Machine learning for exploring microbial inter-kingdom associations in Cystic Fibrosis and Bronchiectasis

L. Cuthbertson*[1], J. Ish-Horowicz*[4], I.C. Felton[2], P. James[1,3], E. Turek[1], M.J. Cox[1], M.R. Loebinger[1,2], N.J. Simmonds[1,2], S. Filippi [4], M.F. Moffatt**[1], W.O.C. Cookson**[1,2]

[1]National Heart and Lung Institute Imperial College London, Genomic Medicine, London, United Kingdom, [2]Royal Brompton & Harefield NHS Trust, London, United Kingdom, [3]Oxford Nanopore Technologies Ltd, Oxford, United Kingdom, [4]Department of Mathematics, Imperial College London, London, United Kingdom

* Authors contributed equally to this manuscript

** Joint senior authors

Corresponding author: L. Cuthbertson, National Heart and Lung Institute, Imperial College London, London SW3 6LY, UK.

E-mail: l.cuthbertson@imperial.ac.uk

Abstract

Background

Cystic fibrosis (CF) and non-CF bronchiectasis (BX) are lung diseases characterised by severe chronic infections. Fungal and bacterial components of infection are both recognized. Recent molecular investigation of sputum from patients with CF and BX has revealed a complex mycobiome. However, little is known about how fungal and bacterial organisms interact or whether the interactions impact on disease outcomes.

Methods

Quantitative PCR and next generation sequencing of ITS2 and 16S rRNA gene was carried out on 107 patients with CF and BX and defined clinical fungal infection status. Fungal and bacterial communities were explored using supervised and unsupervised machine learning to understand associations between fungal and bacterial communities and their relationship to disease.

Results

Fungal and bacterial communities both had significantly higher biomass and lower diversity in CF compared to BX patients. Random forest modelling demonstrated that the fungal and bacterial communities were distinct between CF and BX patients. Within the CF group, bacterial communities contained no predictive signal for fungal disease status. Neither bacterial nor fungal community composition were predictive of the presence of CF pulmonary exacerbation (CFPE). Intra-kingdom correlations were far stronger than those between the two kingdoms. Dirichlet mixture components analysis identified two distinct clusters of bacteria related to the relative abundance of *Pseudomonas*. Fungal community composition contained no predictive signal for bacterial clusters.

Conclusions

Clear changes in diversity were observed between patients with different clinical disease status. Although our results demonstrate that bacterial community composition differs in the presence of fungal disease, no direct relationship between bacterial and fungal OTUs was found.

Keywords: (10 words)

Mycobiome, Microbiome, machine learning, random forest, cystic fibrosis, bronchiectasis

Introduction

Chronic respiratory infections are the leading causes of morbidity and mortality for the chronic suppurative lung diseases (CSLD); cystic fibrosis (CF) and non-CF bronchiectasis (BX) [1, 2]. Bacterial infections are a major pathophysiological factor in disease progression in these patients [3, 4]. The impact of fungal infections is, however, increasingly being recognised. Fungal infections have been associated with higher disease burdens, increased exacerbation rates and accelerated clinical decline [5, 6].

The development of robust microbial sequencing protocols has revealed complex fungal communities within the lungs of patients with CF and BX, exhibiting a range of clinical disease manifestations [7, 8]. The recognition of these fungal communities has led to the investigation of bacterial and fungal interactions and their roles in disease progression.

Insights into these complex associations and interactions are helped through an evolving statistical methodology. A recent publication by Soret *et al.*, used adapted

Phy-Lasso models to explore the inter-kingdom relationships associated with CF pulmonary exacerbations (CFPE)[8]. The study suggested that fungal genera previously linked to decreasing lung function and poor clinical outcomes were also associated with exacerbations [9, 10].

Machine learning techniques, such as those used by Soret *et al.*, have proven to be effective for host phenotype prediction from microbiome profiles [8]. Studies by Zhou *et al.* and Topçuoğlu *et al.* have reported that random forest modelling achieved the best predictive performance for clinical phenotypes on a range of human gut 16S rRNA gene sequencing datasets, beating penalised linear models and support vector machines [11, 12].

Random forests (RFs) are popular in bioinformatics as they are non-parametric, non-linear and are able to achieve strong predictive performance on many types of data [13]. Furthermore, they often offer improved predictive performance relative to linear models due to their ability to model interactions between variables. In addition, random forests offer a degree of interpretability via variable importance measures such as Mean Decrease Gini and Mean Decrease Accuracy. This has resulted in random forests being used for association testing in genome-wide association studies [14] and biomarker identification in gene microarray studies [15].

In the field of microbiome studies random forests have been used to identify important taxa using both 16S rRNA gene (bacterial) [16, 17] and ITS2 (fungal) sequence data [18, 19]. Recognising the microbial predictors of disease status in chronic respiratory disease may prevent disease progression and allow the development of informed treatment regimens improving patient outcomes.

100 Therefore, in this present study we explored the bacterial and fungal communities
101 in CSLD. Using random forest modelling we aimed to explore the power of
102 bacterial and fungal taxa to predict the presence of CF (n=83) or BX (n=24). Within
103 subjects with CF, we also sought predictors of clinical diagnoses of fungal disease
104 (fungal bronchitis, FB) and the presence of CF pulmonary exacerbations (CFPE).
105 Because of the small number of patients with BX, we confined subgroup analyses
106 of fungal disease or exacerbations to CF patients only (Table 1).
107
108 **Table 1. Patient demographics**. Data for all samples included in this manuscript,
109 only samples with paired fungal and bacterial sequencing that passed all quality
110 control steps (see supplementary information) were taken forward for analysis.
111 Differences in continuous variables were calculated using a one-way t-test, while
112 a Chi-square test was used to explore categorical variables.
113

	BX (n = 24)	**CF** (n = 83)	*P*
Age/yrs (mean (SD))	59.79 (11.06)	32.00 (11.90)	<0.001
Male (%)	10 (41.7)	47 (56.6)	0.288
BMI (mean (SD))	26.51 (8.73)	22.51 (6.50)	0.029
FEV^1 % predicted (mean (SD))	58.24 (25.23)	48.81 (17.79)	0.047
Disease Group (%)			<0.001
ABPA	8 (33.3)	14 (16.9)	
CNPA	5 (20.8)	0 (0.0)	
FB	0 (0.0)	20 (24.1)	
NAFD	8 (33.3)	39 (47.0)	
NTM	3 (12.5)	10 (12.0)	
Exacerbation* (%)	1 (4.2)	36 (43.4)	0.001

114 CF: Cystic fibrosis, BX: Non-CF bronchiectasis, BMI: Body mass index, FEV^1: Forced
115 expiratory volume in 1 second, ABPA: Allergic bronchopulmonary aspergillosis,
116 CNPA: Chronic necrotising pulmonary aspergillosis, FB: Fungal bronchitis, NAFD:
117 No active fungal disease, NTM: Non-tuberculosis mycobacteria.

