Microbiome data enhances predictive models of lung function in people with cystic fibrosis

Conan Y. Zhao[1,2,3], Yiqi Hao[2], Yifei Wang[2,3,4], John J. Varga[2,3,5,6], Arlene A. Stecenko[5,6], Joanna B. Goldberg[5,6], Sam P. Brown[1,2,3*]

Institutions:

[1]Interdisciplinary Graduate Program in Quantitative Biosciences, Georgia Institute of Technology, Atlanta, GA, USA.

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

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

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

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

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

*Corresponding author

E-mails: sam.brown@biology.gatech.edu

Abstract

Background. 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).

Methods. 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.

Results. 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.

Conclusions. These results support a reconsideration of clinical microbiology pipelines to ensure the provision of informative data to guide clinical practice.

Keywords. Microbiome; machine learning; cystic fibrosis

35 **Introduction**

36 Bacterial infections often resolve rapidly given effective immune responses, independent of antibiotic
37 treatment. However, in chronic (long-lasting) cases, infections fail to clear even with appropriate drug
38 treatment. Chronic infections impose an elevated morbidity and mortality risk to the individual [1] and an
39 increasing burden on global healthcare systems as at-risk populations grow [2]. Chronic infections typically arise
40 due to deficits in host barrier defenses and/or immune function, and commonly feature changes in pathogen
41 growth mode (e.g. biofilm formation [3]) and additional microbial species acquisition, forming complex
42 multispecies communities [4].

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

48 Cystic fibrosis is an autosomal recessive disease characterized by decreased lung mucociliary clearance
49 and mucus accumulation [5–7]. The resulting environment provides both nutrients for bacterial growth and
50 protection from host immune responses [8–11], facilitating chronic microbial infections [12–15]. Accessible 16S
51 rDNA microbiome profiling has shifted CF airway microbiology research away from a historically single-pathogen
52 focus, as sequencing expectorated sputum has revealed diverse communities of tens to hundreds of taxa,
53 including numerous non-pathogenic bacteria [13,16,17].

54 Numerous lung microbiome studies have linked community composition to disease progression and
55 overall patient health [18–20]. Cross-sectional studies have shown severe disease is associated with pathogen
56 dominance and loss of taxonomic diversity [18,19,21]. Longitudinal studies have associated decreasing
57 microbiome diversity with declining lung function [22]. Additionally, abundance of non-pathogenic fermentative
58 anaerobes (*Veillonella*, *Prevotella*, *Fusobacterium*) is associated with higher lung function [23,24]. While these

59 associations are observed across multiple studies, their causal interpretation is the subject of some controversy.
60 These results may reflect community ecological processes within the lung, where species interactions govern
61 community structure and subsequent harm to the host [12,25,26]. Conversely, these patterns could result from
62 oral anaerobe contamination during sample collection [27,28]. Under this contamination model, increasing
63 pathogen load compared to a constant background of oral microbiome contamination generates a spurious link
64 between oral microbes, microbiome diversity, and patient health [27]. While recent paired sputum-saliva
65 sampling analysis indicates that oral sample contamination is not a substantial contributor to sputum
66 microbiome profiles in people with established CF lung disease [29], these conflicting hypotheses highlight the
67 uncertainty in the role specific taxa present in sputum.
68 In the current study, we side-step this causal inference problem and instead assess how informative
69 expectorated sputum microbiome data (including potential oral contaminants) is of patient lung health using a
70 machine-learning framework. We hypothesize that the addition of non-pathogen data improves the prediction
71 of patient lung function, compared to established pathogen data alone. To address this hypothesis we train
72 predictive models on both lung microbiome and electronic medical record data for a cohort of CF patients. We
73 find that compared to the benchmark of pathogen data alone, model performance was consistently improved by
74 the addition of non-pathogen taxa.
75
76 **Results**
77 *Clinical and microbiome data summary*
78 In total, we obtained sputum expectorates from 77 CF children and adults. Pulmonary function,
79 measured by percent predicted forced expiratory volume in 1 second (ppFEV1), was stratified into four
80 categories from Severe to Normal. A summary of patient information is presented in Table 1. As expected,
81 increasing age correlated with worsening lung function (ANOVA; $p < 0.01$). Culture-based detection of

82 *Pseudomonas aeruginosa* correlated with decreasing lung function (ANOVA; $p<0.001$), as did (log-scaled)
83 bacterial load ($p < 0.05$).
84 The majority (>90%) of reads from our sequencing analysis mapped to one of 13 genera (Figure 1a),
85 consisting of both recognized CF pathogens (red) and orally derived bacteria (black). *Pseudomonas* sequences
86 accounted for 30.4% of all reads, and were detected in every patient sample. Other established CF pathogens
87 (*Staphylococcus*, *Achromobacter*, *Haemophilus*, *and Burkholderia*) collectively represented 19.3%, while oral
88 taxa account for over 45% (Figure 1). Total pathogenic and non-pathogenic taxa abundance were both found to
89 vary significantly ($p << 0.001$) with lung function (Table 1).
90
91 *Microbiome Composition Varies with Lung Function*
92 We analyzed microbiome compositions across broad lung function categories to examine the
93 relationship between sputum taxonomic profile and patient health. Figure 2a highlights the relative
94 compositions of six canonical CF pathogens. As expected, *Pseudomonas* was more prevalent in lungs with
95 reduced function, whereas in normal lungs *Haemophilus* and non-pathogen taxa (gray) were more prevalent.
96 The non-pathogenic composition is consistently dominated by *Veillonella* and *Streptococcus* regardless of lung
97 health or pathogen status (Figure 2b). Shannon diversity calculated with all taxa present is significantly greater
98 for normal lung function ($p<0.01$, Figure S1a), in line with multiple other studies [30,31]. While Principle
99 Coordinate Analysis (PCA) did not qualitatively separate compositions by lung function category, we found
100 ppFEV1 was significantly associated with microbiome composition (Mantel test, $r=0.195$, $p<0.001$, Figure S1b).
101
102 *Integrating microbiome and patient meta-data*
103 To examine multiple confounding variables such as patient age, BMI or CF-related diabetes (CFRD), we
104 calculated spearman correlations across 14 microbiome, 11 patient metadata, and 6 clinical microbiology
105 features (Figure 3). Hierarchical clustering reveals a complex autocorrelation structure, but with many expected

