bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Microbiome data enhances predictive models of lung function in people with CF

Conan Y. Zhao[1,2], Yiqi Hao[1], Yifei Wang[1,2,3], John J. Varga[1,2,4,5], Arlene A. Stecenko[4,5], Joanna B. Goldberg[4,5], Sam P. Brown[1,2]

Institutions:

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA

[2]Center for Microbial Dynamics and Infection, Georgia Institute of Technology, Atlanta, GA, USA

[3]The Institute for Data Engineering and Science (IDEaS), Georgia Institute of Technology, Atlanta, GA, USA

[4]Division of Pulmonary, Allergy/Immunology, Cystic Fibrosis and Sleep, Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA

[5]Emory + Children's Center for Cystic Fibrosis and Airway Disease Research, Atlanta, GA, 30322, USA

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The polymicrobial context of chronic infection has received increasing attention due to widespread use of microbiome sequencing technology. However, clinical microbiology analysis of infection samples in hospitals continues to focus only on established human pathogens. This disconnect between diverse 'infection microbiomes' and limited clinical microbiology profiling leaves open the possibility that important risk markers are being unexploited during infection management. To address this disconnect, we focus on lung infections in people with Cystic Fibrosis (CF). A cohort of CF patients (N=77) were recruited for this study. We collected health information (age, BMI, lung function) and clinical microbiology records for each patient. We also collected sputum samples during a period of clinical stability, and determined lung microbiome compositions through 16S rDNA sequencing. We use a regularized linear regression algorithm (ElasticNet) to select informative features to predict lung function. We find that models including whole microbiome quantitation outperform models trained on pathogen quantitation alone, with or without the inclusion of patient metadata. Our most predictive models retain key pathogens as negative predictors (*Pseudomonas*, *Achromobacter*) along with established correlates of CF disease state (age, BMI, CF related diabetes). In addition, our models select specific non-pathogen taxa (*Fusobacterium*, *Rothia*) as positive predictors of lung health. Our analysis does not address causality, leaving open whether these non-pathogen taxa are playing an active role in promoting lung health (e.g. by suppressing pathogens), or are simply informative biomarkers of patient health (orthogonal to age, BMI, etc). Our results support a reconsideration of clinical microbiology pipelines to ensure the provision of the most informative data to guide clinical practice.

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Bacterial infections in otherwise healthy people are often rapidly resolved by effective immune responses, independent of antibiotic treatment. In some cases however, infections fail to clear even with appropriate drug treatment, permitting the establishment of chronic (long-lasting) infection and imposing elevated morbidity and mortality risk on affected individuals.[1] Chronic infections are a rising burden on global health-care systems as populations at risk (e.g. diabetics) grow.[2] Deficits in host barrier defenses and/or immune function in these at-risk people provide an opening for the establishment of infection, which are further compounded by changes in pathogen growth mode (e.g. biofilm formation[3] and the accumulation of other pathogens to form complex multispecies communities.[4]

The polymicrobial context of chronic infection has received increasing attention due to advances in microbiome sequencing technology. However, clinical microbiology analysis of infection samples in hospitals continues to focus only on the 'usual suspects' of established human pathogens – a relatively short list of organisms for which there is a long established literature on risk to patient health. This disconnect between diverse 'infection microbiomes' and limited clinical microbiology profiling leaves open the possibility that important risk markers are being unexploited during infection management.

To address this potential disconnect, we focus on chronic lung infections in people with cystic fibrosis (CF). Cystic fibrosis is an autosomal recessive disease characterized by defective lung mucociliary clearance and an accumulation of viscous mucus in the patient's lung.[5–7] This environment provides both nutrients for bacterial growth and protection from host immune responses,[8–11] facilitating long-term microbial infections.[12–15] CF-lung infections have historically been studied as single-species phenomena, focused on a few key pathogens that are routinely identified by clinical microbiology labs (e.g. *Pseudomonas aeruginosa* and *Staphylococcus aureus*). However, the advent of inexpensive 16S rDNA sequencing has caused a major shift in CF lung microbiology research. Sequencing of expectorated sputum samples has revealed diverse communities of tens to hundreds of taxa, including numerous non-pathogenic bacteria.[13,16]

CF lung microbiome studies have linked lung microbiome composition to disease progression and overall patient health[17,18] and found three key patterns: (1) severe disease is associated with pathogen dominance and loss of microbiome diversity in cross-sectional studies;[17–19] (2) loss of microbiome diversity correlates with declining lung function in longitudinal studies;[20] (3) prevalence of non-pathogenic fermentative anaerobes (*Veillonella*, *Prevotella*, *Fusobacterium*) is associated with higher lung function.[21,22]

While these correlative observational results are supported across multiple studies, their causal interpretation is the subject of some controversy. One line of argument proposes that these results reflect community ecological processes at play within the lung, where networks of facilitatory and inhibitory

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interactions among species govern community structure and subsequent harm to the host.[12,23,24] Conversely, the counter argument is that these patterns are simply the result of oral anaerobe contamination during sample collection.[25,26] Under this contamination model, increasing pathogen load compared to a constant background of oral microbiome contamination generates a spurious link between oral microbes, microbiome diversity, and patient health, assuming a causal relationship between pathogen burden and health.[25] These conflicting hypotheses highlight the uncertainty in the role of taxa present in sputum, and the limitations of observational studies in establishing causal inference.

In the current study, we side-step the question of causal inference, and instead focus on the degree to which expectorated sputum microbiome data (inclusive of potential oral contaminants) is informative of patient lung health. We hypothesize that the addition of non-pathogen data provides additional information that can improve the predictability of patient health outcomes, compared to established pathogen data alone. To address this hypothesis we apply machine learning tools to an integrated lung microbiome and electronic medical record dataset for a cohort of 77 CF patients. We find that compared to the benchmark of pathogen data alone, prediction of lung function was improved by the addition of non-pathogen taxa.

Results

Clinical and microbiome data summary

In total, we obtained sputum expectorates from 77 CF patients with varying lung function. We measure lung health by percent predicted forced expiratory volume in 1 second (ppFEV1), and stratify ppFEV1 into four categories from Normal to Severe; a summary of patient information is presented in Table 1. As expected, age varies significantly across lung function category (ANOVA; $p < 0.01$). Blood glucose control (HbA1c levels) and bacterial load (log-scaled) are also significantly associated with lung function (ANOVA, $p < 0.01$, $p < 0.05$ respectively). Unsurprisingly, culture-based detection of *Pseudomonas aeruginosa* also strongly associated with lung function (ANOVA, $p < 0.001$).

	Normal	Mild	Moderate	Severe	P-value
N	23	15	25	14	
ppFEV1*	101.2	74.9	46.5	32.9	
(range)	(80.4-119.5)	(61.6-79.8)	(40.8-59.6)	(19.7-39.2)	
Age*	20	24	32	31.5	0.0068
(range)	(9-66)	(17-51)	(10-63)	(21-61)	
Male	11	7	11	6	

