**PHYLOGENETIC TREE BUILDING IN MEGA X FOR FORENSIC APPLICATIONS IN IDENTIFYING UNKNOWN SPECIMENS**

*Nicholas Lorusso1, Maria Denise Gemmellaro2,*

*1) University of North Texas at Dallas, 7400 University Heights Blvd, Dallas, TX, 75241*

*2) Kean University, 1000 Morris Ave, Union, NJ 07083*

This exercise is designed to guide a learner through the construction of phylogenetic trees as a means of addressing research questions in forensic science such as the identification of previously unidentified species contributing to decomposition. Students are introduced, over the course of roughly one lab period, to: metabarcoding and bioinformatics, databases for retrieving publicly available data, the construction of phylogenetic trees using MEGA, and the interpretation and implications of the trees they construct. Learners are provided with a novel sequence for the Cytochrome C Oxidase gene extracted from a real unknown specimen and are tasked with identifying the most likely identity of the specimen. To accomplish this, the learner will construct their own data file of reference sequences by retrieving sequences from published databases before combining this with the provided record prior to phylogenetic tree construction. By creating this predicted phylogeny, the student will identify the most likely identity of the unknown specimen and will be able to clearly explain their prediction based on the prepared dataset and phylogeny. This will not only expose students to how phylogenetics can be applied to forensic applications but will also demonstrate the use of such methods to real world biological questions.

**Learning objectives**

After successful completion of this exercise, students will be able to:

* Apply the scientific method to questions relating phylogenetics to forensic science
* Identify appropriate computational approaches to address biological questions
* Evaluate evolutionary relationships using graphical methods
* Critically assess results of barcoding datasets
* Find and interpret data from major online databases such as Barcode Of Life.
* Use the software package MEGA X to construct and evaluate phylogenetic trees
* Apply bioinformatics methods to solving biological problems in forensic science

# **Exercise Description**

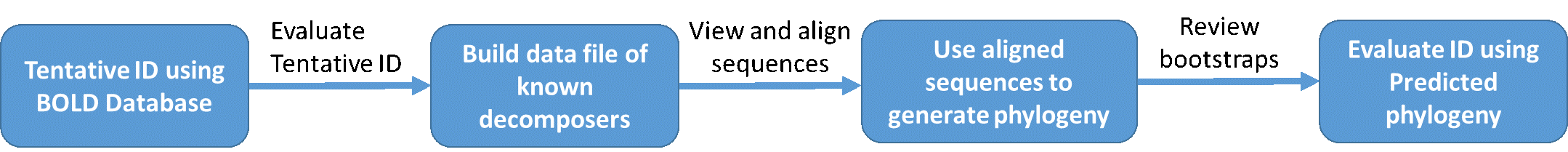
# **Using MEGA and records retrieved from the Barcode of Life database to construct predictive phylogenies for an unknown specimen**

## **Background**

The use of arthropods recovered during forensic investigations can help answer critical questions relating to the timeline of decomposition or time of death, the environment a decomposing body has been exposed to, and cause of death (Joseph et al., 2011). When databases for known arthropod species for a region are incomplete or nonexistent, the application of entomology to forensic investigations may be particularly difficult, since reference data needed to assess the insect evidence collected at a scene are missing. However, this can be remedied by conducting entomological surveys and building databases of forensically important arthropod species (Tomberlin et al., 2011). Building such databases, however, often requires techniques that allow for the identification of previously understudied species to make the information accessible to forensic investigators.

