Distribution Agreement

Signature:

______________________________ ______________

Alexandria Portelli Date

Relationship Between Airway Metabolites and Structural Damage in Young Children with Cystic Fibrosis

By

Alexandria Portelli

Master of Public Health

Department of Biostatistics and Bioinformatics

Limin Peng, Ph.D.

Committee Chair

Joshua Chandler, Ph.D.

Committee Member

Rabindra Tirouvanziam, Ph.D.

Committee Member

Relationship Between Airway Metabolites and Structural Damage

in Young Children with Cystic Fibrosis

By

Alexandria Portelli

B.S.

Boston College

2014

Thesis Committee Chair: Limin Peng, 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 Public Health

in Biostatistics

2018

Abstract

Relationship Between Airway Metabolites and Structural Damage
in Young Children with Cystic Fibrosis

By Alexandria Portelli

Background: Cystic Fibrosis (CF) is a genetic disease affecting over 70,000 people worldwide. Airway disease, the main cause of morbidity and mortality in CF, has been found to begin soon after birth. The Perth-Rotterdam Annotated Grid Morphometric Analysis (PRAGMA-CF) method, used to score chest computed tomography (CT) scans, is a sensitive and reproducible measure of the extent of lung disease in CF children. There is limited knowledge of relationships between PRAGMA-CF scoring of chest CT scans and molecular biomarkers of disease measured by metabolomics of bronchoalveolar lavage fluid (BALF) in CF children.

Objective: Identify significant statistical correlations and relationships between the PRAGMA-CF score of overall structural airway damage (PRAGMA-%Dis, or %Dis) in CF children and specific BALF metabolites.

Methods: Univariate and multivariate biostatistical methods were used to assess a longitudinal dataset from a prospective study of CF children (I-BALL study) to identify BALF metabolites associated with airway damage. Pearson and Spearman correlation coefficients of %Dis and metabolite concentration were calculated for cohorts at ages 1, 3, and 5 years old. Linear mixed models assessed the response of metabolite concentration to covariates including %Dis, age, total BALF protein concentration and % BAL neutrophils.

Results: Significant correlations and linear relationships between the concentration of specific BALF metabolites and %Dis were identified. Trends in the Pearson and Spearman correlations coefficients change in the first 5 years of children born with CF. The linear mixed model including %Dis and total BALF protein concentration was chosen because it had the lowest Akaike information criterion (AIC) overall across most metabolites. %Dis showed significant positive associations with diacyl (aa) and acyl ether (ae) phosphatidylcholines (PCs) aa C30:0, aa C34:1, aa C36:2, ae C34:0 and ae C34:1, and sphingomyelin SM C16:0 when adjusting for total BALF protein concentration.

Conclusions: These results add to our growing understanding of early CF pathogenesis, and how metabolomics can be used to generate clinically-relevant molecular outcomes for disease monitoring at a stage when conventional biomarkers remain at low to undetectable levels. More extensive investigation of age as a covariate is needed on progression of early CF per the %Dis outcome.

Relationship Between Airway Metabolites and Structural Damage in Young Children with Cystic Fibrosis

By

Alexandria Portelli

B.S.

Boston College

2014

Thesis Committee Chair: Limin Peng, 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 Public Health

in Biostatistics

2018

Acknowledgements

I would like to thank the Department of Biostatistics at Emory for their support and development of a community that is centered on learning and personal growth. Specifically, I would like to thank Limin Peng, Ph.D. who advised me through my time at Emory, helped me to reach my full potential, and served as my Thesis Committee Chair.

This thesis would not have been possible without the guidance, support, and data from Rabin Tirouvanziam, Ph.D. and Joshua Chandler, Ph.D. Their enthusiasm and expertise gave my thesis purpose. I am grateful for their patience in explaining the clinical and scientific implications of my research. It has been my pleasure being a part of their team and I feel lucky to have gotten to work with them.

This study would not have been possible without the ongoing collaboration of Hettie Janssens, M.D., Hamed Horati, M.D., and Bob Scholte, Ph.D., at the Erasmus Medical Center / Sophia Children's Hospital in Rotterdam, the Netherlands, where clinical procedures that yielded data for this study were conducted. I would also like to thank the patients and families who engaged in this research.

I would like to thank my fellow Biostatistics cohort. They helped to create a safe and effective learning space which was also enjoyable and memorable. The moral support they gave when writing this thesis and time they took to help me filter through my ideas was invaluable.

Finally, I would like to thank my family and friends for taking my phone calls, encouraging me when I felt discouraged, and motivating me to just keep on writing. Writing a thesis was a learning process which helped me evolve into a better biostastician, researcher, and person. I would not have been able to accomplish this without outside support.

Table of Contents

Introduction 1

Background 1

Methods 5

Study Cohort: I-BALL Study 5

Data Acquisition 6

Targeted Metabolomics 6

PRAGMA-CF Scoring of Chest CT Scans 7

Data Cleaning 7

Metabolomics Data 7

PRAGMA-CF Scores 11

Age 11

Descriptive Characteristics of Study Cohort 11

Univariate Correlation Analysis 12

Pearson Correlation 12

Spearman Correlation 12

Linear Mixed Modeling 13

Assessment of Reproducibility of Metabolite Measurements 15

Computing Environment 15

Results 15

Summary Statistics 15

Pearson Correlations 18

Spearman Correlations 19

Linear Mixed Modeling 21

Akaike Information Criterion 21

Linear Mixed Model with PRAGMA and Total Protein 22

Lin’s Concordance Correlation Coefficient 25

Discussion 26

Interpretation of Results 26

Limitations 29

Further Research 29

Conclusions 30

References 32

Appendix 1: List of Metabolites Detected by AbsoluteIDQ® p180 Kit 36

Appendix 2: Histograms Comparing Distributions 45

Appendix 3: Scatterplots of Significant Metabolites 50

Introduction

Cystic fibrosis (CF) is an autosomal recessive genetic disease that affects over 30,000 people in the US and over 70,000 people worldwide.[1,2] CF is caused by more than 2,000 different mutations affecting the gene coding for the CF transmembrane conductance regulator (CFTR) protein.[3-5] It affects all major secretory organs in the body, including sweat glands and the digestive, reproductive, and respiratory tracts. Symptoms include salty skin, digestive issues such as meconium ileus, hypofertility, as well as wheezing, frequent lung infections, and progressive lung function decline.[2] With enzymatic supplements and improved patient care, the main cause of morbidity and mortality among CF patients has shifted from gastrointestinal to lung disease.[6] Due to improved care, the life expectancy for people with CF born between 2012 and 2016 has increased to 43 years, from 31 years for those born with CF two decades earlier (between 1992 and 1996).[7] Despite the progress made in the last decade, little is known about the specific molecular factors initiating lung disease in asymptomatic infants and fueling its progression at later stages.[8] Furthermore, there is a need for more effective treatments of core symptoms, such as inflammation, to limit disease progression under development in CF patients regardless of genetic background. A critical component of CF research is devoted to understanding the pathogenesis of early lung disease to inform the creation of new therapies and ultimately, a cure.

Background

CF lung disease is characterized by excessive neutrophil infiltration in the lumen of the airways. Therein, these powerful leukocytes actively release primary granules containing enzymes that may exert toxic effects (toward extracellular pathogens and/or host tissue), such as neutrophil elastase (NE), myeloperoxidase (MPO), arginase-I and cathepsins.[9] Consistent with a critical role for neutrophils in CF airway disease pathogenesis, extracellular NE activity and airway neutrophil count are the best available risk factors and correlates of lung disease in children and adults.[9,10] NE and other neutrophil proteins contribute directly to progression of lung damage in early and chronic disease stages.[9] Airway inflammation of early CF lung disease has been found to begin soon after birth.[11] It appears to be the earliest evidence of lung disease associated with CF, as studies have shown that inflammation can be present before the onset of infection and in those that have yet to show symptoms of detectable lung disease.[11-15]

Despite the presence of inflammatory mediators in the airways of all CF patients, the severity of airway damage varies broadly.[11] Therefore, to assess structural damage in young children, sensitive methods such as the Perth-Rotterdam Annotated Grid Morphometric Analysis (PRAGMA-CF) scoring system are needed to interpret chest computed tomography (CT) scans. To conduct PRAGMA-CF scoring, CT scans are overlaid with a grid and each cell is evaluated by an observer to identify bronchiectasis, mucus plugging, bronchial wall thickening, atelectasis, or normal lung structure. This information is then used to obtain a volumetric proportion of total disease (%Dis). Specific proportions such as bronchiectasis (%Bx) and trapped air (%TA) may also be quantified. The PRAGMA-CF method has been validated as a sensitive and reproducible outcome measure for assessing the extent of lung disease in children younger than 6 years with CF.[16]

The key to preserving lung function in CF is early intervention. To develop more effective interventions, the discovery of more sensitive measurements to track early disease progression and the identification of mechanistic molecular and cellular biomarkers are needed.[3] Biomarkers of airway disease and airway inflammation are usually detected through tests of clinical samples that represent airway surface fluids such as sputum, bronchoalveolar lavage (BAL) cells and fluid (BALF), or exhaled breath condensate.[3,11] Currently, BAL samples are the most reliable specimen to examine inflammation occurring in the lungs of young children with CF, because they yield information on the small airway region of the lungs where CF disease is first detectable.[17,18] Sputum is not readily available from young children and exhaled breath condensate contains low biomarker concentrations.[11] Thus, BAL is considered the gold standard method to obtain airway samples and quantify inflammatory markers.[18,19] BAL is performed by flushing the lobes of a patient's lungs with fixed volumes of saline, and retrieving the instilled fluid.[18,20] Caveats of BAL are that it is invasive, requires sedation, and is prone to technical variation among clinical sites and individual practitioners.[17,18,21]

As mentioned earlier, NE is currently one of the best candidate biomarkers for CF airway disease, as it has been correlated with structural damage and lung function decline in multiple studies in infants and adults.[12,17,22] However, detection of NE in the BALF is quite challenging due to the diluted nature of this biological specimen.

To address the gaps in understanding of early CF airway disease biomarkers, metabolomics, the measurement of small molecules (metabolites) seems particularly well-suited. Untargeted metabolomics, which analyzes a large number of chemicals without an *a priori* hypothesis about their relationship to the actual biological mechanisms underlying health or disease in a given cohort of subjects, has been proposed as a method for the identification of pathways associated with CF airway inflammation.[23] Prior studies showed that untargeted metabolomics can enable discovery of inflammatory biomarkers that could be used for early disease detection, monitoring of drug effectiveness, and lead to better understanding of disease progression.[11,20,24] Furthermore, the measurement of metabolites and resolution of associated pathways may lead to the development of new drugs and serve as indicators of efficacy in clinical trials.[11]

Untargeted metabolomics can entail multiplexed quantification of known molecules, unbiased detection of spectral features (sometimes called "discovery metabolomics"), or hybrid approaches that iteratively combine the two. Assessing the presence and abundance of known metabolites has the advantage of providing absolute quantities of all recovered chemicals in a sample, with varying degrees of analytical precision (i.e., quantitative vs. semi-quantitative data). However, these methods have the disadvantage of being inflexible in calibration when they are provided as kits, and it can be difficult to predict their successful application to biological matrices in which they were not first developed (e.g., plasma-validated kits applied to BALF).[25]

So far, only two published studies have described metabolomics analyses of BALF from CF patients in relation to airway inflammation.[20,23] The study by Wolak *et al*. measured metabolites with nuclear magnetic resonance (NMR), not mass spectrometry (MS) as was used in this analysis, and compared these measurements to subjects with high vs. low airway inflammation.[20] Esther *et al.* furthered the understanding of airway disease biomarkers through identification of metabolic pathways in CF children.[23] The latter publication reported untargeted metabolomics of BALF in relation to lobe-specific PRAGMA-CF scoring of chest CT scans, and found several hits using a hierarchical mixed-effects model. Additional studies are critical to confirm hypotheses generated in such discovery analyses, provide coverage of additional metabolites, and advance mechanistic understanding of the discrete steps in the progression of early airway disease in young CF children.

