Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

______________________________ ______________

Ishaan Dave Date

Comprehensive Analysis Examining Metabolites Associated with Severity of Cystic Fibrosis

By

Ishaan Dave
Master of Science in Public Health

Biostatistics and Bioinformatics

Limin Peng, Ph.D.

Committee Chair

Rabindra Tirouvanziam, Ph.D.

Reader

Comprehensive Analysis Examining Metabolites Associated with Severity of Cystic Fibrosis

By

Ishaan Dave

B.A.

Emory University

2016

Thesis Committee Chair: Limin Peng, Ph.D.

Reader: Rabindra Tirouvanziam, Ph.D.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2017

Abstract

Comprehensive Analysis Examining Metabolites Associated with Cystic Fibrosis Severity

By: Ishaan Dave

Introduction: Cystic fibrosis (CF) is an inherited monogenetic disease caused by mutations in the CF transmembrane conductance regulator (CFTR) gene that affects approximately 80,000 people across the world. Altering the gene leads to improper epithelial secretions and multi-organ dysfunction. Discerning changes in airway fluid of CF infants to curb disease is of great interest. The goal of this thesis is to perform a metabolomics analysis on bronchoalveolar lavage fluid (BALF) obtained from an 11 patient cohort aged 35-38 months using untargeted mass spectrometry (MS) profiling.

Methods: Several techniques were employed to analyze data generated by MS of BALF to determine metabolites most significantly associated with severity of CF airway damage, as measured by the sensitive PRAGMA scoring tool for computed tomography images acquired concomitantly with BALF collection. MS data yielded 2,591 features in CF BALF, and Spearman and Pearson correlations for each feature with PRAGMA scores were calculated. Features with a significant Spearman correlation were run through *mummichog* (network analysis tool) to identify pathways in which these metabolites are involved. Penalized regression selected important metabolites among those with significant Pearson correlations to build a parsimonious predictive model for PRAGMA. Finally, a random forest algorithm was run on all the features to identify those most important in predicting PRAGMA.

Results: We identified 105 and 101 features significantly correlated with PRAGMA score using Spearman and Pearson methods, respectively. The 105 Spearman "hits" run through *mummichog* identified several amino acid metabolism pathways, including that of tryptophan, featuring formyl-N-acetyl-5-methoxykynurenamine (AFMK). The random forest algorithm identified 3 important features – GlcCer(d14:1(4E)/20:0(2OH)), tetrahydrocorticosterone, and AFMK. From the penalized regression utilizing the 101 Pearson hits, 11 metabolites were selected to build a predictive model for the PRAGMA score. Among them, nonate and PGF2 alpha-dihydroxypropanylamine, are particularly scientifically insightful.

Discussion: Random forest and *mummichog* identified AFMK – a metabolite of the tryptophan metabolism pathway implicated in the maintenance of mucosal integrity. Penalized regression identified nonate, a succinate derivative – important in the electron transfer chain of mitochondria – and a PGF2 alpha derivative, putatively linked to inflammatory signaling. These results bring novel insights into mechanisms underlying airway disease development in CF infants.

Comprehensive Analysis Examining Metabolites Associated with Severity of Cystic Fibrosis

By

Ishaan Dave

B.A.

Emory University

2016

Thesis Committee Chair: Limin Peng, Ph.D.

Reader: Rabindra Tirouvanziam, Ph.D.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2017

Acknowledgements

I would like to thank the department of Biostatistics and Bioinformatics at Emory University for their support and guidance during the past 2 years. I would like to especially thank Limin Peng for keeping an open door willing to be my advisor for my thesis and for the help and feedback she gave me on my thesis. In addition, I am grateful to Joshua Chandler for unparalleled support and patience in overcoming obstacles I faced through research, including any questions that arose regarding data, analysis, and interpretation of results. Thank you to Rabindra Tirouvanziam for taking the time to be the reader of this thesis and I am gratefully indebted to his valuable comments and feedback on this thesis. Lastly, I'd like to thank my friends, my parents and my brother for their love and encouragement throughout writing this thesis.

