B-528-SP21
5/1/2021
Jack Kaefer

Characterization of Pathways Associated with COPD and NSCLC

Since the height of tobacco use in the United States in the 60s the notion of smoking has slowly been changing in the public's eye. In recent decades it has become less frequent in the United states thanks to stigma in the media. Despite campaigns focused on emphasizing negative aspects of smoking, in 2014 it was recognized by the CDC as the 3rd leading cause of death in the United States.

I chose to look into the studies of two different labs, and study the correlation between lung diseases like COPD and NSCLC and genetic pathways.

Datasets:

Paper 1:

This study performed a reanalysis of microarray datasets collected in COPD studies. The datasets were from blood samples and consisted of: 574 control samples and 688 COPD samples

Paper 2:

This study was an effort to see certain genes implicated in transformation from COPD into NSCLC. The datasets were from tissue samples and consisted of: 58 NSCLC Samples and 58 COPD samples.

Pre-processing(done with projectScript.r at root of project folder):

The first step I took was to normalize the data with the RMA method from different packages, affy, oligo, and limma, affy for *AffymetrixHumanGenomePlus2.0,* oligo for *Affymetrix Human Gene 1.1 ST Array* and limma for *Agilent-014850.* I made the assumption that the RMA function from each of the packages function in relatively the same way. I applied avereps to the Microarrays in order to remove duplicate probes and replace them with a single averaged probe.
In accessing the phenotype sdrf file for GSE106899 or the Illumina dataset from Paper 2 there were no phenotypes other than disease so I decided not to proceed with it.

Having some preprocessing steps out of the way I considered what else could refine the data to emphasize similar genes. Genes uncommon between the sets would not be used to associate changes with other studies so I decided to get rid of those. In R I first resolved the probe ids to gene symbols using different steps for the different data. For the agilent datasets from Paper 2, annotation files were included which associated the probe numbers to gene information stored in GFF like format. This required converting the annotation file into a dataframe in R. All of the Affymetrix Genome Plus 2.0 probes could be successfully resolved to their gene symbols by querying various bioconductor database packages in R. Probe Ids which were now Gene Symbols could be directly compared to genes represented on different chip platforms and. In order to keep track of the intersecting (or common genes) I created an intersection file. This intersection file was saved in python pickle format which was allowed by the R reticulate embedded python package. This allows exchange of dataframes from one language to another and increased functionality of the data structure(standalone R has variable to binary packages, but pickle works for dataframes).

Having found intersected genes, I plotted the first and second principal components of the different sample sets merged. In order to correct for the variance introduced by the different samples

ComBat was used with the parametric(assumption of normal distribution within samples) correction option(**Figure 1.**) to reduce the variance in principal components. This effectively created a new dataframe where the samples could still be retrieved.

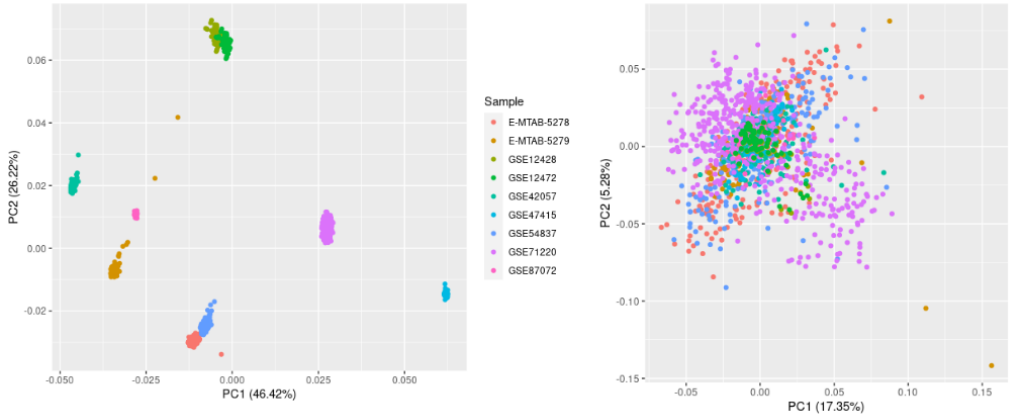


Figure 1. PCA Plots Pre(Left) and Post Combat

There is the possibility of removing as the output of I first took into account the different kinds of data present from either of the studies. The Chinese study used publicly available datasets that were from the source of the disease, samples of the affected tissue, whereas the American study used entirely blood cell gene expression data. If anything this means that the genes discovered in the extracellular matrix were a lot less representative than the cells at the source.