The main goal of this thesis is to fill this critical gap, by identifying significant statistical correlations that exist between specific BALF metabolites and the PRAGMA-CF score of structural airway damage in young CF children. To accomplish this, we used a multiplex metabolomics platform developed for human plasma samples (Biocrates AbsoluteIDQ® p180 Kit; *vide infra*), which provides absolute concentrations of for nearly 200 endogenous metabolites. The remainder of this thesis will go through details of univariate and multivariate biostatistical methods used to assess a longitudinal dataset and identify metabolites that may be associated with airway inflammation. Key results will be highlighted and discussed to draw conclusions about the contribution of metabolomics to the understanding of the progression of early airway disease in children with CF. Identification of metabolites will help clinicians and researchers better understand the biological pathways related to CF to develop better interventions.

Methods

Study Cohort: I-BALL Study

An NIH R01-funded study (R01HL126603, Principal Investigator: Rabindra Tirouvanziam) is currently ongoing in collaboration between our team at Emory and the Erasmus University / Sophia Children's Hospital group in Rotterdam, the Netherlands, combining chest CT scans with multiple molecular and cellular outcome measures on blood and BAL, which include but are not limited to metabolomics. In what is referred to as the I-BALL study (Principal Investigator: Hettie Janssens, M.D.), children with CF are enrolled and followed over the course of the first 6 years of life. Blood samples, BAL samples, and CT scans are collected during visits based on age [E1 (6 months of age), E2 (1 year old), E3 (2 years old), E4 (3 years old), E5 (4 years old), and E6 (5 years old)]. Demographic (e.g., age, gender) and clinical characteristics (e.g., CTFR mutation, infection status) are recorded at each visit. The data collection and management for this paper was performed using the OpenClinica open source software, version 3.1© (Copyright © OpenClinica LLC and collaborators, Waltham, MA, USA, www.OpenClinica.com). This particular study also collects data from the BALF samples and PRAGMA-CF scores from CT scans of children enrolled in the I-BALL cohort.

Data Acquisition

Targeted Metabolomics

BAL was collected by bronchoscopy by four serial saline instillations, of which an aliquot of the pooled second and third fractions (B2+3) was used for metabolomics studies. Two preparation protocols were used over the course of sample collection. According to the first protocol, BALF was prepared through a 330 x *g* spin at 4 ºC for 10 minutes. According to the second protocol, BALF was prepared by first spinning the BAL at 800 x *g* at 4 ºC for 10 minutes, followed by spinning the supernatant a second time at 3000 x *g* at 4ºC for 10 minutes. Aliquots were stored immediately after single or dual centrifugation at -80 ºC and shipped frozen to Emory for analysis. BALF samples were then analyzed with an AbsoluteIDQ® p180 Kit (Biocrates, Innsbruck, Austria) on a triple quadrupole mass spectrometer at the Emory Lipidomics Core. This commercial kit created a sample profile for 188 metabolites classified as amino acids, biogenic amines, acylcarnitines, glycerophospholipids, sphingolipids, or hexose (generically; glucose may not be discriminated from other species of equal mass). Briefly, samples were distributed into 96-well plates, extracted and run through ultra-high performance liquid chromatography combined with tandem mass spectrometry (UPLC-MS/MS) analysis and flow injection analysis (FIA)-MS/MS (in parallel) to yield an absolute concentration for each measured metabolite (full list of metabolites in **Appendix 1**).

PRAGMA-CF Scoring of Chest CT Scans

At or around the same visit that the BAL was obtained, a CT scan was also performed to assess structural lung disease. From these scans, an overall score (PRAGMA-%Dis) was calculated and used in the analyses presented here. To minimize variability, PRAGMA-CF scores were done in batches by a qualified clinician. The subjects in this analysis were scored in two separate batches. Intra-observer reliability of the measurements was found to be high.

Data Cleaning

Metabolomics Data

Data were obtained in Excel sheets as numerical values or “NA”. “NA” (not available) was recorded whenever an analyte was not detected. As detailed in **Table 1** below, each reading was classified as Valid, <LOD, LLOQ, ULOQ, Semi Quant, ISTD out of range, or No Interception.

Table 1: Classification of metabolomics data

Classification	Definition
Valid	Internal standard intensity is in range, valid concentrations can be read, concentration is between LLOQ and ULOQ
<LOD	Value is below limit of detection (LOD), and treated as missing value
LLOQ	Value is above LOD but below lower limit of quantification (LLOQ), value can be used as semi-quantitative
ULOQ	Value is above the upper limit of quantification (ULOQ), value can be used as semi-quantitative
Semi Quant	Applicable to FIA data, calibration curve for metabolite is determined by a compound similar in composition to what is being measured
ISTD Out of Range	Internal Standard Out of Range, calibration curve for metabolite cannot be determined, therefore measurement is not valid, and treated as missing value
No Interception	Concentration of metabolite cannot be calculated because the value is too high and does not intersect with calibration curve, and treated as missing value

We further classified each data point as “usable” or “unusable”, per the definitions above in **Table 1**, for analysis. If a value was <LOD, ISTD out of range, No Interception, or NA the data was “unusable”. Otherwise, the value was classified as “usable”. The output of LC

and FIA runs included 188 metabolites measured over 79 experimental samples along with a blank, standards, quality control (QC) samples, and a pooled BALF reference sample. Data were converted from µM to nM by using a multiplication factor of 1,000 nmol per µmol. “Unusable” data were imputed to a value of NA, if not already NA. Metabolites to be used for analysis were then filtered based on quality metrics. Data were filtered based on percent unusable data, values that were classified as “Valid” or “Semiquant”, and by comparing the sample mean to the blank concentration. **Figure 1** shows the attrition of study metabolites occurring upon implementation of these data cleaning steps.

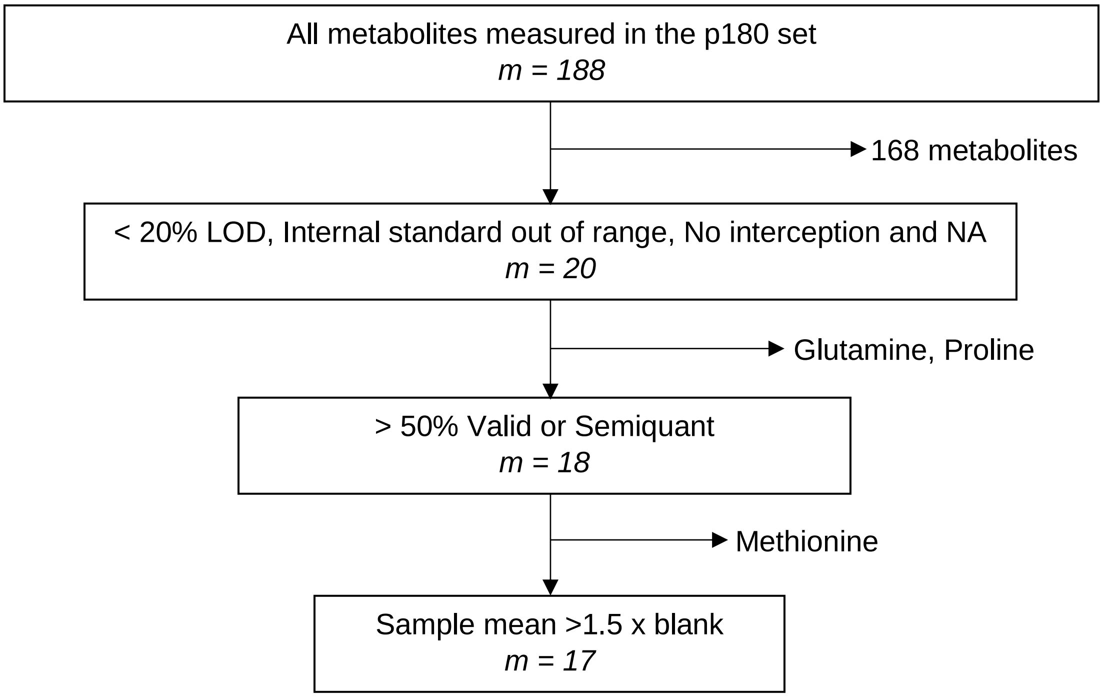


Figure 1: Attrition of metabolites

After data filtering, 17 of 188 metabolites remained to be used in the analysis. **Appendix 1** specifies which 17 metabolites remained. This high rate of attrition is not entirely surprising for two reasons: first, BALF is known to be a dilute sample from the instillation of saline required to collect it. Second, the kit uses plasma values of the metabolites measured for calibration. Therefore, it may be less sensitive to the lower levels expected in BALF.

Of the 17 metabolites being investigated, glycerophospholipids and a sphingomyelin were included. Their nomenclature refers to their chemical structure where phosphatidylcholine and sphingomyelin are abbreviated by PC and SM, respectively. Their name then specifies if they are diacylated (aa) or have one acyl and one ether (ae) bond to describe the glycerol-bound fatty acids. The first number (e.g., 30 in PC aa 30:0) represents the amount of total fatty acyl/ether carbons while the second number (:0) specifies the number of double bonds present in either fatty acid. For example, PC aa C30:0 corresponds to a diacylated phosphatidylcholine with 30 acyl groups and no carbon double bonds. Additionally, lysolipids were analyzed which only have one fatty acid, and are denoted with a single “a” or “e”. Many of the lipid measurements can encompass multiple unique lipid compounds with various structural arrangements of the fatty acids resulting in the same exact mass.

Thirty-six unique patients were included in this study. A total of 44 samples were used, 9 patients had two visits, and 26 had only one visit. Samples were not included if they had inaccurate PRAGMA scores (n=3) or if all metabolite concentrations were NA (n=1); the latter case was suspected to be a mis-labeled sample vial, and was corroborated by several orthogonal methods not the subject of this thesis.

PRAGMA-CF Scores

PRAGMA-CF scores were generated in two batches. The total disease score (%Dis, on a 0-100 scale) was used for this analysis. Based on the methods developed by Rosenow *et al.*, if a CT scan was scored in each batch, the value from the first batch was taken because it had a larger sample size.[16] If there was no score from batch 1, the score from batch 2 was used.