Contents

Introduction

Cystic fibrosis is an inherited monogenetic disease caused by a mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. CF affects approximately 30,000 patients in the United States and more than 80,000 people across the world, with about 2,500 babies born with CF in the United States every year. CFTR gene mutations lead to abnormal folding, expression, translocation or function of the CFTR protein and consequently, normal function of this protein is altered. Under normal circumstances, the CFTR protein acts as a channel for chloride and other low molecular mass anions to pass the epithelial cell membrane, but altering the gene leads to improper function of this channel resulting in abnormal hydration and/or acidification and/or redox properties of epithelial secretions (Muhlebach, 2015). In particular, chloride transport is essential in maintaining a properly hydrated epithelial surface. If the transport of chloride ions is defective, airways and ducts in other organs are clogged or blocked entirely (Wetmore, 2010). For example, CF airways are prone to high mucus viscosity, which increases the propensity to infection and inflammation. Similarly in the gastrointestinal system, CF leads to blockade of the pancreatic duct. Consequently, pancreatic enzymes are not able to travel into the intestine. Pancreatic enzymes normally aid in the breakdown of food, and without them, the body is unable to fully absorb fats and proteins – leading to nutrient deficiency and malnutrition (Muhlebach, 2015). In addition to severe airway and digestive disease manifestations directly linked to abnormal mucus function, several other complications can arise in CF patients, including high sweat osmolarity and propensity to heat exhaustion, sinusitis, nasal polyps, abdominal pain, liver disease, diabetes, and pancreatic inflammation, among others (CDC, 2017).

Problem Statement

The purpose of this thesis is to perform a metabolomics analysis on bronchoalveolar lavage fluid (BALF) obtained from a cohort of 11 CF infants aged 35-38 months old, using untargeted mass spectrometry profiling. Using a multi-pronged biostatistical approach, the analyses presented are used to determine features significantly associated with disease severity based on scoring of chest computed tomography (CT) images with the PRAGMA-total disease (PRAGMA-TD) method. A recent published study (Esther Jr., 2015) attempted to identify metabolites associated with early CF disease development, providing proof-of-concept for our effort. A comprehensive biostatistical approach will be used determine what metabolites and combination of features affect PRAGMA scores. Identifying specific metabolites that are clinically and/or biochemically relevant and correlated with the PRAGMA-TD score will allow researchers and clinicians to better monitor disease progression and design new interventions.

BACKGROUND / LITERATURE REVIEW

The need for early detection and intervention in CF

In 1985, the median survival age of cystic fibrosis patients was 25 years. As of 2007, median life expectancy of CF increased to 37.5 years. This 50% increase in survival age is indeed dramatic, but there is always need for additional new therapies to further increase the survival age of CF patients (Wetmore, 2010). With CF typically being diagnosed in children, as patients with the condition age, male and female adults experience challenges not seen during childhood. For example, 98% of CF males of reproductive age are infertile due to improper development of the *vas deferens* (CF Trust). In females, puberty is delayed approximately 1 to 2 years, and this is coupled with a higher probability of missed period. For CF adults, it is difficult to get adequate energy through natural feeding, so supplementary nutrition is often necessary, in addition to changing medication regimens as they transition from pediatric to adult care (CF Trust). Moreover, transitioning into a “normal life” is difficult because they are hindered by a lack of independence, which in turn, affects forming relationships in the office, at college, or with a partner (CF Trust). Lastly, the symptoms of patients with CF become more severe with age, which makes normal life increasingly difficult (CF Trust).

Early detection and new therapies are vital for attaining better outcomes for patients with CF. Recent scientific advances have led to a better understanding of the 1,800 or so disease-causing mutations of the CFTR gene and their effects. These advances have significantly improved the outlook for patients with CF. Newborns are screened in many countries for CFTR mutations. The lungs of children with CF all appear anatomically normal at birth. However, the signs and symptom of CF become apparent very early in life.

If early disease is mild, then typical symptoms of CF do not appear until later in life. However, if early disease is severe, then symptoms are apparent very early in life. For this thesis, data were collected from infants aged 35-38 months. To measure the severity of CF airway disease, the PRAGMA ("Perth-Rotterdam Annotated Grid Morphometry Analysis") score was computed on chest CT images, reflecting structural airway damage (Rosenow, 2014). We hypothesized that specific metabolites would be significantly increased or decreased as CF severity increased in infants.