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

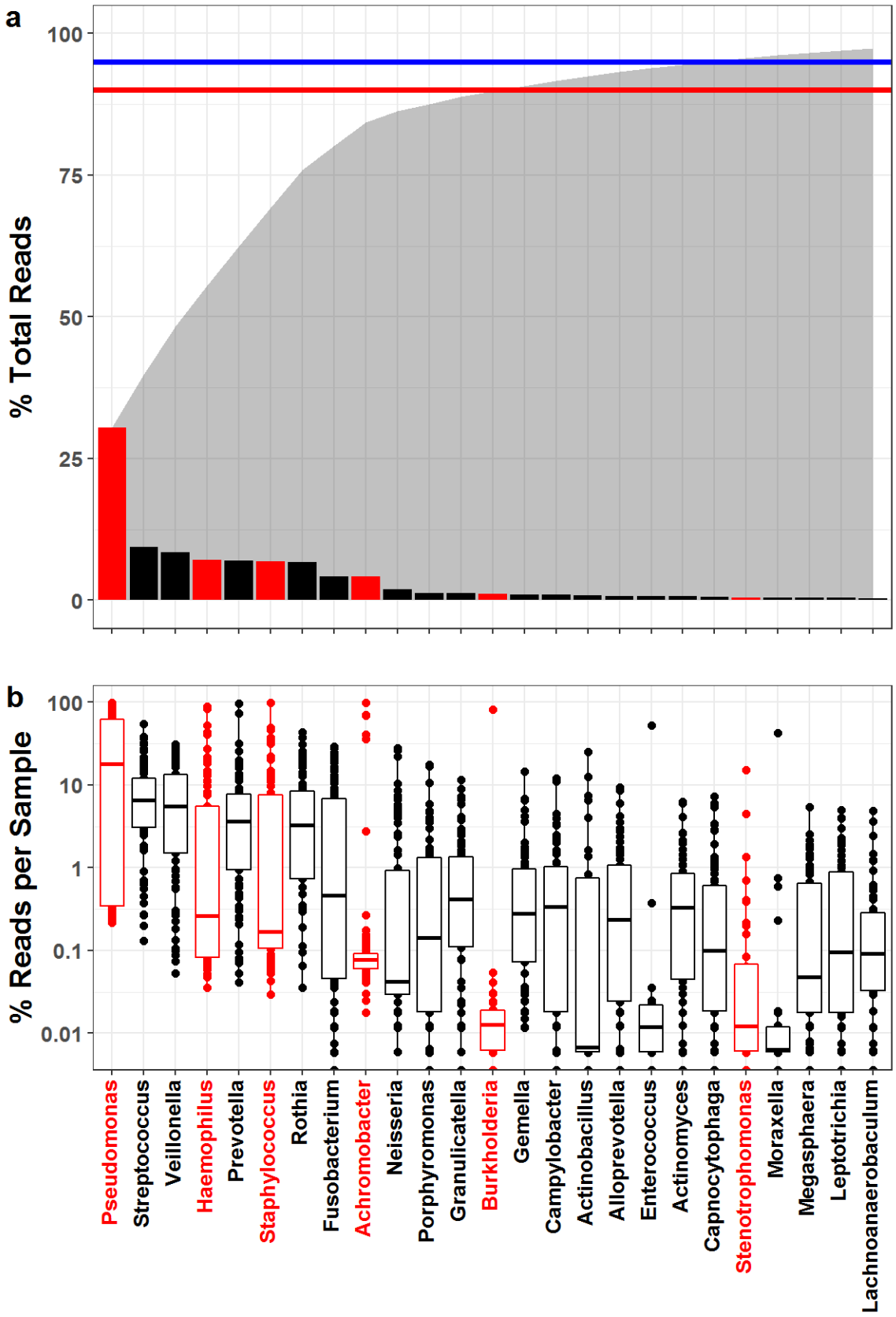
CFTR Genotype					
Homo-dF508	13	7	12	5	
Hetero-dF508	10	8	10	9	
Other/other	0	3	0	0	
BMI*	21.51	22.23	20.73	19.43	0.0943
(range)	(16.65-33.91)	(19.38-26.07)	(16.70-29.81)	(16.27-25.69)	
CF-related diabetes (CFRD)	6	8	14	11	
(%)	(26.1)	(53.3)	(56.0)	(78.6)	
HbA1c*	5.5	5.7	5.9**	6.25	0.00933
(range)	(5.1-7.1)	(5.0-7.6)	(4.9-8.4)	(5.3-11.9)	
Clinical Micro					
PA (%)	5 (21.7)	10 (66.7)	20 (80.0)	10 (71.4)	0.000119
SA (%)	16 (69.6)	10 (66.7)	12 (48.0)	8 (57.1)	0.454
MRSA (%)	4 (17.4)	6 (40.0)	6 (24.0)	4 (28.6)	0.486
BC (%)	0 (0.0)	1 (6.7)	1 (4.0)	0 (0.0)	0.553
AX (%)	2 (8.7)	0 (0.0)	1 (4.0)	3 (21.4)	0.147
STE (%)	2 (8.7)	3 (20.0)	1 (4.0)	0 (0.0)	0.191

Table 1. Summary of patient clinical data, stratified by lung function. *Lung function classes are defined as follows: Normal (ppFEV1 > 80); Mild (60 < ppFEV1 ≤ 80); Moderate (40 < ppFEV1 ≤ 60); and Severe (ppFEV1 < 40). Quantitative metrics are reported using the median and ranges. *Median reported values. **Two patients did not have reported HbA1c values. Significant differences between lung function categories tested by ANOVA, p-values shown.*

Turning to our 16S microbiome analyses, we found that the majority (>90%) of reads mapped to one of 13 genera (Fig 1a), highlighting the relatively species-poor nature of sputum microbiome. Fig 1a illustrates that these 13 dominant genera are a mixture of recognized CF pathogens (red) and orally derived bacteria (black).

The predominant pathogenic species was *Pseudomonas aeruginosa*, accounting for 30.4% of all reads. PA sequences were detected in every patient sample. Other established CF pathogens (*Staphylococcus*, *Achromobacter*, *Haemophilus*, *and Burkholderia*) collectively represented a further 19.3% of all reads, while oral taxa account for over 45% of all reads (Fig 1). Total pathogen and oral taxa abundance were both found to vary significantly ($p \ll 0.001$) with lung function (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig 1. CF lung microbiome is species poor, and dominated by oral anaerobes and opportunistic pathogens *We analyzed CF sputum expectorate (N=77) using 16S sequencing and an in-house QIIME 2-based bioinformatics pipeline to resolve strain-level OTUs. Samples were rarefied to 17000 reads. We identified 217 OTUs across 81 species and 59 genera. Overall, we find that CF sputum samples are species-poor, dominated by oral anaerobes and opportunistic pathogens.* ***a)*** *Sequences mapped to 14 genera comprised 90% (red line) of the total reads obtained. 95% (blue line) of all reads mapped to 21 genera. Total cumulative read fraction represented in shaded region. Pseudomonas was the most prevalent genus, followed by Streptococcus and Veillonella.* ***b)*** *Binning reads by sample shows variation in relative abundance. Pseudomonas comprises >10% of reads in the majority of our samples. While over 6% of the total reads mapped to Achromobacter, only 4 samples were comprised of >10% Achromobacter.*

Microbiome Composition Correlates with Lung Function

To explore the relationship between the composition of sputum samples and patient health, we next analyzed microbiome composition across broad lung function categories. Fig 2a shows composition plots highlighting the relative abundance of six canonical CF pathogens. As expected, *Pseudomonas* was more prevalent in people with severe lung function, whereas samples from people with normal lung function contain more *Haemophilus* and non-pathogen taxa (grey). However, the data also illustrate that identification of pathogens with lower lung function is not clear cut – there are multiple individuals whose sputum is dominated by pathogens and yet have normal lung function. Conversely, there are multiple individuals with low prevalence of any or all pathogens and yet suffer from heavily impaired lung function. We next examined the composition of the non-pathogen component of sputum samples (the gray bar in Fig 2a), and found a striking consistency across individuals regardless of lung function (Fig 2b). *Veillonella* and *Streptococcus* consistently dominate the non-pathogen microbiome component, regardless of lung health or pathogen status. Integrating across pathogen and non-pathogen components, we find that normal lungs are more diverse than severe ones ($p<0.01$, Fig 2c), in line with multiple other studies.[27,28] Turning to ordination plots (principle coordinates analysis, Fig 2d) we find ppFEV1 was significantly associated with microbiome composition (Mantel test, $r=0.195$, $p<0.001$).

