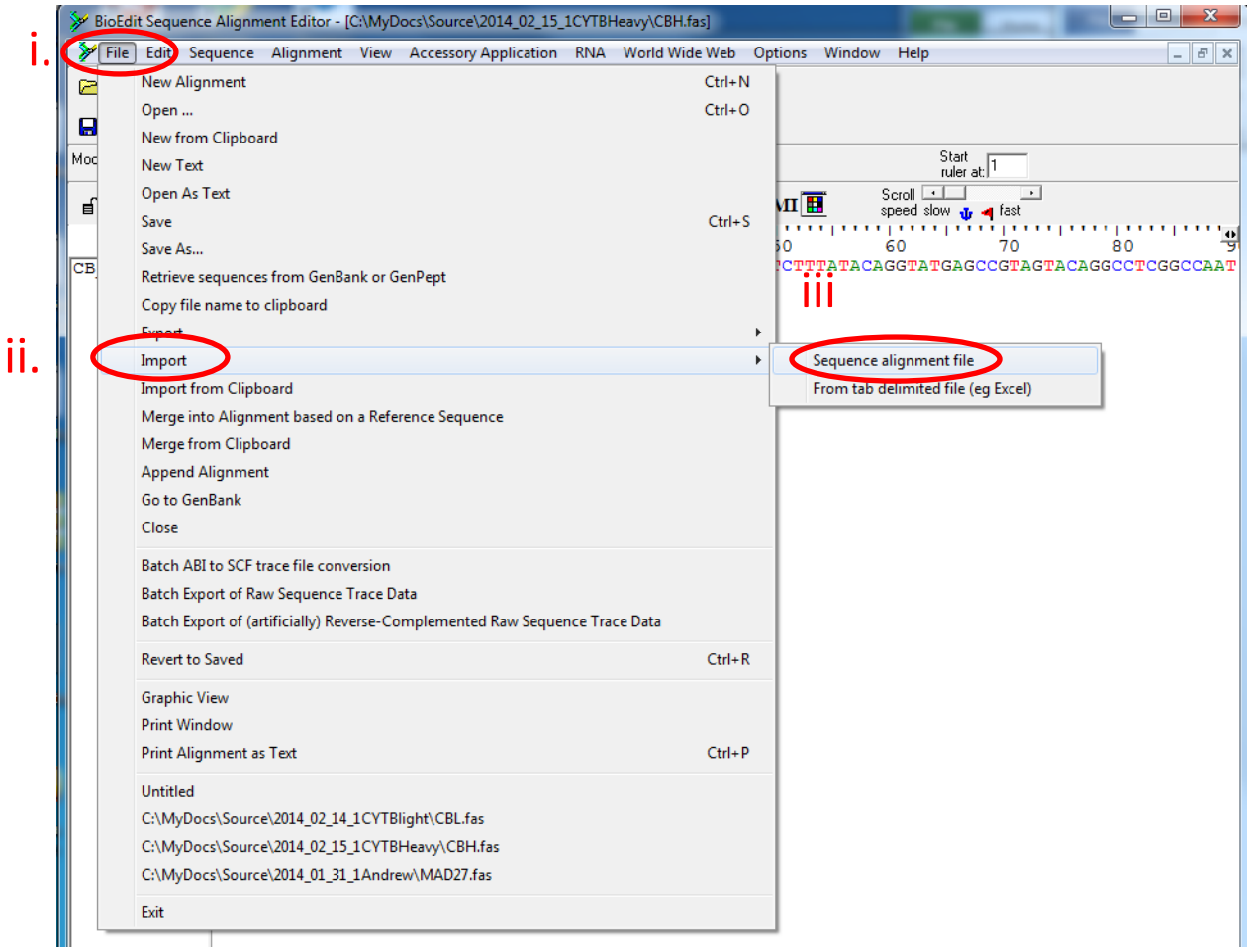
http://www.boldsystems.org/index.php/resources/handbook?chapter=3_submissions.html&section=trace_submissions

7.7 Generating a consensus sequence

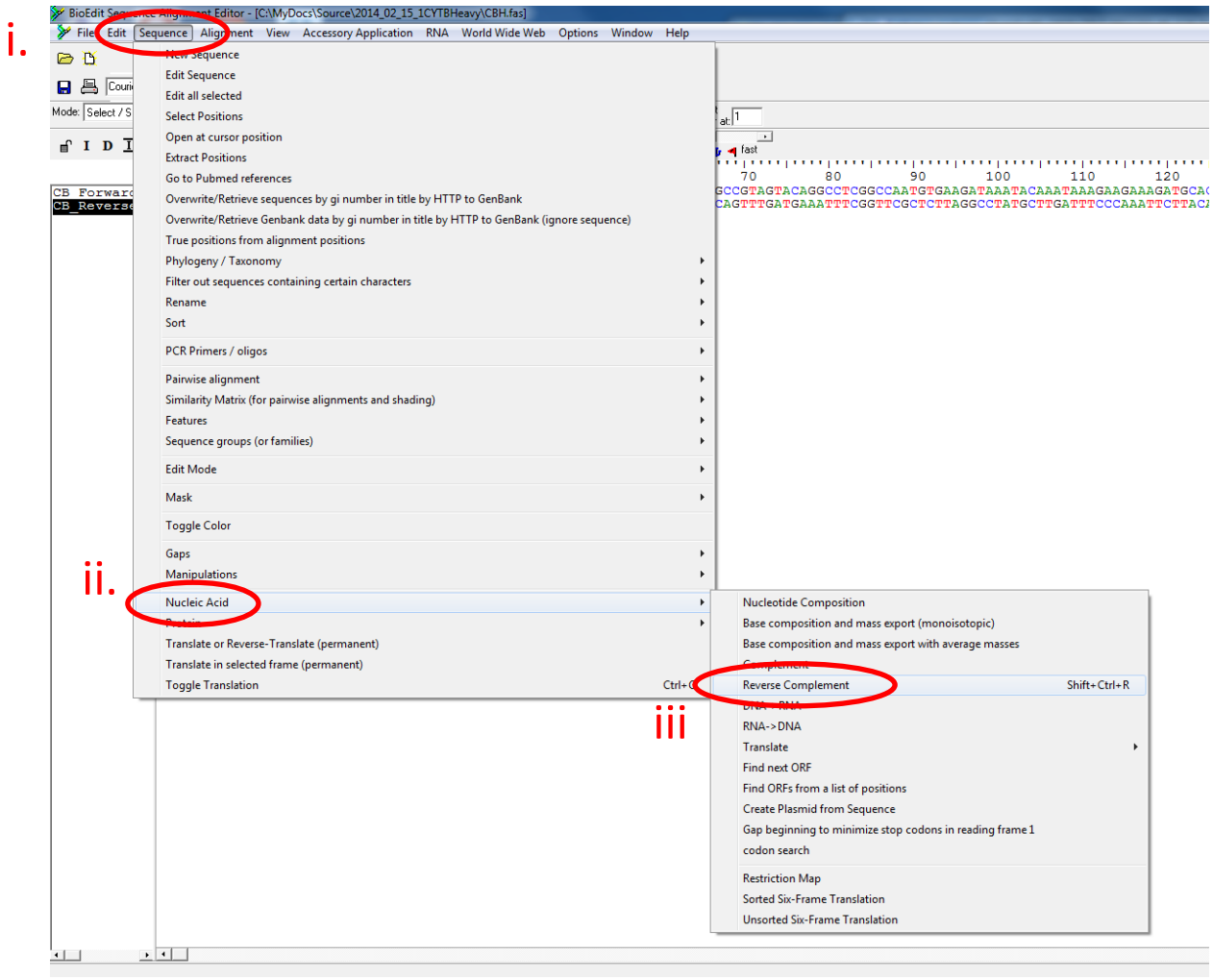
Each of the samples should have been sequenced in both the forward and reverse directions, meaning these complementary/overlapping sequences can be combined into a consensus. This serves as an important way of checking the accuracy of the sequence, and can help remove any ambiguous bases and generate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