Metabolomics Studies of Early CF Disease

A new and innovative approach to characterize all of the metabolites from samples (e.g., bronchoalveolar lavage fluid – BALF) is metabolomics. Metabolomics can be used in studies geared at toxicology, disease diagnosis, and biomarker identification by profiling the metabolites in small volumes of bodily secretions or tissue samples. In particular, the power of metabolomics methods to detect biomarkers has enabled significant advances in disease monitoring and therapeutic approaches. Biological samples contain a multitude of lipids, carbohydrates, and proteins. Untargeted metabolomics is used to profile these molecules and assess pathways, compare patient groups, or gain knowledge about the pathologic processes involved. Among the most commonly used methods are nuclear magnetic resonance (NMR) and mass spectrometry (MS), which is generally more sensitive than NMR. Wolak et al. (2009) performed the first metabolomics analyses on BALF from cystic fibrosis patients in 2013 using NMR. Esther Jr. et al. went on to use MS and showed that the concentration of several metabolites increases or decreases with increased inflammation in the airways of CF infants. These metabolites were involved in cellular energy metabolism, protein catabolism, and lipid signaling (Esther Jr., 2015). Thus, processes involved in the early stages of CF alter the metabolism of cells and affect bioactive mediators of inflammation. Data included in this thesis have been acquired by BALF analysis using MS.

Metabolomics studies can be conducted as either targeted or untargeted approaches. Targeted metabolomics focuses on a defined set of metabolites and generally provides unequivocal molecular identification, while untargeted metabolomics identifies "features" which precise identification can constitute a significant challenge (Bartel, 2013). Compounds' masses found by the MS approach are usually compared to known metabolites on databases such as HMDB, LipidMaps, and METLIN (Bartel, 2013). Network analysis of metabolomics data can be utilized to explore complex metabolite-metabolite interactions and metabolite-gene/transcript/protein interactions. The use of metabolomics analytical technologies under defined conditions have helped illuminate changes in pathways. MS has become the method of choice for such efforts, due to its range of metabolites it can cover as well as its reproducibility in analysis (Zhang, 2012). To fully leverage MS data, multiple complementary techniques are necessary to thoroughly analyze all the properties and molecular contents in a collection of samples.

Review of Statistical Methods for Metabolomics Data

In the early days of metabolomics, analytical approaches focused on a limited number of metabolites, so the results could easily be navigated by investigators. But the rapid advancement of research in the field and evolution of modern technology allowed researchers to analyze thousands of metabolites simultaneously. However, the widened capabilities of metabolomic technology complicated findings, which made it exponentially more difficult to give meaningful results that are interpretable clinically and biologically.

To analyze these increasingly complex data, concurrent use of several robust statistical methods is required so as to examine complex interactions between thousands of analytes usually present in biofluids (Bartel, 2013).

Currently, the two main techniques for the analysis of high-throughput metabolomics data are a) network modeling based on Gaussian graphical models (GGMs) and b) higher order correlation analysis denoted as independent component analysis (ICA). Classical approaches to analyzing "*-omics*" data aimed to find differences in groups by either using univariate methods in a parameter-to-parameter manner (t-tests or analysis of variance -ANOVA) or multivariate methods (multivariate ANOVA – MANOVA, principal component analysis – PCA, or partial least squares regression – PLS – among others). The aim of univariate methods is to reduce the number of analytes measured to only those that showed a strong response to the conditions researched. In this thesis, over 2,500 analytes were measured, and the response was the CF disease severity index represented by the PRAGMA score. However, univariate methods fail to differentiate between groups if only minor differences are present. The goals of multivariate methods are both to capture changes of a single metabolite between groups while assessing the dependency structures between metabolites. The most commonly used multivariate methods in metabolomics are PCA and PLS (Bartel, 2013).