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

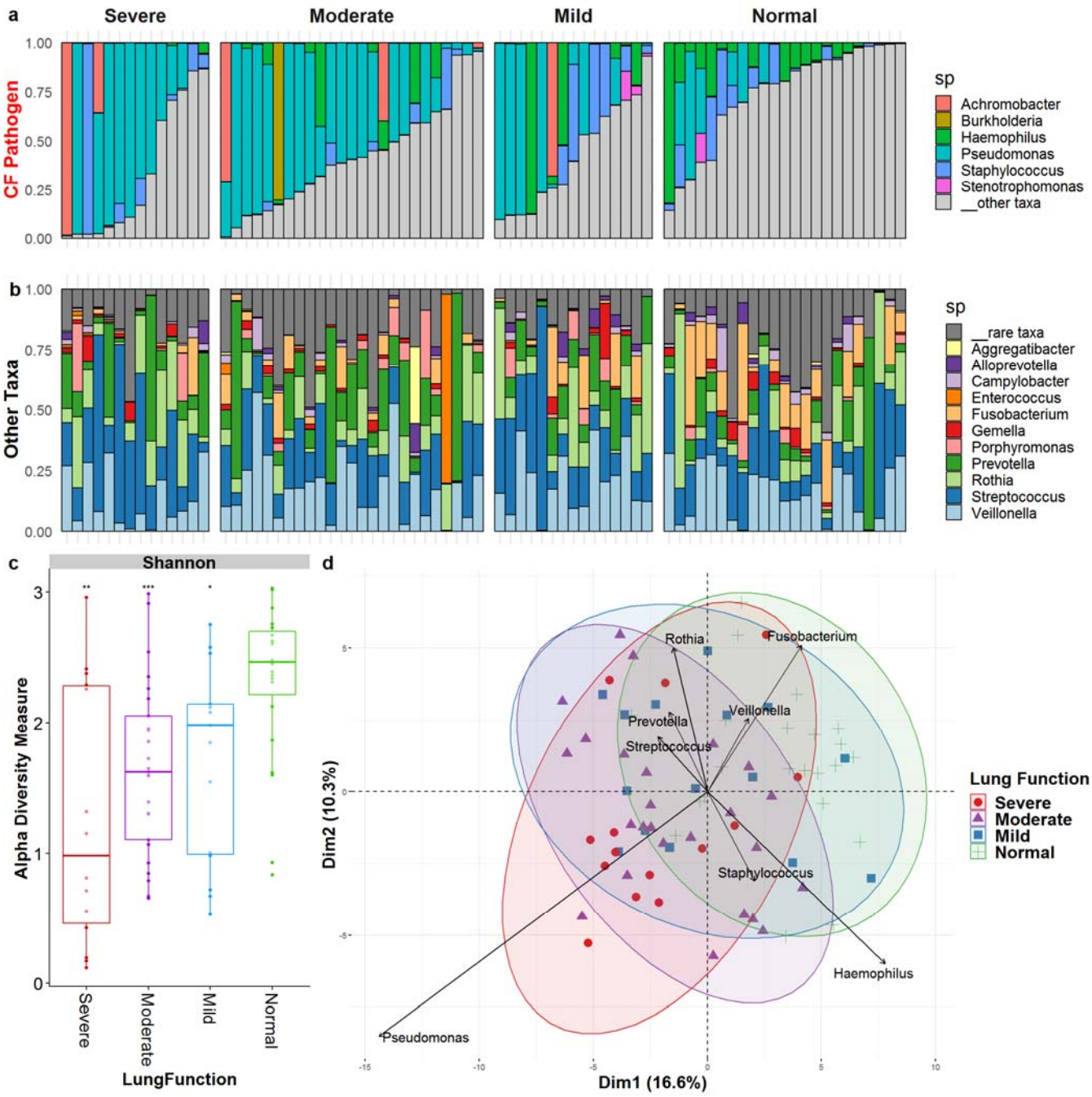


Fig 2. CF Lung microbiome composition varies with lung function and pathogen dominance. *Relative abundances of* ***(a)*** *6 canonical CF pathogens and* ***(b)*** *other taxa (the grey bar taxa in (a)). Microbiome compositions grouped by disease severity, classified by ppFEV1 score: normal (80+), mild (60-80), moderate (40-60), and severe (<40).* ***c)*** *Within-sample diversity (Shannon index) is lower in severe disease states (Kruskal-Wallis, p<0.01).* ***d)*** *Between-sample diversity (Bray-Curtis PCoA on top 25 genera, centered log-ratio transformed). PCs 1 and 2 combined explain ~40% of the microbiome variance, and weakly clusters patients by Lung Function.*

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Integrating microbiome and patient meta-data

Fig 2 illustrates patterns of association between microbiome data and a critical patient health outcome. However multiple confounding variables are not addressed, such as patient age, BMI or CF-related diabetes (CFRD). To look more globally at the associations between our multiple clinical and microbiome metrics, we generated a clustered correlation matrix across all variable pairs (Fig 3). We found a complex autocorrelation structure, with many expected consistencies. Heirarchical clustering notably groups 16S quantitation variables with patient metadata and clinical microbiology results. Unsurprisingly, FEV1 and ppFEV1 cluster together and are anticorrelated with ppFEV1 decline rate (average rate of decline in ppFEV1 since birth). Additionally, 16S quantitation results for *Pseudomonas*, *Staphylococcus*, *Burkholderia*, and *Achromobacter* cluster with their respective culture-based clinical microbiology results. This does not hold for *Stenotrophomonas,* which may be due to its infrequent detection.

Hierarchical clustering identified two large clusters of correlated variables. One correlated with ppFEV1, and included alpha diversity as well as 16S quantitation of *Fusobacterium*, *Haemophilus*, and *Neisseria*. The other anticorrelated with ppFEV1, and included ppFEV1 decline, pathogen abundance, CFRD and 16S quantitation of *Pseudomonas* and *Achromobacter*.

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

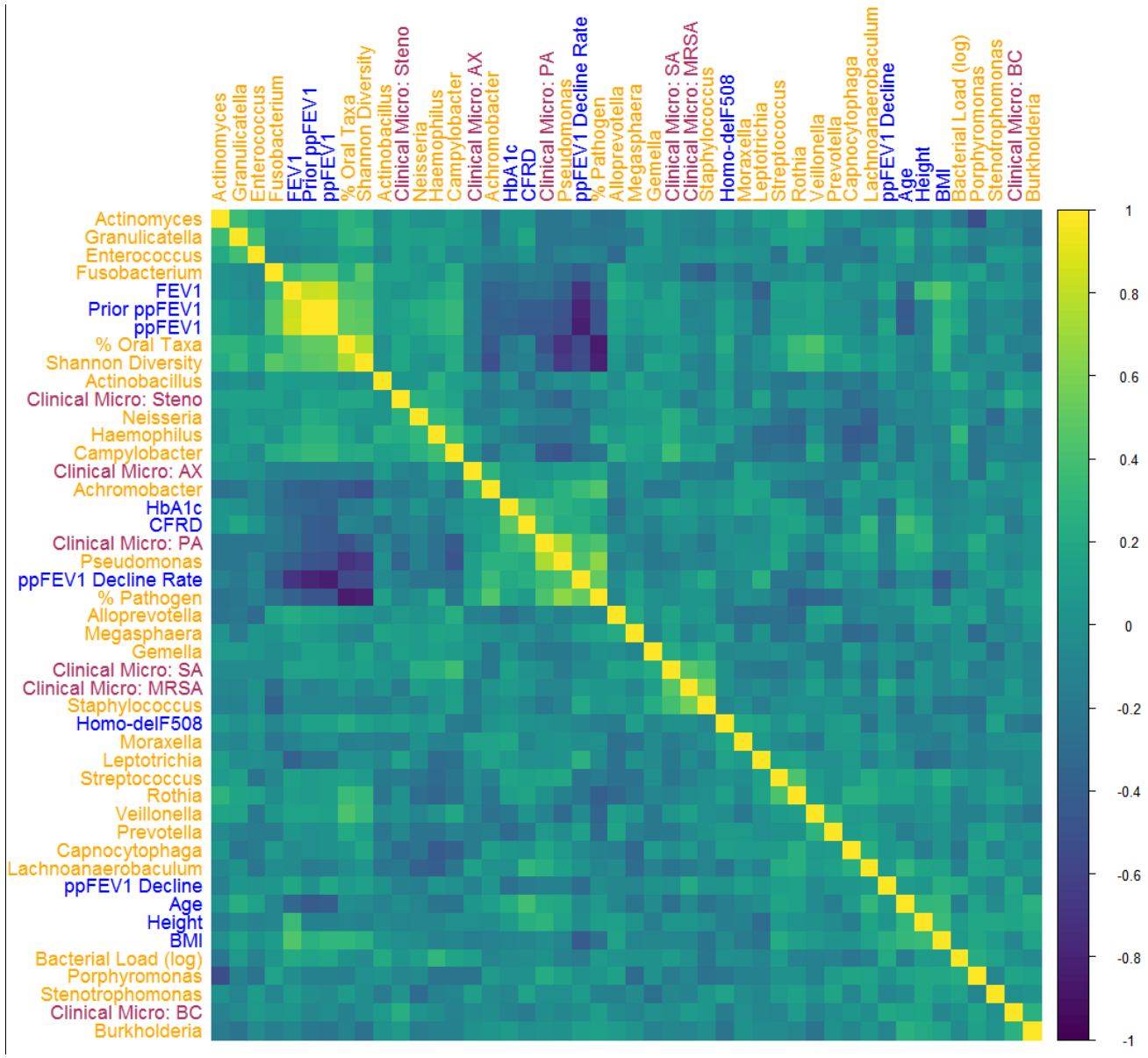


Fig 3. Lung function varies with patient meta-data. *Spearman correlations (R::corrplot) across all patient metadata (blue), clinical micro results (maroon), and microbiome data (orange, clr-transformed) reveal a complex correlation structure. We used a centered-log transform on 16S data to mitigate compositional effects. Rows and columns were ordered by hierarchical clustering, which identified clusters of metadata and microbiome variables with similar correlation patterns.*

Predicting Lung Function

The hairball correlation matrix in Fig 3 highlights the statistical challenges in addressing our underlying question of identifying meaningful predictors of patient outcomes. First, there are many potential predictors (Fig 3 shows 44 patient parameters out of 86 total, including 59 bacterial genera), and second, there are substantial

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and at times strong correlations among these parameters. Further compounding the challenge, we have relatively few independent patient observations (N=77) compared to the number of potential predictors.

To address these challenges, we first restrict our microbiome analysis to only the top 23 genera in our dataset, to focus on commonly encountered taxa only (Fig 1). We also calculate three additional features: % pathogen, % oral taxa, and Shannon diversity. Next, we use machine learning methods to provide a principled basis for the retention of meaningful predictors. To address compositionality of 16S data, we incorporate total bacterial load (universal 16S primer qPCR) as a predictor. In addition, we use a centered log-ratio (clr) transform on our genus-level relative abundance data before standardizing to mean zero, unit variance inputs.

