**BIO00058M: Sequence Analysis (2019)**

**Workshop 1: Analysis of “low” throughput sequence**

Kanchon Dasmahapatra (kanchon.dasmahapatra@york.ac.uk)

This workshop demonstrates use of Sanger sequencing. This type of sequence data is generally used for careful, precise analysis of small regions of DNA, such as checking that genetically engineered plants/bacteria/plasmids/animals have the desired sequence. Sanger sequencing usually generates 600-1000bp of sequence that generally has a very low sequence error rate. In this case we will use this technology to construct phylogenies.

You will get a chance to practice how to:

* Clean and edit Sanger sequence chromatograms
* Make and check DNA alignments
* Make phylogenies and calculate bootstrap support
* Search Genbank for sequences

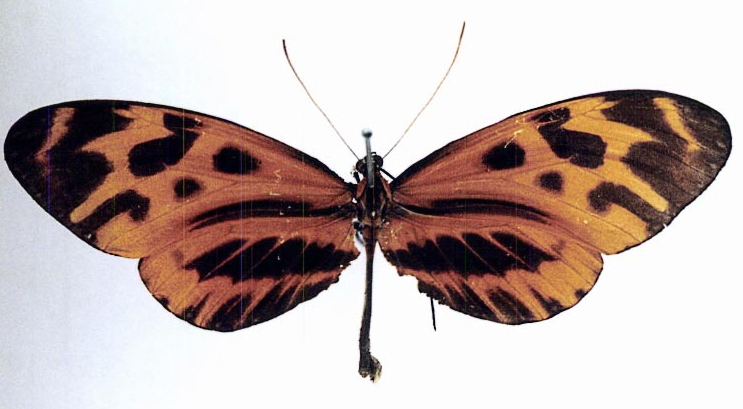
You will need the following software:

**MEGA7**: A new piece of software (that is also sometimes buggy) that can be used for several types of sequence analyses. Here you will use it for making alignments and building phylogenetic trees.

**SeqTrace**: Software for visualising and editing chromatograms.

There are other superior software for these tasks, but usually these are not free. For example, ChromasPro is good for editing chromatograms.

You will be analysing sequences from a neotropical genus of butterflies called *Mechanitis*. These butterflies have black/orange/yellow markings. While four species are recognised, mimicry between species combined with intraspecific variation across their ranges (the bottom four butterflies are the same species, but from different parts of South and Central America) makes it difficult to identify species. It is possible that there are cryptic taxa (individuals that are morpholigically identical to each other but belong to different species) within this species.

You will be able to download 12 Sanger Sequence chromatograms from the VLE. There are sequences from two species, *Mechanitis mazaeus* and *Mechanitis lysimnia*, both from the same area in Peru. For each specimen there are three sequences; one sequence covering ~800bp of a mtDNA gene, and two sequences (generated using forward and reverse primers) for ~1000bp of a nuclear gene called TPI.

**File naming:** Files are names like this: D\_maz\_TPI\_R.ab1.

The letter (D here) denotes the individual, maz = *Mechanitis mazaeus,* lys *= Mechanitis lysimnia,*

TPI/mtDNA denotes the gene and \_F or\_R denoted forward and reverse reads

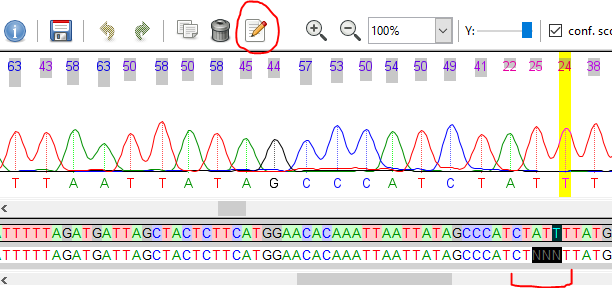
There are also two fasta files called Mechanitis\_mtDNA.fas and Mechanitis\_TPI.fas which you will use in section C of this workshop.

