Integrating plasma proteomics and airway microbiome data to identify multi-omic networks associated with cystic fibrosis pulmonary exacerbation treatment response

**Introduction**

Pulmonary exacerbations (PEx) are the leading cause of morbidity and decreased quality of life in cystic fibrosis (CF). Characterized by an acute decrease in lung function, PEx can occur repeatedly in people with CF and are associated with a long-term decline in lung function and shortened survival [1-3]. Intravenous (IV) antibiotic treatment is typically administered at the onset of a CF PEx in order to recover lost lung function. However, lung function is often not fully recovered after treatment [4-5]. There is therefore interest in characterizing the factors that contribute to incomplete PEx recovery. Identification of such determinants may improve our understanding of CF PEx disease mechanisms, provide evidence for the effectiveness and inadequacies of current CF therapies, as well as point to new therapeutic targets.

Blood and airway inflammatory biomarkers have been shown to decrease after treatment of a PEx, suggesting that airway infection in CF results in a robust host immune response [6-9]. Identifying associations between specific airway bacteria or bacterial communities and host-response may be critical to understanding the pathogenicity of CF bacteria. Further, exploring microbiota—protein relationships in the context of PEx treatment response may provide novel insights into the mechanisms of PEx and the efficacy of PEx treatments.

In this study, we use an extension of canonical correlation analysis (CCA) called sparse multiple canonical correlation network (SmCCNet) to integrate plasma proteomics and airway microbiome data from the onset of PEx (i.e., pre-treatment CF PEx samples) [10]. Specifically, these omics data are integrated into interpretable microbiota—protein networks that are associated with PEx treatment response. We use change in PEx score (ΔPExS) between hospital admission (i.e., PEx onset) and hospital discharge (i.e., late in PEx, post-treatment) as a clinical measure of PEx treatment response. Our aim was to identify potential biomarkers at PEx onset that are indicative of PEx recovery.

**Materials & Methods**

**Study Design & Population**

This study includes 34 PEx events from a cohort of 30 subjects aged 10 to 22 years old with a confirmed diagnosis of CF. Participants could be reenrolled in the study if PEx events were separated by at least six months. Accordingly, the study population includes four subjects with two PEx events and 26 subjects with one PEx event. Participants were recruited prospectively and enrolled at the time of hospital admission for IV antibiotic therapy of a clinically diagnosed PEx. A PEx was defined by a PEx score > 5 and at least three out of 11 PEx criteria being met, as defined by the CF foundation (CFF) Consensus Conference guidelines. Cohort characteristics, stratified by PEx time point, are described in Table 1. Subjects were treated with IV antibiotics targeting their specific CF pathogens as determined through microbial culture.

The study consists of two time points: hospital admission (i.e., PEx onset, day 0-2) and hospital discharge (i.e., after the IV treatment had been administered, day 4-21). Study procedures at each visit included a physical, a spirometry test following standard American Thoracic Society guidelines [ref], a standardized PEx score [ref], a validated quality-of-life measure, and specimen collection (blood plasma and sputum).

**Sample Collection & Processing**

*Blood Samples*

Blood was collected and analyzed for complete blood count (CBC), C-reactive protein (CRP), erythrocyte sedimentation rate (ESR) and calprotectin (S100A8/9; ALPCO-Buhlman, LLD 0.4mcg/mL). An aliquot of 3.5 mL was processed following a standard operating procedure, frozen and batch shipped on dry ice to SomaLogic in Boulder, Colorado for proteomics analysis.

*Airway Samples*

Spontaneously expectorated sputum was collected into a sterile container for microbiologic analysis. Participants unable to spontaneously expectorate underwent sputum induction using a standardized, published protocol with 3% hypertonic saline [ref]. Standard CF culture including quantitative bacterial culture and fungal culture was performed on sputum (expectorated and induced) samples following CFF guidelines [ref]. Respiratory viral panel was performed on admission sputum specimen (Luminex Molecular Diagnostics X-Tag® RVP). Residual sputum samples were frozen at −70°C for microbiome analyses.

**Proteomics & Data Preprocessing**

*SomaLogic Proteomics Assay*

Proteomics data were measured using the SomaScan multiplex proteomics assay, an aptamer-based quantitative proteomic biomarker discovery platform. Approximately 4,000 SOMAmers® (i.e., aptamers or single-stranded deoxyribonucleic acid (ssDNA) molecules that bind specific protein targets) were measured in this assay. SOMAmers targeted 3,579 unique proteins.