Our machine learning pipeline is outlined in Fig 4, illustrating our approach to assessing the relative predictive power of different subsets of patient data (patient electronic records and sputum 16S microbiome data). Our null hypothesis, following the work of Jorth et al. and others[25,26] is that clinical microbiology provides an adequate explanatory basis for lung function outcomes, and more specifically that the addition of non-pathogen 16S data does not improve predictive ability. We expect that the addition of patient metadata (age, BMI etc) will improve our ability to predict lung function due to the progressive nature of CF, however our null hypothesis predicts that the addition of non-pathogen microbiome data will not improve predictive power, with or without the inclusion of meta-data.

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

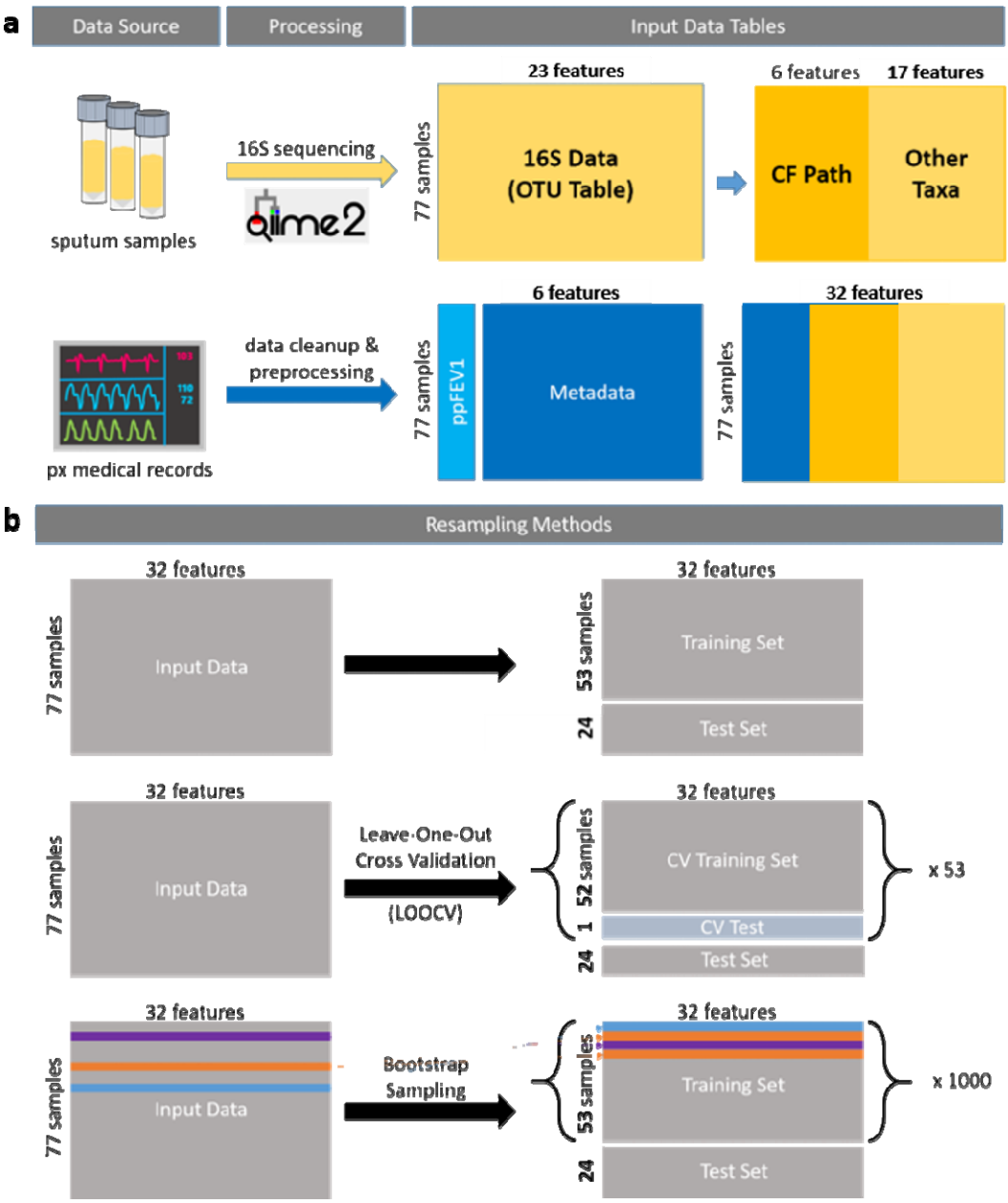


Fig 4. Machine Learning Data Preparation. *Machine learning models are trained on different input data tables using varying data resampling methods. (**a**) Features are categorized by information source (16S, patient metadata, or culture results). 16S data is further split into pathogens and other taxa in agreement with Fig 2. Machine learning models are trained on each input data table individually (CF Pathogens, Other Taxa, Metadata, Clinical Micro) and combined (All 16S Data, All Data). (**b**) We randomly selected 24 patient samples to withhold as a test set and train our models on the remaining 53 samples. To assess overfitting, we use leave-one-out cross validation on our training set. We estimate the variance of our model fits using 1000-fold bootstrap resampling.*

To illustrate our machine learning approach, we begin with the model output trained on the full dataset (all 16S and metadata predictors, Fig 5). Fig 5a plots predicted versus observed lung function, for both the

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

training dataset (data on 53 patients used to train model parameters) and the test dataset (data on 24 patients held back during model training). The consistency of training set and test set R^2 values suggests the model is not overfitting the training sets. Fig 5b highlights the parameters retained in the predictive model and their weighting (blue for positive predictors, red for negative predictors). In Fig. S1 we illustrate the performance of models trained on subsets of the data, all of which show lower R^2 values than the model trained on all data (Fig 5). However the predictive features selected in the integrated model are broadly consistent with models trained on each dataset individually. Positive and negative predictors selected in the pathogen-only and all 16S models (Fig S1b) were also selected in the all data model. While bacterial load and CFTR mutation type were informative in a metadata-only model (metadata, Fig S1b), the all data model does not select these features. We hypothesize that mutation type and bacterial load, share information with *Rothia* quantitation, and indeed find these features closely clustering in our correlation analysis (Fig 3).

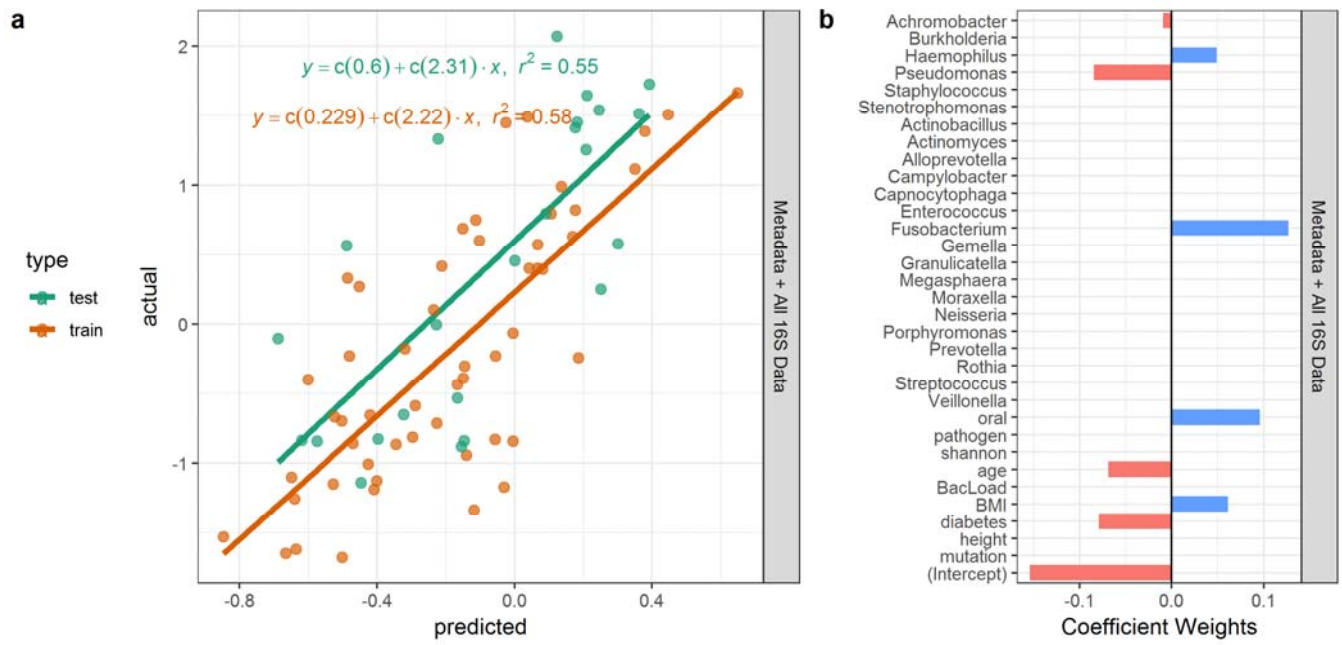


Fig 5. ElasticNet-identified predictors of lung function. *We train a predictive model of ppFEV1 using the ElasticNet algorithm (alpha = 0.5) to perform feature selection. We assess the train-test holdout method on metadata + all 16S data. The train-test uses a standard 70-30 split (53 patient training set, 24 patient test set).* ***(a)*** *We plot model-predicted ppFEV1 values (scaled) against actual ppFEV1 values and calculate the R^2 of the fit. We find that the model trained with the full dataset has the highest performance (see Fig S1 for prediction subset model performance) and selects features across different input data sources.* ***(b)*** *Model coefficients from the train-test holdout show general agreement with CF heuristics. Age, diabetes, and Pseudomonas abundance are selected as negative predictors of age whereas oral taxa abundance and BMI are positive predictors.*

