Mild CF Lung Disease is Associated with Bacterial Community Stability

Thomas H. Hampton[a1], Devin Thomas[b2], Christopher van der Gast[c3], George A. O'Toole[a4] (*), and Bruce A. Stanton[a5] (*)

(*) Co-Corresponding authors

[a]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA
[b]Hubbard Center for Genomic Studies, University of New Hampshire, Durham, New Hampshire, USA
[c]Department of Life Sciences, Manchester Metropolitan University, Manchester, UK

[1]Thomas.H.Hampton@Dartmouth.edu
[2]Devin.W.Thomas@gmail.com
[3]C.vanderGast@mmu.ac.uk
[4]George.A.O'Toole.Jr@dartmouth.edu
[5]Bruce.A.Stanton@dartmouth.edu

Abstract

Microbial communities in the airways of persons with CF (pwCF) are variable, may include genera that are not typically associated with CF, and their composition can be difficult to correlate with long-term disease outcomes. Leveraging two large datasets characterizing sputum communities of 167 pwCF and associated metadata, we identify five bacterial community types. These communities explain 24% of the variability in lung function in this cohort, far more than single factors like Simpson diversity, which explains only 4%. Subjects with *Pseudomonas*-dominated communities tended to be older and have reduced percent predicted FEV_1 ($ppFEV_1$) than subjects with *Streptococcus*-dominated communities, consistent with previous findings. To assess the predictive power of these five communities in a longitudinal setting, we used random forests to classify 346 additional samples from 24 subjects observed 8 years on average in a

range of clinical states. Subjects with mild disease were more likely to be observed at baseline, that is, not in the context of a pulmonary exacerbation, and community structure in these subjects was more self-similar over time, as measured by Bray-Curtis distance. Interestingly, we found that subjects with mild disease were more likely to remain in a mixed *Pseudomonas* community, providing some support for the climax-attack model of the CF airway. In contrast, patients with worse outcomes were more likely to show shifts among community types. Our results suggest that bacterial community instability may be a risk factor for lung function decline and indicates the need to better understand factors that drive shifts in community composition.

Introduction

Cystic fibrosis (CF) is a genetic disease affecting about 30,000 people in the United States and 70,000 world-wide (1). It has long been recognized that airway infection by *Pseudomonas aeruginosa* predicts morbidity and mortality in young children with CF (2). Direct sampling of the CF airway demonstrates that a small set of established CF pathogens such as *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Haemophilus influenzae* and species from genera *Burkholderia*, *Stenotrophomonas,* and *Achromobacter* dominate early infections in children with CF (3). The prevalence of *Pseudomonas* infection in the CF community increases with age (4), and it has been reported that few pathogens other than *Pseudomonas* are detectable in end stage disease (5). Not surprisingly, the prevailing thought was once that CF lung disease might be caused by a single pathogen (6-7), although not necessarily the same pathogen in all people (8). Consequently, treatments were directed at specific pathogens such *Pseudomonas aeruginosa*, *Staphylococcus aureus* and *Haemophilus influenzae* (9). Evidence from the last 15 years

52 demonstrates that genera outside the established set of CF pathogens, such as *Streptococcus*, and
53 anaerobes like *Prevotella* and *Veillonella* also play important roles in lung disease for persons
54 with CF (pwCF) (10-12).
55
56 CF lung disease is now considered a polymicrobial phenomenon (13) that can be described using
57 ecological concepts like species diversity. Diversity has been associated with improved lung
58 function (14-17) but increased diversity may also signal the onset of exacerbation (18).
59 Longitudinal analysis, in which a pwCF is followed over time, generally reveals that the CF
60 airway microbiome as relatively stable (18-21) and changes modestly during cycles of antibiotic
61 treatment and pulmonary exacerbations, acute phenomena characterized by a rapid decline in
62 overall health, accompanied by decreased lung function (19, 22-24) during exacerbations pwCF
63 experience rapid decline in overall health frequently accompanied by permanently decreased
64 lung function (19). More frequent exacerbations, and failure to recover fully from these
65 exacerbations, are thought to differentiate people with moderate or severe disease from those
66 with mild CF lung disease (26, 27). Intravenous antibiotic treatment in a hospital setting is
67 recommended during a pulmonary exacerbation (20) and recovery usually occurs over a period
68 of weeks (21). Complete recovery of lung function to levels prior to exacerbation is more
69 common in younger pwCF (22) and those with fewer exacerbations. Recovery may be mediated
70 by response to antibiotics like tobramycin, which is at least partly predictable from polymicrobial
71 signatures (23). It is possible that some exacerbations are triggered by increasing relative
72 abundance of anaerobes in the two weeks leading up to an exacerbation (24) or by specific
73 species of *Streptococcus*, including *anginosus* group streptococci (AGS), also known as

74 the *milleri* group streptococci (25-27) . Collectively, recent findings paint a complex picture of
75 how exacerbations occur in CF.
76
77 One model of how dynamic changes in the CF airway community might be associated with
78 changes in patient health suggests that the CF airway microbiome is normally a stable climax
79 community (28) dominated by a classic pathogen such as *Pseudomonas*, *Achromobacter*,
80 *Staphylococcus*. From time to time, the intrusion of facultative or obligate anaerobes, via an
81 undescribed mechanism, leads to a fermentative "attack community" that lowers pH and favors
82 further expansion of anaerobes (28). This population shift would increase bacterial species
83 diversity, consistent with the increased diversity during exacerbation that has been previously
84 observed (18). Interestingly, this is one instance where increased diversity is associated with a
85 worse clinical outcome, rather than better outcomes as is typically reported (14, 17, 29, 30, 31).
86 Furthermore, this climax-attack model is also consistent with the concept of stable core and
87 dynamic satellite taxa (16) because these satellite taxa are largely the same genera that are
88 thought to constitute the attack community, and it has been shown that satellite genera are more
89 dynamic across disease states (32).
90
91 The primary aim of this study was to improve our understanding of the baseline community
92 types, including 1) identifying discrete CF airway community types, 2) assessing whether
93 community types are associated with different health outcomes, 3) describing how community
94 types in pwCF change over time, and 4) whether community stability is associated with disease
95 severity. This analysis included a variety of machine learning techniques applied to both cross-
96 sectional and longitudinal data sets. Collectively, our results are consistent with a recent report

97 that bacterial communities in stable patients have less variation in community structure over time
98 (33), and our results support the climax-attack model of CF airway disease (28). In addition, we
99 identify a five-community model that accounts for 24% of the variability in lung function, far
100 more than any single factors explored previously.
101

102 ## Methods

103

104 **Data processing**. Publicly available 16S rRNA gene amplicon library sequence reads and
105 associated metadata from recent studies by Carmody et al. (PRJNA423040) (18) and
106 Cuthbertson et al. (PRJEB30646) (17) were downloaded from Bioproject
107 (https://www.ncbi.nlm.nih.gov/bioproject/). Reads in fastq format were mapped to unique
108 bacterial genera using QIME2 (34).
109

110 **Cross-sectional data set:** We assembled a cross-sectional data set by combining baseline data
111 (i.e., not exacerbating / receiving treatment for exacerbation) set from 167 unique pwCF, 84 from
112 the Cuthbertson et al. study (17) and 83 from the Carmody study (18). In those instances where
113 more than one baseline sample was available from a given subject, the first sample taken was
114 used, together with metadata (e.g. $ppFEV_1$). The two data sets were similar, but not identical in
115 terms of age and lung function. The 84 subjects included in the cross-sectional data set from the
116 Carmody et al. (17) were slightly younger, with a mean age of 25.6 ± 10.0 SD compared to a
117 mean age of 30.27 ± 10.9 SD. OTU level mapping and metadata for the cross-sectional two data
118 sets (17, 18) is provided in Supplemental Table 1

119 (https://doi.org/10.6084/m9.figshare.14261084.v1) and Supplemental Table 2
120 (https://doi.org/10.6084/m9.figshare.14261219.v1)
121
122 **Longitudinal data set:** The longitudinal data set was a subset of samples from the Carmody et
123 al. (18) study, specifically, 346 16S rRNA gene amplicon datasets from 24 subjects that were
124 observed at least 10 times. Age at time of sampling ranged from 11 to 54, with median of 27 and
125 an interquartile range of 8. Lung function ($ppFEV_1$) of subjects in the longitudinal data set
126 ranged from 9 to 123%, with a median of 50 and an interquartile range of 36. Longitudinal
127 subjects are more fully described in Table 1. OTU level mapping and metadata for the
128 longitudinal data set (18) is provided in Supplemental Table 3
129 (https://doi.org/10.6084/m9.figshare.14261432.v1).
130
131 **Statistical analyses**. Data analysis was performed in R 4.0.1 (https://www.R-project.org/). Bray-
132 Curtis dissimilarity was calculated using ecodist (35). Simpson diversity was calculated with
133 vegan 2.5-7 (36). Clustering was performed using kmeans in base R and clusters were visualized
134 and gap analysis was performed using factoextra 1.0.7 (37). Random forests were generated
135 using randomForest 4.6-14 (38). Discrete Time Markov chains were determined using the
136 markovchain package 0.8.5-4 (39).
137
138 ## Results
139
140 *Five Clusters to Explain Two Large Cross-sectional Data Sets*
141

142 As the emphasis in CF airway microbiome investigation has shifted toward a polymicrobial
143 interpretation (13), data analysis has increasingly focused on multivariate comparisons. We
144 therefore began our analysis to identify polymicrobial community clusters by calculating Bray-
145 Curtis dissimilarity (40) between genus level taxonomic assignments in bacterial 16S rRNA gene
146 amplicon sequencing data from 167 sputum samples from people with CF. Bray-Curtis similarity
147 (40) is a common measure used to summarize the similarity between microbial communities, and
148 places a larger weight on species that are commonly observed, partly mitigating the problem of
149 random sampling errors associated with rare species. Samples for this analysis were chosen as a
150 cross sectional subset of two recently published studies (17, 18), and included clinically stable
151 subjects aged 8 to 69 from 14 CF centers in the US and Europe creating a snapshot that should
152 include many of the community types present in pwCF. We calculated the Bray-Curtis similarity
153 across 167 cross-sectional samples from these two independent studies: a cross-sectional study
154 from Cuthbertson et al. colleagues (17), from which 83 samples were analyzed, and a
155 longitudinal study by Carmody et al. (18), from which we selected a single representative sample
156 from each patient for this analysis (84 samples).
157
158 We next used k-means to cluster samples using the Bray-Curtis similarity matrix we generated
159 above. The k-means machine learning algorithm requires the user to identify how many clusters
160 (k) are to be identified in the data, then uses an iterative approach to assign samples to each
161 cluster to minimize distances between samples in the same cluster and maximize distances
162 between samples in different clusters. Such machine learning approaches are commonly applied
163 to microbiome studies (41). For example, unsupervised techniques like hierarchical clustering
164 (42) have been used to identify core taxa in the CF airway (16) and gut (43) and to demonstrate

