IDENTIFYING MICROBIAL BIOMARKERS OF CYSTIC FIBROSIS HEALTH AND DISEASE

A Dissertation
Presented to
The Academic Faculty

By

Conan Ying-Yi Zhao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Quantitative Biosciences
School of Biology

Georgia Institute of Technology

August 2022

IDENTIFYING MICROBIAL BIOMARKERS OF CYSTIC FIBROSIS HEALTH AND DISEASE

Thesis committee:

Prof. Sam P. Brown, Ph.D., Advisor
School of Biological Sciences
Georgia Institute of Technology

Prof. Joshua S. Weitz, Ph.D.
School of Biological Sciences
Georgia Institute of Technology

Prof. Peng Qiu, Ph.D.
Department of Biomedical Engineering
Georgia Institute of Technology

Prof. Arlene A. Stecenko, M.D.
Department of Pediatrics
Emory University School of Medicine

Prof. Rishikesan Kamaleswaran, Ph.D.
Department of Biomedical Informatics
Emory University

Date Approved: 06/24/2022

It is important to draw wisdom from many different places.
Iroh

To my family,
especially my Dad.

ACKNOLWEDGEMENTS

It is not often I get to sit at a major checkpoint in my life and reflect on all the people who helped make it possible for me to get this far. I've been blessed to have the support of so many friends, colleagues, mentors, teachers, and family along this path. It seems like there's an infinite list of names I should add here, and an uncountably infinite list of reasons behind each one. For those reading (for whatever reason that might be), know that perhaps the hardest part of writing this thesis (besides wrangling a citation manager) was paring down this list:

Thank you to the members of the Brown Lab and CMDI, past and present. Especially Jennifer Rattray, Kristofer Waldetoft, and Tim O'Sullivan for welcoming me to the group, and to Jennifer Farrell, Juan Castro, Jelly Vanderwoude, Madeline Mei, Kathleen O'Connor, Juan Barazza, Alex Klementiev, and Julia Schap for getting me through all the ups and downs of research, life, and graduate school.

Thank you to my QBioS Ph.D. cohort for getting me through the first few years – and especially for their company during those early mornings when Overleaf was down for maintenance. A special thank you to Pablo Bravo for all the networking/"not working" coffee chats, and for letting me loop you into helping me run QBioS SGA.

Thank you to Lisa Redding and Will Ratcliff for all the incredible programmatic support.

Thank you to my advisor, Sam Brown, for giving me the freedom to pursue the projects that interested me, and the patience to wait as I came up with excuses for why I still hadn't written a paper. And to Joshua Weitz for "coincidentally" showing up to my office with Sam on the morning that grad school decisions were due to see how my everything was going.

Thank you to Sophia Wiesenfeld for taking care of me while I wrote this thesis.

And finally, I thank my parents, to whom this work is dedicated:

To Mom, for keeping me humble.

And to Dad, for igniting my love of science.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Summary

Chronic, polymicrobial respiratory infections remain the primary driver of morbidity and mortality in cystic fibrosis (CF). This thesis leverages experimental data and large-scale public datasets to investigate the relationships between microbiome structure, pathogen abundance and host health.

First, using a machine learning framework, we show that off-the-shelf machine learning methods can recover known clinical and microbial predictors of lung function from a set of 77 sputum composition profiles. These methods recover known demographic predictors of lung function and further identify novel taxonomic predictors, highlighting the utility of simple machine learning methods for microbial biomarker discovery.

Second, we develop a synthetic infection microbiome model representing CF metacommunity diversity, and benchmark on clinical data. Using this synthetic microbiome system, we provide evidence that commonly used CF antibiotics can drive the expansion (via competitive release) of previously rare opportunistic pathogens and offer a path towards microbiome-informed treatment strategies.

Last, we manually curated a microbiome dataset of over 4000 sputum samples representing more than 1000 people with CF (pwCF), matching samples with corresponding metadata from 36 publications and standardizing bioinformatic analyses with a single common pipeline. We fit Sloan Neutral Community Models to each study and find a consistent set of neutral and non-neutral taxa. Using Dirichlet Multinomial Mixture modeling, we partition non-neutral CF lung microbiomes into 14 distinct pulmotypes. Integrating longitudinal data, we find that not all Pseudomonas-dominated pulmotypes are dynamically equivalent, which carries important implications for infection management in cystic fibrosis

Chapter 1: Introduction

1.1 Background and Significance

Chronic bacterial infections are characterized by the persistence of bacterial pathogens in a host, despite ongoing antibiotic treatment. From a host perspective, chronic infections impose elevated risks of morbidity and mortality (Persoon et al., 2004). These infections often fail to clear even with appropriate antibiotic treatment (as defined by antibiotic susceptibility testing), and the expansion of at-risk populations coupled with a drying antibiotic pipeline pose an increasing burden to global healthcare systems(Guest et al., 2017). From a microbial perspective, these infections commonly feature changes in pathogen growth mode promoting antibiotic tolerance (e.g. biofilm formation (Malone et al., 2017)) and additional microbial species acquisition, forming complex multispecies communities (Stacy et al., 2016) or infection microbiomes.

Cystic fibrosis (CF) is an autosomal recessive disease characterized by decreased lung mucociliary clearance and mucus accumulation (Henke & Ratjen, 2007; Perez-Vilar & Boucher, 2004; Rubin, 2010). The resulting environment provides both nutrients for bacterial growth and protection from host immune responses(Bals et al., 1999; Dickson et al., 2014; Rieber et al., 2014; Yonker et al., 2015), facilitating chronic microbial infections (Conrad et al., 2013; Fodor et al., 2012; Frayman et al., 2017; Lucas et al., 2018). Accessible 16S rDNA microbiome profiling has shifted CF airway microbiology research away from a historically single-pathogen focus. Sequencing expectorated sputum has revealed diverse communities of tens to hundreds of taxa, including numerous non-pathogenic bacteria (Frayman et al., 2017; Huang & LiPuma, 2016; Whelan et al., 2020). However, in practice clinical microbiology analysis continues to focus only on the "usual suspects" of established human pathogens such as *Pseudomonas aeruginosa*, *Staphylococcus aureus,* and other organisms with well-established health risks. This disconnect

between diverse infection microbiomes and limited clinical microbiology profiling may overlook important risk markers.

1.2 High-Dimensional Infection Data and Machine Learning

Healthy human microbiomes are diverse and person-specific. Microbiome composition varies across body sites and over both short and long timescales. Despite numerous microbiome association studies linking community composition with disease phenotypes(Anand & Mande, 2018; Russell et al., 2013; Stokholm et al., 2020), in most contexts the clinical implications of this variation remain unexplained. A number of studies have used machine learning methods to predict clinically-relevant host phenotypes from microbiome variation(Beck & Foster, 2014; Feng et al., 2015; Stokholm et al., 2020; C. Y. Zhao et al., 2021), and interpretable algorithms show promise in tackling the complexities and high dimensionality of microbiome datasets. Identifying phenotype predictors generates a set of candidate taxa to experimentally interrogate for underlying causal pathways governing associations, potentially offering insight for individual-tailored microbiome therapies. These methods fall under two general categories of either classification (e.g. identifying gut microbiome enterotypes (Arumugam et al., 2011) or classifying "diseased microbiomes") or regression (i.e. linking microbiome compositions to continuous disease metrics(Subramanian et al., 2014)).

Microbiome datasets are high-dimensional – typically having few samples but many taxonomic features. Additionally, microbiome data has challenging characteristics, such as zero-inflation, overdispersion, and compositionality. While there are numerous methods to address the challenges of analyzing high dimensional datasets, few have been rigorously benchmarked on microbiome-like data. Regularization and feature selection methods (ElasticNet, LASSO) or dimensionality reduction methods (PCA, NMDS) are often used in microbiome analyses but these challenging characteristics may influence model accuracy and robustness. To mitigate the challenge of compositionality, many pipelines recommend non-linear data transformations such

as centered-log-transformations (Gloor et al., 2017), to transform microbiome count data into a form more compatible with off-the-shelf statistical analyses. To address zero-inflation and overdispersion, algorithms analyzing RNAseq datasets, which share some distribution characteristics such as dropouts (Qiu, 2020), may also be applied to polymicrobial infection data.

1.3 Polymicrobial Infections in Cystic Fibrosis

Microbiome analyses indicate that CF respiratory communities consist of tens to hundreds of species. Cross-sectional and longitudinal studies in CF tend to be sample-limited, with the largest cohorts consisting of around 100 individuals with CF (Li et al., 2016; Zhao et al., 2021). Recently, Hampton et al. reported a cross-sectional study characterizing the sputum of 167 pwCF and used clustering analysis to identify five pulmotypes (Hampton, Thomas, van der Gast, O'Toole, & Stanton, 2021). However, the identification of potential causal links between CF lung microbial communities and overall clinical health metrics still presents a unique set of challenges. In this thesis, I propose to overcome the challenge of limited microbial samples in two ways: dataset augmentation (**Chapter 2**) and dataset augmentation (**Chapter 4**). In addition, I pursue causal investigation via experimental manipulation of synthetic CF microbiome communities (**Chapter 3**).

There are numerous published observational studies of the CF lung microbiome, but only a handful have taken a large-scale data science approach to interrogating patterns across CF sputum profiles (C. Y. Zhao et al., 2021). There is a striking lack of curated aggregate microbiome datasets in CF for benchmarking and algorithm development. Such datasets should include dense longitudinal sampling to characterize between-person variations and daily fluctuations. These datasets would allow for microbiome-specific algorithm development, analogous to CIFAR-10 or MNIST datasets for computer vision research.

Data augmentation is a common practice in machine learning pipelines. Computer vision problems often incorporate image flips and rotations to expand limited datasets, as such

transformations are information-invariant (Giuste et al., 2020). The analogous transformations in microbiome contexts are poorly characterized. Recently, Rong et al., have developed a generative adversarial network method augment microbiome datasets (Rong et al., 2021). Additional methods include subsampling or resampling, although method comparisons are needed to determine best practices.

1.4 Thesis Overview

1.4.1 Chapter 2: The relationship between microbiome structure and lung function

We begin by mapping microbiome structure to one of the central clinical indicators of health in pwCF (lung function) to identify potential microbial biomarkers. Expectorated sputum samples were obtained from 77 pwCF attending Georgia Tech and Emory-affiliated CF Care Centers and16S rDNA sequenced for airway microbiome compositions. De-identified clinical information including age, sex, height, BMI, CFTR genotype, degree of glucose control (HbA1c), and lung function (ppFEV1) for each sample were also obtained.

We employed simple machine learning methods to predict lung function from augmented microbiome and clinical information and compared extracted informative features from the best-performing models. We found that models trained on clinical vs microbiome subsets returned comparable performance, while training on all available data led to the best model performance. Our analysis shows that non-pathogen data improves prediction of lung health, with the most accurate models selecting a combination of clinical data, pathogen quantitation, and non-pathogen features. 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*. Interestingly, our models select *Haemophilus*, a canonical CF pathogen, as a positive predictor of health. We consider alternate hypotheses for this effect, including confounding by age (although age is retained in our best model) and protection by *Haemophilus* against more damaging

pathogens. However, we are unable to further assess causality given our limited sample size and lack of longitudinal data.

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 clinical outcomes, as evidenced by substantial improvement in predictive outcomes whenever pathogenic microbiome data is included over the relatively weak improvements from the addition of non-pathogen taxa. Our results suggest that the composition of non-pathogenic taxa contains important information about the lung health of pwCF, and support full-microbiome screening over the current pathogen-centric clinical microbiology focus.

1.4.2 Chapter 3: Analysis of synthetic microbiome responses to antibiotic perturbation

From Chapter 2, we identify putative associations between microbes and health, including a puzzling links between *Haemophilus*, *Rothia*, and *Fusobacterium* with better lung function. To experimentally investigate microbial interactions our group has developed a synthetic 10-species CF microbiome community, guided by data generated in Chapter 2. Such *in vitro* approaches complement observational data analyses by allowing for hypothesis testing and validation of the putative interactions proposed by data mining algorithms.

Using this *in-vitro* system, we find evidence that under certain conditions, commonly used CF antibiotics may drive expansion of otherwise rare pathogens via competitive release. Specifically, we show that in the absence of antibiotics, communities tend towards states dominated by oral commensals, with very low variability between replicates. In contrast, antibiotic perturbations generate alternate pathogen-dominant end states, enriched with drug-resistant taxa. The results highlight the potential importance of non-evolutionary (community-ecological) processes in driving the growing global crisis of increasing antibiotic resistance and offer a path towards microbiome-informed conditional treatment strategies.

1.4.3 Chapter 4: Revisiting clinical data on a larger scale

One critique of our synthetic community in Chapter 3 is that while our observed end states have clinical relevance, the initial conditions used do not reflect the composition of any individual sputum sample or pwCF. This leads to the question: what is a representative community in CF? Given the challenges of inter- and intra-individual heterogeneity, the aim of Chapter 4 is to establish the largest microbiome database of pwCF to date, and use this data to identify common community types in CF. We hypothesize that taxa driven by non-neutral ecological processes can be grouped into meaningful classes of microbial communities, or pulmotypes, based on their co-occurrence patterns.

Pulmotype identification using unsupervised learning or clustering algorithms require large, representative datasets. To date, the largest publicly available CF sputum dataset represents 299 pwCF across 13 CF centers worldwide (Cuthbertson et al., 2020). We curate a microbiome dataset of over 4000 sputum samples representing more than 1000 people with CF (pwCF) from 36 published studies on the NCBI Short Read Archive (SRA). We matched SRA studies with corresponding publications and standardized sequence analysis using a common pipeline.

We identify a common set of ecologically relevant CF taxa by removing taxa that are neutrally distributed as predicted by the Sloan Neutral Community Model (SNCM) (Sloan et al., 2006). We find that across studies, common CF pathogens are identified as non-neutral more often than neutral.

Dirichlet Multinomial Mixture modeling on non-neutral taxa partitions CF lung microbiomes into 13 pulmotypes, which we further group pulmotypes into three categories: *Pseudomonas*-dominant (PA), oral anaerobe dominant (OA), and other pathogen dominant (OP). We show that these pulmotypes are clinically and compositionally distinct with unique transition patterns. Specifically, we find differing transition frequencies between each PA pulmotypes and pulmotypes dominated by end-stage CF pathogens such as *Burkholderia* and *Achromobacter*. We

find that across a broad cohort of pwCF, *Pseudomonas*-dominated samples are not equivalent. These differing pulmotype transition frequencies provide important insight for infection management in cystic fibrosis. Chapter four represents the transition between my current and future work on CF microbiome dynamics, as the unique dataset established in this chapter will provide the baseline for continued microbiome benchmarking and algorithm development.

1.5 Other Work

In addition to my research focus on CF microbiomes, I have engaged in a range of more distantly related research during the course of my Ph.D. This work spans one publication in medical AI (Giuste et al., 2020), a manuscript each on CF infection microbiology (O'Connor et al., 2021) and a CF case study, and three contributing authorships on COVID epidemiological modeling publications (Farrell et al., 2021; Kraay et al., 2021; Weitz et al., 2020), and manuscripts *in prep* on infection microbiomes in the pediatric ICU as well as additional synthetic community modeling work.

Chapter 2: Microbiome data enhances predictive models of lung function in people with cystic fibrosis[1]

2.1 Summary

Microbiome sequencing has brought increasing attention to the polymicrobial context of chronic infections. However, clinical microbiology continues to focus on canonical human pathogens, which may overlook informative, but non-pathogenic, biomarkers. We address this disconnect in lung infections in people with cystic fibrosis (CF). We collected health information (lung function, age, BMI) and sputum samples from a cohort of 77 children and adults with CF. Samples were collected during a period of clinical stability and 16S rDNA sequenced for airway microbiome compositions. We use Elastic Net regularization to train linear models predicting lung function and extract the most informative features.

Models trained on whole microbiome quantitation outperform models trained on pathogen quantitation alone, with or without the inclusion of patient metadata. Our most accurate 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 non-pathogen taxa (*Fusobacterium*, *Rothia*) as positive predictors of lung health. These results support a reconsideration of clinical microbiology pipelines to ensure the provision of informative data to guide clinical practice.

[1] This chapter was adapted from the following reference: Zhao, C. Y., Hao, Y., Wang, Y., Varga, J. J., Stecenko, A. A., Goldberg, J. B., & Brown, S. P. (2021). Microbiome Data Enhances Predictive Models of Lung Function in People With Cystic Fibrosis. *JID*, 223(Supplement_3), S246–S256. https://doi.org/10.1093/infdis/jiaa655. Reused with permission. I was the primary author of this work.

2.2 Introduction

Bacterial infections often resolve rapidly given effective immune responses, independent of antibiotic treatment. However, in chronic (long-lasting) cases, infections fail to clear even with appropriate drug treatment. Chronic infections impose an elevated morbidity and mortality risk to the individual (Persoon et al., 2004) and an increasing burden on global healthcare systems as at-risk populations grow (Guest et al., 2017). Chronic infections typically arise due to deficits in host barrier defenses and/or immune function, and commonly feature changes in pathogen growth mode (e.g. biofilm formation (Malone et al., 2017)) and additional microbial species acquisition, forming complex multispecies communities (Stacy, McNally, Darch, Brown, & Whiteley, 2016).

Microbiome sequencing has increasingly underscored the polymicrobial context of chronic infection. However, clinical microbiology analysis continues to focus only on the 'usual suspects' of established human pathogens – a relatively short list of organisms with well-established patient health risks. This disconnect between diverse 'infection microbiomes' and limited clinical microbiology profiling may overlook clinically important risk markers. To address this, we focus on chronic lung infections in people with cystic fibrosis (CF).

Cystic fibrosis is an autosomal recessive disease characterized by decreased lung mucociliary clearance and mucus accumulation (Henke & Ratjen, 2007; Perez-Vilar & Boucher, 2004; Rubin, 2010). The resulting environment provides both nutrients for bacterial growth and protection from host immune responses (Bals, Weiner, & Wilson, 1999; Dickson, Martinez, & Huffnagle, 2014; Rieber, Hector, Carevic, & Hartl, 2014; Yonker, Cigana, Hurley, & Bragonzi, 2015), facilitating chronic microbial infections (Conrad et al., 2013; Fodor et al., 2012; Frayman, Armstrong, Grimwood, & Ranganathan, 2017; Lucas, Yang, Dunitz, Boyer, & Hunter, 2018). Accessible 16S rDNA microbiome profiling has shifted CF airway microbiology research away from a historically single-pathogen focus, as sequencing expectorated sputum has revealed diverse communities of tens to hundreds of taxa, including numerous non-pathogenic bacteria (Frayman et al., 2017; Huang & LiPuma, 2016; Whelan et al., 2020).

Numerous lung microbiome studies have linked community composition to disease progression and overall patient health (Acosta et al., 2018; Coburn et al., 2015; Hahn et al., 2020). Cross-sectional studies have shown severe disease is associated with pathogen dominance and loss of taxonomic diversity (Acosta et al., 2018; Coburn et al., 2015; Muhlebach et al., 2018). Longitudinal studies have associated decreasing microbiome diversity with declining lung function (Zhao et al., 2012). Additionally, abundance of non-pathogenic fermentative anaerobes (*Veillonella*, *Prevotella*, *Fusobacterium*) is associated with higher lung function (O'Neill et al., 2015; Zemanick et al., 2015).

While these associations are observed across multiple studies, their causal interpretation is the subject of some controversy. These results may reflect community ecological processes within the lung, where species interactions govern community structure and subsequent harm to the host (Conrad et al., 2013; Klepac-Ceraj et al., 2010; Quinn et al., 2016). Conversely, these patterns could result from oral anaerobe contamination during sample collection (Goddard et al., 2012; Jorth et al., 2019). 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 (Jorth et al., 2019). While recent paired sputum-saliva sampling analysis indicates that oral sample contamination is not a substantial contributor to sputum microbiome profiles in people with established CF lung disease (Lu et al., 2020), these conflicting hypotheses highlight the uncertainty in the role specific taxa present in sputum.

In the current study, we side-step this causal inference problem and instead assess how expectorated sputum microbiome data (including potential oral contaminants) can predict patient lung health using a machine-learning framework. We hypothesize that the addition of non-pathogen data improves the prediction of patient lung function, compared to established pathogen data alone. To address this hypothesis we train predictive models on both lung microbiome and electronic medical record data for a cohort of CF patients. We find that compared to the

benchmark of pathogen data alone, model performance was consistently improved by the addition of non-pathogen taxa.

Table 2.1. 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. BC: Burkholderia, AX: Achromobacter, STE: Stenotrophomonas

	Severe	Moderate	Mild	Normal	P
N	14	25	15	23	
ppFEV1*	32.9	46.5	74.9	101.2	
(RANGE)	(19.7-39.2)	(40.8-59.6)	(61.6-79.8)	(80.4-119.5)	
Age*	31.5	32	24	20	0.007
(RANGE)	(21-61)	(10-63)	(17-51)	(9-66)	
Male	6	11	7	11	
CFTR Genotype					
Homo-dF508	5	12	7	13	
Hetero-dF508	9	10	8	10	
Other/other	0	0	3	0	
BMI*	19.43	20.73	22.23	21.51	0.094
(RANGE)	(16.27-25.69)	(16.70-29.81)	(19.38-26.07)	(16.65-33.91)	
CF-related diabetes	11	14	8	6	0.015
(%)	(78.6)	(56.0)	(53.3)	(26.1)	
HbA1c*	6.25	5.9**	5.7	5.5	0.009
(Range)	(5.3-11.9)	(4.9-8.4)	(5.0-7.6)	(5.1-7.1)	
Clinical Micro					
PA (%)	10 (71.4)	20 (80.0)	10 (66.7)	5 (21.7)	<1.1e-4
SA (%)	8 (57.1)	12 (48.0)	10 (66.7)	16 (69.6)	0.454
MRSA (%)	4 (28.6)	6 (24.0)	6 (40.0)	4 (17.4)	0.486
BC (%)	0 (0.0)	1 (4.0)	1 (6.7)	0 (0.0)	0.553
AX (%)	3 (21.4)	1 (4.0)	0 (0.0)	2 (8.7)	0.147
STE (%)	0 (0.0)	1 (4.0)	3 (20.0)	2 (8.7)	0.191
16S metadata					
% pathogen	0.857	0.589	0.532	0.195	9.05e-5
% Non-pathogen	0.135	0.404	0.597	0.783	3.01e-5

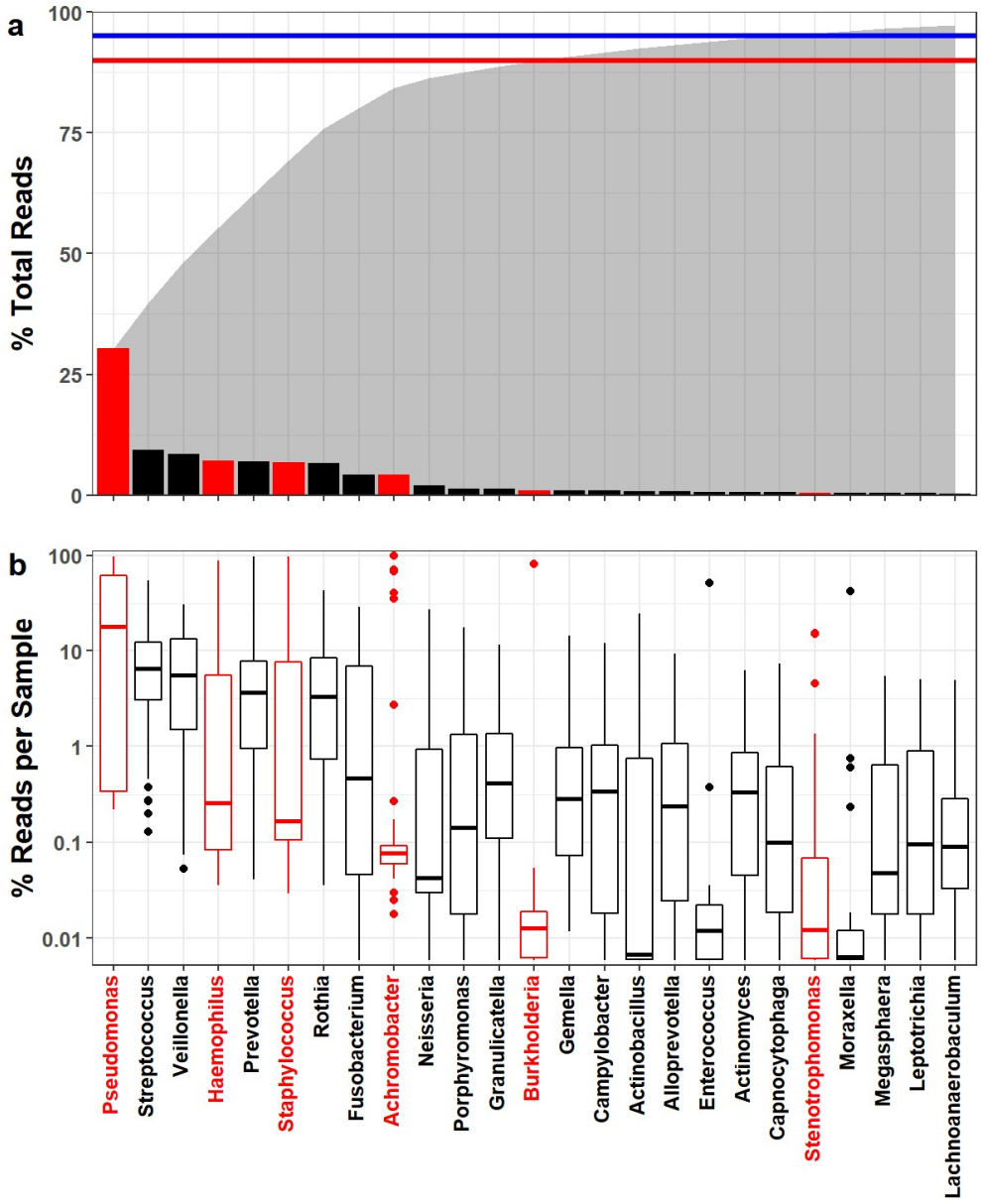

Figure 2.1. CF lung microbiomes are dominated by oral anaerobes and opportunistic pathogens We analyzed CF sputum expectorate (N=77) using 16S sequencing and an in-house QIIME 2-based bioinformatics pipeline to resolve strain-level OTUs. Samples were rarefied to 17000 reads. We identified 217 OTUs across 59 genera and at least 81 species. Overall, we find that CF sputum samples are 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.

2.3 Results

2.3.1 Clinical and microbiome data summary

In total, we obtained sputum expectorates from 77 CF children and adults. Pulmonary function, measured by percent predicted forced expiratory volume in 1 second (ppFEV1), was stratified into four categories from Severe to Normal. A summary of patient information is presented in Table 2.1. As expected, increasing age correlated with worsening lung function (ANOVA; $p < 0.01$). Culture-based detection of *Pseudomonas aeruginosa* correlated with decreasing lung function (ANOVA; $p<0.001$), as did (log-scaled) bacterial load ($p < 0.05$).

The majority (>90%) of reads from our sequencing analysis mapped to one of 13 genera (**Fig 2.1a**), consisting of both recognized CF pathogens (red) and orally derived bacteria (black). *Pseudomonas* sequences accounted for 30.4% of all reads, and were detected in every patient sample. Other established CF pathogens (*Staphylococcus*, *Achromobacter*, *Haemophilus*, *and Burkholderia*) collectively represented 19.3%, while oral taxa account for over 45% (**Fig 2.1**). Total pathogenic and non-pathogenic taxa abundance were both found to vary significantly ($p << 0.001$) with lung function (Table 2.1).

2.3.2 Microbiome Composition Varies with Lung Function

We analyzed microbiome compositions across broad lung function categories to examine the relationship between sputum taxonomic profile and patient health. **Figure 2.2a** highlights the relative compositions of six canonical CF pathogens. As expected, *Pseudomonas* was more prevalent in lungs with reduced function, whereas in normal lungs *Haemophilus* and non-pathogen taxa (gray) were more prevalent. The non-pathogenic composition is consistently dominated by *Veillonella* and *Streptococcus* regardless of lung health or pathogen status (**Fig 2.2b**). Shannon diversity calculated with all taxa present is significantly greater for normal lung function ($p<0.01$, **Fig 2.S1a**), in line with multiple other studies (Carmody et al., 2013; Flight et al., 2015). While Principle Coordinate Analysis (PCA) did not qualitatively separate

compositions by lung function category, we found ppFEV1 was significantly associated with microbiome composition (Mantel test, r=0.195, p<0.001, **Fig 2.S1b**).

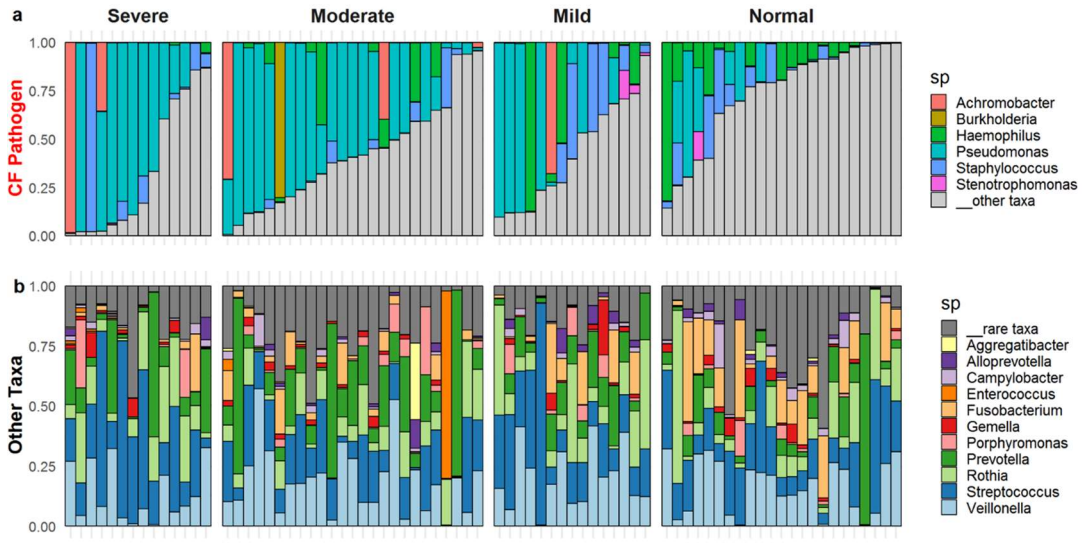

Figure 2.2. CF Lung microbiome composition varies with lung function and pathogen dominance. Relative abundances of **(a)** 6 canonical CF pathogens and **(b)** other taxa (the grey bar taxa in (a)). Microbiome compositions grouped by disease severity, classified by ppFEV1 score: normal (80+), mild (60-80), moderate (40-60), and severe (<40).

2.3.3 Integrating microbiome and patient meta-data

To examine multiple confounding variables such as patient age, BMI or CF-related diabetes (CFRD), we calculated Spearman correlations across 14 microbiome, 11 patient metadata, and 6 clinical microbiology features (**Fig 2.3**). Hierarchical clustering reveals a complex autocorrelation structure, but with many expected consistencies. Overall, there are two main clusters of correlated variables. One correlated with ppFEV1, and included Shannon diversity index 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*. Unsurprisingly, FEV1 and ppFEV1 cluster together and are inversely correlated with ppFEV1 decline rate (an average per year loss in ppFEV1 since birth). Additionally, 16S results for *Pseudomonas*, *Staphylococcus*,

Burkholderia, and *Achromobacter* cluster with their respective culture-based clinical microbiology results. This does not hold for *Stenotrophomonas,* potentially due to its infrequent detection.

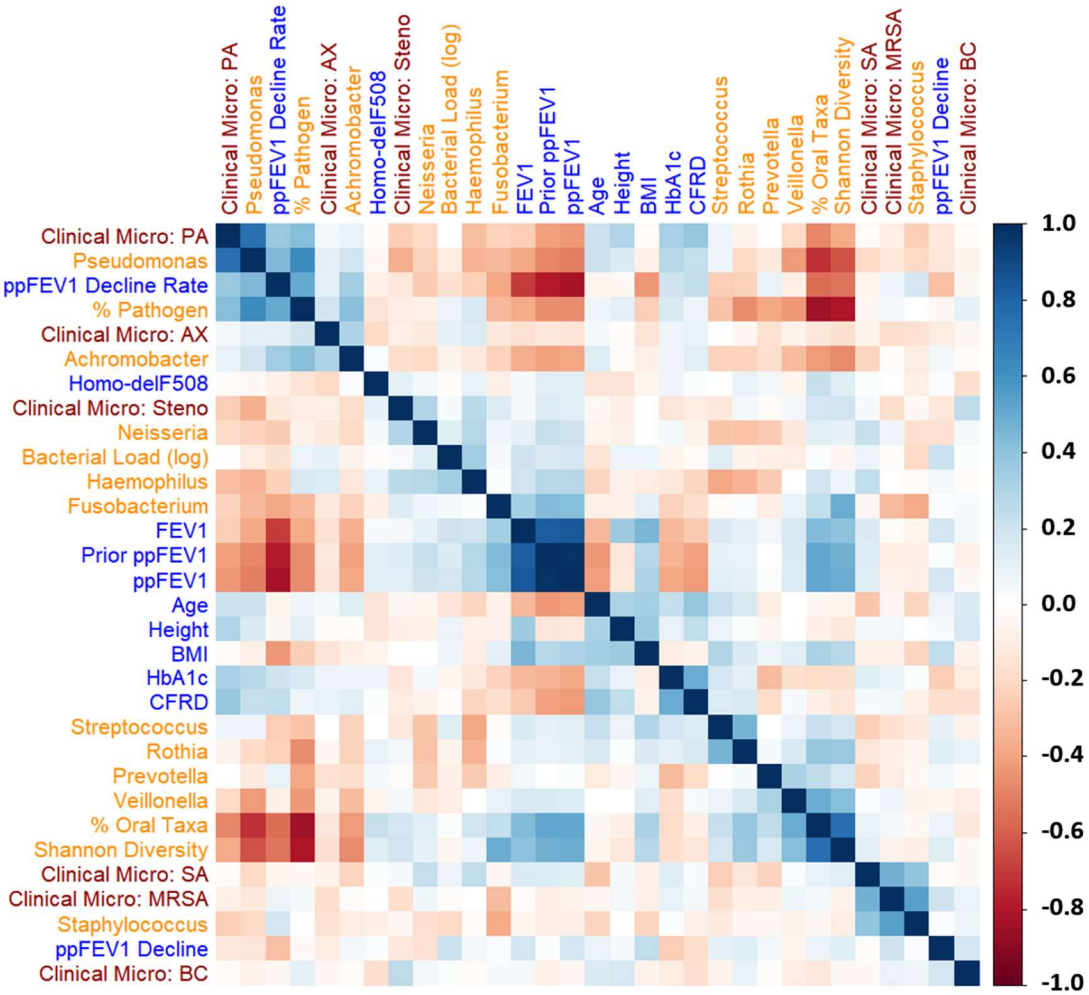

Figure 2.3. Lung function varies with patient meta-data. Spearman correlations (R::corrplot) across all patient metadata (blue), clinical micro results (maroon), and microbiome data (orange, clr-transformed) reveal a complex correlation structure. We used a centered-log transform on 16S data to mitigate compositional effects. Rows and columns were ordered by hierarchical clustering, which identified clusters of metadata and microbiome variables with similar correlation patterns.