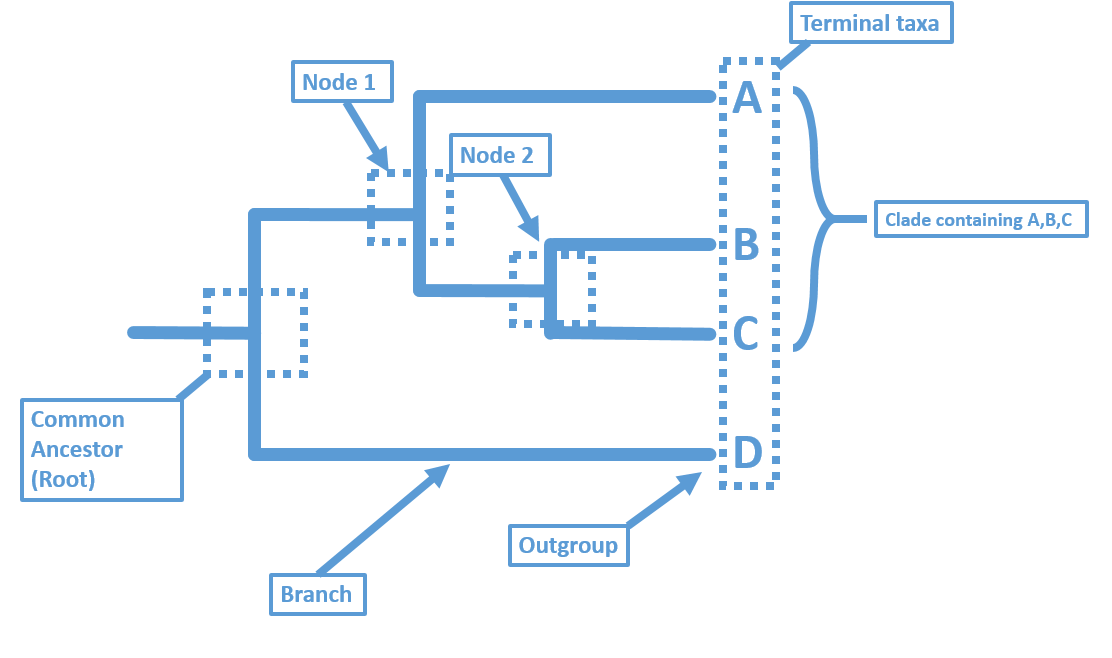
Molecular techniques, such as DNA barcoding, are useful in cases where specimens are difficult to identify morphologically, have been damaged, or were reported, but only environmental traces remain (DeSalle and Goldstein, 2019). DNA barcoding and metabarcoding are molecular techniques that use DNA collected from samples of interest, like unknown arthropod specimens, and uses specific genes as a type of identifier (or barcode) to identify species (Pečnikar and Buzan, 2014). This approach relies on specific genes identified by the scientific community that, once sequenced, are compared to known sequences from different biological species stored in public databases, and can so provide information on the identity of the unknown specimen. Genes in bacteria (16s rRNA gene), fungi (ITS gene), metazoa (COI gene), and plants (rbcL) have been identified as being unique enough across taxa to allow for barcoding (though barcoding genes can vary by application). Once a database for the barcoding genes has been created, an unknown specimen’s DNA can be assessed by comparing it to this database with the goal to create a tentative identification.

This exercise is intended for anyone interested in how barcoding data can be used to answer questions in applied fields such as forensic entomology. You will be provided with cytochrome c oxidase (COI) sequence data from a specimen that was collected as part of a real-world forensic study in the North Eastern United States. Depending on circumstances, there may be a number of ways of identifying an unknown specimen, but this sample comes from a fly larva that does not match entries in morphological forensic databases. This means we may be able to expand our knowledge of species that contribute to decomposition! **You will use the provided gene sequence for the unknown to develop a tentative ID, retrieve comparison sequences from the Barcode of Life Database (BOLD), compare the unknown sequence to other dipteran (fly) species using phylogenetic analysis, then use that to evaluate your tentative ID (Figure 1).** We will start by querying the sample in BOLD to create a prediction, and then begin creating a set of sequences for comparison.



*Figure 1: Conceptual flow diagram for the stages of predicting and evaluating our tentative ID for the unknown sample*

Phylogenetic analysis using barcode data allows researchers to examine and verify predictions from samples of DNA containing barcode sequences. While querying databases can generate strong predictions for species – researchers often want to verify the strength of the position and, in the case of many taxonomic analyses, to evaluate the position of the predicted taxa ID compared to other taxa. **Phylogenies are graphical representations of taxonomic relationships** that can be constructed using various types of data as input (e.g. morphological, genetic, behavioral, etc.) and connect taxa using **nodes and branches** (Figure 2). Interpreting phylogenetic trees can take practice – but in general the taxa that are connected more closely (e.g. have nodes less far to the right in Figure 2) are predicted to be more closely related. These phylogenetic trees are only predictions, and scientists/software packages will often test and compare hundreds if not thousands of possible trees to create the best prediction possible. To evaluate how strong a prediction is for given relationships, **bootstrap values** are often shown at nodes to indicate how strong the support is for a given divergence.