106 consistencies. Overall, there are two main clusters of correlated variables. One correlated with ppFEV1, and
107 included Shannon diversity index as well as 16S quantitation of *Fusobacterium*, *Haemophilus*, and *Neisseria*. The
108 other anticorrelated with ppFEV1, and included ppFEV1 decline, pathogen abundance, CFRD and 16S
109 quantitation of *Pseudomonas* and *Achromobacter*. Unsurprisingly, FEV1 and ppFEV1 cluster together and are
110 inversely correlated with ppFEV1 decline rate (an average per year loss in ppFEV1 since birth). Additionally, 16S
111 quantitation results for *Pseudomonas*, *Staphylococcus*, *Burkholderia*, and *Achromobacter* cluster with their
112 respective culture-based clinical microbiology results. This does not hold for *Stenotrophomonas,* potentially due
113 to its infrequent detection.
114
115 *Dimensionality Reduction*
116 The hairball correlation matrix in Figure 3 highlights the statistical challenges for identifying meaningful
117 lung function predictors. Such challenges include high between-feature correlations and relatively few
118 independent patient observations (N=77) compared to the initial number of available predictors (86 total,
119 including 59 bacterial taxa). To mitigate this dimensionality problem, we first restrict our microbiome analysis to
120 only the top 23 genera in our dataset. These top 23 encompass 97% of the total sequenced reads (Figure 1). We
121 also calculate three additional summary statistics: % pathogen, % oral taxa, and Shannon diversity. As our
122 clustering analysis shows reasonable agreement between clinical microbiology detection and rDNA sequencing,
123 we exclude the binary detection results in favor of 16S quantitation. To address compositionality of 16S data, we
124 incorporate total bacterial load (universal 16S primer qPCR) as a predictor. In addition, we use a centered log-
125 ratio (clr) transform on our genus-level relative abundance data before standardizing to mean zero, unit
126 variance inputs. We refer to this final combination of metadata and 16S data as our “All Features” dataset.
127
128 *Training Machine Learning Models*

129 To assess if non-pathogenic taxa contain informative biomarkers, we split our samples into 53 training
130 and 24 testing samples. ElasticNet was used to train predict lung function while performing feature selection
131 (see methods, Figure 4). We expect that the addition of patient metadata (age, BMI etc) will improve our ability
132 to predict lung function given the progressive nature of CF. Our null hypothesis, following the work of Jorth et al.
133 and others[27,28] is that the taxa targeted by clinical microbiology provide adequate explanatory basis for lung
134 function outcomes, and that the addition of non-pathogen 16S data will not improve model predictions.
135 We test this hypothesis by generating four additional feature subsets (CF Pathogens, All 16S Data,
136 Metadata, and Metadata + Pathogens) and comparing the performance of models trained on each datasets.
137 Initial-pass, non-bootstrapped model training results are shown in Figures S2 and S3.
138
139 *Model Generalizability*
140 We assess overfitting using leave-one-out cross-validation and compare the prediction error across folds
141 against the test set error. For model robustness, we use 1000-fold bootstrap resampling to fit both a baseline
142 and ensemble of models. Robust features selected by the baseline model will also be selected by a large portion
143 of the bootstrapped ensemble. We additionally standardize all features (mean=0, S.D.=1) to allow for cross-
144 feature comparability. As an additional point of comparison, we generate a non-informative control dataset
145 from the All Features set using within-feature shuffling, scrambling between-feature correlations while
146 preserving the mean zero, unit variance within-feature structure. Figure 5 shows the results of our baseline
147 (black points) and ensemble (boxplots) approaches. All models using patient metadata or microbiome data
148 outperform the negative control.
149
150 *Addition of non-pathogen data improves model performance*
151 To address the key question of relative model performance, we find that the addition of non-pathogen
152 taxa significantly improves performance (significantly reduces bootstrapped MSE; Figure 5a), with or without the

153 addition of patient meta-data. Models trained on all 16S quantitation significantly outperform models trained
154 only on pathogen quantitation. Interestingly, while microbiome-only and metadata-only models achieve
155 comparable performance, the combined model achieves greater model performance. Looking broadly across
156 models, we find reasonable consistency in positive and negative predictor selection between our baseline and
157 bootstrapped models (Figure S4).

158 We find multiple features selected across all training sets. *Pseudomonas*, *Achromobacter*, age, and
159 diabetic status are consistently selected as negative predictors, while *Haemophilus*, *Fusobacterium*, *Rothia*, oral
160 taxa abundance, and *BMI* are consistently positive predictors. All informative features selected in the
161 independent models (Figure S4c) were also selected in the All Features model (Figure S4g). A small subset (<
162 50%) of the bootstrapped models also selected a handful of oral taxa, bacterial load, and CFTR mutation type as
163 positive predictors of lung function (Figure 5c, gray boxplots). However, a majority of bootstrapped models and
164 the train/test model did not select these as informative features.

165 As an additional check against overfitting, we obtain ranges of model errors (measured by mean squared
166 error of predicted ppFEV1 values) using leave-one-out cross validation (Figure 5b). We do not find significant
167 differences between cross-validated model errors across our training sets, suggesting that despite the difference
168 in number of available predictors, our models are not overfitting.

169

170 **Discussion**

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

177 to an integrated 77 patient lung microbiome and electronic medical record dataset. Our analysis revealed that
178 the addition of non-pathogen data improves prediction of patient health, with the most accurate models
179 selecting patient metadata, pathogen quantitation, and non-pathogen information. Our inclusive 'all data'
180 models additionally point to a predictive role for specific non-pathogen taxa, in particular the oral anaerobe
181 genera *Rothia* and *Fusobacterium*.

182 Despite the significant contribution of non-pathogen data, our results are still broadly consistent with
183 what might be termed the 'traditional' view of CF microbiology. Established CF pathogens (*P. aeruginosa, S.*
184 *aureus, H. influenzae, B. cenocepacia*) are the major drivers of patient outcomes, as evidenced by substantial
185 improvement in predictive outcomes whenever we include pathogen data (Figure 5a), and the by comparison
186 relatively weak contribution of the addition of non-pathogen taxa. Note that we specifically use quantitative 16S
187 measures of pathogen composition to provide a level playing field in the comparison of pathogen and non-
188 pathogen predictive contribution. Figure 3 highlights that quantitative 16S and qualitative (presence/absence)
189 clinical microbiology data are in general agreement.