165 that samples from the same person tend to be similar to each other, even across disease and
166 treatment (44). K-means analysis has also been used to identify groups of patients with similar
167 patterns in inflammation, microbiota, or clinical factors (45). What makes k-means unsupervised
168 is that similarity between clusters is based only on observed data, in this case, counts of bacterial
169 genera. Unsupervised learning does not take sample labels (e.g., "severe disease" or
170 "exacerbation") into consideration. Therefore, when unsupervised clustering places samples
171 labeled "severe disease" together, it is likely that a non-random process (e.g., the result of a
172 biological difference) caused the samples to cluster.
173
174 As described above, the k-means algorithm requires the user to identify how many clusters (k)
175 are to be identified in the data. Since we had no prior information to justify a particular choice of
176 k, we used another machine learning approach, the gap statistic (46), to estimate the number of
177 distinct clusters in our cross-sectional data. The gap statistic has recently been used to evaluate
178 the optimal number of clusters in CF airway microbiome data (29). The gap statistic identifies
179 the smallest value of k that minimizes distances between samples in the same cluster and
180 maximize distances between samples. Gap analysis suggested the existence of five distinct
181 community types in our cross-sectional data set of the 167 baseline samples analyzed here. We
182 found that subjects belonging to different communities also differed significantly in lung
183 function (see below); this finding suggests that clusters are based on biological differences (47)
184 and that community types may be clinically relevant. The optimal number of clusters (five)
185 chosen by the gap statistic algorithm is shown by the vertical dashed line in Figure 1A. The basis
186 of this choice is driven by the break in the data after k=5.
187

A principal coordinate analysis that captures >70% of the variance of these data (Figure 1B) shows how the clusters separate into 5 largely non-overlapping groups. It is important to note that this analysis raises the possibility of alternative, potentially equally valid grouping of the samples and/or potential overlap among these communities.

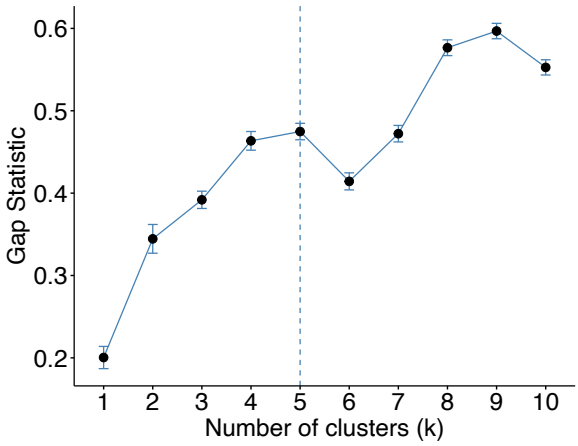


Figure 1A: Gap statistic for different numbers of k clusters, ranging from 2 to 10. Larger values indicate greater separation between clusters. An optimal number of clusters based on gap statistic is shown by a dashed vertical line.

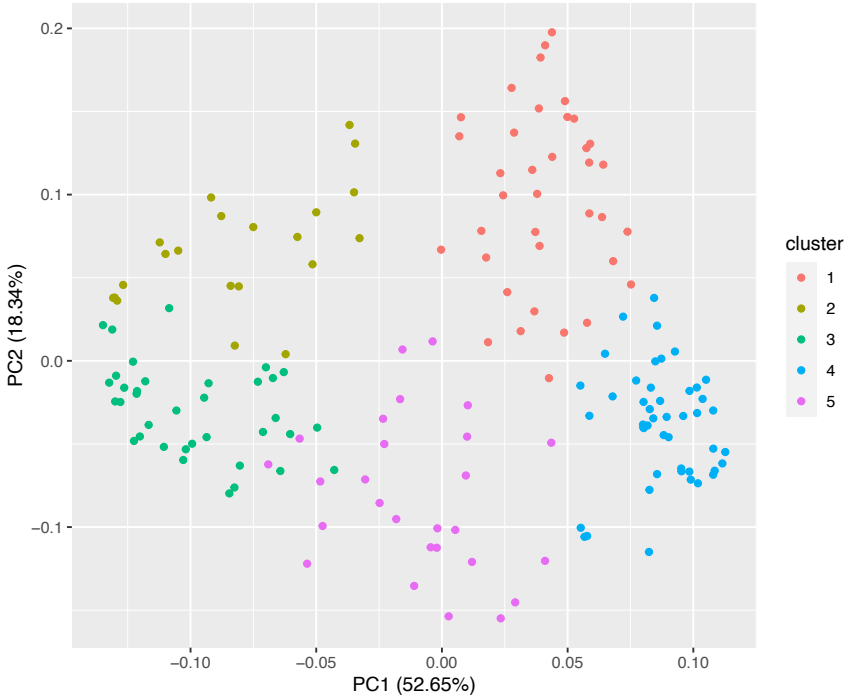


Figure 1B: Principal coordinate analysis of 5 clusters, based on Bray-Curtis dissimilarity matrix. The legend showing the color code of each cluster is on the right. Together, the first two principal coordinates capture >70% of the variance in these data. Each dot represents one of 167 samples in the cross-sectional analysis described in Methods.

It is reassuring that choosing 5 clusters also optimizes the ability of communities to explain patient differences in lung function. Figure 1C shows the value of adjusted r^2 from linear models of $ppFEV_1$, a measure of lung health, as a function of cluster membership. Five clusters explain about 24% (adjusted r^2= 0.24) of patient differences; using six clusters reduces r^2. In summary, two independent lines of evidence suggest that 5 clusters may be optimal for summarizing these data.

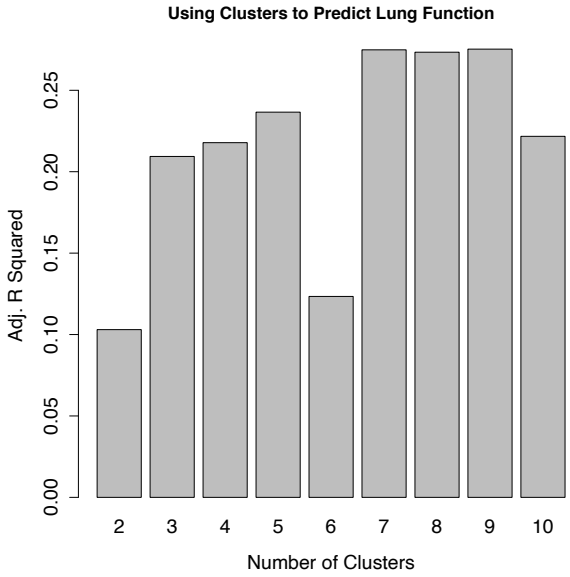
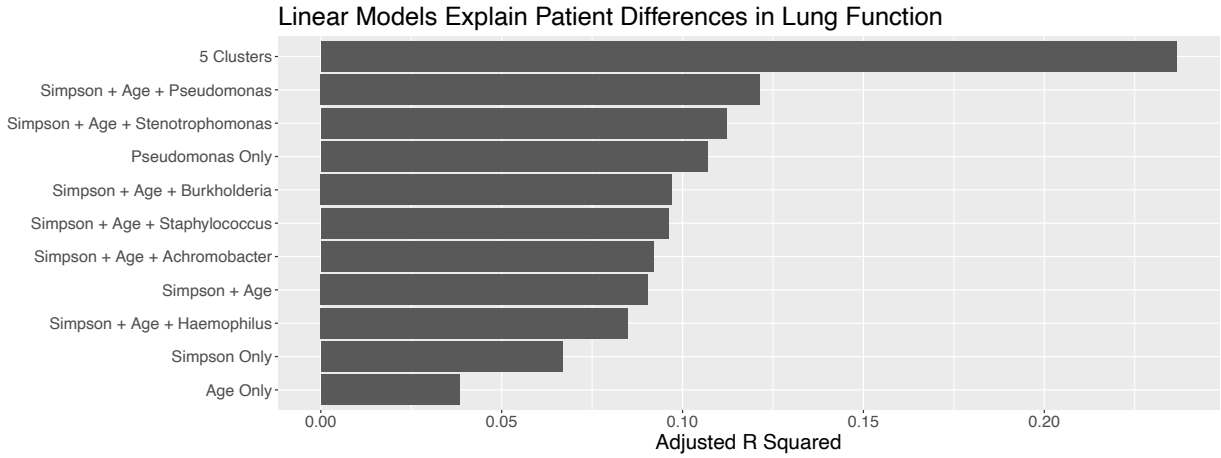


Figure 1C: Adjusted r squared from linear models of lung function (ppFEV1) using membership in a given cluster as the independent variable. Models were run for different numbers of k clusters, ranging from 2 to 10.

Clusters Membership Predicts Lung Function Better than Pseudomonas Relative Abundance, Age or Community Diversity

Despite a reframing of the conversation surrounding the CF airway microbiome away from single pathogens toward ecological descriptions involving species diversity (14, 17-18, 48, 49), the relationship between *Pseudomonas* and lung function is still frequently cited (14, 23, 49, 50). We used linear models to assess the level of association between lung function and commonly identified explanatory variables such as the relative abundance of *Pseudomonas* as well as Simpson Diversity and age. As shown in Figure 2, single variables like relative abundance of

228 *Pseudomonas*, Simpson diversity and age explain less than 11% of patient variability in lung
229 function (adjusted $r^2 < 0.11$). In other words, over 89% of variability in lung function is
230 explained by factors other than *Pseudomonas*, Simpson diversity or age. The explanatory power
231 of these models improves with increasing complexity, and a combined model that incorporates
232 Simpson diversity, age and the relative abundance of *Pseudomonas* or other classical CF
233 pathogens in a single model still explains less than 12% of patient variability in lung function. In
234 contrast, membership in one of the 5 communities identified above explains 24% of patient
235 variability in lung function. In other words, knowing which cluster a person's airway
236 microbiome belongs to provides a better estimate of their lung function than knowing the relative
237 abundance of Pseudomonas in their airway microbiome (Figure 2).