118 * Patients were defined as exacerbating by treating clinician as a change in clinical
119 status from baseline that results in a new pulmonary antibiotic treatment.
120
121 **Methods**
122 ***Study Design***
123 A prospective, cross-sectional study of spontaneously expectorated sputum from
124 adults with CF and BX was carried out at the Royal Brompton NHS trust between
125 April 2013 and July 2014 [7]. Ethical approval was obtained from the Royal
126 Brompton and Harefield Hospital Biomedical Research Unit ethics committee
127 (Advanced Lung Disease Biobank study number: 10/H0504/9).
128 Patient details, including treatments were collected and subjects were partitioned
129 into defined clinical subgroups according to the diagnostic criteria defined in
130 Cuthbertson *et al*. 2020 [7](Table 1).
131 Sputum samples were collected and processed as previously described with half
132 the sample sent for routine clinical microbiological culture and the other half
133 stored at -80 °C for DNA analysis [7]. DNA extraction was performed using the
134 DNA fast spin kit for soil (MPBio, California, USA) according to the manufacturer's
135 instructions. Extraction controls were blinded and processed along with patient
136 samples [7].
137
138 ***Quantitative PCR***
139 Bacterial biomass was quantified by SYBR green quantitative PCR (qPCR) [20].
140 Fungal biomass was estimated using a modified Taqman based qPCR assay as
141 previously described [7]. All qPCR reactions were performed in triplicate.
142

143 ***DNA sequencing***

144 16S rRNA gene and ITS2 sequencing were performed on the Illumina MiSeq
145 platform using dual barcode fusion primers. Bacterial sequencing was performed
146 on the V4 region of the 16S rRNA gene as previously described [20]. ITS2
147 sequencing was performed using the primers, ITS2F (5'-CAR CAA YGG ATC TCT
148 TGG-3') and ITS2R (5'-GAT ATG CTT AAG TTC AGC GGG T-3') with ligated adaptors
149 [20]. Extraction controls, PCR negative and mock communities were included on
150 all sequencing runs.

151 Sequence processing of 16S rRNA gene and ITS2 data was carried out using QIIME
152 1.9 as described previously in Cuthbertson *et al*. 2017 [20] and Cuthbertson *et al*.
153 2020 [7] respectively.

154 All sequences were submitted to the European nucleotide database. Bacterial data
155 can be accessed under project number PRJEB33064 and the fungal data is
156 accessible under project number PRJEB33434.

157

158 ***Statistical analysis***

159 Statistical analysis was carried out in R version 3.5.1. Data was analysed in
160 phyloseq version 1.24.2 [21]. Decontamination of the data was carried out using
161 decontam version 1.1.2 [22]. Full details are available in Supplementary Materials.
162 Differences between categorical variables were calculated using Wilcoxon rank
163 sum test and Kruskal-Wallis. Pearson correlation was used for tests between
164 continuous variables. Differences in microbial community composition were
165 tested with PERMANOVA using the Adonis function from vegan version 2.5-6 [23].
166 Random forests models were fitted using the caret (version 6.0) and ranger
167 (version 0.12) packages [24, 25]. Differential abundance was performed using the

168 ALDEx2 (version 1.26) [26], while the Dirichlet Multinomial package (version
169 1.36) was used for sample-wise clustering [27].
170 All *P*-values were adjusted for multiple corrections using false discovery rate
171 (FDR).

172 ***Predictive modelling using random forests***

173 Random forests were used to investigate associations between microbial
174 community composition and the three clinical labels given and described in Table
175 2. The covariates for these models were the relative abundance of any bacterial
176 genera or fungal species with prevalence over 20% across all 107 samples. Taxa
177 accounting for fewer than 0.0001% of reads were also removed, with the
178 remaining 20 fungal species and 45 bacterial taxa accounting for 99.6% and 96.8%
179 of bacterial and fungal total reads respectively. Whittaker plots (Figure S1) show
180 that fungal species and bacterial genus provide similar resolutions with which to
181 identify microbes. Raw counts were normalised by dividing by the total number
182 of reads per sample before being transformed using log(x+1) for the random
183 forest analysis.

184

185 **Table 2.** Phenotypes investigated in this study using random forest. Phenotypes
186 were chosen to maximise sample sizes and power of the analyses.
187 The Group phenotype was limited to the FB (n = 20) and NAFD (n = 39) groups
188 within the CF population. BX samples were removed from the group analysis due
189 to the low numbers.

190

Phenotype	Description	Classes	n
Disease	Whether a patient has CF or BX	CF or BX	83 CF, 24 BX (107 total)

Group	Whether a CF patient has been diagnosed with FB or NAFD	FB or NAFD	20 FB, 39 NAFD (59 total)
CFPE	Whether a CF patient was experiencing pulmonary exacerbation when the sample was collected (clinician defined).	Yes or No	36 Yes, 47 No (83 total)

CF: cystic fibrosis, BX: non-cystic fibrosis bronchiectasis, FB: fungal bronchitis, NAFD: no active fungal disease.

The predictive performance of each RF model was estimated using 5-fold nested cross-validation with 5 folds in the inner loop. The inner loop performed model selection using a random hyperparameter search of ten combinations of mtry and splitrule. Folds were sampled in a stratified manner meaning that the class proportions in each fold reflected the proportions of the entire dataset. The predictive performance of the models were evaluated using area under receiver-operating characteristic curves (AUC) and 95% confidence intervals on the mean AUC were calculated using the method of LeDell *et al*. [28].

