Gene expression and DNA methylation signatures in whole blood to distinguish risk for IPF

Background

Idiopathic pulmonary fibrosis (IPF) is a scarring disease of the lungs with an unknown cause. IPF affects 5-million individuals world-wide and is inexplicably increasing in prevalence (Evans *et al.* 2016), but it is still likely underdiagnosed. IPF is characterized by thickened and scarred insterstitium which negatively impacts gas exchange. Many studies have contributed to the understanding of the disease such that many consider "idiopathic" to be a misnomer. Many risk factors have been associated with IPF, including genetic and environmental factors. The basic pathobiology of IPF involves recurrent epithelial injury which leads to aberrant repair programs that results in irreversible fibrosis. This replacement of healthy lung tissue with dense collagen leads to respiratory failure. The disease takes years to develop, but most patients are not diagnosed until the late stage. Affected individuals have a median prognosis of 3-5 years upon diagnosis. There are currently no curative treatments, and current therapies target symptom alleviation and quality of life improvement. There is a distinct need to be able to consistently identify risk for developing IPF earlier, in order to provide the potential to treat the lung damage and extend the prognosis.

The CADET population is a large population including individuals with IPF, their relatives at risk for preclinical pulmonary fibrosis (PrePF), and unaffected, control individuals. This population has been used to assess the prevalence and risk factors for PrePF. From the whole CADET population, 48 participants with established IPF, 48 with PrePF, and 96 controls were selected for DNA methylation and RNA-sequencing analysis from whole blood samples. All individuals are from unique families. This data will be used to produce models to predict PrePF from control individuals, with the goal of identifying biomarkers to detect risk for IPF earlier.

Methods

RNA-sequencing differential testing

The RNA paired-end reads have been aligned at the transcript level to the human genome. Independent filtering was performed using DESeq2 to filter out transcripts from differential expression analysis with low expression. The data was normalized using Variance Stabilizing Transformation (VST). Principal component analysis (PCA) was then used to identify and remove outliers more than two and a half standard deviations from the mean of the first two principal components. The normalized read counts were used to perform differential testing using DESeq2, adjusting for age, sex and RNA-sequencing plate.

DNA methylation differential testing

The DNA methylation beta values were first processed to remove problematic and missing data. The beta values were then transformed into M-values to better approximate a normal distribution, which allows application of parametric modeling techniques. PCA was then used on the M-values to identify outliers and patterns in the data. The CpGs from sex chromosomes were then removed to eliminate stratification of the data by sex. Differential testing was then performed using limma, adjusting for age, sex, and position on the array. The data was then also run in GLINT to reduce inflation and look for significant age associations. Finally, prior to modelling, known age-associated probes were removed based on literature (Horvath *et al.*).

Machine learning model training and evaluation

The modeling was performed in R using the packages caret and XGBoost. Predictive models were trained for both RNA-sequencing data as well as the DNA methylation data. The data was separated into controls vs PrePF, as well as the same controls vs IPF. Although the PrePF vs control models were the primary focus, IPF vs control models were also trained to ensure the same models were performing well, if not better, since the stratification between IPF and controls should be larger than that of PrePF vs control given the disease pathobiology. The features were sorted by P-value from differential testing, and subsets were selected for the top 1000, 100, 50, 15, 10, and 5 features. The datasets were then split 70:30 into training and test datasets, respectively. The training data was then used to train models. The models used were Classification and Regression Trees (CART), glmnet (generalized linear model with an elastic net), Support Vector Machines (with a linear kernel and a radial based kernel), Random Forest, and XGBoost. The best tuned model was then tested by inputting the test data, and examining the predicted results compared to the true diagnoses. Within each dataset (RNA-sequencing or DNA methylation for IPF vs control or PrePF vs control) the models with the best predicted performance were selected.

These best models were then tested for robustness by slightly altering the hyperparameters up or downward. Additionally, IPF diagnoses are confounded with age and sex, with the number of cases being higher in men and older aged individuals. To determine whether these demographics affected the best performing models, age and sex were added to the data as features both individually and together. The best tuned model with age, sex, or age and sex were then tested on the test data and predicted performance examined.

Results

RNA-sequencing differential testing

Eight PCA outlier samples were removed which included one IPF sample, five PrePF samples, and two controls. Significantly more inflation was seen in the IPF vs control testing than the PrePF vs control testing. There was a strong differential testing signal for IPF vs controls. There was a weaker signal for PrePF vs controls, but there was a good distribution of P-values.