*Data Preprocessing*

SomaScan measurements, reported in relative fluorescent units (RFUs), were normalized using internal hybridization controls added to the assay prior to hybridization. Normalization was performed to adjust for inter-sample, inter-plate, and inter-run variation. Normalized RFU values were transformed using a log base 2 transformation and standardized prior to downstream statistical analysis.

**Microbiome Sequencing & Data Preprocessing**

*16S rRNA Sequencing*

Bacterial profiles were determined by broad-range amplification and sequence analysis of 16S rDNA following previously described methods [ref]. Amplicons were generated using primers targeting approximately 300 base pairs of the V1/V2 variable region of the 16S rRNA gene. Illumina paired-end sequencing was performed on the MiSeq platform using a 500 cycle v2 reagent kit.

*Analysis of Illumina Paired-End Reads*

Quality control procedures were performed on paired-end sequences as described in the online supplement. Assembled sequences were aligned and classified with SINA (1.2.11) [ref] using the Silva 111 database [ref] as reference configured to yield the Silva taxonomy. Sorted paired-end sequence data were deposited in the NCBI Short Read Archive under accession number xxx. The software package Explicet (v2.10.5) [ref] was used for analysis of Good’s coverage and the calculation of richness, evenness, and Shannon diversity, a composite measure of evenness and richness, at the rarefaction point of xxx sequences.

*Data Preprocessing*

Microbiome count data were filtered to include only prevalent taxa. Detection and prevalence thresholds were set to 0.1% and 10%, respectively (i.e., taxa were filtered out if they did not exceed 0.1% relative abundance (RA) in at least 10% of samples). Count data were transformed using the centered log-ratio (CLR) transformation and standardized prior to downstream statistical analysis. A pseudocount (RAmin/2) was applied to exact zero RA entries before CLR transformation was performed.

**Clinical Phenotype Definition**

The phenotype of interest in this study is change (i.e., difference) in PEx score from hospital admission to hospital discharge. PEx scores were measured using a standardized scoring system [7]. Negative change in PEx score values represent an improved PEx score between hospital admission and discharge, with larger negative values indicating greater improvement (or PEx treatment response). The distribution of our phenotype of interest is shown in Supplemental Figure 1.

**Statistical Analysis**

*Statistical Software*

All statistical analysis work, including multi-omic integration, network identification with, gene ontology (GO) enrichment analysis, and sensitivity analysis, was performed using R version 4.1.1.

*SmCCNet*

Airway microbiome (Χ1) and plasma proteomics (Χ2) data from PEx onset were integrated using SmCCNet (version 0.99.0). SmCCNet is an extension of canonical correlation analysis (CCA) that incorporates a quantitative phenotype to construct interpretable multi-omic networks. Given our omic data types and phenotype of interest, we use SmCCNet to construct microbiota—protein networks at PEx onset that are correlated with PEx treatment response (i.e., ΔPExS, or Υ).

CCA is an unsupervised learning approach that aims to identify relationships between two multivariable data sets by finding the linear combination of variables that maximizes correlation between the sets. In the most general form, CCA seeks to maximize Cor(Χ1*w*1, Χ2*w*2) where *w*1 and *w*2 are canonical weights corresponding to Χ1 and Χ2, respectively. The canonical weight objective function is defined in Equation 1.

A picture containing text, font, white, algebra

Description automatically generated

SmCCNet uses an adaptation of CCA, called sparse multiple CCA (SmCCA), to incorporate a third data type (i.e., the phenotype of interest) into the integration task. SmCCA accomplishes multimodal integration by accounting for phenotype—omic correlation within the canonical weight objective function. Sparsity is imposed on the canonical weight vectors by further adding penalization terms to Equation 1 (functions *P*1(•) and *P*2(•)). The least absolute shrinkage and selection operation (LASSO) is used within the SmCCA implementation used by SmCCNet. The extended canonical weight objective function that allows for sparse, multivariate integration is defined in Equation 2. In this objective function, *c*1 and *c*2 are the pre-specified penalty constants and *a*, *b*, and *c* are scaling constants that can be used to adjust the importance of specific bivariate relationships when determining canonical weights (i.e., microbiota—protein, microbiome—ΔPExS, or protein—ΔPExS correlation). Shi et al suggest that weighted SmCCA (i.e., the case in which *a*, *b*, and *c* are not all equal) might be preferred when the correlation between phenotype and omic data is weaker than the correlation between the omic data sets.