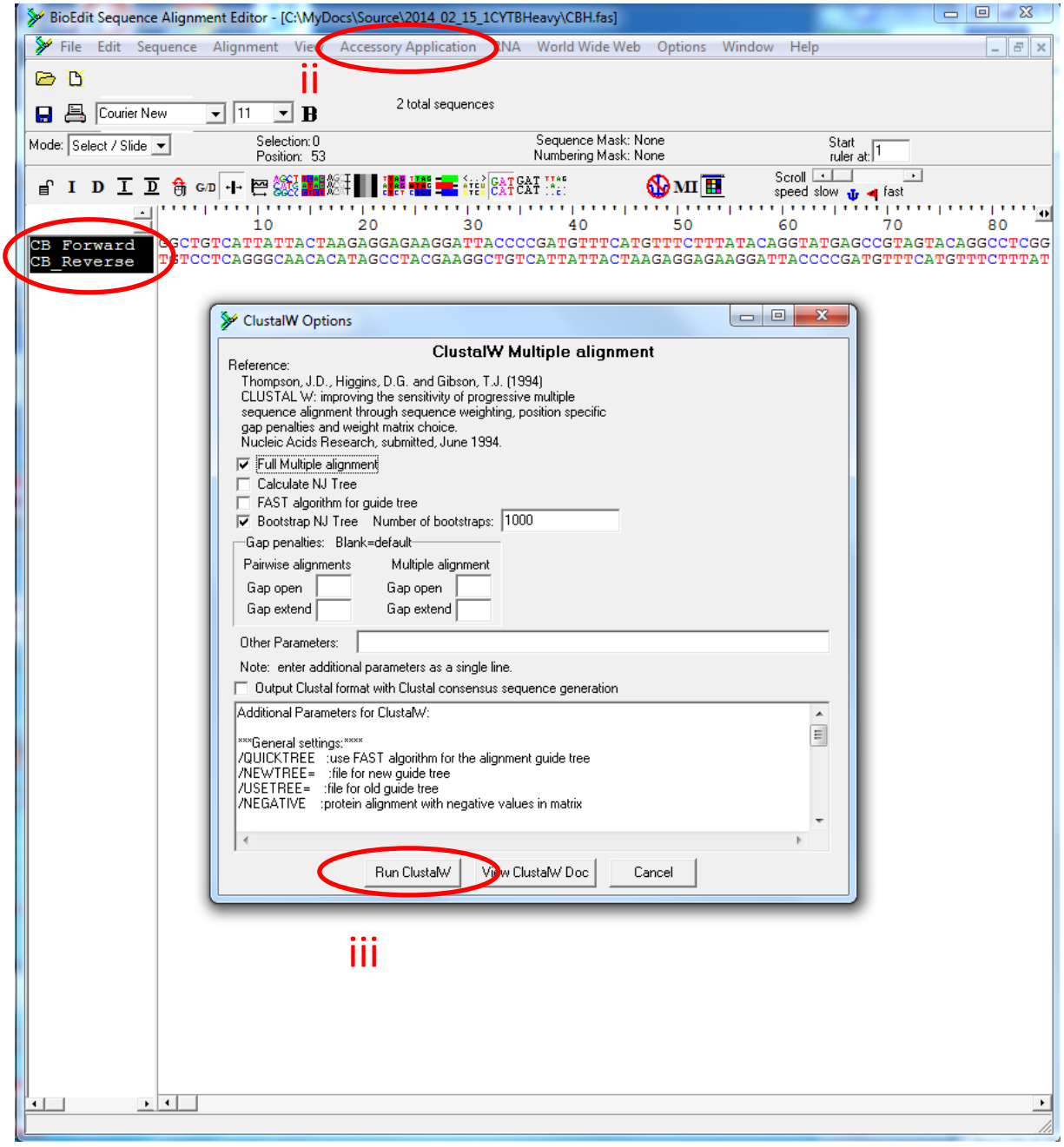
2. It is then necessary to import the reverse fasta file into the software. This is done in the file menu in the uppermost toolbar and selecting import, then sequence alignment file and locating your complementary reverse sequence fasta file.



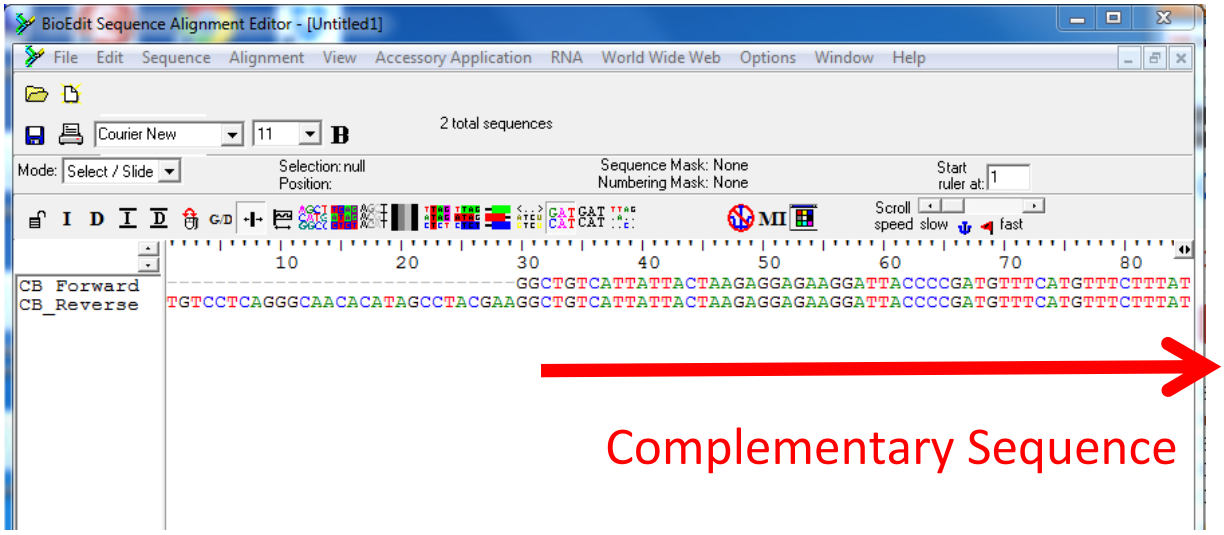
3. Select the reverse sequence within the nucleotide sequence window, just by clicking on its name on the far left. Then, in the sequence menu in the uppermost toolbar, select Nuleic Acid, followed by Reverse Complement.