Age

The age of each study subject at the time of sample collection was found by linking data with clinical information entered in OpenClinica. Each patients' birth month and year and date of bronchoscopy of which the BAL sample was retrieved were obtained from the clinical dataset. Age was calculated as the difference in months between the BAL date and the date of birth (DOB). A more precise age could not be calculated as DOB was only recorded as month and year, but this modest level of imprecision is not expected to skew results and (critically) enables patient health information to remain hidden during analysis. Data checking was done to ensure correct dates were used in the age calculation.

Descriptive Characteristics of Study Cohort

Summary statistics of the study cohort across visits were examined. Means and standard deviations of PRAGMA scores and age were summarized along with the number and percent of the cohort gender, infection status, and CF mutation type.

Univariate Correlation Analysis

Pearson and Spearman tests were performed for data corresponding to separate age-based visits (E2, E4, E6, corresponding to ages 1, 3 and 5, respectively) to assess the correlation between the metabolic concentration and PRAGMA-CF score. The cross-sectional correlations at multiple visits provide an intuitive view of how the relationship between the metabolic concentration and PRAGMA score change over time, prior to implementation of more complex models (*vide infra*).

Pearson Correlation

Normality of the distribution for each metabolite concentration was assessed through histograms, which suggest using natural logarithm transformed metabolite concentration to better approximate normal distributions. **Appendix 2** shows histograms before and after such data transformation. Pearson correlation analysis used the natural log transformation of the values. Missing data were left as NA and not imputed to prevent outlier-driven bias in results. The Pearson correlation coefficient between each log-transformed metabolite concentration and the PRAGMA was calculated. The corresponding p-values for testing zero Pearson correlation coefficient were generated.

Spearman Correlation

As the Spearman correlation is a non-parametric, rank-based test, the data were not log transformed. The missing data in our dataset represent cases where no signal was detected for the analyte. We impute the missing data as 0. Another common imputation strategy adopted in metabolomics analysis is to impute a value to ½ LOD. These two different imputation approaches are expected to produce similar results on Spearman correlation coefficient, which is a rank-based measure. Spearman correlation coefficients between PRAGMA-%Dis and the concentration of each of the 17 metabolites, and corresponding p-values for testing zero Spearman correlation were computed by visit.

Linear Mixed Modeling

We fit the data with linear mixed models to assess the relationships between metabolic concentrations and PRAGMA scores simultaneously across different visits. We set the outcome variable in each linear mixed model as the concentration of each metabolite to address the interest in how metabolic perturbation activities are responsive to lung phenotype progression (assessed by PRAGMA scores). We also specified the correlation structure in each linear mixed model as the autoregressive model AR(1) to properly account for the intra-person correlation of longitudinal measurements of the same subject at different visits. %Dis was included as a variable in every model evaluated, as the purpose of this study was to assess the relationship between %Dis and metabolomic concentrations. Multiple models were run for each metabolite to select the covariates to include in the final linear mixed model. We considered covariates including age at each visit, total protein concentration in the sample, and percent BAL neutrophils, as well as their interaction terms. Incorporating age at visit can account for the potential temporal variation in metabolic concentrations. Total BALF protein concentration and percent BAL neutrophil were considered because they have been linked to the pathogenesis of CF lung disease, and may display covariance with significant metabolites, particularly those that reflect inflammatory pathways. The best model was determined based on the Akaike Information Criterion (AIC). The AIC is a measure of the ability of a model to predict the observed data, utilizing a specific measure such as log-likelihood, restricted maximum likelihood

(REML), or conditional log-likehood. This analysis estimated the AIC using REML. The optimal model is the one that minimizes the AIC.[26] **Table 2** lists the 11 models which include %Dis and other covariates. We evaluated and compared these models based on AIC.

Table 2: Descriptions of candidate models

Model	Variables
p	PRAGMA
p+a	PRAGMA, age
p+pr	PRAGMA, total protein
p+ne	PRAGMA, neutrophil count
p*a	PRAGMA, age, PRAGMA*age
p*a+pr	PRAGMA, age, PRAGMA*age, total protein
p*a+pr*a	PRAGMA, age, PRAGMA*age, total protein, total protein*age
p*a+ne	PRAGMA, age, PRAGMA*age, neutrophil count
p*a+ne*a	PRAGMA, age, PRAGMA*age, neutrophil count, neutrophil count*age
p*a+pr+ne	PRAGMA, age, PRAGMA*age, total protein, neutrophil count
p*pr	PRAGMA, total protein, PRAGMA*total protein
p+pr+a	PRAGMA, total protein, age
p+pr+ne	PRAGMA, total protein, neutrophil count

A boxplot was made to compare the AIC across all metabolites in the different models. The model chosen was the one with the overall lowest AIC value across the majority of metabolites, as opposed to the best model for each metabolite separately.

Assessment of Reproducibility of Metabolite Measurements

Of the samples in the dataset, 11 were run at two unique sites (Emory, the primary analytical site for the study; and headquarters of Biocrates, Innsbruck, Austria, for a pilot feasibility study). Here, targeted metabolomics was intended to provide robust quantification of a large number of metabolites without *a priori* knowledge of pathway-specific perturbations. However, this kit was not optimized for BALF. We used Lin's concordance correlation coefficient (CCC) to assess inter-site agreement of the 11 twice-analyzed BALF samples for the 17 metabolites that withstood statistical attrition. Lin's CCC characterizes the strength of the agreement between two paired measurements of the same sample. It is a scaled measure bounded between -1 and 1; a value 1 (or -1) means perfect agreement (or disagreement) and a value 0 signifies a purely random relationship.

Computing Environment

All analyses were performed using R Statistical Software (Version 3.4.0) for Windows 10, implemented in RStudio version 0.99.903. All raw data, code, and processed results have been archived on Emory's storage system within Dr. Tirouvanziam's ISILON server.

Results

Summary Statistics

Summary statistics for the samples stratified by visit were calculated (see **Table 3**). There were 11 samples at E2, 1 at E3, 17 at E4, 1 at E5, and 14 at E6. The average age at each visit aligned with the age range specified in the protocol. Mean %Dis scores at E2, E4, and

E6 were 1.50 (SD = 0.83), 2.99 (SD = 1.84), and 4.34 (SD = 2.86), respectively. 55% of samples at E2, 59% at E4 and 43% at E6 were from females. The percent of samples with infections increased from 18% at E2 to 41% at E4 to 64% at E6. In general, more than half the samples at each visit were from patients heterozygous for the F508del mutation of the CF gene (i.e., carrying one F508del, and one other mutation).

Table 3: Descriptive characteristics stratified by visit

	Visit									
Characteristic	**E2**		**E3**		**E4**		**E5**		**E6**	
	n = 11		*n = 1*		*n = 17*		*n = 1*		*n = 14*	
	mean	**(SD)**	**mean**	**(SD)**	**mean**	**(SD)**	**mean**	**(SD)**	**mean**	**(SD)**
Age (months)	13.0	(0.9)	25.0	(NA)	37.1	(0.8)	49.0	(NA)	60.9	(0.6)
%Dis	1.5	(0.8)	3.9	(NA)	3.0	(1.8)	2.3	(NA)	4.3	(2.9)
	n	**(%)**	**n**	**(%)**	**n**	**(%)**	**n**	**(%)**	**n**	**(%)**
Female	6	(55)	1	(100)	10	(59)	1	(100)	6	(43)
Infection present	2	(18)	0	(0)	7	(41)	1	(0)	9	(64)
CTFR mutation										
F508del homozygous	5	(45)	0	(0)	8	(47)	0	(0)	5	(36)
F508del heterozygous	6	(55)	1	(100)	8	(47)	1	(100)	8	(57)
Other	0	(0)	0	(0)	1	(6)	0	(0)	1	(7)

Pearson Correlations

Figure 2 presents the Pearson correlation coefficients between %Dis and the natural log of the concentration (nM) for each of the 17 metabolites at visits E2, E4, and E6. It shows that metabolites and PRAGMA scores have rather low correlations at the young age (e.g., 1 year) and demonstrate much stronger correlations at older age (e.g., 3 and 5 years). More specifically, all but one of the Pearson correlation coefficients at E2 were negative, and all were not significant ($p \geq 0.05$). At E4, all of the correlation coefficients were positive, and 15 of the 17 metabolites analyzed were statistically significant ($p < 0.05$). The change in the Pearson correlation coefficient per metabolite at E6 was not as uniform, as some coefficients increased compared to E4 while others decreased. At E6, some remained significant ($p < 0.05$), while others did not. The metabolite that never had a significant correlation with PRAGMA was acetylcarnitine (C2).

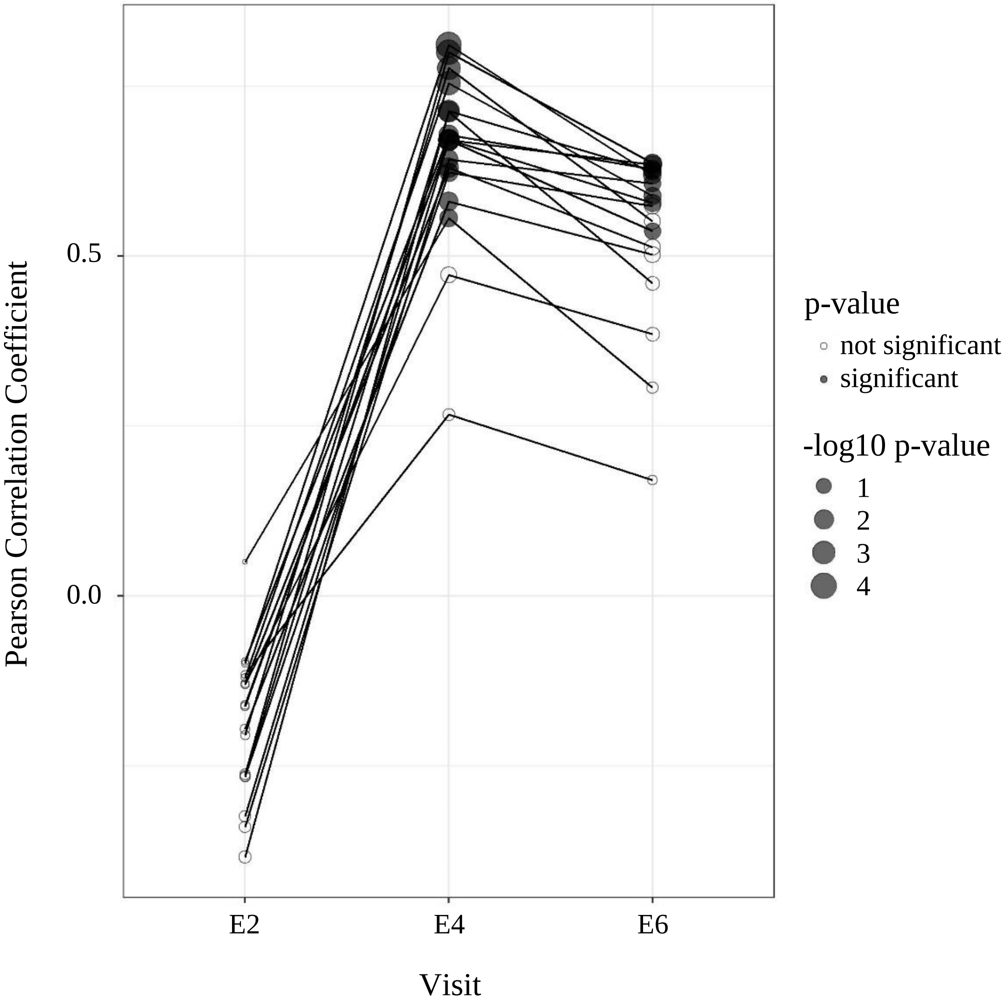


Figure 2: Trend in Pearson correlation coefficients between natural log metabolite concentration and %Dis per each metabolite at visits E2, E4, and E6.