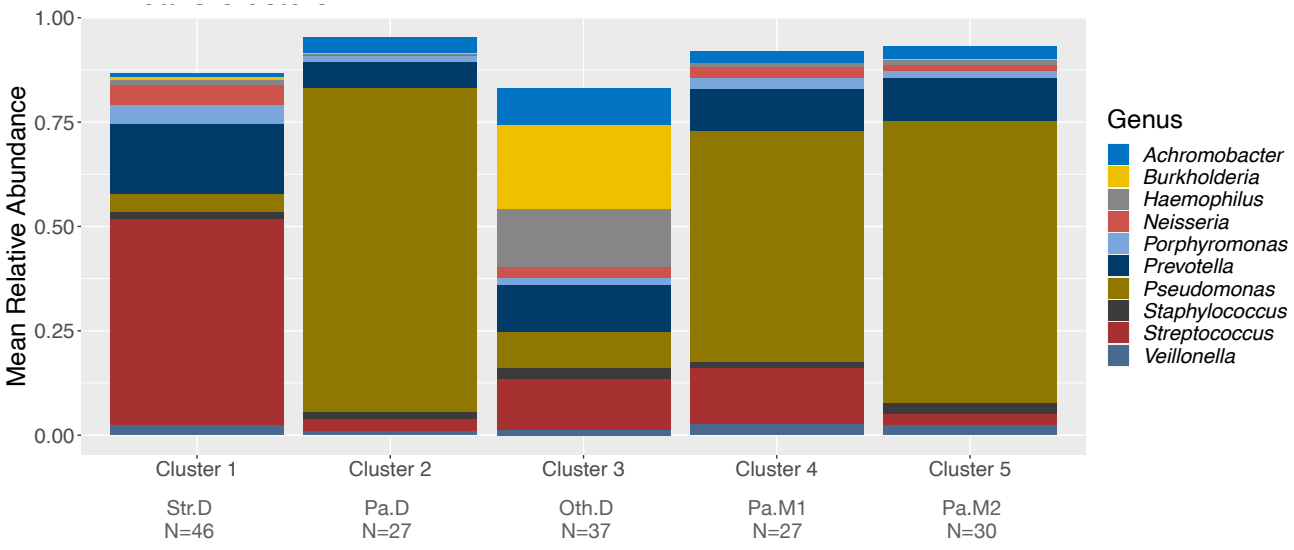
238

239 Figure 2: The ability of linear models to predict $ppFEV_1$, as estimated by adjusted R squared.
240 Individual factors or combinations (denoted by "+") of covariates shown on y axis: Simpson
241 Diversity only, Age only, *Pseudomonas* only, the combination of Simpson and Age, or Simpson
242 and Age and the relative abundance of other classic CF pathogens, compared to membership in
243 one of the 5 clusters.

244

Top 10 Bacterial Genera Across Clusters

We identified the top 10 genera with the highest mean relative abundance across the 167 cross-sectional samples and averaged their relative abundance in each of the 5 clusters, as shown in Figure 3A. Choosing the top 10 genera for visualization still includes the vast majority of the reads (mean 90%) that mapped to a genus. The first cluster (Str.D), which contained 46 samples was, on average, dominated by *Streptococcus* (brick red). Clusters 2, 4 and 5 had large amounts of *Pseudomonas* (olive), but were distinguished by the differences in abundance and/or type of minor genera. Cluster 2 (Pa.D) was dominated by *Pseudomonas*. While Cluster 4 (Pa.M1) and Cluster 5 (Pa.M2) were similar, they clustered separately, perhaps because Cluster 5 had more *Achromobacter* and *Burkholderia* than cluster 4 ($P < .001$ asymptotic Wilcoxon-Mann-Whitney test), and Cluster 4 had more *Streptococcus* ($P < .001$, asymptotic Wilcoxon-Mann-Whitney test). Cluster 3 (Oth.D) was dominated by a genus other than *Streptococcus* or *Pseudomonas* and could be described as having an even mix of genera based on averages.

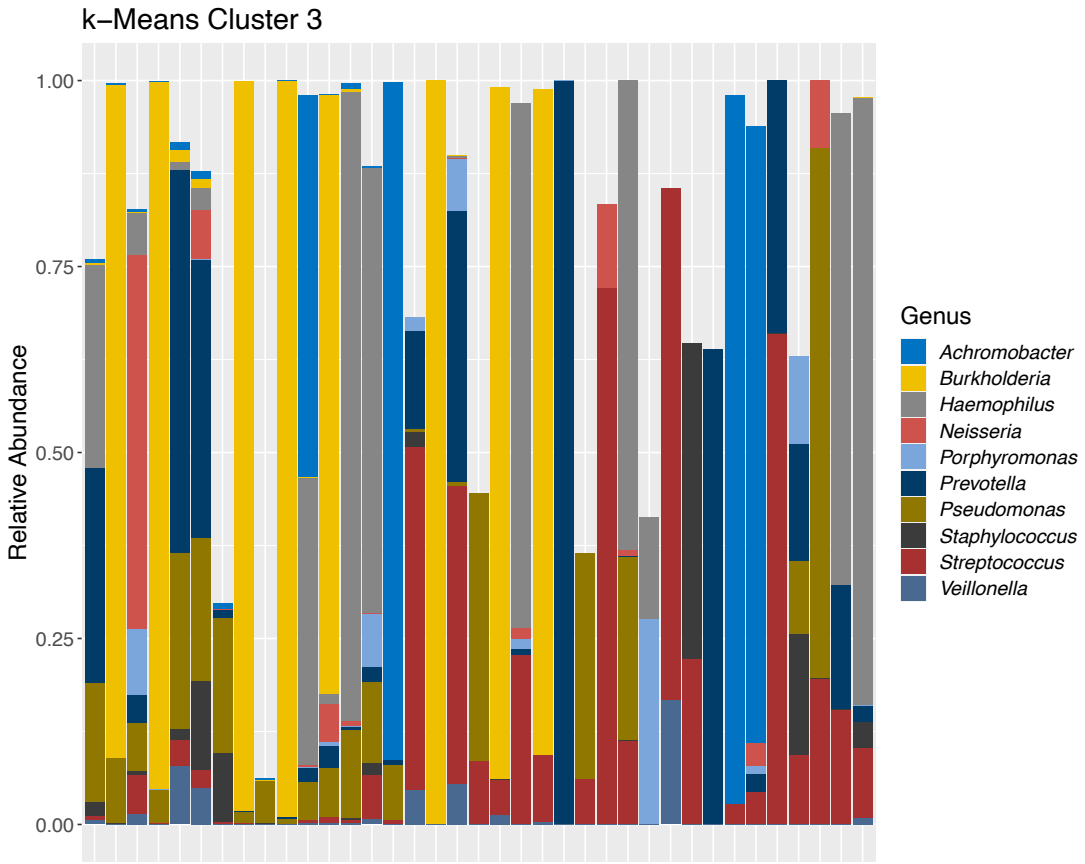


260 Figure 3A: Mean abundances of top 10 genera in each of the 5 clusters. Mnemonic
261 abbreviations: Str.D = *Streptococcus* dominated, Pa.D = *Pseudomonas* dominated, Oth.D =
262 Other dominated, Pa.M1 = *Pseudomonas* mixed community #1, Pa.M2 = *Pseudomonas* mixed
263 community #2. The number of samples that belonged to each cluster appears beneath each
264 cluster mnemonic.
265
266 We then mapped each of these clusters back to the two cross-sectional datasets, as shown in
267 Table 2. The Carmody data set (18) was from as single CF center in the US, and the Cuthbertson
268 data set (17) included subjects from Europe and the US, spanning 13 different CF centers. Some
269 clusters, but not others, were enriched in data from particular studies perhaps as a consequence
270 of biases in sample collection/processing, differences in subject age, and/or systematic
271 differences in patient care. Of note, the two *P. aeruginosa* mixed communities, Pa.M1 and
272 Pa.M2, largely clustered as a function of the dataset from which they were derived.
273 Interestingly, the Carmody (18) data set had no patients in the *P. aeruginosa* dominant cluster
274 (Cluster 2, Pa.D), while the Cuthbertson (17) samples, collected over 13 centers had ~1/3 of its
275 sample map to this community cluster. In contrast, for the communities assigned to the Carmody
276 dataset (18) , more than half were described as *Streptococcus*-dominated, while the remaining
277 samples, while containing *P. aeruginosa*, were more evenly mixed (Pa.M1) or dominated by
278 other pathogens (Oth.D).
279

Cluster Number	Mnemonic	Carmody	Cuthbertson
1	Str.D	45	1
2	Pa.D	0	27
3	Oth.D	19	18
4	Pa.M1	19	8
5	Pa.M2	1	29

Table 2: Number of subjects in cross sectional data set mapping to various clusters shown in Figure 2, for the Carmody (18) and Cuthbertson (17) data sets.

Interestingly, averages do not capture the specifics of a group perfectly, as shown in Figure 3B. Individual samples that mapped to cluster 3 (Oth.D; Other dominated), which when averaged appears to have an even mix of genera, tended to be dominated by individual genera, including *Stenotrophomonas* which was not among the top 10 most abundant genera across all 167 samples. We therefore complemented our analysis of the 10 most abundant genera (on average) with an analysis of the 10 most dominant genera, that is, the top 10 genera achieving the highest relative abundance in any of the 167 samples.



293

294 Figure 3B: Abundances of top 10 genera relative abundance of cluster 3 (Oth.D) in each of the

295 37 samples belonging to this group.

296

297 *Haemophilus*, *Stenotrophomonas*, *Burkholderia*, *Prevotella* and *Achromobacter* achieved high

298 levels of dominance (relative abundance greater than 85%) in samples in cluster 3, as shown in

299 Figure 3C, but not in other clusters. *Streptococcus* achieved dominance in cluster 1, but not

300 elsewhere. *Pseudomonas* achieved high levels in clusters 2, 4 and 5.

301

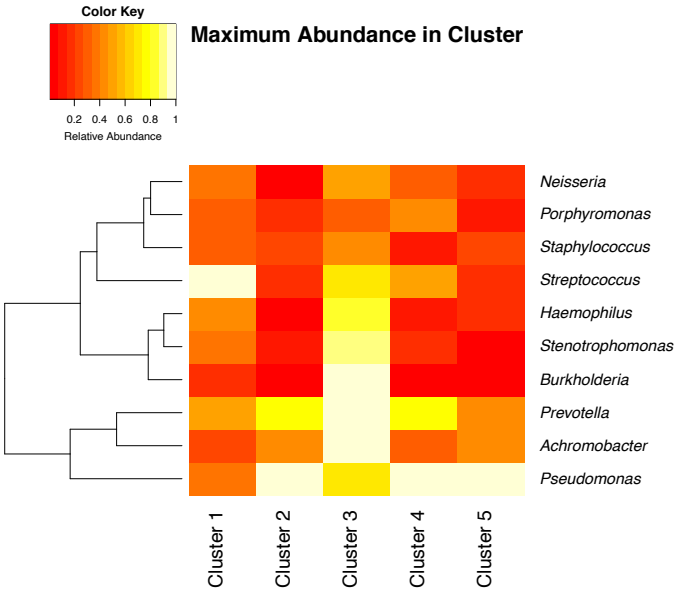
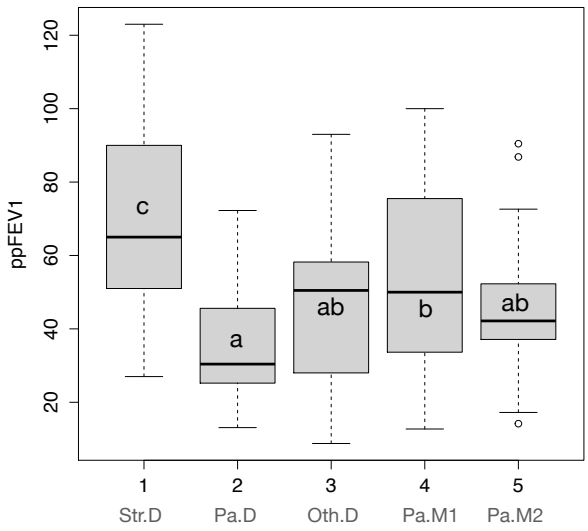


Figure 3C: Maximum abundance of 10 genera achieving the highest relative abundance in any of the 167 samples, as a function of cluster.