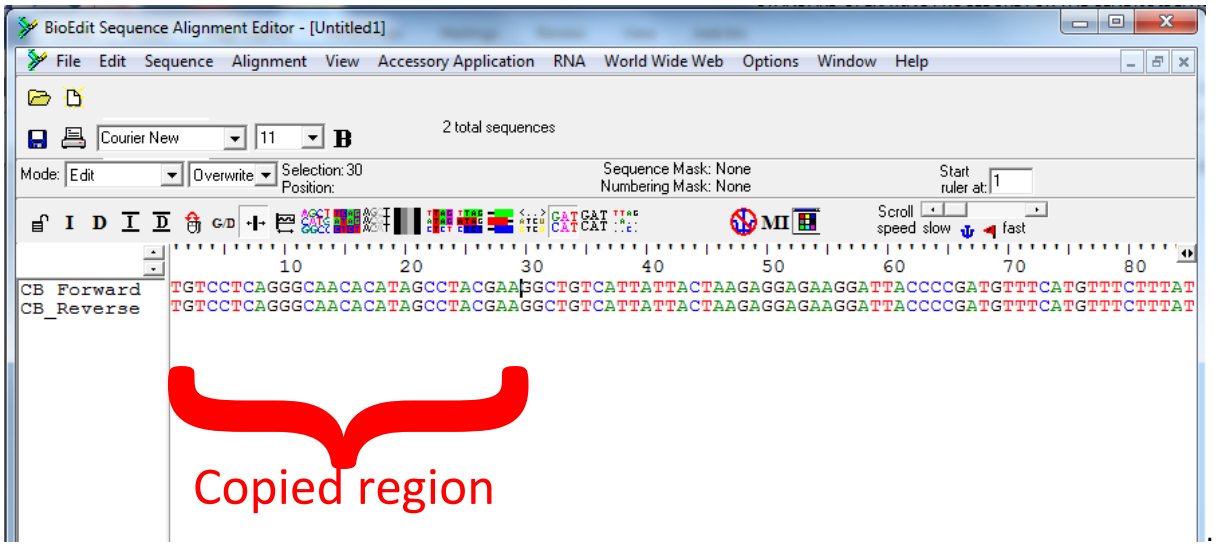
4. Select both the forward and reverse sequence within the nucleotide sequence window, using shift and select (i). Then, in the sequence menu in the uppermost toolbar, select Accessory Application (ii), followed by ClustalW Multiple Alignment. Leave the settings as defaults and click on the Run ClustalW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