Spearman Correlations

Figure 3 presents the Spearman correlation coefficient (rho) between %Dis and concentration (nM) of each of the 17 metabolites at visits E2, E4, and E6. The observations for **Figure 3** are similar to those for **Figure 2**. The Spearman correlation coefficients at E2 were all non-significant (p≥0.05) and all but 2 were negative. At E4, the Spearman correlation coefficient of 16 of the 17 metabolites analyzed were significant (p<0.05) and all were positive. Generally, at E6, the Spearman correlation coefficients decreased but remained significant. As in the Pearson correlation, the metabolite that never had a significant Spearman correlation with %Dis was C2.

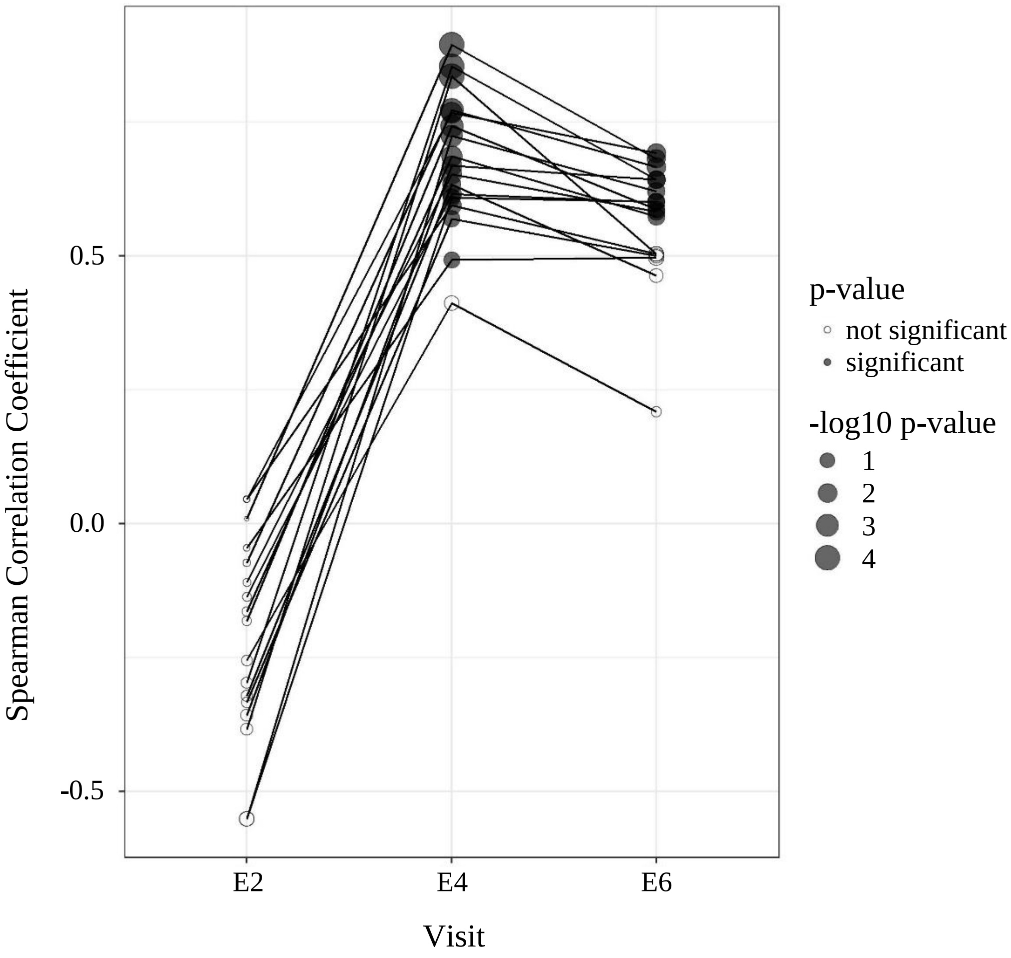


Figure 3: Trend in Spearman correlation coefficient between metabolite concentration (nM) and %Dis per each metabolite at E2, E4, and E6

Linear Mixed Modeling

Akaike Information Criterion

In **Table 4**, we list the model with the smallest AIC corresponding to each metabolite. The various models assessed were defined in **Table 2**. Metabolites of the same class [i.e., amino acids (glycine), biogenic amines (taurine), acetylcarnitines (C2), and glycerophospholipid] had the same model with lowest AIC.

Table 4: Model with lowest AIC for each metabolite

Metabolite	**Model with lowest AIC**	**Metabolite**	**Model with lowest AIC**
Gly	p+ne	PC aa C34:3	p+pr
Taurine	p	PC aa C36:2	p+pr
C2	p+a	PC aa C36:3	p+pr
PC aa C30:0	p+pr	PC aa C36:4	p+pr
PC aa C32:0	p+pr	PC ae C32:1	p+pr
PC aa C32:1	p+pr	PC ae C34:0	p+pr
PC aa C32:2	p+pr	PC ae C34:1	p+pr
PC aa C34:1	p+pr	SM C16:0	p+pr
PC aa C34:2	p+pr		

The box plot in **Figure 4** shows the distribution of the AIC across metabolites for each linear mixed model. The overall model including %Dis and total protein concentration was chosen because it had the lowest AIC overall across most of metabolites, as suggested by **Table 4** and **Figure 4**.

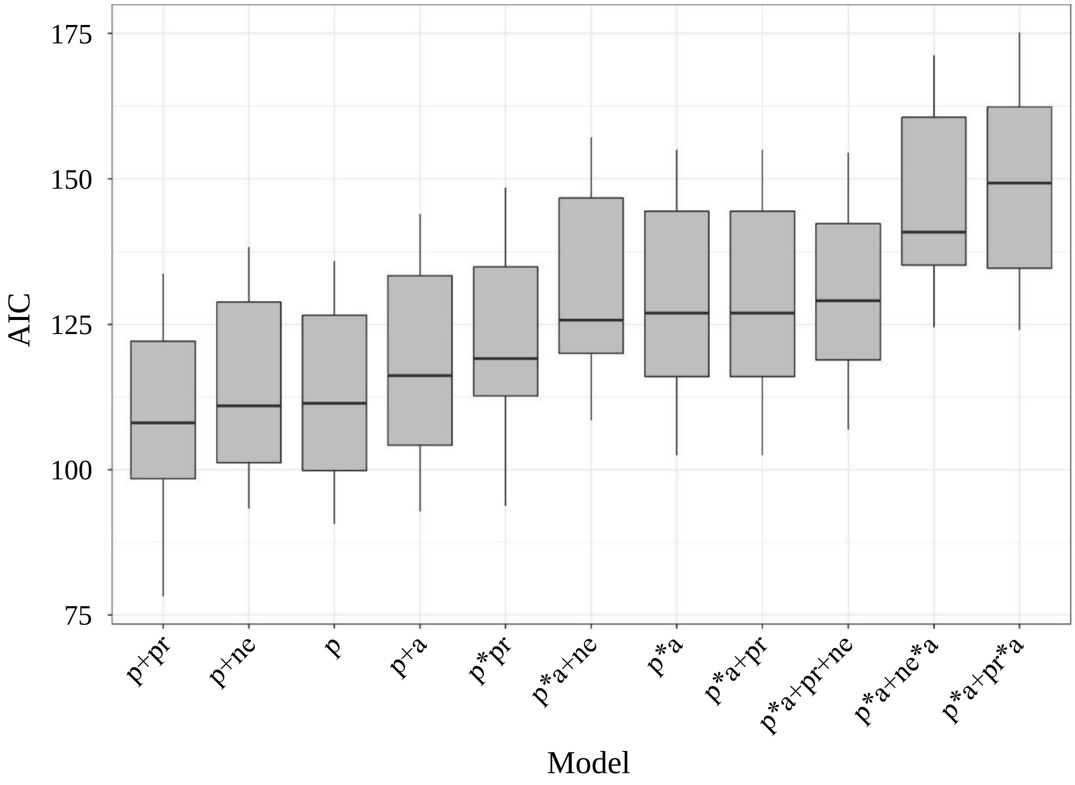


Figure 4: Boxplot of distribution of AIC for metabolites in various models

Linear Mixed Model with PRAGMA and Total Protein

The linear mixed model chosen for evaluating all metabolites is expressed below in Equation (1):

$$(1) \quad \log(\mathrm{Concentration}_{i,j}) = \beta_0 + \theta_i + \beta_1 * \mathrm{PRAGMA}_{i,j} + \beta_2 * \mathrm{total\ protein}_{i,j} + \epsilon_{i,j}$$

where θ_i represents the random intercept specific to subject i, and $\epsilon_{i,j}$ represents the error term for subject i at visit j (j=1, 2, 3, 4, 5). The correlations among ($\epsilon_{i,j}$; j=1, 2, 3, 4, 5) is assumed to follow the AR(1) structure. **Table 5** below summarizes the estimates and respective p-values of the estimated %Dis and total protein parameters in the models for each metabolite.

Table 5: Linear mixed model results in model with %Dis and total protein

Metabolite	Total samples	Unique subjects	Subjects with one repeat	%Dis		Total protein	
				Estimate	*p-value*	*Estimate*	*p-value*
Gly	42	35	7	-0.0177	0.7547	0.0018	0.0518
Taurine	44	35	9	-0.0090	0.8185	0.0019	0.0080*
C2	44	35	9	-0.0533	0.2278	0.0017	0.0120*
PC aa C30:0	44	35	9	0.1129	0.0495*	0.0034	0.0019*
PC aa C32:0	44	35	9	0.1037	0.0887	0.0036	0.0017*
PC aa C32:1	44	35	9	0.0910	0.0823	0.0037	0.0011*
PC aa C32:2	38	32	6	0.0246	0.6012	0.0027	0.0060*
PC aa C34:1	44	35	9	0.1140	0.0379*	0.0037	0.0009*
PC aa C34:2	44	35	9	0.0952	0.0820	0.0034	0.0011*
PC aa C34:3	37	31	6	0.0082	0.8696	0.0029	0.0058*
PC aa C36:2	42	34	8	0.1303	0.0165*	0.0034	0.0010*
PC aa C36:3	40	33	7	0.0700	0.1758	0.0031	0.0024*
PC aa C36:4	38	32	6	0.0200	0.4973	0.0023	0.0014*
PC ae C32:1	35	29	6	0.0424	0.3406	0.0033	0.0017*
PC ae C34:0	40	33	7	0.1306	0.0178*	0.0032	0.0029*
PC ae C34:1	41	33	8	0.1352	0.0091*	0.0037	0.0007*
SM C16:0	40	33	7	0.1218	0.0261*	0.0036	0.0007*

***** denotes significant value (p-value <0.05)

All metabolites except glycine had an estimate where the total protein was significant (p<0.05). The estimate for PRAGMA score was significant (p<0.05) in models for the metabolites PC aa C30:0, PC aa C34:1, PC aa C36:2, PC ae C34:0, PC ae C34:1, and

SM C16:0. A significant estimate signifies a potential linear relationship between %Dis and metabolic activity, after adjusting for total protein concentration.