**A picture containing text, font, white, algebra

Description automatically generated**

SmCCNet uses SmCCA and a feature subsampling approach to obtain a set of canonical weight vectors, each represented by *u* = (*w*1, *w*2). The canonical weight vectors are used to construct a similarity matrix *A* following *A* = *u* ⊗ *u* and subsequently a dissimilarity (or distance) matrix following *D* = 1 − *A*. *D* is used as input for hierarchical clustering and the resultant hierarchical tree is cut to form distinct networks that are associated with the phenotype of interest. 500 feature subsampling iterations and default parameters were used for SmCCNet analysis unless stated otherwise.

*SmCCNet Hyperparameter Tuning*

As mentioned, SmCCNet induces sparsity into the CCA framework by incorporating two LASSO-based penalty parameters (i.e., one penalty parameter for each of Χ1 and Χ2; *c*1 and *c*2, respectively). Penalty parameters were selected using five-fold cross-validation (CV) and a randomized grid search approach. The selected penalty pair corresponds to the penalty pair that minimizes the prediction error between the CV training and test sets. Unlike traditional LASSO-based methods, increasing the value of the penalty parameter *c* decreases the amount of regularization applied to a given omic set. We searched a range of larger values for *c*1 (ranging from 0.40 to 0.60) and a range of larger values for *c*2 (ranging from 0.10 to 0.40) since the dimensionality of Χ1 (60 taxa) was much smaller than the dimensionality of Χ1 (4,001 SOMAmers). The aim was to impose a greater penalization on Χ2 in order to generate networkswith an improved balance of proteins and taxa. We further used imbalanced feature subsampling proportions (i.e., 0.90 and 0.70 for Χ1 and Χ2, respectively) to additionally account for the feature imbalance between omics sets. Ten feature subsampling iterations were used during CV due to computational limitations.

SmCCA objective function scaling coefficients *b* and *c* were additionally tuned using a manual hyperparameter tuning process similar to the one used by Mastej et al [ref]. We explored this *weighted* version of SmCCNet since the correlation between omic sets was found to be stronger than the correlation between the phenotype and each of the omic sets (Supplemental Figure 2). Separate five-fold CVs were performed to select penalty parameters for each of the tested weighting schemes. We considered the number of networks, the strength of network–phenotype correlations, network sizes (i.e., number of nodes, or features), and protein–microbiota balance when determining the optimal weighting scheme (Supplemental Table 1). The aim was to generate robust and reasonably sized multi-omic networks that are strongly correlated with PEx treatment response (ρ > 0.30).

*Network Summarization*

The association between each network and ΔPExS was assessed by performing principal component analysis (PCA) and calculating the correlation between ΔPExS and the first principal component (PC1) of the network-specific matrix of taxa and proteins (i.e., a merged matrix of subnetwork-specific taxa and proteins in Χ1 and Χ2, respectively). Absolute correlation values are reported since the use of PC1 obscures the interpretability of negative and positive correlations.

*Network Pruning*

The maximum node capacity of each network was set to 200 nodes, which represents just less than 5% of the total feature space. The goal of setting this limit was to preserve interpretability. Further, we speculated that not every node in a large network (i.e., a network with more than 200 nodes) contributed heavily to network—phenotype relatedness. We therefore carried out the following network pruning procedure which aims to both yield reasonable network sizes and maximize network—phenotype correlations. First, nodes are ranked by importance using the PageRank algorithm [ref]. Generally, the PageRank algorithm measures node importance by assessing the number and importance of connected nodes. Second, we determine network—phenotype correlations when including 100 to 200 of the top ranked nodes. Third, and finally, we select *k* (i.e., the optimal number of nodes) by determining the *k* top-ranked nodes that result in maximum correlation with the phenotype. This procedure was only performed on networks with more than 200 nodes.