Clusters Dominated by Streptococcus and Pseudomonas Differ in Lung Function and Age

It has previously been observed that species from the *Streptococcus* genus are part of the core CF airway microbiome (16, 31, 50), that the prevalence of *Streptococcus* spp. is sometimes associated with less severe lung function loss (12, 14, 31, 51), but that some genera of *Streptococcus*, including anginosus group streptococci (AGS; also known as the milleri group streptococci) may facilitate exacerbation (25-27). Our cross-sectional analysis highlights the positive aspect of patients carrying the community dominated by *Streptococcus*; subjects whose airway microbiomes were characterized as *Streptococcus*-dominated tended to have better lung function, as shown in Figure 4A, and be younger, as shown in Figure 4B.



317

318 Figure 4A: Association of communities with lung function. Subject-Lung Function profiles of 5

319 distinct communities. Box and whisker plot comparing ppFEV1 between *Streptococcus*-

320 dominated (Str.D), *Pseudomonas*-dominated (Pa.D), communities dominated by other genera

321 (Oth.D), *Pseudomonas* mixed community #1 (Pa.M1), and *Pseudomonas* mixed community #2

322 (Pa.M2), as shown in Figure 3A. Communities sharing a common letter are not significantly

323 different; those with no letter in common differ significantly (Tukey Honest significant

324 difference $P < .05$). For example, subjects with Str.D communities had higher $ppFEV_1$ than those

325 with any other community type.

326

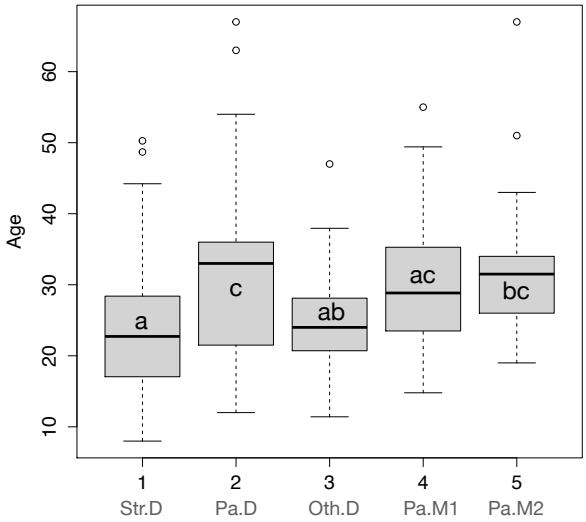


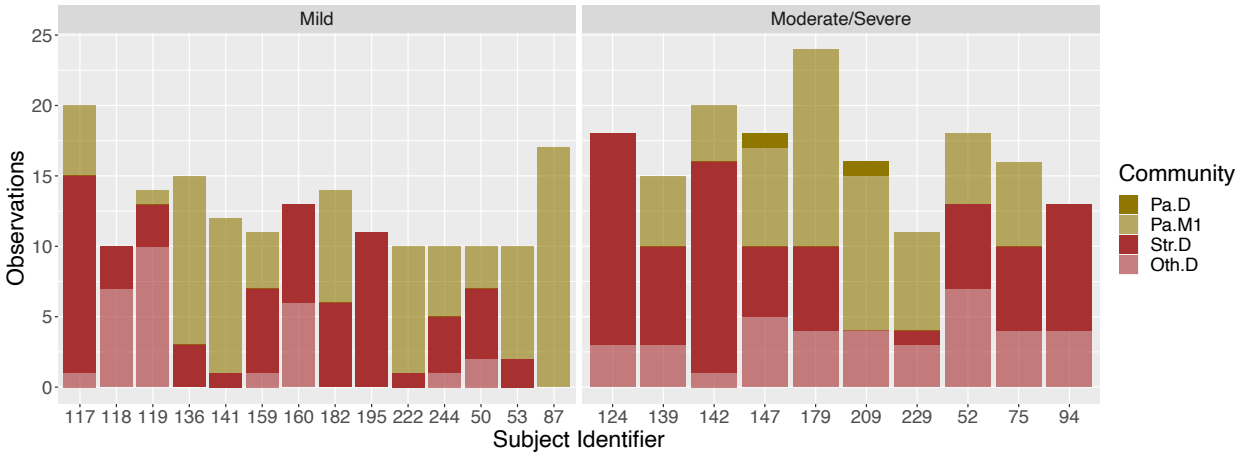
Figure 4B: Association of communities with age. Subject-age profiles of 5 distinct communities. Box and whisker plot comparing age between *Streptococcus*-dominated (Str.D), *Pseudomonas*-dominated (Pa.D), communities dominated by other genera (Oth.D), *Pseudomonas* mixed community #1 (Pa.M1), and *Pseudomonas* mixed community #2 (Pa.M2), as shown in Figure 3A. Communities sharing a common letter are not significantly different; those with no letter in common differ significantly (Tukey Honest significant difference $P < .05$). For example, subjects with Str.D communities were younger than those with Pa.D communities or Pa.M2 communities but did not differ significantly in age from subjects with Oth.D or Pa.M1 communities.

Communities Observed in Subjects with Mild versus Moderate/Severe Disease are Similar

Findings in the cross-sectional data (Figure 4A, B) are consistent with previous reports showing that communities dominated by *Pseudomonas* tend to be associated with older subjects and reduced lung function (36, 39, 44). Given that young, healthier pwCF tend to have sputum communities dominated by *Streptococcus* led to the hypothesis that people with such

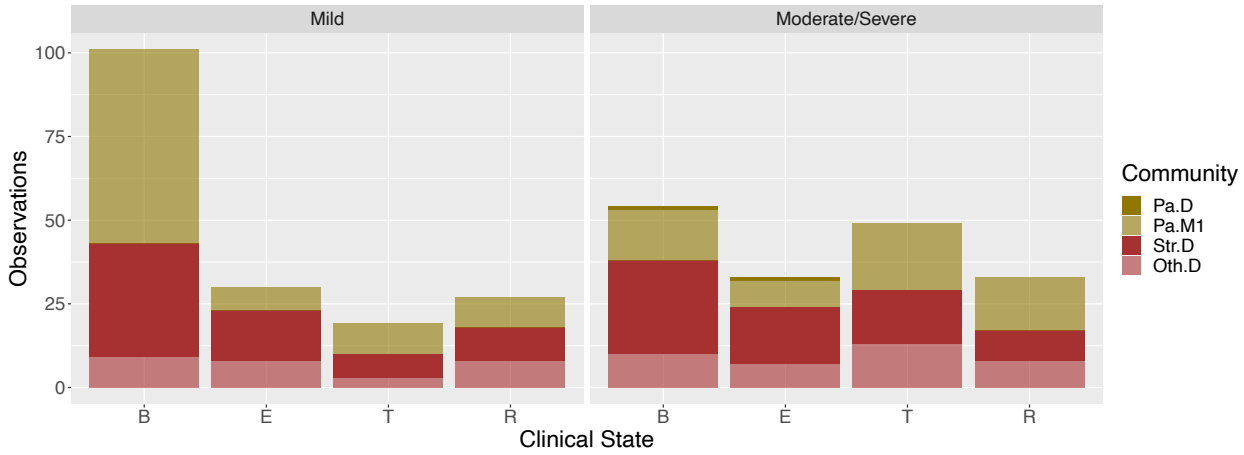
343 *Streptococcus*-dominated communities ultimately transition to communities dominated by
344 *Pseudomonas* as they age. To probe this idea, we used a longitudinal data set to assess how
345 communities changed in the 24 subjects detailed in Table 1, for whom there are 10 or more
346 longitudinal samples collected for times periods ranging from ~6-10 years (average of ~8 years).
347
348 For this analysis we used supervised machine learning to track community type changes in
349 individuals with CF through time. Briefly, we used random forests (52) to identify features that
350 reliably classify a sample as belonging to one of the five previously identified communities.
351 This classifier enabled us to identify the community type of 346 16S rRNA gene amplicon
352 datasets from 24 subjects that were observed across many years (18). This longitudinal data set
353 included 16S rRNA gene amplicon dataset measurements made in different clinical states
354 (baseline, exacerbation, treatment and recovery) and includes clinical metadata such as whether a
355 subject had been diagnosed as having mild or more severe disease. This approach mapped the
356 346 16S rRNA gene amplicon datasets to the four communities shown in Figure 3, specifically,
357 Str.D., Pa.D, Pa.M1 and Oth.D. No samples in the Carmody (18) longitudinal data mapped to
358 Pa.M2, consistent with the description of Pa.M2 cluster being specific Cuthbertson (17) dataset
359 (see Table 2).
360
361 Roughly two or three community types were required to describe any given subject over time,
362 recapitulating previous reports that the CF airway microbiome is fairly stable over time (15, 18,
363 19, 38, 45, 46), but may shift in structure around the time of exacerbation or in response to
364 antibiotic therapy (23, 45). We did not observe large differences in community types between
365 subjects with mild disease compared to moderate/severe disease either as a function of subject

366 (Figure 5A) or disease state (Figure 5B). For example, subject 117 was observed 15 times to
367 show a *Streptococcus* dominated community (Str.D) and 5 times to show the *Pseudomonas*
368 mixed community number 1 (Pa.M1).
369
370 We do note some associations. For example, domination by a pathogen other than *Pseudomonas*
371 (Oth.D) was associated with decreased probability of being observed in the baseline state (p-
372 value = 0.004, odds ratio 0.43, Fisher's exact test). Furthermore, despite similarities, subjects
373 with mild disease did differ from those with moderate/severe disease in specific ways. Subjects
374 with mild disease were more likely to be observed having a community classified as Pa.M1 (p-
375 value = 0.02, odds ratio = 1.69, Fisher's exact test) and more likely to be observed in the baseline
376 state (p-value = 3.1e-06, odds ratio = 2.8, Fisher's exact test) than people with moderate/severe
377 disease.
378



379
380 Figure 5A: Frequency of observing various community types in subjects designated as having
381 mild or moderate/severe disease. Subjects with mild disease (N=14, left) are compared to
382 subjects with moderate/severe disease (N=10, right). "Subject identifier" corresponds to "isolate"

383 in Table 1, a field provided in the Carmody (18) data set. Community types identified in these
384 longitudinal samples were deemed similar to those in Figure 3A by random forest classifier.
385
386
387



388
389 Figure 5B: Frequency of observing various clinical states for the 14 subjects with mild disease
390 aggressiveness (left) and 10 subjects with moderate/severe disease (right) as shown in Table 1.
391 Clinical states are defined as (B = baseline, E = exacerbation, T = treatment, R = Recovery).
392 Community types identified in these longitudinal samples were deemed similar to those in Figure
393 3A by random forest classifier.
394
395
396 Interestingly, we note that people with mild disease have bacterial communities that are more
397 stable over time based on being observed in a specific community category (Figure 5). We tested
398 this association using a continuous variable by calculating Bray-Curtis distances for all pairwise
399 comparison of samples taken from any given subject over time. As shown in Figure 6A,

communities from subjects with mild disease were less dissimilar to each other than communities from subjects with moderate/severe disease, suggesting that variation in community composition is associated with worse disease outcomes.