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Our analysis in Figures 5 and S1 suggest that the addition of non-pathogen 16 data improves model performance as evidenced by improvements in R^2, and flags specific non-pathogen taxa as potential predictors. To more carefully assess these suggestions, we computationally augment our training datasets using 1000-fold bootstrap resampling and train models on each of the bootstrapped datasets.

Figure 6a shows the relative model prediction performance (measured by mean squared error) for each of the five input data sources, and plots bootstrap-generated confidence intervals (boxplots) in addition to single points for the non-bootstrapped train/test model approaches from Figure 5 and S1 (black points). To establish a performance baseline, we design a non-informative (randomized) input dataset by within-feature shuffling the entries from the 'all data' input set, scrambling between-feature correlations while preserving the mean zero, unit variance within-feature structure. All models using patient metadata or microbiome data outperform our negative control baseline.

To address the key question of relative model performance, we find that the addition of non-pathogen taxa significantly improves predictive ability (significantly reduces bootstrapped MSE; Figure 6a), with or without the addition of patient meta-data. Models trained on all 16S quantitation overall significantly outperform models trained only on pathogen quantitation. Interestingly, while microbiome data and metadata-trained models perform equivalently, combining the two datasets permits greater model performance. Looking broadly across models, we find reasonable consistency in positive and negative predictor selection between our non-bootsrapped train/test (black dots) and our bootstrapped (boxplots) models (Fig 6c-g).

We find multiple features selected across all training sets. *Pseudomonas*, *Achromobacter*, age, and diabetic status are consistently selected as negative predictors, while *Haemophilus*, *Fusobacterium*, *Rothia*, oral taxa abundance, and *BMI* are consistently positive predictors. All informative features selected in the independent models (Fig 6c-e) were also selected in the integrated model (Fig 6g). A small subset (< 50%) of the bootstrapped models also selected a handful of oral taxa, bacterial load, and CFTR mutation type as positive predictors of lung function (Fig 6g, gray boxplots). However, a majority of bootstrapped models and the train/test model did not select these as informative features.

As an additional check against overfitting, we obtain ranges of model errors (measured by mean squared error of predicted ppFEV1 values) using leave-one-out cross validation (Fig 6b). We do not find significant differences between cross-validated model errors across our training sets. Median cross-validated errors were consistently lower than the train/test split model error (Figure 6b), suggesting that our models are not overfitting.

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig 6. ML identification of lung function predictors.

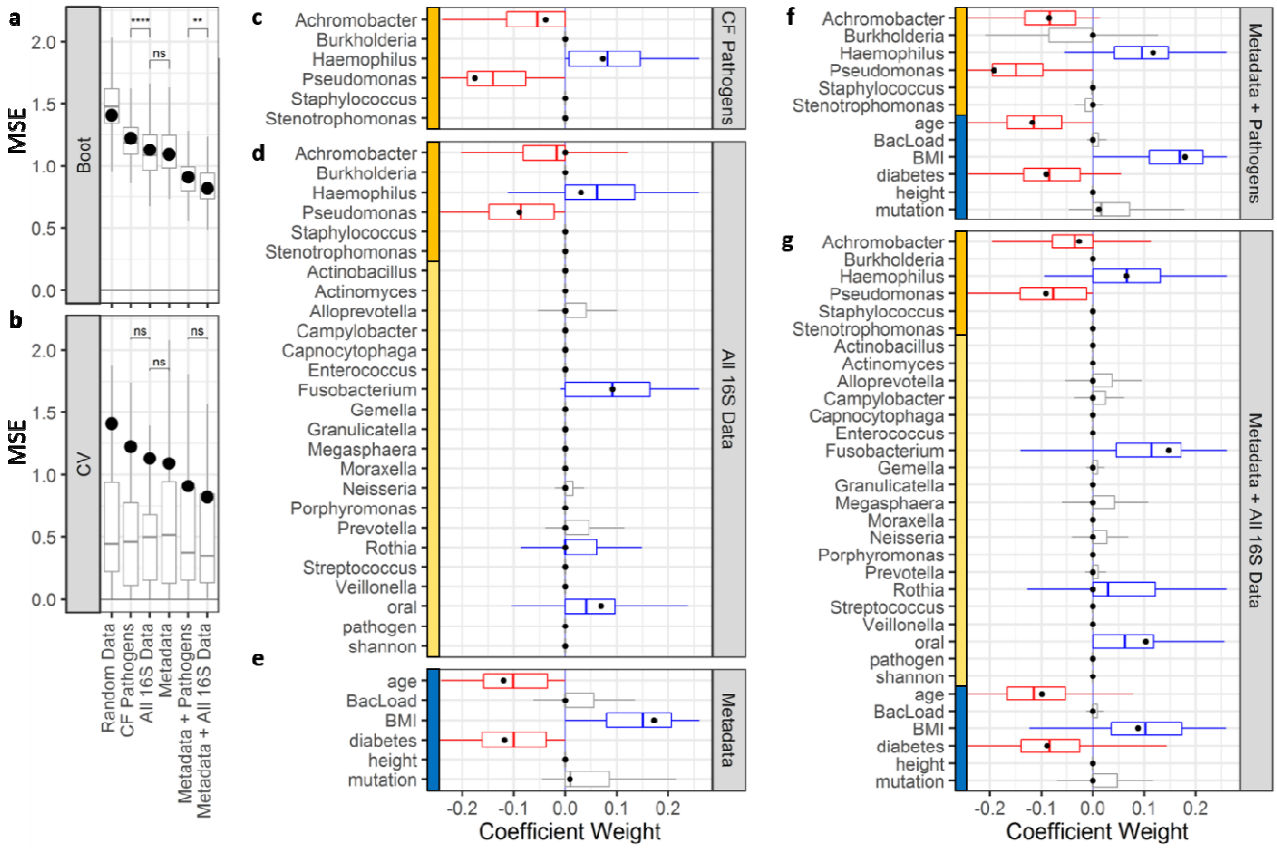


Fig 6. Bootstrapped ElasticNet-identified predictors of lung function. *ML models were trained using varying input datasets.* ***a*** *1000-fold bootstrapping and* ***b)*** *LOOCV were used to generate prediction error (MSE) ranges across feature subsets. Models trained on all of the data show lower error compared to other feature subsets. Adding 16S pathogen quantitation decreases model error. Models trained on all 16S data outperform models using only 16S quantitation ($p < 0.01$, t test). Regardless of input features, models trained on the full sample set (black points) are greater than median LOOCV MSEs (boxplots).* ***c-g)*** *Coefficient ranges for train/test (black points) and bootstrapped models (boxplots) trained on varying input datasets (blue: metadata, orange: 16S pathogens, yellow: 16S other taxa) show consistency between both machine learning strategies. Both cases select* Pseudomonas *and* Achromobacter *as negative predictors.*

Discussion

People with CF face the challenge of managing long-term chronic infections. Current management practice is driven by clinical microbiology identification of specific pathogens in expectorated sputum samples, alongside measures of overall health status (lung function, BMI, CF-RD). In the current study, we used 16S sequencing to assess sputum microbiome content more broadly, and ask whether the addition of non-pathogen taxa improves our ability to predict patient lung health, with or without the inclusion of patient health data. To address this question we applied machine learning tools to an integrated 77 patient lung microbiome and

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

electronic medical record dataset. Our analysis revealed that the addition of non-pathogen data improves prediction of patient health, with the most accurate models selecting patient metadata, pathogen quantitation, and non-pathogen information. Our inclusive 'all data' models additionally point to a predictive role for specific non-pathogen taxa, in particular the oral anaerobe genera *Rothia* and *Fusobacterium*.