190 The traditional role of CF pathogens as the central predictors of patient outcomes has been challenged
191 over the past decade by the advent of microbiome sequencing. Extensive surveys have documented an
192 association between CF lung function and microbiome diversity, also evident in the current study (Figure 2). At
193 face value, these results suggest a biological role for these non-pathogen taxa, potentially competing with [32]
194 or facilitating [33] pathogen taxa and therefore indirectly shaping disease outcomes. Jorth et al. recently
195 published a forceful rejection of this 'active microbiome' view, stressing the potential causal role of changing
196 pathogen densities in shaping disease outcomes and viewing shifting diversity metrics as a simple statistical
197 'relative composition' artifact of shifting pathogen numbers against a roughly constant oral contamination
198 background [27]. While our analyses provide some support for this view, in particular the constancy of the non-
199 pathogen microbiome across patients (Figure 2b) and the lack of substantial predictive improvement on addition
200 of non-pathogen data (Figure 5b), we also see lines of evidence against the contamination hypothesis. First, our

201 use of center log transformations mitigates the risk of spurious associations due to compositionality (refs) and
202 yet non-pathogen taxa are still consistently retained. Second, the contamination hypothesis predicts total
203 bacterial burden to be an important predictor, and yet burden was not retained in our models. Third, our
204 observation of a consistent retention of specific non-pathogen taxa across multiple models (with and without
205 the addition of potentially confounding EMR features, including age and BMI) points to the potential for a
206 distinct causal pathway that is orthologous to age or BMI. We note that the interpretation that oral bacteria are
207 active players in the lung environment is further buttressed by a recent study on people with established CF
208 disease (Lu et al. 2020) that used paired sputum and saliva samples to infer the presence of substantial
209 populations of oral bacteria in the lung.

210 Our 'all data' models highlight *Rothia* and *Fusobacterium* as positive predictors of lung function across
211 our 77 patients, in models that already take into account pathogen data. When we include features already
212 known to correlate with lung health, such as age, BMI, and CFRD status, our models not only these features, but
213 additionally retain *Rothia* and *Fusobacterium* as positive predictors. The retention of these specific taxa in both
214 this full model and in partial models (Figure S4) suggests that these taxa provide potentially valuable predictive
215 information on current patient health. Of course, this analysis does not allow inference to causal mechanism or
216 even direction of causality. It is entirely possible that these taxa are simply bio-markers of dimensions of
217 improved health that are largely independent of age, BMI, and other established positive predictors that are
218 already accounted for in the model. It is also possible that these specific taxa play a more active causal role, for
219 instance holding specific pathogens at bay via competitive interspecific mechanisms [34].

220

221 Interestingly, our All Features models also highlight *Haemophilus*, a canonical CF pathogen, as a positive
222 predictor of lung function. *Haemophilus influenzae* infections are most common in younger CF patients [8,35],
223 hence we would expect a positive association in a model that is not controlled for age (Figure S4c, S4d).
224 However we see that the positive weighting on *Haemophilus* is retained in models that also account for age as a

225 positive predictor of lung function. A second possibility is that the positive weighting of *Haemophilus* is due to
226 pathogen-pathogen competition and the relatively less severe nature of *Haemophilus* infections in adults (i.e.,
227 *Haemophilus* is 'best of a bad job'). Figure 2a illustrates that we only appreciably detect two and rarely three
228 coexisting pathogens of the six we find across all patients. The relative scarcity of multi-pathogen communities
229 implies that *Haemophilus* presence coincides with the absence of other more severe pathogens – and indeed we
230 see a dominance of negative correlations among pathogens (Figure 3). In this context we cannot preclude a
231 protective role of *Haemophilus* against more severe pathogens in older patients.

232 A caveat of this analysis is the dependency of machine learning performance and robustness on
233 particular distributions of data, and the failure of linear algorithms such as LASSO and ElasticNet on microbiome-
234 like data [36–38]. This is in part due to the compositionality constraint of microbiome data, which can be
235 mitigated by using absolute quantitation [39]. However, training on absolute abundances introduces additional
236 caveats, as order-of-magnitude differences in qPCR sample quantitation can in turn over-represent samples with
237 higher bacterial loads. We address these issues by using a centered-log transform on relative abundance data
238 and including log-scaled bacterial load as a potential feature to select. While some bootstrapped models
239 selected bacterial load as a positive predictor (Figure 5c), the majority of models did not. This further suggests
240 that the majority of microbiome information is encoded in the relative ratios of taxa abundance, which is
241 broadly consistent with previous findings [27,28].

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

248 We note that the major predictors identified in our models have been identified in various studies, and
249 taken piecemeal there is less insight. The value of this work lies in the systematic integration of these multiple
250 data sources (from both EMR and microbiome data sources). Our model comparisons (with / without EMR
251 predictors) allow an assessment of the impact of oral bacteria, with and without key potential confounds.
252 Ignoring these confounds could lead to spurious retention of microbiome taxa that correlate strongly with e.g.
253 age or BMI. In addition, our analyses allows assessment of disparate factors on a common predictive scale –
254 indicating for example that the impact of 1 standard deviation shift in *Fusobacterium* abundance is comparable
255 to a 1 S.D. shift in BMI. Our model comparison approach lends more confidence to the conclusion that the
256 retained oral taxa are associated with patient outcomes via causal pathways that are largely independent of age
257 or BMI, being robust to their presence or absence in the predictive models. The research agenda of pursuing the
258 nature of the causal pathways linking oral bacteria in the lung with patient outcomes is now on a firmer footing
259 as a result of our study.
260 In summary, our study finds that inclusion of non-pathogenic taxa significantly improves model
261 prediction accuracy of patient health status. We identify two oral-derived taxa (*Fusobacterium, Rothia*) that are
262 independently informative of lung function, which may be either biomarkers or potential probiotics. Our results
263 call attention to the potential predictive utility of oral microbes (regardless of their functional roles) in the
264 clinical assessment of CF patient health.
265
266
267 **Methods**
268 *Subjects*
269 All procedures performed in studies involving human participants were in accordance with the ethical
270 standards of the institutional and national research committees. Authorization was obtained from each patient
271 enrolled according to the protocol approved by the Emory University Institutional Review Board (IRB00042577).

272

273 *Sample collection and 16S analysis*

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

279 All patients were clinically stable, defined as having no increase in respiratory symptoms compared to
280 baseline, and no acute illness or new medication for three weeks prior to sputum collection. Upon collection,
281 sputum samples were diluted 1:3 (mass:volume) with PBS supplemented with 50 mM EDTA. Diluted samples
282 were then homogenized by being repeatedly drawn through a syringe and 18-gauge needle. The resulting
283 sputum homogenates were aliquot and stored at -80 °C until all 77 samples were collected. Microbiology culture
284 results were obtained for sputum samples sent to the Clinical Microbiology laboratory on the same day as
285 samples for sequencing were collected.

286 DNA was purified from sputum homogenate with the MoBio Power Soil kit (MoBio, Carlsbad, CA). The
287 16S V4 region was amplified and sequenced using Illumina MiSeq, yielding an average of 137,708 sequences per
288 sample. Sequences were quality filtered and amplicon sequence variants were obtained using the QIIME2 deblur
289 plugin. Taxonomic assignments were classified against both SILVA and Greengenes 16S reference databases and
290 assigned based on highest taxonomic resolution. To mitigate compositional effects, 16S data were center-log
291 transformed prior to all analyses. Nucleotides are uploaded to BioProject accession no. PRJNA666192.