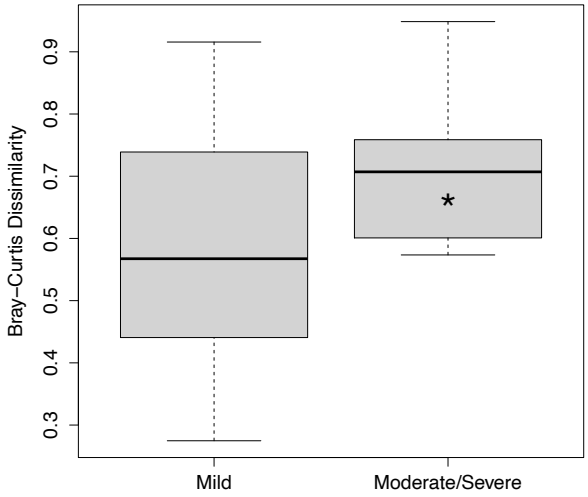


Figure 6A: Median within-subject pairwise distances (Bray-Curtis dissimilarity) for N=14 subjects with mild disease aggressiveness (left) and N=10 subjects with moderate/severe disease (right) as shown in Table 1. Subjects with mild disease had samples that were significantly more similar to each other ("*" signifies $P < .05$, Wilcoxon rank sum exact test).

As a second approach to assess the association between community stability and clinical outcome, a Markov chain analysis was performed to track subjects and their associated community types across time. Markov models simplify dynamic systems by representing them as a set of finite states that are connected to each other enabling transitions between the states. With enough data, the geometry of these systems (i.e., states that are connected to each other, enabling

417 a transition) can be determined. In addition, the probability of transitioning between any two
418 connected states can be estimated. Markov modeling of this system suggests that this self-
419 similarity arises at least partly because subjects with mild disease are repeatedly observed with a
420 Pa.M1 community type (Figure 6B). Specifically, given that a subject was characterized as
421 having mild disease and was last observed with community type Pa.M1, they had a 73%
422 probability of being observed in Pa.M1 again. In contrast, the geometry associated with
423 moderate/severe disease was more complex, and subjects were less likely to remain in the Pa.M1
424 state (60%) and also found more often in the Pa.D community (Figure 6C). Overall, repeated
425 observation of Pa.M1 was significantly more likely in subjects with mild disease compared to
426 moderate/severe disease (red arrow in 6B: "*" indicates P = .03, Fisher's exact test, odds ratio
427 1.78). Together, these data indicate that patients with stable communities tend to have better
428 airway function.

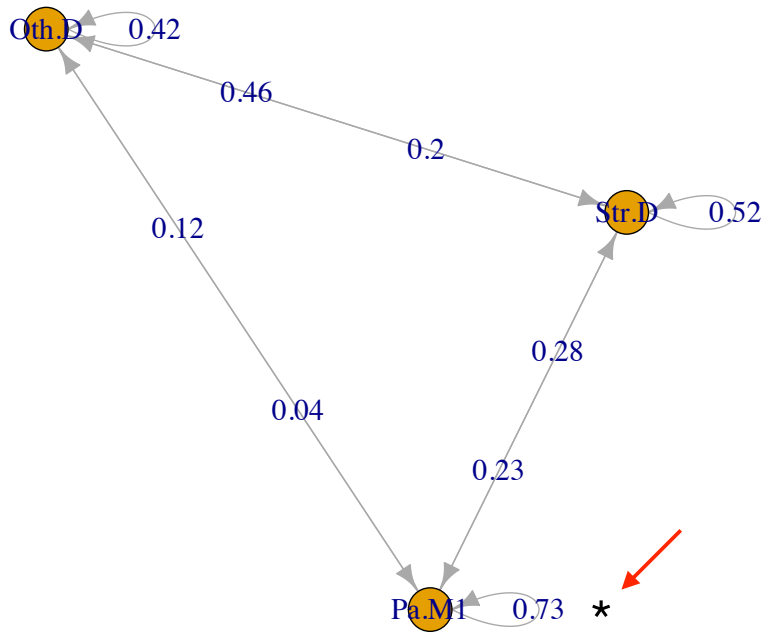


Figure 6B: Markov chain diagram for community transitions in subjects with mild disease (see Table 1). Each node in the diagram (orange circles) represents the state of being observed with a specific bacterial community. Edges leading from each node represent possible transitions,

including the possibility of observing the same community again (loop). Numbers on arrows are estimates of the probability of taking a specific path.

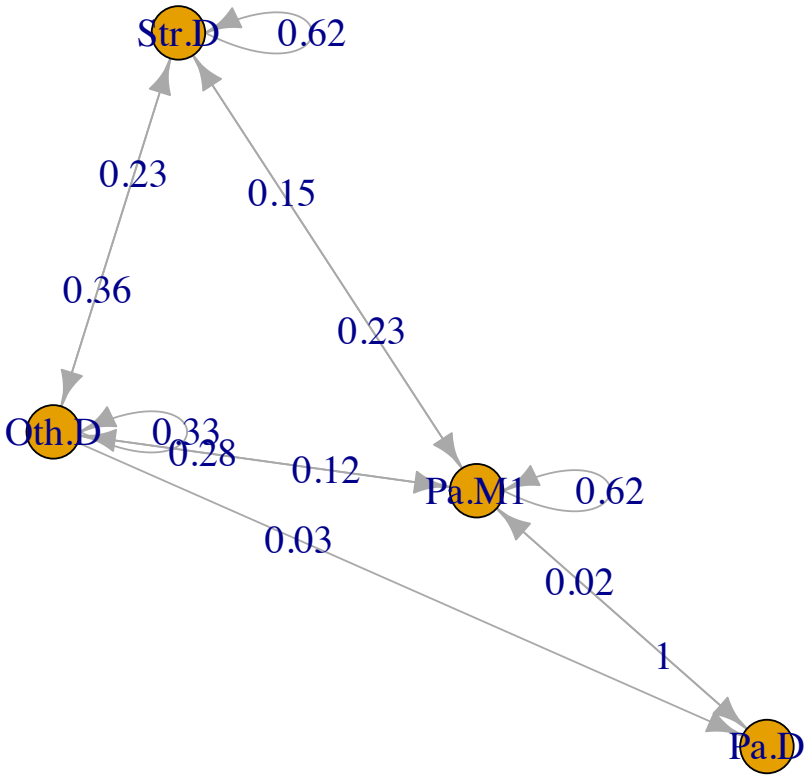


Figure 6C: Markov chain diagram for community transitions in subjects with moderate/severe disease (see Table 1). Each node in the diagram (orange circles) represents the state of being observed with a specific bacterial community. Edges leading from each node represent possible

442 transitions, including the possibility of observing the same community again (loop). Numbers on
443 arrows are estimates of the probability of taking a specific path.
444
445 DISCUSSION
446
447 We present two major findings in this report. First, the previously reported climax-attack model
448 postulates that an attack community composed of genera like *Streptococcus*, *Veillonella* and
449 *Porphyromonas*, perturbs a climax community dominated by *Pseudomonas* and *Staphylococcus*
450 (28), resulting in a metabolic shift toward fermentation and decreasing pH. These conditions are
451 more favorable to fermentative bacteria like *Prevotella*, *Veillonella*, *Streptococcus*, *Rothia* and
452 *Granulicatella* (28). Our observation that mild disease is associated with community stability,
453 and that stability is associated with a *Pseudomonas*-dominated community are therefore
454 consistent with the climax-attack model of CF.
455
456 Second, we have shown that a relatively small number of CF microbiome types explain patient
457 differences in lung function, and that they explain these differences much better than single
458 variables like the relative abundance of *Pseudomonas*, age or Simpson diversity. This finding is
459 important because developing interventions to treat the community as the pathogenic unit will
460 require a reasonable number of experimentally tractable model communities. Without such
461 models, we would be faced by a dizzying array of patient differences. Fortunately, the airway
462 community in a pwCF over multiple samplings generally looks more like "self" than any other
463 patient when using relative abundance of bacteria as a metric of what defines a community
464 increasing the likelihood of identifying an in vitro model that might help design more effective

465 treatment regimens for a pwCF that has a similar community composition. Our findings raise the
466 question of what combinations of abundant microbes (*Pseudomonas*, *Streptococcus*,
467 *Staphylococcus*) and less abundant community members (i.e., anaerobic, fermentative
468 organisms) really matter? Does it matter if a patient has 90% *Pseudomonas* versus 80% or 70%?
469 Intriguingly, *Pseudomonas*, often listed as the quintessential CF pathogen, is highly abundant in
470 long lived people with CF, an aspect of a "survivor phenotype" (53). Alternatively, does it matter
471 if the dominant microbe in a community is *Streptococcus* rather than *Pseudomonas*, or if the
472 minor players are *Prevotella* versus *Veillonella* versus *Gemella*? Although unsupervised
473 machine learning cannot answer these questions with precision, we have demonstrated that as
474 few as five communities may represent a useful starting point as pwCF with a particular
475 community type experience significantly similar lung function as measured by $ppFEV_1$.
476
477 There are several critical factors and caveats to consider in light of the interpretations we make
478 here. First, although a gap statistic suggests five communities may be optimal for the cross-
479 sectional data set used here, one should avoid construing "five" as the number of distinct
480 populations in the CF community. Children younger than eight were not included in this study,
481 and it is possible that broad sampling of young pwCF would have enabled us to identify
482 additional clusters. In the same vein, although cluster 3 (Oth.D) contains subjects that were
483 dominated by genera other than *Pseudomonas,* it is understood that broader sampling of this
484 group would create additional clusters for each genus in cluster 3. Nonetheless, some of the
485 findings here mirror the conclusions from previous studies with smaller data sets (12, 15, 44).
486 Second, this study includes both cross-sectional and longitudinal observations of airway
487 microbiomes and associated clinical parameters, enabling us to identify associations between