Lin’s Concordance Correlation Coefficient

It is shown in **Table 6** that the estimated Lin’s CCC measures for all metabolites are 0.7 or higher, with the exception of PC aa 32:0 (Lin’s CCC=0.6923). This demonstrates high agreement between the two sites of analysis and provides a confirmation of the robustness of quantitative and semi-quantitative results in the BALF sample matrix.

Table 6: Lin’s CCC per metabolite

Metabolite	Lin CCC
Gly	0.8731
Taurine	0.9054
C2	0.8968
PC aa C30:0	0.8181
PC aa C32:0	0.6923
PC aa C32:1	0.7001
PC aa C32:2	0.7475
PC aa C34:1	0.7506
PC aa C34:2	0.8169
PC aa C34:3	0.8057
PC aa C36:2	0.7533
PC aa C36:3	0.8174
PC aa C36:4	0.8204
PC ae C32:1	0.7624
PC ae C34:0	0.7021
PC ae C34:1	0.7338
SM C16:0	0.8625

Discussion

Interpretation of Results

Our analyses identified significant correlations and linear relationships between the nM concentration of specific BALF metabolites and %Dis in young children with CF. In the assessments of the trends of both the Pearson and Spearman correlation coefficients, the E2 visit did not see any significant relationships while the E4 visit showed multiple significant positive correlations. This likely follows the progression (and increase in range between patients) of airway disease between the ages of 1 and 3 years in children with CF. Targeted prevention strategies may be most effective during this time. Although every effort was made to control for batch effects, we cannot rule out a systematic procedural difference between the three age groups as a potential confounder. For example, the difference in correlations and significance could derive from a systematic error in PRAGMA-CF scoring or BAL procedure methods between the different age groups. However, the natural history of early CF (worsening lung disease over time) seems to support our delayed detection of metabolic perturbations in these patient samples. At the E6 visit, the correlations seemed to be somewhat stable with respect to the previous E4 subset. Further analysis could examine whether these correlations are stable or undergo more change as patients age into late childhood, adolescence and adulthood.

Linear mixed models were used to assess the relationships between PRAGMA scores and metabolites allowing us to fully utilize the multiple longitudinal measurements from the same subject while properly accounting for their within-subject correlations. In the process of choosing the best model, different classes of metabolites had different models with the lowest AIC. The separation of metabolites by class indicates a cross-species disequilibrium in airway regulation except according to biochemical relatedness. This was anticipated due to certain intra-class overlap of function, transporters, removal enzymes and scavenger receptors. The model with the lowest AIC for glycine included %Dis and % BAL neutrophils. However, this was not the case for glycerophospholipids or sphingomyelins, signifying that different covariates could influence different classes of metabolites. This supports the notion that even closely related physiological processes, such as inflammation and bronchiectasis, may exert their most potent influences on different subsets of metabolites through the kinetic non-equilibrium of biochemical reactions by various enzymes.

Age was only included in the model with the lowest AIC for C2. Scatter plots of the concentrations of the metabolites with significant parameter estimates compared to %Dis can be seen in **Appendix 3**. The points are grouped by visit (E2, E4, E6) and a 95% CI statistical ellipse is drawn around each visit. The overlap of the ellipses supports the idea that age was not an important variable to include in this analysis, although a larger sample size may have reached a different conclusion. As early CF progresses with age, it is interesting that age was not chosen for the final model based on AIC for other metabolites. This could be due to the small sample size of children that had more than one visit (only 9 of the 36 unique patients included). Concluding that age is not an important variable to include in the model does not help to explain the reasoning for the trend in Pearson and Spearman correlation results. In fact, the large change in correlations of metabolites between E2 and E4 visits indicates age is a very important factor, so that lack of statistical power is the most likely cause that it was not more important in our linear mixed-effects model.

Upon assessment of all metabolites with the linear mixed models including %Dis and total protein count as variables, the estimate associated with total protein concentration was significant for all except glycine. This supports the claim that total protein concentration is associated with activity of the metabolites analyzed, specifically biogenic amines, acetylcarnitines, and glycerophospholipids/sphingomyelins. This model, however, cannot give insight into causality.

In this study, the only metabolites that were linked to a significant coefficient for PRAGMA score in the linear mixed model were several glycerophospholipids and the sole analyzed sphingomyelin. We found that %Dis had a positive, significant estimate for PCs aa C30:0, aa C34:1, aa C36:2, ae C34:0, ae C34:1, and SM C16:0 when modeled with total protein as a covariate. This could mean that the abundance of these types of compounds have the potential to be responsive to the severity of early airway disease when accounting for the influence of total protein, which is a crude proxy for burden of inflammatory exudate in the airway lumen. Therefore, they may reflect the extent of a particular biochemical activity of inflammatory cells (or possibly occurring independent of inflammation) which is increased in patients with worse disease. Such a biochemical mechanism should be active in lipid pathways (synthesis, transport or catabolism), and may reflect such activities in neutrophils. However, further investigation needs to be performed to confirm these findings.

Limitations

There were multiple limitations in this study that may have affected the results. First, the relationship between metabolite concentration and %Dis could not be analyzed for all metabolites because of the stringent filtering of the Absolute*IDQ*® p180 results we performed to ensure optimal quality of data. Because of this, a significant biological relationship could have been missed. Second, the PRAGMA-CF score was computed manually by a clinical researcher and is thus prone to observational error. Third, the process of preparing the BALF used for the metabolomics analysis was done in two different ways over the course of sample collection and was not accounted for in the analysis. The samples done following the second protocol that included a second centrifugation step to remove extra contaminants, including airway bacteria, that could influence BALF composition. However, the likelihood of a bacterial influence on the results is not certain as conditions (i.e., samples were stored at -80 ºC) were not favorable for bacteria metabolism to occur *ex vivo*. After isolation from BAL, BALF was always maintained at either 4 or -80 ˚C.

Further Research

There is room for further analysis to expand on the results from this study. For example, a logistic regression may be conducted to assess if the severity of early CF airway disease is associated with the presence or absence of certain metabolites. This approach would have greater sensitivity than the current approach for any metabolites which are altered in binary fashion according to progression of disease. Ideally, a control cohort should be included to assess the difference in concentrations or trends compared to a cohort with CF to potentially identify CF specific metabolites (however, healthy control BAL sampling is extremely rare; generally, some form of disease must be present for patients to justify this clinical procedure). Models to account for interaction between metabolites, including but not limited to metabolites associated with the same pathway, should be assessed. Also, as the linear mixed model with the lowest AIC was not the same for all metabolites, different models for different classes of metabolites may be necessary to use in future analyses. However, more information is needed to confirm and better understand this result these findings. Finally, more extensive investigation of age as a potential covariate is needed on progression of early CF per the %Dis outcome.

Conclusions

Investigating how the severity of early airway disease relates to metabolomics in patients with early CF can lead to a greater understanding of the progression of the disease. These results add to our growing understanding of early CF pathogenesis, and how metabolomics can be used to generate clinically-relevant molecular outcomes for disease monitoring at a stage when conventional biomarkers remain at low to undetectable levels. This analysis supports the idea that the progression of early airway disease is related to metabolic pathway activity and steady-state metabolite concentrations as the trends in Pearson and Spearman correlations change in the first 5 years of children born with CF. Utilizing a linear mixed model to account for intra-person correlation is a method to continue to pursue to explore longitudinal analysis in metabolomics. Progression of early airway inflammation and damage in children with CF varies individually and may become measurable at different ages. A linear mixed model can account for these inter-individual differences. As additional data are collected, more robust conclusions can be drawn to further the understanding of the relationship of the progression of early airway disease with metabolomics in children with CF. This information is important for clinicians and researchers in the development of effective early interventions to limit lung function decline in CF.

References

1 Cystic Fibrosis Foundation. Cystic Fibrosis Foundation Patient Registry-2016 Annual Data Report. (Bethesda, Maryland, 2017).

2 *About Cystic Fibrosis*, <https://www.cff.org/What-is-CF/About-Cystic-Fibrosis/>. Accessed 22 March 2018.

3 Ramsey, K. A., Schultz, A. & Stick, S. M. Biomarkers in paediatric cystic fibrosis lung disease. *Paediatric respiratory reviews* **16**, 213-218 (2015).

4 US CF Foundation, J. H. U., The Hospital for Sick Children, The Clinical and Functional & TRanslation of CFTR (CFTR2). http://cftr2.org. Accessed October 22.

5 Rowe, S. M., Miller, S. & Sorscher, E. J. Cystic fibrosis. *The New England journal of medicine* **352**, 1992-2001, doi:10.1056/NEJMra043184 (2005).

6 Schindler, T., Michel, S. & Wilson, A. W. Nutrition Management of Cystic Fibrosis in the 21st Century. *Nutrition in clinical practice : official publication of the American Society for Parenteral and Enteral Nutrition* **30**, 488-500, doi:10.1177/0884533615591604 (2015).

7 Foundation, C. F. 2016 Cystic Fibrosis Foundation Patient Registry Highlights. (2017).

8 Simon, R. (2018). Cystic Fibrosis: Overview of the treatment of lung disease. In A. Hoppin (Ed.), *UpToDate*. Retrieved March 22, 2018, from https://www.uptodate.com/contents/cystic-fibrosis-overview-of-the-treatment-of-lung-disease.

9 Margaroli, C. & Tirouvanziam, R. Neutrophil plasticity enables the development of pathological microenvironments: implications for cystic fibrosis airway disease. *Molecular and cellular pediatrics* **3**, 38, doi:10.1186/s40348-016-0066-2 (2016).

10 Sly, P. D. & Wainwright, C. E. Diagnosis and early life risk factors for bronchiectasis in cystic fibrosis: a review. *Expert review of respiratory medicine* **10**, 1003-1010, doi:10.1080/17476348.2016.1204915 (2016).

11 Giddings, O. & Esther, C. R., Jr. Mapping targetable inflammation and outcomes with cystic fibrosis biomarkers. *Pediatric pulmonology*, doi:10.1002/ppul.23768 (2017).

12 Sly, P. D. *et al.* Lung disease at diagnosis in infants with cystic fibrosis detected by newborn screening. *American journal of respiratory and critical care medicine* **180**, 146-152, doi:10.1164/rccm.200901-0069OC (2009).

13 Balough, K. *et al.* The relationship between infection and inflammation in the early stages of lung disease from cystic fibrosis. *Pediatric pulmonology* **20**, 63-70 (1995).

14 Khan, T. Z. *et al.* Early pulmonary inflammation in infants with cystic fibrosis. *American journal of respiratory and critical care medicine* **151**, 1075-1082, doi:10.1164/ajrccm.151.4.7697234 (1995).

15 Tiddens, H., Silverman, M. & Bush, A. The role of inflammation in airway disease: remodeling. *American journal of respiratory and critical care medicine* **162**, S7-s10, doi:10.1164/ajrccm.162.supplement_1.maic-2 (2000).