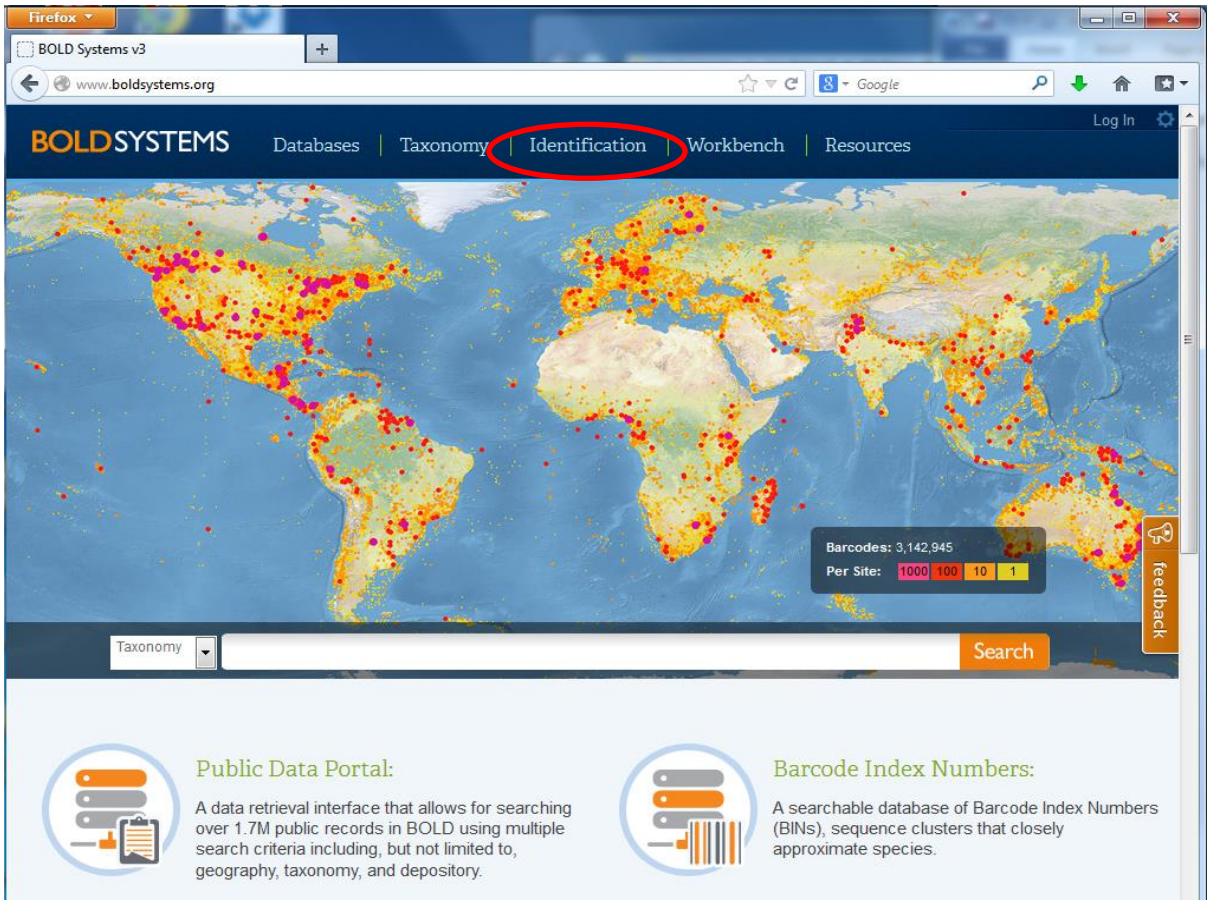
>GadusMorhuaSCFAC839-06

CCTTTATCTCGTATTTGGTGCCTGAGCCGGCATAGTCGGAACAGCCCTAAGCCTGCTCA
TTCGAGCAGAGCTAAGTCAACCTGGTGCACTTCTTGGTGATGATCAAATTTATAATGTGA
TCGTTACAGCGCACGCTTTCGTAATAATTTTCTTTATAGTAATACCACTAATAATTGGAGG
CTTTGGGAACTGACTCATTCCTCTAATGATCGGTGCACCAGATATAGCTTTCCCTCGAAT
AAATAACATAAGCTTCTGACTTCTTCCTCCATCTTTCCTGCTCCTTTTAGCATCCTCTGGT
GTAGAAGCTGGGGCTGGAACAGGCTGAACTGTCTATCCACCTTTAGCCGGAAACCTCG
CTCATGCTGGGGCATCTGTTGATCTCACTATTTTTTCTCTTCATCTAGCAGGGATTTCAT
CAATTCTTGGGGCAATTAATTTTATTACCACAATTATTAATATGAAACCTCCGGCAATTTC
ACAGTACCAAACACCCCTATTTGTTTGAGCAGTACTAATTACAGCTGTGCTTCTACTATT
ATCTCTCCCCGTCTTAGCAGCTGGTATCACAATACTTCTAACTGACCGTAATCTTAATAC
TTCTTTCTTTGACCCTGCTGGAGGAGGTGATCCCATTTTATACCAACA

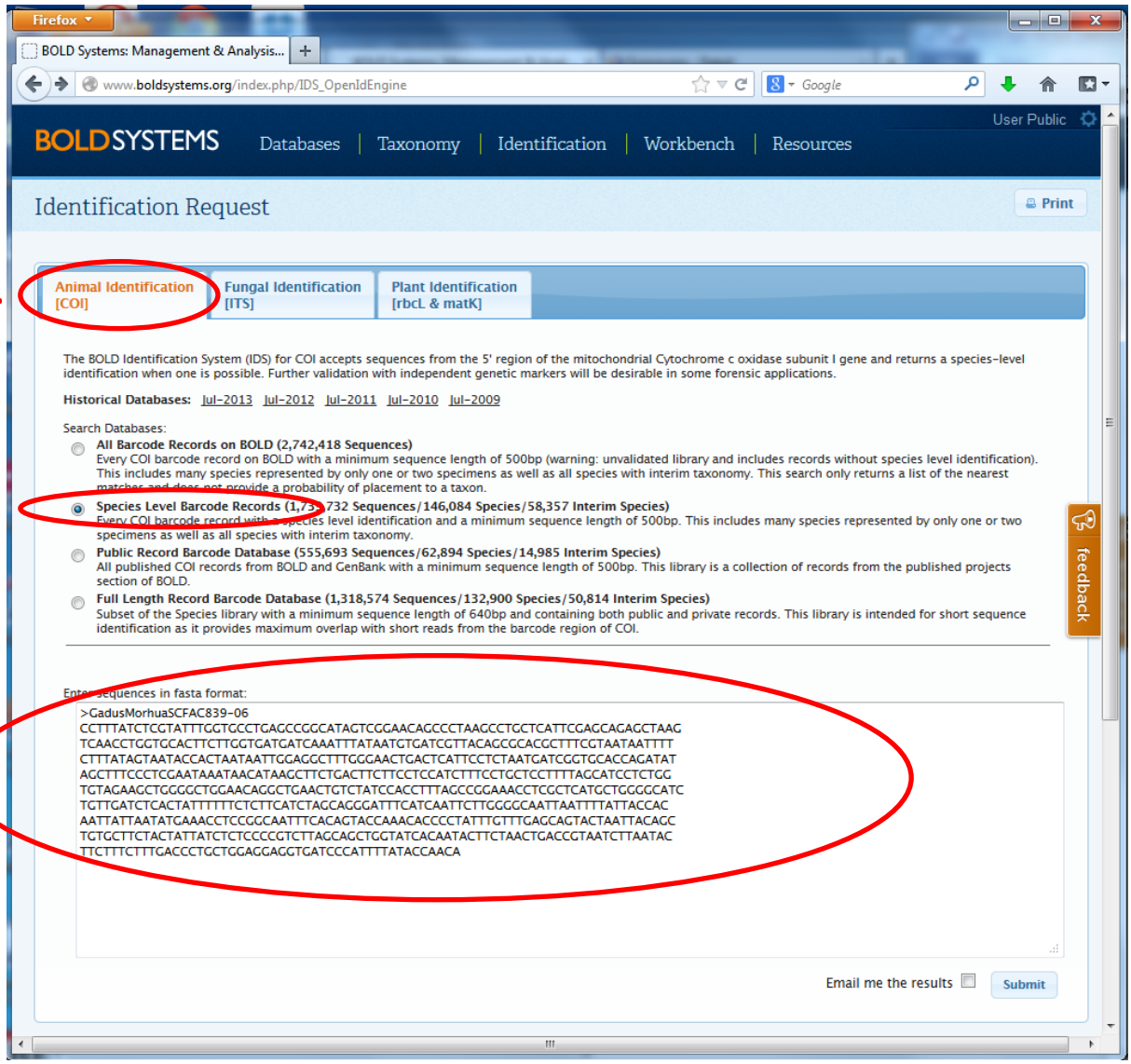
7.8 Identifying the species on the Barcode of Life Database

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.

