**SCOPP**

Single Cope Orthology Pipeline for Phylogenetics

Contributors:

Adnan Moussalli, Liz Milla

This workflow is essentially a BASH wrapper around a selection of established programs and in-house scripts that screens multiple transcriptome assemblies for single copy othologous genes. There are across multiple take you from raw high-throughput (e.g. illumina) sequence data to final alignments ready for downstream phylogenetic analyses.

For those interested in exploring the currently stable release, I have provided a bash script for the installation of all dependencies.

SCOPP\_1.0.0\_install.sh

This pipeline has many dependencies, and indeed dependencies on specific versions. As much as possible I do try to keep up with updates and upgrades, but please do pay attention to the versions recommended for installation. An ideal approach is to install onto a new virtual machine, using vmware or virtualbox for instance. Ultimately, the objective is that wrap all this up in a DOCKER image.

Things you need:

1. Raw sequences, demultiplex. It is a good idea to rename them to something meaningful, e.g. 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 file must end in \*\_R1.fastq.gz or \*\_R2.fastq.gz.
2. A reference file containing a single representative sequence for each exon, translated.
3. A list of the samples you want to process. These will be the names associated with the raw sequence files, but without “\_R1.fastq.gz”.

**SCOPP** has 5 modules and are run sequentially:

1. **ASSEMBLE** - this module quality trims and removes adapters from PE Illumina raw reads using Trimmomatic (ref), followed by assembly using Trinity (Ref).
2. **LAST** – The assemblies are then screened for local alignment hit to a provided translated CDNA reference.
3. **HOMOLOGY** – Based on LAST local alignment coordinates, all homologous local hit are extracted from the assemblies, producing a separate fasta file of all homologs per reference gene.
4. **CONSENSUS** – Homolog alignments are then collapsed based on a provided threshold. This step also produces a summary table detailing the number of homologs remaining per species per gene.

All cases where only a single homolog remains per species are considered high priority candidates for single copy orthologs suitable for phylogenetic analysis. Filtering based on taxa completeness is recommended at this stage (e.g. 80% of the study taxa have this single copy ortholog assembled). These high priority gene alignments need to visually inspected and assessed for hidden paralogs using, for instance, pipelines such as TreSPex (REF).

1. **SUBSET** - this module will produce a seprate alignment for each targeted exon (fasta files), and 'all exons combined' supermatrix files in fasta/nexus (with CHAR SET defined)/phylip format.