**A) Editing chromatograms**

Download and install the SeqTrace program from <https://code.google.com/archive/p/seqtrace/downloads>, choosing the seqtrace-win-0.9.0.zip option.

Extract the downloaded zip file, and then start the SeqTrace program (seqtrace.exe).

1. In SeqTrace, create a new project (*File > New Project*). In the window that pops up, copy and paste the folder location that contains your downloaded sequences in the *Location of trace files*, and press *OK*.
2. The mtDNA sequences are simpler to edit, so start with these. Add the two mtDNA sequences to your project by clicking the *+* button.
3. To view a particular chromatogram, highlight a sequence and then go to *Traces > View selected trace file*. Above each peak you will see a number, this is a Phred score, and is an assessment of the quality of that particular base (as perceived by the sequencer). Find out what Phred scores mean.
4. You can adjust the x- and y-axis zoom levels. Why does the height of the peaks generally fall further along the sequence?
5. You will notice that SeqTrace has already trimmed the beginning and end of the sequence as well as replacing some bases with Ns when it judges that the quality is low. Usually you will need to do some trimming of the beginning and ends of the sequences.
6. You will see that in some cases Ns have been put in when the sequence actually look clean. In these cases you should change the Ns to appropriate base. To edit a base you need to first select the bases to be edited in the bottom row, followed by the Edit button as shown below:



1. When you are happy with the cleaning/editing, press the save button, and move on to the next sequence.
2. Once both sequences have been edited, export them by going to *Sequences > Export sequences > From all trace files.* This creates a fasta file with all your sequences.
3. Now have a go at editing the nuclear sequences in a new SeqTrace project. When the PCR product is large, it may be useful to have the product sequenced in both forward and reverse directions to reduce the number of errors. This is what has been done for the nuclear gene. Each of the four pairs of sequences can be automatically grouped by going to *Traces > Autogroup trace files > Autogroup trace files*. You can then view both forward and reverse chromatograms simultaneously.
4. Start with sample 4-279. You will see how sequencing from both ends reduces erroneous base calls. Edit the sequences as you see fit.
5. Next move on to sample 4-273. This appears to have some double peaks (two peaks of very similar heights) in the same place. What are these caused by, and what do you do with them?
6. 4-267 and 05-854 are more problematic. There are huge numbers of double peaks in these samples. What has happened here? Is there any salvageable sequence?
7. Once all the sequences have been edited, export them as a fasta file by going to *Sequences > Export sequences > From all trace files.*

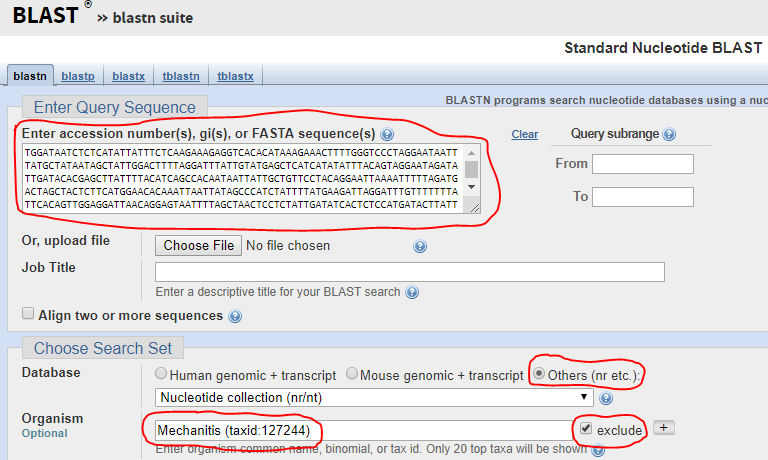
**B) BLAST searches and outgroups**

GenBank is an annotated collection of all publicly available DNA sequences. GenBank shares data on a daily basis with DNA DataBank of Japan (DDBJ) and the European Nucleotide Archive (ENA). One can search this database using a program called BLAST.