2.3.4 Dimensionality Reduction

The correlation matrix in **Figure 2.3** highlights the statistical challenges for identifying meaningful lung function predictors. Such challenges include high between-feature correlations

and relatively few independent patient observations (N=77) compared to the initial number of available predictors (86 total, including 59 bacterial taxa). To mitigate this dimensionality problem, we first restrict our microbiome analysis to only the top 23 genera in our dataset. These top 23 encompass 97% of the total sequenced reads (**Fig 2.1**). We also calculate three additional summary statistics: % pathogen, % oral taxa, and Shannon diversity. As our clustering analysis shows reasonable agreement between clinical microbiology detection and rDNA sequencing, we exclude the binary detection results in favor of 16S quantitation. 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. We refer to this final combination of metadata and 16S data as our "All Features" dataset.

2.3.5 Training Machine Learning Models

To assess if non-pathogenic taxa contain informative biomarkers, we split our samples into 53 training and 24 testing samples. ElasticNet was used to train predict lung function while performing feature selection (see methods, **Fig 2.4**). We expect that the addition of patient metadata (age, BMI etc) will improve our ability to predict lung function given the progressive nature of CF. Our null hypothesis, following the work of Jorth et al. and others(Goddard et al., 2012; Jorth et al., 2019) is that the taxa targeted by clinical microbiology provide adequate explanatory basis for lung function outcomes, and that the addition of non-pathogen 16S data will not improve model predictions.

We test this hypothesis by generating four additional feature subsets (CF Pathogens, All 16S Data, Metadata, and Metadata + Pathogens) and comparing the performance of models trained on each datasets. Initial-pass, non-bootstrapped model training results are shown in **Figures 2.S2** and **2.S3**.

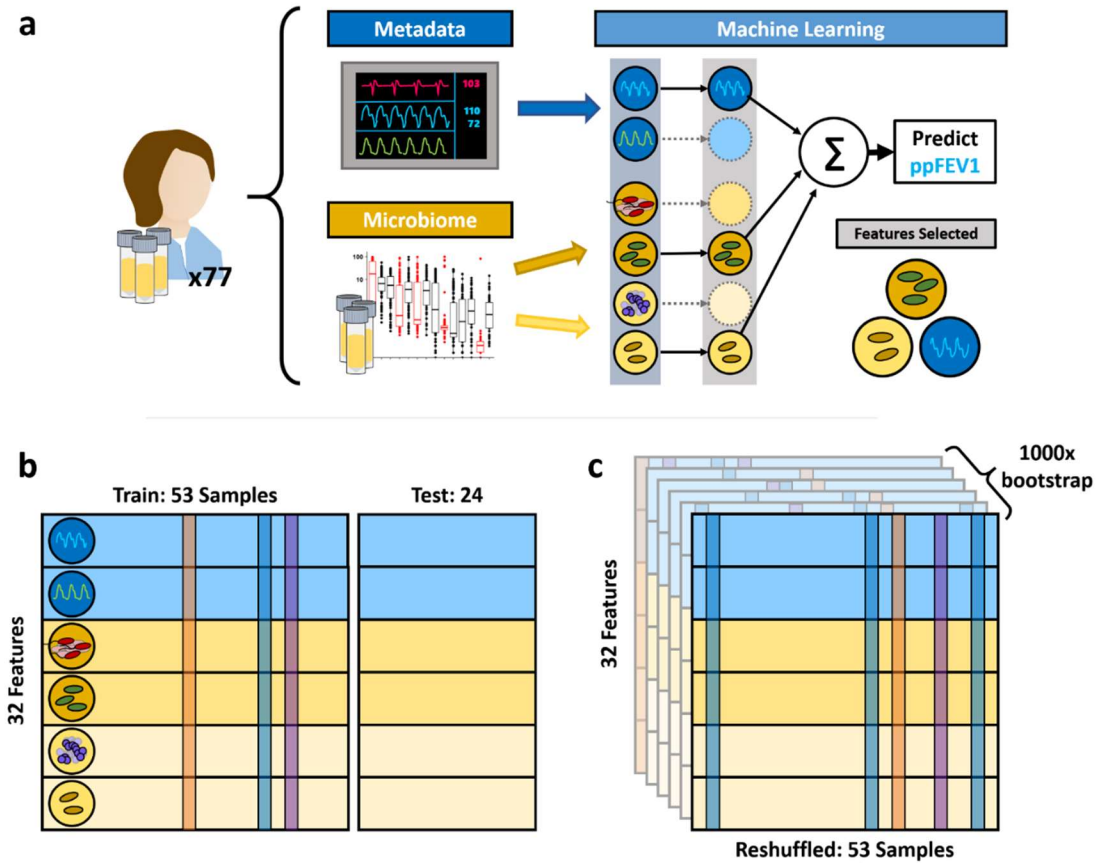

Figure 2.4. Machine Learning Overview. Machine learning models are trained on different input data tables using varying data resampling methods. (**a**) Features are categorized by information source (microbiome or patient metadata). The 16S data is further split into pathogens and other taxa in agreement with **Figure 2.2**. We use elastic net regularization to select informative features that predict ppFEV1. (**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. **(c)** We additionally implement 1000-fold bootstrap resampling to assess the robustness of our model fits.

2.3.6 Model Generalizability

We assess overfitting using leave-one-out cross-validation and compare the prediction error across folds against the test set error. For model robustness, we use 1000-fold bootstrap resampling to fit both a baseline and ensemble of models. Robust features selected by the baseline model will also be selected by a large portion of the bootstrapped ensemble. We additionally standardize all features (mean=0, S.D.=1) to allow for cross-feature comparability. As an additional point of comparison, we generate a non-informative control dataset from the All

Features set using within-feature shuffling, scrambling between-feature correlations while preserving the mean zero, unit variance within-feature structure. **Figure 2.5** shows the results of our baseline (black points) and ensemble (boxplots) approaches. All models using patient metadata or microbiome data outperform the negative control.

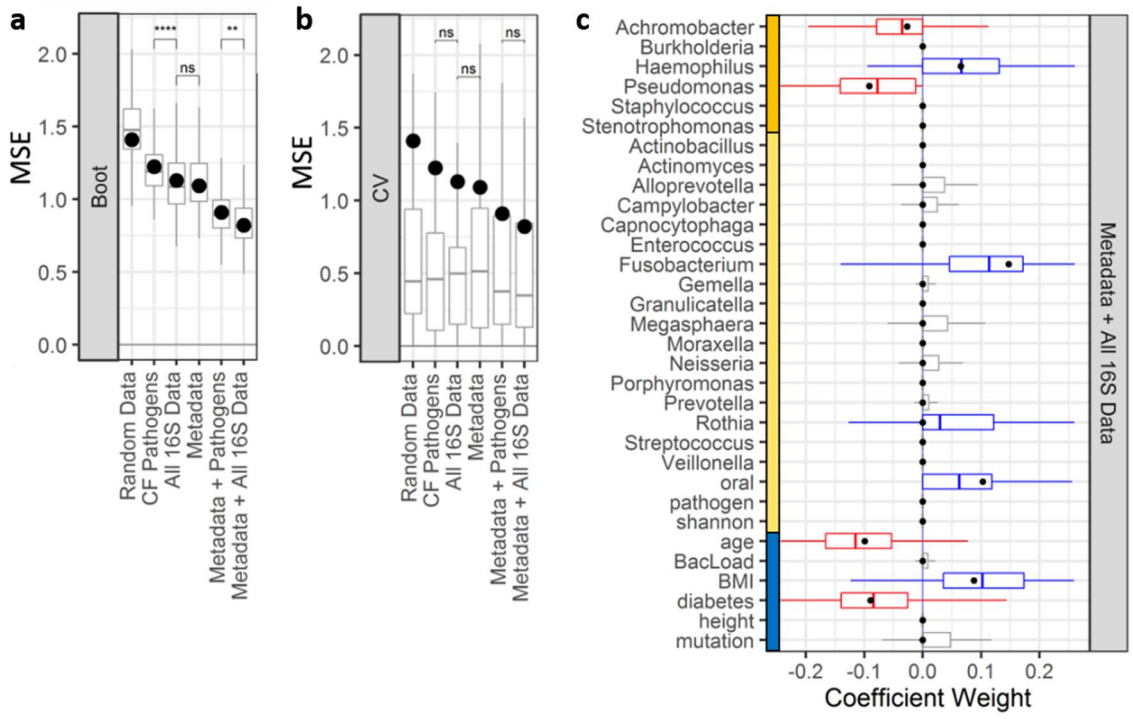

Figure 2.5. Bootstrapped ElasticNet-identified predictors of lung function. ML models were trained using varying input datasets. **a)** 1000-fold bootstrapping and **b)** leave one out cross-validation (LOOCV) were used to generate prediction error (MSE) ranges across feature subsets. Models trained 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)** Coefficient ranges for train/test (black points) and bootstrapped models (boxplots) trained on standardized 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.

2.3.7 Addition of non-pathogen data improves model performance

To address the key question of relative model performance, we find that the addition of non-pathogen taxa significantly improves performance (significantly reduces bootstrapped MSE;

Fig 2.5a), with or without the addition of patient meta-data. Models trained on all 16S quantitation significantly outperform models trained only on pathogen quantitation. Interestingly, while microbiome-only and metadata-only models achieve comparable performance, the combined model achieves greater model performance. Looking broadly across models, we find reasonable consistency in positive and negative predictor selection between our baseline and bootstrapped models (**Fig 2.S4**).

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 2.S4c**) were also selected in the All Features model (**Fig 2.S4g**). 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 2.5c**, 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 2.5b**). We do not find significant differences between cross-validated model errors across our training sets, suggesting that despite the difference in number of available predictors, our models are not overfitting.

2.4 Discussion

Individuals with CF face the challenge of managing long-term chronic infections. Current respiratory management practice is driven by clinical microbiology identification of specific pathogens in throat cultures or expectorated sputum samples, alongside measures of respiratory status (changes in symptoms, signs, and/or lung function). 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 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 2.5a**), and by comparison, the 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. **Figure 2.3** 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. Extensive surveys have documented an association between CF lung function and microbiome diversity, also evident in the current study (**Fig 2.2**). At face value, these results suggest a biological role for these non-pathogen taxa, potentially competing with (Quinn et al., 2016) or facilitating (Flynn, Niccum, Dunitz, & Hunter, 2016) pathogen taxa and therefore indirectly shaping disease outcomes. Jorth et al. recently published a forceful rejection of this 'active microbiome' view, stressing the potential causal role of changing pathogen densities in shaping disease outcomes and viewing shifting diversity metrics as a simple statistical 'relative composition' artifact of shifting pathogen numbers against a roughly constant oral contamination background (Jorth et al., 2019).

While our analyses provide some support for this view, in particular the constancy of the non-pathogen microbiome across patients (**Fig 2.2b**) and the lack of substantial predictive improvement on addition of non-pathogen data (**Fig 2.5b),** we also see lines of evidence against the contamination hypothesis. First, our use of center log transformations mitigates the risk of spurious associations due to compositionality (refs) and yet non-pathogen taxa are still consistently retained. Second, the contamination hypothesis predicts total bacterial burden to be an important predictor, and yet burden was not retained in our models. Third, our observation of a consistent retention of specific non-pathogen taxa across multiple models (with and without the addition of potentially confounding EMR features, including age and BMI) points to the potential for a distinct causal pathway that is orthologous to age or BMI. We note that the interpretation that oral bacteria are active players in the lung environment is further buttressed by a recent study on people with established CF disease (Lu et al. 2020) that used paired sputum and saliva samples to infer the presence of substantial populations of oral bacteria in the lung.

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. When we include features already known to correlate with lung health, such as age, BMI, and CFRD status, our models not only these features, but additionally retain *Rothia* and *Fusobacterium* as positive predictors. The retention of these specific taxa in both this full model and in partial models (**Fig 2.S4**) 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 (McNally & Brown, 2015).

Interestingly, our All Features 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 (Bals et al., 1999; Bogaert et al., 2011), hence we would expect a positive association in a model that is not controlled for age (**Fig 2.S4c, S4d**). 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. 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'). **Figure 2.2a** illustrates that we only appreciably detect two and rarely three coexisting pathogens of the six we find across all patients. The relative scarcity of multi-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 2.3**). 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 (Banerjee, Garai, Mallick, Chowdhury, & Chatterjee, 2018; Leng, Tran, & Nott, 2014; Rush, Lee, Mio, & Kim, 2016). This is in part due to the compositionality constraint of microbiome data, which can be mitigated by using absolute quantitation (Jian, Luukkonen, Yki-Jarvinen, Salonen, & Korpela, 2018). 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 some bootstrapped models selected bacterial load as a positive predictor (**Fig 2.5c**), 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 (Goddard et al., 2012; Jorth et al., 2019).

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 patient medical record 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.

We note that the major predictors identified in our models have been identified in various studies, and taken piecemeal there is less insight. The value of this work lies in the systematic integration of these multiple data sources (from both EMR and microbiome data sources). Our model comparisons (with / without EMR predictors) allow an assessment of the impact of oral bacteria, with and without key potential confounds. Ignoring these confounds could lead to spurious retention of microbiome taxa that correlate strongly with e.g. age or BMI. In addition, our analyses allows assessment of disparate factors on a common predictive scale – indicating for example that the impact of 1 standard deviation shift in *Fusobacterium* abundance is comparable to a 1 S.D. shift in BMI. Our model comparison approach lends more confidence to the conclusion that the retained oral taxa are associated with patient outcomes via causal pathways that are largely independent of age or BMI, being robust to their presence or absence in the predictive models. The research agenda of pursuing the nature of the causal pathways linking oral bacteria in the lung with patient outcomes is now on a firmer footing as a result of our study.

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.

2.5 Methods

2.5.1 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 (IRB00042577).

2.5.2 Sample collection and 16S analysis

Expectorated sputum samples were obtained from CF patients attending the Children's Healthcare of Atlanta and Emory University CF Care Center from January 2015 to August 2016. De-identified patient information including age, sex, height, BMI, CFTR genotype, degree of glucose control (HbA1c), and ppFEV1) were obtained (Table 2.1). 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 no increase in respiratory symptoms compared to baseline, and no acute illness or new medication for three weeks prior to sputum collection. Upon collection, sputum 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. Microbiology culture results were obtained for sputum samples sent to the Clinical Microbiology laboratory on the same day as samples for sequencing were collected.

DNA was purified from sputum homogenate with the MoBio Power Soil kit (MoBio, Carlsbad, CA). The 16S V4 region was amplified and sequenced using Illumina MiSeq, yielding an average of 137,708 sequences per sample. Sequences were quality filtered and amplicon sequence variants were obtained using the QIIME2 deblur plugin. Taxonomic assignments were classified against both SILVA and Greengenes 16S reference databases and assigned based on highest taxonomic resolution. To mitigate compositional effects, 16S data were center-log transformed prior to all analyses. Nucleotides are uploaded to BioProject accession no. PRJNA666192.

2.5.3 Statistical and Quantitative Analysis

Patient samples were binned by ppFEV1 (Normal: >80%, Mild: 80-60%, Moderate: 60-40%, Severe: <40%). Variance across lung function categories in patient metadata 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 agglomerated to the genus level (McMurdie & Holmes, 2013). Associations between continuous variables were tested using Spearman correlations. A full pairwise correlation matrix was calculated, with rows and columns ordered by hierarchical clustering (Wei et al., 2017).

2.5.4 Machine Learning

We use ElasticNet to fit regularized linear models predicting lung function (ppFEV1) from patient metadata, microbiome composition, and clinical microbiology results (Yuan, Ho, & Lin, 2011). ElasticNet solves a penalized linear regression model using a weighted average of L1 (LASSO) and L2 (ridge regression) penalties. This limits over-fitting by penalizing non-zero coefficients. We split our samples using a simple 70:30 train-test holdout, where models are

trained on 53 samples and used to predict on the remaining 24. All input features were standardized (mean=0, S.D.=1) prior to model training to allow between-feature interpretability. From our full dataset, we create 4 additional data subsets: CF Pathogens, All 16S Data, Metadata, and Metadata + Pathogens. We include within-feature shuffling on the full set as a non-informative negative control.

We employ two methods to assess model robustness, and compare model performance using mean squared error (MSE). We generate 1000 bootstrap resampled sets from the training set and fit an ensemble of regularized linear models to obtain distributions for each model coefficient. We identify key metadata and taxa robustly selected (nonzero coefficients) across the ensemble of models. We assess model generalizability using leave-one-out cross-validation on the training set and compare resulting MSE ranges.

2.6 Acknowledgements

We thank Karan Kapuria and Eunbi Park for help with bio-informatic pipeline development, and Peng Qiu for machine learning guidance. We acknowledge the technical support and provision of the clinical data by the CF Biospecimen Registry at the Children's Healthcare of Atlanta and Emory University CF Discovery Core. This research was supported in part through research cyber-infrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at Georgia Tech.

Chapter 3: Antibiotics drive expansion of rare pathogens in a chronic infection microbiome model[2]

3.1 Summary

Chronic (long-lasting) infections are globally a major and rising cause of morbidity and mortality. Unlike typical acute infections, chronic infections are ecologically diverse, characterized by the presence of a polymicrobial mix of opportunistic pathogens and human-associated commensals. To address the challenge of chronic infection microbiomes, we focus on a particularly well-characterized disease, cystic fibrosis (CF), where polymicrobial lung infections persist for decades despite frequent exposure to antibiotics. Epidemiological analyses point to conflicting results on the benefits of antibiotic treatment yet are confounded by the dependency of antibiotic exposures on prior pathogen presence, limiting their ability to draw causal inferences on the relationships between antibiotic exposure and pathogen dynamics. To address this limitation, we develop a synthetic infection microbiome model representing CF metacommunity diversity, and benchmark on clinical data. We show that in the absence of antibiotics, replicate microbiome structures in a synthetic sputum medium are highly repeatable and dominated by oral commensals. In contrast, challenge with physiologically relevant antibiotic doses leads to substantial community perturbation characterized by multiple alternate pathogen-dominant states and enrichment of drug-resistant species. These results provide evidence that antibiotics can drive the expansion (via competitive release) of previously rare opportunistic pathogens and offer a path towards microbiome-informed conditional treatment strategies.

[2] This chapter was adapted from the following reference: Varga, J. J., Zhao, C., Davis, J. D., Hao, Y., Farrell, J. M., Gurney, J. R., Voit, E., & Brown, S. P. (2021). Antibiotics drive expansion of rare pathogens in a chronic infection microbiome model. BioRxiv, 2021.06.21.449018. Reused with permission. John Varga, Jacob Davis, and I were joint primary authors of this work.

3.2 Introduction

Physicians face two growing crises that impact their ability to treat bacterial infections. The first is widely recognized – the evolution of antibiotic resistance (Laxmin arayan et al., 2013). The second receives less attention – chronic (long-lasting) infections that are more difficult to treat (Filkins et al., 2015; Siddiqui et al., 2010; Young et al., 2002). Chronic infections are globally a rising burden on health-care systems due to increases in populations at risk (e.g., the elderly, people with diabetes or other chronic diseases) (Guest et al., 2017). At-risk populations have deficits in host-barrier defenses and/or immune function that provide an opening for the establishment of infections, and these chronic infections are further complicated by changes in pathogen growth mode (e.g., formation of multicellular biofilm-like aggregates (Bjarnsholt et al., 2013; Darch et al., 2017; Kragh et al., 2014)) and development of complex multispecies communities (Stacy et al., 2016).

To address the global challenge of chronic infections, we focus on a particularly well-characterized disease, cystic fibrosis (CF), where bacterial infections can persist for decades. CF is caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR), an ion channel that conducts chloride and thiocyanate ions across epithelial cell membranes, leading to defective mucociliary clearance and polymicrobial infection (Henke & Ratjen, 2007; Perez-Vilar & Boucher, 2004), resulting in eventual pulmonary failure (Surette, 2014; Yonker et al., 2015).

Traditionally, CF research and patient care have focused on a small cohort of opportunistic pathogens, highlighting a distinct successional pattern (CFF, 2019) characterized by peak prevalence of *Haemophilus influenzae* in childhood, *Staphylococcus aureus* during adolescence and *Pseudomonas aeruginosa* in adulthood. In addition to the core pathogen species, 16S rDNA amplicon sequencing of expectorated sputum samples has revealed much more diverse communities including numerous bacteria that are considered non-pathogenic in CF and that are normally associated with oral and upper-respiratory environments (Filkins et al., 2012; Fodor et

al., 2012; Frayman et al., 2017; Huang & LiPuma, 2016; Lucas et al., 2018). The functional role of these non-pathogenic taxa in CF lungs is currently disputed (Caverly & LiPuma, 2018). Epidemiological analyses have identified potentially positive roles, as higher lung function correlates with higher relative abundance of oral bacteria in sputum samples from both cross-sectional (Acosta et al., 2018; Coburn et al., 2015; Muhlebach, Zorn, et al., 2018) and longitudinal studies (Zhao et al., 2012). In contrast, *in vitro* experimental studies have suggested health risks of specific oral bacteria in the lung, due to the potential facilitation of pathogen growth (Adamowicz et al., 2018; Flynn et al., 2016). A third interpretation is that oral bacteria found in sputum are simply the result of sample contamination with oral microbes during expectoration (Goddard et al., 2012; Jorth et al., 2019). A number of approaches to address the sputum contamination issue have been taken, including mouth cleaning and sputum rinsing (Rogers et al., 2006), as well as more invasive sampling techniques (subject to clinical need (Hogan et al., 2016; Jorth et al., 2019; Muhlebach, Hatch, et al., 2018; Zemanick et al., 2015)). Most recently, computational analysis of paired sputum and saliva samples from adults with established CF lung disease has demonstrated that saliva contamination during sample collection has a minimal quantitative impact on the community profile (Lu et al., 2020).

As a result of long-term bacterial infection, people with CF are exposed to high levels of antibiotics (Chmiel et al., 2014), both as maintenance therapy (Waters, 2018) and as treatment for exacerbations. In the context of a critical health challenge (an acute pulmonary exacerbation), health outcomes are variable – lung function can rapidly increase back to baseline values or remain at a new, lower baseline following antibiotic intervention. Unfortunately, a recent systematic review of 25 articles indicated little correlation between these variable clinical outcomes and antibiotic susceptibility test results for the target pathogen (Somayaji et al., 2019). Several factors for this disconnect have been proposed, including differences in bacterial physiology (Kidd et al., 2018), non-representative infection sampling (Darch et al., 2015; Jorth et al., 2015) and polymicrobial interactions (Waters et al., 2019). In a microbiome context,

epidemiological studies indicate variable outcomes of antibiotic treatment, ranging from minimal impact on microbiome structure (Cuthbertson et al., 2016; Fodor et al., 2012; Price et al., 2013) to target pathogen declines, microbiome structural changes (Hahn et al., 2019; Nelson et al., 2020; Smith et al., 2014; Zemanick et al., 2013) and risk of subsequent infection (Taccetti et al., 2012). However, there is a fundamental confounding factor in these epidemiological studies, as antibiotic exposures are themselves dependent on the microbiome state of the patient. Specifically, the detection of pathogens within the microbiome will dictate antibiotic choice (Morley et al., 2019).

Here we seek to overcome this confounding impact of pathogen detection through the development and benchmarking of a clinically relevant experimental infection microbiome model. Using this model we seek to address a number of broad and overlapping questions concerning the determinants of infection microbiome structure: (1) Can a single experimental model generate multiple alternate infection microbiome states? (2) What are the impacts of independent pathogen and antibiotic manipulations on microbiome structure? (3) Can antibiotics drive pathogen expansion and community diversification, via competitive release? (4) Do antibiotics promote facilitatory species interactions?

While most experimental polymicrobial models of CF have focused on two species pathogen interactions (Hotterbeekx et al., 2017; Limoli et al., 2017; Nguyen & Oglesby-Sherrouse, 2016), some studies have developed up to 6-species models (Vandeplassche et al., 2017; Vandeplassche et al., 2020). These more complex models have demonstrated that species antibiotic susceptibility is not impacted by community context (Vandeplassche et al., 2020), but their use of rich media (to facilitate single-species comparisons) raises the issue of relevance to the *in vivo* context of growth in sputum (Jean-Pierre et al., 2021). Our experimental approach begins with a "synthetic sputum" that recreates the biochemical and physical conditions of the sputum found in CF lungs (Palmer et al., 2005; Turner et al., 2015). We then add defined combinations of the 10 most abundant bacterial species on the meta-community scale: 10 species

that together account for over 85% of the observed bacterial diversity within the CF lung in a 77 person cohort (Zhao et al., 2020). Five of these species are established human pathogens (*S. aureus*, *P. aeruginosa*, *H. influenzae, Burkholderia cenocepacia*, and *Achromobacter xylosoxidans*), while the rest are oral microbes frequently found in CF lungs. To underline that our 10 bacterial species model captures observed CF diversity at the meta-community (multi-patient) scale, we refer to this model as the CF meta-community model (see schematic **Fig 3.1**). We hypothesize that this single experimental model can self-organize into multiple alternate community states that approach the diversity of microbiome states observed across individuals with CF.

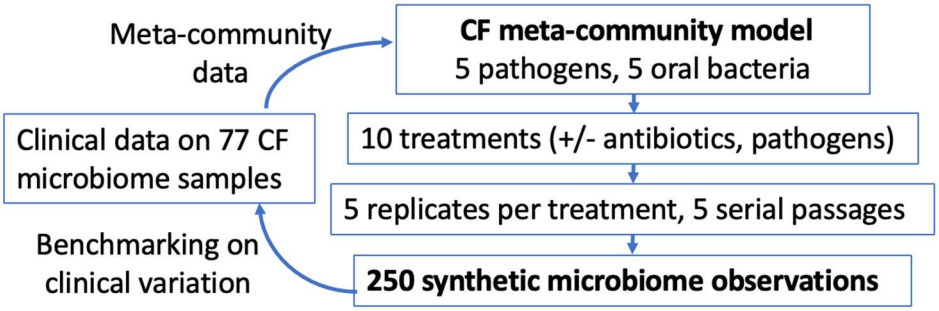

Figure 3.1 Schematic outline of the CF meta-community approach. All experiments are derived from a 10 species menu that captures the majority of CF microbiome diversity across a cohort of 77 people with CF (Zhao et al., 2020), and is consistent with microbiome content across the CF literature (Acosta et al., 2018; Coburn et al., 2015; Filkins et al., 2012; Fodor et al., 2012; Frayman et al., 2017; Huang & LiPuma, 2016; Lucas et al., 2018; Muhlebach, Zorn, et al., 2018; Zhao et al., 2012). The 10 species meta-community is exposed to 10 treatments (in 5x replication), propagated for 5 serial passages. The experimental design results in 250 individual synthetic microbiome observations.

Replicate communities are cultured anaerobically to capture oxygen-depleted conditions within mucus plugs (Cowley et al., 2015; Kolpen et al., 2010; Worlitzsch et al., 2002). We show that under our *in vitro* model infection conditions, oral bacteria form stable communities that suppress the growth of multiple pathogen species, and this competitive suppression is reduced by controlled antibiotic exposures, leading to multiple alternate pathogen-dominant outcomes, the emergence of facilitatory species interactions and the non-evolutionary enrichment of antibiotic resistance.

3.3 Results

3.3.1 In the absence of antibiotics, commensal anaerobes dominate over CF pathogens

Experiments performed in the absence of antibiotics demonstrated a consistent and reproducible community structure, characterized by population expansion during the initial 48 h and a composition primarily consisting of *P. melaninogenica*, *H. influenzae*, and *V. parvula* **(Fig 3.2**). At 48 h, the total bacterial density averaged about ~7.7×10^6 CFU/ml (+/- 2.0×10^6 SD), which falls within the broad range of reported bacterial densities in sputum in clinical studies (typically between 10^4 and 10^9 CFU/ml (Tunney et al., 2008; Wong et al., 1984; Zhao et al., 2020)). From passage 3 onwards, each replicate showed a high degree of stability through time, both in terms of total abundance and relative composition. Across replicates, we also see a striking convergence in microbiome structure. To assess consistency across the 5 replicates, we calculated coefficients of variation (CV = standard deviation / mean) for each species' total abundance, all showing under-dispersion (i.e. standard deviation less than the mean, with an average species CV of 0.46 at end of the experiment, see **Fig 3.S1**), consistent with stabilizing ecological forces limiting variation in species densities across replicates.

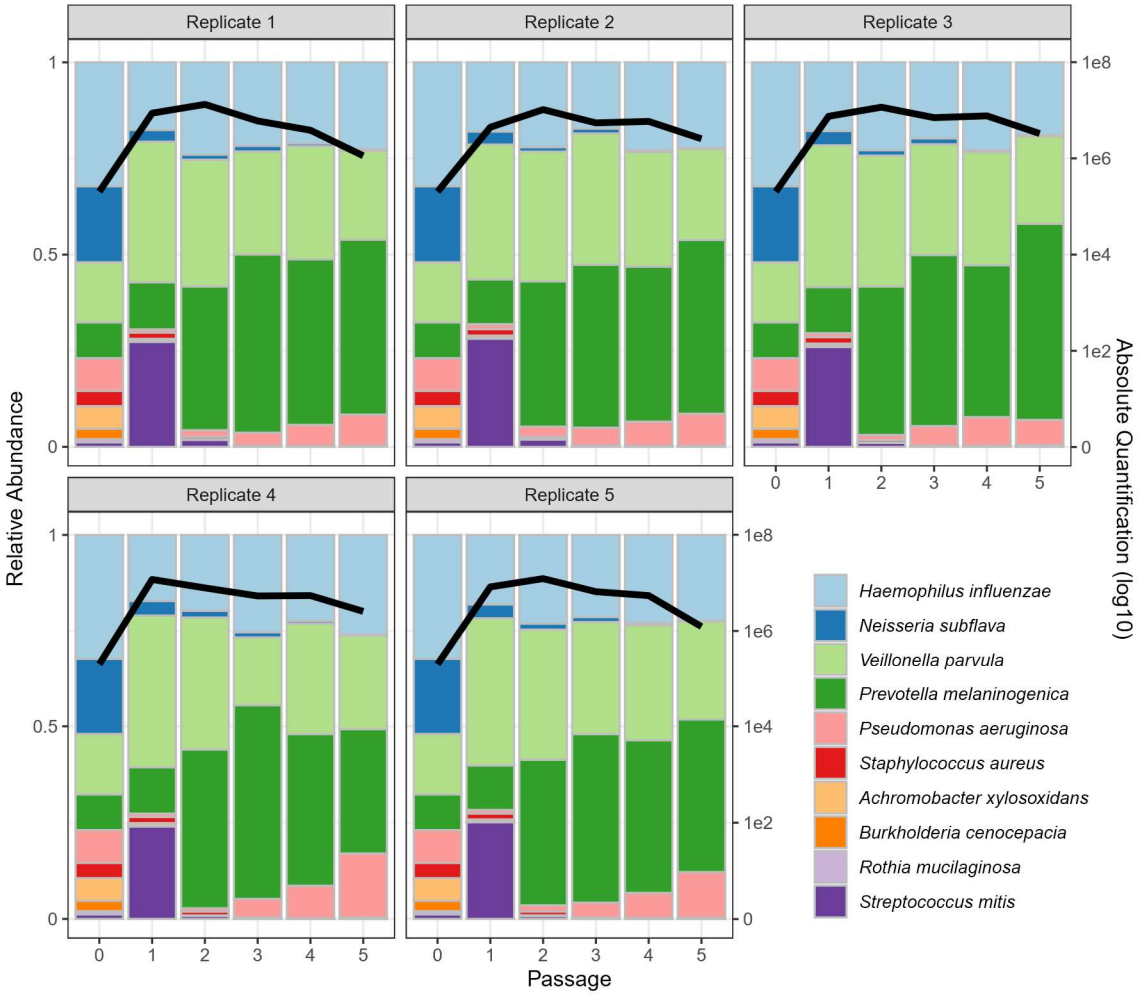

Figure 3.2. Five-fold replicated synthetic CF microbiomes converge toward a single stable state in the absence of antibiotic perturbations. 5 replicate synthetic microbiomes were grown anaerobically in artificial sputum medium. The community composition was estimated by 16S rDNA amplicon sequencing at time zero and at every two-day passage (x-axes) into fresh medium (10% transfer of 2 ml culture volume). The colored bars represent relative abundance of each species in the community (left y-axis), while the black line represents the total bacterial abundance per mL (right y-axis, log scale). Each panel represents a separate replicate experiment. Strain information is provided in Table 3.1 (our default *P. aeruginosa* strain is mucoid PDO300).

The results in **Figure 3.2** point to a robust community structure in the absence of perturbations, consistent with the frequent dominance of oral bacteria in individuals with higher lung function but far from capturing the diversity of microbiome structures observed across the broader CF community, in particular our results do not recapitulate the common observation of variably pathogen-dominated microbiomes (Coburn et al., 2015; O'Neill et al., 2015; Zemanick et

al., 2015; Zhao et al., 2020; Zhao et al., 2012). To assess the role of variable pathogen strain identity or presence / absence, we repeated the experiments in **Figure 3.2** with wildtype (non-mucoid) PAO1, and also with variation in the presence or absence of *S. aureus*. In light of previous experimental work demonstrating that single-locus changes impacting biofilm phenotypes (like mucoidy) can have dramatic community ecological impacts exceeding the impact of species removal (McClean et al., 2015), we hypothesized that the presence/absence of mucoidy would generate substantial community shifts, exceeding removal of *S. aureus* and/or *P. aeruginosa.* In contrast to this hypothesis, we found very small quantitative variations in community structure under all pathogen manipulations **(Fig 3.3**), and no support for the prediction of a greater impact of mucoidy versus species removal (**Table 3.S1**). Across all pathogen treatments we observed overall the same qualitative pattern as in **Figure 3.2** with consistent dominance by *H. influenzae*, *P. melaninogenica*, and *V. parvula **(Fig*** **3.3**, **Figure 3.S2**).

3.3.2 Antibiotics skew community structure toward pathogen expansion and dominance

Having established the repeatability and stability of the community in the absence of antibiotics, we then assessed the impact of antibiotic treatment on community structure. To test our hypothesis that antibiotic exposure will induce substantial community perturbations, communities were continually challenged with 3 individual antibiotics and 2 pairs commonly used in the CF clinic (tobramycin, meropenem, ciprofloxacin, tobramycin and meropenem, tobramycin and ciprofloxacin) (Chmiel et al., 2014; Doring et al., 2012) in physiologically relevant concentrations (Cipolla et al., 2016; Kuti et al., 2004; Moriarty et al., 2007; Ruddy et al., 2013; Wenzler et al., 2015). Consistent with our hypothesis, antibiotic exposures resulted in dramatically different outcomes across treatments and replicates, compared to the antibiotic free communities **(Fig 3.4**, S1). To quantify community-scale impacts of antibiotic perturbations (compared to to the no antibiotic control treatment, **Figure 3.2**) we use the analysis of similarity

(ANOSIM) *R* metric, revealing significant impacts on community structure, exceeding the impacts of pathogen treatments **(Fig 3.S3**). Antibiotic effect sizes range from modest impacts of tobramycin (mirroring clinical data (Heirali et al., 2020; Nelson et al., 2020) to substantial impacts for treatments involving meropenem.