292

293 *Statistical and Quantitative Analysis*

294 Patient samples were binned by ppFEV1 (Normal: >80%, Mild: 80-60%, Moderate: 60-40%, Severe:
295 <40%). Variance across lung function categories in patient metadata and 16S metadata was tested using ANOVA.

296 Variation between microbiome composition and ppFEV1 was tested using Mantel tests on Bray-Curtis distances
297 at 9999 permutations. Within-sample and among-sample diversity was calculated using the Shannon diversity
298 index and Bray-Curtis based PCoA on 16S quantitation agglomerated to the genus level [40]. Associations
299 between continuous variables were tested using Spearman correlations. A full pairwise correlation matrix was
300 calculated, with rows and columns ordered by hierarchical clustering [41].
301
302 *Machine Learning*
303 We use ElasticNet to fit regularized linear models predicting lung function (ppFEV1) from patient
304 metadata, microbiome composition, and clinical microbiology results [42]. ElasticNet solves a penalized linear
305 regression model using a weighted average of L1 (LASSO) and L2 (ridge regression) penalties. This limits over-
306 fitting by penalizing non-zero coefficients. We split our samples using a simple 70:30 train-test holdout, where
307 models are trained on 53 samples and used to predict on the remaining 24. All input features were standardized
308 (mean=0, S.D.=1) prior to model training to allow between-feature interpretability. From our full dataset, we
309 create 4 additional data subsets: CF Pathogens, All 16S Data, Metadata, and Metadata + Pathogens. We include
310 within-feature shuffling on the full set as a non-informative negative control.
311 We employ two methods to assess model robustness, and compare model performance using mean
312 squared error (MSE). We generate 1000 bootstrap resampled sets from the training set and fit an ensemble of
313 regularized linear models to obtain distributions for each model coefficient. We identify key metadata and taxa
314 robustly selected (nonzero coefficients) across the ensemble of models. We assess model generalizability using
315 leave-one-out cross-validation on the training set and compare resulting MSE ranges.
316
317 **Acknowledgements.** We thank Karan Kapuria and Eunbi Park for help with bio-informatic pipeline development,
318 and Peng Qiu for machine learning guidance. We acknowledge the technical support and provision of the clinical
319 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.

Disclaimer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Financial Support. This work was funded by the CDC (BAA 2016-N-17812, BAA 2017-OADS-01), NIH (HR56L142857, R21AI143296) and the Cystic Fibrosis Foundation (BROWN19I0, BROWN19P0), as well as the Emory and Children's Center for Cystic Fibrosis and Airways Disease Research for a pilot award to Drs. Stecenko and Goldberg.

Potential Conflicts of Interest. The authors declare that there is no conflict of interest.

330 **References**

331 1. Persoon A, Heinen MM, Vleuten CJM van der, Rooij MJ de, Kerkhof PCM van de, Achterberg T van. Leg
332 ulcers: a review of their impact on daily life. J Clin Nurs [Internet]. **2004**; 13(3):341–54. Available from:
333 http://www.ncbi.nlm.nih.gov/pubmed/15009337
334 2. Guest JF, Ayoub N, McIlwraith T, et al. Health economic burden that different wound types impose on
335 the UK's National Health Service. Int Wound J. **2017**; 14(2):322–330.
336 3. Malone M, Bjarnsholt T, McBain AJ, et al. The prevalence of biofilms in chronic wounds: a systematic
337 review and meta-analysis of published data. J Wound Care [Internet]. **2017**; 26(1):20–25. Available from:
338 http://www.magonlinelibrary.com/doi/10.12968/jowc.2017.26.1.20
339 4. Stacy A, McNally L, Darch SE, Brown SP, Whiteley M. The biogeography of polymicrobial infection. Nat
340 Rev Microbiol [Internet]. **2016**; 14(2):93–105. Available from:
341 http://www.nature.com/articles/nrmicro.2015.8
342 5. Perez-Vilar J, Boucher RC. Reevaluating gel-forming mucins' roles in cystic fibrosis lung disease. Free
343 Radic Biol Med. **2004**; 37(10):1564–1577.
344 6. Henke MO, Ratjen F. Mucolytics in cystic fibrosis. Paediatr Respir Rev. **2007**; 8(1):24–29.
345 7. Rubin BK. Mucus, Phlegm, and Sputum in Cystic Fibrosis. Respir Care. **2010**; 54(6):726–732.
346 8. Bals R, Weiner DJ, Wilson JM. The innate immune system in cystic fibrosis lung disease. J Clin Invest.
347 **1999**; 103(3):303–307.
348 9. Dickson RP, Martinez FJ, Huffnagle GB. The role of the microbiome in exacerbations of chronic lung
349 diseases. Lancet. **2014**; 384(9944):691–702.
350 10. Rieber N, Hector A, Carevic M, Hartl D. Current concepts of immune dysregulation in cystic fibrosis. Int J
351 Biochem Cell Biol [Internet]. Elsevier Ltd; **2014**; 52:108–112. Available from:
352 http://dx.doi.org/10.1016/j.biocel.2014.01.017
353 11. Yonker LM, Cigana C, Hurley BP, Bragonzi A. Host-pathogen interplay in the respiratory environment of

354 cystic fibrosis. J Cyst Fibros. **2015**; 14(4):431–439.

355 12. Conrad D, Haynes M, Salamon P, Rainey PB, Youle M, Rohwer F. Cystic fibrosis therapy: A community
356 ecology perspective. Am J Respir Cell Mol Biol. **2013**; 48(2):150–156.

357 13. Frayman KB, Armstrong DS, Grimwood K, Ranganathan SC. The airway microbiota in early cystic fibrosis
358 lung disease. Pediatr Pulmonol. **2017**; 52(11):1384–1404.

359 14. Lucas SK, Yang R, Dunitz JM, Boyer HC, Hunter RC. 16S rRNA gene sequencing reveals site-specific
360 signatures of the upper and lower airways of cystic fibrosis patients. J Cyst Fibros [Internet]. European
361 Cystic Fibrosis Society; **2018**; 17(2):204–212. Available from: https://doi.org/10.1016/j.jcf.2017.08.007

362 15. Fodor AA, Klem ER, Gilpin DF, et al. The Adult Cystic Fibrosis Airway Microbiota Is Stable over Time and
363 Infection Type, and Highly Resilient to Antibiotic Treatment of Exacerbations. PLoS One. **2012**; 7(9).