488 them. However, this type of study cannot prove causation. Indeed, our interpretation of the data
489 here should rightly be viewed as hypotheses to be tested either via additional clinical studies or
490 in the laboratory. Third, and related to the previous point, the most useful aspect of this work is
491 that it does set a framework for hypothesis testing, for example, can we build the communities
492 we observe here *in vitro* and assess their function (i.e., are some clusters of microbes more/less
493 cytotoxic or more/less recalcitrant to antibiotics)? Alternatively, are similar communities
494 associated with positive/negative health outcomes in other cohorts of pwCF or in other disease
495 settings? For example, a recent report by Frey et. al. (29) concludes that airway diversity is a
496 biomarker of lung disease severity that links dysbiosis to lung function decline in pwCF. They
497 observed a negative correlation between inflammatory cytokine secretion and airway community
498 diversity, suggesting that more diverse communities are for some reason more salubrious than
499 less diverse communities (a common but unexplained finding across many microbiome studies).
500 Using a slightly different unsupervised approach to identify CF airway communities (i.e.,
501 hierarchical clustering of Unifrac distances), these investigators found that high diversity maps to
502 an oropharyngeal-like community typically found in younger people. The authors carefully avoid
503 making the claim that bacterial diversity reduces inflammation, although others have reported
504 this association (50) or found a positive association between higher lung function and increased
505 bacterial diversity (14, 17). The diverse, oropharyngeal-like community identified by Frey is
506 characterized by anaerobes like *Prevotella* and *Veillonella* and aerobic commensals including
507 *Neisseria* and *Streptococci*, making it very similar to the *Streptococcus* dominated community
508 (Str.SD) identified in this report (Figure 3); our Str.D community was also associated with higher
509 lung function (Figure 4A) and relative youth (Figure 4B) in pwCF. Thus, orthogonal approaches

510 appear to converge on similar conclusions as to community composition associated with less
511 severe health outcomes.
512
513 The findings here and reported by Frey et. al. (29) and others (14, 50, 51) raise the question of
514 whether remodeling CF airway communities to make them more diverse and more similar to the
515 oropharyngeal communities found in younger people with higher lung function, if possible,
516 might improve outcomes in CF. There are several reasons to question the wisdom of attempting
517 to turn back the hands of time in older subjects and enhance the abundance of oropharyngeal
518 flora in their airways. First, as described above, it is possible that the introduction of diversity
519 into an established community (for example, one dominated by *Pseudomonas*) would destabilize
520 the community and lead to enhanced inflammation, as suggested by the climax-attack model of
521 CF (54). Second, as demonstrated by Muhlebach (55), inflammation increases when pediatric
522 pwCF are first colonized by oropharyngeal flora; although this finding should not be surprising,
523 it is worth pointing out that the best airway community is probably no community at all. Third,
524 even in adults, increased diversity is not uniformly benign. Studies by Raghuvanshi et al. (24)
525 show a linear increase in the microbiome alpha-diversity and in the log ratio of anaerobes to
526 pathogens two weeks before a pulmonary exacerbation, confirming similar observations by
527 Carmody et al. (18) showing that bacterial community diversity and the relative abundance of
528 anaerobes increased during exacerbations, i.e., before antibiotic treatment. Others have shown
529 that treating adults with antibiotics primarily affects taxa other than classic CF pathogens (56,
530 57), and that antibiotics reduce the abundance of *Streptococcus*, *Prevotella* and *Veillonella,* but
531 not *Pseudomonas,* the ostensible target of treatment. Collectively then, the results cited support

532 the somewhat surprising conclusion that oropharyngeal flora are associated with worse airway
533 disease, and that treatments work because they reduce the abundance of oropharyngeal genera.
534
535 Finally, as mentioned above, we present just one approach for analyzing the extant microbial
536 community data associated with CF airway infections, but an approach that has both statistical
537 support and practical utility. That is, it is likely that only controlled experiments using model
538 communities in laboratory or animals can unravel the mechanistic role played by oropharyngeal-
539 like flora in an established adult microbiome (14, 25, 50, 58) and to understand the functional
540 consequences of community composition. That is, we have identified 5 communities comprised
541 of 10 genera, which presents an experimentally tractable numbers of communities and microbes.
542 As another example, Crabbe et al. (59) recently reported using a nine member community for in
543 vitro studies. Going forward, this and other studies can help inform the nature of laboratory
544 model communities used to probe the hypotheses we and others have developed by analyzing
545 clinical datasets.
546
547 Acknowledgements: Support for these studies was provided by the Cystic Fibrosis Foundation
548 (CFF STANTO19R0 and STANTO19GO) and NIH (R01HL151385, R37 AI083256-06, and P30
549 DK117469). We would like to thank Dr. John LiPuma for access to his team's data and many
550 helpful insights about this project.
551
552

553 **Table 1. Characteristics of 24 subjects with at least 10 longitudinal samples**.

isolate	Aggressiveness	N	Period (yr)	Median Age	Median FEV1	Clinical States	FEV1/yr
75	Moderate/Severe	16	8.99	14.9	32.5	R,E,T,T,E,R,B,B,R,T,E,B,B,R,B,R	-2.23
139	Moderate/Severe	15	6.64	15.1	34	B,B,B,T,R,T,B,B,B,B,R,B,T,T,R	-5.17
94	Moderate/Severe	13	9.16	17.2	39	T,T,E,E,T,B,E,B,B,B,B,R,R	1.56
229	Moderate/Severe	11	7.27	19.2	34	R,R,R,B,B,T,R,B,T,B,B	-4.50
52	Moderate/Severe	18	9.74	23.0	43.5	T,E,R,R,T,R,R,B,R,R,B,E,E,B,B,B,E,T,R,R	-1.84
179	Moderate/Severe	24	8.66	23.8	41	B,E,T,T,B,E,T,R,B,B,E,B,R,R,E,R,T,B,T,E,R,B,E,E	-3.08
159	Mild	11	7.97	24.3	82	B,B,B,E,B,B,B,B,B,B,R	-1.20
87	Mild	17	7.54	25.0	82	T,T,R,B,E,B,B,B,B,B,R,T,R,B,R,B,B	1.84
160	Mild	13	8.01	25.2	73	B,T,B,B,R,B,R,B,R,B,T,E,B	0.79
142	Moderate/Severe	20	10.14	25.4	67	B,T,T,T,T,E,B,R,T,T,E,E,R,B,T,E,B,R,T,E	-1.16
141	Mild	12	8.91	26.3	85.5	R,E,R,B,B,B,R,T,B,R,B,T	-0.20
124	Moderate/Severe	18	10.24	26.3	54.5	B,B,T,B,T,B,E,T,B,E,B,E,B,E,T,B,E,B,B	-1.00
147	Moderate/Severe	18	8.17	27.2	36.5	B,R,E,E,B,B,B,B,B,T,R,T,T,T,B,B,R,E	-3.73
209	Moderate/Severe	16	7.87	28.0	37.5	T,B,T,B,E,T,R,E,R,R,R,T,R,R,B,R	-6.97
53	Mild	10	5.35	28.6	69.5	E,R,T,E,B,E,B,E,B,NA,NA	1.55
182	Mild	14	5.69	29.0	83.5	T,B,R,E,E,B,B,B,B,B,B,E,E,E	-2.28
118	Mild	10	7.33	30.3	58.5	E,T,B,R,B,B,B,T,R,E	-1.33
244	Mild	10	6.42	30.8	74.5	E,R,B,B,R,B,B,B,E,B	-2.23
222	Mild	10	7.47	33.5	97	B,R,B,T,B,T,B,B,B,NA	-0.21
136	Mild	15	7.6	36.8	57	B,R,B,E,B,B,T,R,B,R,E,E,B,T,B	-2.99
119	Mild	14	7.72	37.1	36	E,B,T,B,B,B,B,T,B,B,B,R,B,B	-3.42
195	Mild	11	8.24	41.4	54	E,B,E,E,B,E,B,T,R,B,R	0.45
50	Mild	10	6.27	43.5	70.5	B,E,R,B,E,B,B,B,B,R	-2.27
117	Mild	20	9.72	49.6	51	E,E,B,R,B,E,B,E,T,T,B,T,B,E,B,E,T,T,B,B	-0.31

554 [a]Disease aggressiveness for each subject was based on publicly available metadata.

555 [b]N represents the number of sputum microbiomes available for each subject, observed over the

556 period, given in years.

557 [c]Clinical states observed: B = baseline, E = exacerbation, T = treatment, R = Recovery.

558 [d]Linear regression estimates of annual change in FEV1 percent predicted for each subject shown

559 at far right. The table is sorted by increasing median age.

560

561

562 ## References

563

564 1. **MacKenzie T**, **Gifford AH**, **Sabadosa KA**, **Quinton HB**, **Knapp EA**, **Goss CH**,
565 **Marshall BC**. 2014. Longevity of Patients with Cystic Fibrosis in 2000 to 2010 and
566 Beyond: Survival Analysis of the Cystic Fibrosis Foundation Patient Registry. Ann
567 Intern Med **161**:233–241.

568 2. **Emerson J**, **Rosenfeld M**, **McNamara S**, **Ramsey B**, **Gibson RL**. 2002.
569 *Pseudomonas aeruginosa* and Other Predictors of Mortality and Morbidity in Young
570 Children with Cystic Fibrosis. Pediatr Pulmonol **34**:91–100.

571 3. **Jorth P**, **Ehsan Z**, **Rezayat A**, **Caldwell E**, **Pope C**, **Brewington JJ**, **Goss CH**,
572 **Benscoter D**, **Clancy JP**, **Singh PK**. 2019. Direct Lung Sampling Indicates that
573 Established Pathogens Dominate Early Infections in Children with Cystic Fibrosis.
574 Cell Reports **27**:1190–1204.e3.

575 4. **Salsgiver EL**, **Fink AK**, **Knapp EA**, **LiPuma JJ**, **Olivier KN**, **Marshall BC**,
576 **Saiman L**. 2016. Changing Epidemiology of the Respiratory Bacteriology of Patients
577 with Cystic Fibrosis. Chest **149**:390–400.