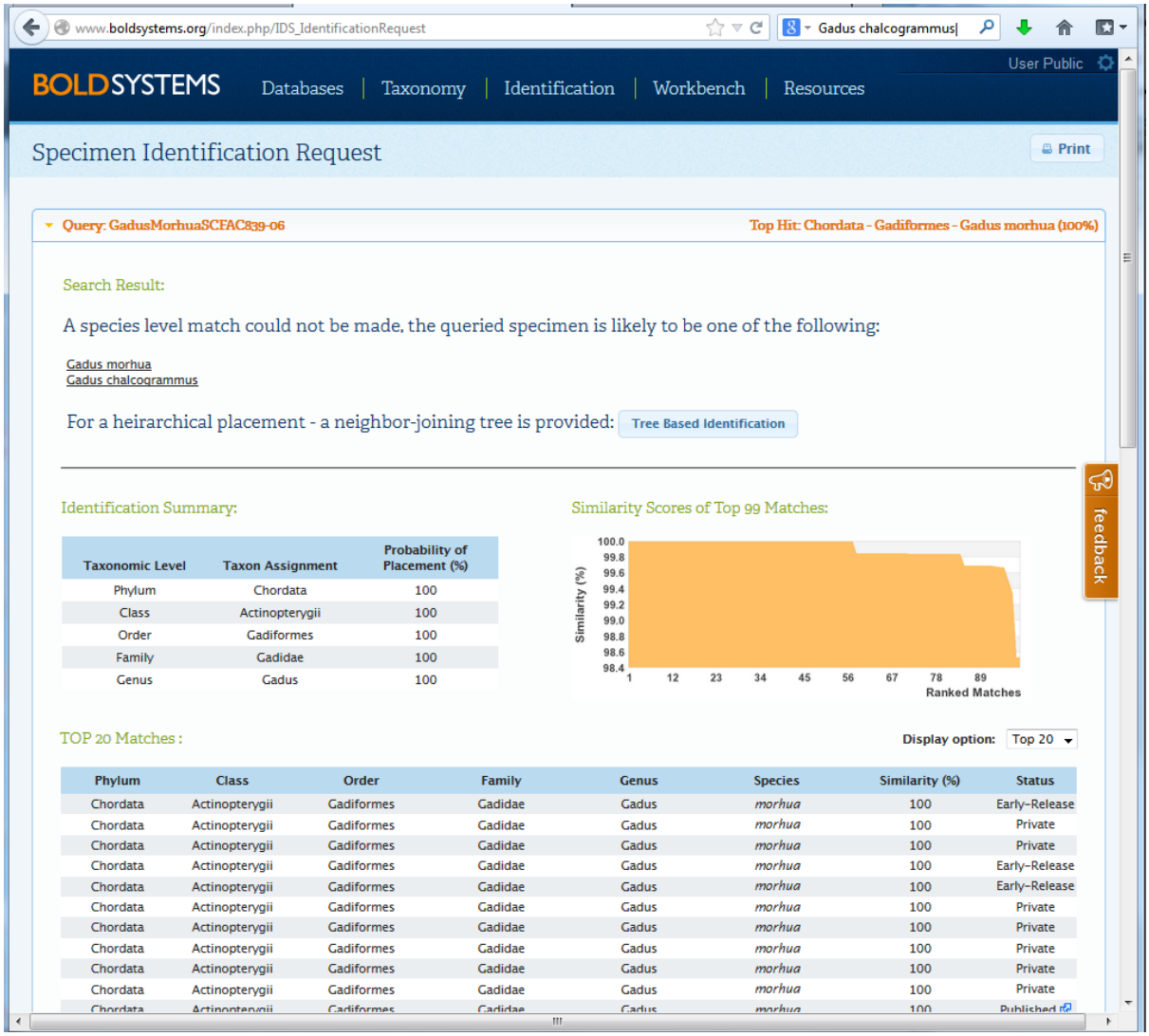
1. Start by navigating to the **BOLD Systems** webpage (http://www.boldsystems.org/) and select the “Identification” tab at the top of the webpage.



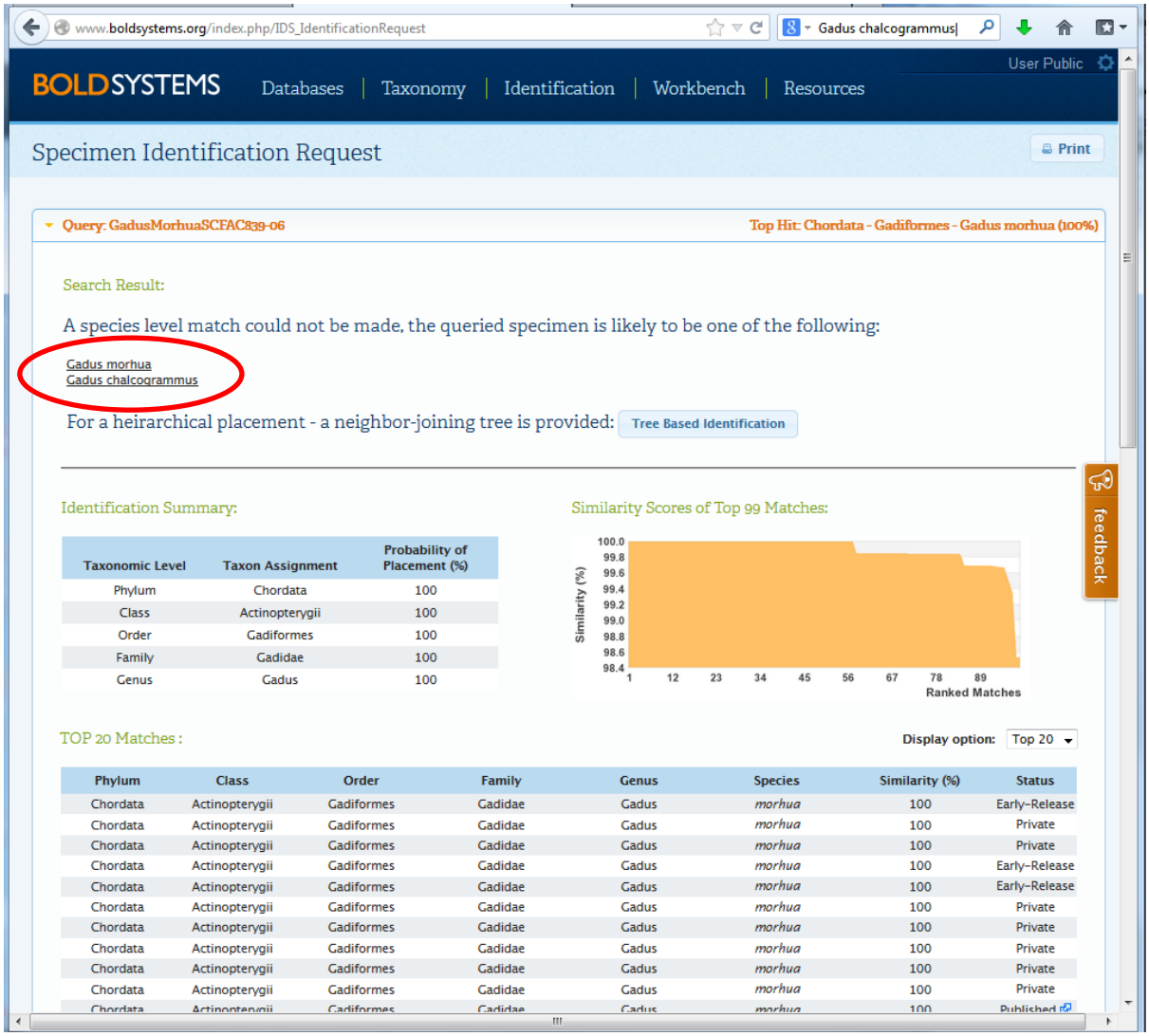
2. This page acts as a portal allowing the consensus sequence generated in the laboratory to be referenced against the entire BOLD database of reference data, i.e. from known species. **Various search options** are possible that relate to different collections of reference data, but the **default settings** 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.