DNA methylation differential testing

In the PCA for the DNA methylation data, the first two PC's were stratified by sex, due to CpGs on the sex chromosomes. Once these CpGs were removed, the data were no longer stratified by sex. Of important note, the PC's were also separated by position on the array. This is a common problem in DNA methylation array data, and was included in the differential testing. Again significantly more inflation was seen in the IPF vs control data. Using limma, there was a strong signal for IPF vs control and a weak signal for PrePF vs control. After running the data in GLINT, the inflation was reduced.

Machine learning model training and evaluation

Many models with strong predictive abilities were obtained from the modeling (Figure 1). When first examining the model performance and receiver operating characteristic (ROC) curves, the algorithms were narrowed. CART models did not perform well and were removed. Additionally, SVM Linear models did not perform well on RNA-sequencing data. Random Forest models did not perform very well nor consistently. XGBoost models performed well, especially on IPF vs control data; however, when the feature importance was examined, it was observed that XGBoost models prioritized almost completely unique features from the best performing caret models, so these models were also excluded. This left glmnet and SVM Radial for RNA-sequencing data and glmnet, SVM Radial, and SVM Linear for DNA methylation data. The common machine learning tradeoff was observed between number of features and predictive capability, such that models with more features perform better. A large number of features often makes interpreting biological relevance of features more challenging. To balance this predicament, models using 100 features were selected. Models were achieved with overall accuracy above 90% and balanced accuracy about 80%.

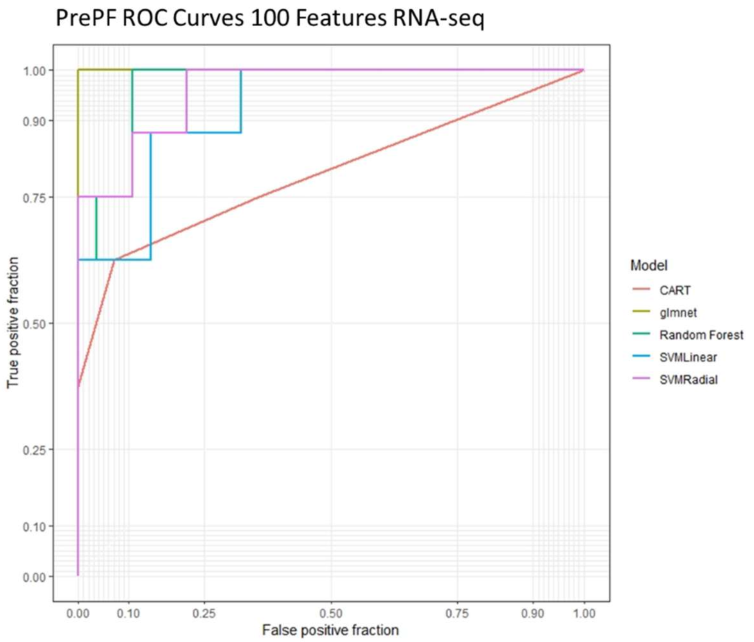

Figure 1) the ROC curves for each of the shown models trained on the top differentially expressed RNA-sequencing features. CART and SVM Linear did not perform well on RNA-sequencing. Random Forest also did not perform well. Similar patterns were seen for methylation data, with glmnet, SVMR and SVML all performing well with high balanced accuracies.

The best models and the associated hyperparameters were ascertained. When the hyperparameters were adjusted up or down, there was minimal effect on the predictive performance of the models. This helped show that these best models were relatively robust. Age and sex demographics were then also added into these models. Adding in age and sex individually and together did not change the predictive performance of the models. When the feature importance plots were examined, age was sometimes a feature of high importance in the model. This is not surprising, as it is known that age and diagnosis are highly confounded. However, since it did not change the predictive performance of the model, we feel comfortable in keeping the models agnostic of demographics.

Discussion

IPF is a heterogeneous disease with a short 3-5 year prognosis upon diagnosis. The ability to detect risk for development of IPF before irreversible lung damage and respiratory failure is necessary. The CADET population designed to look for PrePF risks is ideal for building predictive models. These models enable classification of PrePF individuals from DNA methylation or RNA-sequencing data obtained from blood samples. These models can also be used to identify potential biomarkers which could be used for diagnostic purposes in the clinic. Using blood samples would provide an easier clinical tools than lung samples as they are much easier to obtain and therefore ideal for this study. An earlier detection of PrePF would allow the opportunity to treat the lung fibrosis before it became irreversible and ideally extend the prognosis.

