1. **Downloading Sequence Data**

* If all the data is available in one paper, check to see if they have uploaded their alignment or phylogeny to a web database e.g., treebase.
* To download all the available sequences for a species/taxon, go to the online database GenBank. https://www.ncbi.nlm.nih.gov/genbank/

In Genbank you can search for species/taxon of interest. E.g. *Cracticus* (see below)

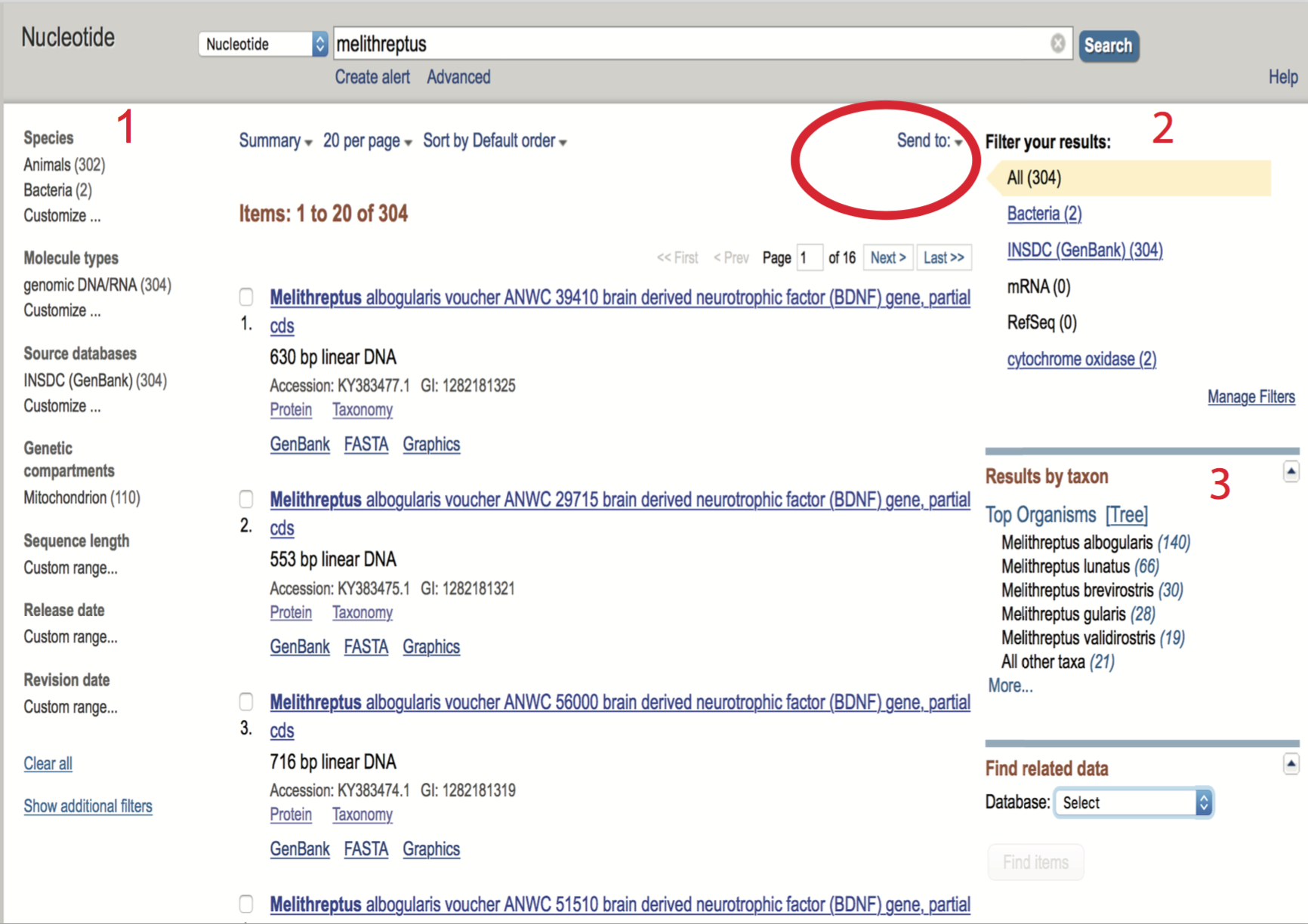
* Check if there is a PopSet (all sequences from a published study) available for your group of interest.