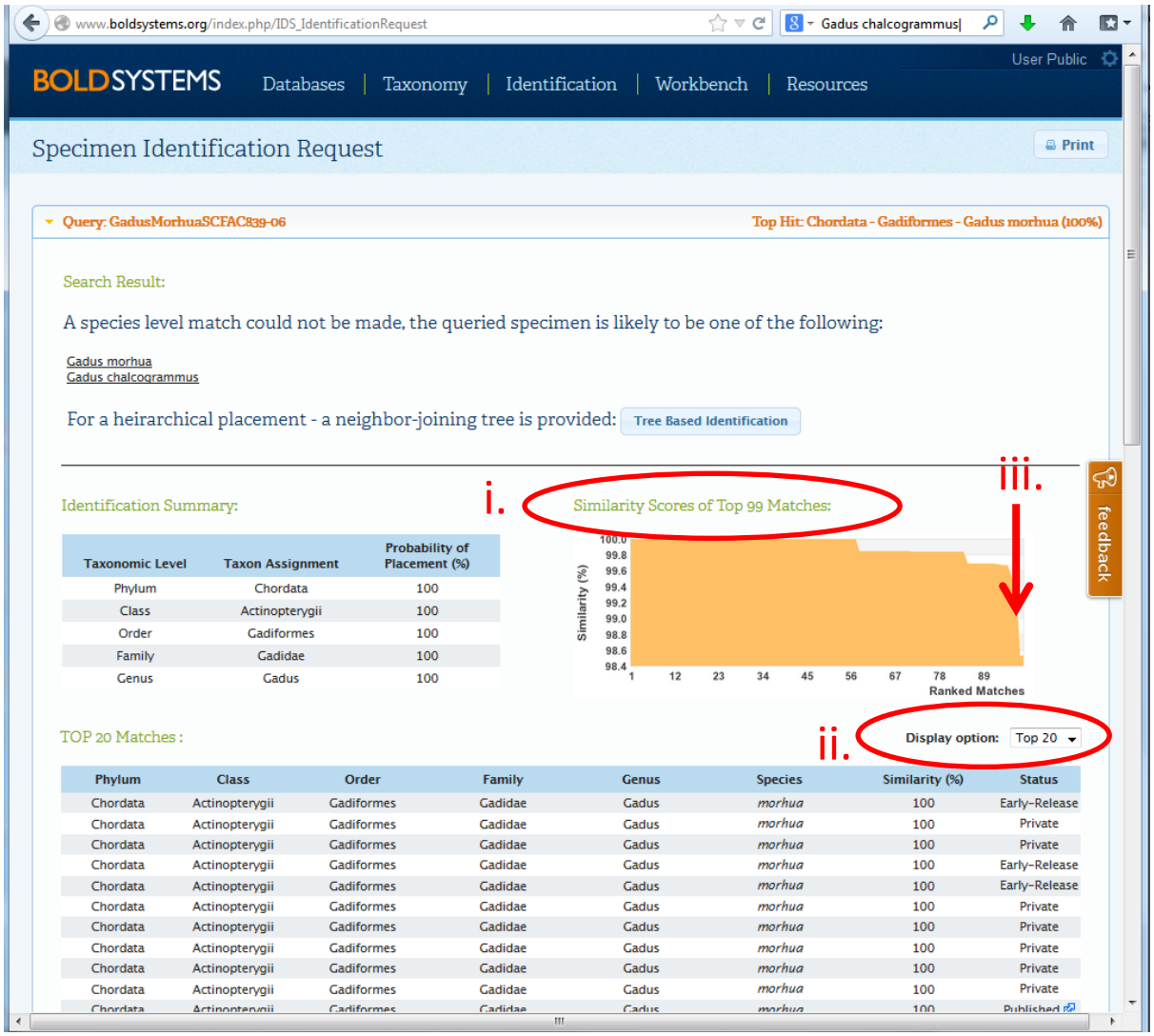
3. In a few seconds the browser will update and give you the results of the search, revealing the records contained in the database that **yields the closest match** in terms of sequence similarity. First, it is important to save a screen grab of the results as proof of the result, something similar 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).



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.



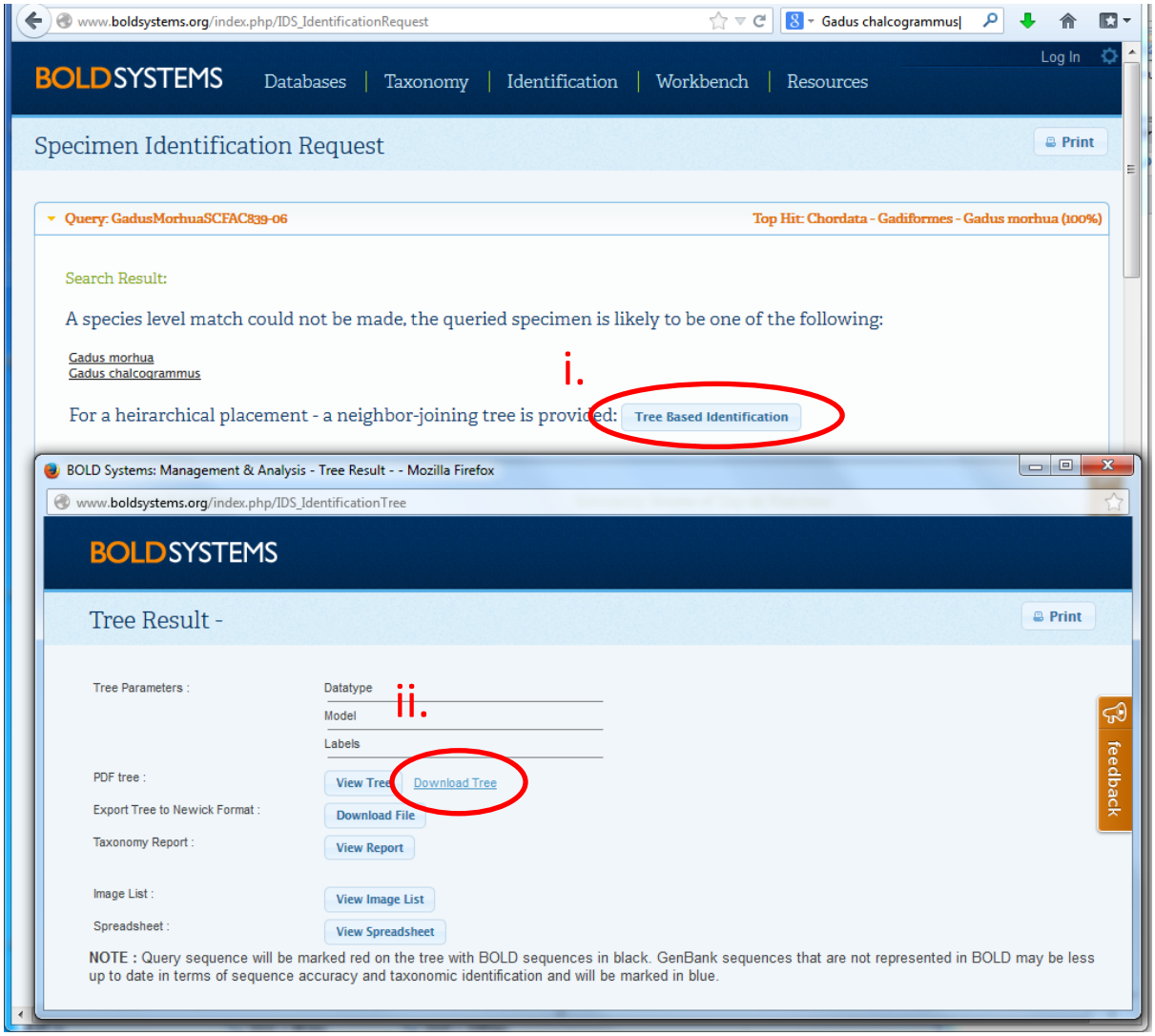
5. Next examine the graph entitled “Similarity Scores of Top 99 Matches” that shows the percent similarity for each of 99 top matching records in the database against your consensus sequences (i.). Also alter the display options in the drop down menu (ii.), to make BOLD show the full records for these corresponding top 99 matches.
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*.