This project successfully developed multiple predictive models that could accurately classify PrePF cases from controls. These models were robust and not heavily influenced by demographic confounders (age or sex). From these models a few genes and DNA methylation sites were identified as potential biomarkers for IPF risk. From the RNA-seq models, a few highly important features were found to have IPF relevance. For example, SPEG is a gene that is required for myocyte cytoskeletal development, and myofibroblasts are characteristic of IPF. Additionally, PCSK5 was a top feature in our predictive models that has been shown to have 7-fold gene expression increase in pulmonary hypertension pericytes and smooth muscle cells. From the methylation models a quite a few IPF relevant CpG sites were found to be highly predictive as well. For example the site cg12961080 lies within an active enhancer within the BCL11B first intron, which is a gene shown to be differentially expressed in IPF lung tissue.

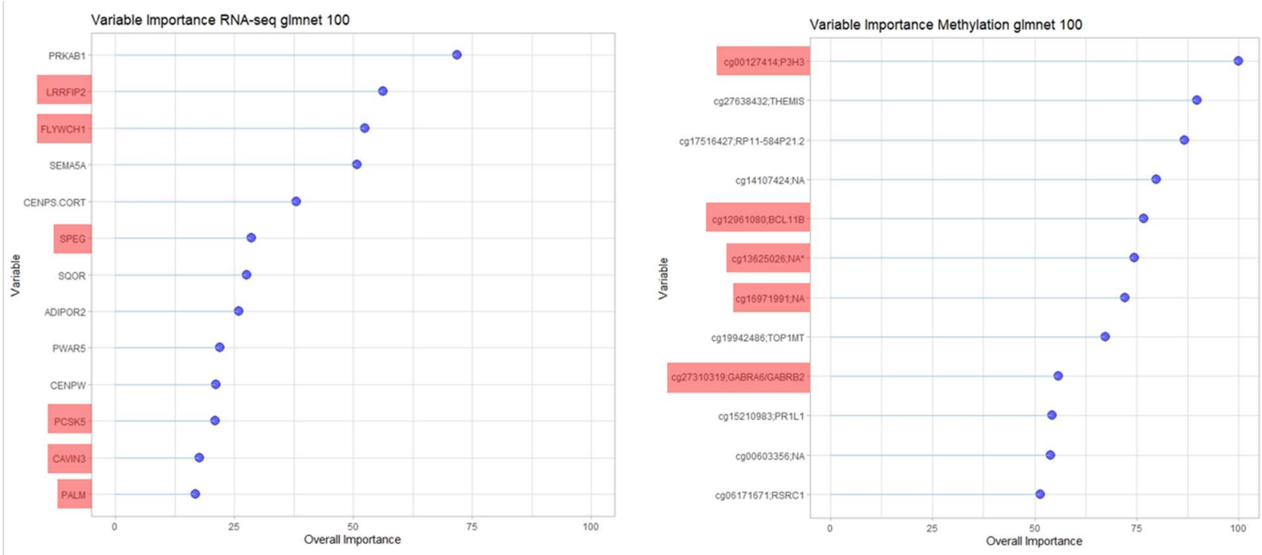

Figure 2) Feature importance plots for RNA-sequencing (left) and DNA methylation (right) with IPF relevant genes or sites highlighted.

Additionally two sites, cg13625026 and cg1697199 are just upstream of MUC5B, within an ENCODE predicted enhancer. MUC5B is highly overexpressed in IPF patients, and contains a common variant that contributes a significant amount of risk for IPF.

This study is limited in statistical power by the number of samples. However, the model robustness was tested in multiple ways to try to ensure model performance was not limited to this specific data set. Additionally, models were selected that were not as prone to overfitting, to maximize chances of generalizability.

In the future, more samples should be obtained from individuals with IPF and PrePF to create a second, larger dataset. This dataset could then be used to test the models created within this study. Additionally, there is a high possibility that features within the RNA-sequencing or DNA methylation datasets are colinearly expressed, thus confounding the models. the R packet caret has methods that could be utilized to remove collinearity from the features. This could reduce the feature space and make the models easier to interpret through a pathobiological lens.

In all, this rotation project allowed me to vastly expand my analytical skills. I analyzed DNA methylation data for the first time, and performed differential testing on the cleaned dataset. This gave me insight into the patterns of methylation data and the tools needed to analyze it. I also gained significant experience with machine learning. My prior experience with machine learning was limited to writing algorithms to perform the machine learning. This project provided me with the opportunity to apply different algorithms to real data. I learned how to evaluate the performance of different models and how to tune them. I also was able to learn about the advantages and disadvantages of the different algorithms. This project was an excellent learning opportunity and could be very useful in the clinic as a stepping stone toward a new diagnostic tool.