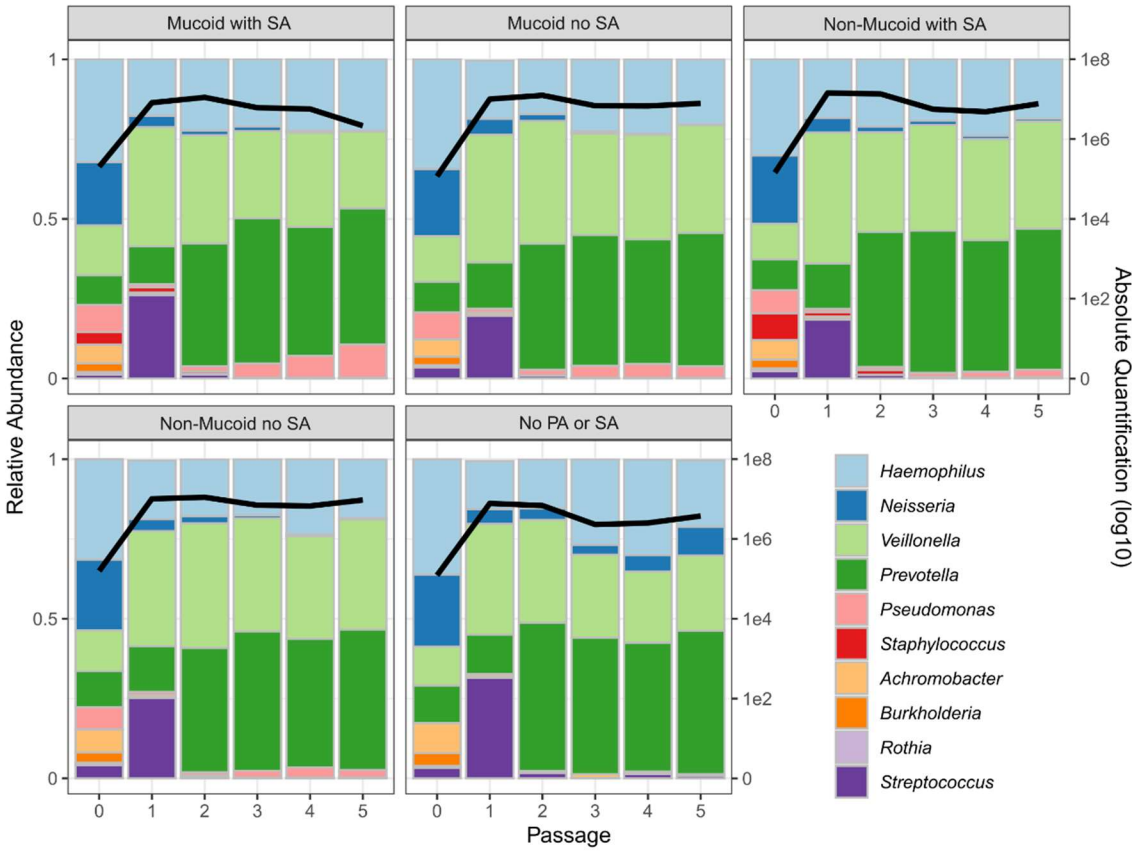

Figure 3.3. Varying the pathogen composition has minimal impact on community composition. Each panel represents the average of 5 replicates in the absence of antibiotics, the mucoid PA with SA panel is the average of **Figure 3.1**. Figure details are the same as described for **Figure 3.1.** Data on individual replicates per treatment are presented in **Figure 3.S2**. Mucoid and non-mucoid PA = *P. aeruginosa* strains PAO1 and PDO300, respectively. SA = *S. aureus*.

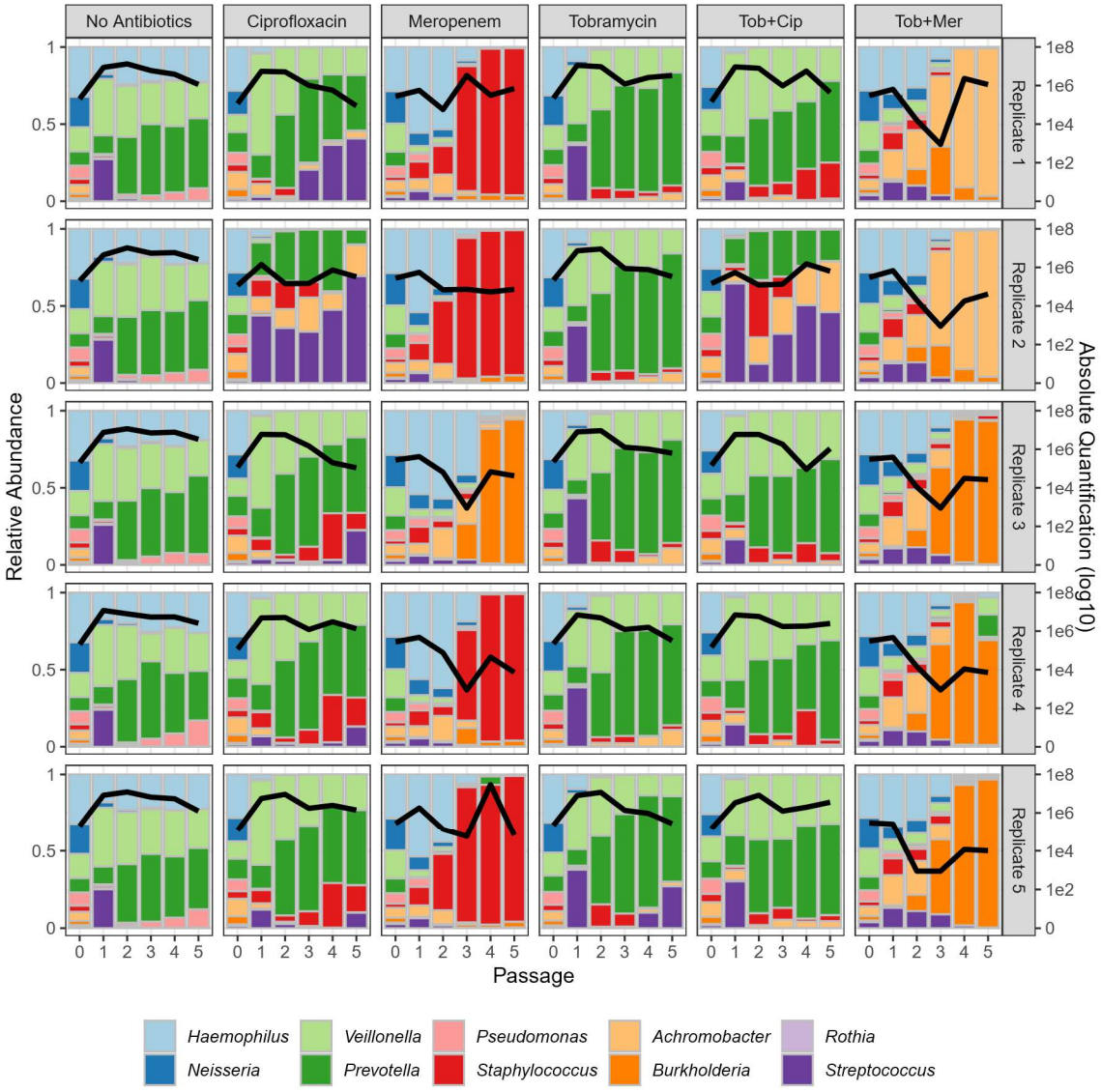

Figure 3.4. Antibiotic treatments produce large community fluctuations and alternative community states. Columns represent distinct antibiotic treatments (the first 'no antibiotics' control column is reproduced from **Figure 3.1**), rows represent 5 replicates. The left axes measure community composition (bar charts), the right axes measure total bacterial abundance per mL (black lines). Experimental procedures, sampling, and analysis were performed as described in **Figure 3.1**. Fresh antibiotics were re-supplemented at each passage. Total abundance data by species is presented for each treatment and timepoint in **Figure 3.S3**.

The compositional presentation in **Figure 3.4** highlights that the same antibiotic treatment often leads to distinct pathogen dominance across replicates. For example, under meropenem 4 out of 5 replicates result in persistent *S. aureus* dominance, while one replicate

shows persistent *B. cenocepacia* dominance. One possibility is that these distinct endpoints represent alternative *stable* states, implying stabilizing ecological forces sending separate replicates towards *B. cenocepacia* dominance (and *S. aureus* absence) or *S. aureus* dominance (and *B. cenocepacia* absence), dependent on fluctuations in initial conditions (Estrela et al., 2022). To further investigate this claim, we turn to taxon absolute abundance data to test the 'alternative stable states' prediction of a negative correlation between *B. cenocepacia* and *S. aureus* absolute abundances, across replicates. **Figure 3.S4** presents absolute abundance data for each taxon, treatment and replicate through time. Under the meropenem data **(Fig 3.S4**C) we can see substantial variation in the final abundance of *B. cenocepacia* and *S. aureus (*see also **Figure 3.S1**). However, across final abundances of these two taxa, there is no negative correlation between the absolute density of *B. cenocepacia* and *S. aureus (*Pearson's correl coeff = 0.026, p = 0.967). Under the meropenem/tobramycin treatment we see a similar pattern of variable dominance between *A. xylosidans* and *B. cenocepacia* (see **Figs 3.4, 3.S1** and **3.S4F**), but again no negative correlation across replicates in final absolute abundances (Pearson's correl coeff = 0.310, p = 0.611). In light of this analysis, our data rules against alternative stable states, and indicates that the variable dominance across replicates under drug exposure is a result of relaxed ecological regulation resulting in increased cross-replicate noise.

Figure 3.4 indicates large shifts in response to antibiotic treatments, but compositional analysis alone cannot separate the relative importance of differential survival versus differential expansion. Using absolute abundances **(Fig 3.S4**), we can now test whether pathogens undergo competitive release (expansion, following removal of competitors (Aspenberg et al., 2019; de Roode et al., 2004; Wale et al., 2017)) in response to antibiotic exposure, by assessing whether the final pathogen density is greater in the presence of antibiotic compared to its absence **(Fig 3.5**). Comparing densities in the presence/absence of antibiotics, we find evidence for significant and substantial (>100-fold in some replicates) antibiotic-dependent amplification via competitive release of *S. aureus*, *B. cenocepacia* and *A. xylosoxidans* under specific antibiotic exposures **(Fig**

3.5). In contrast, there is evidence of significant suppression of *H. influenzae* and *P. aeruginosa* in all antibiotic treatments **(Fig 3.5**; two-tailed Wilcoxon test, $p < 0.01$), together with *N. subflava* in all treatments as well as *V. parvula* and *P. melaninogenica* in all meropenem treatments **(Fig 3.S5**).

3.3.3 Antibiotic susceptibility explains community composition on a functional scale, but not on a taxon scale.

The simplest hypothesis to account for the substantial impacts of antibiotic exposures on community structure (**Figs 3.4, 3.5, 3.S4**) is that antibiotics present a survival filter through which only resistant organisms can pass. Under this model, the community structure after antibiotic treatment is simply the product of whether or not each taxon can grow in the antibiotic (s) administered.

To assess the survival filter hypothesis, we derived antibiotic susceptibility measures (minimal inhibitory concentrations, MICs) under standard growth conditions that allowed the more fastidious strains to grow independently (**Table 3.S2**) and used these data to predict experimental responses to defined antibiotic exposures **(Fig 3.6**). **Figure 3.6** illustrates that the drug susceptible *P. aeruginosa* lab strain PDO300 behaves as predicted by the survival filter hypothesis – it is present in the absence of treatment but then absent (average relative abundance is <1%) in the presence of antibiotics. The same is true for *H. influenzae*.

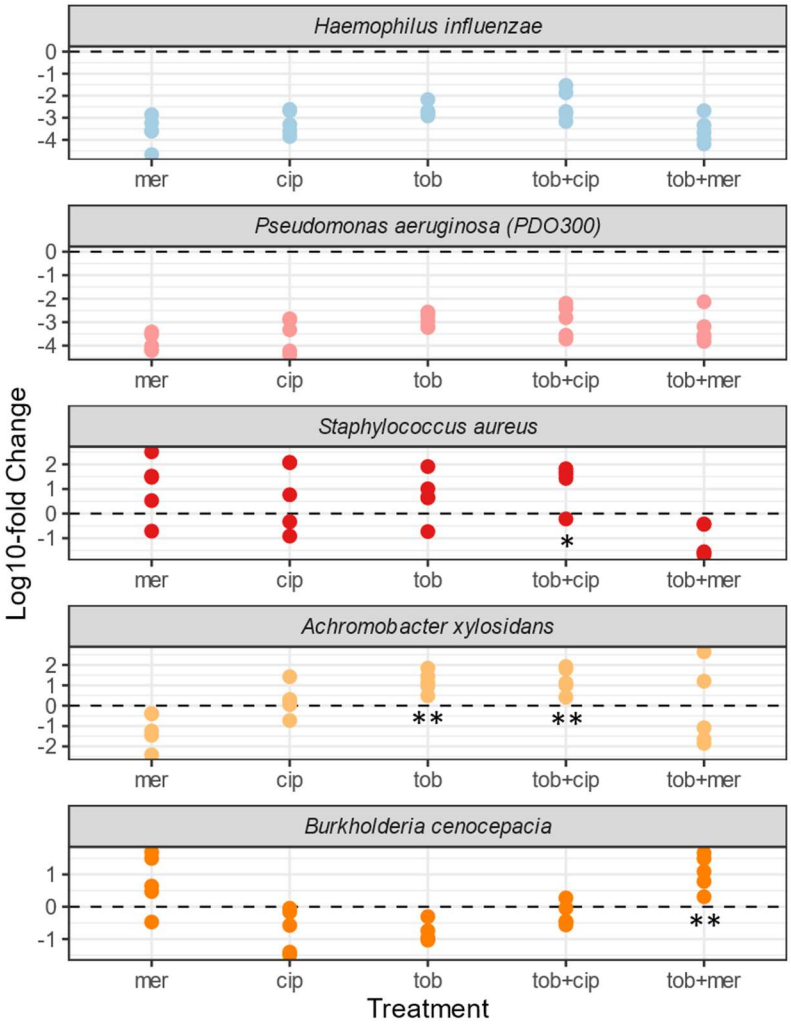

Figure 3.5. Absolute pathogen densities are variable and often increased under antibiotic exposures. Each dot corresponds to the fold-change difference of an individual replicate of species-specific final time-point absolute density under defined antibiotic treatments, compared to the mean value of the no antibiotic control (data redrawn from **Figure 3.3**). Mer = meropenem, cip = ciprofloxacin, tob = tobramycin. Asterisks denote significantly higher final densities in presence of antibiotic, compared to antibiotic-free controls (competitive release; one-tailed Wilcoxon test * $p < 0.05$, ** $p < 0.01$).

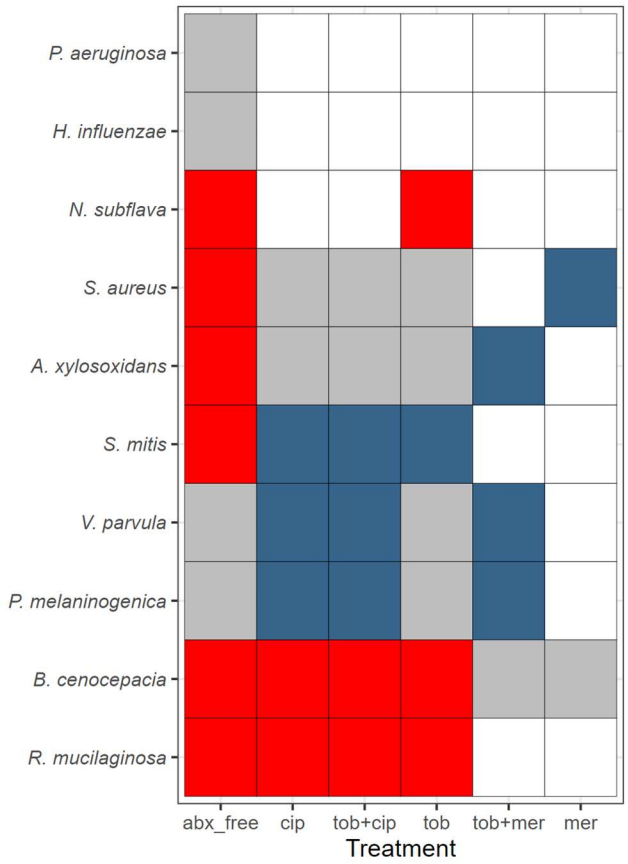

Figure 3.6. Antibiotic resistance testing does not consistently predict species presence/absence in a community context. For each species / drug combination, we assessed predicted survival (MIC in rich medium (Table 3.S2) > experimental concentration) and observed survival (relative abundance of at least 1% averaged across all five replicates at the final time point **(Fig 3.4)**). True positive cases (predicted and observed present) are coded in grey, true negatives (predicted and observed absent) in white. False positives (predicted present, observed absent – evidence for competition) are in red, and false negatives (predicted absent, observed present – evidence for facilitation) are in blue. Species order was determined through clustering via stringdist (van der Loo, 2014). In the discussion we address the caveat that single species MIC measures are taken under distinct growth conditions.

However, for multiple examples the ability to resist antibiotics (in a standard clinical assay (European Committee for Antimicrobial Susceptibility Testing of the European Society of Clinical & Infectious, 2003; McKenney et al., 2012)) did not predict the presence / absence of the species after treatment. In red, **Figure 3.6** displays cases where the species was predicted to be present (given MIC resistance data, **Table 3.S2**), but was nevertheless absent in the final community. This pattern is suggestive of an additional role for microbe-microbe competitive

interactions in shaping community structure, and was observed for 6 of the 10 taxa, and most often in the absence of antibiotics. Conversely, blue regions in **Figure 3.6** identify cases where the pathogen was predicted from MIC data to be unable to grow in the allocated antibiotic, and yet was present in the multi-species community experiment in at least 1 community. This pattern is indicative of antibiotic-dependent faciliatory interactions, where other species aid the focal species to survive under antibiotic insult, for example via antibiotic detoxification (Brook, 2004; Brook et al., 1983; Dugatkin et al., 2005; Estrela & Brown, 2018; Hackman & Wilkins, 1975; Tacking, 1954).

To assess the role of antibiotic-dependent facilitation, we focus on the meropenem treatment (far right column, **Fig 3.6**), which indicates that the ability of *S. aureus* to grow in an otherwise lethal dose of meropenem is due to facilitation by *B. cenocepacia*. *B. cenocepacia* encodes multiple β-lactamase enzymes (Holden et al., 2009) that are potentially capable of degrading meropenem and therefore enable *S. aureus* to grow in this environment. To test the facilitation hypothesis, we culture *B. cenocepacia* and *S. aureus* alone and in co-culture (using rich media to allow monoculture comparisons), in both the presence and absence of meropenem **(Fig 3.7**). In monoculture we find that *S. aureus* growth is limited by meropenem (paired one-tailed t-test on *S. aureus* final density +/- meropenem, $p = 0.003$), consistent with MIC data (Table 3.S2). In contrast in co-culture we find that *S. aureus* growth in meropenem is rescued by co-culture with *B. cenocepacia (*paired one-tailed t-test on *S. aureus* final density in meropenem, +/- *B. cenocepacia*, $p = 0.020$), consistent with antibiotic-dependent facilitation.

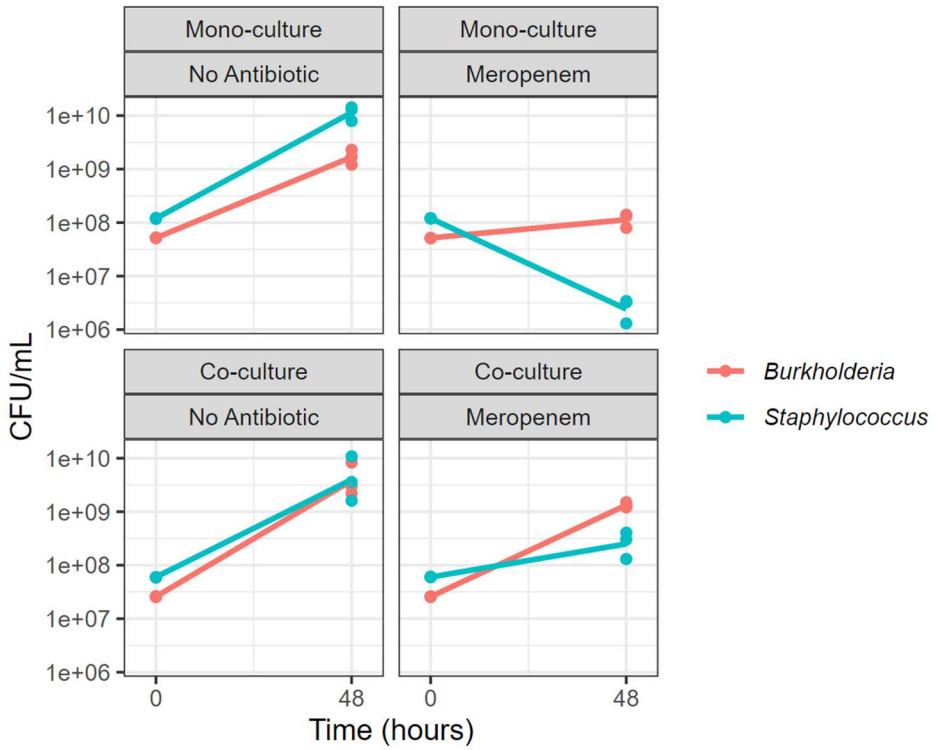

Figure 3.7. *S. aureus* growth in meropenem is facilitated by co-culture with *B. cenocepacia.* Experiments were conducted in rich media (TSYE broth) in room air, in the presence or absence of 10 µg/ml meropenem and for each species either grown alone (mono-culture) or together (co-culture) in a 96 well plate with hourly shaking. At zero and 48 hours, cells were serially diluted and plated at concentrations of 10^{-2} to 10^{-7} onto either Mannitol Salt Agar (for *S. aureus*) or LB Agar with 500 mg/L Gentamicin (for *B. cenocepacia*).

In light of the inability of antibiotic resistance data to reliable predict community structure at the species scale **(Fig 3.6)**, we next asked whether the resistance data is predictive at a broader, functional scale. Pooling drug resistant pathogens together (*S. aureus, B. cenocepacia*, and *A. xylosoxidans*) we find consistent enrichment (19- to 41-fold on average per treatment) across all drug exposures **(Fig 3.8)**, indicating a consistent enrichment of more problematic organisms following antibiotic exposure. In the discussion we explore potential contributing reasons other than community ecological interactions for the disconnect between MIC predictions on the species scale and observed community presences **(Fig 3.5)**.

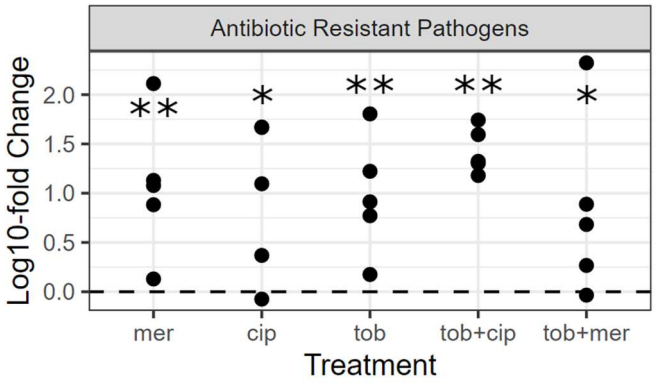

Figure 3.8. Drug-resistant pathogens are consistently enriched as a functional class, across all drug treatments. Fold-change differences for the sum of drug resistant pathogens (*B. cenocepacia, A. xylosoxidans,* and *S aureus*) compared to no antibiotic control, details as in **Figure 3.4**. Asterisks mark significant competitive release; one-tailed Wilcoxon test * $p < 0.05$, ** $p < 0.01$.

3.3.4 Community compositions across all antibiotic treatments are consistent with diversity across clinically observed *in vivo* communities

We finally ask, how do our *in vitro* synthetic microbiomes compare with the diversity of microbiome structures observed in people with CF? We begin with a PCA ordination plot to visualize experimental data (initial and final timepoints from **Figure 3.4**) alongside clinical data **(Fig 3.9**).

Figure 3.9A illustrates that our initial 10-species inocula (colored squares) are not representative of individual patient microbiome states (grey/black circles), reflecting their derivation from the typical meta-community state of populations with CF (**Fig 3.1**). **Figure 3.9B** highlights the repeatable end-point microbiomes in the absence of antibiotics (blue triangles), approaching commonly observed oral microbe dominated states (Zhao et al., 2020). **Figure 3.9C** illustrates the more divergent states resulting from antibiotic perturbations. Contrasting clinical versus pooled experimental data we see intermediate levels of community differentiation (ANOSIM $R = 0.28$), intermediate between the impacts of pathogen **(Fig 3.3**) and antibiotic **(Fig 3.4**) manipulations **(Fig 3.S3).**

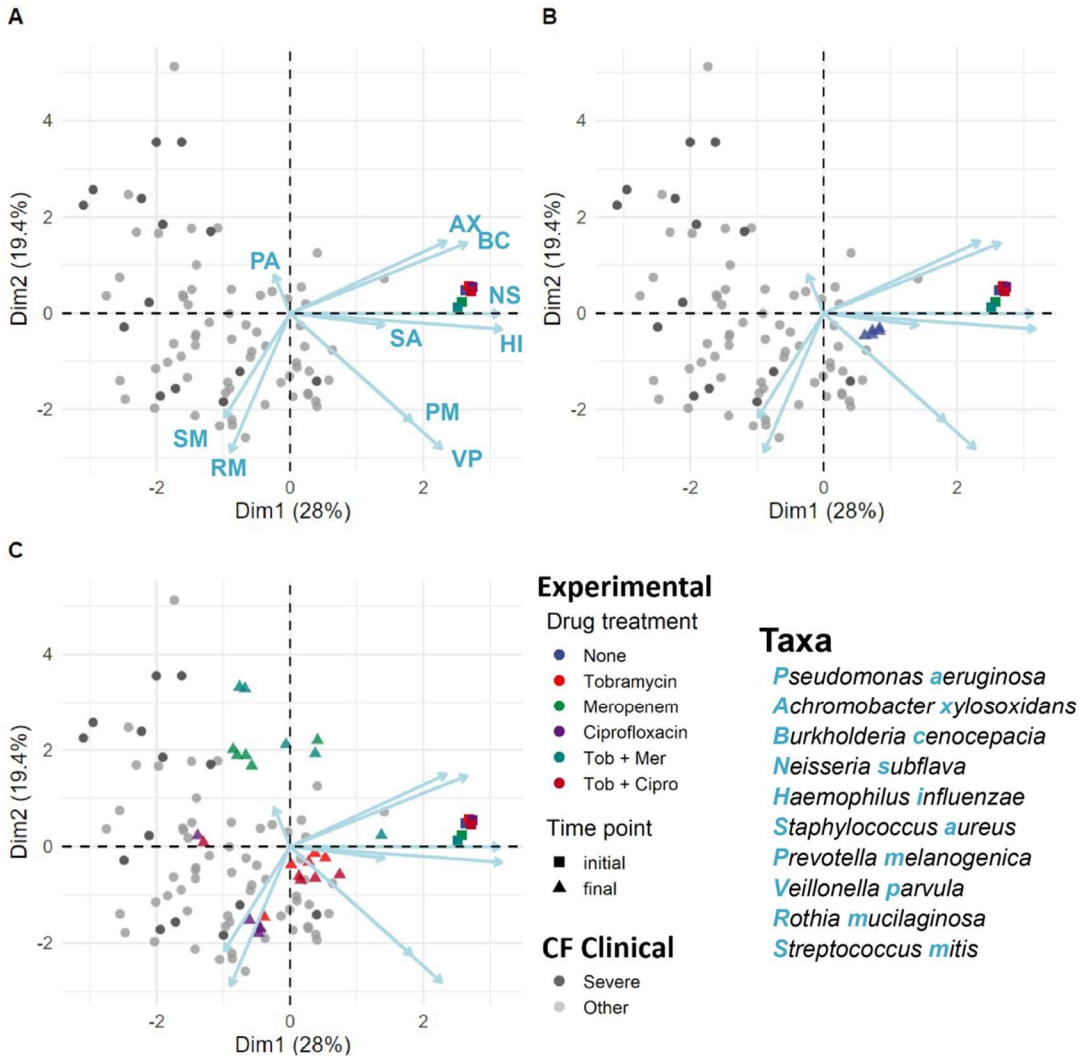

Figure 3.9. Antibiotics drive pathogen enrichment in experimental microbiomes, producing community structures that overlap with clinical sputum communities. PCA visualization of experimental microbiome data (colored triangles and squares, summarizing data in **Figure 3.4**) plus clinical microbiome data across a cohort of 77 people with CF (grey/black circles, black/severe signifies low lung function (Zhao et al., 2020)). (A) Squares illustrate experimental initial conditions, (B,C) triangles are final compositions after 5 serial passages (10 days), in the absence (B) or presence (C) of antibiotics. Colors denote experimental condition (see key). Each experimental treatment is replicated 5-fold, producing highly repeatable dynamics in the absence of antibiotics (blue triangles, B) and variable pathogen enriched outcomes following antibiotic treatment (C). Antibiotics were supplemented at each passage at clinically relevant concentrations (meropenem, 15 μg / ml; tobramycin, 5 μg / ml; ciprofloxacin, 2.5 μg / ml). Each point is a single microbiome sample (species resolution for clinical samples via the DADA2 plugin in QIIME 2 (Callahan et al., 2016; Zhao et al., 2020)). Ordination is PCA of centered log-ratio transformed relative abundances.

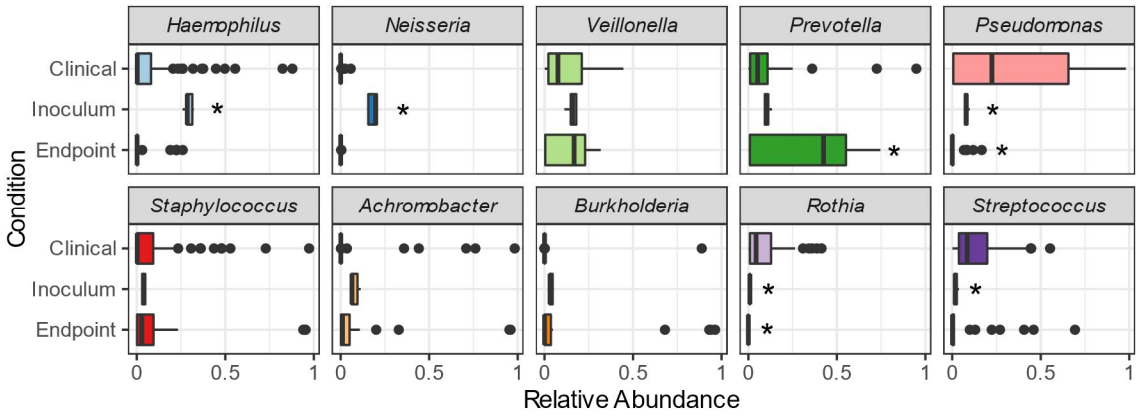

Figure 3.10. Most end-point experimental taxa fall within the range of clinically observed relative frequencies. The relative abundances of taxa in synthetic microbiome inocula and end points (30 samples) compared to 77 clinical cohort observations (Zhao et al., 2020). The box represents the interquartile range (from 25%-75% of samples) with the horizonal line at the median. Outliers are represented as dots (two-tailed Welch's t-test versus clinical data with Bonferroni multiple testing correction: * corrected $p < 0.001$).

Building on the overview provided by **Figure 3.9**, we now look at a more granular level and ask for each taxon whether species relative abundances in our experimental model fall within the range of clinical variation from our previous clinical study (Zhao et al., 2020). In **Figure 3.10** we first assess our metacommunity inoculum condition (time zero in Figures 2,4) against the yardstick of clinical variation, and unsurprisingly see a substantial number of taxon 'misses' (5 out 10 species abundance in the inocula is distinct from clinical data; Welch's t-test, $p < 0.001$), reflecting that our metacommunity initial conditions are not well matched to the typical profiles of individual sputum samples **(Figs 3.1, 3.9**). We next assess experimental community states after 5 serial passages (final timepoints across all treatments) and find a better match with clinical data. 3 of the 5 taxon 'misses' move within clinical variation, while one taxon (*P. melaninogenica*) moves outside of clinical variation, resulting in 7 out of 10 taxa where our experimental model produces ranges of relative abundances that do not significantly differ from benchmark clinical data **(Fig 3.10**). We find that our model significantly over-represents *Prevotella* and under-represents *Pseudomonas* and *Rothia.* These misses provide an opportunity to improve our model in future work, by pointing towards an environmental mismatch on oxygenation (with the strict

anaerobe *P. melaninogenica* benefitting and the facultative anaerobes *P. aeruginosa* and *R. mucilaginosa* suffering from the anaerobic atmosphere). This pattern of misses suggests that the distribution of oxygenation experienced clinically by CF microbiomes is more oxygenated than that provided by our anaerobic chambers with only brief exposures to oxygen every 48 h.

3.4 Discussion

Our results show that in the absence of antibiotic perturbations, our defined 10-species synthetic CF microbiome community follows a highly repeatable path to a stable community composition **(Fig 3.2**, **3.3, 3.9B**). In contrast, antibiotic treatments resulted in substantial community shifts **(Fig 3.4**), featuring both competitive release of previously rare pathogens **(Fig 3.5**) and emergent facilitatory interactions **(Fig 3.7**). Under antibiotic treatment we observed distinct trajectories across both drugs and replicates **(Figs 3.S2, 3.S4**), dispersed through a broad range of observed CF community structures, including alternate pathogen-dominant states **(Fig 3.9**). **Table 3.S3** summarizes our results in light of motivating hypotheses.

Our results highlight that standard antibiotic resistance MIC data (**Table 3.S2**) often fails to predict individual species presence following antibiotic exposure **(Fig 3.6**), with predictions showing both false positives and false negatives. A simple general explanation for departures from the ecological filter hypothesis is the presence of significant ecological interactions among species. Under this framework, false positives are evidence for suppressive interactions, suggesting that for example *S. aureus* fails to grow in the antibiotic free environment because it is out-competed by one or more of the other taxa. Conversely, false negatives are evidence of facilitation, suggesting for instance that the ability of *S. aureus* to grow in an otherwise lethal dose of meropenem is due to facilitation from another species in the community. In agreement with this hypothesis we find that *B. cenocepacia* shows meropenem-dependent facilitation of *S. aureus* growth, of sufficient magnitude to rescue *S. aureus* growth in super-inhibitory concentrations of antibiotic **(Fig 3.7**).

While the combination of antibiotic resistance and species interactions is a candidate explanation for our results **(Fig 3.7**), other factors are potentially at play. First, we again note the important caveat that the MIC estimates were derived using standard growth-promoting rich culture assays, which are known to generate estimates that tend to under-estimate the resistance of cells under more physiologically relevant conditions (Brown et al., 1990; Gilbert et al., 1990). If our MICs are under-estimates of resistance, then we would anticipate more 'false positive' evidence of competition in our experimental community. A second possibility for divergent results is the presence of physiological or evolutionary adaptation to the community conditions, across the 10 days of serial passaging. The stability and repeatability across replicates in **Figure 3.2** argue against a major role for genetic evolution in steering community dynamics – consistent with recent work on the suppressive impact of community interactions on bacterial evolution (Baumgartner et al., 2020).

