#Transcriptomics noteook

Author: Gwen Ellis

Affiliation: University of Vermont, Dept. of Biology

E-mail contact: gwen.ellis@uvm.edu

Start Date: 2023.10.09

End Date: 2023.XX.XX

Project Descriptions:

This notebook will be used to document my workflow during the population genomics module of the 2023 Ecological Genomics course.

Table of Contents:

- Entry 1: 2023-10-09
- Entry 2: 2023-10-11
- Entry 3: 2023-10-16
- Entry 4: 2023-10-18
- Entry 5: 2023-10-23
- Entry 6: 2023-10-25
- Entry 7: 2023-10-30
- Entry 8: 2023-11-01

Entry 1: 2023-10-09. - experimental set-up - Illumina RNAseq library prep protocols (TruSeq3) - (2 x 150bp) w/ the Illumina Novoseq 6000

Entry 2: 2023-10-11. - Acidification + Warming (HH) F11 (HH_F11) - fastp to clean data, not trimming beginning of sequence –> trying to preserve as much info as possible and we could be accidentally removing an important part of the sequence (fastp version 0.23.4) - phred >20, 6bp window, keeping reads >35bp - 96.9% complete BUSCO score for assembly

Entry 3:

2023-10-16. - assessing read quality -> >20M reads/library is usually a good threshold for RNA-seq, on average we had 45.34M reads per library with 44.48M passing (98.14% passing) - summary of refer- ence assembly at ahud_Trinity_assembly_stats.txt - BUSCO stats for assembly: 96.9% complete BUSCOs (7.1% single, 89.8% duplicate) - prepping for mapping our sequencing reads to reference assembly outputs ahud_Trinity.fasta.gene_trans_map and ahud_Trinity.fasta.salmon.idx - HH_F11_Rep1: 91.672% mapping rate

Entry 4:
2023-10-18. -
ahud_total_mapping_rates.txt,
most samples
had >90%
mapping rate
but
AA_F11_rep3
had a mapping
rate of ~68% -
compiling
quant.sf files
(reads for each
transcript for
each sample) for
each sample
into one matrix

Entry 5: 2023-10-23. - redoing assembly: using CD-HIT_EST to cluster the initial assembly based on sequences with 95% similarity, Transdecoder to filter down to only open reading frames - median contig length is now 792bp, and contig N50 is 1069 - BUSCO single increased from 7.1% to 61.8% - However, only 66% mapping now (instead of 98%) because we're losing the reads with variation from the clustered assembly –> losing isoforms because of Transdecoder? losing alleles? losing splice variants? - OA_F2_Rep2 (ocean acidification) could also be an outlier based on read count plot and heat map - plotting PCAs - testing for differential gene expression across generations and treatments - checkking a few of the most DEGs show that expression under O_3A and OW is not additive for OWA - making

---

Entry 6: 2023-10-25. - making a WGCNA for samples: original module clustering is first agnostic of meta-data, but can test if there is a correlation between modules and sample groups - before starting WGCNA, remove all genes with counts < 15 in more than 75% of samples using DESeq - detect outliers and make basic PCA - normalize gene counts - choose soft threshold power –> 6 (strength of correlation) - signed WGCNA (up or down regulation matters for modules) - visualize cluster dendrogram, blue and turquoise have the most genes

Entry 7: 2023-10-30. - continue working on WGCNA –> Do any of the eigengenes from our modules associate with our measured traits? (Pearson's correlation coefficient) - the yellow module for example has a few significant associations. When we look at the eigengene values across samples for this module, we see that there is a split between OWA+OW and OA+AM - fitness meta data = (EPR (egg production rate) + HS (hatching succession) + survival + development time) - with 6 power threshold, we see significant correlation between the grey module and mean fitness indicating that there could be something functionally interesting in that module that could be separated out by increasing power threshold'

Entry 7:
2023-11-01.