*Network Visualization*

Networks were visualized using CytoScape (version 3.9.1). The following additional pruning steps were applied to networks prior to visualization in order to generate interpretable visualizations reflective of the underlying biology. First, networks often contain a set of insubstantial node-to-node connections (or edges) that might incorrectly suggest erroneous associations. Our intention was to remove these edges to improve biological interpretability. We used between-node pairwise correlations for systematic removal of weak edges, removing edges between nodes that were weakly correlated (ρ < 0.30). Second, interpretation of large network visualizations is difficult due to the number of nodes, edges, and network attributes. We aimed to reduce large networks into their core components to create coherent and comprehensible visualizations. To accomplish this, we determined the maximum weight connected to every node. Between-node weights, which are calculated in the previously described *A* matrix, are reflective of the mutual importance of two nodes in maximizing the correlation between all data sets (Χ1, Χ2, and Υ). We arbitrarily selected all nodes associated with a maximum weight of greater than 0.20 for visualization. This pruning step was only performed for large networks (e.g., Network 1 in Figure 1).

*Gene Ontology Enrichment Analysis*

GO enrichment analysis was performed on select networks using network-specific protein/gene sets and the clusterProfiler R package (version 4.2.2). The aim was to identify over-represented GO pathways (specifically biological processes, or BP) associated with networks of interest. The full set of proteins/genes targeted by the SomaScan proteomics assay was used as the background gene list. *org.Hs.eg.db* was used as the human genome wide annotation data base.

*Network Sensitivity Analysis*

Work in progress.

**Results**

A picture containing text, screenshot, font, number

Description automatically generated

**Table 1.** Study cohort demographics and clinical measures at hospital admission. ΔPExS, the phenotype of interest in this study, is reported as well. ΔPExS represents change in PEx Score from hospital admission to hospital discharge. Continuous demographic and clinical information is reported as median [range].

*Selected SmCCNet Configuration & Networks*

A weighting scheme of (1, 2, 2) (i.e., *a* = 1, *b* = *c* = 2) with a CV selected penalty parameter pair of (0.25, 0.40) was found to generate the most desirable networks when considering the following network characteristics: network–phenotype correlation, network size (i.e., number of nodes), and protein–microbiota balance (Supplemental Table 1). Eight networks, summarized in Figure 1, resulted from this SmCCNet configuration, including six networks with a strong correlation to ΔPExS (ρ > 0.30). Comprehensive analyses are provided for Networks 1 and 3 in the subsequent sections, including network visualizations, GO enrichment analyses, and network sensitivity analyses. Network 1 is a large and robust network comprised of many bacteria that are not typically found in CF airways. Contrastingly, Network 3 is a smaller network comprised of traditional CF pathogens that we might expect.

A screenshot of a graph

Description automatically generated with low confidence

**Figure 1.** SmCCNet network summarizations for the selected SmCCNet configuration. The scatterplot (left) shows correlations between individual network nodes and ΔPExS. Protein nodes are represented as blue circles and taxon nodes are represented as pink triangles. The heatmap (right) reports the absolute correlation observed between a given network’s PC1 and ΔPExS. Bold typeface is used to specify the networks that are analyzed in-depth.

*The Traditional CF Pathogen Network*

The *Traditional CF Pathogen* *Network* (Network 3 in Figure 1) is a small network consisting of 12 nodes (2 taxa and 10 proteins). Taxa include *Pseudomonas* and *Prevotella*.

*The Robust Network*

A picture containing diagram, line, circle

Description automatically generated

**Figure 2.** The *Traditional CF Pathogen Network* identified by SmCCNet. Protein nodes are represented as circles and taxon nodes are represented as triangles. A divergent color palette is used for node color. Dark shades of brown and green indicate strong negative and positive correlations between the node and ΔPExS, respectively. Lighter shades of brown and green indicate weak negative and positive correlations between the node and ΔPExS. Node size corresponds to the absolute correlation observed between a node and ΔPExS. Edge thickness corresponds to between-node canonical weight-based connectivity (i.e., pairwise weights found in matrix *A*). Positively associated nodes are connected by purple edges and negatively associated nodes are connected by grey edges.

**A picture containing circle, diagram, line, design

Description automatically generated**

**Figure 3.** The *Robust* *Network* identified by SmCCNet. Protein nodes are represented as circles and taxon nodes are represented as triangles. A divergent color palette is used for node color. Dark shades of brown and green indicate strong negative and positive correlations between the node and ΔPExS, respectively. Lighter shades of brown and green indicate weak negative and positive correlations between the node and ΔPExS. Node size corresponds to the absolute correlation observed between a node and ΔPExS. Edge thickness corresponds to between-node canonical weight-based connectivity. Positively associated nodes are connected by red edges and negatively associated nodes are connected by blue edges. Weakly connected nodes are indicated by thin and near-white edges.