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Use Nucleotide BLAST on Genbank (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to find out the identity of the two sequenced genes.

1. Paste one of the mtDNA sequences into the box labelled:  
   **Enter accession number(s), gi(s), or FASTA sequence(s)**Change the database to **Others**

exclude Mechanitis sequences (otherwise all of your top hits will be *Mechanitis*, and you will not be able to find a suitable outgroup for the next section).The click on **BLAST** and wait a bit.  


1. What is a suitable outgroup species to use?   
   An outgroup a species that not within the group of species you are studying (the ‘ingroup’), but sufficiency closely related so you can get a good alignment.  
   This link will help: <http://tolweb.org/Ithomiini/27570>
2. Download the sequence from an appropriate outgroup species by clicking on the Accession Number and then Send to > File > Format FASTA
3. Repeat steps a) to c) using one of the TPI sequences.
4. You will be able to add these outgroup sequences in the next section.

(Note that there are various flavours of BLAST: blastn, blastp, blastx, tblastx, tblastn)

**C) Making alignments**

You need to have a reliable alignment if you want to generate a meaningful phylogeny.

You should now have two fasta files that you have created, each containing a small number of sequences that you edited in the previous section. You will also be making use of the Mechanitis\_mtDNA.fas and Mechanitis\_TPI.fas fasta files which contain additional mtDNA and TPI sequences of *Mechanitis* individuals.

1. Open the mtDNA fasta file that you created with MEGA. *File > Open a file*

Choose the *Align* option

You will see that by default MEGA thinks this is a protein alignment. Change this by going to *Data > DNA sequences*

1. Add in the outgroup sequence/s by going to *Edit > Insert Sequence from File* and choosing the downloaded Genbank fasta file.

Also add in additional *Mechanitis* sequences from the Mechanitis\_mtDNA.fas

1. Within MEGA, use the program called ClustalW. Go to *Alignment > Align by ClustalW* to automatically align the sequences you have just edited. Accept the default penalties (you can find out more about them by clicking on the *Help* button).

Visually inspect the alignment that has been generated. Do you agree with the computer’s alignment? Are there parts of the alignment that you think are dodgy (e.g. near the ends, or near gaps)? You can manually edit the alignment if you think there are mistakes. For example, you can delete the ends.

If necessary, go back to your pre-aligned data and further clean/trim your sequences until you are happy with your alignment.

1. Save the alignment *Data > Export Alignment > FASTA format*
2. Repeat the above steps a) to d) to merge your TPI fasta file with Mechanitis\_TPI.fas

You should end up with two alignments, one for the mtDNA sequences and one for the nuclear sequences.

**D) Phylogenetic analysis**

You can carry out some phylogenetic analysis within MEGA.

1. Use the *Phylogeny* button and the *Construct/Test Neighbor-Joining tree* option to open the mtDNA alignment that you have just made (do you understand some of the many options that come up?).
2. Make a phylogeny from the nuclear gene
3. Add bootstraps using the *Test of phylogeny* option (how many bootstrap replicates are needed?)
4. Try making a Maximum Likelihood phylogeny
5. Are there differences between the two phylogenetic methods

Remember that Neighbor-Joining (NJ) trees are fast and relatively crude, while Maximum Likelihood (ML) trees are better but take more time, particularly when you have many sequences.

There are faster and better programs for making NJ and ML trees such as PhyML and RAxML. However, these usually need to be run from the command line. The advantage to software that can be run on a command line is that it can be automated by ‘scripts’ (small programs) that allow us to run the same methods on hundreds or thousands of genes.

**E) Biological inferences**

The real interest in making phylogenetic trees is to answer biologically interesting questions. You may want to think about the following questions:

1. Are there differences between the mtDNA and nuclear trees?
2. Why might there be differences between these trees
3. Is there any evidence for cryptic species? If so, how many cryptic species do you think there are?
4. What evidence would you need to determine whether or not these were truly different species?