16 Rosenow, T. *et al.* PRAGMA-CF. A Quantitative Structural Lung Disease Computed Tomography Outcome in Young Children with Cystic Fibrosis.

American journal of respiratory and critical care medicine **191**, 1158-1165, doi:10.1164/rccm.201501-0061OC (2015).

17 Brennan, S., Gangell, C., Wainwright, C. & Sly, P. D. Disease surveillance using bronchoalveolar lavage. *Paediatric respiratory reviews* **9**, 151-159, doi:10.1016/j.prrv.2008.01.002 (2008).

18 De Blic, J. *et al.* Bronchoalveolar lavage in children. ERS Task Force on bronchoalveolar lavage in children. European Respiratory Society. *European Respiratory Journal* **15**, 217-231 (2000).

19 Stafler, P., Davies, J. C., Balfour-Lynn, I. M., Rosenthal, M. & Bush, A. Bronchoscopy in cystic fibrosis infants diagnosed by newborn screening. *Pediatric pulmonology* **46**, 696-700, doi:10.1002/ppul.21434 (2011).

20 Wolak, J. E., Esther Jr, C. R. & O'Connell, T. M. Metabolomic analysis of bronchoalveolar lavage fluid from cystic fibrosis patients. *Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals* **14**, 55-60 (2009).

21 Sofia, M. *et al.* Exploring airway diseases by NMR-based metabonomics: a review of application to exhaled breath condensate. *Journal of biomedicine & biotechnology* **2011**, 403260, doi:10.1155/2011/403260 (2011).

22 Pillarisetti, N. *et al.* Infection, inflammation, and lung function decline in infants with cystic fibrosis. *American journal of respiratory and critical care medicine* **184**, 75-81, doi:10.1164/rccm.201011-1892OC (2011).

23 Esther, C. R., Jr. *et al.* Metabolomic biomarkers predictive of early structural lung disease in cystic fibrosis. *The European respiratory journal* **48**, 1612-1621, doi:10.1183/13993003.00524-2016 (2016).

24 Nobakht, M. G. B. F., Aliannejad, R., Rezaei-Tavirani, M., Taheri, S. & Oskouie, A. A. The metabolomics of airway diseases, including COPD, asthma and cystic fibrosis. *Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals* **20**, 5-16, doi:10.3109/1354750x.2014.983167 (2015).

25 Dunn, W. B. *et al.* Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **9**, 44-66 (2013).

26 Müller, S., Scealy, J. L. & Welsh, A. H. Model selection in linear mixed models. *Statistical Science* **28**, 135-167 (2013).

Appendix 1: List of Metabolites Detected by AbsoluteIDQ® p180 Kit

Note: Metabolites used in correlation and linear analysis are marked by an asterisk (*)

Class	Compound
Amino acids & Biogenic amines	Alanine (Ala)
Amino acids & Biogenic amines	Arginine (Arg)
Amino acids & Biogenic amines	Asparagine (Asn)
Amino acids & Biogenic amines	Aspartate (Asp)
Amino acids & Biogenic amines	Citrulline (Cit)
Amino acids & Biogenic amines	Glutamine (Gln)
Amino acids & Biogenic amines	Glutamate (Glu)
Amino acids & Biogenic amines	Glycine (Gly)*
Amino acids & Biogenic amines	Histadine (His)
Amino acids & Biogenic amines	Isoleucine (Ile)
Amino acids & Biogenic amines	Leucine (Leu)
Amino acids & Biogenic amines	Lysine (Lys)
Amino acids & Biogenic amines	Methionine (Met)
Amino acids & Biogenic amines	Ornithine (Orn)
Amino acids & Biogenic amines	Phenylalanine (Phe)
Amino acids & Biogenic amines	Proline (Pro)
Amino acids & Biogenic amines	Serine (Ser)
Amino acids & Biogenic amines	Threonine (Thr)

Class	Compound
Amino acids & Biogenic amines	Tryptophan (Trp)
Amino acids & Biogenic amines	Tyrosine (Tyr)
Amino acids & Biogenic amines	Valine (Val)
Amino acids & Biogenic amines	Acetylornithine (Ac-Orn)
Amino acids & Biogenic amines	Asymmetric dimethylarginine (ADMA)
Amino acids & Biogenic amines	alpha-Aminoadipic acid (alpha-AAA)
Amino acids & Biogenic amines	cis-4-Hydroxyproline (c4-OH-Pro)
Amino acids & Biogenic amines	Carnosine
Amino acids & Biogenic amines	Creatinine
Amino acids & Biogenic amines	DOPA
Amino acids & Biogenic amines	Dopamine
Amino acids & Biogenic amines	Histamine
Amino acids & Biogenic amines	Kynurenine
Amino acids & Biogenic amines	Methioninesulfoxide (Met-SO)
Amino acids & Biogenic amines	Nitrotyrosine (Nitro-Tyr)
Amino acids & Biogenic amines	Phenylethylamine (PEA)
Amino acids & Biogenic amines	Putrescine
Amino acids & Biogenic amines	Symmetric dimethylarginine (SDMA)
Amino acids & Biogenic amines	Serotonin
Amino acids & Biogenic amines	Spermidine
Amino acids & Biogenic amines	Spermine
Amino acids & Biogenic amines	trans-OH-Pro (t4-OH-Pro)

Class	Compound
Amino acids & Biogenic amines	Taurine*
Amino acids & Biogenic amines	total DMA
Acylcarnitines	Carnitine (C0)
Acylcarnitines	Acetylcarnitine (C2)*
Acylcarnitines	Propionylcarnitine (C3)
Acylcarnitines	Propenonylcarnitine (C3:1)
Acylcarnitines	Hydroxybutyrylcarnitine (C3-DC (C4-OH))
Acylcarnitines	Hydroxypropionylcarnitine (C3-OH)
Acylcarnitines	Butanoylcarnitine (C4)
Acylcarnitines	Butenylcarnitine (C4:1)
Acylcarnitines	Valerylcarnitine (C5)
Acylcarnitines	Tiglylcarnitine (C5:1)
Acylcarnitines	Glutaconylcarnitine (C5:1-DC)
Acylcarnitines	Glutarylcarnitine (Hydroxyhexanoylcarnitine) (C5-DC (C6-OH))
Acylcarnitines	Methylglutarylcarnitine (C5-M-DC)
Acylcarnitines	Hydroxyvalerylcarnitine (Methylmalonylcarnitine) (C5-OH (C3-DC-M))
Acylcarnitines	Hexanoylcarnitine (Fumarylcarnitine) (C6 (C4:1-DC))
Acylcarnitines	Hexenoylcarnitine (C6:1)

Class	Compound
Acylcarnitines	Pimelylcarnitine (C7-DC)
Acylcarnitines	Octanoylcarnitine (C8)
Acylcarnitines	Nonaylcarnitine (C9)
Acylcarnitines	Decanoylcarnitine (C10)
Acylcarnitines	Decenoylcarnitine (C10:1)
Acylcarnitines	Decadienylcarnitine (C10:2)
Acylcarnitines	Dodecanoylcarnitine (C12)
Acylcarnitines	Dodecenoylcarnitine (C12:1)
Acylcarnitines	Dodecanedioylcarnitine (C12-DC)
Acylcarnitines	Tetradecanoylcarnitine (C14)
Acylcarnitines	Tetradecenoylcarnitine (C14:1)
Acylcarnitines	Hydroxytetradecenoylcarnitine (C14:1-OH)
Acylcarnitines	Tetradecadienylcarnitine (C14:2)
Acylcarnitines	Hydroxytetradecadienylcarnitine (C14:2-OH)
Acylcarnitines	Hexadecanoylcarnitine (C16)
Acylcarnitines	Hexadecenoylcarnitine (C16:1)
Acylcarnitines	Hydroxyhexadecenoylcarnitine (C16:1-OH)
Acylcarnitines	Hexadecadienylcarnitine (C16:2)
Acylcarnitines	Hydroxyhexadecadienylcarnitine (C16:2-OH)
Acylcarnitines	Hydroxyhexadecanoylcarnitine (C16-OH)
Acylcarnitines	Octadecanoylcarnitine (C18)
Acylcarnitines	Octadecenoylcarnitine (C18:1)

Class	Compound
Acylcarnitines	Hydroxyoctadecenoylcarnitine (C18:1-OH)
Acylcarnitines	Octadecadienylcarnitine (C18:2)
Lysophosphatidylcholines	lysoPC a C14:0
Lysophosphatidylcholines	lysoPC a C16:0
Lysophosphatidylcholines	lysoPC a C16:1
Lysophosphatidylcholines	lysoPC a C17:0
Lysophosphatidylcholines	lysoPC a C18:0
Lysophosphatidylcholines	lysoPC a C18:1
Lysophosphatidylcholines	lysoPC a C18:2
Lysophosphatidylcholines	lysoPC a C20:3
Lysophosphatidylcholines	lysoPC a C20:4
Lysophosphatidylcholines	lysoPC a C24:0
Lysophosphatidylcholines	lysoPC a C26:0
Lysophosphatidylcholines	lysoPC a C26:1
Lysophosphatidylcholines	lysoPC a C28:0
Lysophosphatidylcholines	lysoPC a C28:1
Phosphatidylcholines	PC aa C24:0
Phosphatidylcholines	PC aa C26:0
Phosphatidylcholines	PC aa C28:1
Phosphatidylcholines	PC aa C30:0*
Phosphatidylcholines	PC aa C30:2
Phosphatidylcholines	PC aa C32:0*

Class	**Compound**
Phosphatidylcholines	PC aa C32:1*
Phosphatidylcholines	PC aa C32:2*
Phosphatidylcholines	PC aa C32:3
Phosphatidylcholines	PC aa C34:1*
Phosphatidylcholines	PC aa C34:2*
Phosphatidylcholines	PC aa C34:3*
Phosphatidylcholines	PC aa C34:4
Phosphatidylcholines	PC aa C36:0
Phosphatidylcholines	PC aa C36:1
Phosphatidylcholines	PC aa C36:2*
Phosphatidylcholines	PC aa C36:3*
Phosphatidylcholines	PC aa C36:4*
Phosphatidylcholines	PC aa C36:5
Phosphatidylcholines	PC aa C36:6
Phosphatidylcholines	PC aa C38:0
Phosphatidylcholines	PC aa C38:1
Phosphatidylcholines	PC aa C38:3
Phosphatidylcholines	PC aa C38:4
Phosphatidylcholines	PC aa C38:5
Phosphatidylcholines	PC aa C38:6
Phosphatidylcholines	PC aa C40:1
Phosphatidylcholines	PC aa C40:2