Random forest variable importance and differential abundance analysis

Phenotype-taxa associations were quantified using the debiased Mean Decrease Gini (MDG) variable importance measure proposed by Nembrini *et al*. [29]. Uncorrected MDG scores are known to be artificially inflated in many settings and this increases the likelihood of false positive associations [30]. Each random forest was grown using 5,000 trees to achieve stability in the variable importance scores [31, 32]. The statistical significance of each score was assessed using *P*-values

211 calculated using the permutation method of Altmann *et al*. with 5,000
212 permutations [33]. Random forest variable importance scores are unsigned (they
213 do not give an effect direction) although an approximate effect direction, using
214 partial dependence plots which visualise the marginal effect of a variable on the
215 predictions of a model, can be obtained [34].

216 ALDEx2 effect sizes and Welch's t-tests were used to complement the random
217 forest scores for each taxa having differential abundance. Unlike random forest
218 ALDEx2 explicitly models the compositional nature of sequencing data [35].
219 Combining multiple computational tools with different underlying assumptions
220 has been shown to lead to more accurate conclusions in various biological
221 contexts by reducing the chance of findings being due to statistical or technical
222 artefacts [36, 37].

223

224 **Correlation analysis and sample-wise clustering**

225 The correlation structure amongst the taxa included in the RF analysis was also
226 analysed. Samples were Centre Log Ratio (CLR)-transformed prior to calculating
227 correlations to account for the compositional nature of 16S rRNA gene and ITS2
228 data. The CLR transformation was applied separately to the 16S rRNA gene and
229 ITS2 samples.

230

231

232 **Results**

233 ***Description of data***

234 After decontamination and removal of samples with less than 2,000 reads (n = 2),
235 107 samples were retained for downstream analyses (for demographics see Table
236 1). For analyses based on relative abundances, bacterial reads were rarefied to
237 2,357 while fungal reads were rarefied to 2,542.
238
239 ***Ecological analysis***
240 *Differences between diseases*
241 Both bacterial and fungal diversity were significantly higher in patients with BX
242 than CF (Bacteria; richness, W = 1,507, effect size = 0.369, $P < 0.001$, Shannon, W
243 = 1,545, effect size = 0.396,$P < 0.001$, Simpson, W = 1,549, effect size = 0.399, $P <$
244 0.001. Fungi; richness, W = 1,603, effect size = 0.438, $P < 0.001$, Shannon, W =
245 1,522, effect size = 0.379, $P < 0.001$, Simpson, W = 1,465, effect size = 0.338, $P <$
246 0.001). No significant difference in dominance however was observed in either
247 bacterial or fungal communities between diseases (Bacteria; W = 786, effect size
248 = 0.151, $P = 0.118$. Fungi; W = 781, effect size = 0.155, $P = 0.108$). Although
249 bacterial biomass was significantly lower in the CF group compared to BX (W =
250 1,100, effect size = 0.241, $P = 0.018$), no significant difference in fungal biomass
251 was observed (W = 706, effect size = 0.112, $P = 0.126$). Differences in both bacterial
252 and fungal community composition, calculated by PERMANOVA, showed
253 significant differences between CF and BX (bacterial, $R^2 = 0.066$, $P < 0.001$; fungal
254 $R^2 = 0.028$, $P = 0.004$).
255
256 *Differences between clinical fungal disease groups*
257 Participants within the study cohort were classified into four clinically defined
258 fungal disease groups (see Table 2): (i) fungal bronchitis (FB), (ii) Allergic

259 bronchopulmonary aspergillosis (ABPA), (iii) chronic necrotising pulmonary
260 aspergillosis (CNPA, BX only and (iv) non-tuberculosis mycobacteria (NTM). NTM
261 patients were included as a subgroup due to the strong link between NTM
262 infection and fungal disease [7, 38, 39]. A control group of patients presenting with
263 no active fungal disease (NAFD) were also included. Due to the small number of
264 patients with BX in the study (n = 24) the analysis was confined to the CF patient
265 samples.
266 We observed significant differences in fungal alpha diversity between the 39 CF
267 patients with no active fungal disease (NAFD) and the 20 with fungal bronchitis
268 (FB) (biomass: W = 522, effect size = 0.275, $P = 0.03$, richness; W = 211, effect size
269 = 0.374, $P < 0.004$; Shannon: W = 204, effect size = 0.387, p = 0.002; Simpson: W =
270 213, effect size = 0.368, $P = 0.004$).
271 Fungal beta diversity using Bray-Curtis similarity was also significantly different
272 between groups (PERMANOVA $R^2 = 0.077$, $P = 0.004$). We found no significant
273 differences in bacterial biomass or diversity between fungal disease groups.
274
275 *CFPE*
276 Thirty-six of the 83 CF patients in the study were defined as experiencing CFPE at
277 the time of sampling. There was no significant difference in bacterial or fungal
278 biomass or alpha diversity measures ($P > 0.1$) between CFPE subjects and the rest.
279
280 ***Random forest and differential abundance analysis***
281 We investigated differences in fungal and bacterial community composition
282 between groups of patients using random forest modelling. Models with mean
283 AUC >0.7 that were trained on a single kingdom were then used for association

284 testing. The strength of association was measured using de-biased random forest
285 variable importance scores while statistical significance was assessed using the
286 method of Altmann *et al*. (see Methods). Effect directions for the RF variable
287 importance scores were assigned using partial dependence plots (Figure S2). We
288 note that RF variable importance scores cannot be compared across models and
289 so are best used for ranking taxa from the same model.