* Downloading a PopSet
  + Click Find Items (see above)
  + Choose the PopSet (study) of interest from the list
  + Under the tab “Send to” (see screenshot below)
    - Complete Record
    - Choose Destination: File
    - Format: Fasta
* A PopSet might include more sequences than you want – you can subset these by opening the file in a text editor and removing the ones you don’t want. Or use the program Mesquite (<https://www.mesquiteproject.org>) to delete the sequences you want to remove.

Selecting specific sequences directly in GenBank. If a PopSet is not available or if you want to use sequences from many different studies - use filters to refine results: e.g. all Melithreptus species (see screenshot below)

1. Sometimes other species will be associated with yours and need to be excluded here e.g. select animals to remove bacteria (see 1 in screenshot)
2. Select the gene or organelle you are interested here if you want to select only a single gene (e.g. select cytochrome oxidase will give you all COI sequences)
3. Select the species you are interested in here.



To download the sequences:

Send to (see screenshot above)

Complete Record

Choose Destination: File

Format: Fasta

1. **Aligning your sequences**

Sequences will need to be aligned using Mafft or in Mega (download to your computer). Either software to fine to use.

Alignment in Mafft online server:

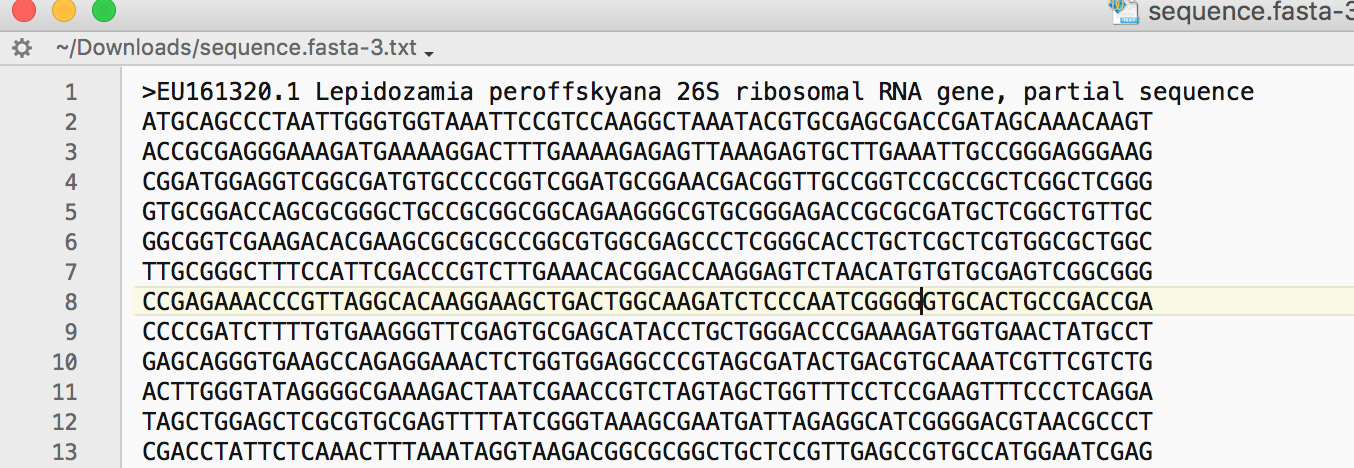
https://mafft.cbrc.jp/alignment/server/

* Paste your fasta file contents here and as long as the data is **coding**, use the default settings. If not come and talk to me and we can discuss parameters to use.
* After running the file, you can use Mafft to make a basic NJ tree. This is a distance method and might not be appropriate for the analysis you need to do but it’s a good way to have a quick look and see if your data/alignment look ok.
* Save the output as Fasta and copy it into your text document and save as ‘Filename.fas’

1. **Editing Taxa names**

* Taxa names will need to be exactly the same across data sets for subsequent analyses. That means if you are going to make a tree with more than one gene, taxa names in each gene data set need to be identical. Or if you want to run a PGLS, your taxa names will need to be identical in the corresponding traits dataset. The names can be edited in Mesquite or in a text editing program. Open the fasta file in a text editor (e.g. for mac use Bbedit). In the example below “>” is part of the fasta format and needs to be retained. The rest of the first line can be edited so that it only reads:

>Lepidozamia\_peroffskyana



1. **Editing Sequence Data**

* Open fasta file in Mesquite and check if data needs to be trimmed. Columns can be selected and deleted. Make your choice based on minimising missing data while maximising sequence length. For example, if one or a few sequences have a longer sequence, trim those nucleotides so all sequences start at the same place. You want to make sure you still have around 600+ base pairs. If one or a few taxa are missing some base pairs and those taxa are important, consider leaving them in and making note that they have missing data.
* Export the file as a simplified nexus file. This can now be read into BEAST.