A commonly used way to interpret the highly intertwined pathways of metabolites is by using a graph or network. Networks are usually made up of nodes that are linked to each other, with each node typically representing a molecule, or in this study, a metabolite. Several databases exist whose focus is to reconstruct these pathways (KEGG, Human Recon I, and EHMN) and the networks contained in these databases can be used as a guide to design appropriate statistical analysis approaches. Pathway databases are far from complete, however. One proposed method is to reconstruct the pathways straight from the data. To that end, statistical methods “exploit the naturally occurring biological variation in the abundance of metabolites between biological replicates” (Bartel, 2013). This variation could come from temperature or pH fluctuations (extrinsic factors) or enzyme level changes linked to different regulatory states (intrinsic factors). Bayesian networks are the most commonly used method for reconstructing pathways. Bayesian networks show random variables (metabolite concentrations) as nodes and dependencies as edges.

Gaussian graphical models (GGMs) are based on partial correlation coefficients and eliminate interactions by taking the pairwise correlations and conditioning them against every other variable in the data. GGMs have recently gained traction in metabolomics data, but the calculation of full-order partial correlation usually requires a higher number of samples than variables. Owing to the fact that untargeted metabolomics can generate thousands of measurable variables, this is rarely attainable. Bartel et al. (2013) recently used GGMs in a metabolomics study to show that using this approach effectively identified metabolic reactions from human plasma metabolomics data. Comparing the output from this analysis to existing pathway databases, these authors found that high partial correlation coefficients correspond to known metabolic reactions, as well as other candidates for pathway interactions. While GGMs are powerful methods able to graphically depict metabolite pathways and interactions, they are limited by covariance between variables. Indeed, higher-order dependencies are ignored by GGMs.

Independent component analysis (ICA) is a method that is mathematically similar to PCA, but with the capability to capture higher-order dependencies. ICA was initially used in the neurobiological field, then transitioned to transcriptomics data, and most recently has been applied to metabolomics data. ICA decomposes the data matrix of measured metabolomics profiles into *k* statistically independent components (ICs). Biologically, these ICs all contribute to the overall metabolic profile. However, determining how many ICs there exist is a major challenge. In addition, no prior information can be incorporated into each independent component. Therefore, the Bayesian ICA approach was devised to tackle both of these problems. With this approach, the Bayesian information criterion can be used to determine the optimal number of independent components. To show that the ICA approach was more reliable than a basic PCA, Bartel et al. (2013) analyzed metabolomics data from a cohort of 1,764 patients and 218 measured metabolites and showed that ICA outperformed PCA in providing more biologically sound data decomposition. In addition, ICA identified distinct metabolic pathways, while PCA showed inconsistent distribution of metabolites among pathways.

METHODS

PRAGMA Data Collection

Subjects in this study were children with CF (n=11) followed at Erasmus University Rotterdam (EUR) in the Netherlands. Children had their chest CT scored according to the PRAGMA method by EUR researchers. In brief, to compute the PRAGMA score, a square grid was overlaid on ten slices equidistant of the chest CT from each other. Each cell in the grid was then annotated according to whether or not it featured bronchiectasis (Bx), generating a specific score. Bx and other anomalies (atelectasis, trapped air) were added toward calculation of a total disease (TD) score, also expressed as a percentage.

Metabolomics Data Collection

The process of obtaining the metabolomics profile consists of several steps. First, the BALF collected from the CF infants in the study is mixed with a solution of acetonitrile, vortexed for proper mixing, incubated on ice for 30 minutes and centrifuged at 16,000 G for 10 minutes at 4ºC. The resulting product has two distinct layers. On top is a liquid layer, which contains small molecules which are analyzed for metabolomics. On the bottom is a solid layer which contains large molecular aggregates of lipids, DNA, and proteins. A small amount of the liquid layer (5 µl) is placed in a vial into the mass spectrometer (Orbitrap). First, the liquid is run through a column for separation by liquid chromatography. The hydrophilic interaction liquid chromatography (HILIC) column that the solution runs through is polar, so polar compounds in the solution are retained longer, while non-polar compounds come off the column quicker. As the sample runs through the column, a mobile phase is mixed into the liquid at different ratios depending on how the user wants metabolites to elute. While the sample runs down the column, the mobile phase/supernatant mixture contacts a positively charged needle. This needle will either activate an electrospray that produces ions, or remain sitting at a high voltage. After the electrospray is activated, the liquid is converted to as gas, which causes particles to fly around the chamber until they are caught by the ion trap. This ion trap can be calibrated to a specific range of mass to charge ratio (m/z) values to capture and retain analytes. The ions formed are variations of a given mass (M), created by the generation of adducts ($M+H-H_20$, $[M+H]^+$, or $M+Na^+$), and the m/z is derived from the adducts and resulting charges. Graphs are generated for all the individual ions (“features”) measured by the MS, in which areas under each peak are proportional to the amount of metabolites present in the sample.