290

291 **Disease status**

292 Fungal and bacterial community composition were both good predictors of
293 Disease status (CF or BX, bacterial mean AUC=0.78, fungal mean AUC=0.74, Figure
294 1a, left). Including both kingdoms as predictors in the model increased the mean
295 AUC to 0.79. The confidence intervals on the mean AUC however overlap for all
296 three models, so we cannot conclude that any one set of covariates were
297 significantly better predictors of Disease status.
298 The Gram-negative bacteria genus *Neisseria* was the highest ranked bacterial taxa
299 by the RF model predicting Disease status from bacterial genera (Figure 1b, left)
300 and was statistically significantly associated with Disease status at 1% FDR.
301 Partial dependence plots show that a higher relative abundance of *Neisseria*
302 implies a higher probability of BX. Other highly ranked bacterial genera included
303 *Megasphaera, Lactobacillus* and *Treponema* (direction BX) as well as *Pseudomonas*
304 (direction CF).
305 *Pseudomonas* was highly ranked by the RF model and was also the most
306 differentially abundant bacterial taxa according to ALDEx2, with a higher
307 abundance being significantly associated with CF at 1% FDR (effect size=2.5,

308 Figure 1c, left panel) Fusobacterium, Gemella and *Leptrotrichia* were the bacterial
309 taxa for which a higher abundance was significantly associated with BX by ALDEx2
310 at 10% FDR.
311 Three *Penicillium* species, - *Penicillium thomii*, *Penicillium commune* and
312 *Penicillium psychrosexualis* - were highly ranked by the RF models and
313 significantly associated with disease status by the RF model at 5% FDR (direction
314 BX), as was *Malassezia restricta. Candida parapsilosis* was the only one of the five
315 top-ranked fungal taxa where an increased relative abundance increased the
316 probability of CF according to the RF model. In addition, it was significantly
317 differentially abundant according to ALDEx2 at 10% FDR (Figure 1c, centre panel)
318 The pathogenic *Exophiala dermatitidis* was also identified by ALDEx2 but was not
319 highly ranked by the corresponding RF model.

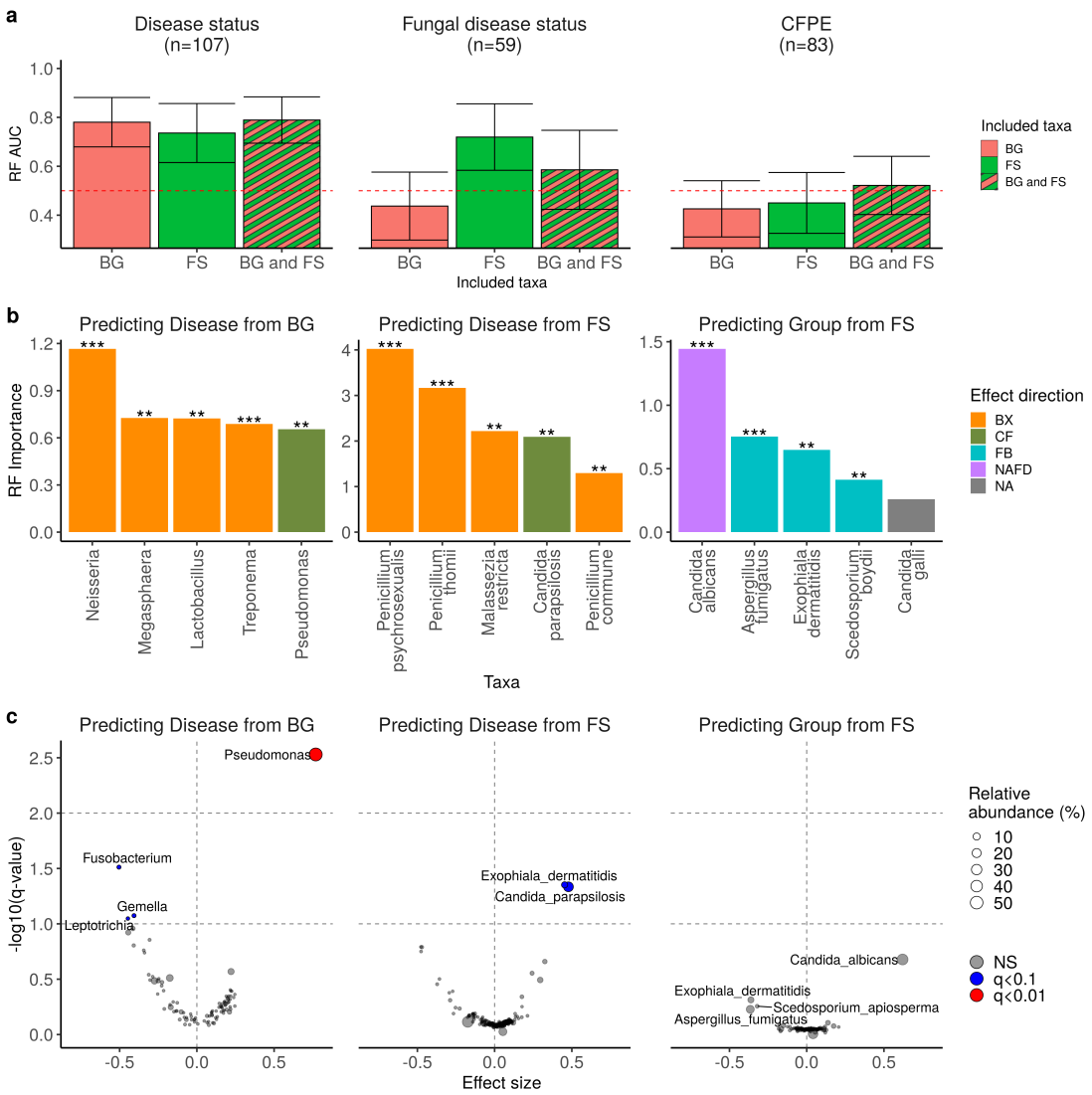

Figure 1. (a) Bacterial genera (BG), fungal species (FS) and a combination of both are all equally good predictors of CF/BX disease status (left panel), while fungal species are the best predictors of fungal bronchitis (centre panel). Neither fungal species nor bacterial genera are good predictors of CFPE (right panel). Individual fold AUCs are available in Figure S3. (b) Top five taxa by RF (unsigned) importance scores. Effect directions are taken from partial dependence plots (Figure S3) with NA indicating that the partial dependence is non-monotonic. Scores marked with **/*** are statistically significant at 5%/1% FDR. (c) ALDEx2 volcano plot showing effect size magnitudes and FDR-corrected *P*-values from Welch's t-test. NS: not statistically significant.

