Analysis Plan: GRADS PBMC transcriptome

*<Once have a cleaned data set>*

Xiting, here are the analyses I recommend we start looking at:

**1. “unsupervised” analyses**

**Remaining questions related to the Analysis plan:**

* How many features to include: 500-100 most variable genes across samples??
* What is the reference group? for me, the goal here is to look at the variation of GE across clinical phenotypes. We should be using healthy control samples as the denominator so we can look at the variation across all the clinical phenotypes. Worth a discussion about this so I can explain my perspective.

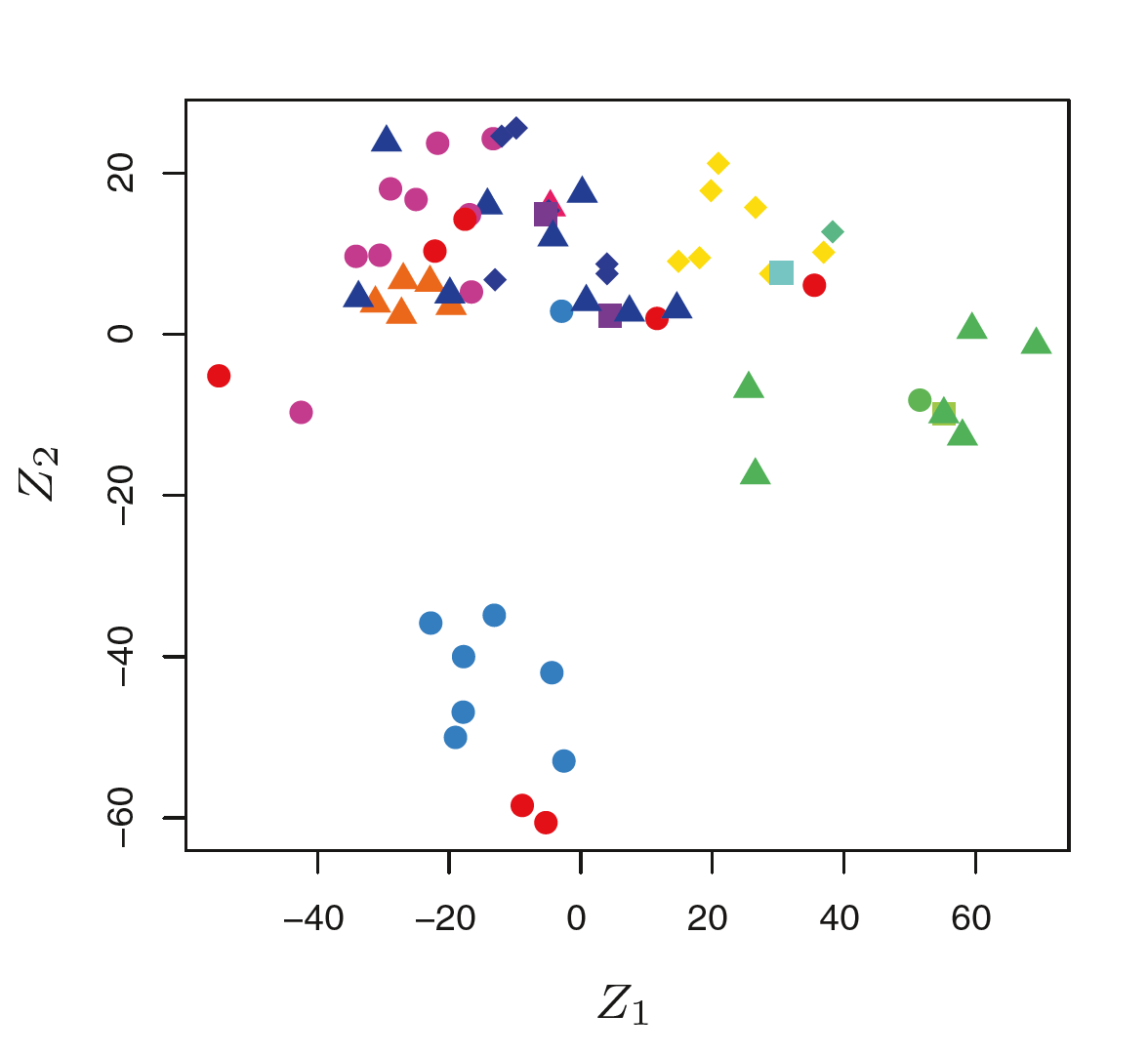
*1a. heatmap*

i. most variable genes: Addresses the question of how distinct are clinical phenotypes from a gene/biology perspective

