Analysis Plan 20200601

Primary Goals:

1. Generate annotation table across tissues to aid in goals
   1. Output Table format:
      1. Gene, tissue, C0vsC7, C0vsE7,SOM,K-means,Circ\_yan, circ\_kegg
2. Bin gene signatures in these categories
   1. Circadian
   2. Exercise Responsive
   3. Ambiguous
   4. Irrelevant
   5. Ambiguous Exercise Response
3. Generate Summary Plots
   1. R2 x R2 plots for pooled tissues
      1. Circles around dots: som clusters
      2. Color of dots: up or down genes
         1. First attempt, make binary, then adjust to stronger log2FC
      3. Size of dots: Shared genes (Up/down) across tissues
   2. R2 x R2 plot for individual tissues
      1. Circles: som clusters
      2. Color: up/down genes
   3. Matched with R2 x R2 plot, we should have pathway enrichment analysis plots
4. Generate model plots of individual genes by bin
   1. Strong Exercise Response
   2. Strong Circadian Response
   3. Ambiguous Gene
   4. Ambiguous gene with exercise response (plot pseudo Bollinger bands)
5. Design a scoring system that attempts to bin genes based on test results (tests described below)
   1. Output will be a table of all scores
6. Perform experiment: Remove all circ genes and perform pathway enrichment analysis
7. Annotate the “Circadian Genome” and visualize with a density plot

Tests to bin genes by Circadian/Exercise Status (ordered by priority)

1. C0 -> C7, C0 -> E7, C7vsE7 comparisons (Jiayu)
   1. Method: Create a table with results, place in one script
   2. Tissue specific: Outliers, duplicates, tissue-specific variance
   3. Output Table:
      1. Columns: Gene, Tissue, C0vsC7, C0vsE7, Inference
2. Phase Comparisons (Jiayu & Alec)
   1. Visualize phase distributions
      1. Datasets:
         1. SIN/COS model
         2. Yan Sets(tissue specific)
         3. Jun’s Set
      2. Annotations:
         1. Dataset
         2. Up/down/NS
   2. Apply KS test to see if pools come from similar distributions
      1. <https://towardsdatascience.com/when-to-use-the-kolmogorov-smirnov-test-dd0b2c8a8f61>
   3. Output table:
      1. TODO:
      2. Columns: Gene, tissue, SINmod-vs-YanSet, ks-value (D, p)
3. Generate pseudo Bollinger curves around model (Alec)
   1. Bollinger test for circadian model:
      1. Purpose: Evidence of exercise response amongst circadian genes
      2. There are n timepoints that represent pseudo exercise response (E0, E0.5, E1, E4, E7), perform median on these time points and plot, if median crosses time points, then infer exercise response.
         1. Essentially a t-test between model prediction and median expression and exercise hours
         2. Correct for multiple hypothesis tests
   2. Output table:
      1. Gene, Tissue, E0\_med, E0.5\_med, E1\_med, E4\_med, E7\_med, E0\_E7\_med\_sum, FDR
   3. During modeling:
      1. Recalculate the x-axis
      2. Investigate the effect of covariants (Sex, Aging)
4. Perform modeling on residuals (Alec & Jiayu)
   1. Perform circadian model, collect residual, and apply exercise model to residual
   2. Perform exercise model, collect residual, and apply circadian model to residual
   3. Output table:
      1. Columns: Gene, Tissue, CIRCMOD-C1 Metrics, EXMOD-C1 Metrics,ModelAC2, ModelBC2