331

332 **Clinically defined fungal disease**

333 The random forest models were unable to predict clinically defined fungal disease
334 from bacterial community composition, as an AUC of 0.5 was within the 95%
335 confidence interval on the mean AUC (Figure 1a, centre panel). As expected, fungal
336 community composition was a good predictor of fungal disease status (mean
337 AUC=0.72). In this case, inclusion of both bacterial genera and fungal species in
338 the RF model reduced the power of the model to predict the fungal disease group.
339 RF models highlight three previously identified fungi as being associated with
340 fungal disease status: *Aspergillus fumigatus* ($P < 0.05$); *Exophiala dermatitidis* ($P <$
341 0.05) and *Scedosporium boydii* ($P < 0.05$) [7]. *Candida albicans* was the highest
342 ranked fungal taxa and was significantly associated with the absence of fungal
343 disease ($P < 0.05$). *Aspergillus fumigatus, Exophiala dermatitidis* and *Candida*
344 *albicans* also had the largest ALDEx2 effect sizes, however the associations were
345 not significant at 10% FDR (Figure 1c, left panel).

346

347 **Exacerbations**

348 Neither the bacterial nor fungal taxa contained strong predictive signal for CFPE
349 (Figure 1a, right panel). RF models predicting CFPE from bacterial genera had a
350 mean AUC<0.5, indicating that the random forest model is exhibiting worse-than-
351 random performance on unseen data. This may be a result of stratification bias in
352 a small sample, low-signal dataset (the presence of different proportions of the
353 two classes in different cross-validation folds) [40]. The poor predictive
354 performance of these models meant that no differential abundance analysis was
355 performed for CFPE.

356

357 **Unsupervised learning/clustering analysis and clustering**

358 Bacterial and fungal communities are structured, with strong positive and
359 negative correlations between the abundance of different organisms. Correlation
360 analysis found large blocks of co-correlated taxa from the same kingdom with few
361 strong correlations between taxa in different kingdoms (Figure 2a and 2b).
362 Amongst bacterial genera, the largest block of dependent taxa included
363 *Steptococcus*, *Veillonella*, *Rothia*, *Selemonas*, *Prevotella*, *Fusobacterium* and
364 *Atopobium*.
365 Two clear groups of fungal taxa were observed. One contained four Penicilluim
366 species (*psychrosexualis*, *thomii*, *commune* and *brevicompactum*) and the second
367 contained the pathogenic species *Aspergillus flavus*, *Aspergillus subversicolor*,
368 *Filobasidiella depauperata*, *Talaromyces marneffei* and *Trichosporon dematis*.
369 There were fewer significant inter-kingdom correlations than intra-kingdom
370 correlations for all FDR thresholds. This further indicates that the main
371 dependencies in this dataset exist within kingdoms rather than between them
372 (Figure 2c).
373

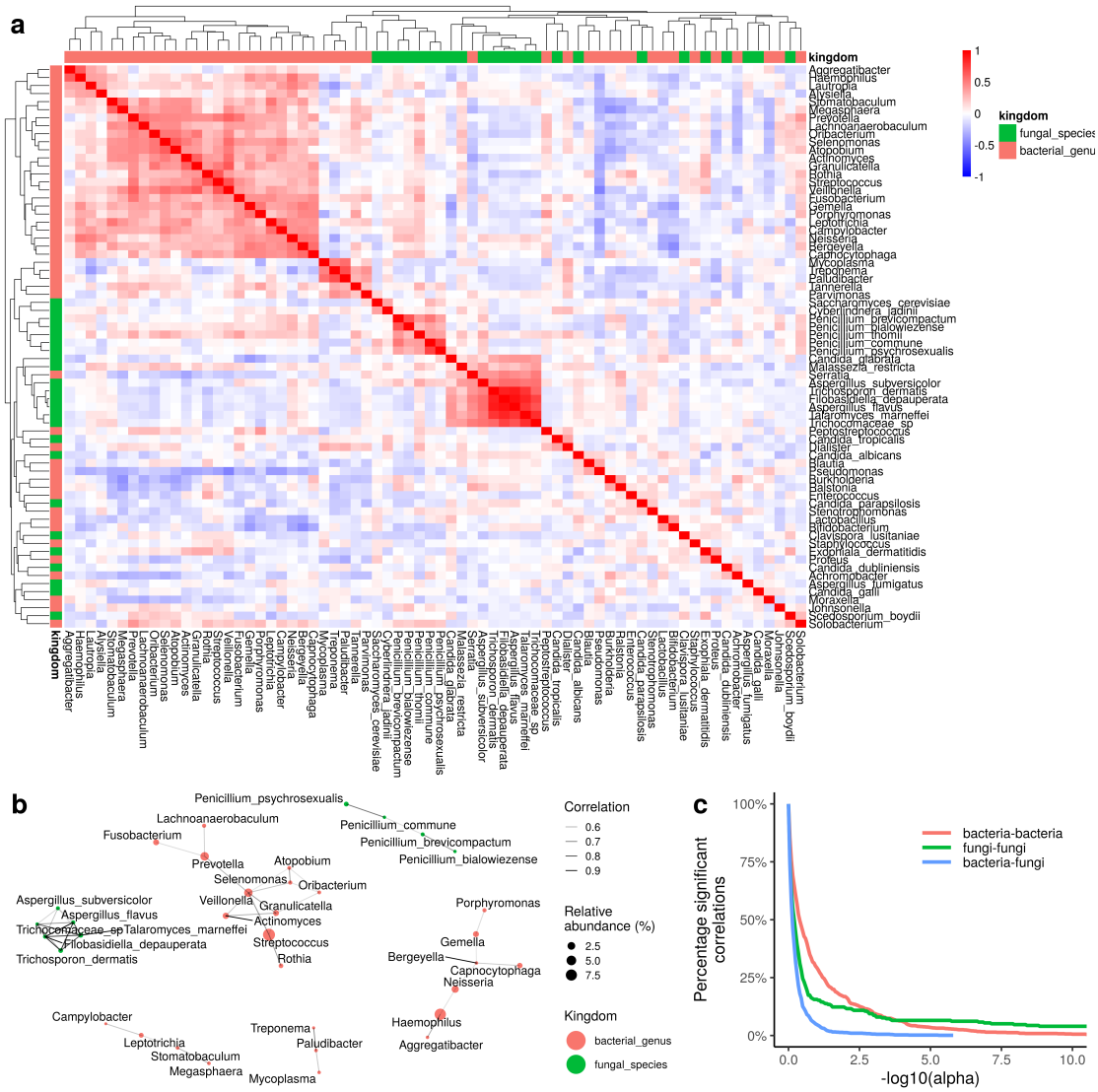

Figure 2: (a) Pearson correlation between the taxa included in the RF analysis (those with prevalence over 20%). The bacterial and fungal abundances were separately centred log-ratio transformed prior to calculating correlations. (b): The same correlations visualised as a dependence network. Correlations with magnitude less than 0.5 are not shown and all displayed correlations are positive(c): There are a higher percentage of significant correlations within the bacterial and fungal communities (red and green lines) than between them (blue line) for any FDR threshold alpha.