In order to develop an experimentally tractable model, we made a number of choices regarding specific experimental conditions (e.g. nutrients, initial community structure, strain identity) that likely influenced our specific results. The healthy lung is evidently an oxygen rich environment, however during the course of tissue degradation in the CF airways, the sputum environment can become oxygen deprived due to the combined forces of mucus plugs, along with oxygen consumption by immune cells and microbes (Cowley et al., 2015; Kolpen et al., 2010; Wu et al., 2018). To capture an oxygen stressed environment, we performed our experiments under static anaerobic conditions that were only subjected to oxygenation during bench passaging every 48 h. While all bacteria in the community are capable of either fermentation, anaerobic respiration, or both, the largely anaerobic condition represents a potential to bias the results towards strictly anaerobic bacteria. Our clinical benchmarking exercise indicates that the distribution of oxygen exposures in the clinic is less biased towards anaerobic conditions, as our three endpoint taxon 'misses' **(Fig 3.10**) consisted of over-representation of an anaerobe (*P. melaninogenica*) and under-representation of two aerobes (*P. aeruginosa*, *R. mucilaginosa*). This

pattern is also consistent with recent transcriptomic analyses of *P. aeruginosa* from CF sputum, highlighting a transcriptional response indicative of reduced oxygen, but not necessarily anaerobic conditions (Cornforth et al., 2020). In future work we will investigate synthetic community dynamics in static communities with partial exposure to room air, following recent experimental *ex vivo (*patient sputum) models (Flynn et al., 2020; Quinn et al., 2018).

Turning to our choices regarding synthetic community composition, by focusing on the most abundant bacterial taxa, we ignored the potential for rare keystone species to shape community dynamics (Banerjee et al., 2018). We also overlooked the potential importance of interactions among strains within each species (Allen et al., 2016; Pollitt et al., 2014). Concerning specific strain choices, **Figure 3.3** illustrates that replacing mucoid *P. aeruginosa* PDO300 with an otherwise isogenic non-mucoid strain (PAO1) produces little dynamical change. However other studies in different environmental contexts have demonstrated substantial dependency of interactions on strain identity (Limoli et al., 2016; Maliniak et al., 2016), leaving open the importance of specific strain identities in governing community outcomes. More broadly, we did not include other potentially critical players in the lung microbiome, spanning human epithelial and immune cells, fungal species, and viruses of all the above. We note that our experimental platform is amenable to the addition of these players and additional manipulation of timing and order of introductions in future controlled experiments.

Our results demonstrate the power of a model 10-species system for the study of chronic lung infection dynamics. This model provides a platform to assess the community ecological impacts of currently deployed antibiotic treatments **(Figs 3.4-3.8**) and novel treatments – from different compounds to different strategies of their implementation. Current practice is to 'hit hard' with an antibiotic that is effective against a target pathogen (Read et al., 2011). In the context of our model community, detecting drug susceptible *P. aeruginosa* would typically trigger combination treatments that lead in our example to rapid emergence of more dominant and more resistant pathogen replacements **(Figs 3.4, 3.8**) (Halpin et al., 2016). One avenue to

improve on this picture is to run community-scale resistance diagnostics, and in turn use this diagnostic information to optimize antibiotic (and probiotic) choices (McAdams et al., 2019). While simple in outline, identifying optimal treatment choices in the context of complex multi-species communities poses a substantial computational and experimental challenge.

3.5 Materials and Methods

3.5.1 Bacterial strains

Table 3.1 outlines the specific strains in our 10-species community. Species choices were initially informed based on our previous study of a 77-person CF cohort with samples taken during periods of clinical stability (Zhao et al., 2020). Our 10 species represent the most abundant genera from our 16S rDNA analyses (together accounting for over 85% of reads). Note that these species are collectively representative of the 'metacommunity' (the community of communites (Leibold et al., 2004)) of microbes across a population of people with CF, and are not necessarily representative of individual community states. We view this metacommunity as the menu of organisms from which individual communities are sampled.

Table 3.1. Experimental model organisms used in synthetic community experiments. * indicates pulmonary source, [x] indicates oral source. Red font indicates established CF pathogen (CFF, 2019). [#] Isolate from Children's Hospital of Atlanta. Collectively, these organisms represent over 85% of clinical sequence reads across a 77-person CF lung microbiome study (Zhao et al., 2020).

Species	*Experimental Strain*	Relative abundance of the genus in clinical samples (%)
Pseudomonas aeruginosa	PDO300 (mucoid) PAO1 (wildtype)	29.7
Veillonella parvula	Clinical[#]	9.8
Rothia mucilaginosa	ATCC49042*	9.1
Prevotella melaninogenica	ATCC25845*	8.4
Streptococcus mitis	ATCC49456[x]	7.9
Haemophilus influenza	ATCC10211	5.8
Staphylococcus aureus	SAJE2	5.6
Achromobacter xylosoxidans	ATCC27061	4.8
Neisseria subflava	ATCC49275[x]	1.8
Burkholderia cenocepacia	K56-2	1.1

To guide our experimental species choices, we turned to existing CF metagenome sequencing data (Moran Losada et al., 2016), which provided high confidence for all but one of our species calls (Table 3.1). The exception is *Streptococcus*, where reads are distributed across a range of species. We chose *S. mitis* because it is present in sputum metagenomic profiles (Moran Losada et al., 2016), and it is an experimentally tractable organism that is typically considered to be non-pathogenic (Mitchell, 2011).

Within each species, we focused on well-characterized reference strains, as far as these were available, including American Type Culture Collection (ATCC) strains. For the dominant pathogen *P. aeruginosa (*PA), we used both the reference strain PAO1 and its mucoid derivative PDO300 (Mathee et al., 1999). Our default experimental choice is PDO300, as this strain better reflects the mucoid phenotype prevalent in chronic CF (Martin et al., 1993; Mathee et al., 1999).

3.5.2 Community growth medium

Our Artificial Sputum Medium (ASM) is based on the benchmarked synthetic CF sputum medium 2 (SCFM2 (Palmer et al., 2005; Turner et al., 2015)), but with differences in the preparation of the mucin and DNA macro-molecules. Specifically, mucins were ethanol washed and autoclaved (not UV sterilized, due to larger volume requirements), and the entire medium was filter sterilized following addition of DNA. Given the potential for differences in preparation methods to impact the results, we refer to our medium under the more generic name of ASM to underline these differences from the reference recipe for SCFM2 (Palmer et al., 2005; Turner et al., 2015).

3.5.3 Bacterial pre-culture and community construction

Before the experiment, all bacterial strains were revived from frozen stocks by streaking on rich media agar plates (chocolate or BHI agar, depending on the species, see **Table 3.S4**) and cultured at 37°C for 48 hours (microaerophilically (for *H. influenzae* and *N. subflava*) or

anaerobically (for *P. melaninogenica* and *V. parvula*, in GasPak jars). Five colonies were then picked from each plate and used to inoculate specific monoculture rich medium, which was cultured for a further 48 h; specific culture conditions are detailed in Table 3.S4.

The bacterial cultures were then washed in a defined ASM buffer base (ASM minus all carbon sources), OD_{600} values were measured with a Hidex plate reader (Hidex Oy, Finland) and adjusted to 0.5 for each species and diluted 10-fold in ASM. These standardized bacterial dilutions of equal volume were mixed and antibiotic stocks were added according to the experimental design for each treatment. The bacterial mixtures (plus antibiotics, dependent on treatment) were homogenized with pipetting, then divided into five replicates of 2 ml each in 24-well plates. An additional 0.5 ml of the initial inoculum mixture was stored at -80°C to assess community composition at time zero by subsequent genomic analysis.

3.5.4 Treatments and passaging

To measure the impact of exposure to antibiotics, we tested three antibiotics that are widely used in CF therapy: tobramycin (5 μg / ml); meropenem (15 μg / ml); ciprofloxacin (2.5 μg / ml), and two widely used combinations; tobramycin and meropenem; tobramycin and ciprofloxacin (adding the concentrations above). The specific concentrations used reflect measurements of antibiotic concentrations in CF sputum (Cipolla et al., 2016; Kuti et al., 2004; Moriarty et al., 2007; Ruddy et al., 2013; Wenzler et al., 2015). Our choice of 5 μg / ml tobramycin is low when compared to peak concentrations measured immediately following inhaled therapy (Lam et al., 2013; Somayaji & Parkins, 2015). Even in this immediate post-treatment context, the concentrations we used are within the range of their reported measured concentrations at 30 minutes point treatment (Somayaji & Parkins, 2015). Concentrations are not reported for any longer duration in these studies. All experiments were performed with 5 replicates of 2 mL cultures in 24-well plates cultured at 37°C in anaerobic GasPak jars. Every 48 h, bacterial cultures were mixed by pipetting, and 10% of the volume was transferred to fresh

ASM (with fresh antibiotics as defined by the treatment). 0.5 ml of the culture was stored at each passage at -80°C for later DNA purification and amplicon sequencing. Each experimental line was maintained for 5 passages (10 days).

To assess the role of pathogen characteristics, we conducted 5 pathogen manipulations (presence/absence of *P. aeruginosa* mucoidy (PDO300 versus PAO1) x presence/absence of *S. aureus*, plus a no *P. aeruginosa* + no *S. aureus* treatment). These experiments were done in the absence of antibiotics, but otherwise with the same conditions as above.

3.5.5 16S rDNA sequencing and qPCR.

DNA purification, sequencing, and qPCR were performed by MR DNA Lab (Shallowater, TX). Briefly: DNA was purified from sputum homogenate after mechanical lysis with the MoBio Power Soil kit (MoBio, Carlsbad, CA). The 16S V4 region of the resulting DNA was amplified with 515F and 806R primers incorporating the barcode in the forward primer and subjected to Illumina sequencing (Caporaso et al., 2011). The sequence data were generated in a total of 6 MiSeq runs. Total 16S abundance in each sample was determined by qPCR using standard 515F/806R primers (Caporaso et al., 2011).

3.5.6 16S rDNA sequence analysis

To generate taxa counts from the sequence data, we processed each run independently and combined the results. Across the 6 sequencing runs, a total of 15,347,658 sequence reads were generated, with a median of 59,686 sequences per sample (minimum 22,707, maximum 126,680). All sequence processing was done through QIIME2 2019.10.0. Unless otherwise noted, we left parameters as defaults based on the Moving Pictures workflow. Samples were demultiplexed using the cutadapt plugin in QIIME2. We found that some of the barcode sequences were also found in the 16S region of several taxa. To mitigate this confounder, we removed from each metadata file the first four nucleotides in the 515F primer and added it to the

barcode. For example, the barcode "GAGATGTG" was remapped as "GAGATGTGGTGC" and the primer became "CAGCMG…".

Reads were denoised using the deblur plugin, and resulting sequences were trimmed to 250 bp. Taxonomic assignments were classified against the greengenes 16S database. Some assignments were not possible at a level of genus resolution, so we interpreted reads mapping to "o__Lactobacillales" to "g__Streptococcus", "f__Burkholderiaceae" as "g__Burkholderia", and "f__Pseudomonadaceae" as "g_Pseudomonas". Finally, for each sample we removed spurious (and rare) taxon calls that did not map onto our experimentally defined communities. Sequence data have been deposited to the SRA (Accession # deposit pending). The analysis pipeline is available on GitHub (github.com/GaTechBrownLab/Varga-et-al_CompetitiveAbxRelease_SRA-upload).

Absolute abundances were determined by proportion of the total 16S count and then normalized to species-specific 16S rDNA copy counts (Stoddard et al., 2015; Zhao et al., 2020).

3.5.7 Statistical analyses

All analyses and plots used the R programming language (R Core Team, 2018; Wickham, 2016). Tables and scripts can be found at (https://github.com/GaTechBrownLab/Varga-et-al_CompetitiveAbxRelease_SRA-upload). A nonparametric Wilcoxon rank sum test was used to test for differences in absolute species abundances across experimental conditions, using a two-tailed test for change in abundance and a one-tailed test to assess competitive release (testing for increases only). A t-test with a Bonferroni multiple testing correction was performed to compare relative abundances of the 10 species under experimental conditions with clinical samples from the 77-patient cohort. To compare experimental treatments (and clinical benchmark data) at a community scale, we calculated ANOSIM R values on Bray-Curtis dissimilarity matrices for each treatment using the vegan package (Bray & Curtis, 1957; Clarke, 1993; Oksanen et al., 2019). The R statistic is a ratio of within-treatment differences to between-treatment differences

on a scale of -1 to 1, where a value of 1 would mean that all dissimilarity is between treatments, indicating completely different communities.

To visualize community-scale differences we constructed ordination plots for combined clinical and experimental compositional data. Clinical and experimental observations were center-log-transformed first (Kassambara & Mundt, 2020; Van den Boogaart et al., 2020), then standardized before principal component analysis (Kassambara & Mundt, 2020; R Core Team, 2018).

3.6 Acknowledgments

We thank the Brown lab, the Center for Microbial Dynamics and Infection, and Drs. Sylvie Estrela, Ellinor Alseth, and John LiPuma for valuable discussion and feedback. We thank Drs. Joanna Goldberg and Bob Jerris for providing bacterial strains.

This work was supported in part by CDC contracts BAA 2016-N-17812, BAA 2017-OADS-01, and 75D30120C-09782 to SPB, CFF awards BROWN19I0 and BROWN21P0 to SPB, GURNEY20F0 to JRG and DAVIS21H0 to JDD and NIH awards 1R21AI143296 and 1R21AI156817 to SPB.

Chapter 4: Non-neutral Taxa Partition Into 13 Pulmotypes Across 1000 people with CF[3]

4.1 Summary

We manually curated a microbiome dataset of over 4000 sputum samples representing more than 1000 people with CF (pwCF), matching every sample with corresponding metadata from 36 publications and standardizing bioinformatic analyses with a single common pipeline. We apply the Sloan Neutral Community Model (SNCM) to each dataset and find a consistent set of neutral and non-neutral taxa. We find that common CF pathogens are generally identified as non-neutral across studies, even though neutrality varies by study. We hypothesize that taxa driven by non-neutral ecological processes can be grouped into meaningful classes of microbial communities, or pulmotypes, based on their co-occurrence patterns. Dirichlet Multinomial Mixture modeling on non-neutral taxa partitions CF lung microbiomes into 13 distinct pulmotypes. Overall, we find that these pulmotypes differ by composition and clinical associations. Transition patterns between pulmotypes from longitudinal data further reveal different transition frequencies between pulmotypes. Specifically, we find that the five *Pseudomonas*-dominated pulmotypes are differentially linked to pulmotypes dominated by the end-stage CF pathogens *Burkholderia* and *Achromobacter*. Our findings suggest that *Pseudomonas*-dominated samples are not necessarily equivalent in community trajectory over time, which carries important implications for infection management in cystic fibrosis.

[3] Coauthors: Elijah Mehlferber, Haojun Song, Jinyeong Eum, & Sam P. Brown

4.2 Introduction

Cystic fibrosis (CF) is a genetic disease affecting more than 30,000 people in the US and 70,000 globally. CF is characterized by chronic, polymicrobial infections of the respiratory tracts. Culture-independent sequencing methods of sputum samples from people with CF (pwCF) have identified thousands of taxa including common CF pathogens (most notoriously, *Pseudomonas aeruginosa* (Cystic Fibrosis Foundation, 2021)) in addition to taxa often associated with commensal oral and nasopharyngeal microbiomes (Huang & LiPuma, 2016; Lucas et al., 2018).

Numerous cross-sectional and longitudinal CF studies have drawn associations between lung microbiome composition and disease progression, finding that loss of taxonomic diversity is consistently associated with severe disease and antibiotic use (Coburn et al., 2015; Cuthbertson et al., 2020; Widder et al., 2022; J. Zhao et al., 2012a). Lower diversity is also associated with increasing pathogen load and decreased abundance of oral anaerobes (Cuthbertson et al., 2020). Further, high-density and long-term longitudinal sampling studies have found large between-individual variation in microbiome composition (Blainey et al., 2012; Stressmann et al., 2011, 2012). This variation is expected, given the clinical and treatment differences found across pwCF including variation in therapy (antibiotics, CF correctors), disease state (Cystic Fibrosis Foundation, 2021; Konstan et al., 2009), or lifestyle factors (Caverly et al., 2019; J. Zhao et al., 2012b). However, this makes inferring ecological rules governing microbial dynamics in the lung difficult, as broad signals are often washed out by this individual-specific variation.

One common approach to handle high variability across individuals is to focus analyses on ‘core’ taxa that exceed defined thresholds of prevalence and/or abundance (Cuthbertson et al., 2020; Van Der Gast et al., 2011). Here we evaluate a distinct approach, focusing instead on taxa that show evidence of non-neutral ecological dynamics. Hubbell’s neutral biodiversity theory poses that ecological communities can be well-explained by stochastic processes, where taxonomic distributions are governed by random dispersal, birth and death events (Hubbell, 2001). Sloan extended this theory to prokaryotes and showed that these stochastic neutral

processes can explain the patterns of abundance found in many (but not all) microbial communities (Sloan et al., 2006). We propose that the neutral model can be used to identify taxa that are more likely to be ecologically significant members of the CF microbiota, based on whether they are non-neutral or neutral respectively. This approach may additionally allow for a reduction in the between-individual variation, as the bacteria that are randomly acquired from the environment via neutral processes will be removed, and only those with non-neutral patterns will be evaluated.

Employing this neutral model, Venkataraman et al. found that while non-CF airway microbiota distributions were consistent with neutral biodiversity, CF lung communities poorly fit neutral models (Venkataraman et al., 2015). They further hypothesized that departures from neutral distributions are associated with pulmonary disease states, and that the extent of departure correlates with disease severity. However, samples from only nine pwCF were considered for this analysis, and this sample-limitation is common across CF microbiome studies.

While individual studies generally have low sample size, the field of CF respiratory microbiomics is highly active and has produced well over 100 lung microbiome studies, primarily focused on 16S amplicon sequencing of expectorated sputum samples. These include numerous association studies tracking community composition as well as various health metrics, interventions, and disease states. However, studies vary in methodological details, such as study inclusion criteria, 16S region and primers used, and pipelines used for downstream bioinformatic analysis. This poses further challenges to extracting broadly applicable conclusions from each individual study.

In this study, we leverage this community effort by building a single standardized and publicly available dataset of 16S CF lung microbiome studies published on the NCBI Short Read Archive. For each study, we identify their corresponding publications, match all samples with existing metadata, and process all reads using a single common bioinformatic pipeline. To examine the value of our expanded dataset, we revisit central questions in the CF field: what are

the key functional taxa in the CF lung microbiome (Layeghifard et al., 2019; Quinn, Whiteson, et al., 2016)? Are there distinct microbiome profiles (or pulmotypes) among pwCF, and do these pulmotypes correspond to different clinical trajectories? (Hampton, Thomas, van der Gast, O'Toole, & Stanton, 2021; Widder et al., 2022) pulmotypes across pwCF. We find consistent signatures of non-neutrality among all leading CF pathogens together with other non-pathogenic organisms, indicating they are governed by deterministic (non-neutral) forces. Using these taxa we partition our 1000+ pwCF into 13 pulmotypes that collectively identify leading pathogen-dominant states and are separable by compositional, clinical, and transitional differences, indicating that there are distinct community composition types across CF patients.

4.3 Results

4.3.1 A Standardized CF Microbiome Database

We curated a database of 4171 sputum compositions across 1175 individuals with CF from 36 published studies (median 53 samples per study; 1-163 samples per individual pwCF) representing 26 CF centers across 14 countries (**Fig 4.1A**). In light of variation in inclusion criteria and methodological choices, we hypothesized that methodological and regional differences will result in compositional differences among studies. Consistent with this hypothesis, we see significant differences in composition in multiple pairwise study comparisons (**Fig. 4.1C**, Analysis of Similarity ANOSIM R $p < 0.05$ Bonferroni corrected), yet effect sizes rarely met our threshold for a substantial effect (ANOSIM R > 0.4). In one-versus-all comparisons (**Fig 4.1B,D**) we did not identify samples from any individual study or center with substantial (ANOSIM R > 0.4) and significantly different ($p < 0.05$ Bonferroni corrected) composition from the rest (**Fig 4.1B-D**). The most compositionally distinct study (ANOSIM R = 0.391, $p < 0.05$ Bonferroni corrected) was reported from N.F. Gamaleya in Russia (PRJNA544655 (Voronina et al., 2020)). While we did not identify major methodological differences, the study is characterized by a strikingly high representation of *Burkholderia* in

Russian pwCF. Of the fifteen samples that passed our quality filter, six contained over 75% *Burkholderia* (33.3% samples *Burkholderia* dominant). For comparison, of the 77 samples from a survey across individuals attending Emory and Georgia Tech affiliated clinics analysed in Chapter 2 (PRJNA666192, (C. Y. Zhao et al., 2021)), only one sample contained over 75% *Burkholderia* (1.3% of samples *Burkholderia* dominant).

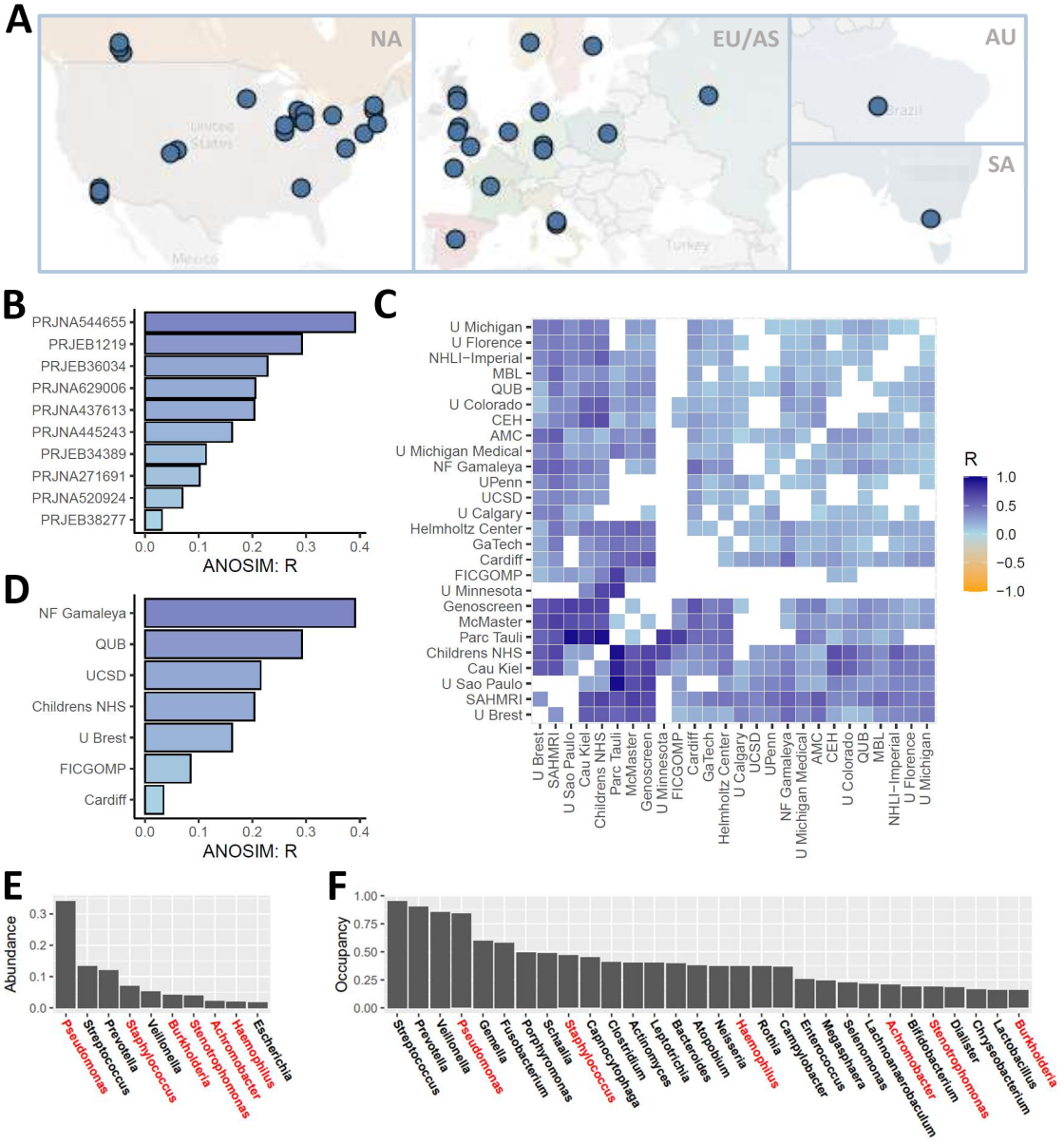

Figure 4.1 Study Characteristics. (**A**) We analyzed 36 studies across 26 CF centers from 14 countries worldwide. Three studies (PRJEB30646, PRJEB38277, and PRJNA207555) are split across more than one study site. (**B**) One-vs-all ANOSIM of CF sputum sequence repositories deposited to NCBI-SRA (BioProject numbers provided) do not identify any studies with

substantial (ANOSIM R > 0.4) and significantly different ($p < 0.05$, Bonferroni corrected) composition. One study, PRJNA544655, returned an ANOSIM R of 0.392 (**C)** Pairwise ANOSIM for studies grouped by CF center that uploaded the study. White spaces denote pairs that were not significantly different ($p > 0.05$, Bonferroni corrected). **(D)** One-vs-all ANOSIM, for studies grouped by CF center. (**E**) Mean relative abundance by genera. (**F**) Mean prevalence by genera. Canonical CF pathogens (*Pseudomonas, Staphylococcus, Burkholderia, Stenotrophomonas, Achromobacter,* and *Haemophilus*) are highlighted in red.

Post-rarefaction, 95% of reads came from 19 taxa. *Pseudomonas* was the most abundant taxa detected, followed by *Streptococcus*, *Prevotella*, and *Staphylococcus* (**Fig 4.1C**). *Streptococcus* was the most prevalent taxa, occurring in 95.3% of all samples, followed by *Prevotella*, *Veillonella*, and *Pseudomonas* (**Fig 4.1D**).

4.3.2 Canonical CF Pathogens are Consistently Non-Neutral

Neutral models of taxa distribution have been found to be consistent with observed community compositions in various microbiome contexts (Burns et al., 2016; Shade & Stopnisek, 2019; Venkataraman et al., 2015). To assess neutrality in CF, we apply the Sloan Neutral Community Model (SNCM) across individual studies in our dataset and recover varying degrees of fit (**Fig 4.2A**). We use $R^2 = 0.7$ as our threshold for a good fit, which is consistent with the model interpretations from Venkataraman et al. While overall we observe poor fits to SNCMs, we identified seven studies with taxon distributions consistent with SNCM predictions (R^2 range: 0.72-0.93; BioProject IDs in order of increasing fit: PRJEB8060, PRJNA645089, PRJNA666192, PRJNA662963, PRJNA360332, PRJNA234009, and PRJEB31332).

Turning to individual taxa, across all studies we find 46 neutrally distributed taxa comprising 20.2% of all reads (grey points in **Fig 4.2A**). The most abundant neutral genera were *Streptococcus*, *Enterococcus*, *Gemella*, and *Escherichia.* In contrast, 28 taxa deviated from neutral model predictions in the majority of studies that they were detected in (blue/red points, **Fig 4.2A**), including all 6 canonical CF pathogens (**Fig 4.2B**).

Additionally, all models were fit with a low immigration probability ($m < 0.1$, **Fig 4.2C**). The immigration parameter *m* represents the probability that a new cell is sourced via immigration, versus local reproduction (Sloan et al., 2006). Our results therefore imply that local reproduction consistently and substantially dominates immigration across all studies, but we note that ecological interpretations of our model fits are sensitive to the adequacy of our model assumptions (see discussion).

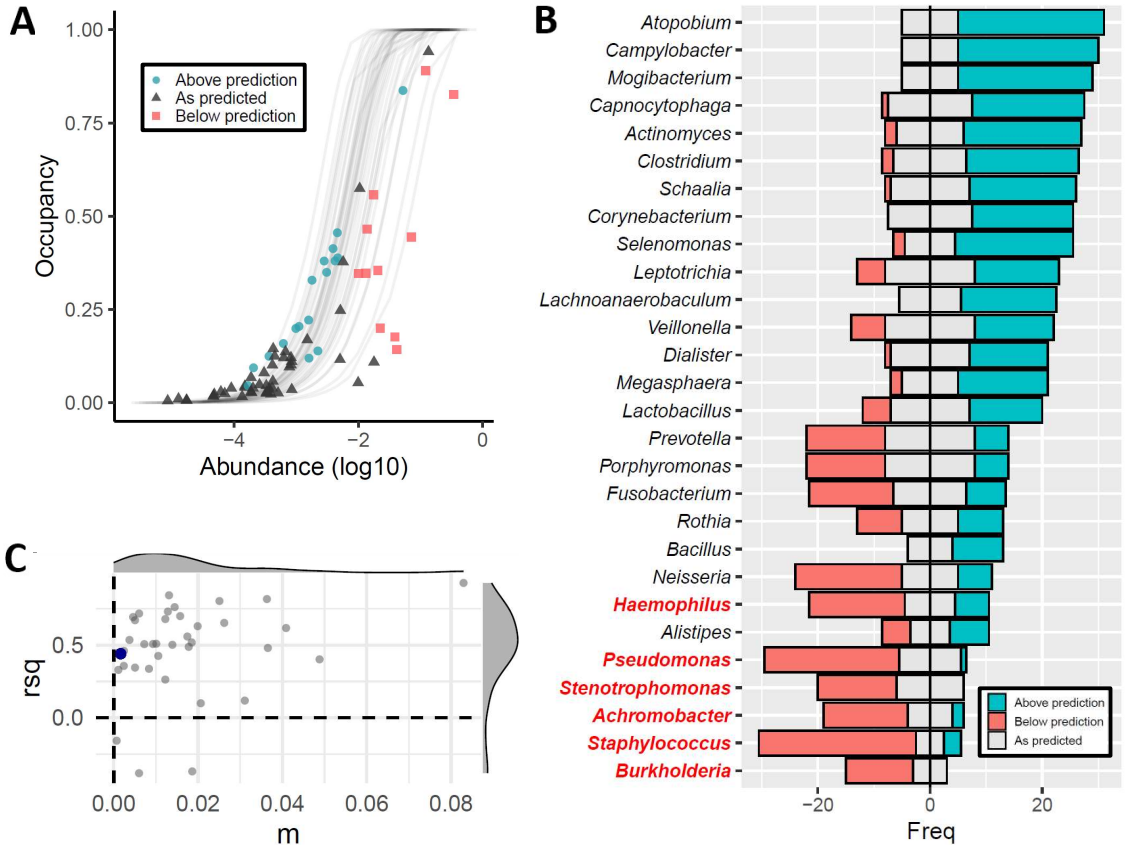

Figure 4.2. Neutral models identify non-randomly distributed taxa across studies. (**A**) Sloan neutral models were fit to abundance-occupancy curves (AOCs) across all 36 studies. Grey lines represent best model fits for each individual study. Only the first available sputum sample for each pwCF was used to generate AOCs. Taxon abundance and occupancy data for the composite dataset (all 1088 samples) are shown. AOC data points for individual datasets are not shown. Colored points represent taxa identified as non-neutral. (**B**) Across all studies, 28 taxa did not fit neutral distributions as predicted by the Sloan model. Taxa that did fit were excluded from further analysis. All 6 canonical CF pathogens (red font) are identified as non-neutral and below prediction. The most common commensal genus *Streptococcus* was identified as neutral. (**C**)

Summary of model fits for all 36 studies. All models were fit with a low immigration probability ($m < 0.1$), consistent with a low dispersal rate.

Many studies partition CF microbiota into core and transient species, defined by prevalence and/or abundance measures. Core taxa are ones that are both highly prevalent and abundant, and therefore more likely to drive disease ecology whereas transient ones drive metacommunity diversity (Cuthbertson et al., 2016, 2020; Van Der Gast et al., 2011). Cuthbertson et al., defined core taxa as those in the top quartile by prevalence and found that core taxa are more highly represented in more severe disease, with a greater proportion of variability in microbial diversity attributed to satellite taxa rather than core.

Rather than applying an arbitrary cut off to assign important taxa, we utilize a model to identify taxa that show patterns of abundance and dispersal that indicate ecological relevance in the CF environment. Specifically, we assume that neutral taxa detected in sputum samples are neither interacting with the host lung environment (e.g. consuming nutrients or being suppressed by host immune activity within the lung), nor with other taxa (e.g. competition or facilitation) to an extent that is enough to overcome the effects of random dispersal and drift in the lung.

In this study, we address partitioning into core and transient taxa by using inferred neutrality as our partitioning criterion. We should note that we are not the first to apply neutral models as a filtering mechanism. Venkataraman et al., used inferred neutrality to filter putative sequencing contaminants by fitting SCNMs to known sequencing controls (Venkataraman et al., 2015). We extend this method and remove all reads from taxa that were identified as neutral taxa in the majority of the included studies. We further excluded samples with fewer than 1000 reads, yielding 3585 sputum samples across 1088 pwCF.

