Funstraveganza Proposal: RNASEQ UPSTREAM RAW READS PROCESSING AND POST-ALIGNMENT ANALYSIS

Using the available pre-processed raw reads and post-alignment expression values from experimental datasets that are housed in the Gene Expression Atlas (<https://www.ebi.ac.uk/gxa/home/)>, this project will attempt to perform a standard RNAseq analysis and initial gene regulatory network construction.

Briefly:

1. The raw sequence reads will be trimmed (trimmomatic and/or other quality control packages) and aligned (bowtie package) to their given genome with proper annotation files (gff3 files god-willing). This will utilize basic Unix skills and data file handling abilities. Running differential expression analysis packages might be an option provided timely completion of the alignment.
2. The Gene Expression Atlas already has the differential expression (DE) files that correspond to the raw reads in Step 1, and these files can be utilized while alignment and/or DE are running simultaneously. Some potential options for analyzing this DE data are:
   1. GO enrichment analysis on genes displaying a DE significance threshold (i.e. >1 log2 fold change compared to control). This will use our Python data manipulation skills and Ontology Enrichment familiarity.
   2. Transcription factor binding site enrichment analysis by gathering the upstream sequence of the DE genes from Step2a and submitting them to the MEME suite program. This will require cultivating gene ids and getting their upstream sequence data.
   3. BLASTing interesting gene groups to identify homologs in different organisms to describe related functions and tie in other gene networks that may have been previously described. This will make use of any ontology databases and Python packages.
   4. Curating co-expression information from our gene list from available databases (STRING or others maybe?).
   5. Piping in our data to programs designed to formulate gene regulatory networks from DE data (ex: WGCNA <https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/)>

Many of these proposed steps (primarily those in Step 2) will utilize a combination of Unix and Python data handling, file conversion and parsing, and installing and running different modules and programs (BioPython, aligners, motif enrichment, etc.) to perform the desired analyses.

This project is flexible towards picking an organism and dataset, but provides useful experience in current expression analysis techniques, essential programming, database querying, and all-around troubleshooting.