**\*\*\* Under Reconstruction \*\*\***

**ECPP**

Exon Capture for Phylogenetics

Contributors:

Adnan Moussalli, Tim O’Hara, Mark Phuong, Luisa Teasdale, and Jose Grau

This workflow is essentially a BASH wrapper around a selection of established programs and in-house scripts that take you from raw sequence data to final alignments ready for downstream phylogenetic analyses.

This pipeline has many many dependencies, hence I have containerised it. You will need APPTAINER installed on you system. See the following for instructions:

<https://apptainer.org/docs/admin/1.0/installation.html>

<https://github.com/apptainer/apptainer/blob/main/INSTALL.md>

Things you need:

1. Raw sequences, demultiplex. It is a good idea to rename them to something meaningful, e.g. EC1\_FAMILY\_GENUS\_SPECIES\_REGNO\_LOCALITY. This name will carry through all the way to the first tree you produce. Create a new folder for each projects, and in that folder place the raw sequences in a subfolder named “0\_raw”. Note, all sequence files must end in \*\_R1.fastq.gz or \*\_R2.fastq.gz.
2. A reference fasta file containing a single representative sequence for each exon, translated. This is typically your bait design before you chopped them up into baits.
3. A list of the samples you want to process as a text file. These will be the names of the raw sequence files, but without “\_R1.fastq.gz”.
4. Please ensure that the reference fasta file and the sample list text file are in Unix format. “DOS uses carriage return and line feed ("\r\n") as a line ending, while Unix uses just line feed ("\n").” It is the number one reason why the pipeline might fail to execute. Simply install “dos2unix” and convert all text based files you bring over from windows before running the pipelines. One day I will automate this step.

ECPP has 5 modules and are run sequentially. Please follow the syntax outlined in the file ../Test

1. **CLEAN** - this module remove duplicates, subsets the raw reads to only those having a hit to the reference (though the threshold is somewhat relaxed) and quality trims and removes adapters using trimmomatic.
2. **TRINITY** - this module assembles using the CLEAN reads. I have tried and tested many assemblers. Based on my experience, Trinity remains the best options. I plan to benchmark it against MEGAHITS shortly, but for now TRINITY it is. It then tblastn the reference (as query) against the sample specific assembly (as subject). The best hit per reference exon is identified and the local hit coordinates used to extract the corresponding exon match from the corresponding assembled contig. This creates a new reference which is "sample specific".
3. **MAP** - this module maps the CLEAN reads onto the new "sample specific" reference using BBMAP, followed by variant calling using VARSCAN.
4. **ALIGN** - this module will produce a separate fasta alignment for each targeted exon, and 'all exons combined' supermatrix files in fasta, nexus (with CHAR SET defined) and phylip format.
5. **SUBSET** – based on summary file from the last module, you identify either exons or taxa that should be removed. Here is you opportunity to do so. Realignment are done on these subsets with some additional polishing. Final alignments are ready for phylogenetic analyses.