ii. IFN related genes: addresses question of whether differences of IFN inflammation across clinical phenotypes (I have list of IFN genes ready)

*1b. clustering:*

Addresses the question of how many different patterns of gene expression are there in PBMC; color code the phenotypes and see if get something like the figure to right example



Need to decide on Gene selection: 500 most variable across samples

Types of approaches:

HOPACH: (advantage of this method is it will generate a cluster number, k, based on the mathematical properties of the data)

PCA using markers to indicate each clinical phenotype like this example to the right

*1c. find clinical meaning in clusters:*

If we identify clusters of patients based on genes, compare clinical variables across clusters

*Overall Goals for this analysis: These analyses will help me understand the overall patterns; I think these approaches are a good thing to do early on to develop ideas about the data; ie, help develop next steps and hypotheses.*

*The types of things I am interested in seeing from the plots above include, are there patterns of blood GE shared within clinical phenotypes but not across phenotypes using most variable genes? If you narrow your analysis to IFN or IRGs do you find some clinical phenotypes show higher or lower levels of interferon GE? How many clusters or PCA groups do there appear to be?*

**2. “Supervised” analyses**

*2a. Classification machine learning*

goal to identify genes predicting different clinical phenotypes. Those of interest might be:

i. cardiac

ii. stage IV

problem with this approach: too few cardiac cases to do cross validation ML approaches (maybe too few for stage IV as well).

notes from papers on the topic:

Recently, there have been limited studies that have assessed RNA-seq data with supervised and unsupervised machine learning techniques (Thompson et al. 2016). However, these studies utilized RNA-seq data by leveraging only gene-level expression data rather than more detailed transcript-level data available for the alternative splicing transcripts (Chen and Manley 2009). Most recently, a study analyzed the utility of RNA-seq transcriptlevel data for the disease/nondisease phenotype classification of the samples, showing the advantage of the transcript expression data for the disease phenotype prediction task (Labuzzetta et al. 2016).

To reduce the dimensionality of the feature space, a feature selection method (Hall and Smith 1998) was applied in a classification-specific and data set-specific manner,

Regardless of the classification task or data set, the normalization of the RNA-seq data did not make a significant difference on the choice of the selected features: Variation in the numbers of selected features was <1%

Next, we hypothesized that because of the observed specificity of alternative splicing across tissues, ages, sexes, and between disease/normal phenotypes, training classifiers with the RNAseq data at the transcript level for the biological classification tasks could increase the classification accuracy (Hall and Smith 1998; Xiong et al. 2015).

The most frequently top performing methods were the random forest and logistic regression classifiers

**3. Candidate gene analyses:**

3a. gene scores as independent predictors of end stage lung disease (IFN factor, TCR factor, IFN-inducible chemokines)

Univariate analysis: candidate gene scores plotted against CXR stage; do genes add anything more than serum CD4 counts?

If there appears to be stage associated differences in levels, move to multivariate modeling

Outcome=stage 4

Predictors: gene score, duration, lymph level, immunosuppression, age, race

**4. Longitudinal analysis:**

4a. identify which subjects had a drop in 6 month FVC, FEV1, TLC or DLCO by 300 cc’s for volume or 3 unit drop

Would it be possible to identify ge changes in this group c/w those without PFT drops. Calculate change in gene expression unit from first to second measurement. Then rank changes largest in magnitude for one group vs the other

**5. Comparison to other fibrotic lung disease:**

Can take a candidate gene approach to comparing genes upregulated in PBMC of IPF patients analyzed by Naftali. Plot gene expression by CXR stage (consider other imaging features depending on the data available)