Pathway Analysis

PRAGMA scores were first assessed for potential correlations with all the features recorded by MS. A Spearman correlation test was used because ranked values of each feature were considered to be more important than the raw data, and there was potential for multimodality between the intensities and the PRAGMA scores. Those features that yielded *p*-values less than 0.05 were used in further analyses. Then, these features were run through a Python script called *mummichog* (Li, 2013) to perform a pathway analysis, in order to identify the pathways in which the identified features are most important.

Statistical Data Mining

A random forest algorithm was run on all MS features to identify those that were the most important in predicting PRAGMA score. A random forest is a data mining method for regression or classification that constructs several decision trees and outputs the mean prediction of the outcome (Ho, 1995). The random forest method reduces the chance of over-fitting a model to its covariates (Ho, 1995). The *randomForest* package in R was used to perform this analysis. To form the forest, 500 individual decision trees were formed. The nodes, or splits, in the decision trees tested 863 variables each time, and those that repeatedly were selected and seemed to reduce the variance of the model were considered most important.

Lasso-penalized linear regression was used to model PRAGMA scores based on features whose Pearson correlation with the PRAGMA score was statistically significant at the 0.05 significance level. Lasso regression is regression analysis that performs variable selection to optimize the accuracy in predicting the outcome (Tibshirani, 1996). The method was implemented using the *GLMNET* package in R. All features were normalized to have a mean of 0 and a standard deviation of 1. Specifically, the sample mean of each variable is subtracted from each observation of the variable, and the resulting value is divided by the sample standard deviation of the variable. This normalization process enables GLMNET to run efficiently and effectively. The GLMNET algorithm produces shrinkage estimation of high-dimensional coefficients. Cross-validation was employed to determine the tuning parameter. The set of selected variables include those with non-zero coefficients, representing a group of features that can jointly predict the outcome of interest – in this case, the PRAGMA score. Statistical data mining techniques were conducted using R software and R studio (R Core, 2015), and pathway analysis was conducted using *mummichog* (Li et al., 2013). The significance level was set at $\alpha = 0.05$.

RESULTS

The average age of the 11 CF infants in the study was 37.2 ± 0.6 months, and of these subjects, five were male and six were females. After the BALF of these children was run through MS, there were 2,591 m/z features that were detected and included for further analysis. The initial analysis included the correlation analysis of every feature against the PRAGMA scores of the 11 children. Using Spearman correlation, 105 out of the 2,591 features were statistically significant at 0.05 significance level. These 105 features were then run through *mummichog*. From the inputted features, *mummichog* output predicted metabolites of different adducts, the activity network of these metabolites, and the main pathways in which they are involved (**Table 1)**. The main pathways identified correspond to amino acid metabolic pathways.

Table 1 – Main pathways in which predicted metabolites are involved in.

Pathway	Overlap Size	Pathway Size	*p*-value (raw)	*p*-value (adjusted)
Tryptophan	3	15	.01857	.00535
Arginine and Proline	2	16	.13209	.02151
Aspartate and Asparagine	2	34	.40582	.07346
Tyrosine	2	47	.5845	.13138

After running the Spearman correlation analysis and *mummichog*, modeling techniques were used to identify the features with the most impact on PRAGMA scores. First, a random forest algorithm was run to determine the most important features. After tuning the algorithm, it was determined that using 1,762 variables at each iteration of the random forest minimized the error in the model. After running the random forest with 500 trees, there were 3 variables that were noticeably more important than the rest (**Figure 1).** The m/z values for the resulting 3 variables and the predicted metabolites with their adducts (found using the METLIN database) are indicated in the table below (**Table 2).**

