**Emily’s pipeline guide doc**

Module 1: Downloading sequences from GenBank

**1a.** **1\_Download\_seqs.R**

Inputs:

* Taxa list, inc. outgroup taxa. Genus works better than species names. CSV format list.
* API key to increase rate of requests allowed. More info: <https://ncbiinsights.ncbi.nlm.nih.gov/2017/11/02/new-api-keys-for-the-e-utilities/>

Outputs:

* FASTA files of DNA sequences, organised by search-term e.g. genus.

**1b. 1\_Download\_seqs\_10\_repeats.R**

This script is essentially the same as the one above (1a), except that it repeats the entrez\_link step (linking between taxonomy & nucleotide databases) 10 times.

During my MSc project, there was a problem with this entrez\_link function - you would get different results each time you used it (see: <https://github.com/ropensci/rentrez/issues/188>). We got around this by running it 10 times and compiling the results together, and this would catch pretty much everything (except for maybe a few of the big genera with loads of species, it would catch ~98%). Not ideal but if it is a recurring issue this is a good workaround.

Unfortunately, it looks like I’m still having this issue and the 10 repeats may still be needed – but check this.

It may be worthwhile merging the script that downloads the sequence metadata from GenBank (**4\_Download\_seq\_metadata.R** ) with this script that downloads the sequences - may save time.

**2. 2\_Bind\_and\_clean\_seqs.R**

Inputs:

* Outputs from previous scripts (FASTA files organised by taxa search-terms)

Outputs:

* Filtered sequences in 1 FASTA file

This script filters sequences to remove duplicated sequences (more than one of the same accession number), seqs with the word “genome” in their annotations and seqs with <200 bp or >5000 bp.

**3. 3\_Bait\_and\_pattern\_sort.R**

This script uses “bait files”, which are template/example sequences for each locus. These bait files are used to test for sequence similarity.

Choosing bait-files

* Previously, I manually selected my bait files by looking for sequences of the Type genera of my families of interest, and selecting sequences that have been used in published research and sequences that also were as long as possible.

Inputs:

* Bait files. Maybe we can remove the need to manually find appropriate bait files?
* Variations of loci names – that will be used in the pattern-sort

Outputs:

* Sequences organised by locus (FASTA files)

**4. 4\_Download\_seq\_metadata.R**

This script retrieves all the extra information entered into GenBank for each sequence from GenBank ‘features’.

This information will later be used to link between different sequences that have come from the same specimen, but the extra info may also be useful for other downstream analyses. E.g. geographical, as there is often location info in the form of country, lat & lon etc..

Inputs:

* Direct output from bait & pattern sort. Locus-sorted sequences (FASTA files).

Outputs:

* Data frame of all associated sequence info.