578 5. **Rudkjøbing VB**, **Thomsen TR**, **Alhede M**, **Kragh KN**, **Nielsen PH**, **Johansen UR**,
579 **Givskov M**, **Høiby N**, **Bjarnsholt T**. 2012. The Microorganisms in Chronically
580 Infected End-Stage and Non-End-Stage Cystic Fibrosis Patients. FEMS Immunology
581 & Medical Microbiology **65**:236–244.

582 6. **Govan JRW**, **Nelson JW**. 1992. Microbiology of Lung Infection in Cystic Fibrosis.
583 British Medical Bulletin **48**:912–930.

584 7. **Burns JL**, **Emerson J**, **Stapp JR**, **Yim DL**, **Krzewinski J**, **Louden L**, **Ramsey BW**,
585 **Clausen CR**. 1998. Microbiology of Sputum from Patients at Cystic Fibrosis Centers
586 in the United States. Clinical Infectious Diseases **27**:158–163.

587 8. **Saiman L**, **Siegel J**. 2004. Infection Control in Cystic Fibrosis. Clinical Microbiology
588 Reviews **17**:57–71.

589 9. **Konstan MW**, **Butler SM**, **Schidlow DV**, **Morgan WJ**, **Julius JR**. 1999. Patterns of
590 Medical Practice in Cystic Fibrosis: Part II. Use of Therapies. Pediatr Pulmonol
591 **28**:248–254.

592 10. **Sibley CD**, **Rabin H**, **Surette MG**. 2006. Cystic fibrosis: A Polymicrobial Infectious
593 Disease. http://dxdoiorg/102217/174609131153 **1**:53–61.

594 11. **Harris JK**, **De Groote MA**, **Sagel SD**, **Zemanick ET**, **Kapsner R**, **Penvari C**, **Kaess**
595 **H**, **Deterding RR**, **Accurso FJ**, **Pace NR**. 2007. Molecular Identification of Bacteria
596 in Bronchoalveolar Lavage Fluid from Children with Cystic Fibrosis. Proc Natl Acad
597 Sci USA **104**:20529–20533.

598 12. **Filkins LM**, **Hampton TH**, **Gifford AH**, **Gross MJ**, **Hogan DA**, **Sogin ML**,
599 **Morrison HG**, **Paster BJ**, **O'Toole GA**. 2012. Prevalence of Streptococci and
600 Increased Polymicrobial Diversity Associated with Cystic Fibrosis Patient Stability.
601 Journal of Bacteriology **194**:4709–4717.

602 13. **O'Toole GA**. 2018. Cystic Fibrosis Airway Microbiome: Overturning the Old,
603 Opening the Way for the New. Journal of Bacteriology **200**:344.

604 14. **Acosta N**, **Heirali A**, **Somayaji R**, **Surette MG**, **Workentine ML**, **Sibley CD**, **Rabin**
605 **HR**, **Parkins MD**. 2018. Sputum Microbiota Is Predictive of Long-Term Clinical
606 Outcomes in Young Adults with Cystic Fibrosis. Thorax **73**:1016–1025.

607 15. **Zhao J**, **Schloss PD**, **Kalikin LM**, **Carmody LA**, **Foster BK**, **Petrosino JF**,
608 **Cavalcoli JD**, **VanDevanter DR**, **Murray S**, **Li JZ**, **Young VB**, **LiPuma JJ**. 2012.
609 Decade-Long Bacterial Community Dynamics in Cystic Fibrosis Airways. Proceedings
610 of the National Academy of Sciences **109**:5809–5814.

611 16. **van der Gast CJ**, **Walker AW**, **Stressmann FA**, **Rogers GB**, **Scott P**, **Daniels TW**,
612 **Carroll MP**, **Parkhill J**, **Bruce KD**. 2011. Partitioning Core and Satellite Taxa from
613 within Cystic Fibrosis Lung Bacterial Communities. The ISME Journal **5**:780–791.

614 17. **Cuthbertson L**, **Walker AW**, **Oliver AE**, **Rogers GB**, **Rivett DW**, **Hampton TH**,
615 **Ashare A**, **Elborn JS**, **De Soyza A**, **Carroll MP**, **Hoffman LR**, **Lanyon C**,
616 **Moskowitz SM**, **O'Toole GA**, **Parkhill J**, **Planet PJ**, **Teneback CC**, **Tunney MM**,
617 **Zuckerman JB**, **Bruce KD**, **van der Gast CJ**. 2020. Lung Function and Microbiota
618 Diversity in Cystic Fibrosis. Microbiome, 4 ed. **8**:45–13.

619 18. **Carmody LA**, **Caverly LJ**, **Foster BK**, **Rogers MAM**, **Kalikin LM**, **Simon RH**,
620 **VanDevanter DR**, **LiPuma JJ**. 2018. Fluctuations in Airway Bacterial Communities
621 Associated with Clinical States and Disease Stages in Cystic Fibrosis. PLoS ONE
622 **13**:e0194060.

623 19. **Goss CH**, **Burns JL**. 2007. Exacerbations in Cystic Fibrosis · 1: Epidemiology and
624 Pathogenesis. Thorax **62**:360–367.

625 20. **Flume PA**, **Peter J Mogayzel J**, **Robinson KA**, **Goss CH**, **Rosenblatt RL**, **Kuhn RJ**,
626 **Marshall BC**, **Committee ATCPGFPT**. 2012. Cystic Fibrosis Pulmonary Guidelines.
627 Am J Respir Crit Care Med **180**:802–808.

628 21. **Flume PA**, **Suthoff ED**, **Kosinski M**, **Marigowda G**, **Quittner AL**. 2019. Measuring
629 Recovery in Health-Related Quality Of Life During and After Pulmonary
630 Exacerbations in Patients with Cystic Fibrosis. Journal of Cystic Fibrosis **18**:737–742.

631 22. **Wagener JS**, **VanDevanter DR**, **Konstan MW**, **Pasta DJ**, **Millar SJ**, **Morgan WJ**.
632 2020. Lung Function Changes Before and After Pulmonary Exacerbation
633 Antimicrobial Treatment in Cystic Fibrosis. Pediatr Pulmonol **55**:828–834.

634 23. **Heirali A**, **Thornton C**, **Acosta N**, **Somayaji R**, **Laforest Lapointe I**, **Storey D**,
635 **Rabin H**, **Waddell B**, **Rossi L**, **Arrieta MC**, **Surette M**, **Parkins MD**. 2020. Sputum
636 Microbiota in Adults with CF Associates with Response io Inhaled Tobramycin.
637 Thorax **5**:thoraxjnl–2019–214191.

638 24. **Raghuvanshi R**, **Vasco K**, **Vázquez-Baeza Y**, **Jiang L**, **Morton JT**, **Li D**, **Gonzalez**
639 **A**, **DeRight Goldasich L**, **Humphrey G**, **Ackermann G**, **Swafford AD**, **Conrad D**,
640 **Knight R**, **Dorrestein PC**, **Quinn RA**. 2020. High-Resolution Longitudinal Dynamics
641 of the Cystic Fibrosis Sputum Microbiome and Metabolome through Antibiotic
642 Therapy. mSystems **5**:5809.

643 25. **Sibley CD**, **Parkins MD**, **Rabin HR**, **Duan K**, **Norgaard JC**, **Surette MG**. 2008. A
644 Polymicrobial Perspective of Pulmonary Infections Exposes an Enigmatic Pathogen in
645 Cystic Fibrosis Patients. Proceedings of the National Academy of Sciences
646 **105**:15070–15075.

647 26. **Parkins MD**, **Sibley CD**, **Surette MG**, **Rabin HR**. 2008. The *Streptococcus milleri*
648 Group—An Unrecognized Cause of Disease in Cystic Fibrosis: A Case Series and
649 Literature Review. Pediatr Pulmonol **43**:490–497.

650 27. **Scott JE**, **O'Toole GA**, **Margolin W**. 2019. The Yin and Yang of *Streptococcus* Lung
651 Infections in Cystic Fibrosis: A Model for Studying Polymicrobial Interactions.
652 Journal of Bacteriology **201**:15070.

653 28. **Quinn RA**, **Whiteson K**, **Lim YW**, **Zhao J**, **Conrad D**, **LiPuma JJ**, **Rohwer F**,
654 **Widder S**. 2016. Ecological Networking of Cystic Fibrosis Lung Infections. NPJ
655 Biofilms Microbiomes **2**:4–11.

656 29. **Frey DL**, **Boutin S**, **Dittrich SA**, **Graeber SY**, **Stahl M**, **Wege S**, **Herth FJF**,
657 **Sommerburg O**, **Schultz C**, **Mall MA**, **Dalpke AH**. 8 January 2021 posting date.
658 Relationship Between Airway Dysbiosis, Inflammation and Lung Function in Adults
659 with Cystic Fibrosis. J Cyst Fibros. doi: 10.1016/j.jcf.2020.12.022.

660 30. **Hahn A**, **Burrell A**, **Fanous H**, **Chaney H**, **Sami I**, **Perez GF**, **Koumbourlis AC**,
661 **Freishtat RJ**, **Crandall KA**. 2018. Antibiotic Multidrug Resistance in the Cystic
662 Fibrosis Airway Microbiome is Associated with Decreased Diversity. Heliyon
663 **4**:e00795.

664 31. **Coburn B**, **Wang PW**, **Caballero JD**, **Clark ST**, **Brahma V**, **Donaldson S**, **Zhang**
665 **Y**, **Surendra A**, **Gong Y**, **Tullis DE**, **Yau YCW**, **Waters VJ**, **Hwang DM**, **Guttman**
666 **DS**. 2015. Lung Microbiota Across Age and Disease Stage in Cystic Fibrosis. Sci Rep
667 **5**:1–12.

668 32. **Cuthbertson L**, **Rogers GB**, **Walker AW**, **Oliver A**, **Green LE**, **Daniels TWV**,
669 **Carroll MP**, **Parkhill J**, **Bruce KD**, **van der Gast CJ**. 2016. Respiratory Microbiota
670 Resistance and Resilience to Pulmonary Exacerbation and Subsequent Antimicrobial
671 Intervention. The ISME Journal **10**:1081–1091.

672 33. **Sherrard LJ**, **Einarsson GG**, **Johnston E**, **O'Neill K**, **McIlreavey L**, **McGrath SJ**,
673 **Gilpin DF**, **Downey DG**, **Reid A**, **McElvaney NG**, **Boucher RC**, **Muhlebach MS**,
674 **Elborn JS**, **Tunney MM**. 2019. Assessment of Stability and Fluctuations of Cultured
675 Lower Airway Bacterial Communities in People with Cystic Fibrosis. J Cyst Fibros
676 **18**:808–816.