Despite the significant contribution of non-pathogen data, our results are still broadly consistent with what might be termed the 'traditional' view of CF microbiology. Established CF pathogens (*P. aeruginosa, S. aureus, H. influenzae, B. cenocepacia*) are the major drivers of patient outcomes, as evidenced by substantial improvement in predictive outcomes whenever we include pathogen data (Fig 6a), and the by comparison relatively weak contribution of the addition of non-pathogen taxa. Note that we specifically use quantitative 16S measures of pathogen composition to provide a level playing field in the comparison of pathogen and non-pathogen predictive contribution. Fig 3 highlights that quantitative 16S and qualitative (presence/absence) clinical microbiology data are in general agreement.

The traditional role of CF pathogens as the central predictors of patient outcomes has been challenged over the past decade by the advent of microbiome sequencing. In a CF context, extensive surveys have documented an association between CF lung function and microbiome diversity, also evident in the current study (Fig 2). These results at face value suggest a biological role for these non-pathogen taxa, potentially competing with[29] or facilitating[30] pathogen taxa and therefore indirectly shaping disease outcomes. Jorth et al. recently published a forceful rejection of this 'active microbiome' view, stressing the causal role of changing pathogen densities in shaping disease outcomes, viewing shifting diversity metrics as a simple statistical artifact of shifting pathogen numbers against a roughly constant oral contamination background.[25] Our analyses provide some support for this view, in particular the constancy of the non-pathogen microbiome across patients (Fig 2b) and the lack of substantial predictive improvement on addition of non-pathogen data (Fig 6b). However on a more fine-scaled analysis we see that specific non-pathogen taxa are retained in our most explanatory models, alongside pathogen taxa.

Our 'all data' models highlight *Rothia* and *Fusobacterium* as positive predictors of lung function across our 77 patients, in models that already take into account pathogen data, age and BMI. The retention of these specific taxa in both this full model and in partial models (Fig 6b-c) suggests that these taxa provide potentially valuable predictive information on current patient health. Of course, this analysis does not allow inference to causal mechanism or even direction of causality. It is entirely possible that these taxa are simply bio-markers of dimensions of improved health that are largely independent of age, BMI, and other established positive predictors that are already accounted for in the model. It is also possible that these specific taxa play a more active causal role, for instance holding specific pathogens at bay via competitive interspecific mechanisms.[31]

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Interestingly, our 'all data' models also highlight *Haemophilus*, a canonical CF pathogen, as a positive predictor of lung function. *Haemophilus influenzae* infections are most common in younger CF patients,[8,32] hence we would expect a positive association in a model that is not controlled for age (Fig 6c, 6d). However we see that the positive weighting on *Haemophilus* is retained in models that also account for age as a positive predictor of lung function (Fig 6f-g). A second possibility is that the positive weighting of *Haemophilus* is due to pathogen-pathogen competition and the relatively less severe nature of *Haemophilus* infections in adults (i.e., *Haemophilus* is 'best of a bad job'). Fig 2a illustrates that we only appreciably detect two and rarely three coexisting pathogens of the six we find across all patients. The relatively depauperate pathogen communities implies that *Haemophilus* presence coincides with the absence of other more severe pathogens – and indeed we see a dominance of negative correlations among pathogens (Fig 3). In this context we cannot preclude a protective role of *Haemophilus* against more severe pathogens in older patients.

A caveat of this analysis is the dependency of machine learning performance and robustness on particular distributions of data, and the failure of linear algorithms such as LASSO and ElasticNet on microbiome-like data.[33–35] This is in part due to the compositionality constraint of microbiome data, which can be mitigated by using absolute quantitation.[36] However, training on absolute abundances introduces additional caveats, as order-of-magnitude differences in qPCR sample quantitation can in turn over-represent samples with higher bacterial loads. We address these issues by using a centered-log transform on relative abundance data and including log-scaled bacterial load as a potential feature to select. While a small minority of bootstrapped models selected bacterial load as a positive predictor (Fig 6c, Metadata + All 16S Data), the majority of models did not. This further suggests that the majority of microbiome information is encoded in the relative ratios of taxa abundance, which is broadly consistent with previous findings.[25,26]

Finally, our study is limited to a cross-sectional analysis, limiting us to making predictions on lung function state at the same time-point as microbiome sample and EMR collection. Assessing and refining our predictive machine learning algorithms on subsequent lung function data is an important future goal. Our primary objective is to predict future disease states and preemptively identify patients in need of medical intervention using early warning microbiome markers. To this effect, we plan to continue our analysis on a cohort of patients across time to evaluate predictive capacity for future health status.

In summary, our study finds that inclusion of non-pathogenic taxa significantly improves model prediction accuracy of patient health status. We identify two oral-derived taxa (*Fusobacterium, Rothia*) that are independently informative of lung function, which may be either biomarkers or potential probiotics. Our results call attention to the potential predictive utility of oral microbes (regardless of their functional roles) in the clinical assessment of CF patient health.

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Subjects

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committees. Authorization was obtained from each patient enrolled according to the protocol approved by the Emory University Institutional Review Board (IRB00010219 for adult and IRB00002161 for pediatric patients).

Sample collection.

Expectorated sputum samples were obtained from the CF-BR at Children's Healthcare of Atlanta and Emory University Pediatric CF Discovery Core from January 2015 to August 2016. De-identified patient information including age, sex, height, BMI, CFTR genotype, degree of glucose tolerance (HbA1c), and ppFEV1) were obtained (Table 1). Among these CF patients, 39 were diagnosed with CF-related diabetes patients (CFRD) by a CF endocrinologist. HbA1c value was missing for one CFRD subject.

All patients were clinically stable, defined as having less than a 10% change in ppFEV1 over the previous year with no medication changes for three weeks prior to sputum collection. Upon collection, sputum samples were stored and transported according to Emory CF-Biospecimen Registry protocols. Briefly, samples were diluted 1:3 (mass:volume) with PBS supplemented with 50 mM EDTA. Diluted samples were then homogenized by being repeatedly drawn through a syringe and 18-gauge needle. The resulting sputum homogenates were aliquot and stored at -80 °C until all 77 samples were collected; these were then sent to MR DNA Lab (Shallowater, TX) for DNA extraction, sequencing library preparation, Miseq sequencing, and absolute 16S quantitation. Clinical microbiology results were additionally obtained on the sputum sampling date.

DNA extraction and 16S sequencing

DNA was purified from sputum homogenate with the MoBio Power Soil kit (MoBio, Carlsbad, CA). The V4 region of the resulting DNA was amplified with the 16S universal primers 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3'). A single-step 30 cycle PCR integrating sequencing amplification and library adapter/barcode attachment was performed using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) by first incubation at 94 °C for 3 minutes, followed by 28 cycles of 94 °C for 30 seconds, 53 °C for 40 seconds and 72 °C for 1 minute, followed by a final elongation step at 72 °C for 5 minutes. Amplification products were then normalized, pooled and purified using calibrated Ampure XP beads for Illumina Miseq sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bioinformatics pipeline

Illumina Miseq sequencing generated in a total of 10,603,544 sequences, with an average of 137,708 sequences per sample (minimum 76,281, maximum 191,868). All sequence processing was done through QIIME2 2018.2.0. Raw sequences were firstly de-multiplexed and quality filtered on a per-nucleotide basis (min quality: 4, window: 3, min length fraction: 0.75, max ambiguous: 0). Reads were denoised using the deblur plugin, and the sequences were trimmed at the length of 250 bp (sample stats: T, mean error: 0.005, indel_prob: 0.01, indel_max: 3, min_reads: 10, min_size: 2, jobs_to_start: 1). Taxonomic assignments were classified against both the SILVA and greengenes database and assigned based on their highest taxonomic resolution. Discrepancies were resolved manually through BLAST and comparing against the non-redundant NCBI sequence database.

Based on taxonomic information, microbiome composition data was obtained for every sputum sample and a phylogenetic tree was constructed via *fasttree*. To correct for the variation 16S rDNA copy number among different taxa, the number of sequences per sample were divided by known 16S rDNA copy number of the genus or divided by four (average number of 16S rDNA copy number) if the information was missing.[37] Samples were rarefied to 17000 reads to guarantee equal sampling for subsequent analysis.[38]

Statistical and Quantitative Analysis

Patient samples were binned by ppFEV1-based lung function (Normal: >80%, Mild: 80-60%, Moderate: 60-40%, Severe: <40%). Variance across lung function categories in patient metadata, clinical microbiology data, and 16S metadata was tested using ANOVA. Variation between microbiome composition and ppFEV1 was tested using Mantel tests on Bray-Curtis distances at 9999 permutations. Within-sample and among-sample diversity was calculated using the Shannon diversity index and Bray-Curtis based PCoA on 16S quantitation data agglomerated to the genus level.[39] Associations between continuous variables were tested using Spearman correlations. To mitigate compositional effects, 16S data were center-log transformed prior to all analyses. A full pairwise correlation matrix was calculated, with rows and columns ordered by hierarchical clustering.[40]

Machine Learning

We use ElasticNet to fit regularized linear models predicting lung function (ppFEV1) from patient metadata, microbiome composition, and clinical microbiology results.[41] All input features were standardized to mean 0 variance 1 prior to model training. We create 6 input datasets based on information source – Clinical Micro, CF Pathogens, Other Taxa, All 16S Data (CF Pathogens + Other Taxa), Metadata, and All Data. We

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

additionally perform within-feature shuffling on the All Data set to create a bootstrap randomized dataset with the same dimensions, serving as a non-informative negative control.