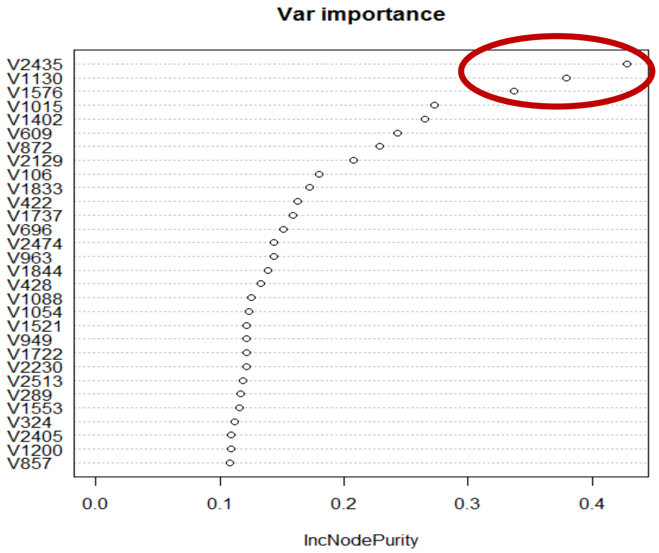

Figure 1 – Variable importance plot from Random Forest algorithm
Table 2 – m/z values of the 3 most important predicted metabolites (random forest)

m/z value	Predicted Metabolite (Adduct)
754.5273	GlcCer(d14:1(4E)/20:0(2OH)) (M+K)
265.1178	Formyl-N-acetyl-5-methoxykynurenamine (M+H)
351.2525	Tetrahydrocorticosterone (M+H)

Among the 3 predicted metabolites listed above, only one, formyl-N-acetyl-5-methoxykynurenamine (AFMK), overlapped with the pathway analysis generated by *mummichog*. This metabolite belongs to the tryptophan metabolism pathway. When examining the scatter plot with the intensities associated with AFMK against the PRAGMA scores, it is important to note that there does not seem to be any linear relationship between the two. However, since the random forest algorithm selects variables to be used together, the intensities for AFMK coupled with those of GlcCer(d14:1(4E)/20:0(2OH)) and tetrahydrocorticosterone provided prediction among all the features. The mean absolute prediction error from the random forest calculated by the *predict()* function in R was 0.462 (Liaw, 2002). Considering that PRAGMA scores ranged from 1% to 6%, this estimation error of 0.462 is very low, and the features selected by random forest can provide quite accurate predictions of PRAGMA scores based on the identified random decision tree.

The penalized linear regression aimed to create a model to predict the severity of CF with the features selected by a significant Pearson correlation. Of the 2,591 features in the dataset, 101 had a correlation whose p-value was less than 0.05 using this method. The Lasso penalized regression selected 11 out of these 101 features that were marginally correlated with PRAGMA at the level of 0.05. Then, we used the METLIN database to search the m/z values selected by *GLMNET*, yielding two features that seemed especially relevant in CF. The first one, with an m/z value of 189.1118, is the $[M+H]^+$ adduct of nonate – a derivative of succinic acid. The second was the $[M+H]^+$ adduct of PGF2 alpha-dihydroxypropanylamine (Smith, 2005). PGF2 alpha-dihydroxypropanylamine had a relatively strong positive linear association with PRAGMA scores (**Figure 2**).

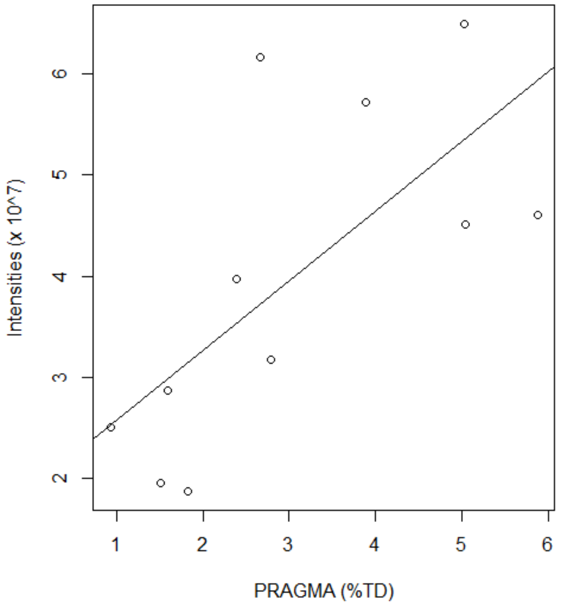

Figure 2 – PRAGMA score plotted against the intensities of PGF2 alpha dihydroxypropanylamine, showing relatively strong positive association (Pearson correlation = .692).