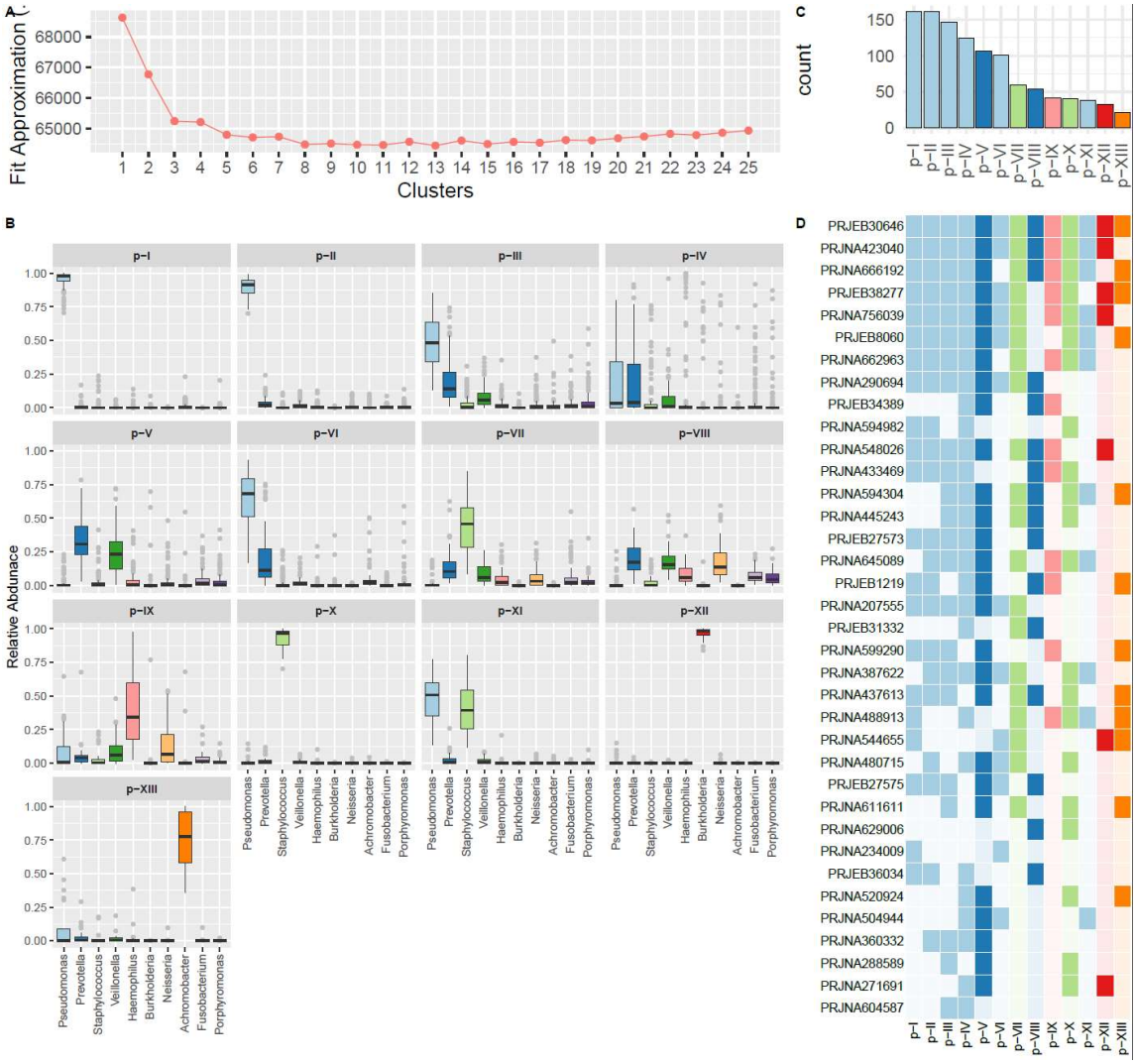

Figure 4.3. Cystic fibrosis sputum microbiomes separate into 13 pulmotypes. Clusters (k=1…25) were calculated using DMMs on individual snapshot data (N=1088) of non-neutral taxa, rarefied to 2000 sequences each. (**A**) Using the Laplace approximation of the negative log model evidence, we find a minimum at k=13 clusters (pulmotypes). (**B**) Boxplots of the top 10 taxa grouped by 13 pulmotypes. (**C**) Overall frequency of each pulmotype across the initial 1088 samples. (**D**) Pulmotypes represented in each study are shaded in. Studies are ordered by number of samples included.

4.3.3 Non-neutral CF microbiomes partition into thirteen pulmotypes

Using the distributions of only non-neutral taxa, we next apply a common clustering method to assess whether our microbiome data can be adequately represented by a limited number of distinct clusters or "pulmotypes" (Holmes et al., 2012; Widder et al., 2022).

Comparing the performance of one to 25 clusters, we find that non-neutral CF lung microbiomes are best partitioned into thirteen pulmotypes **(Fig 4.3)**, in contrast to prior studies identifying 5 (Hampton et al., 2021) or 8 (Widder et al., 2022) pulmotypes**.** To avoid over-representation by longitudinally tracked pwCF, pulmotypes were calculated from a single initial sputum sample per subject with neutral taxa excluded (N=1088 pwCF, **Fig 4.3**). The resulting distribution of samples (one sample per pwCF) across pulmotypes ranges from 21 to 161 samples (**Fig 4.3C**). The relative abundances of the top genera for each pulmotype are shown in **Figure 4.3B**. Using the highest median relative abundance taxa, we further group pulmotypes into three categories: *Pseudomonas*-dominant (PA), oral anaerobe dominant (OA), and other pathogen dominant (OP).

The majority (67.6%) of the 1088 training samples are categorized as one of six PA pulmotypes (I-IV, VI, and XI). The most common pulmotypes represented are p-I and p-II (N=161 samples each), which consist of samples with high *Pseudomonas* relative abundance. Type II is the most homogeneous (DMM homogeneity score, $\theta_{II} = 44.3$; higher θ is more homogeneous) and consists of samples with 70.2-99.0% *Pseudomonas*. Pulmotype IV is the least homogeneous ($\theta_{IV} = 1.52$) and contains samples with varying abundances of *Pseudomonas*. Pulmotype XI is the least frequent PA pulmotype and characterized by *Pseudomonas*-*Staphylococcus* co-domination, a condition that has received significant research and clinical attention (Limoli et al., 2016, 2017). The PA pulmotypes also vary in overall *Pseudomonas* relative abundance (from lowest to highest: IV, III, XI, VI, II, and I). This pattern is consistent with a continuous gradation of *Pseudomonas* across samples arbitrarily clustered into distinct pulmotypes. The two OA pulmotypes (V and VIII), comprise of 161 (14.8%) samples with varying *Prevotella*, *Veillonella*, and *Neisseria* relative abundances. Pathogen levels in the OA pulmotypes are generally low. OP pulmotypes (VII, IX-X and XII-XIII) consist of samples dominated by non-*Pseudomonas* pathogens, namely *Staphylococcus*, *Haemophilus*, *Burkholderia*, and *Achromobacter*. Notably, we do not find a single OP pulmotype comprising of *Stenotrophomonas*-dominated cases, despite loads of higher than 30% in eleven samples and over

90% in five. Instead, ten of these samples are grouped into the PA pulmotype IV and the remaining one into OP pulmotype IX.

For comparison, we applied the same clustering method without a neutral filter (**Fig 4.S1**) and identified 12 all-data pulmotypes. We find five *Pseudomonas*, three *Streptococcus*, one *Veillonella*, one *Burkholderia*, one *Staphylococcus*, and one *Haemophilus* pulmotype. We observe many similar patterns of pulmotypes (including a *Pseudomonas*-*Staph* codominant pulmotype). At face value, non-neutral pulmotyping yields a better model fit (**Fig 4.S1A** and **Fig 4.3A**) as well as a greater number of pathogen-dominated pulmotypes. In particular, we identify a *Achromobacter* dominant pulmotype p-XIII (**Fig 4.3B**), whereas without neutral filtering, *Achromobacter*-dominant samples are distributed throughout multiple pulmotypes (**Fig 4.S1B)**.

4.3.4 Pulmotypes are represented across studies

To examine the validity and distinctness of these pulmotypes, we first analyze their representation across included studies. A risk of our multi-study approach is that clustering will simply separate pwCF by study. In contrast, we find that every pulmotype is represented in at least seven studies, and one study contained all thirteen pulmotypes (**Fig 4.3D**). The top five studies by number of pulmotypes represented (in decreasing order by total number: PRJEB30646, PRJNA423040, PRJNA666192, PRJEB38277, and PRJNA756039) were larger cross-sectional studies (pwCF > 57), including two that were used in prior pulmotyping analyses (Hampton, Thomas, van der Gast, O'Toole, & Stanton, 2021; Widder et al., 2022). To further assess the validity of these pulmotypes, we turn to three additional lines of evidence: compositional differences, clinical similarity, and dynamical transition patterns.

4.3.5 Pulmotypes differ in composition but not generally in lung function

Next, we looked at the similarity between pulmotypes by assessing compositional similarity and lung function (**Fig 4.4**). We used ANOSIM to compare pulmotype compositional

similarities and found that most pairwise comparisons yielded significant differences in composition (**Fig 4.4A**, $p < 0.05$, Bonferroni corrected), largely exceeding an established threshold of $R > 0.4$ for a meaningful effect size.

Turning to differences in lung function between pulmotypes, we found in contrast that most pairwise comparisons did not yield significant differences in lung function (**Fig 4.4B**). In **Fig 4.4C** we offer a summary of both compositional and lung function differences, highlighting 12 pairs of pulmotypes that are both compositionally and clinically similar (yellow boxes) and 9 pairs that are substantially (large effect size) distinct on both measures (blue boxes). The remaining 57 pairs show mixed results, typically reflecting substantial differences in composition (**Fig 4.4A**) combined with insignificant and/or small differences in lung function (**Fig 4.4B**).

Of the twelve pairs of pulmotypes that are compositionally and clinically similar, nine include p-IV, a PA pulmotype with the lowest homogeneity score. Within the rest, we find three sets of two pulmotypes that are compositionally and clinically similar: I-II, III-VI, and V-VIII. Pulmotypes I, II, III, and VI are all PA pulmotypes with relatively high *Pseudomonas* abundance. Pulmotypes V and VIII are both OA pulmotypes.

4.3.6 Transition patterns in longitudinal data differentiate similar pulmotypes

Finally, we assess pulmotype robustness by looking at transition patterns across longitudinal samples. While we can infer ordinality (which sample came first, second, etc…) across all subjects with longitudinal samples, we do not have information on between-sample duration for every subject. Using the 13-component DMM trained on cross-sectional data alone, we classify the remaining 2497 longitudinal samples (representing transitions across 356 pwCF). Across all pairs of consecutive samples, 1456 (58.3%) are the same source and target pulmotype. Across all individual trajectories, we find that 95 pwCF (26.7%) do not transition to a new pulmotype during their surveillance period, and 138 pwCF (38.8%) do not change their pulmotype group (i.e. remaining with the broad classification of PA, OA or OP dominated

pulmotypes). Taken together, these indicate substantial longitudinal stability in our pulmotype classifications.

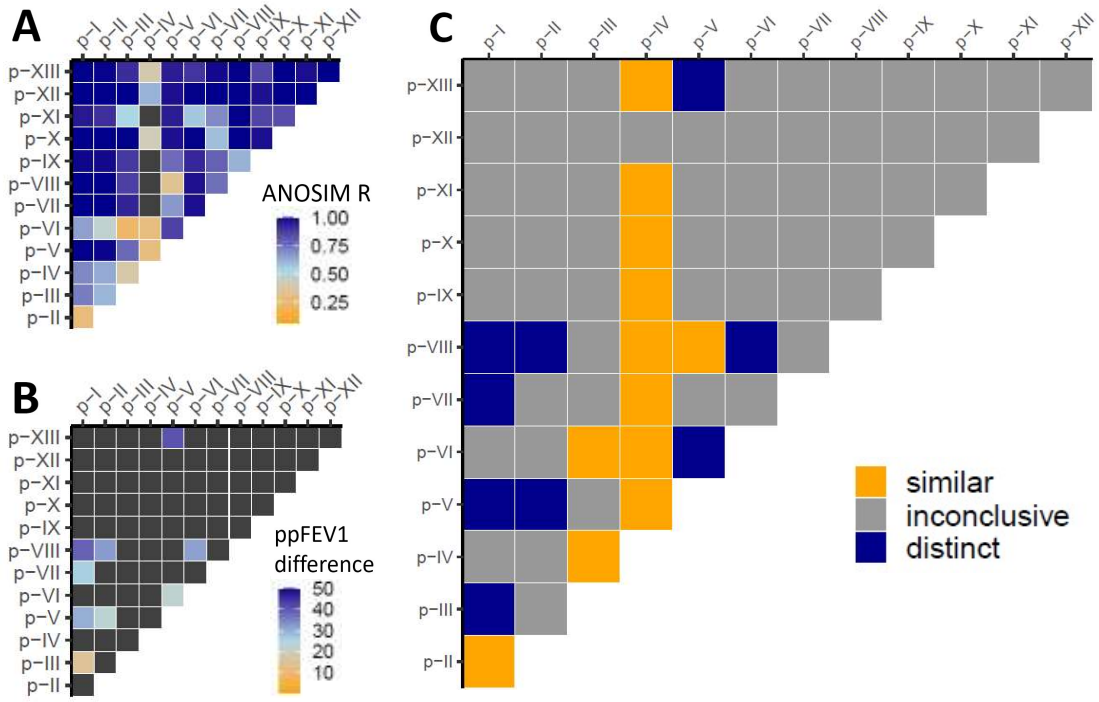

Figure 4.4. Compositional and clinical similarity between pulmotypes. (**A**) We calculated compositional differences between pairs of pulmotypes using ANOSIM. Colors represent magnitude of differences (ANOSIM R values). Dark grey boxes denote pairs that were not significantly different ($p > 0.05$, Bonferroni corrected). (**B**) We calculated clinical differences between pulmotypes by comparing lung function score (ppFEV1, data available for 542 out of 1088 pwCF). Significant differences were determined using the Wilcoxon rank-sum test ($p > 0.05$, Bonferroni corrected). (**C**) We then assessed agreement between clinical and compositional similarities. We define similar pulmotypes (yellow) as pairs that are both compositionally similar ($R < 0.4$ or $p > 0.05$) and clinically similar (ppFEV1 difference < 5% or $p > 0.05$). We define distinct pulmotypes (blue) as pairs that are both substantially compositionally distinct ($R > 0.4$ and $p < 0.05$) and clinically distinct (ppFEV1 difference > 5% and $p < 0.05$). All other pairs are shaded in grey.

We next analyze the characteristics of pulmotype transitions between consecutive samples (**Fig 4.5**). For each pair of pulmotypes we report the number of individuals in which we find that given transition across two consecutive samples (**Fig 4.5A**). We hypothesized that transitions in pulmotype state would result from either misclassification (samples drifting along

pulmotype boundaries) or progressive (directional) community shifts. If misclassification is a substantial issue, we would expect to see a negative association between assignment confidence (the likelihood of observing a given sample in its maximally likely component in the fit DMM model) and the likelihood of a shift in pulmotype in the subsequent sample. In contrast, we find that sample-pulmotype assignment confidence and distance between consecutive samples are unrelated (R = -0.04, **Fig 4.5B**), and that within-pulmotype and between-pulmotype transitions do not differ in sample-pulmotype assignment confidence (p=0.556, Wilcoxon W=18300, **Fig 4.5C**), suggesting that pulmotype transitions are not due to misclassification. We also assess whether our classification supports a proximity model, where transitions are more common among pulmotypes that are more structurally similar (and thus have a smaller between-pulmotype distance). Consistent with this model we find that pairs of pulmotypes with lower Bray-Curtis dissimilarity experience more frequent transitions (R = -0.50, **Fig 4.5D**).

We assess directionality by identifying pairs of pulmotypes with the highest between-pulmotype transition frequency. We visualize these observed transitions by generating a bidirectional network (**Fig 4.6**). We threshold based on pulmotype assignment confidence, only including transitions where the product of source and target confidences exceeded 80%. This removes 878 transitions (35.2% of total transitions). Applying this confidence threshold, we find that the majority of consecutive samples did not change pulmotypes (N=1130 samples, 69.8%), supporting substantial structural persistence on the pulmotype scale. To avoid individual over-representation, edges in the network are scaled by the number of pwCF for which a given transition was observed rather than the total number of instances each transition was observed across our dataset. Node sizes are scaled by frequency across the initial DMM training set (**Fig 4.3C**).

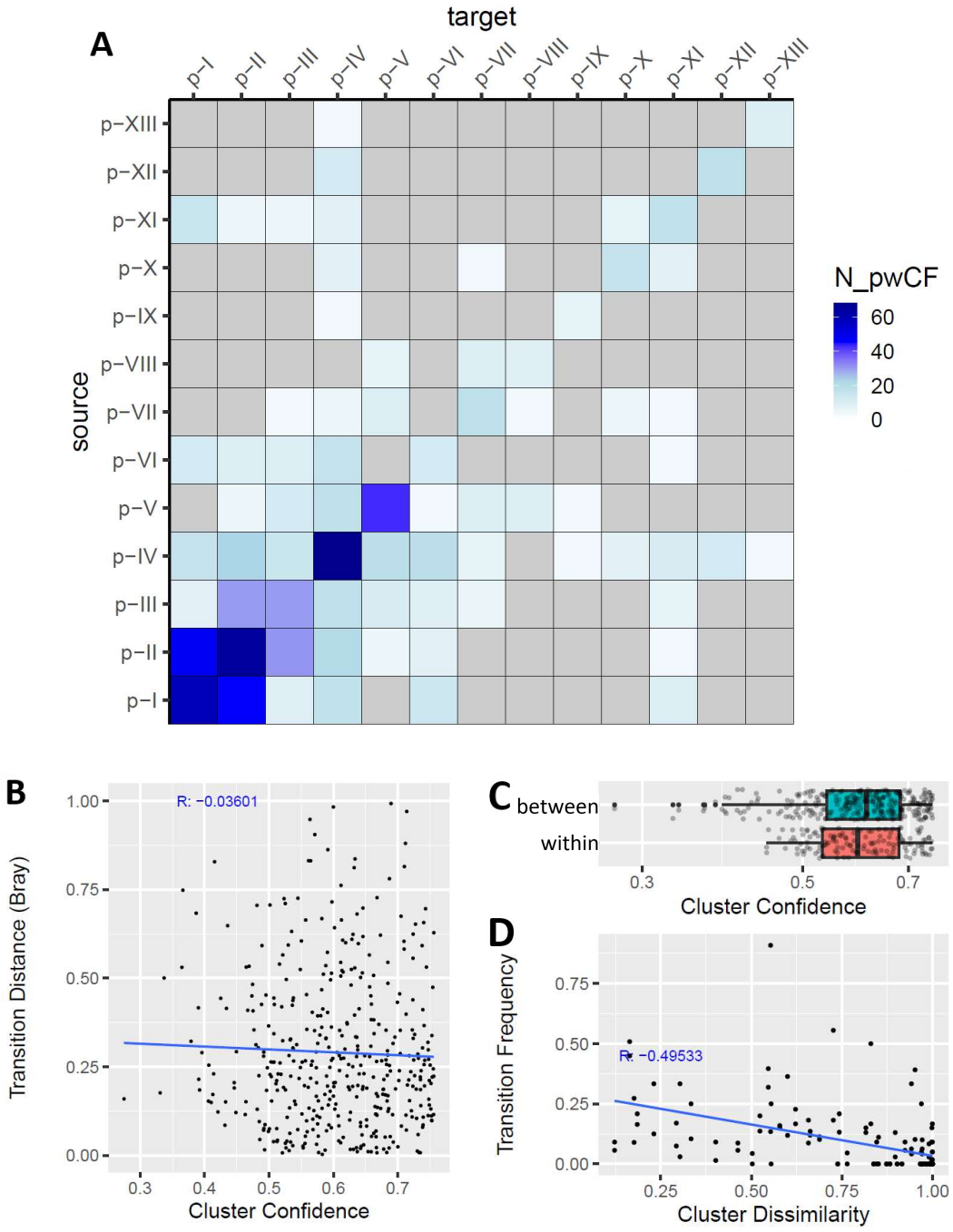

Figure 4.5. Transition frequencies between pulmotypes. We assess transition frequencies between pairs of consecutive samples for each pwCF. (**A**) Matrix of all transitions between pulmotypes, colored by the number of individuals a given transition was observed in. Transitions with fewer than 3 observations are represented by grey boxes. (**B**) Pulmotype assignment confidence (the likelihood of observing a given sample in its maximally likely component in the fit DMM model), does not correlate with distance between consecutive samples (Bray-Curtis dissimilarity) for all transitions. (**C**) Pulmotype assignment confidence is lower in between-pulmotype transitions than within-pulmotype transitions. (**D**) The transition frequency between two pulmotypes decreases with mean dissimilarity (ANOSIM R).

Across all pulmotype transitions, we find a signature of bi-directionality (balanced forward and reverse transition rates), indicating a general absence of progressive, directional change in pulmotypes through time (**Fig 4.5A**). A simple binomial fit to the forward and reverse transition rates for all pulmotypes show that the forward transition is not more likely than the reverse and vice versa. The most frequently observed between-pulmotype transitions were from II to I (N=31 pwCF) and from I to II (N=30 pwCF). The most common between-group transitions were from Pa to OA (N=87 pwCF), OA to Pa (N=85 pwCF), and from Pa to OP (N=54).

Under a progressive, directional model we would expect to see that OP pulmotypes are common end-states or sinks, as these pathogens (along with PA) are often associated with end-stage disease. However, we find 13 pwCF with transitions from OP to OA pulmotypes, suggesting that OP pulmotypes are not sinks. Alternatively, this may also be indicative of non-robust boundaries between pulmotypes, or a result of the substantial heterogeneity within the p-IV pulmotype (**Fig 4.3B**).

At face value the varying levels of *Pseudomonas* in each of the PA pulmotypes was consistent with a continuous gradient of increasing PA dominance. We find some transition patterns consistent with this interpretation. Specifically, we find reciprocal transitions between p-IV and p-III (**Fig 4.5A**). However, p-III and p-IV have distinct transition patterns (**Fig 4.5A, 4.6**). For example, p-III only transitions to two OP pulmotypes, p-VII and p-X whereas we observe p-IV transitions between all five OP pulmotypes. Overall, the PA pulmotypes exhibit balanced reciprocal transitions. Taken together, the differing transition patterns and lack of directionality further refute the continuous gradient hypothesis and suggest that these pulmotypes are biologically distinct and adjacent in both composition (**Fig 4.4**) and transition (**Fig 4.6**) space.

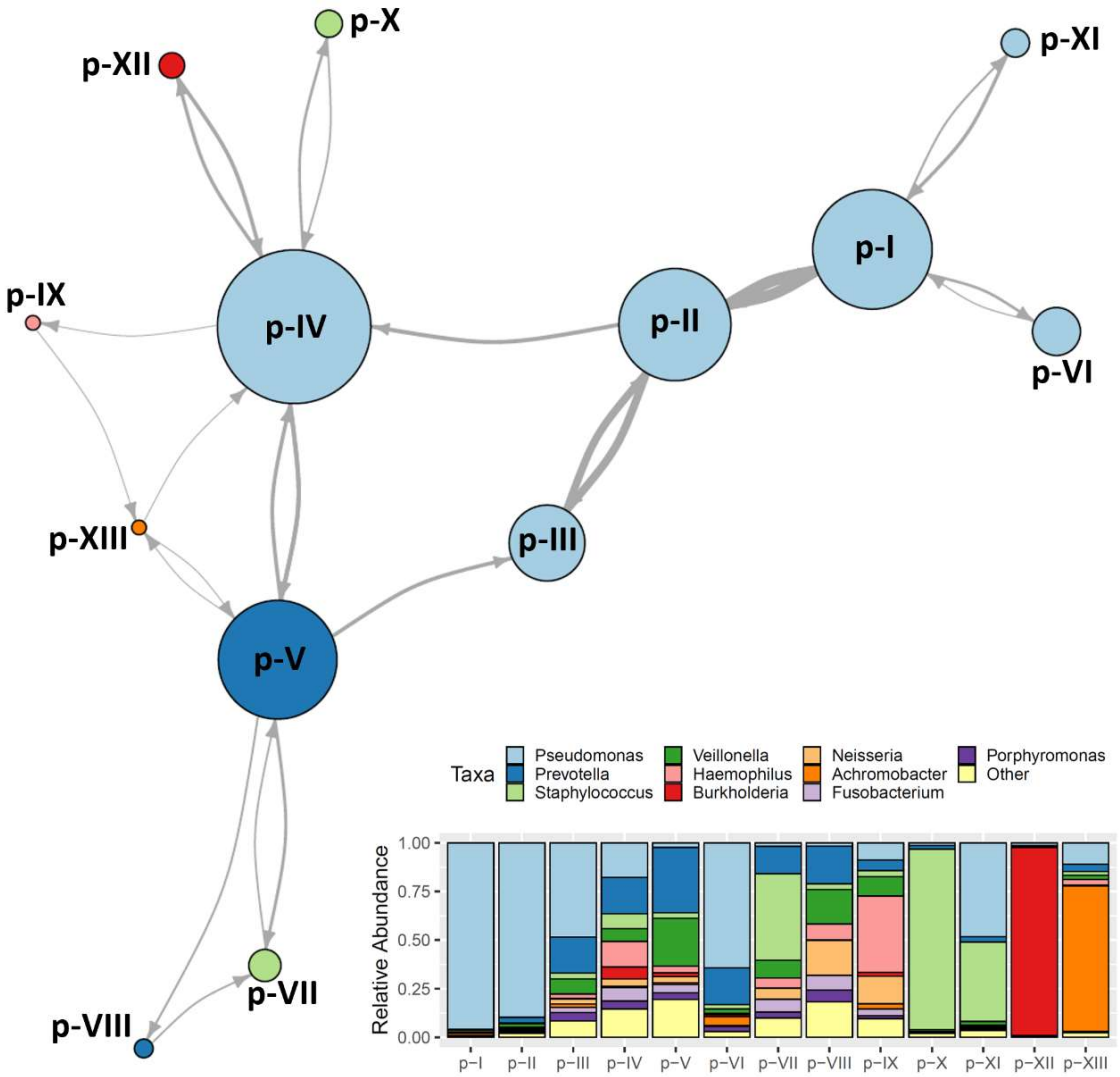

Figure 4.6. Transition network between pulmotypes. We calculate a directed network from transition frequencies with confidence scores (product of source and target pulmotype assignment confidence) greater than 80%. For each node, we only show the most common incoming source node and outgoing target node. Edges are scaled by the number of pwCF for which a given transition is observed. Node sizes are scaled by observed frequency in the overall dataset. (**Lower right**) Mean taxa distributions for each pulmotype.

4.4. Discussion

In this study, we combined publicly available sputum samples with published metadata to identify consistent ecological patterns across over 1000 pwCF. We produce two key findings: (1) patterns of neutrality across CF microbiota are broadly consistent across studies (**Fig 4.1**), indicating that the CF lung is potentially shaped by deterministic forces as it becomes dominated

by pathogens adapted to the lung environment, and (2) CF lung microbiomes partition into 13 pulmotypes (**Fig 4.3**) that are separable by compositional, clinical, and transitional differences, indicating that there are distinct community composition types across CF patients, with potential implications for disease management. The database assembled by this study will be made publicly available via GitHub and represents a significant resource to both the CF community and for scientists engaged in the study of human associated microbiomes in the context of disease.

The field of CF has generated a significant number of microbiome studies. These studies have identified complex microbial communities inhabiting CF airways with high between-person and even within-person variability. In light of the multitude of generally low sample size studies, composite studies are potentially beneficial in identifying broad-scale patterns across numerous cohorts. Li et al., previously performed a data mining analysis across 18 studies with 718 sputum samples and found that overall, antibiotic treatments for exacerbation management had a large impact on commensal taxa but little impact on CF pathogens (Li et al., 2016).

Across all studies in our analysis we find numerous taxa in addition to canonical CF pathogens. Some of these taxa show up at low abundance and prevalence, and drive variability between samples and individuals (Cuthbertson et al., 2020). There is debate as to the role of these additional organisms (Jorth et al., 2019; Lu et al., 2020). Many of these other bacteria are commonly isolated from oral cavities and upper airway sites and are not generally considered pathogenic (Filkins et al., 2012; Fodor et al., 2012; Frayman et al., 2017; Huang & LiPuma, 2016; Lucas et al., 2018). This leads some to postulate that these organisms are contaminants in the sampling process, as both bronchoalveolar lavage and sputum sampling necessarily involve traversal through the oral cavity. Partitioning CF lung communities into core and transient taxa overcomes this issue by focusing only on taxa posited to have large influences in CF pathophysiology, given their substantial presence and/or abundance (van der Gast et al., 2011). In a refinement to this method, we utilize neutral community models to identify taxa that show ecologically relevant patterns of abundance and prevalence, defining transient taxa as those that

are well-fit to neutral processes, and identifying non-neutral taxa which are more likely to interact with either the host or other lung microbiota, and that may influence CF disease ecology.

In evaluating neutral models, we consider two factors: goodness of fit (R^2), and inferred immigration probability (*m*), (**Fig 4.2**). Communities with high fits are consistent with model assumptions of both neutral immigration (random dispersion) and neutral selection (no competitive advantages). However, there are numerous ways to interpret poor fits. For example, non-neutral dispersal (e.g. some more virulent taxa have higher dispersal rates (Chung et al., 2017)) and non-neutral selection (e.g. certain taxa deterministically outcompete other members (Friedman et al., 2017), or host immunity preferentially suppresses certain taxa) are consistent with observed deviations from neutrality.

Application of neutral models across studies identifies canonical CF pathogens as non-neutral. Specifically, pathogens have abundance patterns consistently below those predicted by the neutral model for their given prevalence. In support of this method, we also find that four out of the five non-pathogens that were identified consistently below-neutral (*Prevotella*, *Porphyromonas*, *Fusobacterium*, and *Rothia*) have been reported as potential key taxa in CF lung ecology. *Prevotella* and other anaerobic taxa may increase *Pseudomonas* abundance through putative cross-feeding interactions such as mucin degradation (Flynn et al., 2016). A drop in *Porphyromonas* often predicted future *Pseudomonas* acquisition (Keravec et al., 2019). *Fusobacterium* and *Rothia* were both identified as positive predictors of lung function (C. Y. Zhao et al., 2021), and *Rothia* has been found to modulate host inflammatory responses (Rigauts et al., 2022).

However, the model does not explicitly predict the cause of their non-neutral distribution. These ‘below the curve’ distributions could be a product of selection (e.g. host suppression of pathogens), competitive advantage (pathogen expansion post initial seeding), dispersal (increased pathogen infectivity and virulence), or some combination of the three. With that in mind, we hypothesize that in dispersal-limited (low *m*) environments, larger deviations in non-neutral

dispersal are required to produce the same effect as smaller deviations in non-neutral selection. Given the overall lower *m* values, we therefore suspect that these patterns may be the result of low dispersal between patients, with a higher local pathogen competitive advantage.

We identify two limitations to this analysis. First, we do not have a good comparative interpretation of migration rates across other neutral model systems, as few references publish their *m* values (Burns et al., 2016; Sloan et al., 2006; Venkataraman et al., 2015). Thus, further work, such as metabolic assays, or experimental interrogation (e.g. *in vitro* synthetic communities, using the platform we develop in Chapter 3) is needed to elucidate these mechanisms.

Second, we assume that the source dispersal is from the metapopulation of CF microbiomes, and this source is adequately represented by averaging the taxa abundances across all samples. However, physiologically we suspect dispersal to come from the oral cavity as well as environmental sources of pathogens. Although are numerous paired studies already included in our collated dataset, we've removed all non-sputum samples for this analysis. In future work, we will revisit these paired saliva-sputum samples to more explicitly model source communities.

We find that lung pulmotypes can be separated into *Pseudomonas*-dominant, Oral Anaerobe, and Other Pathogen-dominant communities. These pulmotypes are generally compositionally distinct, and separatable by transition patterns. We are not the first study to apply pulmotyping to CF. Hampton et al. (Hampton, Thomas, van der Gast, O'Toole, Stanton, et al., 2021) (PRJNA420343 and PRJEB30646, **Fig 4.3D**) recently identified five pulmotypes across 167 pwCF: three *Pseudomonas*-dominant, one oral-dominant, and one mixed bag of people with *Achromobacter* or *Burkholderia*-dominant microbiomes. However, only the average of all communities assigned to these clusters is reported, so it is difficult to determine the individual sample composition. The eight pulmotypes identified in Widder et al. (PRJNA423040, PRJNA 756039, **Fig 4.3D**) well-align with ours in both composition and frequency. Their analysis includes *Pseudomonas*, *Pseudomonas*-*Staphylococcus* codominant, *Burkholderia*, and

Staphylococcus pulmotypes. However, this approach does not capture *Achromobacter* and *Haemophilus* in separate pulmotypes, likely due to their relatively lower prevalence. Instead, samples dominated by *Achromobacter* and *Haemophilus*, and potentially other rarer community types, are grouped into more heterogeneous pulmotypes. While *a priori* we may recognize this underfitting given prior knowledge of the importance of these pathogens in CF, this presents a challenge to novel pulmotype discovery that is likely only overcome by pulmotyping on larger and broadly representative sample collections.

Similar to Widder et al., (Widder et al., 2022) our analysis also finds that samples dominated by *Pseudomonas* can be partitioned into multiple different PA pulmotypes (**Fig 4.3**). The PA pulmotypes often transition into each other (**Fig 4.5, Fig 4.6**), but are distinguishable by their broader transition patterns. For example, p-III is primarily linked to both PA and OA pulmotypes, whereas many individuals transition between p-IV and OP states. We speculate that the transitions between PA states with higher *Pseudomonas* loads to lower ones corresponds to antibiotic treatment, as similar transitions were identified in Widder et al. This is consistent with the notion that antibiotic treatment of *Pseudomonas* leads to variable outcomes (Chapter 3), and may either reduce pathogen loads to a more oral anaerobe-dominant state, or potentially allow for competitive release of other pathogens (Varga et al., 2021) and therefore promote transitions into distinct pulmotypes. We were ultimately unable to assess this hypothesis given a lack of published, standardized antibiotic information, and we suggest that the field move towards recording and sharing this data where possible to aid in future studies.

One advantage of pulmotyping analyses is that they establish general benchmarks for synthetic community systems. Reliably inferring community interaction parameters often requires a combination of a combinatorial exploration of initial conditions with high sampling density. Our pulmotyping analysis has identified candidate initial conditions with known transitions. These may provide the foundations for *in vitro* and *in silico* experiments to better understand microbiome dynamics in CF.

In the context of my PhD, this final data chapter represents the threshold from my current to future work in this domain. In this chapter we have developed and applied a robust pipeline to generate an unprecedented combined dataset representing over 1000 pwCF. Using the dataset, we identify ecologically relevant taxa in CF lung microbiota and distill complex, high-dimensional microbiome data into distinct, globally robust lung community types. These pulmotypes form the basis of a universal CF microbiome categorization schema, potentially enabling the development of more individualized microbiome therapies. In future work building on this thesis, we will use this dataset to assess the value of alternate approaches to CF microbiome classification and prediction.

4.5 Methods

4.5.1 Dataset Curation

On February 9th, 2022 we searched NCBI-SRA for 16S-sequenced sputum surveys of CF lung microbiomes using the following query:

> (((cystic fibrosis OR cf) AND (lung OR respiratory OR sputum OR airway))) AND (amplicon[Strategy] OR other[Strategy]).

This returned 107 potential BioProject numbers. We manually searched for corresponding publication for each repository and excluded all studies that lacked corresponding methods or any information to match sputum samples with sample donor information. BioProjects that were not demultiplexed before SRA submission were excluded from this study.

In total, 36 BioProjects passed our inclusion criteria. For each BioProject, we used the same pipeline. All function parameters were set to defaults unless otherwise noted. First, we pulled all associated sample fastq files. As studies sequenced different 16S regions using different primers, we standardize all sequences by applying a uniform quality filter (maxEE=5) and trimming the first 20bp and truncating to 140bp (trimLeft=20; truncLen=140) using the

dada2::filterAndTrim function in R, with the remaining parameters set to defaults. Samples for which fewer than 20% of the reads passed this sequencing quality filter were discarded.

To optimize the dada2 error inference step, we included all samples that passed our filters, including non-sputum samples. We used dada2::dada to infer sequencing error rates and removed chimeras using dada2::removeBimeraDenovo. We assigned taxonomy by BLASTing each sequence against the NCBI 16S ribosomal RNA database, identifying the top 10 alignments, and reporting the most frequent genus call. All subsequent data handling was built using the phyloseq packages in R. We successfully analyzed 5201 samples using this pipeline and subsequently removed non-sputum samples such as sequencing controls and paired samples from other body sites, as well as non-observational (experimentally manipulated) samples from further downstream analysis.

The resulting dataset contained 4171 sputum samples with manually curated and matched sample metadata. Sputum samples contained a median of 20061 reads (median 19497, range 1045-1264313). DNA sequences were assigned to 1090 distinct genera. To mitigate study-specific sequencing bias, we only analyze genera that were detected across the majority (>18) of studies. This yielded 74 genera across our dataset for downstream analysis. To account for the variability in total read output for each individual study, all samples were then subsampled to a sequencing depth of 2000 reads; 145 samples were below this read threshold and thus discarded, yielding 4026 sputum samples across 1184 subjects.