<u>Prediction and Analysis:</u>

Grouping into further categories:

Expression sets generated by R were used by a python jupyter notebook titled postcombat.ipynb to facilitate separating samples of studies into their respective categories: disease state, gender, or age. Doing this was dependent on a script called genPheno.py which for the Gene Omnibus data was able to call the phenotype data, but with the Affymetrix data had to read sdrf.txt files. The Use of python was simply because of familiarity with the program and thus ease of use. Output from the python scripts were expression sets that were restricted on the top level by presence or absence of COPD. The reason for outputting category refined dataframe/expression sets was for ease of creating a design matrix. (If an expression set containing only copd samples was combined with another containing only control samples then upon merging, the order and location of the samples could be determined).

Finding DEGs:

Having the design matrix in R allowed for classifying differential expression and the significance of the differential expression using the eBayes function. With topTable the most significant DEGs could be displayed and refined to the n topmost DEGs. Different strategies were used to emphasize similarity between and within samples and to classify COPD:

Strategy1: Predicting COPD using generalized expression score

Looking for a way to use general commonalities in overall gene expression to characterize them. (The number of genes being explored were already reduced from ~46-54k per chip to 12685). Having removed uncommon genes, an intersection between differentially expressed genes of the samples could indicate genes that are similarly expressed and possibly be associated with COPD. Taking intersections

between the top 100 differentially expressed genes in each category resulted in no common genes. Increasing this number to 1000 top DEGs allowed for around 4 or 5 intersecting genes to exist between all of the datasets, but these weren't necessarily the topmost DEGs. This lack of intersection persisted even after applying combat to reduce the variance introduced by batch effect.

Using the post ComBat merged (regardless of sample) matrix and comparing all controls to all COPD led to certain significantly differentially expressed genes, however these genes are fairly hypothetically differentially expressed. I decided to come up with a value that emphasized the upregulation of the genes in the merged matrix and the intersection or common upregulation of genes between individual samples:

$$\sum_{merge} \div \sum_{intersect}$$

Figure 2. Equation for characterizing expression matrices: An intersect is the score given to a gene that corresponds to an intersection of upregulation, merge corresponds to a score given to a gene if it is upregulated between the merged(all samples control matrix) and the merged(all samples copd matrix)

The above equation would apply to individual samples in attempts at taking into account the overall differential expression. The goal of creating a summary expression was that adding the effect of the merged dataset would potentially make up for the missing expression information associated with COPD.

```
#Regulation_Merged(logFC sign of DEGs (ignoring sample))
#Regulation_Intersect(logFC sign of DEGs in common between samples)
sampleFactor
numeratorlist
denominator
For gene in sample:
    numerator times Regulation_Merged[gene]
    denominator times Regulation_Intersect[gene]

sampleFactor=numerator divided by denominator
```

Figure 3. Pseudocode example of Calculation: Regulation_Intersect and Regulation_Merged are dictionaries containing multipliers for of the associated genes depending on their logFC sign(- downregulation or + upregulation) for that gene in the dictionary

The work to characterize the microarrays is shown in the jupyter notebook: postcompat.ipynb. The section *Getting Expression Component Of Model Input* shows the methods taken to calculate this Expression factor that would represent the expression set. The section *Training Predictive model* shows the attempt at predicting the presence or absence of COPD using this expression factor. First the data were encoded to 0 or 1 based on whether or not they matched up with the .pheno files. Then an SGD classifier which has the ability to learn from deviation from the correct value was used. Unfortunately both methods resulted in 0 percent accuracy when predicting for COPD.

Strategy 2.: Explaining important genes and samples using Gene Ontology

A Jupyter notebook toptable2degList.ipynb(in project combineExpress directory) was used to convert R differential expression design matrix to a list of top 100 DEGs suitable for upload to grofiler. The top 100 DEGs were taken for multiple eBayes comparisons. By looking at the heatmaps for gene regulation between Males and Females with COPD I discovered that there was in fact a significant intersection between the DEGs in males and females who have covid. Even within the top 20 DEGs it is visible that they are fairly consistent across the studies:

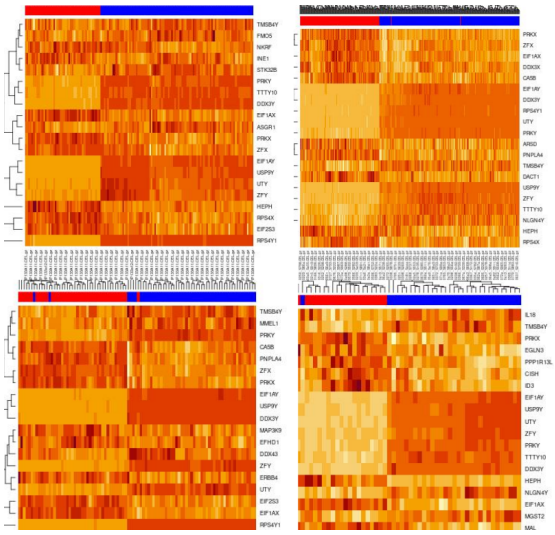
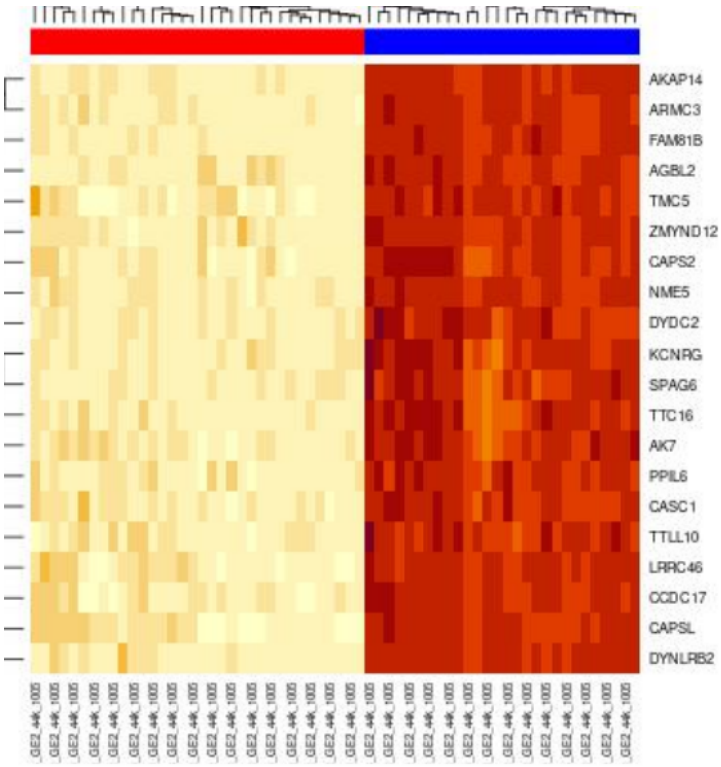


Figure 4. Heat maps across four samples for individuals with COPD(refined by gender): The red block on top represents the female subjects with copd and the blue represents the male subjects

Inputting the top 100 DEGs into g:profiler at 'https://biit.cs.ut.ee/gprofiler/gost' revealed enrichment in the different sets:
GSE71220: showed impaired B cell generation and proliferation and most significantly decreased circulation of IgE.
GSE54837: showed a most significant from the REAC database to be implicated in formation of ternary complex and 43s(important for eukaryotic translation initiation), but was also found in lung cell macrophages expression.
GSE42057: also had genes associated with ternary complex and 43s formation as with the previous sample. It was most enriched for translational initiation.
E-MTAB-5278: was found by the reac database to be involved in dissolution of fibrin clot and lymphotoxin beta receptor complex, but most significantly as molecular function regulator by go:mf

Putting the common 8 DEGs between the studies that are visibly obvious from the heat map: PRKY, TTTY10, DDX3Y, EIF1AY, USP9Y, UTY, ZFY, RPS4Y1 into gprofiler results in : oligospermia and azoospermia pathways which are infertility issues in men. If the heatmap is correct, it is indicating an upregulation in males. Perhaps this is due to the fact that women don't possess sperm cells that there is such a difference, but perhaps the cause of the copd is also drastically affecting their fertility.



Five of these genes also intersect with the cancer dataset's top 100 DEGs: MORN2, BAIAP3, NBEA, LRRC56, TJP3, RIBC1.

Uploading the top 100 DEGs for Cancer resulted in different pathways than characterized by the top 100 DEGs for the intersection between Men,Women and COPD, if anything there were many more disease phenotypes characterized by the HP(Human Phenotype Ontology) category of the pathways including infertility in both males and females .

Conclusion:

Had I found a study dataset(GSE71220,GSE54837,etc.) intersection between COPD and women and the men in each study I wouldn't have tried to create the predictor I did for COPD. It would have been better to just stick with refining the datasets, which would allow better focus on the upregulation and downregulation of the specific genes. I made an incorrect assumption that all of the collected genes from the COPD patients would somehow play a role in causing the COPD. I also realized only using one feature to train a model is a lot less effective than using multiple features. The scripts I have implemented and heatmaps are in this drive folder(projectScript.r is the main R script and heatmaps having the .heatmap extension): https://drive.google.com/drive/folders/1TXNkprnRONCbxRq46NJ3x3OtFDqR9KIv

Works Cited:

Rogers, Lavida R. K., et al. “Gene Expression Microarray Public Dataset Reanalysis in Chronic Obstructive Pulmonary Disease.” *PLOS ONE*, Public Library of Science, journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0224750.

Zhang L;Chen J;Yang H;Pan C;Li H;Luo Y;Cheng T; “Multiple Microarray Analyses Identify Key Genes Associated with the Development of Non-Small Cell Lung Cancer from Chronic Obstructive Pulmonary Disease.” *Journal of Cancer*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/33442399/.