364 16. Huang YJ, LiPuma JJ. The Microbiome in Cystic Fibrosis. Clin Chest Med [Internet]. **2016**; 37(1):59–67.
365 Available from: https://linkinghub.elsevier.com/retrieve/pii/S0272523115001379

366 17. Whelan FJ, Waddell B, Syed SA, et al. Culture-enriched metagenomic sequencing enables in-depth
367 profiling of the cystic fibrosis lung microbiota. Nat Microbiol [Internet]. Springer US; **2020**; 5(February):1–
368 12. Available from: http://www.nature.com/articles/s41564-019-0643-y

369 18. Coburn B, Wang PW, Diaz Caballero J, et al. Lung microbiota across age and disease stage in cystic
370 fibrosis. Sci Rep [Internet]. Nature Publishing Group; **2015**; 5:1–12. Available from:
371 http://dx.doi.org/10.1038/srep10241

372 19. Acosta N, Heirali A, Somayaji R, et al. Sputum microbiota is predictive of long-term clinical outcomes in
373 young adults with cystic fibrosis. Thorax. **2018**; 73(11):1016–1025.

374 20. Hahn A, Burrell A, Ansusinha E, et al. Airway microbial diversity is decreased in young children with cystic
375 fibrosis compared to healthy controls but improved with CFTR modulation. Heliyon [Internet]. Elsevier
376 Ltd; **2020**; 6(6):e04104. Available from: https://doi.org/10.1016/j.heliyon.2020.e04104

377 21. Muhlebach MS, Zorn BT, Esther CR, et al. Initial acquisition and succession of the cystic fibrosis lung

378 microbiome is associated with disease progression in infants and preschool children. PLoS Pathog. **2018**;
379 14(1):1–20.
380 22. Zhao J, Schloss PD, Kalikin LM, et al. Decade-long bacterial community dynamics in cystic fibrosis airways.
381 Proc Natl Acad Sci. **2012**; 109(15):5809–5814.
382 23. Zemanick ET, Wagner BD, Robertson CE, et al. Assessment of airway microbiota and inflammation in
383 cystic fibrosis using multiple sampling methods. Ann Am Thorac Soc. **2015**; 12(2):221–229.
384 24. O'Neill K, Bradley JM, Johnston E, et al. Reduced bacterial colony count of anaerobic bacteria is
385 associated with a worsening in lung clearance index and inflammation in cystic fibrosis. PLoS One. **2015**;
386 10(5).
387 25. Quinn RA, Whiteson K, Lim YW, et al. Ecological networking of cystic fibrosis lung infections. npj Biofilms
388 Microbiomes [Internet]. Springer US; **2016**; 2(1):4. Available from: http://dx.doi.org/10.1038/s41522-
389 016-0002-1
390 26. Klepac-Ceraj V, Lemon KP, Martin TR, et al. Relationship between cystic fibrosis respiratory tract bacterial
391 communities and age, genotype, antibiotics and Pseudomonas aeruginosa. Environ Microbiol [Internet].
392 **2010**; 12(5):1293–1303. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2010.02173.x
393 27. Jorth P, Ehsan Z, Rezayat A, et al. Direct Lung Sampling Indicates That Established Pathogens Dominate
394 Early Infections in Children with Cystic Fibrosis. Cell Rep [Internet]. ElsevierCompany.; **2019**; 27(4):1190-
395 1204.e3. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2211124719304255
396 28. Goddard AF, Staudinger BJ, Dowd SE, et al. Direct sampling of cystic fibrosis lungs indicates that DNA-
397 based analyses of upper-airway specimens can misrepresent lung microbiota. Proc Natl Acad Sci. **2012**;
398 109(34):13769–13774.
399 29. Lu J, Carmody LA, Opron K, et al. Parallel Analysis of Cystic Fibrosis Sputum and Saliva Reveals
400 Overlapping Communities and an Opportunity for Sample Decontamination. mSystems. **2020**; 5(4):1–16.
401 30. Carmody LA, Zhao J, Schloss PD, et al. Changes in cystic fibrosis airway microbiota at pulmonary

402 exacerbation. Ann Am Thorac Soc. **2013**; 10(3):179–187.

403 31. Flight WG, Smith A, Paisey C, et al. Rapid Detection of Emerging Pathogens and Loss of Microbial Diversity
404 Associated with Severe Lung Disease in Cystic Fibrosis. J Clin Microbiol. **2015**; 53(7):2022–2029.

405 32. Quinn RA, Whiteson K, Lim YW, et al. Ecological networking of cystic fibrosis lung infections. npj Biofilms
406 Microbiomes [Internet]. Springer US; **2016**; 2(1):0–1. Available from: http://dx.doi.org/10.1038/s41522-
407 016-0002-1

408 33. Flynn JM, Niccum D, Dunitz JM, Hunter RC. Evidence and Role for Bacterial Mucin Degradation in Cystic
409 Fibrosis Airway Disease. PLoS Pathog. **2016**; 12(8):1–21.

410 34. McNally L, Brown SP. Building the microbiome in health and disease: Niche construction and social
411 conflict in bacteria. Philos Trans R Soc B Biol Sci. **2015**; 370(1675).

412 35. Bogaert D, Keijser B, Huse S, et al. Variability and diversity of nasopharyngeal microbiota in children: A
413 metagenomic analysis. PLoS One. **2011**; 6(2).

414 36. Rush ST, Lee CH, Mio W, Kim PT. The Phylogenetic LASSO and the Microbiome. **2016**; (August). Available
415 from: http://arxiv.org/abs/1607.08877

416 37. Leng C, Tran MN, Nott D. Bayesian adaptive Lasso. Ann Inst Stat Math. **2014**; 66(2):221–244.

417 38. Banerjee P, Garai B, Mallick H, Chowdhury S, Chatterjee S. A Note on the Adaptive LASSO for Zero-
418 Inflated Poisson Regression. J Probab Stat. **2018**; 2018:1–9.

419 39. Jian C, Luukkonen P, Yki-Jarvinen H, Salonen A, Korpela K. Quantitative PCR provides a simple and
420 accessible method for quantitative microbiome profiling. bioRxiv [Internet]. **2018**; :478685. Available
421 from: https://www.biorxiv.org/content/early/2018/11/27/478685

422 40. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of
423 Microbiome Census Data. Watson M, editor. PLoS One [Internet]. **2013**; 8(4):e61217. Available from:
424 https://dx.plos.org/10.1371/journal.pone.0061217

425 41. Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. corrplot: Visualization of a correlation matrix. R Packag

[Internet]. **2017**; . Available from: https://github.com/taiyun/corrplot%0Ahttps://github.com/taiyun/corrplot/issues

42. Yuan G-X, Ho C-H, Lin C-J. An improved GLMNET for l1-regularized logistic regression. Proc 17th ACM SIGKDD Int Conf Knowl Discov data Min - KDD ’11 [Internet]. New York, New York, USA: ACM Press; 2011. p. 33. Available from: http://dl.acm.org/citation.cfm?doid=2020408.2020421

434 **Tables & Figures**

	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 (78.6)	14 (56.0)	8 (53.3)	6 (26.1)	0.015
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

435

436 ***Table 1. Summary of patient clinical data, stratified by lung function.*** *Lung function classes are defined as*