4.5.2 Sloan's Neutral Community Model

The crux of the Sloan neutral model is the following relationship between the occupancy of a species, i.e. how many samples contain reads from a given taxa, and the overall relative abundance of that species in the source community:

$$x_i \sim Beta[N_T m p_i, N_T m(1 - p_i)]$$

where x_i is the occupancy frequency of the i-th species, p_i is the relative abundance of the i-th species in the source community, N_T is the total number of microbes across all samples, and m is the immigration frequency. Because we can directly measure N_T, our models effectively fit a single parameter.

We take inspiration from the fitting procedures outlined in Venkataraman et al. and Burns et al. (Burns et al., 2016; Venkataraman et al., 2015). For each BioProject, we calculate x_i for all taxa with a limit of detection of 1/2000. We assume that species abundance in the source community, p_i, is well approximated by the mean abundance across all samples in a community. We estimate m such that the joint probability $\Pi_i[p(x_i|N_T m, p_i)]$ is maximized. We also estimate 95% confidence intervals. Taxa that lie above and below the interval are identified as non-neutral. Overall goodness-of-fit is often reported using R^2 (values close to 1 imply the taxa distributions are consistent with a neutral generative process, but given the nonlinearity of the model, negative values can occur).

4.5.3 Dirichlet Multinomial Modeling

We construct non-neutral microbiomes by removing all reads assigned to neutral taxa before rarefaction. Following the procedure outlined in Holmes et al. (Holmes et al., 2012), we partition samples into pulmotypes using Dirichlet Multinomial Mixture (DMM) modeling. DMMs have been used to perform unsupervised clustering on microbiome datasets across numerous contexts (Costea et al., 2017; Holmes et al., 2012; Wang et al., 2022; Widder et al., 2022). We fit DMMs using the DirichletMultinomial R package (Morgan, 2020). To avoid individual over-representation, we only include the earliest available sputum sample for each subject during model training. Cluster assignment confidence was calculated as the likelihood of observing a given sample in its maximally likely component in the fit DMM model.

4.5.4 Compositional and Clinical differences

We test for compositionally distinct studies using a leave-one-out approach. Analysis of similarity (ANOSIM) tests were performed on Bray-Curtis distances between each study and the composite of all other studies using the anosim function in the vegan R package (Oksanen et al., 2019). We then aggregate studies by reporting center and perform both leave-one-out and pairwise center comparisons using ANOSIM. We also test pulmotype compositional differences using pairwise ANOSIM tests. For each set of comparisons, significance was assessed at p=0.05 (Bonferroni-corrected). Meaningful effect sizes were assessed at R>0.4, following previously established convention (Quinn, Lim, et al., 2016; Roberts et al., 2008).

To test for clinical differences between pulmotypes, we compare the distributions of reported lung function scores (ppFEV1) using a Wilcoxon rank-sum test. Significance was also assessed at p=0.05 (Bonferroni-corrected), with meaningful effect size set as a difference in mean ppFEV1 > 5%.

4.5.5 Bidirectionality of pulmotype transitions

We calculate transition pulmotypes by first assigning all longitudinal samples to pulmotypes using our trained 14-component DMM. As samples are taken at variable intervals, we infer sample order using reported sampling date. We calculate pulmotype transition frequencies by assessing changes in pulmotype across consecutive samples. We discard transitions where the product of the source and target cluster assignment confidences was less than 0.8. Given the variable number of total samples per individual, we avoid overfitting by reporting as the number of individuals in which a given transition is observed for each possible pulmotype transition pair.

To test for bidirectionality of pulmotype transitions, we calculated the likelihood of observing the distribution of forward and reverse transitions frequencies as a binomial distribution with $p = 0.5$. All implementation was performed in base R.

Chapter 5: Discussion

5.1 Summary of Work

The overarching goal of this work is to offer clinicians a robust method to integrate CF microbiome information into clinical decision-making and ultimately improve treatment outcomes. Although advances in sequencing have brought increasing attention to the polymicrobial context of chronic infections, clinical integration of microbiome information is made difficult by the large between-individual variability. Thus, accomplishing this goal requires understanding the array of host-microbiome relationships and ecological processes that shape individual CF microbiomes and their impacts on the health of people with CF (pwCF).

Understanding generalizable host-microbiome relationships requires large, annotated datasets as well as robust algorithms to tackle the high-dimensionality of microbiome data. As more clinical departments are investing in machine learning (ML) and medical AI resources, our approach is to apply machine learning to biomarker discovery.

In **Chapter 2**, we show that basic machine learning techniques predicting lung function from microbiome composition can identify both known and novel predictors of health, despite sample-size limitations. While we use the discovery of known predictors as performance benchmarks, the lack of a broad set of gold standards – a common feature of ML/AI applications in other fields – limits our ability to draw conclusions from our novel predictors. Consequently, it is difficult to estimate the representativeness of our training dataset. Nevertheless, we find one of these novel predictors, *Rothia*, has recently been shown to mitigate host inflammatory reactions to *Pseudomonas* (Rigauts et al., 2022), further supporting the utility of these hypothesis-generating methods.

While our ML approaches can identify host-microbiome associations, these methods are unable to directly determine causality. Thus, we turn to experimental interrogation. In **Chapter 3**,

we develop a synthetic microbiome system to study the ecological interactions of common CF taxa as well as their community response to antibiotic perturbation. We find that in a controlled, antibiotic-free *in vitro* setting, community dynamics are highly repeatable and tend towards commensal-dominated states. In contrast, antibiotic perturbation produces alternate pathogen-dominated end states. Our *in vitro* system provides an established baseline for future work that may more accurately represent CF lung environment conditions and treatment strategies (e.g. addition of inflammatory biomarkers or testing cycling between antibiotics rather than continuous exposure).

In **Chapter 2**, we use bootstrapping methods under the assumption that variation across the resampled datasets adequately represents the process of sampling from the full population of individuals who visit Emory/Georgia Tech-affiliated CF clinics. In lieu of testing additional models or applying techniques to estimate out-of-sample variation, we propose additional sample collection and collation of published sputum samples into a single standardized repository. In **Chapter 4**, we manually curate such a large CF sputum sample database and demonstrate the utility of such an approach. Using this dataset, we identify ecologically-relevant taxa found across published studies and show that an unsupervised clustering algorithm partitions microbiome data into 13 pulmotypes. Using available clinical metadata, we show these pulmotypes are represented across studies, and represent compositionally, clinically, and transitionally distinct communities. Our goal is for this dataset to serve as a general benchmarking tool for future CF-specific machine learning algorithm development.

There are additional limitations to this ensemble dataset approach. Non-standardized publication procedures created a need for time-intensive manual curation. We developed a standardization protocol and automated as much of the analysis pipeline as possible but found numerous study-specific record-keeping schema that could not be parsed without user intervention.

In light of these challenges we suggest a series of practical suggestions to improve data re-use. We propose a publication requirement for data columns specifying a unique, HIPAA-compliant subject identification number for each individual that a given sample came from and a separate sample number column. Inclusion of sample collection dates are imperative as well, given the changing landscape of available CFTR modulator or effector drugs and their effects of CF microbiomes (Sosinski et al., 2021). These standardizations will allow for easier data pooling, enabling clinicians and researchers to better access and utilize the existing body of published CF microbiome literature.

5.2 Future Work

We identify three immediate areas of future work from this thesis. First, in **Chapter 4**, we take a neutral model approach to identifying ecologically-relevant taxa. One of our primary model assumptions is that dispersal is from the metapopulation of CF microbiomes. However, we suspect that dispersal is from a combination of proximal sources, including the nasopharynx, oral cavity, and environmental sources. While we can identify pathogens and other taxa that we would expect to follow non-neutral distributions, this averaging method presents a limitation to how we can interpret our results.

Venkataraman et al., use paired oral samples as the source community. In our collated dataset, we've included studies that contain paired saliva-sputum samples, as well as paired samples from other sources. We propose using these paired samples to compare dispersal from alternate sources to better identify ecologically relevant taxa in CF lung communities.

Second, our algorithmic results (**Chapters 2** and **4**) are aimed at elucidating broad patterns in the field of CF. However, given the high between-individual variation, these results require additional robustness analyses to be clinically applicable. For example, in **Chapter 2** we utilize a 70-30 train-test split before data analysis. Re-running the initial splits or utilizing nested

cross validation and comparing resultant ensemble-selected features for each split or fold would lend further credibility to the out-of-sample applicability of the features we selected.

In **Chapter 4**, we recognize an additional robustness challenge in the total number of pulmotypes selected. In this chapter, we discuss validation methods for individual pulmotypes and conclude that while some of the discrete categorizations make sense in light of known CF microbiological patterns (e.g. rare, opportunistic pathogens dominate individual pulmotypes), others seem to be arbitrarily discretizing a continuous space (e.g. all the *Pseudomonas*-dominant pulmotypes). The latter is consistent with overfitting our data. Looking at fits across all *k*-component DMMs, we see that fit values plateau between 8 and 20 (**Fig 4.3**). We find a similar pattern when running our analysis on all taxa, as fits for either 6 or between 9 and 15 have similar values (**Fig 4.S1**).

These plateaus challenge the notion that there is a singular best fit, and that a measure of uncertainty should be included when reporting the number of pulmotypes our analyses return. Thus, future work involves applying methods to assess the robustness of the total number of pulmotypes we identify. For example, k-fold cross validation or bootstrap resampling may be used to generate sets of DMMs which can be compared for overall consistency. Alternatively, we may apply machine learning techniques such as bootstrap aggregation to stabilize our results while also reducing overfitting. Given the size of our datasets, these methods demand significant computational resources and will likely require further dimensionality reduction to achieve reasonable run times.

Lastly, in this thesis, we've provided the foundations for applying machine learning to CF microbiomics for supervised biomarker discovery and unsupervised pulmotype classification. While this approach begins to address sample size limitations, there is still a lack of microbiome-specific machine learning benchmarks. While these benchmark-driven approaches have come under criticism for promoting metric improvements over scientific discovery (Raji et al., 2021), they still provide a common ground to evaluate and cross-compare model performances.

Microbiome research often lacks these benchmarks, but the tools and datasets we developed in this work can be refined to address this. We propose mapping microbiome composition onto future lung health as one of these standardized questions. While overall individual health is due to a multi-faceted combination of factors, ppFEV1 is the most commonly reported health metric. In future work, as publicly available datasets grow, we can begin to assess better clinical disease metrics. Nevertheless, this lung-function centric approach provides foundational benchmarks for the field. From these foundations, we may begin to assess algorithmic improvements while testing the utility other common ML/AI methods, such as dataset augmentation or transfer learning, and move towards developing individualized microbiome-informed therapies for people with CF.

APPENDICES

Appendix A: Supplemental Methods - Microbiome data enhances predictive models of lung function in people with cystic fibrosis

A.1.1 Detailed Sequencing Analysis

Samples were sent to MR DNA Lab (Shallowater, TX) for DNA extraction, sequencing library preparation, Miseq sequencing, and absolute 16S quantitation. Microbiology culture results were obtained for sputum samples sent to the Clinical Microbiology laboratory on the same day as samples for sequencing were collected.

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.

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. Samples were rarefied to 17000 reads to guarantee equal sampling for subsequent analysis.

A.1.2: Machine Learning

To illustrate our machine learning approach, we begin with the model output trained on the full dataset (all 16S and metadata predictors, **Fig 2.S2**). **Figure 2.S2a** plots predicted versus observed lung function, for both the training dataset (data on 53 patients used to train model parameters) and the test dataset (data on 24 patients held back during model training). **Figure 2.S2b** highlights the parameters retained in the predictive model and their weighting. Our initial machine learning analysis (**Fig 2.S2**) suggest that the addition of non-pathogen 16S data improves model performance as evidenced by the retention of non-pathogen predictors in a penalized regression, and flags specific taxa as potential predictors.

Appendix B: Supplemental Tables and Figures

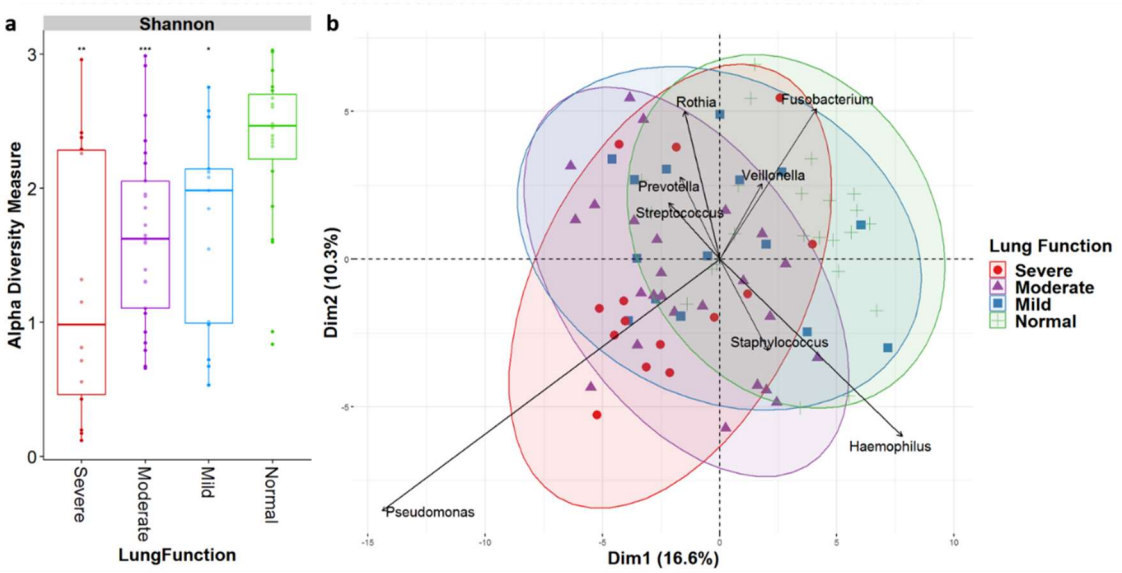

Figure 2.S1. Shannon diversity and ordination. a) Within-sample diversity (Shannon index) is lower in severe disease states compared to normal (Kruskal-Wallis, $p<0.01$). **b)** Between-sample diversity (Bray-Curtis PCoA on top 25 genera, centered log-ratio transformed). PCs 1 and 2 combined explain ~27% of the microbiome variance, and weakly clusters patients by Lung Function.

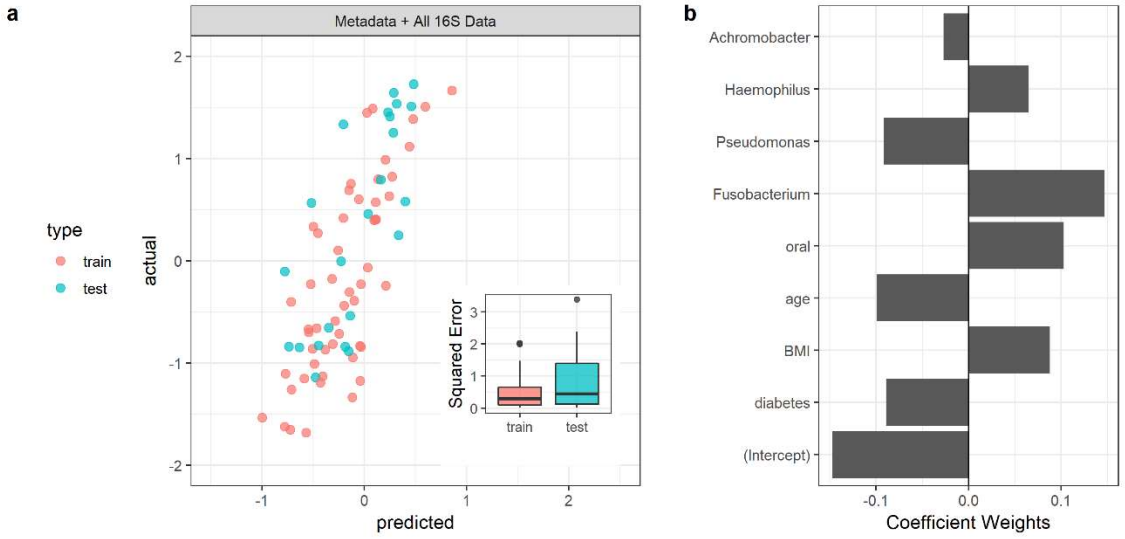

Figure 2.S2. ElasticNet-identified predictors of lung function. We train a baseline predictive model of ppFEV1 using the ElasticNet algorithm (alpha = 0.5) to perform feature selection. We assess the train-test holdout method on metadata + all 16S data. The train-test uses a standard 70-30 split (53 patient training set, 24 patient test set). (**a**) We plot model-predicted ppFEV1 values (scaled) against actual ppFEV1 values and calculate squared errors for each data point. We find that the model trained with the full dataset has the highest performance (see **Fig 2.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, Achromobacter and Pseudomonas abundance are selected as negative predictors of age whereas Haemophilus, Fusobacterium, oral taxa abundance, and as BMI are positive predictors.

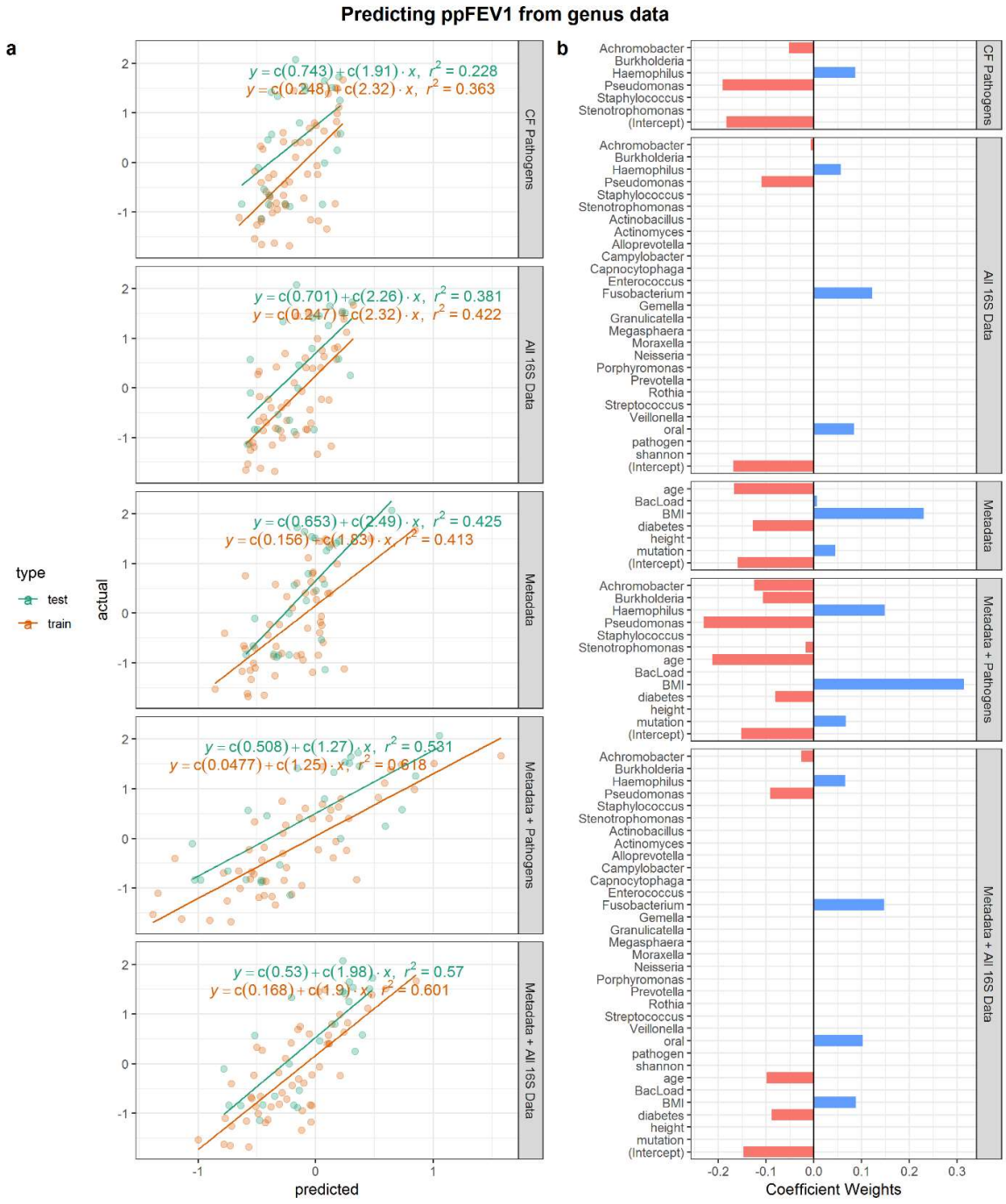

Figure 2.S3. Predicting ppFEV1 from genus data. To obtain baseline models, we assess the train-test holdout method on five input data sources: 16S quantitation of CF Pathogens (clr-transformed), all 16S data (clr-transformed), metadata, metadata + pathogens, and metadata + all 16S data.

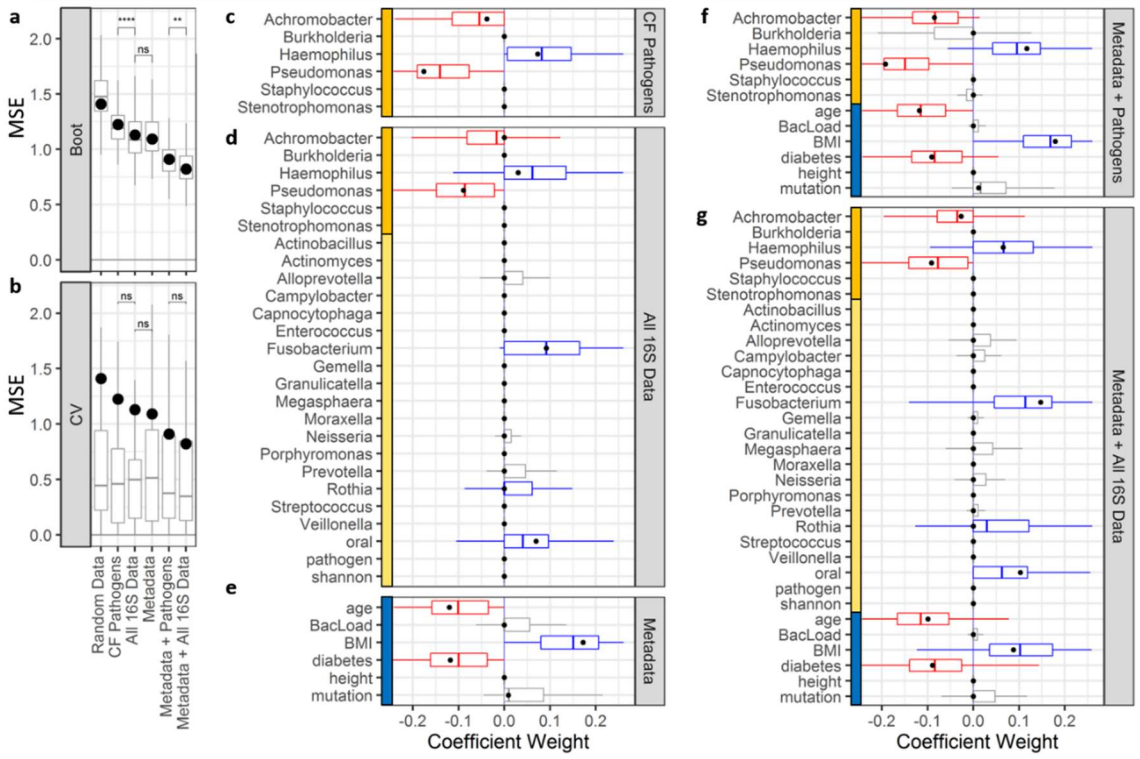

Figure 2.S4. Bootstrapped ElasticNet-identified predictors of lung function. ML models were trained using varying input datasets. **a)** 1000-fold bootstrapping and **b)** leave one out cross-validation (LOOCV) were used to generate prediction error (MSE) ranges across feature subsets. Models trained 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.

Table 3.S1. Differences in community structures across pathogen treatments. (Fig 3.3 data). The analysis of similarity (ANOSIM R) statistic captures the ratio of between-group to within-group variances; as R approaches 1, more variance is found between groups than within groups (Clarke, 1993). Passage 0 (inoculum) is omitted from the analyses. All but one R value is significant via permutation tests at $p < 0.05$. Applying the more conservative convention of $R > 0.4$ for significance (Quinn et al., 2016; Roberts et al., 2008), we find no significant effects of any pathogen treatment.

Treatment contrast	**ANOSIM *R* (permutation *p* value)**
Effect of mucoidy (SA present)	$R = 0.135$ ($p = 0.003$)
Effect of mucoidy (SA absent)	$R = 0.019$ ($p = 0.164$)
Effect of SA removal (mucoid PA)	$R = 0.068$ ($p = 0.026$)
Effect of SA removal (non-mucoid PA)	$R = 0.080$ ($p = 0.010$)
Effect of SA and PA removal (mucoid PA)	$R = 0.193$ ($p = 0.001$)

Table 3.S2. Antibiotic susceptibility in rich medium. Minimal Inhibitory Concentrations (MICs, in μg / ml) of synthetic community members were determined in rich medium.

	Tobramycin	**Meropenem**	**Ciprofloxacin**
P. aeruginosa PAO1	1	1	0.125
P. aeruginosa PDO300	2	1	<0.125
S. aureus	8	1	64
B. cenocepacia	≥ 128	32	32
A. xylosoxidans	≥ 128	4	8
S. mitis	4	0.25	2
N. subflava	8	0.125	0.125
R. mucilaginosa	128	0.125	16
H. influenzae	4	0.125	0.125
P. melaninogenica	≥ 128	0.125	1
V. parvula	32	0.25	1

Table 3.S3. Summary of hypothesis tests conducted in this study. Hypotheses (italic text) are organized by topic area (underlined text). Hypotheses in grey are rejected under the specific experimental conditions outlined in our study and may still apply in other contexts.

Topic and *hypotheses* (grey/black – rejected/supported under our specific experimental conditions)	**Previous literature**	**Key figures**
Synthetic microbiome structure and diversity *Single-locus biofilm mutations can produce large-scale community shifts.*	(McClean et al., 2015)	3.3
Competitive release / survival filter *Antibiotics enrich for resistant species (taxon enrichment).* *Antibiotics enrich for the sum of resistant species (functional enrichment).*	(Aspenberg et al., 2019; de Roode et al., 2004; Wale et al., 2017)	3.4, 3.6, 3.8
Variation and alternative stable states *Drug exposure increases variability across replicates* *Drug exposure produces alternate stable states*	(Estrela et al., 2022)	3.S1, 3.S3
Role of oral bacteria in CF microbiomes *Oral bacteria facilitate CF pathogens.* *Oral bacteria suppress CF pathogens.*	(Caverly & LiPuma, 2018; Flynn et al., 2016)	3.2, 3.4
Microbial interactions *Stressors increase inter-specific facilitation.* *B cenocepacia* facilitates *S. aureus* in a meropenem-dependent manner.	(Piccardi et al., 2019)	3.7
Models of CF microbiomes *Distinct experimental platforms are necessary to produce distinct 'pulmotypes'.* *A single 'meta-community' experimental platform can approach diverse CF 'pulmotypes', contingent on antibiotic exposures.*	(Jean-Pierre et al., 2021)	3.4, 3.5, 3.9, 3.10

Table 3.S4. Monoculture pre-culture conditions. The atmospheric environment specifies the oxygenation used for both the agar plate (Brain Heart Infusion (BHI) or chocolate agar) and liquid culture steps. The liquid medium supplements had the following concentrations: hemin, 15 mg / L; NAD, 15 mg / L; vitamin K1, 1 mg / L; L-lactate, 50 mM. Note that in subsequent experiments we simplified the protocol so that all bacteria were cultured first on chocolate agar plates, and then in a common medium of TSYE supplemented with hemin, NAD, vitamin K, and lactic acid

Genus	**Species**	**Oxygen tolerance**	**Liquid medium**	**Agar plates**
Pseudomonas	*aeruginosa*	Aerobic	TSYE	BHI
Staphylococcus	*aureus*	Aerobic	TSYE	BHI
Achromobacter	*xylosoxidans*	Aerobic	TSYE	BHI
Haemophilus	*influenzae*	Microaerophilic	TSYE+hemin+NAD	Chocolate agar
Streptococcus	*Mitis*	Aerobic	TSYE	BHI
Rothia	*mucilaginosa*	Aerobic	TSYE	Chocolate agar
Burkholderia	*cenocepacia*	Aerobic	TSYE	BHI
Neisseria	*subflava*	Microaerophilic	TSYE	Chocolate agar
Prevotella	*melaninogenica*	Anaerobic	TSYE+hemin+VK1	Chocolate agar
Veillonella	*parvula*	Anaerobic	TSYE+lactate	Chocolate agar

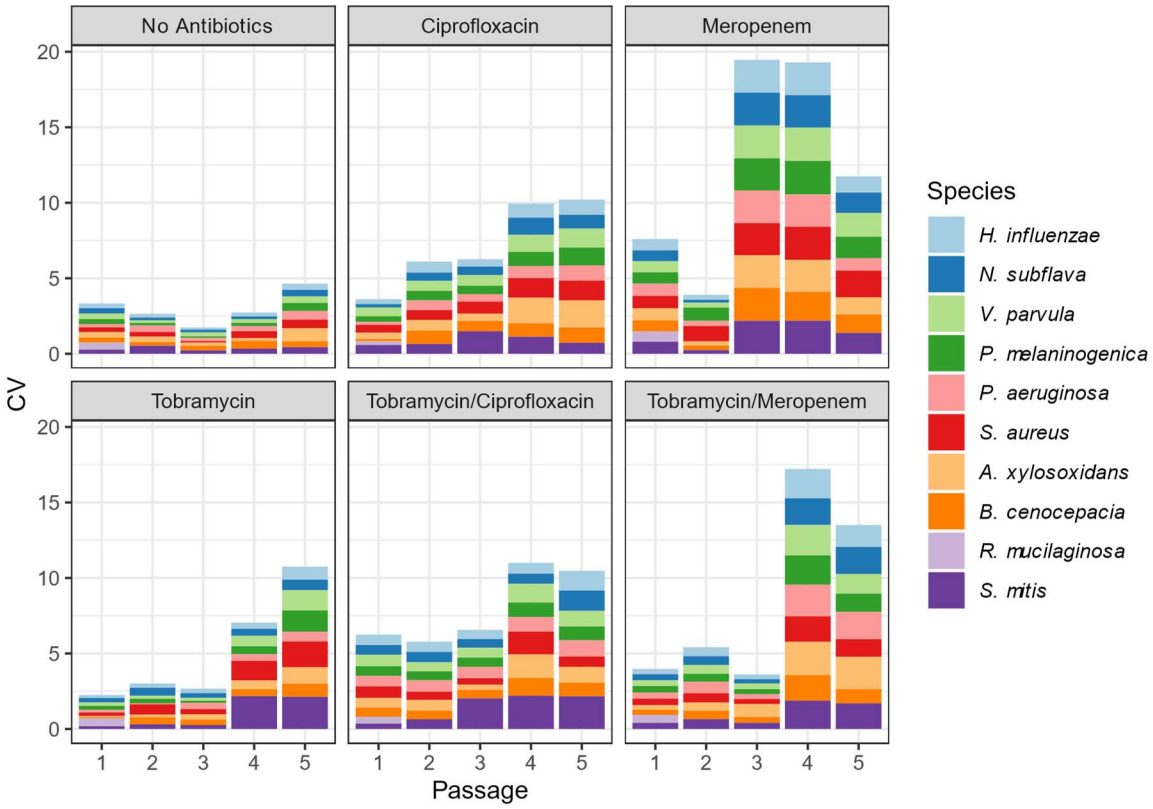

Figure 3.S1. Coefficient of Variation (CV) in species abundances, across replicates. CVs under different treatments through time. Stacked bars represent the CV of each individual species. CV is calculated as the standard deviation (across replicates) at each passage divided by its mean.

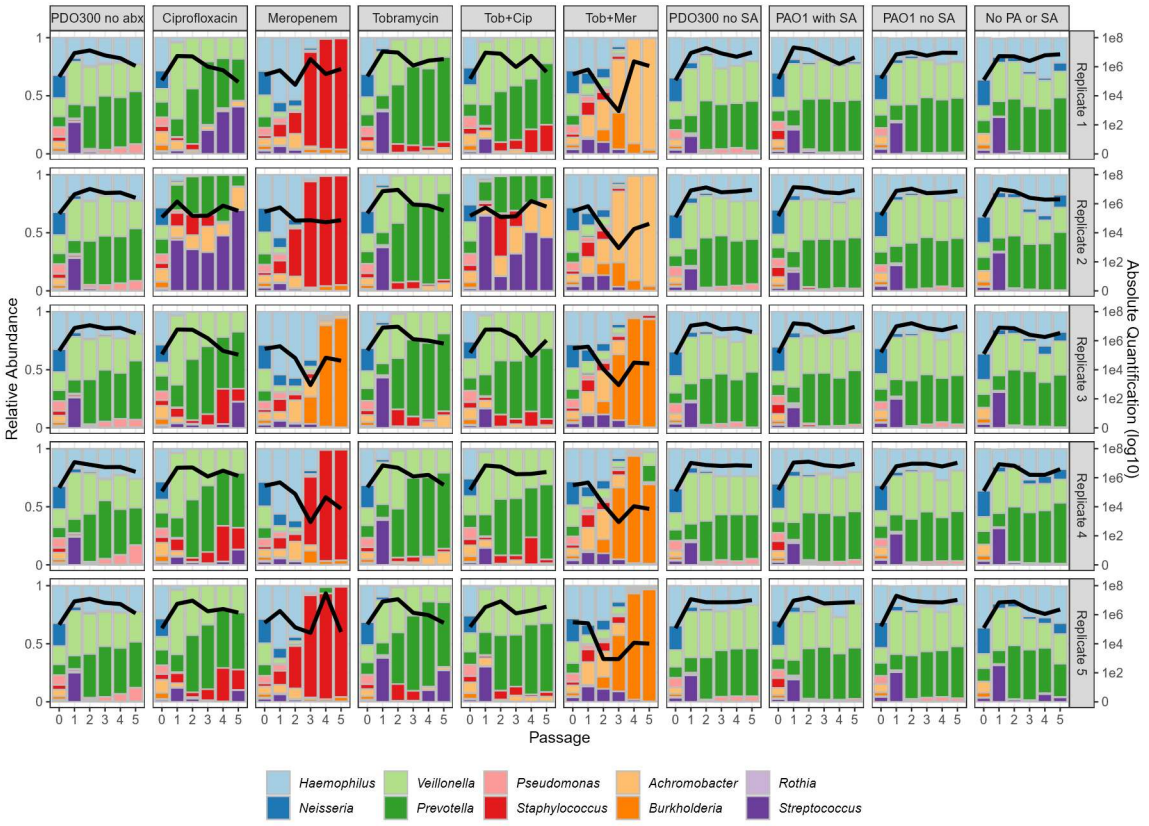

Figure 3.S2. Compositional and total abundances across all treatments (columns) and replicates (rows). For details, see legend of **Figure 3.1**.