384 Community structure within the bacterial and fungal kingdoms across samples
385 was further analysed with Dirichlet mixture components [41]. The latter groups
386 samples into distinct clusters with similar composition. In this context two
387 samples are similar if they are well-described by the same Dirichlet multinomial
388 distribution. This unsupervised approach can provide insight into the community-
389 level structure that may exist across all the samples and may or may not
390 correspond to pre-defined clinical labels. This complements the supervised RF
391 analysis in the previous section, which identified taxa that differentiated between
392 specified groups.
393 All 107 samples were clustered using Dirichlet Multinomial Mixture models. This
394 was done separately using the fungal species and bacterial genera, giving two sets
395 of cluster labels. Using information-theoretic goodness of fit measures (Figure S4)
396 2 distinct bacterial clusters (Figure 3a) and a single fungal cluster were identified.
397 Random forest variable importance scores and ALDEx2 effect sizes show that the
398 two bacterial clusters are characterised by the presence or absence of
399 *Pseudomonas* domination (Figure 3c). The two bacterial clusters did not
400 correspond to any of the clinical labels in the dataset (Table 3 and Figure 3d). The
401 dependence between the fungal and bacterial communities was then further
402 investigated using RF modelling with the bacterial cluster as the class label. Fungal
403 community composition contained low levels of predictive signal for bacterial
404 cluster (mean AUC=0.56, Figure 3b), indicating that this important feature of
405 bacterial community composition (*Pseudomonas* domination) is largely
406 independent of the fungal community.
407

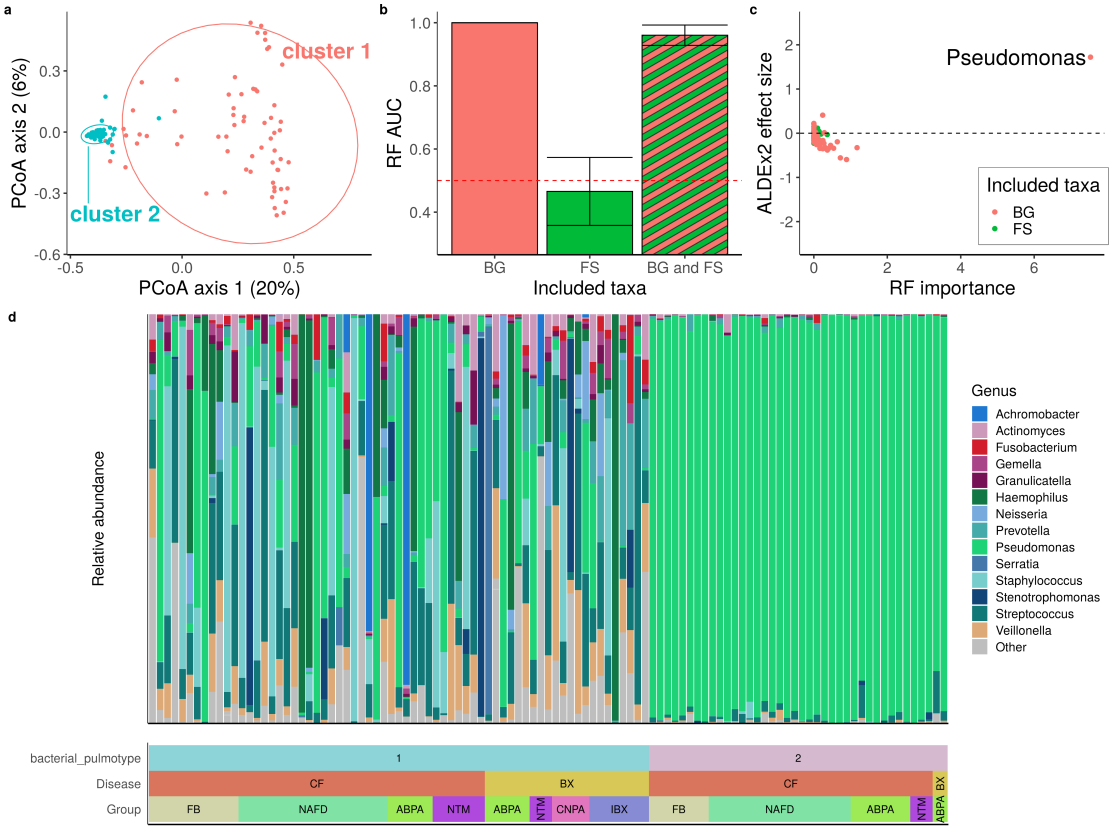

Figure 3: (a) Clustering of samples based on bacterial genera abundances identifies two clusters that are separable in Bray-Curtis Principle co-ordinates analysis (PCoA) space. The clusters do not correspond to clinical labels (see Table 3). Clustering of samples based on fungal species abundance did not identify any distinct clusters. (b) Fungal species are poor predictors of bacterial cluster, suggesting limited interactions between the kingdoms in this dataset. Error bars denote 95% confidence interval on mean validation area under the curve (AUC). (c) Random Forest (RF) modelling and ALDEx2 differential expression analysis show that *Pseudomonas* abundance differentiates the two clusters, with Cluster 1 containing low-*Pseudomonas* samples and Cluster 2 containing high-*Pseudomonas* samples. Negative RF variable importance scores are set to zero. (d) The bacterial composition of the samples when ordered by cluster clearly shows that they correspond to presence or absence of domination by *Pseudomonas* species.

422

423 **Table 3.** Adjusted Rand Index (ARI) between bacterial cluster and clinical labels

424 show that there is no similarity between the bacterial cluster labels and other

425 clinical labels used in this study**.**

	Disease	Group	CFPE
ARI(bacterial cluster, clinical label)	-0.01	-0.02	0.02

426

427

428 **Discussion**

429 Identifying factors predicting the prevalence and severity of chronic respiratory

430 infections may be crucial for improving clinical outcomes in CSLDs [1, 2]. To date

431 most research has focused on bacterial pathogens., Clinical evidence however

432 suggests that fungal infection plays a key role in chronic disease progression [5,

433 6]. As such, understanding the inter-kingdom association present within the lungs

434 is an essential step towards effective antimicrobial treatments.