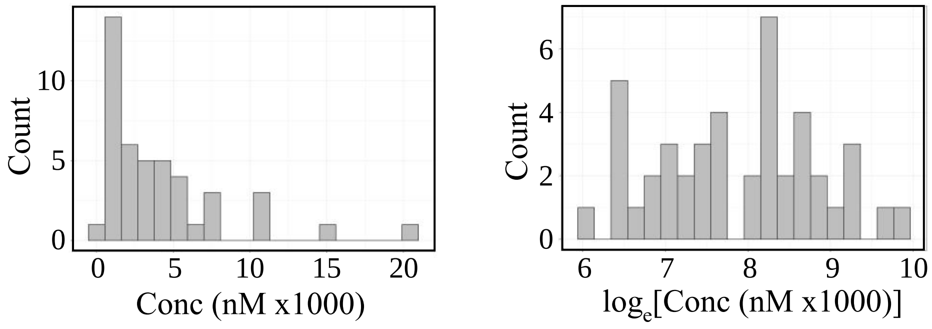
Class	**Compound**
Phosphatidylcholines	PC aa C40:3
Phosphatidylcholines	PC aa C40:4
Phosphatidylcholines	PC aa C40:5
Phosphatidylcholines	PC aa C40:6
Phosphatidylcholines	PC aa C42:0
Phosphatidylcholines	PC aa C42:1
Phosphatidylcholines	PC aa C42:2
Phosphatidylcholines	PC aa C42:4
Phosphatidylcholines	PC aa C42:5
Phosphatidylcholines	PC aa C42:6
Phosphatidylcholines	PC ae C30:0
Phosphatidylcholines	PC ae C30:1
Phosphatidylcholines	PC ae C30:2
Phosphatidylcholines	PC ae C32:1*
Phosphatidylcholines	PC ae C32:2
Phosphatidylcholines	PC ae C34:0*
Phosphatidylcholines	PC ae C34:1*
Phosphatidylcholines	PC ae C34:2
Phosphatidylcholines	PC ae C34:3
Phosphatidylcholines	PC ae C36:0
Phosphatidylcholines	PC ae C36:1
Phosphatidylcholines	PC ae C36:2

Class	Compound
Phosphatidylcholines	PC ae C36:3
Phosphatidylcholines	PC ae C36:4
Phosphatidylcholines	PC ae C36:5
Phosphatidylcholines	PC ae C38:0
Phosphatidylcholines	PC ae C38:1
Phosphatidylcholines	PC ae C38:2
Phosphatidylcholines	PC ae C38:3
Phosphatidylcholines	PC ae C38:4
Phosphatidylcholines	PC ae C38:5
Phosphatidylcholines	PC ae C38:6
Phosphatidylcholines	PC ae C40:1
Phosphatidylcholines	PC ae C40:2
Phosphatidylcholines	PC ae C40:3
Phosphatidylcholines	PC ae C40:4
Phosphatidylcholines	PC ae C40:5
Phosphatidylcholines	PC ae C40:6
Phosphatidylcholines	PC ae C42:0
Phosphatidylcholines	PC ae C42:1
Phosphatidylcholines	PC ae C42:2
Phosphatidylcholines	PC ae C42:3
Phosphatidylcholines	PC ae C42:4
Phosphatidylcholines	PC ae C42:5

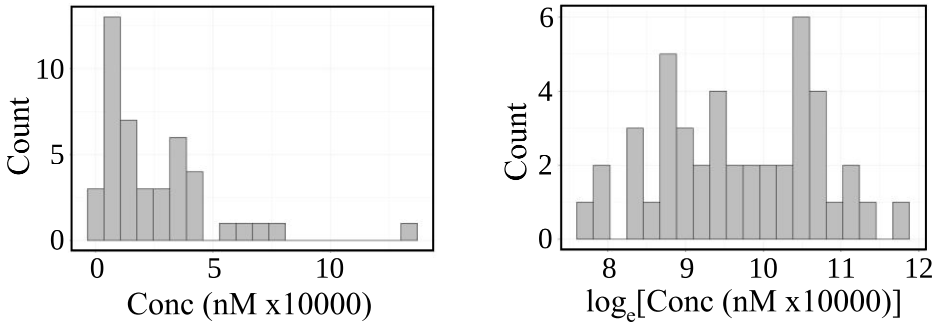
Class	**Compound**
Phosphatidylcholines	PC ae C44:3
Phosphatidylcholines	PC ae C44:4
Phosphatidylcholines	PC ae C44:5
Phosphatidylcholines	PC ae C44:6
Sphingomyelins	SM (OH) C14:1
Sphingomyelins	SM (OH) C16:1
Sphingomyelins	SM (OH) C22:1
Sphingomyelins	SM (OH) C22:2
Sphingomyelins	SM (OH) C24:1
Sphingomyelins	SM C16:0*
Sphingomyelins	SM C16:1
Sphingomyelins	SM C18:0
Sphingomyelins	SM C18:1
Sphingomyelins	SM C20:2
Sphingomyelins	SM C22:3
Sphingomyelins	SM C24:0
Sphingomyelins	SM C24:1
Sphingomyelins	SM C26:0
Sphingomyelins	SM C26:1
Hexoses	H1