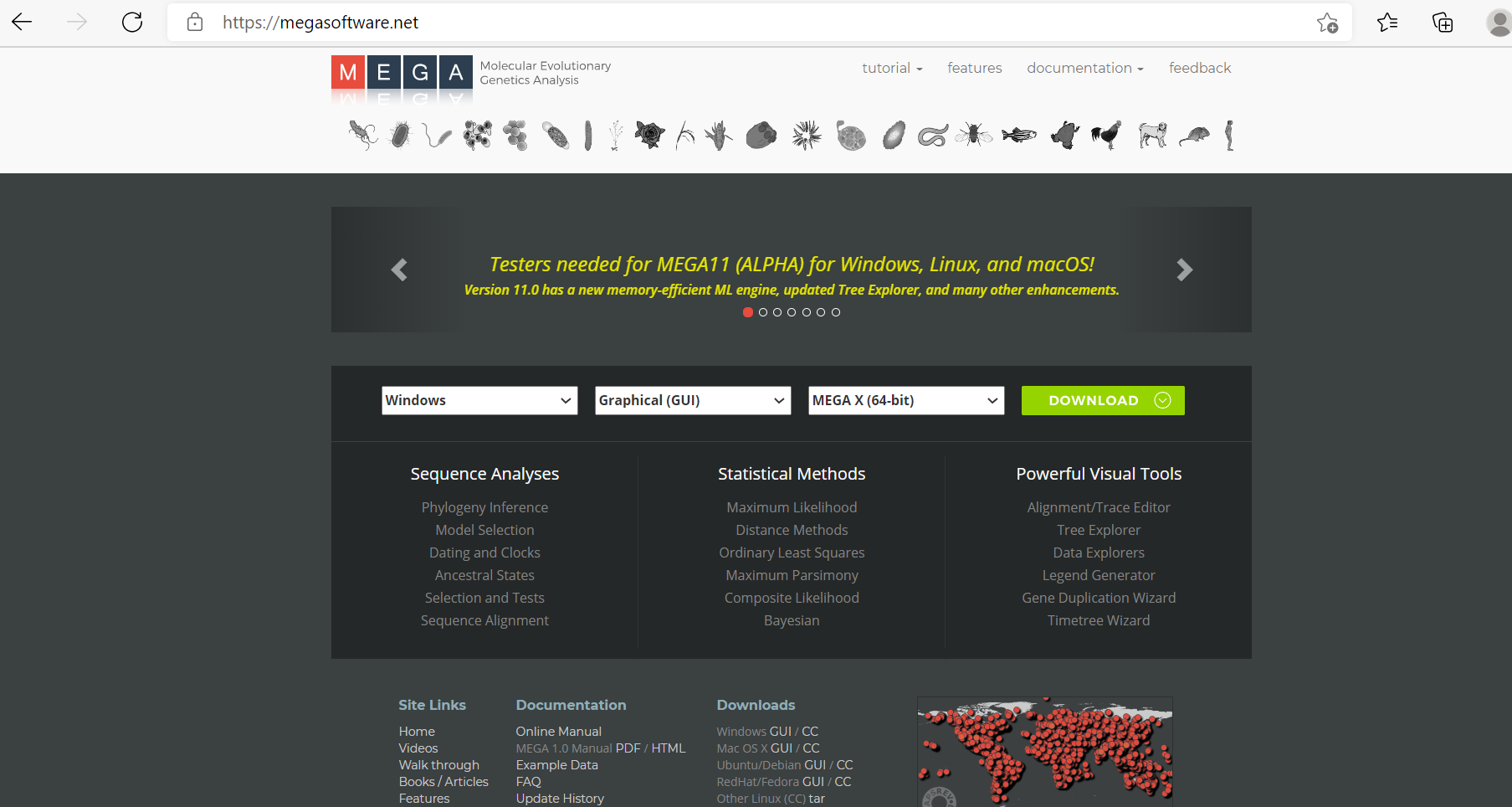


*Figure 2. Labeled diagram of a phylogenetic tree with parts commonly used for comparison labeled.*

Once we have created a predicted phylogeny we will evaluate our initial prediction from our BOLD query and see if we still accept the result.