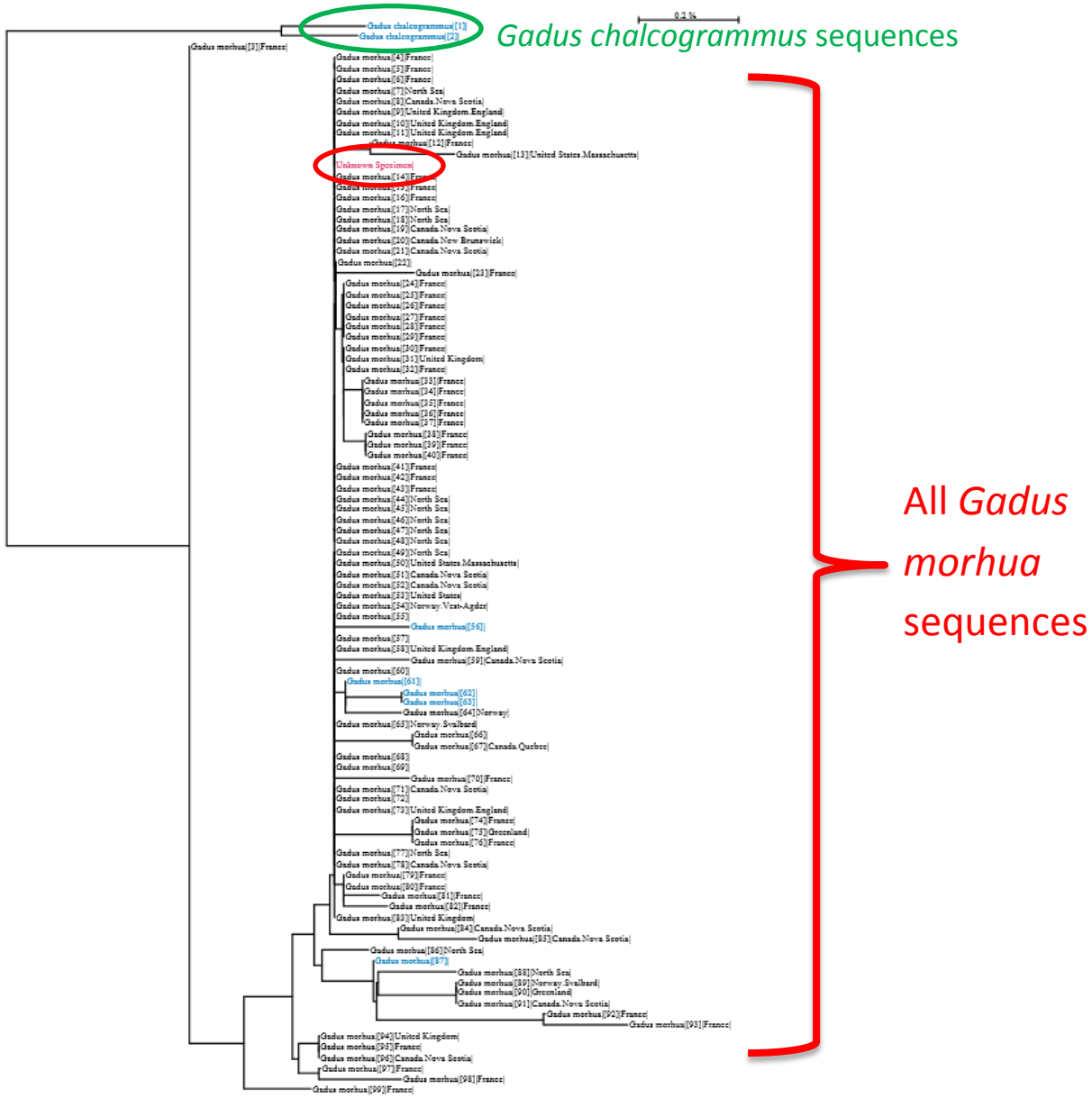
Taxonomic Level	Taxon Assignment	Probability of Placement (%)
Phylum	Chordata	100
Class	Actinopterygii	100
Order	Gadiformes	100
Family	Gadidae	100
Genus	Gadus	100