Appendix 2: Histograms Comparing Distributions

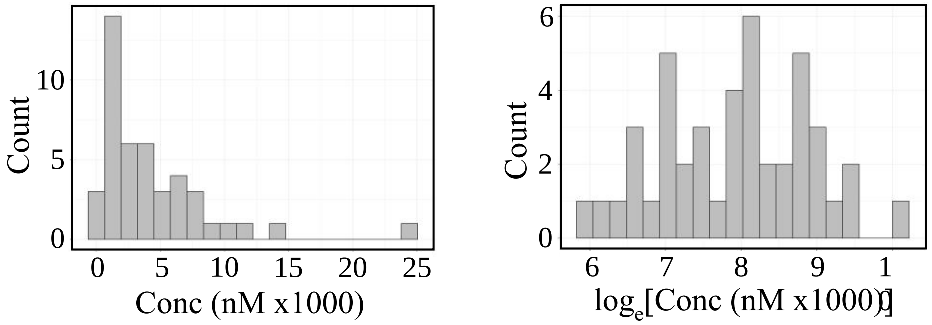
PC aa C30:0



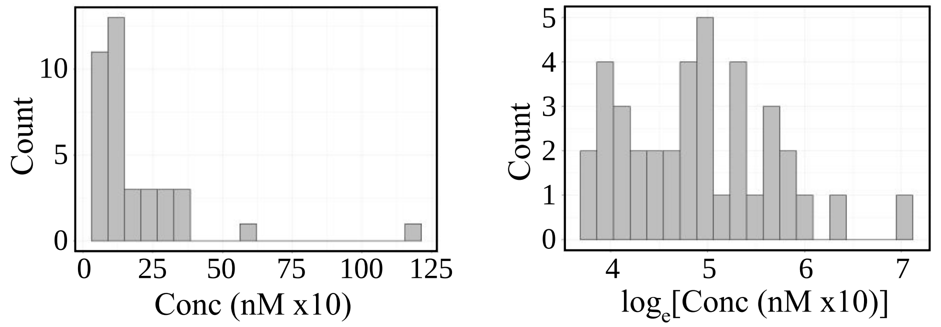
PC aa C32:0



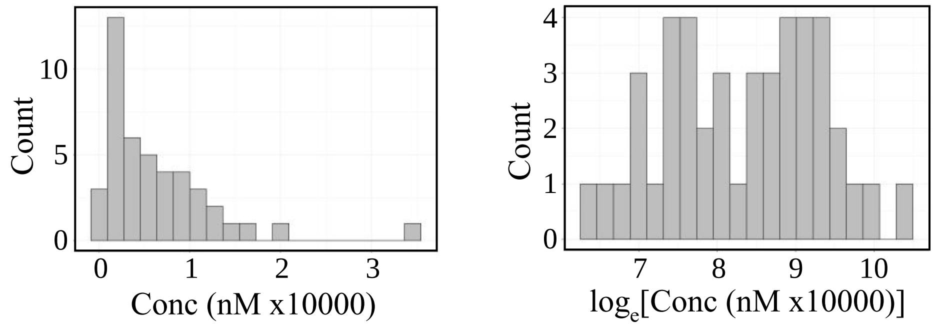
PC aa C32:1



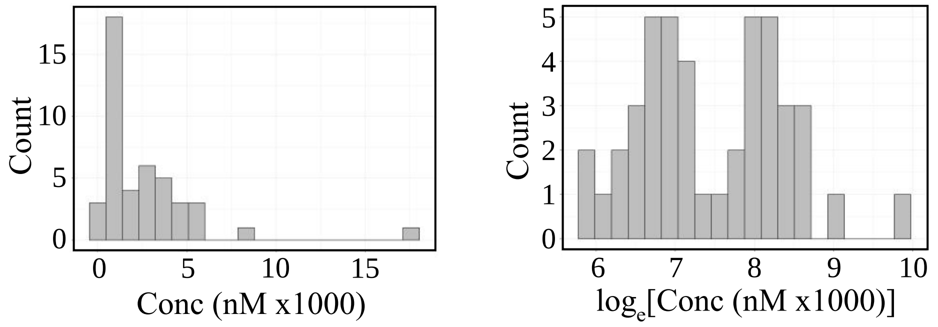
PC aa C32:2



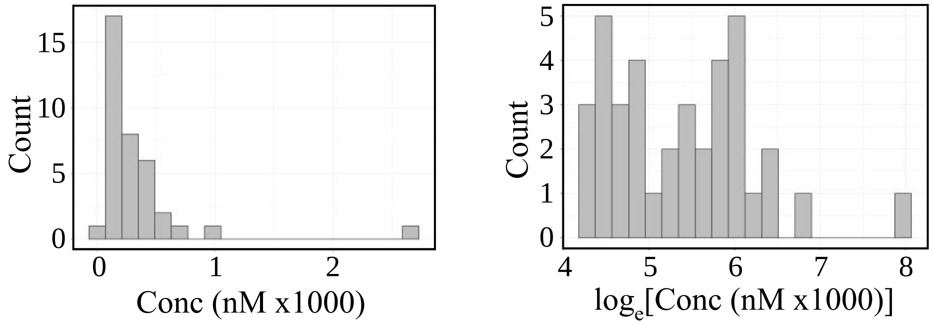
PC aa C34:1



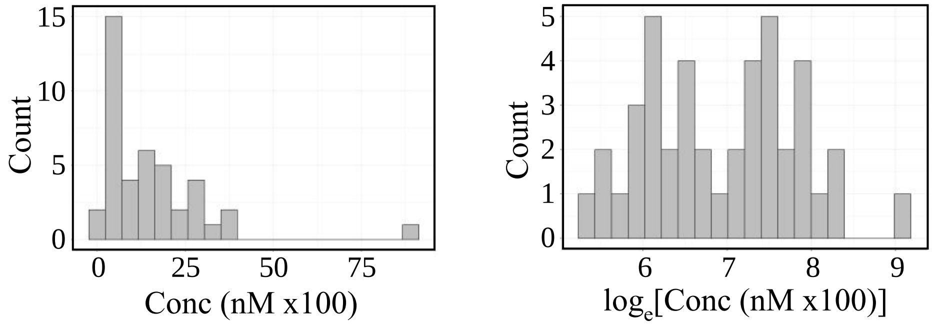
PC aa C34:2



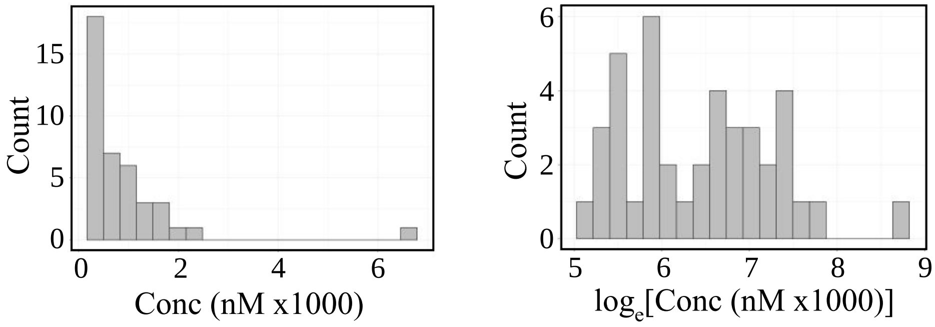
PC aa C34:3



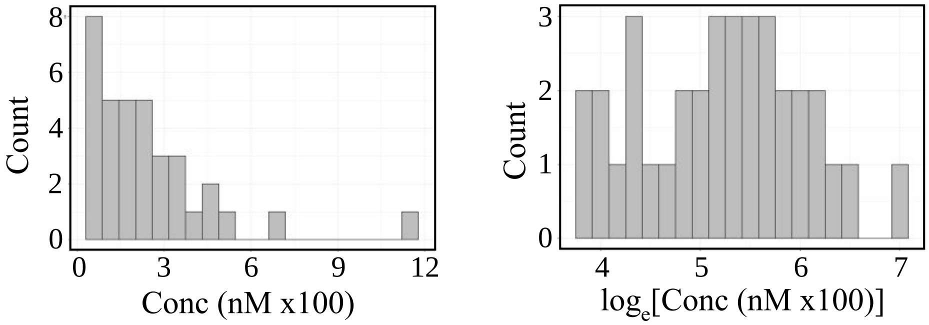
PC aa C36:2



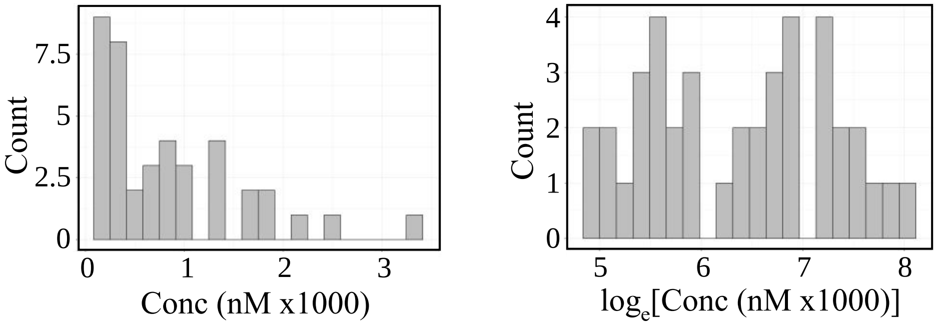
PC aa C36:3



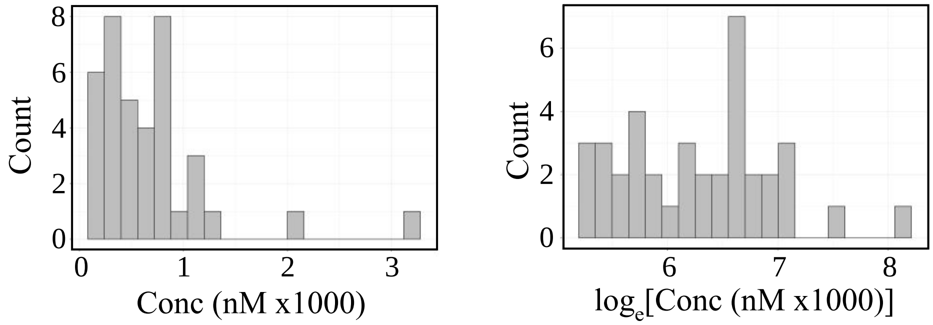
PC ae C32:1



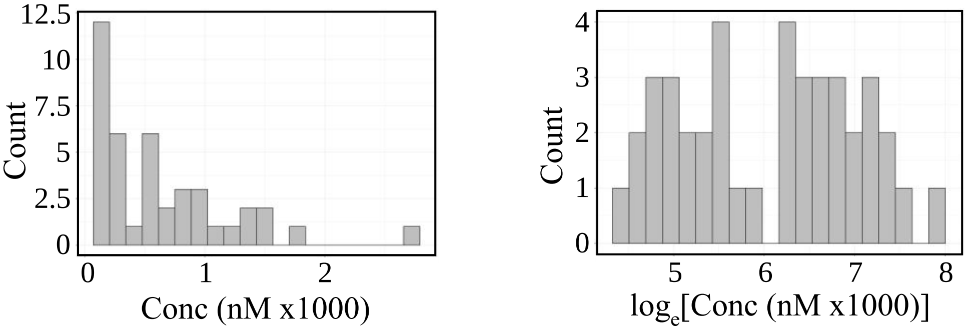
PC ae C34:0



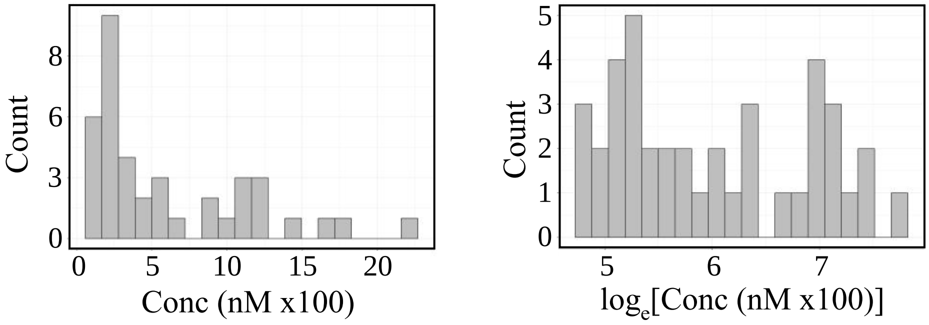
PC ae C36:4



PC ae C34:1

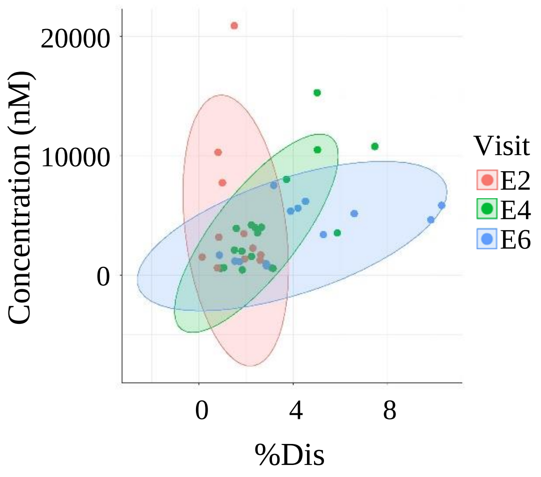


SM C16:0

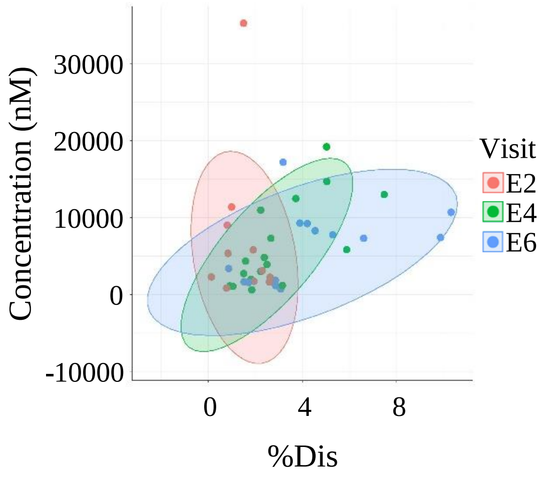


Appendix 3: Scatterplots of Significant Metabolites

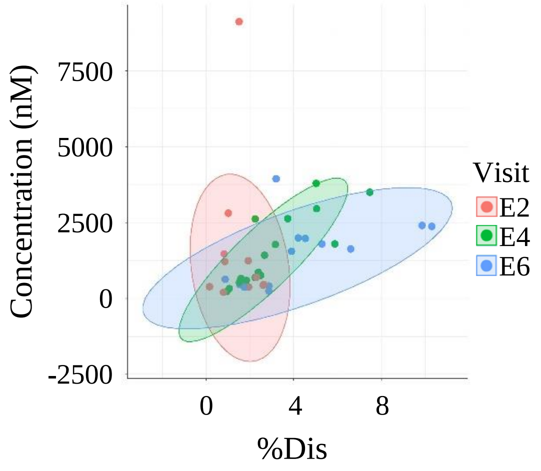
PC aa C30:0



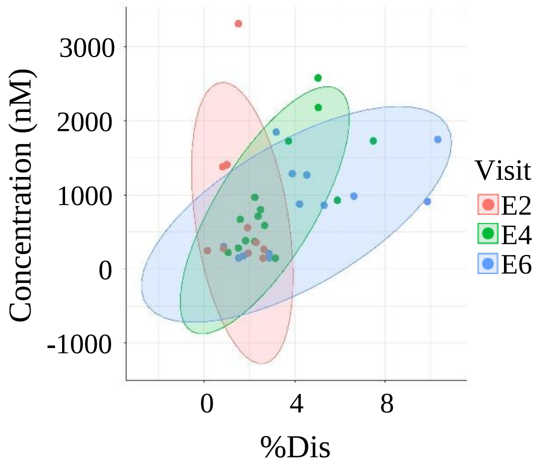
PC aa C34:1



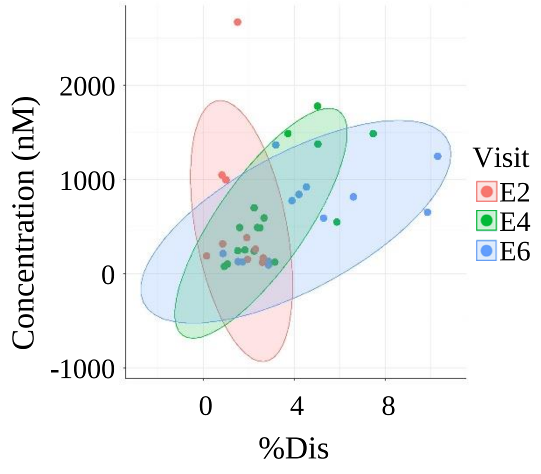
PC aa C36:2



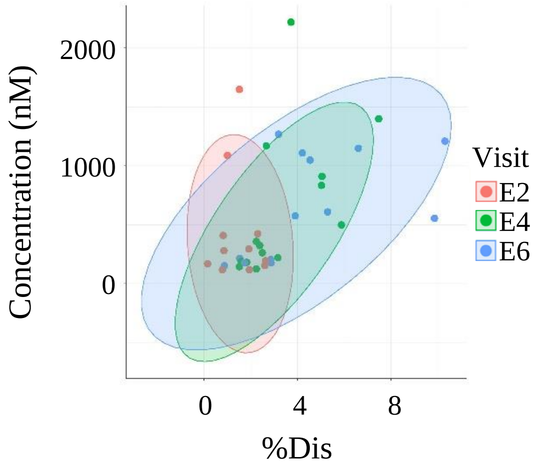
PC ae C34:0



PC ae C34:1



SM C16:0


Concentration (nM)
%Dis
Visit
● E2
● E4
● E6