We will use the mitochondrial Cytochrome C Oxidase (COI) gene as our barcode both because it is what we have from our sample and also because it is a commonly used barcoding gene for the animals due to its commonality and slow evolutionary rate. Genes that change gradually over time like COI are good barcode targets because they are likely to give good resolution (separation) between different species, without members of a species differing so much that they don’t group together in our phylogenetic analyses. There are multiple software options available to construct phylogenies but we will use the freely available MEGA X software (supplemental instructions are included for creating comparison trees using R as an extension exercise). Make sure you have access to: 1) MEGA X (download instructions below), 2) Notepad or another text editor, 3) the BOLD database.

### **Install MEGA X and download data files**

* 1. Go to <http://www.megasoftware.net/>. Choose your operating system and ensure the graphical version of the program is selected in the second dropdown. Install MEGA X from the downloaded file and confirm it launches.  
     
  2. Download and save the data file “unknown.fas” to work with in MEGA (provided by your instructor). Open the file using a program like notepad to see what the entry looks like – FASTX (FASTA, FASTQ, etc.) files are commonly used for saving data from sequencing projects for bioinformatics that includes:

### Record information: This follows a “>” symbol that identifies a name for the sequence. This will often have a species name in taxonomic data.

* Sequence: This will show as a collection of C,T,A,G indicating nucleotide identification from sequencing. Cases where nucleotides were missing (-) or unidentified (N) are also recorded.

### **Accessing the BOLD database**

Now that we have the required software we can make our prediction of the specimen ID and retrieve comparison sequences.

* 1. Access <https://www.boldsystems.org/> and select “Identification” from the top menu



* 1. Select “All Barcode Records on BOLD” and copy and paste the contents of the unknown FASTA file into the search box and hit submit. *Note: BOLD is a public database so it may take some time for your query to be completed.*

**Question 1.** Reviewing the matching hit on BOLD – what is the likely identity of the specimen and how likely is the ID?

* 1. The ID you’ve made is actually really interesting because not only is it not just unknown as a decomposer locally – there might not be records *anywhere*. Try performing a google/google scholar search for something like “*GENUS NAME* decomposer” or “*GENUS NAME* necrophagous”

**Question 2.** Can you find a record of the genus of your ID being necrophagous? Given that this sample was taken from a decomposing body, does this make you less certain of the tentative ID?

* 1. Click the “Tree” button in the BOLD result and wait for the results to populate. Choose either a few of the published top matches by clicking the blue icon next to “Published” in the status column OR scroll down to see an interactive map. Be sure to examine the PDF phylogeny produced in BOLD.

*Note: You might notice that, compared to how strong the match is in the hits from your search, the tree is less clear. Don’t worry! That’s why we’re going through the steps of building these phylogenies. Services like BOLD offer great options but some things (like predicted phylogenies) are automated and can need further investigations like the ones we will conduct.*

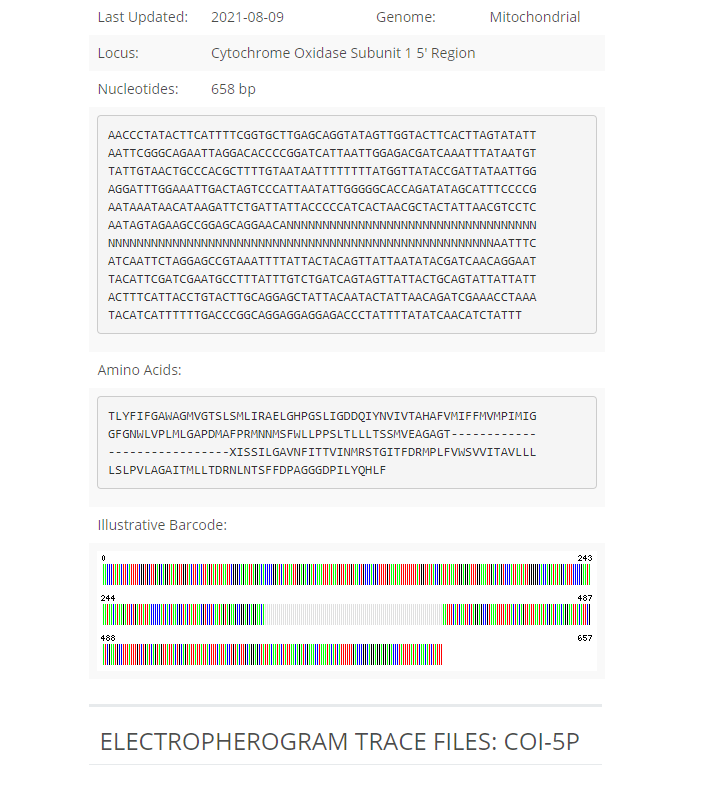
**Question 3.** Where, geographically are most of the hits from and does that match our sample collection information?