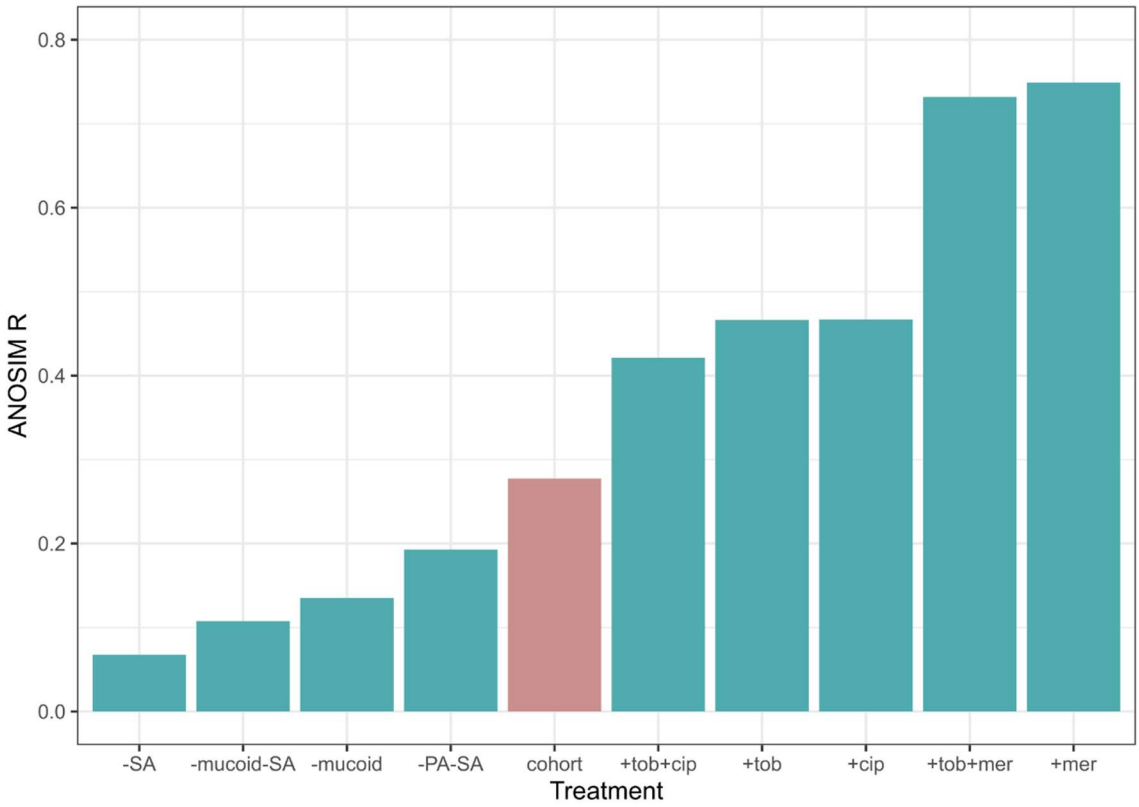

Figure 3.S3. Differences in community structures across experimental treatments and clinical data. The analysis of similarity (ANOSIM R) statistic captures the ratio of between-group to within-group variances; as R approaches 1, more variance is found between groups than within groups (Clarke, 1993). Green boxes are comparisons between absolute abundances of the 'no drug' reference treatment (PDO300 with SA, **Figure 3.1**) and experimental manipulations (pathogen treatments, **Figure 3.3**, and drug treatments, **Figure 3.4**). The pink box is the comparison between relative abundances of all experimental conditions **(Figs 3.2, 3.3, 3.4**) and all clinical conditions. Passage 0 (inoculum) is omitted from the analyses. All R values are significant with $p < 0.05$ (permutation test). Applying the more conservative convention of $R >$ 0.4 for significance (Quinn et al., 2016; Roberts et al., 2008), we find that only the antibiotic treatments achieve significant levels of community differentiation.

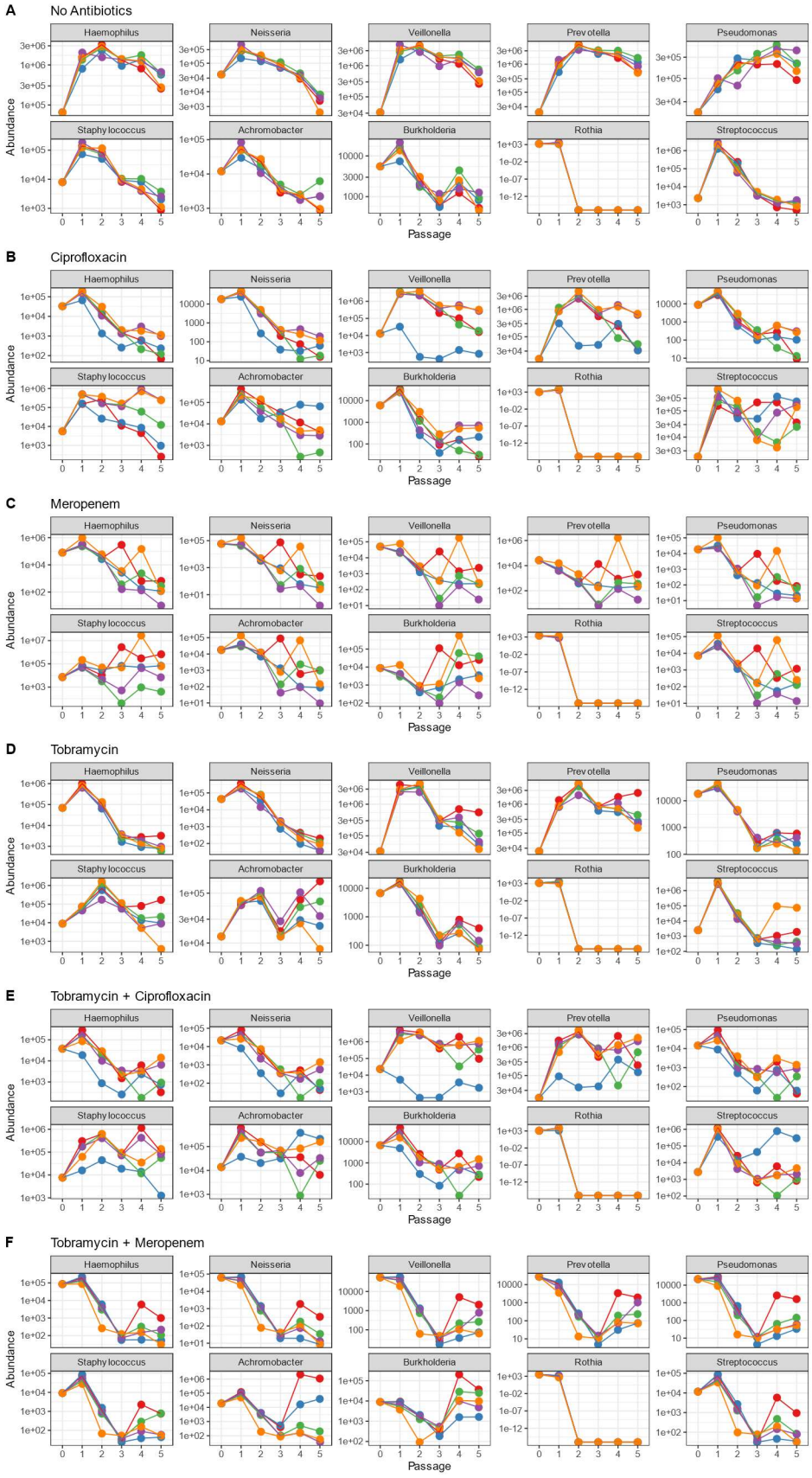
A
No Antibiotics
B
Ciprofloxacin
C
Meropenem
D
Tobramycin
E
Tobramycin + Ciprofloxacin
F
Tobramycin + Meropenem
Haemophilus
Neisseria
Veillonella
Prevotella
Pseudomonas
Staphylococcus
Achromobacter
Burkholderia
Rothia
Streptococcus
Abundance
Passage

Figure 3.S4. Temporal absolute abundances for all treatments (panels A-F), taxa (sub-panels) and replicates (colored lines). Data are plotted for each species under each antibiotic condition (A-F). The X-axis represents passage and Y-axis represents absolute abundance per ml. Individual replicates are connected by individually colored lines.

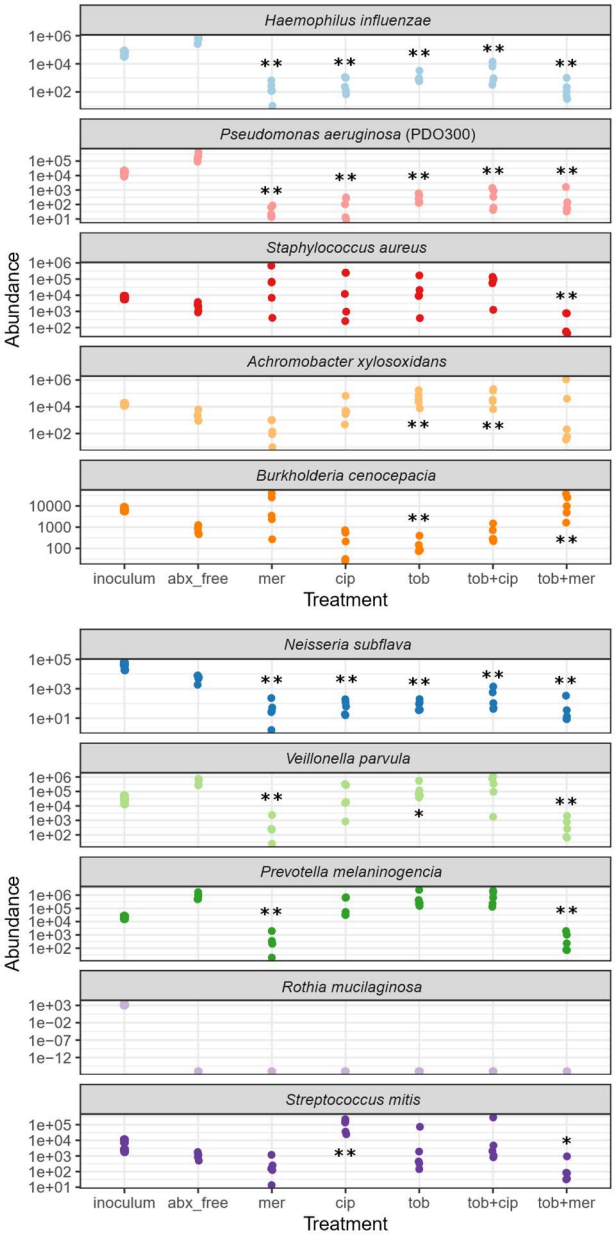

Figure 3.S5. Absolute microbe densities across antibiotic exposures. Each dot corresponds to an individual replicate of species-specific initial (inoculum) and final time-point absolute density under defined antibiotic treatments (data redrawn from **Figure 3.4**). abx_free = no antibiotic condition, mer = meropenem, cip = ciprofloxacin, tob = tobramycin. Asterisks denote significantly higher/lower final densities in presence of antibiotic, compared to antibiotic-free controls (two-tailed Wilcoxon test, * $p < 0.05$, ** $p < 0.01$).

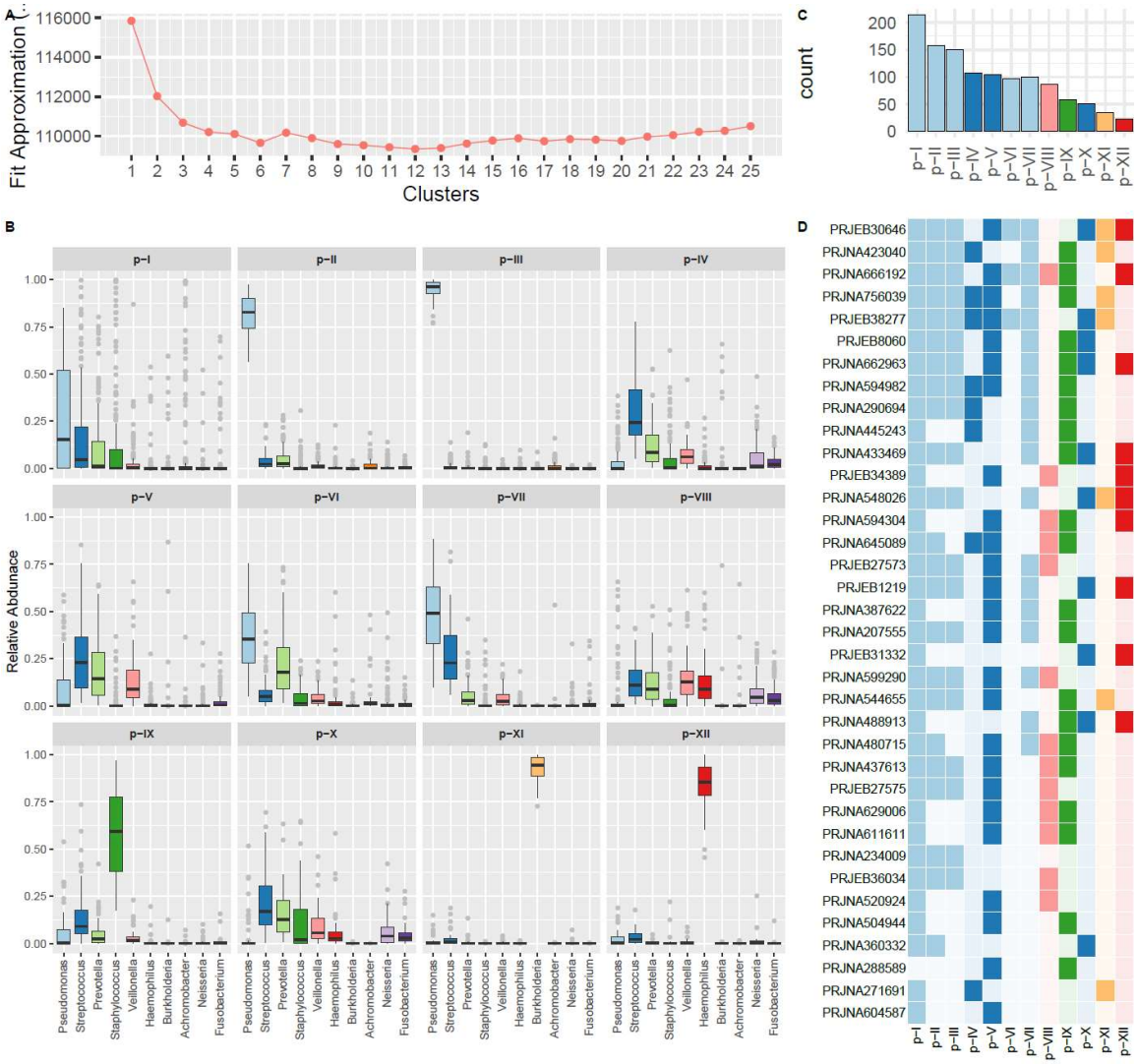

Figure 4.S1. All-data clustering identifies 12 pulmotypes. Clusters (k=1…25) were calculated using DMMs on individual snapshot data (N=1184) across 74 taxa, rarefied to 2000 sequences each. (**A**) Using the Laplace approximation of the negative log model evidence, we find a minimum at k=12 clusters (pulmotypes). (**B**) Boxplots of the top 10 taxa grouped by 12 pulmotypes. (**C**) Overall frequency of each pulmotype across the initial 1184 samples. (**D**) Pulmotypes represented in each study are shaded in. Studies are ordered by number of samples included.

References

Acosta, N., Heirali, A., Somayaji, R., Surette, M. G., Workentine, M. L., Sibley, C. D., Rabin, H. R., & Parkins, M. D. (2018). Sputum microbiota is predictive of long-term clinical outcomes in young adults with cystic fibrosis. Thorax, 73 (11), 1016-1025. https://doi.org/10.1136/thoraxjnl-2018-211510

Adamowicz, E. M., Flynn, J., Hunter, R. C., & Harcombe, W. R. (2018). Cross-feeding modulates antibiotic tolerance in bacterial communities. Isme j, 12 (11), 2723-2735. https://doi.org/10.1038/s41396-018-0212-z

Allen, R. C., McNally, L., Popat, R., & Brown, S. P. (2016). Quorum sensing protects bacterial co-operation from exploitation by cheats. Isme j, 10 (7), 1706-1716. https://doi.org/10.1038/ismej.2015.232

Anand, S., & Mande, S. S. (2018). Diet, microbiota and gut-lung connection. *Frontiers in Microbiology*, *9*(SEP). https://doi.org/10.3389/fmicb.2018.02147

Arumugam, M., Raes, J., Pelletier, E., le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., … Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, *473*(7346), 174–180. https://doi.org/10.1038/nature09944

Aspenberg, M., Sasane, S. M., Nilsson, F., Brown, S. P., & Waldetoft, K. W. (2019). Hygiene Hampers Competitive Release of Resistant Bacteria in the Commensal Microbiota. BioRxiv.

Bals, R., Weiner, D. J., & Wilson, J. M. (1999). The innate immune system in cystic fibrosis lung disease. *Journal of Clinical Investigation*, *103*(3), 303–307. https://doi.org/10.1172/JCI6277

Banerjee, S., Schlaeppi, K., & van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. Nature Reviews Microbiology, 16 (9), 567-576. https://doi.org/10.1038/s41579-018-0024-1

Baumgartner, M., Bayer, F., Pfrunder-Cardozo, K. R., Buckling, A., & Hall, A. R. (2020). Resident microbial communities inhibit growth and antibiotic-resistance evolution of Escherichia coli in human gut microbiome samples. Plos Biology, 18 (4), e3000465. https://doi.org/10.1371/journal.pbio.3000465

Beck, D., & Foster, J. A. (2014). Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0087830

Bjarnsholt, T., Alhede, M., Alhede, M., Eickhardt-Sørensen, S. R., Moser, C., Kühl, M., Jensen, P., & Høiby, N. (2013). The in vivo biofilm. Trends Microbiol, 21 (9), 466-474. https://doi.org/10.1016/j.tim.2013.06.002

Blainey, P. C., Milla, C. E., Cornfield, D. N., & Quake, S. R. (2012). Quantitative Analysis of the Human Airway Microbial Ecology Reveals a Pervasive Signature for Cystic Fibrosis. *Science Translational Medicine*, *4*(153), 1–7. https://doi.org/10.1126/scitranslmed.3004458

Bray, J. R., & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs, 27 (4), 325-349. https://doi.org/doi.org/10.2307/1942268

Brook, I. (2004). Beta-lactamase-producing bacteria in mixed infections. Clin Microbiol Infect, 10 (9), 777-784. https://doi.org/10.1111/j.1198-743X.2004.00962.x

Brook, I., Pazzaglia, G., Coolbaugh, J. C., & Walker, R. I. (1983). In-vivo protection of group A beta-haemolytic streptococci from penicillin by beta-lactamase-producing Bacteroides species. J Antimicrob Chemother, 12 (6), 599-606. https://doi.org/10.1093/jac/12.6.599

Brown, M. R., Collier, P. J., & Gilbert, P. (1990). Influence of growth rate on susceptibility to antimicrobial agents: modification of the cell envelope and batch and continuous culture studies. Antimicrob Agents Chemother, 34 (9), 1623-1628. https://doi.org/10.1128/aac.34.9.1623

Burns, A. R., Stephens, W. Z., Stagaman, K., Wong, S., Rawls, J. F., Guillemin, K., & Bohannan, B. J. M. (2016). Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME Journal*, *10*(3), 655–664. https://doi.org/10.1038/ismej.2015.142

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. https://doi.org/10.1038/nmeth.3869

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A, 108 Suppl 1, 4516-4522. https://doi.org/10.1073/pnas.1000080107

Caverly, L. J., & LiPuma, J. J. (2018). Good cop, bad cop: anaerobes in cystic fibrosis airways. Eur Respir J, 52 (1). https://doi.org/10.1183/13993003.01146-2018

Caverly, L. J., Lu, J., Carmody, L. A., Kalikin, L. M., Shedden, K., Opron, K., Azar, M., Cahalan, S., Foster, B., VanDevanter, D. R., Simon, R. H., & LiPuma, J. J. (2019). Measures of cystic fibrosis airway microbiota during periods of clinical stability. *Annals of the American Thoracic Society*, *16*(12), 1534–1542. https://doi.org/10.1513/AnnalsATS.201903-270OC

CFF. (2019). Cystic Fibrosis Foundation patient registry 2019 annual data report. Cystic Fibrosis Foundation. https://www.cff.org/Research/Researcher-Resources/Patient-Registry/2019-Patient-Registry-Annual-Data-Report.pdf

Chmiel, J. F., Aksamit, T. R., Chotirmall, S. H., Dasenbrook, E. C., Elborn, J. S., LiPuma, J. J., Ranganathan, S. C., Waters, V. J., & Ratjen, F. A. (2014). Antibiotic management of lung infections in cystic fibrosis. I. The microbiome, methicillin-resistant Staphylococcus aureus, gram-negative bacteria, and multiple infections. Ann Am Thorac Soc, 11 (7), 1120-1129. https://doi.org/10.1513/AnnalsATS.201402-050AS

Chung, H., Lieberman, T. D., Vargas, S. O., Flett, K. B., McAdam, A. J., Priebe, G. P., & Kishony, R. (2017). Global and local selection acting on the pathogen Stenotrophomonas maltophilia in the human lung. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14078

Cipolla, D., Blanchard, J., & Gonda, I. (2016). Development of Liposomal Ciprofloxacin to Treat Lung Infections. Pharmaceutics, 8 (1). https://doi.org/10.3390/pharmaceutics8010006

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, 18 (1), 117-143. https://doi.org/doi.org/10.1111/j.1442-9993.1993.tb00438.x

Coburn, B., Wang, P. W., Diaz Caballero, J., Clark, S. T., Brahma, V., Donaldson, S., Zhang, Y., Surendra, A., Gong, Y., Elizabeth Tullis, D., Yau, Y. C. W., Waters, V. J., Hwang, D. M., & Guttman, D. S. (2015). Lung microbiota across age and disease stage in cystic fibrosis. *Scientific Reports*, *5*, 1–12. https://doi.org/10.1038/srep10241

Conrad, D., Haynes, M., Salamon, P., Rainey, P. B., Youle, M., & Rohwer, F. (2013). Cystic fibrosis therapy: A community ecology perspective. *American Journal of Respiratory Cell and Molecular Biology*, *48*(2), 150–156. https://doi.org/10.1165/rcmb.2012-0059PS

Cornforth, D. M., Diggle, F. L., Melvin, J. A., Bomberger, J. M., & Whiteley, M. (2020). Quantitative Framework for Model Evaluation in Microbiology Research Using Pseudomonas aeruginosa and Cystic Fibrosis Infection as a Test Case. MBio, 11 (1). https://doi.org/10.1128/mBio.03042-19

Costea, P. I., Hildebrand, F., Manimozhiyan, A., Bäckhed, F., Blaser, M. J., Bushman, F. D., De Vos, W. M., Ehrlich, S. D., Fraser, C. M., Hattori, M., Huttenhower, C., Jeffery, I. B., Knights, D., Lewis, J. D., Ley, R. E., Ochman, H., O'Toole, P. W., Quince, C., Relman, D. A., … Bork, P. (2017). Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, *3*(1), 8–16. https://doi.org/10.1038/s41564-017-0072-8

Cowley, E. S., Kopf, S. H., LaRiviere, A., Ziebis, W., & Newman, D. K. (2015). Pediatric Cystic Fibrosis Sputum Can Be Chemically Dynamic, Anoxic, and Extremely Reduced Due to Hydrogen Sulfide Formation. MBio, 6 (4), e00767. https://doi.org/10.1128/mBio.00767-15

Cuthbertson, L., Rogers, G. B., Walker, A. W., Oliver, A., Green, L. E., Daniels, T. W., Carroll, M. P., Parkhill, J., Bruce, K. D., & van der Gast, C. J. (2016). Respiratory microbiota resistance and resilience to pulmonary exacerbation and subsequent antimicrobial intervention. *Isme j*, *10*(5), 1081–1091. https://doi.org/10.1038/ismej.2015.198

Cuthbertson, L., Walker, A. W., Oliver, A. E., Rogers, G. B., Rivett, D. W., Hampton, T. H., Ashare, A., Elborn, J. S., de Soyza, A., Carroll, M. P., Hoffman, L. R., Lanyon, C., Moskowitz, S. M., O'Toole, G. A., Parkhill, J., Planet, P. J., Teneback, C. C., Tunney, M. M., Zuckerman, J. B., … van der Gast, C. J. (2020). Lung function and microbiota diversity in cystic fibrosis. *Microbiome*, *8*(1), 1–13. https://doi.org/10.1186/s40168-020-00810-3

Cystic Fibrosis Foundation. (2021). Cystic Fibrosis Foundation 2020 Annual Data Report. *Cystic Fibrosis Foundation Patient Registry*, 1–96.

Darch, S. E., Kragh, K. N., Abbott, E. A., Bjarnsholt, T., Bull, J. J., & Whiteley, M. (2017). Phage Inhibit Pathogen Dissemination by Targeting Bacterial Migrants in a Chronic Infection Model. MBio, 8 (2). https://doi.org/10.1128/mBio.00240-17

Darch, S. E., McNally, A., Harrison, F., Corander, J., Barr, H. L., Paszkiewicz, K., Holden, S., Fogarty, A., Crusz, S. A., & Diggle, S. P. (2015). Recombination is a key driver of genomic and phenotypic diversity in a Pseudomonas aeruginosa population during cystic fibrosis infection. Sci Rep, 5, 7649. https://doi.org/10.1038/srep07649

de Roode, J. C., Culleton, R., Bell, A. S., & Read, A. F. (2004). Competitive release of drug resistance following drug treatment of mixed Plasmodium chabaudi infections. Malar J, 3, 33. Dickson, R. P., Martinez, F. J., & Huffnagle, G. B. (2014). The role of the microbiome in exacerbations of chronic lung diseases. *The Lancet*, *384*(9944), 691–702. https://doi.org/10.1016/S0140-6736(14)61136-3

Doring, G., Flume, P., Heijerman, H., & Elborn, J. S. (2012). Treatment of lung infection in patients with cystic fibrosis: current and future strategies. J Cyst Fibros, 11 (6), 461-479. https://doi.org/10.1016/j.jcf.2012.10.004

Dugatkin, L. A., Perlin, M., Lucas, J. S., & Atlas, R. (2005). Group-beneficial traits, frequency-dependent selection and genotypic diversity: an antibiotic resistance paradigm. Proc Biol Sci, 272 (1558), 79-83. https://doi.org/10.1098/rspb.2004.2916

Estrela, S., & Brown, S. P. (2018). Community interactions and spatial structure shape selection on antibiotic resistant lineages. PLoS Comput Biol, 14 (6), e1006179. https://doi.org/10.1371/journal.pcbi.1006179

Estrela, S., Vila, J. C. C., Lu, N., Bajić, D., Rebolleda-Gómez, M., Chang, C. Y., Goldford, J. E., Sanchez-Gorostiaga, A., & Sánchez, Á. (2022). Functional attractors in microbial community assembly. Cell Syst, 13 (1), 29-42.e27. https://doi.org/10.1016/j.cels.2021.09.011

European Committee for Antimicrobial Susceptibility Testing of the European Society of Clinical, M., & Infectious, D. (2003). Determination of minimum inhibitory concentrations (MICs) of antibacterial agents by broth dilution [https://doi.org/10.1046/j.1469-0691.2003.00790.x]. Clinical Microbiology and Infection, 9 (8), ix-xv. https://doi.org/doi.org/10.1046/j.1469-0691.2003.00790.x

Farrell, J. M., Zhao, C. Y., Tarquinio, K. M., & Brown, S. P. (2021). Causes and Consequences of COVID-19-Associated Bacterial Infections. *Frontiers in Microbiology*, *12*(July), 1–6. https://doi.org/10.3389/fmicb.2021.682571

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., Su, L., Li, X., Li, X., Li, J., Xiao, L., Huber-Schönauer, U., Niederseer, D., Xu, X., Al-Aama, J. Y., … Wang, J. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms7528

Filkins, L. M., Hampton, T. H., Gifford, A. H., Gross, M. J., Hogan, D. A., Sogin, M. L., Morrison, H. G., Paster, B. J., & O'Toole, G. A. (2012). Prevalence of streptococci and increased polymicrobial diversity associated with cystic fibrosis patient stability. *J Bacteriol*, *194*(17), 4709–4717. https://doi.org/10.1128/jb.00566-12

Filkins, L. M. & O'Toole, G.A. (2015). Cystic Fibrosis Lung Infections: Polymicrobial, Complex, and Hard to Treat. PLoS Pathogens, 11, e1005258. https://doi.org/10.1371/journal.ppat.1005258

Flynn, J. M., Cameron, L. C., Wiggen, T. D., Dunitz, J. M., Harcombe, W. R., & Hunter, R. C. (2020). Disruption of Cross-Feeding Inhibits Pathogen Growth in the Sputa of Patients with Cystic Fibrosis. mSphere, 5 (2). https://doi.org/10.1128/mSphere.00343-20

Flynn, J. M., Niccum, D., Dunitz, J. M., & Hunter, R. C. (2016). Evidence and Role for Bacterial Mucin Degradation in Cystic Fibrosis Airway Disease. *PLoS Pathog*, *12*(8), e1005846. https://doi.org/10.1371/journal.ppat.1005846

Fodor, A. A., Klem, E. R., Gilpin, D. F., Elborn, J. S., Boucher, R. C., Tunney, M. M., & Wolfgang, M. C. (2012). The Adult Cystic Fibrosis Airway Microbiota Is Stable over Time and Infection Type, and Highly Resilient to Antibiotic Treatment of Exacerbations. *PLoS ONE*, *7*(9). https://doi.org/10.1371/journal.pone.0045001

Frayman, K. B., Armstrong, D. S., Grimwood, K., & Ranganathan, S. C. (2017). The airway microbiota in early cystic fibrosis lung disease. *Pediatric Pulmonology*, *52*(11), 1384–1404. https://doi.org/10.1002/ppul.23782

Friedman, J., Higgins, L. M., & Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology and Evolution*, *1*(5). https://doi.org/10.1038/s41559-017-0109

Gilbert, P., Collier, P. J., & Brown, M. R. (1990). Influence of growth rate on susceptibility to antimicrobial agents: biofilms, cell cycle, dormancy, and stringent response. Antimicrob Agents Chemother, 34 (10), 1865-1868. https://doi.org/10.1128/aac.34.10.1865

Giuste, F., Venkatesan, M., Zhao, C., Tong, L., Zhu, Y., Deshpande, S. R., & Wang, M. D. (2020). Automated Classification of Acute Rejection from Endomyocardial Biopsies. *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2020*. https://doi.org/10.1145/3388440.3412430

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, *8*(NOV), 1–6. https://doi.org/10.3389/fmicb.2017.02224

Goddard, A. F., Staudinger, B. J., Dowd, S. E., Joshi-Datar, A., Wolcott, R. D., Aitken, M. L., Fligner, C. L., & Singh, P. K. (2012). Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. Proc Natl Acad Sci U S A, 109 (34), 13769-13774. https://doi.org/10.1073/pnas.1107435109

Guest, J. F., Ayoub, N., McIlwraith, T., Uchegbu, I., Gerrish, A., Weidlich, D., Vowden, K., & Vowden, P. (2017). Health economic burden that different wound types impose on the UK's National Health Service. *International Wound Journal*, *14*(2), 322–330. https://doi.org/10.1111/iwj.12603

Hackman, A. S., & Wilkins, T. D. (1975). In vivo protection of Fusobacterium necrophorum from penicillin by Bacteroides fragilis. Antimicrob Agents Chemother, 7 (5), 698-703. https://doi.org/10.1128/aac.7.5.698

Hahn, A., Fanous, H., Jensen, C., Chaney, H., Sami, I., Perez, G. F., Koumbourlis, A. C., Louie, S., Bost, J. E., van den Anker, J. N., Freishtat, R. J., Zemanick, E. T., & Crandall, K. A. (2019). Changes in microbiome diversity following beta-lactam antibiotic treatment are associated with therapeutic versus subtherapeutic antibiotic exposure in cystic fibrosis. Sci Rep, 9 (1), 2534. https://doi.org/10.1038/s41598-019-38984-y

Hampton, T. H., Thomas, D., van der Gast, C., O'Toole, G. A., Stanton, B. A., O'Toole, G. A., & Stanton, B. A. (2021). Mild Cystic Fibrosis Lung Disease Is Associated with Bacterial Community Stability. *Microbiology Spectrum*, *9*(1), e0002921. https://doi.org/10.1128/Spectrum.00029-21

Halpin, A. L., de Man, T. J., Kraft, C. S., Perry, K. A., Chan, A. W., Lieu, S., Mikell, J., Limbago, B. M., & McDonald, L. C. (2016). Intestinal microbiome disruption in patients in a long-term acute care hospital: A case for development of microbiome disruption indices to improve infection prevention. Am J Infect Control, 44 (7), 830-836. https://doi.org/10.1016/j.ajic.2016.01.003

Heirali, A., Thornton, C., Acosta, N., Somayaji, R., Laforest Lapointe, I., Storey, D., Rabin, H., Waddell, B., Rossi, L., Arrieta, M. C., Surette, M., & Parkins, M. D. (2020). Sputum microbiota in adults with CF associates with response to inhaled tobramycin. Thorax, 75 (12), 1058-1064. https://doi.org/10.1136/thoraxjnl-2019-214191

Henke, M. O., & Ratjen, F. (2007). Mucolytics in cystic fibrosis. *Paediatric Respiratory Reviews*, *8*(1), 24–29. https://doi.org/10.1016/j.prrv.2007.02.009

Hogan, D. A., Willger, S. D., Dolben, E. L., Hampton, T. H., Stanton, B. A., Morrison, H. G., Sogin, M. L., Czum, J., & Ashare, A. (2016). Analysis of Lung Microbiota in Bronchoalveolar Lavage, Protected Brush and Sputum Samples from Subjects with Mild-To-Moderate Cystic Fibrosis Lung Disease. PLoS One, 11 (3), e0149998. https://doi.org/10.1371/journal.pone.0149998

Holden, M. T., Seth-Smith, H. M., Crossman, L. C., Sebaihia, M., Bentley, S. D., Cerdeno-Tarraga, A. M., Thomson, N. R., Bason, N., Quail, M. A., Sharp, S., Cherevach, I., Churcher, C., Goodhead, I., Hauser, H., Holroyd, N., Mungall, K., Scott, P., Walker, D., White, B., . . . Parkhill, J. (2009). The genome of Burkholderia cenocepacia J2315, an epidemic pathogen of cystic fibrosis patients. J Bacteriol, 191 (1), 261-277. https://doi.org/JB.01230-08 [pii]

Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE*, *7*(2). https://doi.org/10.1371/journal.pone.0030126

Hotterbeekx, A., Kumar-Singh, S., Goossens, H., & Malhotra-Kumar, S. (2017). In vivo and In vitro Interactions between Pseudomonas aeruginosa and Staphylococcus spp. Front Cell Infect Microbiol, 7, 106. https://doi.org/10.3389/fcimb.2017.00106

Huang, Y. J., & LiPuma, J. J. (2016). The Microbiome in Cystic Fibrosis. *Clinics in Chest Medicine*, *37*(1), 59–67. https://doi.org/10.1016/j.ccm.2015.10.003

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press. https://doi.org/10.1515/9781400837526

Jean-Pierre, F., Vyas, A., Hampton, T. H., Henson, M. A., & O'Toole, G. A. (2021). One versus Many: Polymicrobial Communities and the Cystic Fibrosis Airway. MBio, 12 (2). https://doi.org/10.1128/mBio.00006-21

Jorth, P., Ehsan, Z., Rezayat, A., Caldwell, E., Pope, C., Brewington, J. J., Goss, C. H., Benscoter, D., Clancy, J. P., & Singh, P. K. (2019). Direct Lung Sampling Indicates That Established Pathogens Dominate Early Infections in Children with Cystic Fibrosis. *Cell Rep*, *27*(4), 1190-1204.e3. https://doi.org/10.1016/j.celrep.2019.03.086

Jorth, P., Staudinger, B. J., Wu, X., Hisert, K. B., Hayden, H., Garudathri, J., Harding, C. L., Radey, M. C., Rezayat, A., Bautista, G., Berrington, W. R., Goddard, A. F., Zheng, C., Angermeyer, A., Brittnacher, M. J., Kitzman, J., Shendure, J., Fligner, C. L., Mittler, J., . . . Singh, P. K. (2015). Regional Isolation Drives Bacterial Diversification within Cystic Fibrosis Lungs. Cell Host Microbe, 18 (3), 307-319. https://doi.org/10.1016/j.chom.2015.07.006

Kassambara, A., & Mundt, F. (2020). factoextra: extract and visualize the results of multivariate data analyses R package version 1.0.7. In: R Foundation for Statistical Computing. Vienna.