For model validation, we employ three methods. First, we use a simple 70/30 train/test holdout, where models are trained on 53 samples and used to predict on the remaining 24. Model accuracy is measured using mean squared error (MSE). Second, we perform leave-one-out cross-validation on the training set to simulate model performance on new data, and compare the resulting MSE ranges to the holdout method to assess model overfitting. Finally, we perform bootstrap reshuffling on the training set to generate 1000 new training sets, fit a new regularized linear model to each, and obtain ranges and variances for the selected model coefficient weights. We perform each validation method on each input dataset to compare between information sources. We identify important predicting variables as ones with nonzero median coefficient weights selected across all bootstrap models.

Acknowledgements

We would like to thank Karan Kapuria and Eunbi Park for help with the development of our bio-informatic pipeline, and Peng Qiu for advice on our machine learning approach. We also thank the CDC (BAA 2016-N-17812, BAA 2017-OADS-01), the NIH (HR56L142857, R21AI143296) and the Cystic Fibrosis Foundation (BROWN19I0) for funding, as well as the Cystic Fibrosis and Airways Disease Research and Children's Healthcare of Atlanta for Initial pilot funding to Drs. Stecenko and Goldberg. Sputum samples were obtained from the CF-BR at Children's Healthcare of Atlanta and Emory University Pediatric CF Discovery Core.

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Persoon, A. *et al.* Leg ulcers: a review of their impact on daily life. *J. Clin. Nurs.* **13**, 341–54 (2004).
2. Guest, J. F. *et al.* Health economic burden that different wound types impose on the UK's National Health Service. *Int. Wound J.* **14**, 322–330 (2017).
3. Malone, M. *et al.* The prevalence of biofilms in chronic wounds: a systematic review and meta-analysis of published data. *J. Wound Care* **26**, 20–25 (2017).
4. Stacy, A., McNally, L., Darch, S. E., Brown, S. P. & Whiteley, M. The biogeography of polymicrobial infection. *Nat. Rev. Microbiol.* **14**, 93–105 (2016).
5. Perez-Vilar, J. & Boucher, R. C. Reevaluating gel-forming mucins' roles in cystic fibrosis lung disease. *Free Radic. Biol. Med.* **37**, 1564–1577 (2004).
6. Henke, M. O. & Ratjen, F. Mucolytics in cystic fibrosis. *Paediatr. Respir. Rev.* **8**, 24–29 (2007).
7. Rubin, B. K. Mucus, Phlegm, and Sputum in Cystic Fibrosis. *Respir. Care* **54**, 726–732 (2010).
8. Bals, R., Weiner, D. J. & Wilson, J. M. The innate immune system in cystic fibrosis lung disease. *J. Clin. Invest.* **103**, 303–307 (1999).
9. Dickson, R. P., Martinez, F. J. & Huffnagle, G. B. The role of the microbiome in exacerbations of chronic lung diseases. *Lancet* **384**, 691–702 (2014).
10. Rieber, N., Hector, A., Carevic, M. & Hartl, D. Current concepts of immune dysregulation in cystic fibrosis. *Int. J. Biochem. Cell Biol.* **52**, 108–112 (2014).
11. Yonker, L. M., Cigana, C., Hurley, B. P. & Bragonzi, A. Host-pathogen interplay in the respiratory environment of cystic fibrosis. *J. Cyst. Fibros.* **14**, 431–439 (2015).
12. Conrad, D. *et al.* Cystic fibrosis therapy: A community ecology perspective. *Am. J. Respir. Cell Mol. Biol.* **48**, 150–156 (2013).
13. Frayman, K. B., Armstrong, D. S., Grimwood, K. & Ranganathan, S. C. The airway microbiota in early cystic fibrosis lung disease. *Pediatr. Pulmonol.* **52**, 1384–1404 (2017).
14. Lucas, S. K., Yang, R., Dunitz, J. M., Boyer, H. C. & Hunter, R. C. 16S rRNA gene sequencing reveals site-specific signatures of the upper and lower airways of cystic fibrosis patients. *J. Cyst. Fibros.* **17**, 204–212 (2018).
15. Fodor, A. A. *et al.* The Adult Cystic Fibrosis Airway Microbiota Is Stable over Time and Infection Type, and Highly Resilient to Antibiotic Treatment of Exacerbations. *PLoS One* **7**, (2012).
16. Huang, Y. J. & LiPuma, J. J. The Microbiome in Cystic Fibrosis. *Clin. Chest Med.* **37**, 59–67 (2016).
17. Coburn, B. *et al.* Lung microbiota across age and disease stage in cystic fibrosis. *Sci. Rep.* **5**, 1–12 (2015).
18. Acosta, N. *et al.* Sputum microbiota is predictive of long-term clinical outcomes in young adults with

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cystic fibrosis. *Thorax* **73**, 1016–1025 (2018).

19. Muhlebach, M. S. *et al.* Initial acquisition and succession of the cystic fibrosis lung microbiome is associated with disease progression in infants and preschool children. *PLoS Pathog.* **14**, 1–20 (2018).
20. Zhao, J. *et al.* Decade-long bacterial community dynamics in cystic fibrosis airways. *Proc. Natl. Acad. Sci.* **109**, 5809–5814 (2012).
21. Zemanick, E. T. *et al.* Assessment of airway microbiota and inflammation in cystic fibrosis using multiple sampling methods. *Ann. Am. Thorac. Soc.* **12**, 221–229 (2015).
22. O'Neill, K. *et al.* Reduced bacterial colony count of anaerobic bacteria is associated with a worsening in lung clearance index and inflammation in cystic fibrosis. *PLoS One* **10**, (2015).
23. Quinn, R. A. *et al.* Ecological networking of cystic fibrosis lung infections. *npj Biofilms Microbiomes* **2**, 4 (2016).
24. Klepac-Ceraj, V. *et al.* Relationship between cystic fibrosis respiratory tract bacterial communities and age, genotype, antibiotics and Pseudomonas aeruginosa. *Environ. Microbiol.* **12**, 1293–1303 (2010).
25. Jorth, P. *et al.* Direct Lung Sampling Indicates That Established Pathogens Dominate Early Infections in Children with Cystic Fibrosis. *Cell Rep.* **27**, 1190-1204.e3 (2019).
26. Goddard, A. F. *et al.* Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. *Proc. Natl. Acad. Sci.* **109**, 13769–13774 (2012).
27. Carmody, L. A. *et al.* Changes in cystic fibrosis airway microbiota at pulmonary exacerbation. *Ann. Am. Thorac. Soc.* **10**, 179–187 (2013).
28. Flight, W. G. *et al.* Rapid Detection of Emerging Pathogens and Loss of Microbial Diversity Associated with Severe Lung Disease in Cystic Fibrosis. *J. Clin. Microbiol.* **53**, 2022–2029 (2015).
29. Quinn, R. A. *et al.* Ecological networking of cystic fibrosis lung infections. *npj Biofilms Microbiomes* **2**, 0–1 (2016).
30. Flynn, J. M., Niccum, D., Dunitz, J. M. & Hunter, R. C. Evidence and Role for Bacterial Mucin Degradation in Cystic Fibrosis Airway Disease. *PLoS Pathog.* **12**, 1–21 (2016).
31. McNally, L. & Brown, S. P. Building the microbiome in health and disease: Niche construction and social conflict in bacteria. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, (2015).
32. Bogaert, D. *et al.* Variability and diversity of nasopharyngeal microbiota in children: A metagenomic analysis. *PLoS One* **6**, (2011).
33. Rush, S. T., Lee, C. H., Mio, W. & Kim, P. T. The Phylogenetic LASSO and the Microbiome. (2016).
34. Leng, C., Tran, M. N. & Nott, D. Bayesian adaptive Lasso. *Ann. Inst. Stat. Math.* **66**, 221–244 (2014).
35. Banerjee, P., Garai, B., Mallick, H., Chowdhury, S. & Chatterjee, S. A Note on the Adaptive LASSO for Zero-

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Inflated Poisson Regression. *J. Probab. Stat.* **2018**, 1–9 (2018).

36. Jian, C., Luukkonen, P., Yki-Jarvinen, H., Salonen, A. & Korpela, K. Quantitative PCR provides a simple and accessible method for quantitative microbiome profiling. *bioRxiv* 478685 (2018). doi:10.1101/478685

37. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K. & Schmidt, T. M. rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* **43**, D593–D598 (2015).