**Question 4.** Do most of the top 20 matches appear to be identified to the species level? How many are at the species level (*Genus species*)?

* 1. Since this does not seem like a commonly reported decomposing arthropod – let’s create a tree just to evaluate our predicted ID. Select “Databases” from the top menu in BOLD then click “Public Data Portal”
  2. We need a set of sequences from dipteran species (preferably known necrophages and non-necrophagous) to compare to AND a collection of species from the genus we found in our search. Enter that genus name into the search box and hit submit.

**Question 5**. How many records are there on BOLD for that genus?

**2.7.** Review one of the entries by clicking the record number and review the available information (especially the FASTA sequence and the illustrated barcode).



**2.8.** Navigate back to the list of all records for your genus anddownload sequences for one example for each species from the list. Do this by marking the checkboxes next to the entries and then clicking the “FASTA” button at the top right of that list.

**2.9.** Open the FASTA file you downloaded using notepad or another text editor and copy the sequences into the “unknown.fas” file from your instructor and use “File” -> “Save as” to save the file as “GENUSMEGA.fas”  
**2.10.** Select some (or all) of the genera below to search in BOLD and download the entries as in steps 2.6-2.9. *Note: choose at least two other genera but note that the more genera you choose the more computational power/time will be needed to make the trees. There likely be many entries for each genus and we recommend choosing ~10 different species (or as many different species as there are in the database)* ***with only one example of each species selected****.*

**Potential other genera of necrophagous diptera:**

* + - * *Calliphora*
      * *Cyanus*
      * *Dexosarcophaga*
      * *Helicobia*
      * *Lucilia*
      * *Metopia*
      * *Senotainia*
      * *Taxigramma*

**Potential non-necrophagous diptera:**

* *Aedes*
* *Drosophila*
* *Stenotabanus*

**2.11.** Once you’ve downloaded the genus identified in your initial BOLD search **and** at least two other genera and copied the FASTA entries into “unknown.fas” save the file as “FLYMEGA.fas”

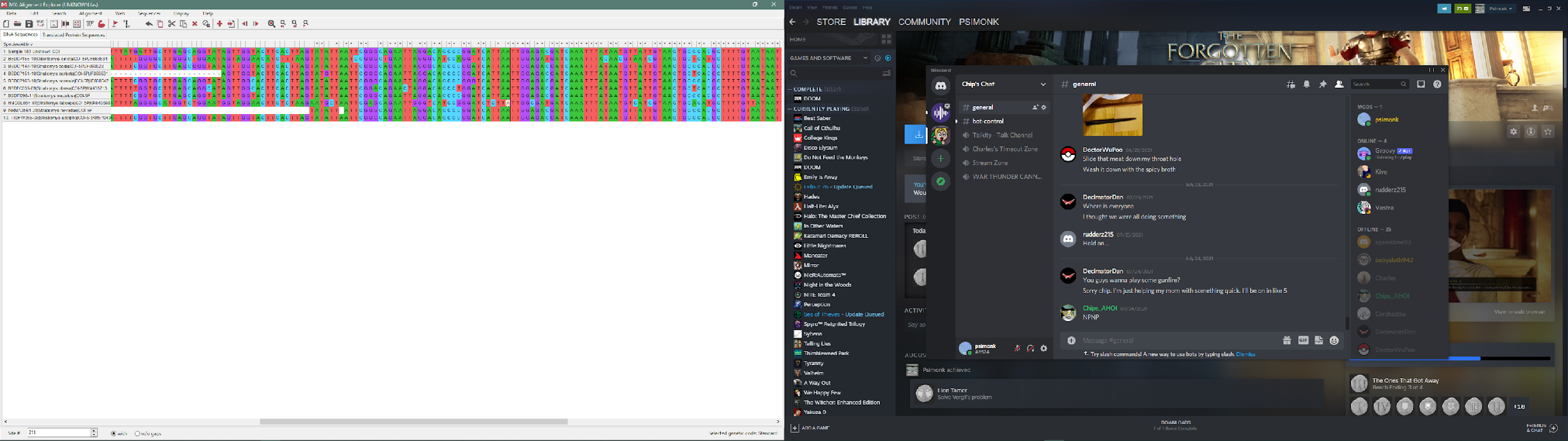
*Note: at this point you should have 3 FASTA files total, the unknown file we’ve been working in, “GENUSMEGA.fas” and “FLYMEGA.fas” – view these in notepad to make sure they saved properly.*

