**Notes from 5/20/20**

1. **Notes on TGFb project**
   1. Likely seeing batch effects causing high variability  
      **CL**: not necessarily a batch effect, but this set of assays does seem remarkably more variable than the original assay set
   2. Sample sizes are small and harder to get a good significance rating
   3. Could be that the differences in expression in the control group is captured because the averaging of gene expression between replicates. Perhaps these difference aren’t truly as significant  
      **CL:** correct, or outliers. You might want to plot many of the top genes to double check.
   4. We could potentially attempt to adjust batch effects(SVA, others?)
   5. When we acquire new data we can see figures cluster by run
      1. Run1 and run2 include into the design model
         1. Create PCA
         2. Heat map (cluster by genes and do another with clustering by genes and samples)
         3. See if run1 and run2 associate with any SVs found
         4. We can see if dendrograms clusters with heatmap aside from the heatmap itself   
            **CL:** right, see if clustering the heatmap by column reveals any patterns
         5. We can also see how expression differs in one gene from DMSO1 vs DMSO2 using y = expression X = group
         6. Stick to P-value < .05
   6. Dendrograms cluster according to similarity in expression based on hierarchical clustering. The shorter the branch the more the expression values are similar to each other.
   7. Typically we view maybe a histogram of pvalues/FDR or some form of the results and look at what we get then decide on a precise significance scores to use. FDR < .05 means 5% of ours results could be false positives  
      **CL:** yep, or I’d phrase this as, “we anticipate around 5% of the genes we call significant to be false positives”
   8. Also pay attention to the replicate # 3 group and see if these replicates cluster which might indicate batch effects of this run. Perhaps the drug wasn’t administered adequately.
2. Notes on p20003 menard Katcher EoE project
   1. When writing code and generally try to keep separate objects for each model or rather have another script for each model. Also keep the objects concise, consistent and informative.   
      **CL:** as a reminder, I like this because it’s more reproducible (allows colleagues and other researchers to review your code) and will ultimately result in less manual work from you.
   2. SVA code is approved
   3. Regarding associations of SVs and variables  
      **CL:** please have a first pass attempt at this for our meeting on Monday (table is sufficient, heatmap is visually nice but not necessary)
      1. Gather pearson correlation coefficients of each comparison and format the table into a way that allows for a heat map. Separate one for categorical and continuous. One set of variables(cat and continuous) will be based on Control, EoE, and FS EoE, while one will be only EoE and FS EoE. (4 heatmaps total)
         1. For the categorical variables use anova for gathering Pvalues the represent associations between the metadata variables and the SVs
            1. Something like: anova\_obj <- lm(SV1 ~ Phen, metadata) & then : Summary(anova\_obj)$Coeff[“Phenotype”, “Pvalue”]   
                 
               **CL:** here’s example code that isn’t off the top of my head--  
               anova\_obj <- lm(SV1 ~ Phenotype, metadata)  
               summary(anova\_obj)$coef  
               summary(anova\_obj)$coef[2, 4]  
                 
               These steps respectively  
               1) run an ANOVA  
               2) look at anova solution (note that the second row output here has the p-value we want -- the “intercept” term will make sense to you after you get more training in modeling  
               3) grab row 2, column 4, our ANOVA p-value

**Notes from 5/27/20**

1. Regarding Calies report
   1. SVA and potential confounders
      1. Redo using SVs from model without Age adjustment
      2. Also include RIN and phenotype
      3. Turn of clustering by SV so we can see them in their order
      4. Redo the color scale for Categorical as a one direction representation of pvalues (monochromatic legend)
      5. Potentially include the pvalues in the respective boxes
   2. Analysis may need to be redone entirely with RUV so we can compare SVA to RUV and see if the results are similar so that we can be more comfortable with SVA in general for the project
   3. General info
      1. As far as direction of correlation goes with meta data and SVs, it does not matter quite as much as the fact that there is a potential association at all
      2. Let Pvalues from table 1 somewhat guide you but do consider the actual biology of the question at hand an initial glance at the pvalues may be providing a false interpretation   
         CL: this is because we may not have statistical power to detect true differences.