677 34. **Bolyen E**, **Rideout JR**, **Dillon MR**, **Bokulich NA**, **Abnet CC**, **Al-Ghalith GA**,
678 **Alexander H**, **Alm EJ**, **Arumugam M**, **Asnicar F**, **Bai Y**, **Bisanz JE**, **Bittinger K**,
679 **Brejnrod A**, **Brislawn CJ**, **Brown CT**, **Callahan BJ**, **Caraballo-Rodríguez AM**,
680 **Chase J**, **Cope EK**, **Da Silva R**, **Diener C**, **Dorrestein PC**, **Douglas GM**, **Durall**
681 **DM**, **Duvallet C**, **Edwardson CF**, **Ernst M**, **Estaki M**, **Fouquier J**, **Gauglitz JM**,
682 **Gibbons SM**, **Gibson DL**, **Gonzalez A**, **Gorlick K**, **Guo J**, **Hillmann B**, **Holmes S**,
683 **Holste H**, **Huttenhower C**, **Huttley GA**, **Janssen S**, **Jarmusch AK**, **Jiang L**,
684 **Kaehler BD**, **Bin Kang K**, **Keefe CR**, **Keim P**, **Kelley ST**, **Knights D**, **Koester I**,
685 **Kosciolek T**, **Kreps J**, **Langille MGI**, **Lee J**, **Ley R**, **Liu Y-X**, **Loftfield E**,

686 **Lozupone C**, **Maher M**, **Marotz C**, **Martin BD**, **McDonald D**, **McIver LJ**, **Melnik**
687 **AV**, **Metcalf JL**, **Morgan SC**, **Morton JT**, **Naimey AT**, **Navas-Molina JA**, **Nothias**
688 **LF**, **Orchanian SB**, **Pearson T**, **Peoples SL**, **Petras D**, **Preuss ML**, **Pruesse E**,
689 **Rasmussen LB**, **Rivers A**, **Robeson MS**, **Rosenthal P**, **Segata N**, **Shaffer M**, **Shiffer**
690 **A**, **Sinha R**, **Song SJ**, **Spear JR**, **Swafford AD**, **Thompson LR**, **Torres PJ**, **Trinh P**,
691 **Tripathi A**, **Turnbaugh PJ**, **Ul-Hasan S**, **van der Hooft JJJ**, **Vargas F**, **Vázquez-**
692 **Baeza Y**, **Vogtmann E**, **Hippel von M**, **Walters W**, **Wan Y**, **Wang M**, **Warren J**,
693 **Weber KC**, **Williamson CHD**, **Willis AD**, **Xu ZZ**, **Zaneveld JR**, **Zhang Y**, **Zhu Q**,
694 **Knight R**, **Caporaso JG**. 2019. Reproducible, Interactive, Scalable and Extensible
695 Microbiome Data Science Using QIIME 2. Nat Biotechnol **37**:852–857.

696 35. **Goslee SC**, **Urban DL**. 2007. The Ecodist Package for Dissimilarity-Based Analysis
697 of Ecological Data. Journal of Statistical Software, 22(7):1–19.

698 36. **Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre**
699 **Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson,**
700 **Peter Solymos, M. Henry, H. Stevens, Eduard Szoecs and Helene Wagner**. 2020.
701 vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-
702 project.org/package=vegan.

703 37. **Kassambara A**, **Mundt F**. 2020. factoextra: Extract and Visualize the Results of
704 Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-
705 project.org/package=factoextra.

38. **Andy Liaw and Matthew Wiener**. 2002. Classification and Regression By randomForest. R package version 4.6-14. https://CRAN.R-project.org/package=randomForest

39. **Giorgio Alfredo Spedicato**. 2017. Discrete Time Markov Chains with R. R package version 0.8.5-4. https://CRAN.R-project.org/package=markovchain

40. **Bray JR**, **Curtis JT**. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs **27**:325–349.

41. **Namkung J**. 2020. Machine Learning Methods for Microbiome Studies. J Microbiol **58**:206–216.

42. **Sneath, P., Sokal, R.** 1962. Numerical Taxonomy**.** Nature**. 193**: 855–860

43. **Madan JC**, **Koestler DC**, **Stanton BA**, **Davidson L**, **Moulton LA**, **Housman ML**, **Moore JH**, **Guill MF**, **Morrison HG**, **Sogin ML**, **Hampton TH**, **Karagas MR**, **Palumbo PE**, **Foster JA**, **Hibberd PL**, **O'Toole GA**. 2012. Serial Analysis of the Gut and Respiratory Microbiome in Cystic Fibrosis in Infancy: Interaction Between Intestinal and Respiratory Tracts and Impact of Nutritional Exposures. mBio **3**:e00251–12–e00251–12.

44. **Price KE**, **Hampton TH**, **Gifford AH**, **Dolben EL**, **Hogan DA**, **Morrison HG**, **Sogin ML**, **O'Toole GA**. 2013. Unique Microbial Communities Persist in Individual Cystic Fibrosis Patients throughout a Clinical Exacerbation. Microbiome **1**:27–1.

725 45. **Harris JK**, **Wagner BD**, **Zemanick ET**, **Robertson CE**, **Stevens MJ**, **Heltshe SL**,
726 **Rowe SM**, **Sagel SD**. 2020. Changes in Airway Microbiome and Inflammation with
727 Ivacaftor Treatment in Patients with Cystic Fibrosis and the G551D Mutation. Annals
728 of the American Thoracic Society **17**:212–220.

729 46. **Tibshirani R**, **Walther G**, **Hastie T**. 2001. Estimating the Number of Clusters in a
730 Data Set Via the Gap Statistic. Journal of the Royal Statistical Society: Series B
731 (Statistical Methodology) **63**:411–423.

732 47. **Tibshirani R**, **Walther G**. 2012. Cluster Validation by Prediction Strength. Journal of
733 Computational and Graphical Statistics.

734 48. **Stearns JC**, **Davidson CJ**, **McKeon S**, **Whelan FJ**, **Fontes ME**, **Schryvers AB**,
735 **Bowdish DME**, **Kellner JD**, **Surette MG**. 2015. Culture and Molecular-Based
736 Profiles Show Shifts in Bacterial Communities of the Upper Respiratory Tract that
737 Occur with Age. The ISME Journal **9**:1246–1259.

738 49. **Stressmann FA**, **Rogers GB**, **van der Gast CJ**, **Marsh P**, **Vermeer LS**, **Carroll MP**,
739 **Hoffman L**, **Daniels TWV**, **Patel N**, **Forbes B**, **Bruce KD**. 2012. Long-Term
740 Cultivation-Independent Microbial Diversity Analysis Demonstrates that Bacterial
741 Communities Infecting the Adult Cystic Fibrosis Lung Show Stability and Resilience.
742 Thorax **67**:867–873.

743 50. **Zemanick ET**, **Wagner BD**, **Robertson CE**, **Ahrens RC**, **Chmiel JF**, **Clancy JP**,
744 **Gibson RL**, **Harris WT**, **Kurland G**, **Laguna TA**, **McColley SA**, **McCoy K**,
745 **Retsch-Bogart G**, **Sobush KT**, **Zeitlin PL**, **Stevens MJ**, **Accurso FJ**, **Sagel SD**,

746 **Harris JK**. 2017. Airway Microbiota Across Age and Disease Spectrum in Cystic
747 Fibrosis. Eur Respir J **50**:1700832.

748 51. **Flight WG**, **Smith A**, **Paisey C**, **Marchesi JR**, **Bull MJ**, **Norville PJ**, **Mutton KJ**,
749 **Webb AK**, **Bright-Thomas RJ**, **Jones AM**, **Mahenthiralingam E**, **McAdam AJ**.
750 2015. Rapid Detection of Emerging Pathogens and Loss of Microbial Diversity
751 Associated with Severe Lung Disease in Cystic Fibrosis. Journal of Clinical
752 Microbiology **53**:2022–2029.

753 52. **Breiman L**. 2001. Random Forests. Machine Learning **45**:5–32.

754 53. **Conrad DJ**, **Billings J**, **Teneback C**, **Koff J**, **Rosenbluth D**, **Bailey BA**, **Jain R**.
755 2021. Multi-Dimensional Clinical Phenotyping of a National Cohort of Adult Cystic
756 Fibrosis Patients. Journal of Cystic Fibrosis **20**:91–96.

757 54. **Conrad D**, **Haynes M**, **Salamon P**, **Rainey PB**, **Youle M**, **Rohwer F**. 2013. Cystic
758 Fibrosis Therapy: A Community Ecology Perspective. Am J Respir Cell Mol Biol
759 **48**:150–156.

760 55. **Muhlebach MS**, **Zorn BT**, **Esther CR**, **Hatch JE**, **Murray CP**, **Turkovic L**,
761 **Ranganathan SC**, **Boucher RC**, **Stick SM**, **Wolfgang MC**. 2018. Initial Acquisition
762 and Succession of The Cystic Fibrosis Lung Microbiome is Associated with Disease
763 Progression in Infants and Preschool Children. PLoS Pathog **14**:e1006798.

764 56. **Nelson MT**, **Wolter DJ**, **Eng A**, **Weiss EJ**, **Vo AT**, **Brittnacher MJ**, **Hayden HS**,
765 **Ravishankar S**, **Bautista G**, **Ratjen A**, **Blackledge M**, **McNamara S**, **Nay L**,
766 **Majors C**, **Miller SI**, **Borenstein E**, **Simon RH**, **LiPuma JJ**, **Hoffman LR**. 2020.

767 Maintenance Tobramycin Primarily Affects Untargeted Bacteria in the CF Sputum
768 Microbiome. Thorax **75**:780–790.

769 57. **Heirali AA**, **Acosta N**, **Storey DG**, **Workentine ML**, **Somayaji R**, **Laforest**
770 **Lapointe I**, **Leung W**, **Quon BS**, **Berthiaume Y**, **Rabin HR**, **Waddell BJ**, **Rossi L**,
771 **Surette MG**, **Parkins MD**. 2019. The Effects ff Cycled Inhaled Aztreonam on the
772 Cystic Fibrosis (CF) Lung Microbiome. J Cyst Fibros **18**:829–837.

773 58. **Grinwis ME**, **Sibley CD**, **Parkins MD**, **Eshaghurshan CS**, **Rabin HR**, **Surette MG**.
774 2010. Characterization of *Streptococcus milleri* Group Isolates from Expectorated
775 Sputum of Adult Patients with Cystic Fibrosis. Journal of Clinical Microbiology
776 **48**:395–401.

777 59. **Vandeplassche E**, **Sass A**, **Ostyn L**, **Burmølle M**, **Kragh KN**, **Bjarnsholt T**, **Coenye**
778 **T**, **Crabbé A**. 2020. Antibiotic Susceptibility of Cystic Fibrosis Lung Microbiome
779 Members in a Multispecies Biofilm. Biofilm **2**:100031.