### **Aligning sequences retrieved from available databases**

Prior to using the sequences we have saved in FLYMEGA.fas, we need to **align** them so that MEGA is comparing each nucleotide sequence locus by locus. This step is important so that MEGA and other software can interpret differences between the nucleotide sequences we retrieved meaningfully. Without this step, even if we had sequences for the same gene for all of our different species, the software might not be properly comparing them when we move on to creating our phylogenies.

* 1. Open MEGA and use the following procedure to create an alignment file to use for creating our phylogenies.
  2. Open the “GENUSMEGA.fas” by clicking the “ALIGN” button in the MEGA menu bar, selecting “edit/build alignment”, and then “Retrieve a sequence from a file” before clicking OK to open that file for alignment.
  3. The new window that opens will contain sequences of all of the taxa you retrieved for alignment. To use MEGA to perform an automated alignment, navigate to the “Alignment” menu in the alignment editor, select “Align by MUSCLE”, and click “OK” (use default parameters).

One reason MEGA is a useful program for learning about these techniques is its user interface. Notice that each of the four nucleotides is color-coded and that MEGA has inserted gaps to make the sequences line up (an example of MEGA alignment output can be seen below). One notable feature is that some of the loci (columns) will be highly variable by species (rows) while others will be more conserved. Try scrolling left and right to move through the sequence.



**Question 6.** Do there appear to be many gaps in your alignment (not the photo above)?

**Question 7.** Do most Loci seem conserved or variable across species?

* 1. As you move left and right you will find there may be more gaps/variable segments at the end of the aligned sequence that might interfere with our attempt to make a phylogeny. Find a good starting point (move left past the worst of the gaps) and end point (moving from the extreme right) and select the columns you want to delete. Remove the columns using the edit menu. Finally, if any species (rows) are almost entirely gaps – delete these rows by selecting them and using the edit menu to delete them.
  2. Click the “Data” drop down menu and click “Phylogenetic Analysis” and click Yes. Save the alignment by navigating to “Data” then “Export alignment” and select “MEGA” as the format. Name your file “genus\_alignment” and close the alignment window.

### **Creating predictive phylogenies using maximum parsimony.**

We will now take the alignment we created in the previous steps and use it to compare taxa by their sequence similarity. To do this – we will use **Maximum Parsimony**. This is only one approach to phylogeny construction that makes very specific assumptions, but we can use it to quickly generate hypotheses.

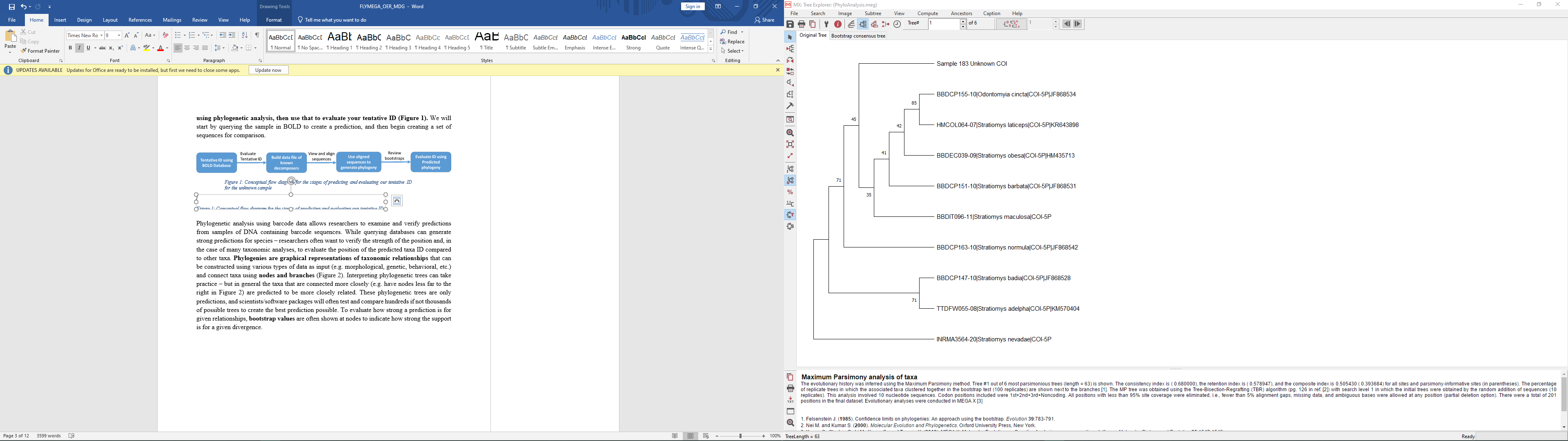
**Maximum Parsimony** methods evaluate a large (ideally infinite) number of trees that could all be possible, and then considers which of those trees requires the least number of genetic changes across the members. The tree which has the fewest number of changes to explain the data is considered most likely.