**Notes from 5/29/2020**

1. Calies report
   1. SVA and potential confounders
      1. Fix the legend and color for continuous map (equidistance)
      2. Fix the color scale for categorical variables and try to emphasize under Pvalue<.05
      3. Create a standalone report with this information   
         **CL:** Note that our due date was Tues June 2nd @ 5PM.
         1. Must have
            1. Table1
            2. Heatmaps + respective table
            3. Repeated using RUV
      4. SV associations
         1. SV1 -> inflamm
         2. SV2 -> RIN
         3. SV3 -> Age (somewhat)
         4. SV4 -> not really associating with variables provided
   2. General Info
      1. Eosinophil represents white bloods cells representing inflammation
      2. Fgsea (github) is faster for ontology
      3. GO -> webGestalt is a good GO tool
      4. Choo also thinks gsea output is limited by small sample sizes  
         CL: this note is a bit misleading -- rather both I and Lauren do not suggest limiting the gene sets going into GSEA, as the permutation step is the most computationally intensive part

CL: example code for RUV attached. I use tidyverse, vegan, and the RUVSeq R packages. This is copied out of an existing project for me, so let me know ASAP if you need more guidance.

# pre-RUV formula

preruv\_formula <-

"~ Phenotype" %>%

as.formula

# DESeq2 object --> VST transform

vst\_counts <-

DESeqDataSetFromMatrix(

countData =

final\_counts %>%

select(-alias\_id) %>%

t(),

design = preruv\_formula,

colData = phenodata\_final

) %>%

estimateSizeFactors() %>%

estimateDispersions() %>%

varianceStabilizingTransformation()

# grab residuals

vst\_residuals <-

vst\_counts %>%

assay %>%

apply(., 1,

function(y) {

tmp <-

phenodata\_final %>%

bind\_cols(y = y)

lm(update(preruv\_formula, y ~ .),

tmp) %>%

resid() }

)

# check # of RUV components -- including this code here for you to run

# but just use k = 4 for now for concordance with SVA

check\_ruv\_kvals <-

function(k) {

ruv\_tmp <-

RUVr(x = assay(vst\_counts),

k = k,

residuals = vst\_residuals %>% t,

isLog = T)$W

adonis(

vst\_residuals ~ ruv\_tmp,

method = "manhattan") %>%

.$aov.tab %>%

.$R2 %>%

.[1]

}

r2\_by\_k <- sapply(1:20, check\_ruv\_kvals)

ggplot(data = NULL, aes(x = 1:20, y = r2\_by\_k)) +

geom\_point() +

geom\_line() +

xlab("# of RUVr Components (k)") +

ylab("% of Residual Variance Explained\n(Multivariate PERMANOVA)") +

theme\_classic()

ruv\_soln <-

RUVr(x = assay(vst\_counts),

k = 4,

residuals = vst\_residuals %>% t,

isLog = T)

phenodata\_final$RUV1 <- ruv\_soln$W[ , 1]

phenodata\_final$RUV2 <- ruv\_soln$W[ , 2]

phenodata\_final$RUV3 <- ruv\_soln$W[ , 3]

phenodata\_final$RUV4 <- ruv\_soln$W[ , 4]

**6/8/20**

1. **Choosing # of inferred covariates from RUV** 
   1. **Sometimes it over estimates**
   2. **Typically it plateaus**
2. **Model 3 and Model 2 are most likely best choice**
   1. **Model 2 -> ~Phenotype + RIN (RIN adjusted)**
   2. **Model 3 -> ~Phenotype + RUV1 + RUV2 (RUV Adjusted)**
3. **Not intended to go over recent report with the PI**
   1. **Miscommunication with Choo**
   2. **Type out notes during meeting**
   3. **Review before meeting ends**
   4. **Determine who the current information is intended for**
4. **Don’t send anything without review ever!!!**
5. **SVA should not be capturing phenotype (what’s inputted into the model)**
   1. **RUV looks better**
6. **Always look at diagnostics post variation removal method (determines if functioning properly)**
7. **Sometimes run both to compare (Ideally, or if seeing something clearly wonky)**
8. **Moving forward with RUV!**
9. **Create venn diagrams with new models**
10. **Report with essential results (end of day wednesday)**
    1. **Be clear about models were comparing (describe too)**
    2. **Venn diagrams for comparing these models and contrasts between them**
    3. **Have all gene lists in nice clean excel document (clearly labeled nicely)**
11. **Share (report, code and excel document prior to meeting) with Choo and Katerina before meeting on wednesday at 2pm (end of day wednesday)**
12. **Respond to all emails even something as simple as quoting the previous email and acknowledging that I’ve seen it.**