437 *follows: Normal (ppFEV1 > 80); Mild (60 < ppFEV1 ≤ 80); Moderate (40 < ppFEV1 ≤ 60); and Severe (ppFEV1 < 40).*

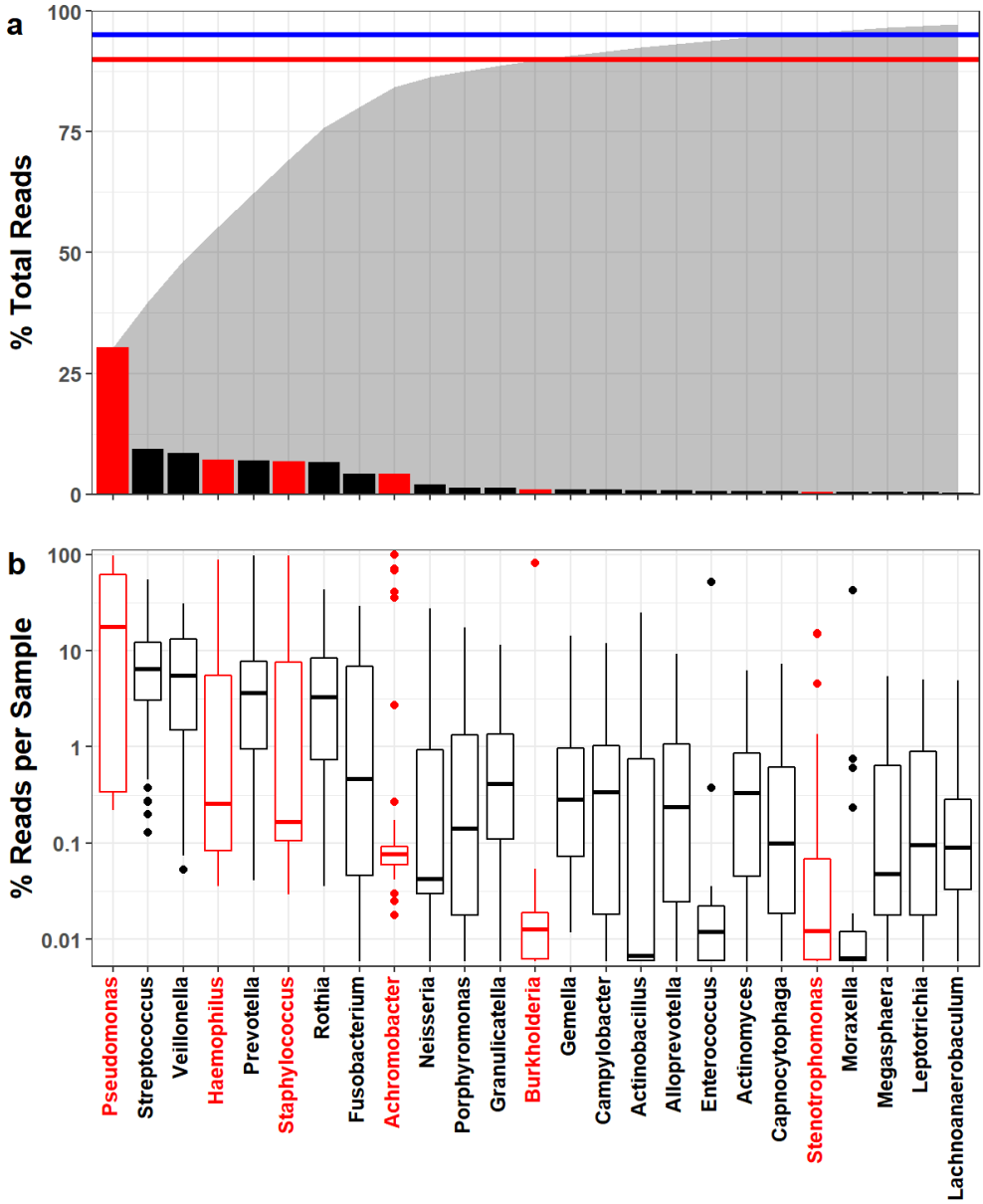
438 *Quantitative metrics are reported using the median and ranges. *Median reported values. **Two patients did*

439 *not have reported HbA1c values. Significant differences between lung function categories tested by ANOVA, p-*

440 *values shown. BC: Burkholderia, AX: Achromobacter, STE: Stenotrophomonas*

441

442



443

444 ***Figure 1. CF lung microbiomes are dominated by oral anaerobes and opportunistic pathogens*** *We analyzed CF*

445 *sputum expectorate (N=77) using 16S sequencing and an in-house QIIME 2-based bioinformatics pipeline to*

446 *resolve strain-level OTUs. Samples were rarefied to 17000 reads. We identified 217 OTUs across 59 genera and at*

447 *least 81 species. Overall, we find that CF sputum samples are dominated by oral anaerobes and opportunistic*

448 *pathogens.* ***a)*** *Sequences mapped to 14 genera comprised 90% (red line) of the total reads obtained. 95% (blue*

449 *line) of all reads mapped to 21 genera. Total cumulative read fraction represented in shaded region.*

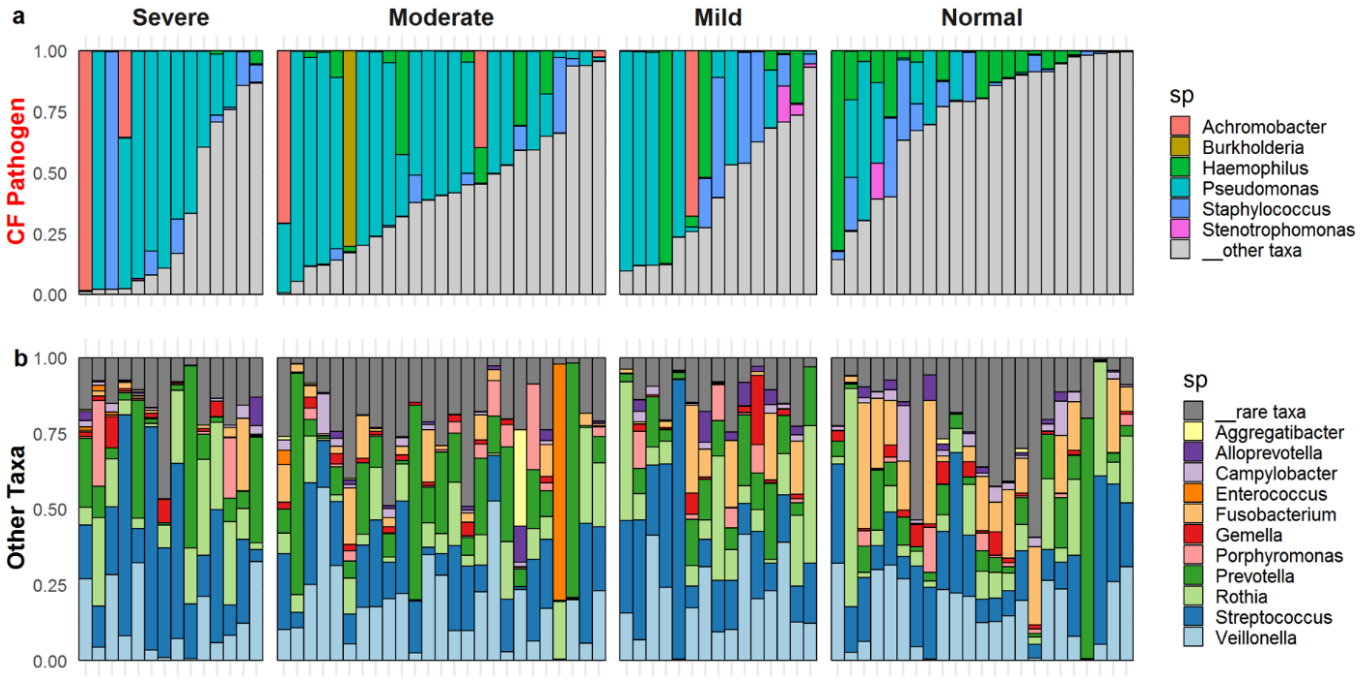
450 *Pseudomonas was the most prevalent genus, followed by Streptococcus and Veillonella.* ***b)*** *Binning reads by*