435 Despite many similarities in the symptoms and treatments of the underlying

436 diseases (CF and BX), there are clearly fundamental differences in their microbial

437 communities. These are likely to be driven by the physiological differences

438 underlying the individual diseases [42].

439 To understand the microbial drivers that differ between CF and BX, we used

440 random forest modelling to explore the power of bacterial and fungal

441 communities to distinguish between CF and BX patients. In addition, we used

442 equivalent analyses to investigate differences between CF patients with and

443 without clinically defined fungal disease, and indicators of those experiencing

444 CFPE at the time of sampling.

445 Using RF modelling, bacterial and fungal communities were both independently
446 found to be good predictors of underlying disease status. Including both fungal
447 and bacterial OTUs in the model did not however significantly improve the
448 predictive power, suggesting independence between the fungal and bacterial
449 communities. We then used the models to identify taxa associated with either CF
450 or BX. We found that the variable importance scores for fungal covariates were
451 more concentrated in a small number of highly discriminative taxa (*Penicillium*
452 *thomii*, *Penicillium commune, Penicillium psychrosexualis, Malassezia restricta* and
453 *Candida parapsilosis)*. In contrast, RF models identified *Neisseria* as being highly
454 discriminative (direction BX) followed by 18 taxa with similar importance scores
455 (Figures S5(a) and S6). This point is further demonstrated by the fact that RF
456 models predicting Disease status identified 4 fungal species with $P < 0.05$, while
457 corresponding models identified 19 bacterial genera with $P < 0.05$ (Figure S5(b)).
458 RF models revealed *Neisseria* to have the highest variable importance scores and
459 to be associated with BX. Members of the *Neisseria* genus are commonly isolated
460 in the upper respiratory tract with some species being known pathogens [43]. Our
461 results may suggest that a pathogenic role for *Neisseria* spp. could be considered
462 for BX. This warrants further investigation.
463 Both RF modelling and ALDEx2 identified the *Pseudomonas* genus as significantly
464 associated with CF patients. This is consistent with the known role of
465 *Pseudomonas aeruginosa* as one of the most common pathogens associated with
466 CF lung disease. In non-CF bronchiectasis *Pseudomonas* infection may be
467 associated with more severe disease [44], but it was not a prominent factor in our
468 BX patients.

469 Fungal species associated with BX were primarily part of the *Penicillium* genus.
470 Symptomatic infections with *Penicillium* spp. are rare [45] and *Penicillum* spp. are
471 widely present in the air, making it a logical part of the normal respiratory flora.
472 Conversely *Exophaila dermatitidis* and *Candida parapsilosis*, which were identified
473 as significantly associated with CF using ALDEx2, have been shown to be
474 significant pathogens in CF. RF models identified *E. dermatitidis* alongside
475 *Aspergillus fumigatus* and *Scedosporium boydii*, as one of the major pathogens in
476 FB, while *Candida parapsilosis* was shown to be associated with CF patients with
477 NAFD. Despite the association with the NAFD group, *C. parapsilosis* has been
478 identified previously to be an opportunistic human pathogen [46].
479 We also explored differences in the microbiota between patients with and without
480 a clinical fungal infection. The relatively low numbers of patients with BX led us to
481 restrict analysis of fungal disease diagnoses to differences between FB and NAFD
482 in the CF patients. We did not find any predictive signal in the bacterial community
483 for the presence of FB. This provides further evidence for a lack of inter-kingdom
484 interactions driving fungal disease. The fungal pathogens *Aspergillus fumigatus,*
485 *Exophiala dermatitidis* and *Scedosporium boydii* were however associated with FB
486 by RF models, with *Aspergillus fumigatus* and *Exophiala dermatitidis* showing the
487 strongest effects by both the RF model and ALDEx2. Although the RF associations
488 were significant ($P < 0.05$) those from ALDEx2 were not ($P > 0.1$).
489 The most important taxon associated with the NAFD group in the CF cohort was
490 *Candida albicans*. *C. albicans* is a common commensal organism, however, in
491 immunocompromised patients these organisms have been shown to be
492 opportunistic pathogens. In CF adults *Candida spp.* colonization has been shown
493 to be associated with use of inhaled steroids, diabetes mellitus and antibiotic

494 treatment. Despite these observations the virulence potential of *C. albicans* in CF
495 is still being explored[47]. In the present study *Candida spp*. were present with a
496 lower relative abundance in the FB group suggesting that dominance of
497 filamentous fungi may out compete *Candida spp.* in these patients. More work is
498 therefore needed to understand the role *C. albicans* plays in CF disease
499 progression.
500 Pulmonary exacerbations are major clinical events in patients with CF resulting in
501 lung function decline and clinical disease progression [48]. The presence of
502 bacteria and viruses is commonly associated with these poor outcomes during
503 CFPE [8], but defining their exact role is challenging. Recent evidence has
504 suggested fungal infections are also associated with increased CFPE although to
505 date few studies have explored this area [49, 50]. A recent publication by Soret *et*
506 *al*. investigated CFPE using an adapted penalised linear model and cross-sectional
507 data and suggested that two fungal genera, *Aspergillus* and *Malassezia* were
508 associated with CFPE [8]. In our study however random forest models were not
509 able to distinguish between patients with and without CFPE. This may be because
510 the cross-sectional data in our study is unable to show intra-patient variation
511 within the microbial community. Longitudinal data is likely to be more
512 informative for predicting pulmonary exacerbations and should be considered for
513 future studies.
514 We used an unsupervised clustering analysis to detect interactions between
515 different genera, identifying two bacterial clusters which did not correspond
516 simply to the CF/BX groups. The bacterial clusters were however clearly driven
517 by dominance of *Pseudomonas* within individual samples (Figure 2c). This is in
518 agreement with the RF analysis, which found that a higher relative abundance of