#### **Building a maximum parsimony tree for our genus of interest**

* + 1. In the main MEGA window click the “PHYLOGENY” button then select “Construct Tree/Test Maximum Parsimony Trees” – click OK when asked if you want to use the current file for analysis.

**Analysis**=Phylogeny reconstruction  
**Statistical Method**=Maximum parsimony  
**Test of phylogeny**=Bootstrap method  
**No. of Bootstrap Replications**=100  
**MP Search Method**=Tree-Bisection-Reconnection (TBR)  
**No. of initial trees (Random addition)**=10  
**MP Search Level**=1  
**Max No. of Trees to Retain**=100

Click **OK** and wait for the analysis to complete (This can take up to 10 minutes depending on your computer specifications). Each clade (grouping) on the tree is an evolutionary hypothesis that we can evaluate. At the end of the process you should have a predicted phylogeny like the one found in Figure 3 (below). The arrangement of the tips and nodes indicates what relationships MEGA predicts based on the data provided while the numbers at the nodes (show the bootstrapping support for how likely a relationship is (with a range of 0-100)



*Figure 3. Maximum parsimony tree. In this tree the number on the branches represents the likelihood of the relationship shown occurring out of the total number of trees considered (closer to 100 is more likely).* ***Your tree will look different depending on what sequences you use, how you trim, etc. Remember that there’s not a “right” answer – we’re working on making a hypothesis!***

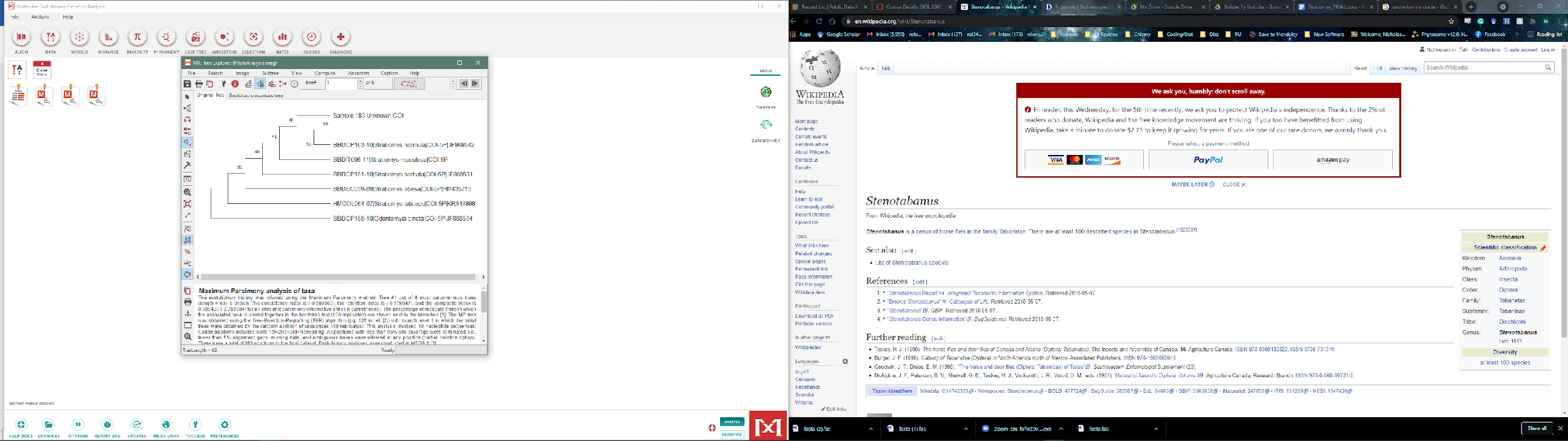
* + 1. Before going further, you should learn how to designate an **outgroup**, and **root** the tree to make sense of it. An outgroup should be an isolate, or small group of isolates, that are known to be more distantly related than all the other isolates being analyzed, but similar enough that the DNA can still be aligned. A good choice of outgroup in our case is an isolate from a distantly related insect species. The other initial sequence from unknown.fas is a different genus of fly that will work for us as an example. Select an outgroup for the tree by clicking the button on the left-hand pane of the tree viewer with the red arrow () and then clicking on the branch of “Odontomyia cincta” to select it as the outgroup. This will root the tree and will examine the relationships between the remaining taxa relative to that taxa. There should only be one tree.
    2. Be sure that in the tree view window you are in the “Original Tree” tab. Save the tree by copying it into your word document for submission – be sure to include a descriptive figure caption!
    3. Close the tree after saving to proceed to the next exercise.

