Dec. 4 2024

RNAseq: methods

- Nf-core on 4 datasets combined was too much for program: compute resources
- Too many nuances that can't be accounted for when everything is pooled

Jan 3 2025

Issue 3:
- Created merged metadata

Issue #4:
- Differential abundance init. Approach - toy example success
- GTF file issues

-Steve ran mutual info on roberto's data
- i applied his script to the perkinsus datasets (meta data trait generalized)
Obtained merged counts table and ran mutual information but didn't see super strong separation (2025-01-03_RNAseq_all_AI_diffexp.ipynb)

Jan 16 2025

Subset data to get better separation: didn't help too much

Things to consider:
- Fixed vs. random effects
- sample size
- taking control group into consideration
- Batch corrections

9: add study 5
#12: decide best methods and execute
- Integrated data analysis: does not work when data is noisy (too much within and/or across study variation) and signal is not strong enough.
- Talked with Erin witkof about BMC paper (Dina co-author) meta analysis of 12 datasets

Feb 2025

- Study specific effects are much stronger than trait effects
- Attempted batch correction (1. COMBAT and 2. RemoveBatchEffect issue#18) and that had little effect on improving separation based on trait (trait that we generally defined)
- Limit variation within study: Subsetting for common time point DEG against control groups

Commented [1]: Big lesson #1

Commented [2]: Big Lesson #2: traits were oversimplified

April 2025

- Running DifferentialAbundance on each dataset independently
- #26: what flags to use; GC bias and figured out johnson data tag seq and initial rnaseq analysis was not good
- #28 rerun johnson data with different params
    - https://resilience-biomarkers-for-aquaculture.github.io/SW-FastPparams4tagseq/
- #29&31: ran differential abundance on all datasets together but haven't yet interpreted these results
    - https://github.com/Resilience-Biomarkers-for-Aquaculture/Cvirg_Pmarinus_RNAseq/issues/31
        - Didn't know what tables to look so moved on

June 2025

- #32: attempted differentialabundance on datasets separately
    - Steve focused on 5 and shelly focused on study 1
    - We were trying to get a handle on parameters and how to best run differential abundance

July 2025

- #34: Combine study 4 injected + study 5 to increase sample size
    - Will study 4 injected group cluster with resistance or susceptible group from study 5?
    - Learned more about the steps in differentialabundance pipeline when normalization happens
        - PCAs are before any differential abundance analysis happens
    - Attempted batch correction and compared PCAs with and without batch correction on the top 500 most variable genes
        - Starting seeing evidence of innate trait
    - Revisit this analysis to see if selecting the 567 genes that showed significantly differential abundance (by DEseq) show greater clustering
- #39: compare DE results from papers and come up with a list of DEGs/markers

August 2025

- Run differentialabundance with GSEA
    - Created GMT file with descriptions
- #41: step-wise approach
    - Step 1: controls vs treated
    - Step 2: resistant vs sensitive
    - https://github.com/Resilience-Biomarkers-for-Aquaculture/Cvirg_Pmarinus_RNAseq/tree/main/analyses/stepwise_differentialabundance

Commented [3]: could return to this

Commented [4]: could return to this

Commented [5]: remaining to be done

- Shelly did dataset1
- Only 1 sig. Gene
  - DEseq isn't ideal for starting with such a pared down set of genes (VST couldn't work well with a set of only 50 genes)
- Are we removing biomarkers that are innate?

- #42: validate sr320 classification results
  - Are the ~50 markers convincing about the difference between sensitive vs. resistant?
  - #43 is sr320's AI model
- #44: combined datasets 1 &5
  - This compared integrated data analysis vs. post-data integration approaches
  - This was completed (6-gene classifier)
    - Pipeline step 1: ranks genes based on reproducibility, consistency of directionality in expression differences, and heterogeneity.
    - Pipeline step 2: Logistic regression (a model that tries to find the minimum number of genes that make a good classifier)
    - SY-gene-classifier-panel
  - There was strong separation between tolerant and sensitive

September 2025

- #36 run differentialabundance independently for each dataset and compare DEGs across all for commonalities
  - Are these completed? Not completed
  - Theme: post-data integration → do we see more overlap from doing this?
- # 45 understand GSEA
  - Not done
- #46 integrate all data and run through differentialabundance pipeline
  - Not done
- #47 not sure where we were going with this, no need to revisit
- #49: plot 6 genes to gain confidence that these distinguish phenotypes
  - #51: replot heatmap with improved clustering and labels
  - #52: coverage density plots
    - Still need a nb entry for this
- #53: innate vs. reactive: https://resilience-biomarkers-for-aquaculture.github.io/SY-innate-gene-expression/
- #54: find more datasets
  - Postponed
-

Commented [6]: Big lesson #3: biomarkers may exist in controls if they are part of an innate process and we don't want to filter these out!

Commented [7]: Revisit: make plots and see if they are convincing

Commented [8]: Big lesson #4: only include training set in test set if exploring within study. If trying to preduce a model that will predict phenotypes in other studies you should definitely not include the training in the test set

Commented [9]: could return to this. good question about post-data integration vs. integrated data analysis. but uncertain about the subsetting mentioned in the issue

Commented [10]: could revisit

Commented [11]: revisit this