Given the coefficients and the data from the features, the estimation error of the penalized regression can be found the same way as that from random forest – summing the absolute value of difference in observed and predicted values across observations. The overall prediction error of the PRAGMA scores for the penalized regression is calculated by: $\sum_{i=1}^{n} |y_i - \hat{y}_i|$, where, n is the observation number, y_i is the observed PRAGMA score for patient i, and $\hat{y}_i$ is the predicted PRAGMA score for patient *i.* With a penalization parameter, λ = .045, and elastic-net mixing parameter α = 1, the estimation error using the coefficients given by the regression analysis was approximately 1.863. This shows that the *GLMNET* procedure was useful in predicting the PRAGMA scores of infants in this study.

DISCUSSION

The purpose of this thesis was to identify the metabolites that are statistically, clinically, and biologically pertinent in children with early CF airway disease. Specifically, those metabolites that significantly correlated to the severity of CF airway damage as measured by the newly developed, sensitive PRAGMA score computed on CT scans. Using a multi-pronged approach, the relevant metabolites were identified using a Spearman correlation analysis, a pathway analysis using *mummichog* using the statistically significant correlated features, a random forest algorithm to determine the most important variables in predicting PRAGMA scores, and a penalized regression to pick the variables most significantly associated with PRAGMA scores.

Results from the random forest and pathway analysis revealed overlap in the most important variables and identified several candidate metabolic pathways. The feature that was present in random forest and *mummichog* was the M+H adduct of AFMK, which belongs to the tryptophan metabolism pathway. AFMK has been suggested to work as a strong oxidant scavenger in neutrophil-rich inflammatory milieu (Silva, 2006), which is relevant to the neutrophil-rich, oxidatively stressed environment of CF airways. The penalized regression identified two particularly interesting features that are important in the human metabolome. The first was likely nonate – a derivative of succinic acid – whose anion, succinate, is able to donate electrons to the electron chain in the Krebs cycle. The second was likely PGF2 alpha-dihydroxypropanylamine, a derivative of PGF2 alpha. Collins et al. (1999) showed the levels of PGF2 alpha, a marker of oxidative stress were higher in patients with CF than in controls, using an ELISA method.

Results from this pilot study can inform future studies assessing the association between early CF disease severity and BALF metabolites identified by MS. The analyses performed in this thesis use an untargeted approach, but the results can be further validated by performing a targeted metabolomics analysis. Future larger studies will also need to assess the impact of key categorical variables on metabolomic profiles. For example, assessing the influence of sex and mutation type would be of great interest.

Limitations

The analysis presented here included data obtained from BALF of CF infants aged 35-38 months. The parent clinical study has only been ongoing for less than two years, so the small sample size of 11 limits the power of the statistics performed. In addition, this cohort of all children aged approximately 3 years old with cystic fibrosis is from a small region of the world and is not a full representation of children with CF. Infants in this cohort are also very young, which limits the clarity of the data because none of the children have lived long enough to develop severe symptoms. This is seen by the low PRAGMA scores observed, which all fall in the range of 1-6%. As we continue to recruit more patients in the parent study, we expect to generate more robustness to the statistical analyses conducted here.

References

1. Bartel, J., Krumsiek, J., Theis, F. Statistical Methods for the Analysis of High-Throughput Metabolomics Data. *Computational And Structural Biotechnology* (2013); Volume No: 4, Issue: 5, DOI: 10.5936/csbj.201301009

2. Collins, C., Quaggiotto, P., Wood, L., O'Loughlin, E., Henry, R., Garg, M. Elevated plasma levels of F2 alpha isoprostane in cystic fibrosis. Abstract. *Lipids* (1999). Jun; 34(6):551-6.