#### **Building a maximum parsimony tree using multiple genera**

Now that we’ve tried to evaluate the genus of interest in terms of which species seems most likely for our unknown – let’s look at that genus in terms of its relationship to other dipteran genera. This can be helpful because closer taxonomic relationships to necrophagous groups may indicate if our identified genus is surprising as a decomposer. In this way we can evaluate if out genus of interest from our ID is a good candidate for adding to forensic databases.

* + 1. Open a new alignment session using the same steps from section 4.1, selecting to discard the current session as you should have already saved the alignment and tree files.
    2. Once the alignment has been created – be sure to trim the sequences and remove any uninformative species as we did for our first tree. Save the alignment as “fly\_alignment”
    3. Use the same options as before for creating our second maximum parsimony phylogeny.

**Analysis**=Phylogeny reconstruction  
**Statistical Method**=Maximum parsimony  
**Test of phylogeny**=Bootstrap method  
**No. of Bootstrap Replications**=100  
**MP Search Method**=Tree-Bisection-Reconnection (TBR)  
**No. of initial trees (Random addition)**=10  
**MP Search Level**=1  
**Max No. of Trees to Retain**=100

* + 1. Root the tree as before – using the “Odontomyia cincta” as the outgroup. Observe this and compare it to the tree produced using maximum parsimony earlier. Save this tree as before.
    2. Bootstrap values offer a convenient way of assessing how much trust we can put in the branching we see in our tree. MEGA offers us the ability to filter or “condense” the tree to only see relationships that most models predict to be true. To do this – select the compress button in the left hand menu () and then click the node you want to collapse. Give the node a name and it should make the tree easier to interpret.
    3. Observe changes to the tree and compare it with your uncondensed tree. Save it like the condensed tree and then save any other files before closing MEGA.

**Question 8**. Did the placement of the unknown change significantly when considering the genus of interest alone versus including more species in the alignment?

**Question 9.** Based on the tree – does the genus of interest seem more closely related to necrophagous or non-necrophagous species?

**Question 10.** Based on both trees and your experience building them – is the initial result (the top hit in the search) in our BOLD query reasonable?

**Question 11.** Based on this – and the collection of this larva from a decomposing specimen – do you think it would be reasonable to expand existing databases of necrophagous species in the North Eastern United States to include the genus/species you found? Why or why not?

**Acknowledgements**

The authors thank Maria Shumskaya and Christopher Zambell for the original tutorials. Special thanks to Gloria Raise for critical reading and testing of the exercise.

**References**

DeSalle, R., & Goldstein, P. (2019). Review and interpretation of trends in DNA barcoding. *Frontiers in Ecology and Evolution*, *7*, 302.

Joseph, I., Mathew, D. G., Sathyan, P., & Vargheese, G. (2011). The use of insects in forensic investigations: An overview on the scope of forensic entomology. *Journal of forensic dental sciences*, *3*(2), 89.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, *35*(6), 1547.

Pečnikar, Ž. F., & Buzan, E. V. (2014). 20 years since the introduction of DNA barcoding: from theory to application. *Journal of applied genetics*, *55*(1), 43-52.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tomberlin, J. K., Benbow, M. E., Tarone, A. M., & Mohr, R. M. (2011). Basic research in evolution and ecology enhances forensics. *Trends in Ecology & Evolution*, *26*(2), 53-55.