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

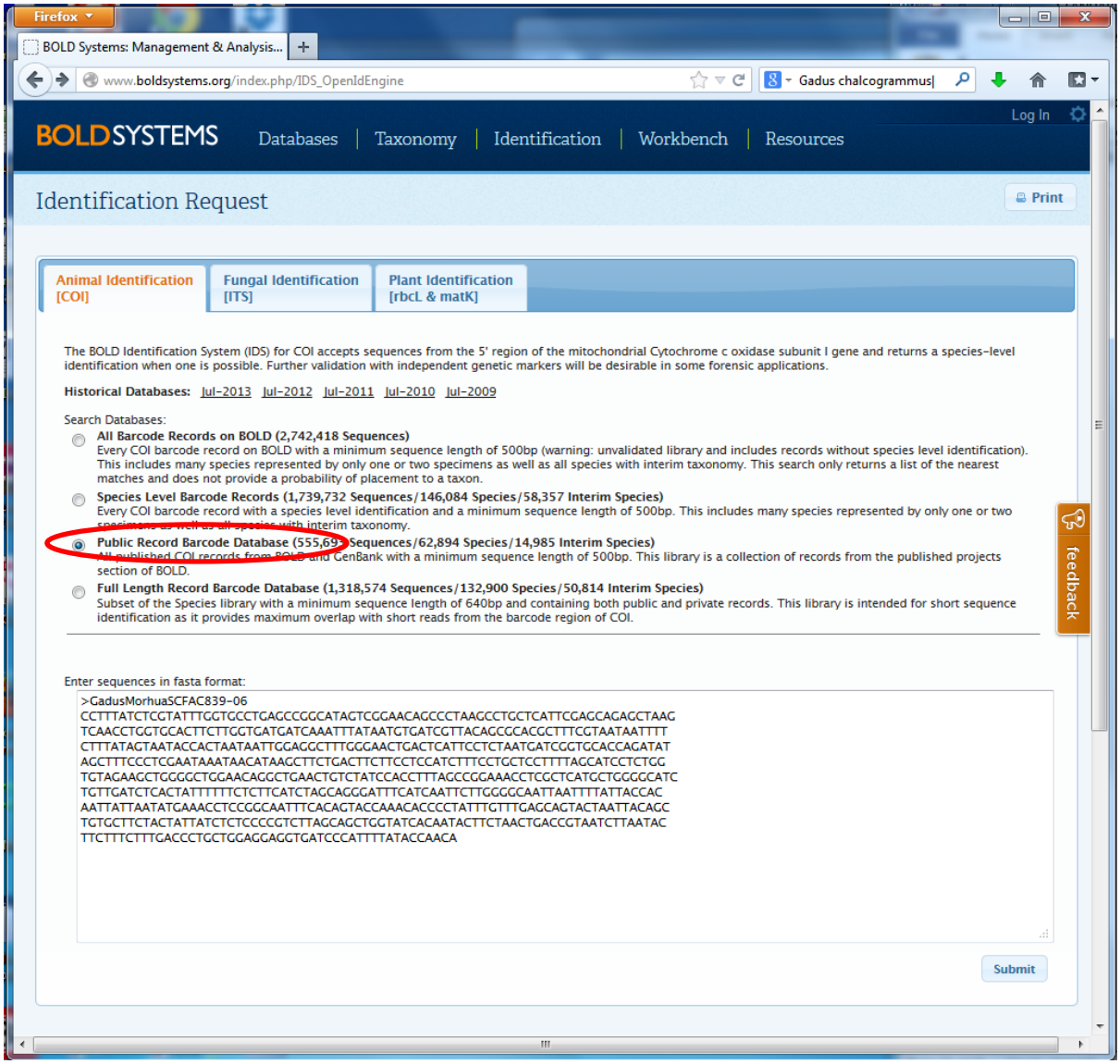
6. Besides referencing your sequence against the BOLD reference database, it is **also important to produce a simple tree** to graphically display the results of the homology search (although this is not a highly robust phylogenetic reconstruction). First click on the "Tree Based Identification" tab (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.



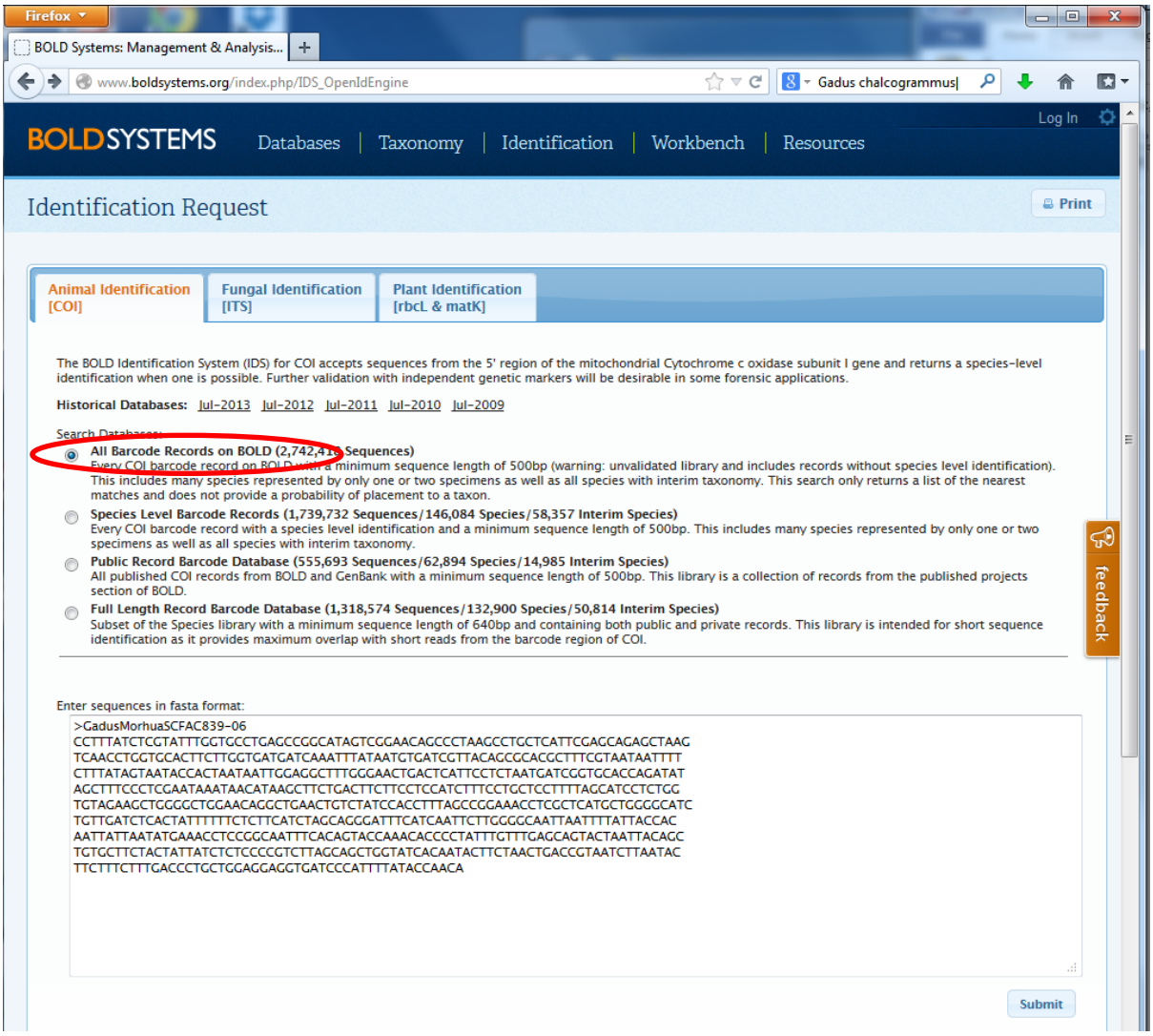
7. In the tree diagram the uploaded sequence is highlighted in red. In order to make a clear identification, this “unknown specimen” should **only cluster with sequences originating from a single species** (i.e. from a *monophyletic* group). The tree generated from our example sequence is below; our 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).



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.



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.

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

The two references cited in this SOP are:

Ivanova, NV, Zemlak, TS, Hanner, R, & Hebert, PDN (2007). Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes*, **7**: 544–548.

Ward, RD, Zemlak TS, Innes BH, Last, PR, & Hebert, PDN (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B*, **360**: 1847–57.