**Meeting with Katerina, Choo, and Calies**

Meeting with Katerina and Choo

5/8/20

Notes for EoE project

ANALYSIS

1. Keep the ‘Main phenotype’ model ( ~Phenotype) but in addition for the actual comparison of age vs phenotype, create a model (~Phenotype + SV1 + SV2 + SV3 + SV4)
   1. Ultimately make the comparison between (~Phenotype + SV1 + SV2 + SV3 + SV4) and (~Phenotype + Age + SV1 + SV2 + SV3 + SV4)
2. Plot genes of interest by expression and age not just expression and phenotype
3. Perform STC analysis by grouping them based on the presence of STC or not and comparing those
   1. Model: ~Phenotype + any SVs that are found? (phenotype in this case would be control and Y or N to the STC)
4. For the ~ Phenotype + SVA model, please present some plots examining whether RIN, peak eos, etc are associated with each surrogate variable. -- Choo
5. Explore number of SVs - KK

PRESENTATION

1. Redo the structure of the report to having the PCAs and Table one information at the top for the exploratory portion of the report
2. Table one should not have the NULL p Values.
3. Run a spell check on the report
4. Decribe the content of the figure and analysis better (i.e. gene/rna expression and not ‘DNA’ expression)
5. Table one can be shown in a better format? (what do u guys think? Should it match the DE tables format?)
6. Learn the proper naming conventions and rename all items
7. Adopt version control principles. At minimum this should have the date that the report was last worked on in the file title, and CIDA header info (e.g., project number) -- Choo

Here is a link with some best practices: <https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming> - KK

5/14/20

Notes for EoE project

To-do now & PRESENTATION

1. Compile the tables and work on a few more edits such as table naming, file naming, move boxplots to respective tables (stand alone) to the report to send to Calies with new title and date.
2. Select date for meeting with Choo
3. **KK:** Schedule separate meeting with Katerina (see below)
4. Select date for next meeting with Calies

ANALYSIS

1. Discuss/analyze issues with SV1’s potential association with the phenotype. Adjustment might be giving us false negatives but at the same time consolidating more precise genes?
2. **KK:** A summary of the SV by variables may be useful, otherwise it’s a lot of plots. A heatmap of SVs BY variable with correlation or p-value of association may give a quick snapshot and then only show scatter plots of the ones that show some association (also change x and y axis,s ee below).
3. Think about an analysis with continuous measurements
4. Wait for Calies and her Clinical group to decide on a final model before proceeding
5. Select final model or models to continue with and move on with the aim 2 portion of the analysis
6. Tackle remaining scope questions

REPORT IMPROVEMENTS

1. **CL** I might retitle what you have so far as something like "EoE Interim Results: Differences by Categorical Phenotype" to emphasize that we only have the categorical or "Aim 1" results done for now.
2. **CL** Under "Q1. Are there differences in gene expression by categorical EoE phenotype?" please label these tables to make it clear which models were used.
3. **CL**  in addition, please move the figures up so the differentially expressed genes by phenotype (boxplots) are right below their appropriate tables (pairwise contrasts). Similarly, keep the plots of expression by age near the age table.
4. **CL** to make document a little bit more standalone -- in particular, label tables clearly to indicate what model was used.
5. **KK** In general, check your language and spell check (e.g. on page 14 “Multiple design model were used …”, should be plural “models”; models are not ‘
6. **KK** Include PCA plots only for EoE subjects (Calies mentioned this)
7. Right after “Comprehensive Results”, you describe the SVAs after Model 1, but you use them in Model 1.
8. **KK** Flip X and Y axis in SV plots - I found what you have harder to see
9. **KK** What is the table “Differential expression results reflecting the age ….. “ are genes that have an age association? And in what model?
10. **KK** Not clear what “Top 8 expression plots by Age” figures are? Clarify model? I don’t see a strong age association for RPS17.
11. **KK** In STC analysis, improve description of “Phenotype2”, it’s confusing.
12. **KK** Improve this sentence “The following PCA indicates color by the STC presence and the follow models are shaped and colored by (KK: what?) to compare the age adjusted Phenotype against the new STC model phenotype
13. **KK** For the STC PCAs, why are there two different legends? Sometimes with phenotype, other times with phenotype2. Clarify/clean up
14. **KK** Be consistent with ordering - your Venn diagrams have age-adjusted model on the left and main phenotype on the right. But later for the STC analysis, you switch these. I would also have model 1 on the left so change the first Venn diagram.

ORAL PRESENTATION IMPROVEMENT

1. Work on language to use to describe results (i'm assuming things)
   1. **KK:** Understand and explain the concept of “Association is not Causation”

in the context of these types of human studies

* + 1. Read this resources and practice explaining in separate meeting with me - schedule one for the next week

<http://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module1A-Populations/index.html>

Focus on module 6+

<https://www.nature.com/articles/nmeth.3587>

* + 1. Do not use “causes”, “depends”, “results in”
    2. Use “associated”, “correlated”, “trends”, “patterns”
  1. **KK:** Understand and correctly explain how PCA works
     1. There are many resources available, read these and practice explaining in separate meeting with me (see above)

<https://blog.bioturing.com/2018/06/14/principal-component-analysis-explained-simply/>

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-194

* 1. **KK:** Understand and correctly explain p-values
     1. There are many resources available, read these and associated links and practice describing in separate meeting with me (see above) <https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-values>

1. Enunciate and project louder (I recorded the video and watched it again. Feel free to ask for it if interested)
2. Start each meeting with a summary and recap of previous work and summary of what’s been done since then.

**Meeting 3**

**6/5/20**

1. **The scaling of the coefficients is not ideal for truly representing the scores**
   1. **Must not scale and but fix legend for non row scaled data**
2. **RUV and SVA are similar approaches and help us conclude a couple things**
   1. **We definitely see some association with Inflammation in the 1st inferred covariates**
   2. **We also see RIN (sample quality) as an association with the 2nd inferred covariate.**
3. **Approach for moving forward**
   1. **We believe that adjusting for sample quality (RIN) is important**
   2. **We’re hesitant on including SVA or RUV approaches due to the first inferred covariates association with the phenotype**
   3. **We would like to lastly consider 3 approaches as Choo trusts RUV more so** 
      1. **Not adjust for anything**
      2. **Adjust for RIN in the DESeq model**
      3. **Adjust for RUV1 and RUV2 in the DESeq model**
   4. **Next steps are to take a look at the results that these models provide and further conclude the final model going forward by monday at 2pm or when the exact meeting is scheduled.**

**6/10/20**

1. Upload all documents to the shared storage (C drive) so that Choo can check my code. Do this by 5pm thursday (6/11/20)
   1. If wrong then the code shall be corrected
      1. Then reevaluate and select the best model to move forward with
2. Use FDR <.05
3. Compare the two models by overlapping the same contrasts from each model
4. Gene by phenotype should be box plots w/ points
5. Create gene lists of the intersection of the two models
6. Use webgestalt for a first pass GO analysis by 5pm monday (6/15)