*Sensitivity Analysis*

Work in progress.

**Discussion**

**Limitations**

* Assuming independent observations even though we have repeats
* Cohort limitations (sample size, generalizability)

**Supplementary Materials**

**Supplemental Figures**

Supplemental Figure 1.

**A picture containing text, screenshot, rectangle, line

Description automatically generated**

Supplemental Figure 2.

**A picture containing diagram, line, plot, origami

Description automatically generated**

**Supplemental Tables**

Supplemental Table 1.

**A screenshot of a table

Description automatically generated with low confidence**

**Citations**

1. Amadori A, Antonelli A, Balteri I, Schreiber A, Bugiani M, De R, V. Recurrent exacerbations affect FEV(1) decline in adult patients with cystic fibrosis. Respir Med. 2009 Mar;103(3):407-13
2. Liou TG, Adler FR, Fitzsimmons SC, Cahill BC, Hibbs JR, Marshall BC. Predictive 5-year survivorship model of cystic fibrosis. Am.J.Epidemiol. 2001 Feb 15;153(4):345-52
3. Sanders DB, Bittner RC, Rosenfeld M, Redding GJ, Goss CH. Pulmonary exacerbations are associated with subsequent FEV1 decline in both adults and children with cystic fibrosis. Pediatr.Pulmonol. 2011 Apr;46(4):393-400. PMCID:20967845
4. Sanders DB, Bittner RC, Rosenfeld M, Hoffman LR, Redding GJ, Goss CH. Failure to recover to baseline pulmonary function after cystic fibrosis pulmonary exacerbation. Am.J.Respir.Crit Care Med. 2010 Sep 1;182(5):627-32
5. Sanders DB, Hoffman LR, Emerson J, Gibson RL, Rosenfeld M, Redding GJ, Goss CH. Return of FEV1 after pulmonary exacerbation in children with cystic fibrosis. Pediatr.Pulmonol. 2010 Feb;45(2):127-34
6. Ordonez CL, Henig NR, Mayer-Hamblett N, Accurso FJ, Burns JL, Chmiel JF, Daines CL, Gibson RL, McNamara S, Retsch-Bogart GZ, et al. Inflammatory and microbiologic markers in induced sputum after intravenous antibiotics in cystic fibrosis. Am.J.Respir.Crit Care Med. 2003 Dec 15;168(12):1471-5
7. Colombo C, Costantini D, Rocchi A, Cariani L, Garlaschi ML, Tirelli S, Calori G, Copreni E, Conese M. Cytokine levels in sputum of cystic fibrosis patients before and after antibiotic therapy. Pediatr.Pulmonol. 2005 Jul;40(1):15-21
8. Chiron R, Grumbach YY, Quynh NV, Verriere V, Urbach V. Lipoxin A(4) and interleukin-8 levels in cystic fibrosis sputum after antibiotherapy. J.Cyst.Fibros. 2008 Nov;7(6):463-8
9. Sagel, S. D., Thompson, V., Chmiel, J. F., Montgomery, G. S., Nasr, S. Z., Perkett, E., Saavedra, M. T., Slovis, B., Anthony, M. M., Emmett, P., & Heltshe, S. L. (2015). Effect of treatment of cystic fibrosis pulmonary exacerbations on systemic inflammation. *Annals of the American Thoracic Society*, *12*(5), 708–717. https://doi.org/10.1513/AnnalsATS.201410-493OC
10. Shi, W. J., Zhuang, Y., Russell, P. H., Hobbs, B. D., Parker, M. M., Castaldi, P. J., et al. (2019). Unsupervised Discovery of Phenotype-specific Multi-Omics Networks. Bioinformatics 35, 4336–4343. doi:10.1093/bioinformatics/btz226
11. Rosenfeld M, Emerson J, Williams-Warren J, Pepe M, Smith A, Montgomery AB, Ramsey B. Defining a pulmonary exacerbation in cystic fibrosis. J.Pediatr. 2001 Sep;139(3):359-65