Keravec, M., Mounier, J., Guilloux, C. A., Fangous, M. S., Mondot, S., Vallet, S., Gouriou, S., Le Berre, R., Rault, G., Férec, C., Barbier, G., Lepage, P., & Héry-Arnaud, G. (2019). Porphyromonas, a potential predictive biomarker of Pseudomonas aeruginosa pulmonary infection in cystic fibrosis. *BMJ Open Respiratory Research*, *6*(1), 1–5. https://doi.org/10.1136/bmjresp-2018-000374

Kidd, T. J., Canton, R., Ekkelenkamp, M., Johansen, H. K., Gilligan, P., LiPuma, J. J., Bell, S. C., Elborn, J. S., Flume, P. A., VanDevanter, D. R., & Waters, V. J. (2018). Defining antimicrobial resistance in cystic fibrosis. J Cyst Fibros, 17 (6), 696-704. https://doi.org/10.1016/j.jcf.2018.08.014

Kolpen, M., Hansen, C. R., Bjarnsholt, T., Moser, C., Christensen, L. D., van Gennip, M., Ciofu, O., Mandsberg, L., Kharazmi, A., Döring, G., Givskov, M., Høiby, N., & Jensen, P. (2010). Polymorphonuclear leucocytes consume oxygen in sputum from chronic Pseudomonas aeruginosa pneumonia in cystic fibrosis. Thorax, 65 (1), 57-62. https://doi.org/10.1136/thx.2009.114512

Konstan, M. W., Wagener, J. S., & VanDevanter, D. R. (2009). Characterizing aggressiveness and predicting future progression of CF lung disease. *Journal of Cystic Fibrosis : Official Journal of the European Cystic Fibrosis Society*, *8 Suppl 1*, S15–S19. https://doi.org/10.1016/s1569-1993(09)60006-0

Kraay, A. N. M., Nelson, K. N., Zhao, C. Y., Demory, D., Weitz, J. S., & Lopman, B. A. (2021). Modeling serological testing to inform relaxation of social distancing for COVID-19 control. *Nature Communications*, *12*(1), 1–10. https://doi.org/10.1038/s41467-021-26774-y

Kragh, K. N., Alhede, M., Jensen, P., Moser, C., Scheike, T., Jacobsen, C. S., Seier Poulsen, S., Eickhardt-Sørensen, S. R., Trøstrup, H., Christoffersen, L., Hougen, H. P., Rickelt, L. F., Kühl, M., Høiby, N., & Bjarnsholt, T. (2014). Polymorphonuclear leukocytes restrict growth of Pseudomonas aeruginosa in the lungs of cystic fibrosis patients. Infect Immun, 82 (11), 4477-4486. https://doi.org/10.1128/iai.01969-14

Kuti, J. L., Nightingale, C. H., Knauft, R. F., & Nicolau, D. P. (2004). Pharmacokinetic properties and stability of continuous-infusion meropenem in adults with cystic fibrosis. Clin Ther, 26 (4), 493-501. https://doi.org/10.1016/s0149-2918 (04)90051-3

Lam, J., Vaughan, S., & Parkins, M. D. (2013). Tobramycin Inhalation Powder (TIP): An Efficient Treatment Strategy for the Management of Chronic Pseudomonas Aeruginosa Infection in Cystic Fibrosis. Clin Med Insights Circ Respir Pulm Med, 7, 61-77. https://doi.org/10.4137/ccrpm.s10592

Laxminarayan, R., Duse, A., Wattal, C., Zaidi, A. K. M., Wertheim, H. F. L., Sumpradit, N., Vlieghe, E., Hara, G. L., Gould, I. M., Goossens, H., Greko, C., So, A. D., Bigdeli, M., Tomson, G., Woodhouse, W., Ombaka, E., Peralta, A. Q., Qamar, F. N., Mir, F., . . . Cars, O. (2013). Antibiotic resistance - the need for global solutions. The Lancet Infectious Diseases, 13 (12), 1057-1098. https://doi.org/10.1016/S1473-3099 (13)70318-9

Layeghifard, M., Li, H., Wang, P. W., Donaldson, S. L., Coburn, B., Clark, S. T., Caballero, J. D., Zhang, Y., Tullis, D. E., Yau, Y. C. W., Waters, V., Hwang, D. M., Guttman, D. S., Zhang, Y., Tullis, D. E., Yau, Y. C. W., Waters, V., Hwang, D. M., & Guttman, D. S. (2019). Microbiome networks and change-point analysis reveal key community changes associated with cystic fibrosis pulmonary exacerbations. *Npj Biofilms and Microbiomes*, *5*(1), 4. https://doi.org/10.1038/s41522-018-0077-y

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., Holt, R. D., Shurin, J. B., Law, R., & Tilman, D. (2004). The metacommunity concept: a framework for multi-scale community ecology. Ecol Lett, 7 (7), 601-613.

Li, J., Hao, C., Ren, L., Xiao, Y., Wang, J., & Qin, X. (2016). Data mining of lung microbiota in cystic fibrosis patients. *PLoS ONE*, *11*(10), e0164510. https://doi.org/10.1371/journal.pone.0164510

Limoli, D. H., Whitfield, G. B., Kitao, T., Ivey, M. L., Davis Jr., M. R., Grahl, N., Hogan, D. A., Rahme, L. G., Howell, P. L., O'Toole, G. A., & Goldberg, J. B. (2017). Pseudomonas aeruginosa Alginate Overproduction Promotes Coexistence with Staphylococcus aureus in a Model of Cystic Fibrosis Respiratory Infection. *MBio*, *8*(2). https://doi.org/10.1128/mBio.00186-17

Limoli, D. H., Yang, J., Khansaheb, M. K., Helfman, B., Peng, L., Stecenko, A. A., & Goldberg, J. B. (2016). Staphylococcus aureus and Pseudomonas aeruginosa co-infection is associated with cystic fibrosis-related diabetes and poor clinical outcomes. *European Journal of Clinical Microbiology and Infectious Diseases*, *35*(6), 947–953. https://doi.org/10.1007/s10096-016-2621-0

Lu, J., Carmody, L. A., Opron, K., Simon, R. H., Kalikin, L. M., Caverly, L. J., & LiPuma, J. J. (2020). Parallel Analysis of Cystic Fibrosis Sputum and Saliva Reveals Overlapping Communities and an Opportunity for Sample Decontamination. *MSystems*, *5*(4). https://doi.org/10.1128/msystems.00296-20

Lucas, S. K., Yang, R., Dunitz, J. M., Boyer, H. C., & Hunter, R. C. (2018). 16S rRNA gene sequencing reveals site-specific signatures of the upper and lower airways of cystic fibrosis patients. *Journal of Cystic Fibrosis*, *17*(2), 204–212. https://doi.org/10.1016/j.jcf.2017.08.007

Maliniak, M. L., Stecenko, A. A., & McCarty, N. A. (2016). A longitudinal analysis of chronic MRSA and Pseudomonas aeruginosa co-infection in cystic fibrosis: A single-center study. J Cyst Fibros, 15 (3), 350-356. https://doi.org/10.1016/j.jcf.2015.10.014

Malone, M., Bjarnsholt, T., McBain, A. J., James, G. A., Stoodley, P., Leaper, D., Tachi, M., Schultz, G., Swanson, T., & Wolcott, R. D. (2017). The prevalence of biofilms in chronic wounds: a systematic review and meta-analysis of published data. *Journal of Wound Care*, *26*(1), 20–25. https://doi.org/10.12968/jowc.2017.26.1.20

Martin, D. W., Schurr, M. J., Mudd, M. H., Govan, J. R., Holloway, B. W., & Deretic, V. (1993). Mechanism of conversion to mucoidy in Pseudomonas aeruginosa infecting cystic fibrosis patients. Proc Natl Acad Sci U S A, 90 (18), 8377-8381. https://doi.org/10.1073/pnas.90.18.8377

Mathee, K., Ciofu, O., Sternberg, C., Lindum, P. W., Campbell, J. I., Jensen, P., Johnsen, A. H., Givskov, M., Ohman, D. E., Molin, S., Hoiby, N., & Kharazmi, A. (1999). Mucoid conversion of Pseudomonas aeruginosa by hydrogen peroxide: a mechanism for virulence activation in the cystic fibrosis lung. Microbiology, 145 (Pt 6), 1349-1357.

McAdams, D., Wollein Waldetoft, K., Tedijanto, C., Lipsitch, M., & Brown, S. P. (2019). Resistance diagnostics as a public health tool to combat antibiotic resistance: A model-based evaluation. Plos Biology, 17 (5), e3000250. https://doi.org/10.1371/journal.pbio.3000250

McClean, D., McNally, L., Salzberg, L. I., Devine, K. M., Brown, S. P., & Donohue, I. (2015). Single gene locus changes perturb complex microbial communities as much as apex predator loss. Nat Commun, 6, 8235. https://doi.org/10.1038/ncomms9235

McKenney, E. S., Sargent, M., Khan, H., Uh, E., Jackson, E. R., Jose, G. S., Couch, R. D., Dowd, C. S., & van Hoek, M. L. (2012). Lipophilic Prodrugs of FR900098 Are Antimicrobial against Francisella novicida In Vivo and In Vitro and Show GlpT Independent Efficacy. PLoS One, 7 (10), e38167. https://doi.org/10.1371/journal.pone.0038167

Mitchell, J. (2011). Streptococcus mitis: walking the line between commensalism and pathogenesis. Mol Oral Microbiol, 26 (2), 89-98. https://doi.org/10.1111/j.2041-1014.2010.00601.x

Moran Losada, P., Chouvarine, P., Dorda, M., Hedtfeld, S., Mielke, S., Schulz, A., Wiehlmann, L., & Tummler, B. (2016). The cystic fibrosis lower airways microbial metagenome. ERJ Open Res, 2 (2). https://doi.org/10.1183/23120541.00096-2015

Morgan, M. (2020). DirichletMultinomial: Dirichlet-Multinomial Mixture Model Machine Learning for Microbiome Data. *Arkansas Academy of Science Proceedings*, *19*, R. package version 1.32.0. https://doi.org/10.1371/journal.pone.0030126.Author(s)Maintainer

Moriarty, T. F., McElnay, J. C., Elborn, J. S., & Tunney, M. M. (2007). Sputum antibiotic concentrations: implications for treatment of cystic fibrosis lung infection. Pediatr Pulmonol, 42 (11), 1008-1017. https://doi.org/10.1002/ppul.20671

Morley, V. J., Woods, R. J., & Read, A. F. (2019). Bystander Selection for Antimicrobial Resistance: Implications for Patient Health. Trends Microbiol, 27 (10), 864-877. https://doi.org/10.1016/j.tim.2019.06.004

Muhlebach, M. S., Hatch, J. E., Einarsson, G. G., McGrath, S. J., Gilipin, D. F., Lavelle, G., Mirkovic, B., Murray, M. A., McNally, P., Gotman, N., Davis Thomas, S., Wolfgang, M. C., Gilligan, P. H., McElvaney, N. G., Elborn, J. S., Boucher, R. C., & Tunney, M. M. (2018). Anaerobic bacteria cultured from cystic fibrosis airways correlate to milder disease: a multisite study. Eur Respir J, 52 (1). https://doi.org/10.1183/13993003.00242-2018

Muhlebach, M. S., Zorn, B. T., Esther, C. R., Hatch, J. E., Murray, C. P., Turkovic, L., Ranganathan, S. C., Boucher, R. C., Stick, S. M., & Wolfgang, M. C. (2018). Initial acquisition and succession of the cystic fibrosis lung microbiome is associated with disease progression in infants and preschool children. PLoS Pathog, 14 (1), e1006798. https://doi.org/10.1371/journal.ppat.1006798

Nelson, M. T., Wolter, D. J., Eng, A., Weiss, E. J., Vo, A. T., Brittnacher, M. J., Hayden, H. S., Ravishankar, S., Bautista, G., Ratjen, A., Blackledge, M., McNamara, S., Nay, L., Majors, C., Miller, S. I., Borenstein, E., Simon, R. H., LiPuma, J. J., & Hoffman, L. R. (2020). Maintenance tobramycin primarily affects untargeted bacteria in the CF sputum microbiome. Thorax, 75 (9), 780-790. https://doi.org/10.1136/thoraxjnl-2019-214187

Nguyen, A. T., & Oglesby-Sherrouse, A. G. (2016). Interactions between Pseudomonas aeruginosa and Staphylococcus aureus during co-cultivations and polymicrobial infections. Appl Microbiol Biotechnol, 100 (14), 6141-6148. https://doi.org/10.1007/s00253-016-7596-3

O'Connor, K., Zhao, C. Y., & Diggle, S. P. (2021). Frequency of quorum sensing mutations in Pseudomonas aeruginosa strains isolated from different environments. *BioRxiv*. https://doi.org/10.1101/2021.02.22.432365

O'Neill, K., Bradley, J. M., Johnston, E., McGrath, S., McIlreavey, L., Rowan, S., Reid, A., Bradbury, I., Einarsson, G., Elborn, J. S., & Tunney, M. M. (2015). Reduced bacterial colony count of anaerobic bacteria is associated with a worsening in lung clearance index and inflammation in cystic fibrosis. PLoS One, 10 (5), e0126980. https://doi.org/10.1371/journal.pone.0126980

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., Szoecs, E., & Wagner, H. (2019). *vegan: Community Ecology Package. R package version 2.5-5. https://CRAN.R-project.org/package=vegan*.

Palmer, K. L., Mashburn, L. M., Singh, P. K., & Whiteley, M. (2005). Cystic fibrosis sputum supports growth and cues key aspects of Pseudomonas aeruginosa physiology. J Bacteriol, 187 (15), 5267-5277. https://doi.org/10.1128/JB.187.15.5267-5277.2005

Perez-Vilar, J., & Boucher, R. C. (2004). Reevaluating gel-forming mucins' roles in cystic fibrosis lung disease. *Free Radical Biology and Medicine*, *37*(10), 1564–1577. https://doi.org/10.1016/j.freeradbiomed.2004.07.027

Persoon, A., Heinen, M. M., van der Vleuten, C. J. M., de Rooij, M. J., van de Kerkhof, P. C. M., & van Achterberg, T. (2004). Leg ulcers: a review of their impact on daily life. *Journal of Clinical Nursing*, *13*(3), 341–354. http://www.ncbi.nlm.nih.gov/pubmed/15009337

Piccardi, P., Vessman, B., & Mitri, S. (2019). Toxicity drives facilitation between 4 bacterial species. Proc Natl Acad Sci U S A, 116 (32), 15979-15984. https://doi.org/10.1073/pnas.1906172116

Pollitt, E. J., West, S. A., Crusz, S. A., Burton-Chellew, M. N., & Diggle, S. P. (2014). Cooperation, quorum sensing, and evolution of virulence in Staphylococcus aureus. Infect Immun, 82 (3), 1045-1051. https://doi.org/10.1128/iai.01216-13

Price, K. E., Hampton, T. H., Gifford, A. H., Dolben, E. L., Hogan, D. A., Morrison, H. G., Sogin, M. L., & O'Toole, G. A. (2013). Unique microbial communities persist in individual cystic fibrosis patients throughout a clinical exacerbation. Microbiome, 1 (1), 27. https://doi.org/10.1186/2049-2618-1-27

Qiu, P. (2020). Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, *11*(1), 1–9. https://doi.org/10.1038/s41467-020-14976-9

Quinn, R. A., Lim, Y. W., Mak, T. D., Whiteson, K., Furlan, M., Conrad, D., Rohwer, F., & Dorrestein, P. (2016). Metabolomics of pulmonary exacerbations reveals the personalized nature of cystic fibrosis disease. *PeerJ*, *4*, e2174. https://doi.org/10.7717/peerj.2174

Quinn, R. A., Whiteson, K., Lim, Y. W., Zhao, J., Conrad, D., Lipuma, J. J., Rohwer, F., & Widder, S. (2016). Ecological networking of cystic fibrosis lung infections. *Npj Biofilms and Microbiomes*, *2*(1), 0–1. https://doi.org/10.1038/s41522-016-0002-1

R Core Team. (2018). R: a language and environment for statistical computing. . In R Foundation for Statistical Computing www.R-project.org

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). *AI and the Everything in the Whole Wide World Benchmark*. *NeurIPS*. http://arxiv.org/abs/2111.15366

Read, A. F., Day, T., & Huijben, S. (2011). The evolution of drug resistance and the curious orthodoxy of aggressive chemotherapy. Proc Natl Acad Sci U S A, 108 Suppl 2 (Suppl 2), 10871-10877. https://doi.org/10.1073/pnas.1100299108

Rieber, N., Hector, A., Carevic, M., & Hartl, D. (2014). Current concepts of immune dysregulation in cystic fibrosis. *International Journal of Biochemistry and Cell Biology*, *52*, 108–112. https://doi.org/10.1016/j.biocel.2014.01.017

Rigauts, C., Aizawa, J., Taylor, S. L., Rogers, G. B., Govaerts, M., Cos, P., Ostyn, L., Sims, S., Vandeplassche, E., Sze, M., Dondelinger, Y., Vereecke, L., van Acker, H., Simpson, J. L., Burr, L., Willems, A., Tunney, M. M., Cigana, C., Bragonzi, A., … Crabbé, A. (2022). Rothia mucilaginosa is an anti-inflammatory bacterium in the respiratory tract of patients with chronic lung disease. *European Respiratory Journal*, *59*(5). https://doi.org/10.1183/13993003.01293-2021

Roberts, J. M., Henry, L. A., Long, D., & Hartley, J. P. (2008). Cold-water coral reef frameworks, megafaunal communities and evidence for coral carbonate mounds on the Hatton Bank, north east Atlantic. *Facies*, *54*(3), 297–316. https://doi.org/10.1007/s10347-008-0140-x

Rogers, G. B., Carroll, M. P., Serisier, D. J., Hockey, P. M., Jones, G., Kehagia, V., Connett, G. J., & Bruce, K. D. (2006). Use of 16S rRNA gene profiling by terminal restriction fragment length polymorphism analysis to compare bacterial communities in sputum and mouthwash samples from patients with cystic fibrosis. J Clin Microbiol, 44 (7), 2601-2604. https://doi.org/10.1128/jcm.02282-05

Rong, R., Jiang, S., Xu, L., Xiao, G., Xie, Y., Liu, D. J., Li, Q., & Zhan, X. (2021). MB-GAN: Microbiome Simulation via Generative Adversarial Network. *GigaScience*, *10*(2), 1–11. https://doi.org/10.1093/gigascience/giab005

Rubin, B. K. (2010). Mucus, Phlegm, and Sputum in Cystic Fibrosis. *Respiratory Care*, *54*(6), 726–732. https://doi.org/10.4187/002013209790983269

Ruddy, J., Emerson, J., Moss, R., Genatossio, A., McNamara, S., Burns, J. L., Anderson, G., & Rosenfeld, M. (2013). Sputum tobramycin concentrations in cystic fibrosis patients with repeated administration of inhaled tobramycin. J Aerosol Med Pulm Drug Deliv, 26 (2), 69-75. https://doi.org/10.1089/jamp.2011.0942

Russell, S. L., Gold, M. J., Willing, B. P., Thorson, L., McNagny, K. M., & Finlay, B. B. (2013). Perinatal antibiotic treatment affects murine microbiota, immune responses and allergic asthma. *Gut Microbes*, *4*(2), 158–164. https://doi.org/10.4161/gmic.23567

Shade, A., & Stopnisek, N. (2019). Abundance-occupancy distributions to prioritize plant core microbiome membership. *Current Opinion in Microbiology*, *49*, 50–58. https://doi.org/10.1016/j.mib.2019.09.008

Siddiqui, A. R. & Bernstein, J. M. (2010). Chronic wound infection: facts and controversies. Clinics in dermatology, 28, 519-526. https://doi.org/10.1016/j.clindermatol.2010.03.009

Sloan, W. T., Lunn, M., Woodcock, S., Head, I. M., Nee, S., & Curtis, T. P. (2006). Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental Microbiology*, *8*(4), 732–740. https://doi.org/10.1111/j.1462-2920.2005.00956.x

Smith, D. J., Badrick, A. C., Zakrzewski, M., Krause, L., Bell, S. C., Anderson, G. J., & Reid, D. W. (2014). Pyrosequencing reveals transient cystic fibrosis lung microbiome changes with intravenous antibiotics. Eur Respir J, 44 (4), 922-930. https://doi.org/10.1183/09031936.00203013

Sosinski, L. M., H, C. M., Neugebauer, K. A., Ghuneim, L. A. J., Guzior, D. v., Castillo-Bahena, A., Mielke, J., Thomas, R., McClelland, M., Conrad, D., & Quinn, R. A. (2021). A restructuring of microbiome niche space is associated with Elexacaftor-Tezacaftor-Ivacaftor therapy in the cystic fibrosis lung. *Journal of Cystic Fibrosis*, *xxxx*. https://doi.org/10.1016/j.jcf.2021.11.003

Somayaji, R., & Parkins, M. D. (2015). Tobramycin inhalation powder: an efficient and efficacious therapy for the treatment of Pseudomonas aeruginosa infection in cystic fibrosis. Ther Deliv, 6 (2), 121-137. https://doi.org/10.4155/tde.14.94

Somayaji, R., Parkins, M. D., Shah, A., Martiniano, S. L., Tunney, M. M., Kahle, J. S., Waters, V. J., Elborn, J. S., Bell, S. C., Flume, P. A., & VanDevanter, D. R. (2019). Antimicrobial susceptibility testing (AST) and associated clinical outcomes in individuals with cystic fibrosis: A systematic review. J Cyst Fibros, 18 (2), 236-243. https://doi.org/10.1016/j.jcf.2019.01.008

Stacy, A., McNally, L., Darch, S. E., Brown, S. P., & Whiteley, M. (2016). The biogeography of polymicrobial infection. *Nature Reviews Microbiology*, *14*(2), 93–105. https://doi.org/10.1038/nrmicro.2015.8

Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R., & Schmidt, T. M. (2015). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res, 43 (Database issue), D593-598. https://doi.org/10.1093/nar/gku1201

Stokholm, J., Thorsen, J., Blaser, M. J., Rasmussen, M. A., Hjelmsø, M., Shah, S., Christensen, E. D., Chawes, B. L., Bønnelykke, K., Brix, S., Mortensen, M. S., Brejnrod, A.,

Vestergaard, G., Trivedi, U., Sørensen, S. J., & Bisgaard, H. (2020). *Delivery mode and gut microbial changes correlate with an increased risk of childhood asthma*. *72*.

Stressmann, F. A., Rogers, G. B., Klem, E. R., Lilley, A. K., Donaldson, S. H., Daniels, T. W., Carroll, M. P., Patel, N., Forbes, B., Boucher, R. C., Wolfgang, M. C., & Bruce, K. D. (2011). Analysis of the bacterial communities present in lungs of patients with cystic fibrosis from American and British centers. *Journal of Clinical Microbiology*, *49*(1), 281–291. https://doi.org/10.1128/JCM.01650-10

Stressmann, F. A., Rogers, G. B., van der Gast, C. J., Marsh, P., Vermeer, L. S., Carroll, M. P., Hoffman, L., Daniels, T. W. V., Patel, N., Forbes, B., & Bruce, K. D. (2012). Long-term cultivation-independent microbial diversity analysis demonstrates that bacterial communities infecting the adult cystic fibrosis lung show stability and resilience. *Thorax*, *67*(10), 867–873. https://doi.org/10.1136/thoraxjnl-2011-200932

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M. A., Benezra, A., Destefano, J., Meier, M. F., Muegge, B. D., Barratt, M. J., VanArendonk, L. G., Zhang, Q., Province, M. A., Petri, W. A., Ahmed, T., & Gordon, J. I. (2014). Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature*, *510*(7505), 417–421. https://doi.org/10.1038/nature13421

Surette, M. G. (2014). The cystic fibrosis lung microbiome. Ann Am Thorac Soc, 11 Suppl 1, S61-65. https://doi.org/10.1513/AnnalsATS.201306-159MG

Taccetti, G., Bianchini, E., Cariani, L., Buzzetti, R., Costantini, D., Trevisan, F., Zavataro, L., Campana, S., & Italian Group for, P. a. E. i. C. F. (2012). Early antibiotic treatment for Pseudomonas aeruginosa eradication in patients with cystic fibrosis: a randomised multicentre study comparing two different protocols. Thorax, 67 (10), 853-859. https://doi.org/10.1136/thoraxjnl-2011-200832

Tacking, R. (1954). Penicillinase-producing bacteria in mixed infections in rabbits treated with penicillin. 2. Studies on penicillinase-producing coli- and pyocyaneus-bacteria. Acta Pathol Microbiol Scand, 35 (5), 445-454.

Tunney, M. M., Field, T. R., Moriarty, T. F., Patrick, S., Doering, G., Muhlebach, M. S., Wolfgang, M. C., Boucher, R., Gilpin, D. F., McDowell, A., & Elborn, J. S. (2008). Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. Am J Respir Crit Care Med, 177 (9), 995-1001. https://doi.org/10.1164/rccm.200708-1151OC

Turner, K. H., Wessel, A. K., Palmer, G. C., Murray, J. L., & Whiteley, M. (2015). Essential genome of Pseudomonas aeruginosa in cystic fibrosis sputum. Proc Natl Acad Sci U S A, 112 (13), 4110-4115. https://doi.org/10.1073/pnas.1419677112

Van den Boogaart, K., Tolosana, R., & Bren, M. (2020). compositions: Compositional Data Analysis. R package version 2.0-0. R Foundation for Statistical Computing., Vienna. https://CRAN.R-project.org/package=compositions

van der Gast, C. J., Walker, A. W., Stressmann, F. A., Rogers, G. B., Scott, P., Daniels, T. W., Carroll, M. P., Parkhill, J., & Bruce, K. D. (2011). Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *ISME Journal*, *5*(5), 780–791. https://doi.org/10.1038/ismej.2010.175

van der Loo, M. P. (2014). The stringdist package for approximate string matching. https://CRAN.R-project.org/package=stringdist. . The R Journal, 6 (1), 111-122.

Vandeplassche, E., Coenye, T., & Crabbé, A. (2017). Developing selective media for quantification of multispecies biofilms following antibiotic treatment. PLoS One, 12 (11), e0187540. https://doi.org/10.1371/journal.pone.0187540

Vandeplassche, E., Sass, A., Ostyn, L., Burmølle, M., Kragh, K. N., Bjarnsholt, T., Coenye, T., & Crabbé, A. (2020). Antibiotic susceptibility of cystic fibrosis lung microbiome members in a multispecies biofilm. Biofilm, 2, 100031. https://doi.org/10.1016/j.bioflm.2020.100031

Varga, J. J., Zhao, C., Davis, J. D., Hao, Y., Farrell, J. M., Gurney, J. R., Voit, E., & Brown, S. P. (2021). Antibiotics drive expansion of rare pathogens in a chronic infection microbiome model. *BioRxiv*, 2021.06.21.449018.

Venkataraman, A., Bassis, C. M., Beck, J. M., Young, V. B., Curtis, J. L., Huffnagle, G. B., & Schmidt, T. M. (2015). Application of a neutral community model to assess structuring of the human lung microbiome. *MBio*, *6*(1). https://doi.org/10.1128/mBio.02284-14

Voronina, O. L., Ryzhova, N. N., Kunda, M. S., Loseva, E. v., Aksenova, E. I., Amelina, E. L., Shumkova, G. L., Simonova, O. I., & Gintsburg, A. L. (2020). Characteristics of the Airway Microbiome of Cystic Fibrosis Patients. *Biochemistry (Moscow)*, *85*(1), 1–10. https://doi.org/10.1134/S0006297920010010

Wale, N., Sim, D. G., Jones, M. J., Salathe, R., Day, T., & Read, A. F. (2017). Resource limitation prevents the emergence of drug resistance by intensifying within-host competition. Proc Natl Acad Sci U S A, 114 (52), 13774-13779. https://doi.org/10.1073/pnas.1715874115

Wang, J., Chai, J., Zhang, L., Zhang, L., Yan, W., Sun, L., Chen, Y., Sun, Y., Zhao, J., & Chang, C. (2022). Microbiota associations with inflammatory pathways in asthma. *Clinical and Experimental Allergy*, *52*(5), 697–705. https://doi.org/10.1111/cea.14089

Waters, V. (2018). Chronic Antibiotic Use in Cystic Fibrosis: A Fine Balance. Ann Am Thorac Soc, 15 (6), 667-668. https://doi.org/10.1513/AnnalsATS.201803-172ED

Waters, V. J., Kidd, T. J., Canton, R., Ekkelenkamp, M. B., Johansen, H. K., LiPuma, J. J., Bell, S. C., Elborn, J. S., Flume, P. A., VanDevanter, D. R., & Gilligan, P. (2019). Reconciling Antimicrobial Susceptibility Testing and Clinical Response in Antimicrobial Treatment of Chronic Cystic Fibrosis Lung Infections. Clin Infect Dis, 69 (10), 1812-1816. https://doi.org/10.1093/cid/ciz364

Weitz, J. S., Beckett, S. J., Coenen, A. R., Demory, D., Dominguez-Mirazo, M., Dushoff, J., Leung, C. Y., Li, G., Măgălie, A., Park, S. W., Rodriguez-Gonzalez, R., Shivam, S., & Zhao, C. Y. (2020). Modeling shield immunity to reduce COVID-19 epidemic spread. *Nature Medicine*, *26*(6), 849–854. https://doi.org/10.1038/s41591-020-0895-3

Wenzler, E., Gotfried, M. H., Loutit, J. S., Durso, S., Griffith, D. C., Dudley, M. N., & Rodvold, K. A. (2015). Meropenem-RPX7009 Concentrations in Plasma, Epithelial Lining Fluid, and Alveolar Macrophages of Healthy Adult Subjects. Antimicrob Agents Chemother, 59 (12), 7232-7239. https://doi.org/10.1128/aac.01713-15

Whelan, F. J., Waddell, B., Syed, S. A., Shekarriz, S., Rabin, H. R., Parkins, M. D., & Surette, M. G. (2020). Culture-enriched metagenomic sequencing enables in-depth profiling of the cystic fibrosis lung microbiota. *Nature Microbiology*, *5*(February), 1–12. https://doi.org/10.1038/s41564-019-0643-y

Wickham, H. (2016). ggplot2-Elegant Graphics for Data Analysis. Springer International Publishing. Cham, Switzerland.

Widder, S., Zhao, J., Carmody, L. A., Zhang, Q., Kalikin, L. M., Schloss, P. D., & LiPuma, J. J. (2022). Association of bacterial community types, functional microbial processes and lung disease in cystic fibrosis airways. *ISME Journal*, *16*(4), 905–914. https://doi.org/10.1038/s41396-021-01129-z

Wong, K., Roberts, M. C., Owens, L., Fife, M., & Smith, A. L. (1984). Selective media for the quantitation of bacteria in cystic fibrosis sputum. J Med Microbiol, 17 (2), 113-119. https://doi.org/10.1099/00222615-17-2-113

Worlitzsch, D., Tarran, R., Ulrich, M., Schwab, U., Cekici, A., Meyer, K. C., Birrer, P., Bellon, G., Berger, J., Weiss, T., Botzenhart, K., Yankaskas, J. R., Randell, S., Boucher, R. C., & Döring, G. (2002). Effects of reduced mucus oxygen concentration in airway Pseudomonas infections of cystic fibrosis patients. J Clin Invest, 109 (3), 317-325. https://doi.org/10.1172/jci13870

Wu, Y., Klapper, I., & Stewart, P. S. (2018). Hypoxia arising from concerted oxygen consumption by neutrophils and microorganisms in biofilms. Pathog Dis, 76 (4). Yonker, L. M., Cigana, C., Hurley, B. P., & Bragonzi, A. (2015). Host-pathogen interplay in the respiratory environment of cystic fibrosis. *Journal of Cystic Fibrosis*, *14*(4), 431–439. https://doi.org/10.1016/j.jcf.2015.02.008

Young, D., Hussel, T., & Dougan, G. (2002). Chronic bacterial infections: living with unwanted guests. Nature immunology, 3, 1026-1032. https://doi.org/10.1038/ni1102-1026.

Zemanick, E. T., Harris, J. K., Wagner, B. D., Robertson, C. E., Sagel, S. D., Stevens, M. J., Accurso, F. J., & Laguna, T. A. (2013). Inflammation and airway microbiota during cystic fibrosis pulmonary exacerbations. PLoS One, 8 (4), e62917. https://doi.org/10.1371/journal.pone.0062917

Zemanick, E. T., Wagner, B. D., Robertson, C. E., Stevens, M. J., Szefler, S. J., Accurso, F. J., Sagel, S. D., & Harris, J. K. (2015). Assessment of airway microbiota and inflammation in cystic fibrosis using multiple sampling methods. Ann Am Thorac Soc, 12 (2), 221-229. https://doi.org/10.1513/AnnalsATS.201407-310OC

Zhao, C. Y., Hao, Y., Wang, Y., Varga, J. J., Stecenko, A. A., Goldberg, J. B., & Brown, S. P. (2021). Microbiome Data Enhances Predictive Models of Lung Function in People With Cystic Fibrosis. *The Journal of Infectious Diseases*, *223*(Supplement_3), S246–S256. https://doi.org/10.1093/infdis/jiaa655

Zhao, J., Schloss, P. D., Kalikin, L. M., Carmody, L. A., Foster, B. K., Petrosino, J. F., Cavalcoli, J. D., VanDevanter, D. R., Murray, S., Li, J. Z., Young, V. B., & LiPuma, J. J. (2012a). Decade-long bacterial community dynamics in cystic fibrosis airways. *Proceedings of the National Academy of Sciences*, *109*(15), 5809–5814. https://doi.org/10.1073/pnas.1120577109