**PASS1C/PASS1A Analysis Plan**

1. Data Structure
   1. Organize and Clean UM data
   2. Batches by PASS, Location, Technology, Named, Targeted
      1. Visualize in colored excel tables
   3. Sample and feature dimensions
      * 1. Documents data dimensions (NxP matrixes by Tissue)
        2. Controls and Experimental samples
        3. Experimental Groups
        4. Control Samples by Tissue
2. QC Analysis for PASS1A and PASS1C
   1. Exploratory Data Analysis
      1. Distribution of Zero/NA values across samples/features
      2. Distribution of Sample Medians
      3. Visualize NxP Matrices
         1. Zero Values
         2. NA Values
         3. Abundances
            1. Raw
            2. Transformed (x,y,z)
            3. Transformed and Clustered (JZL Outlier Insights)
      4. Sample by Sample Correlation Matrices (Spearman)
         1. Sample Order
            1. Run Order (w/ & w/out references)
            2. Likely Sex/Hour (2 time axes)
         2. Transformed
            1. Raw/Log/Additional Transformations
         3. Correlation Medians
            1. Run Order (w/out References)
            2. Sex/Hour (2 time axes)
      5. Sample PCA Plots
3. Determine Shared Samples and Features between Batches
   1. Determine if a reproducibility analysis is appropriate
   2. NxN Correlation Matrices
      1. Shared Samples and Individual Samples
4. After Exploring the data, perform modeling. Whether that’s the modeling we previously performed or a new technique that takes the new data into account.

ADDITIONAL NOTES/THOUGHTS:

-Data Structure (know what we’re working with)

-QC PASS1A and PASS1C

-Examine the data, cluster it and see where the new controls lie in relation to other samples (pca, correlation matrixes)

-I’d also like to inquire how reproducible the PASS1A control are compared to PASS1C. I’d like to measure shared correlation with a reproducibility analysis.

-Let’s image the controls cluster together, but there’s a batch effect

-We want to measure that batch effect, but we want to measure it across gene/metabolites (let’s just say genes)

-We want to see the distribution of that bacth effect across genes, is it uniform across genes?

-Let’s assume that it is more or less

-Then we have another effect we are looking for: that from circadian rhythm and that from feeding

-We have classified genes based 2 competing models: a model to model circadian rhythm and a model for acute exercsie response.

-So we have a group of genes that are (for lack of a better way to put it) genes that fit circadian modeling and those that fit acute exercise modeling.

-We’ll ask for DE genes between different batches to get an idea of gene sets that differ.

-Now, if we correct for a batch effect, we’ll need to ensure we do not inadvertently correct for some biological variable of interest (the models I discussed above)

-Let’s say we are successful in doing that: that is correcting for batch effect but keeping the biological symbols in tact

-Then it’s time to ask the differences again: we’re going to see DE genes again

- I hope that those different genes are the 1-3 percent of genes that are strongly associated with circadian rhythm—but we will have to see

-Finally, we would apply our competing models for circadian rhythm and acute exercsie response and then again bin the genes for how circadian they are and how acute exercise responsive they are.

This will build insights:

-First was our prior modeling accurate?

-Was PASS1C necessary

-This final insight doesn’t get much attention, but I think the biology is exciting. There must be circadian rhythm genes that respond to acute exercise. These data would be limited, but I’d be excited to:

a. Identify those genes

b. Gain insight into how they respond to acute exercise under certain circumstances