38. Heck, K. L., Jr., G. van B. & Daniel Simberloff. Explicit Calculation of the Rarefaction Diversity Measurement and the Determination of Sufficient Sample Size Author (s): Kenneth L . Heck , Jr ., Gerald van Belle and Daniel Simberloff Published by: Wiley on behalf of the Ecological Society of America. *Ecology* **56**, 1459–1461 (1975).

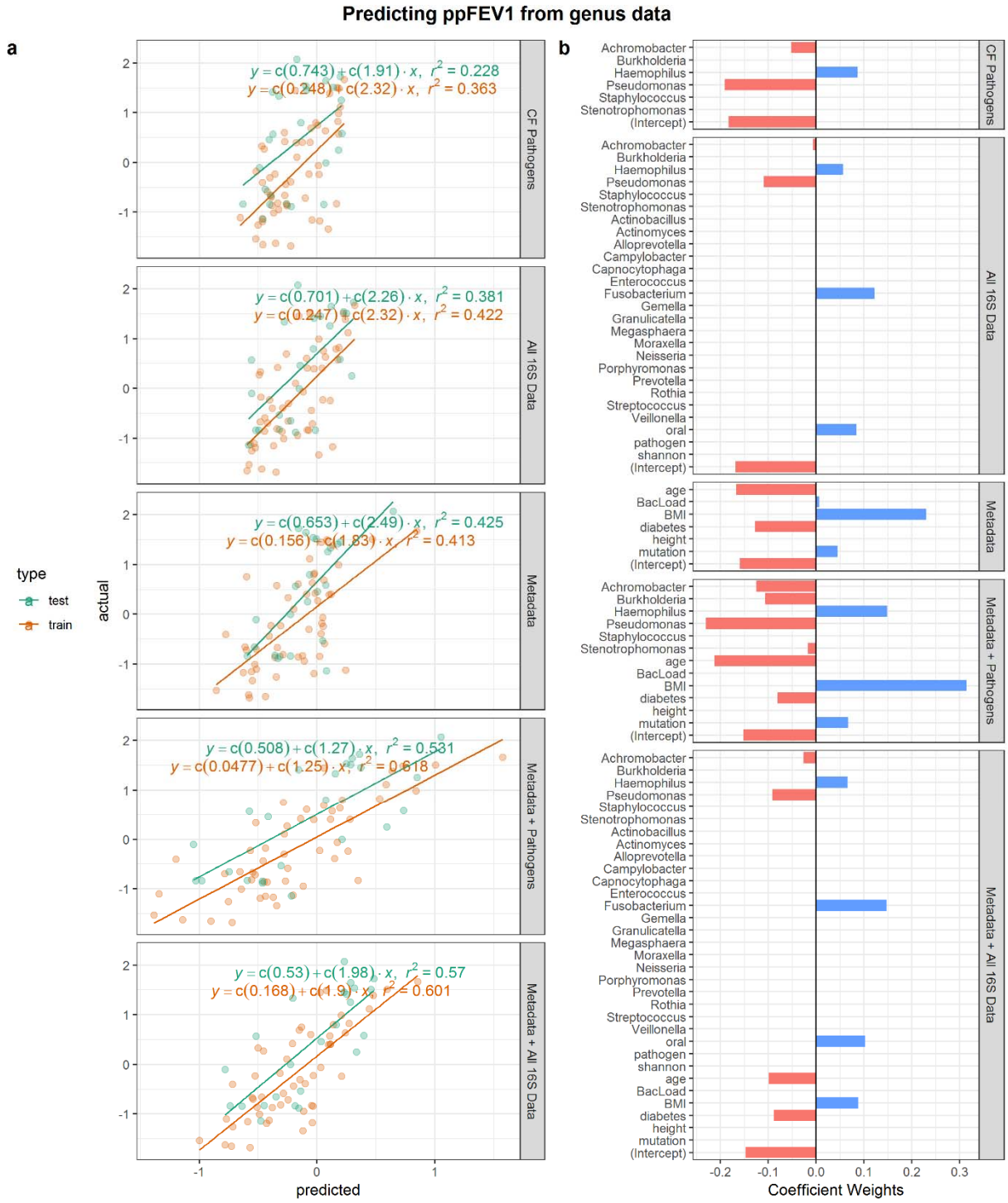
39. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**, e61217 (2013).

40. Wei, T. *et al.* corrplot: Visualization of a correlation matrix. *R Packag.* (2017).

41. Yuan, G.-X., Ho, C.-H. & Lin, C.-J. An improved GLMNET for l1-regularized logistic regression. in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11* **13**, 33 (ACM Press, 2011).

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

468 **Microbiome data enhances predictive models of lung function in people with CF - Supplementary Figures**



469

470 **Fig S1. Predicting ppFEV1 from genus data.** *We assess the train-test holdout method on five input data sources:*

471 *16S quantitation of CF Pathogens (clr-transformed), all 16S data (clr-transformed), metadata, metadata +*

472 *pathogens, and metadata + all 16S data.*

bioRxiv preprint doi: https://doi.org/10.1101/656066; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

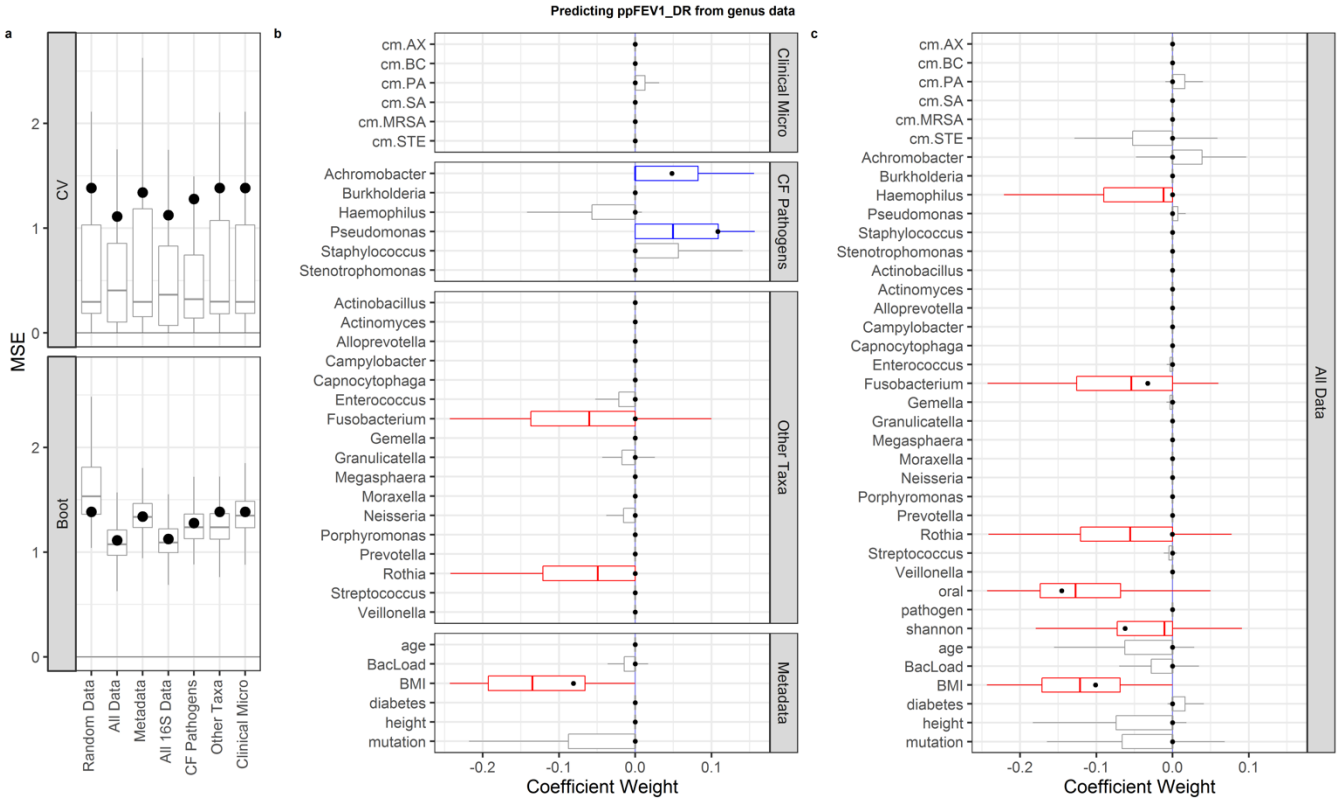


Fig S2. Predicting ppFEV1 decline rate. We find that models trained on varying datasets weakly outperform a random data control. *Rothia* and *Fusobacterium* are selected as negative predictors of decline rate, suggesting a protective role in the CF microbiome.