451 *sample shows variation in relative abundance. Pseudomonas comprises >10% of reads in the majority of our*

452 *samples. While over 6% of the total reads mapped to Achromobacter, only 4 samples were comprised of >10%*

453 *Achromobacter.*

454



455

456 ***Figure 2. CF Lung microbiome composition varies with lung function and pathogen dominance.*** *Relative*

457 *abundances of* ***(a)*** *6 canonical CF pathogens and* ***(b)*** *other taxa (the grey bar taxa in (a)). Microbiome*

458 *compositions grouped by disease severity, classified by ppFEV1 score: normal (80+), mild (60-80), moderate (40-*

459 *60), and severe (<40).*

460

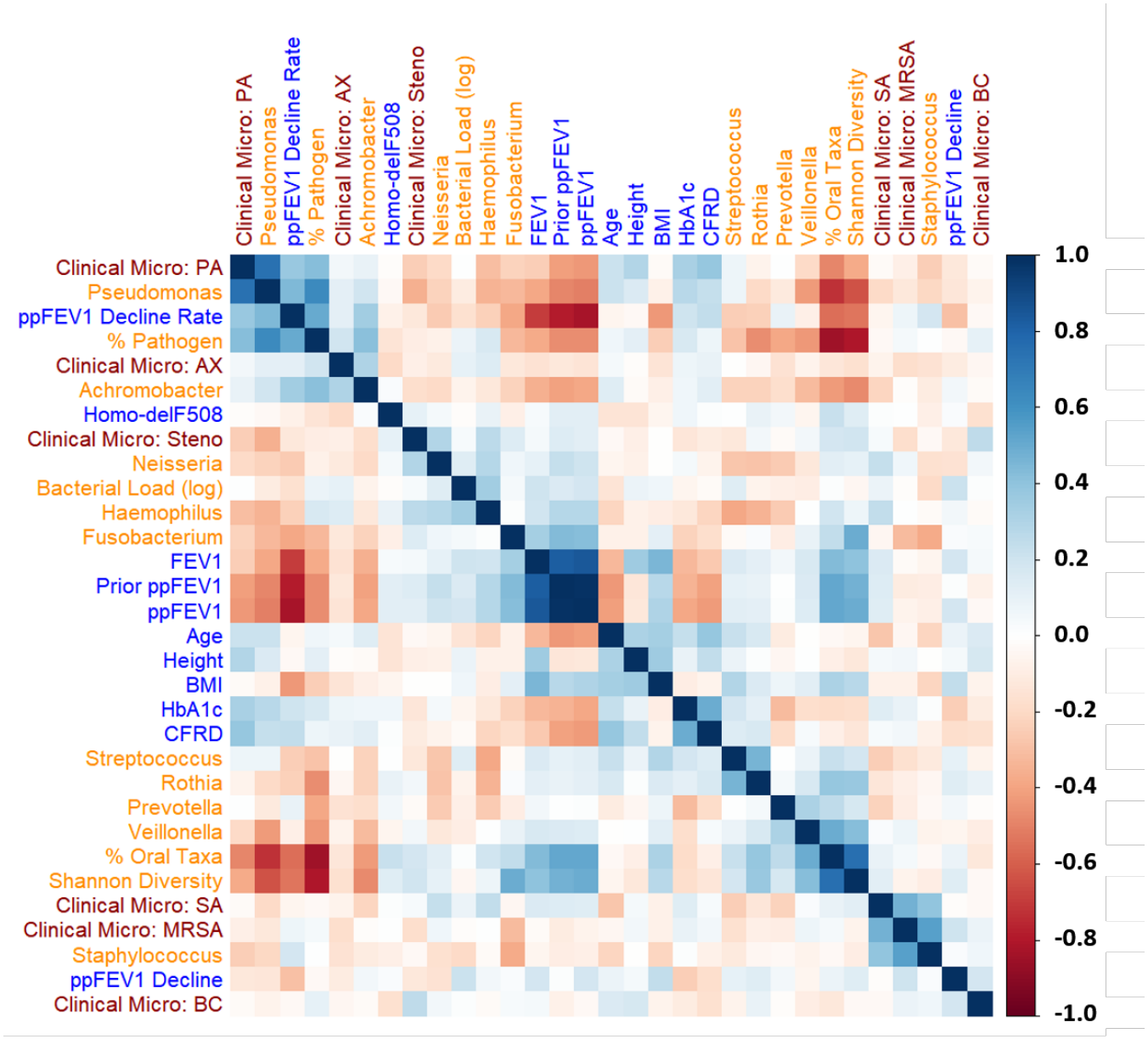


Figure 3. Lung function varies with patient meta-data. *Spearman correlations (R::corrplot) across all patient metadata (blue), clinical micro results (maroon), and microbiome data (orange, clr-transformed) reveal a complex correlation structure. We used a centered-log transform on 16S data to mitigate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