3. "Cystic Fibrosis." *Facts About.* CDC. NIH. NIH Publication No. 95-3650. November 1995.

4. Esther Jr., C., Coakley, R., Henderson, A., Zhou, Y., Wright, F., Boucher, R. Metabolomic Evaluation of Neutrophilic Airway Inflammation in Cystic Fibrosis. *Chest* (2015); 148(2):507-515. doi:10.1378/chest.14-1800

5. Esther Jr, C., Turkovic, L., Rosenow, T., Muhlebach, M., Boucher, R., Ranganathan, S., Stick, S. Metabolomic biomarkers predictive of early structural lung disease in cystic fibrosis. *European Respiratory Journal* (2016); 0: 1–10 | DOI: 10.1183/13993003.00524-2016

6. Harrell Jr., F., Dupont, C. et al. *Hmisc: Harrell Miscellaneous.* (2016); R package version 3.17-4. http://CRAN.R-project.org/package=Hmisc.

7. Friedman, J., Hastie, T., Tibshirani, R. Regularization Path for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* (2010); 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

8. Ho, T. Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (1995); Montreal, QC, 14–16 August 1995. pp. 278–282.

9. Li, S., Park, Y., Duraisingham, S., Strobel, F.H., Khan, N., Soltow, Q.A., Jones, D.P., Pulendran, B. Predicting Network Activity from High Throughput Metabolomics. *PLoS Computational Biology* (2013); 9(7): e1003123. doi:10.1371/journal.pcbi.1003123

10. Liaw, A., Wiener, M. Classification and Regression by randomForest. *R News* (2002); 2(3), 18--22.

11. Muhlebach, M., Sha, W., Lessons learned from metabolomics in cystic fibrosis. *Molecular and Cellular Pediatrics* (2015); 2:9 DOI 10.1186/s40348-015-0020-8

12. National Center for Biotechnology Information. PubChem Compound Database; CID=6305, https://pubchem.ncbi.nlm.nih.gov/compound/6305 (accessed Mar. 14, 2017).

13. R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

14. Revelle, W. (2015) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.5.8.

15. Rosenow, T., Kuo, W., de Bruijne, M., Murray, C., Tiddens, H., Stick, S. A new gold standard for assessing CT in early CF lung disease? *European Respiratory Journal* (2014); 44: 3446;

16. Silva, S, Carvalho, S., Ximenes, V., Okada, S., Campa, A. Melatonin and its kynurenin-like oxidation products affect the microbicidal activity of neutrophils. *Microbes and Infection* (2006) Feb;8(2):420-5.

17. Smith, C., O'Maille, G., Want, E., Qin, C., Trauger, S., Brandon, T., Custodio, D., Abagyan, R., Siuzdak, G. METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring* (2005); 27 :747-51.

18. Tibshirani, R. “Regression Shrinkage and Selection via the lasso”. *Journal of the Royal Statistical Society. Series B (methodological)* (1996); Wiley: 267–88. 58 (1). http://www.jstor.org/stable/2346178.

19. Wetmore, D., Joseloff, E., Pilewski, J., Lee, D., Lawton, K., Mitchell, M., et al. Metabolomic Profiling Reveals Biochemical Pathways and Biomarkers Associated with Pathogenesis in Cystic Fibrosis Cells. *Journal of Biology and Chemistry* (2010)*;* Y VOL. 285, NO. 40, pp. 30516 –30522

20. Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software* (2011); 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

21. Wolak, J., Esther, C., O’Connell, T. Metabolomic analysis of bronchoalveolar lavage fluid from cystic fibrosis patients. *Biomarkers* (2009); 14(1): 10.1080/13547500802688194. DOI: 10.1080/13547500802688194

22. Zhang, A., Sun, H., Wang, P., Han, Y., Wang, X. Modern analytical techniques in metabolomics analysis. *The Analysis* (2012); DOI: 10.1039/c1an15605e

Other References

1. “CF in adulthood.” *Growing old.* Cystic Fibrosis Trust. Web. 24 March 2017.

2. “The Digestive Tract.” *Cystic Fibrosis Foundation.* Web. March 2017