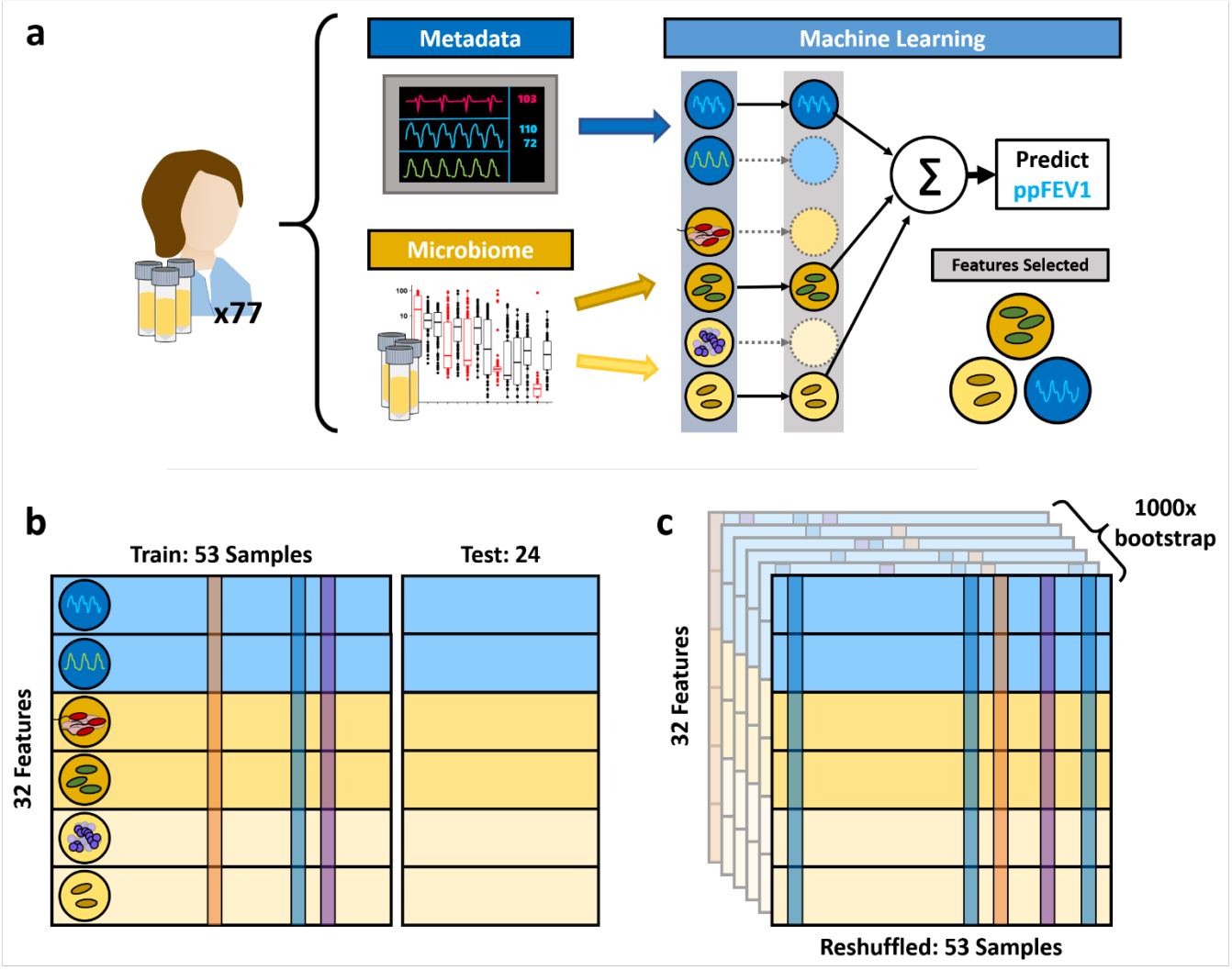


Figure 4. Machine Learning Overview. *Machine learning models are trained on different input data tables using varying data resampling methods. (**a**) Features are categorized by information source (microbiome or patient metadata). The 16S data is further split into pathogens and other taxa in agreement with Figure 2. 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.*

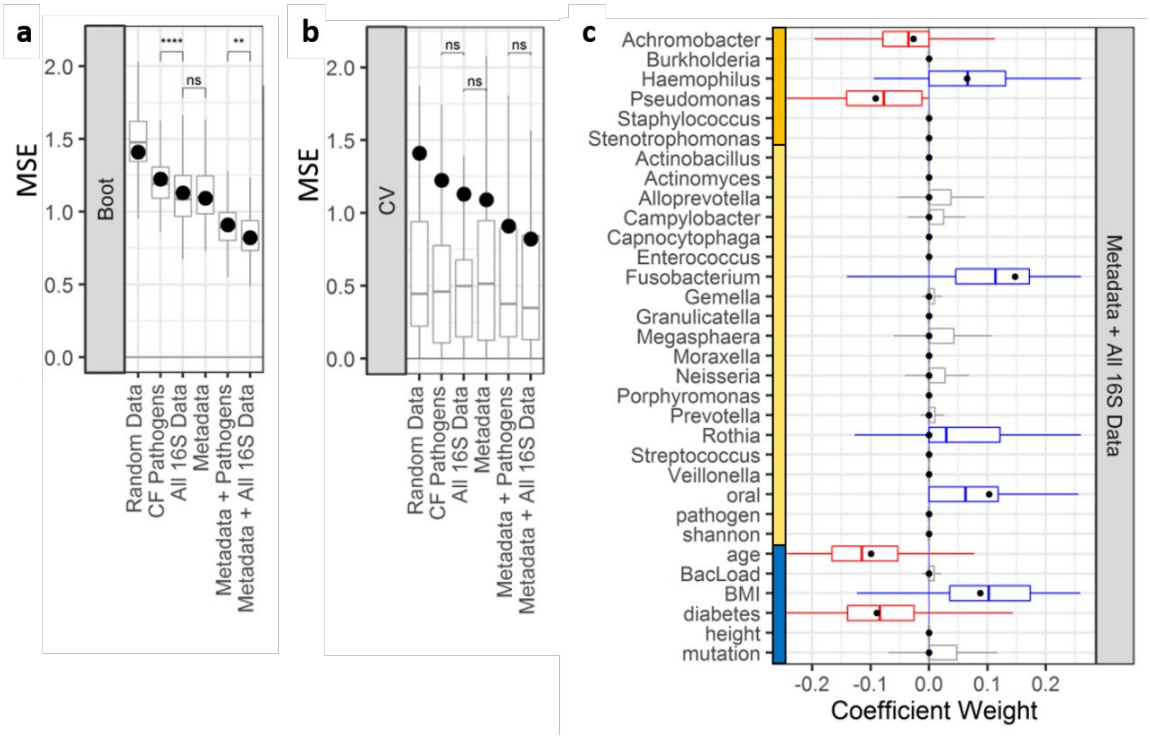


Figure 5. Bootstrapped ElasticNet-identified predictors of lung function. *ML models were trained using varying input datasets.* ***a)*** *1000-fold bootstrapping and* ***b)*** *leave one out cross-validation (LOOCV) were used to generate prediction error (MSE) ranges across feature subsets. Models trained 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.*