519 *Pseudomonas* was associated with increased probability of CF. This effect was
520 weak with partial dependence plots (Figure S2) showing a small increase in
521 probability of CF when the relative abundance of *Pseudomonas* was >70%.
522 Our RF models were unable to predict bacterial clusters from fungal species,
523 providing additional evidence of independence between the bacterial and fungal
524 communities. Inter-kingdom correlations were generally weaker than those
525 within either kingdom (measured by proportion of significant correlations at
526 different significance thresholds). This further indicates that intra-kingdom
527 interactions may play a minor role in these subjects.
528 By utilising distinct statistical methods (RF variable importance scores and
529 ALDEx2 effect sizes) we were able to obtain a consensus result that is more robust
530 than results from either individual tool, as applying two statistical approaches
531 with distinct assumptions reduces the risk that findings are due to technical or
532 statistical artefacts [51]. ALDEx2 explicitly models the compositional nature of
533 sequencing data, while RFs do not make assumptions about the characteristics of
534 the covariates [35]. There was statistically significant agreement between the RF
535 variable importance scores and ALDEx2 effect sizes for all the models used for
536 differential abundance analysis (Figure S7).
537 Our analysis does though have several limitations that should be considered when
538 interpreting the results presented. Most importantly, associations identified by
539 machine learning models such as random forest do not imply causal links.
540 Furthermore, the importance scores from random forests should be interpreted
541 with care. Using debiased Gini importance scores and assessing statistical
542 significance via permutation reduces the danger of spurious associations but does
543 not provide a framework that allows quantitative statements to be made (as is

544 possible with effect sizes in linear models). Machine learning is a powerful tool for
545 exploring microbial interactions and drivers of disease, however, understanding
546 the limitations of the models is vital for interpretation. We have consequently
547 utilised multiple statistical tools to reduce the dangers of over-interpreting the
548 results of one approach.
549 A further limitation of this study is the use of 16S rRNA gene sequencing and ITS2
550 sequencing for the exploration of these communities. This technology allows us to
551 understand the microbial community present within the lung but give us no
552 information on their activity.
553 A recent study carried out by Hughes *et al*. established that using culture-based
554 methodologies, the known CF pathogens, *Pseudomonas aeruginosa* and *Aspergillus*
555 *fumigatus* are rarely cultured from the same sample [52]. Despite this, our culture
556 independent techniques clearly show the common presence of both *Pseudomonas*
557 *aeruginosa* and *Aspergillus fumigatus* reads in the samples. It is possible that
558 microbial interactions within the lungs may not be detected by DNA based
559 methods. Future work may require functional analyses to explore relative
560 microbial gene expression within the lung.
561
562 **Conclusions**
563 Our study suggests that the fungal microbiota play a key role in chronic
564 respiratory disease independently of bacterial infection. Further work is required
565 to understand the full impact of fungal infection in chronic respiratory disease.
566 Improvements in clinical diagnosis of fungal infections, whether by sequence
567 analysis, transcriptomics or advanced cultures, may underpin the improvement of
568 patient outcomes.

Declarations

Availability of data and material

Sequencing data is freely assessable through the European nucleotide database, bacterial data can be accessed under project number PRJEB33064, while the fungal data can be accessed under project number PRJEB33434.

Data analysis scripts are freely available on figshare, under the project "Machine learning for exploring microbial inter-kingdom associations in Cystic Fibrosis and Bronchiectasis" DOIs: https://doi.org/10.6084/m9.figshare.17897708 and https://doi.org/10.6084/m9.figshare.17942543.

Funding

This project was supported by the Asmarley Trust, the Wellcome Trust and the NIHR Respiratory Disease Biomedical Research Unit at the Royal Brompton and Harefield NHS Foundation Trust, Imperial College London. IF was supported by an NIHR PhD studentship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

All authors listed provided substantial contributions to this manuscript. Study concept and design; MFM, WOCC, SF, LC, NJS, MRL, MJC. Acquisition of data; IF, PJ, LC, MJC. Statistical analysis and interpretation of the data; LC, JIH. Drafting of the

manuscript; LC, JIH, WOCC, SF. Critical revision of the manuscript for was carried out by all authors.

References

1. O'Sullivan BP, Freedman SD: **Cystic fibrosis.** *Lancet* 2009, **373:**1891-1904.
2. King PT: **The pathophysiology of bronchiectasis.** *Int J Chron Obstruct Pulmon Dis* 2009, **4:**411-419.
3. Zemanick ET, Hoffman LR: **Cystic Fibrosis: Microbiology and Host Response.** *Pediatr Clin North Am* 2016, **63:**617-636.
4. Maselli DJ, Amalakuhan B, Keyt H, Diaz AA: **Suspecting non-cystic fibrosis bronchiectasis: What the busy primary care clinician needs to know.** *Int J Clin Pract* 2017, **71**.
5. Chotirmall SH, O'Donoghue E, Bennett K, Gunaratnam C, O'Neill SJ, McElvaney NG: **Sputum Candida albicans presages FEV(1) decline and hospital-treated exacerbations in cystic fibrosis.** *Chest* 2010, **138:**1186-1195.
6. Amin R, Dupuis A, Aaron SD, Ratjen F: **The effect of chronic infection with Aspergillus fumigatus on lung function and hospitalization in patients with cystic fibrosis.** *Chest* 2010, **137:**171-176.
7. Cuthbertson L, Felton I, James P, Cox MJ, Bilton D, Schelenz S, Loebinger MR, Cookson WOC, Simmonds NJ, Moffatt MF: **The fungal airway microbiome in cystic fibrosis and non-cystic fibrosis bronchiectasis.** *Journal of Cystic Fibrosis* 2021, **20:**295-302.
8. Soret P, Vandenborght LE, Francis F, Coron N, Enaud R, Avalos M, Schaeverbeke T, Berger P, Fayon M, Thiebaut R, et al: **Respiratory mycobiome and suggestion of inter-kingdom network during acute pulmonary exacerbation in cystic fibrosis.** *Sci Rep* 2020, **10:**3589.
9. Reece E, Segurado R, Jackson A, McClean S, Renwick J, Greally P: **Co-colonisation with Aspergillus fumigatus and Pseudomonas aeruginosa is associated with poorer health in cystic fibrosis patients: an Irish registry analysis.** *BMC Pulm Med* 2017, **17:**70.
10. Parize P, Boussaud V, Poinsignon V, Sitterle E, Botterel F, Lefeuvre S, Guillemain R, Dannaoui E, Billaud EM: **Clinical outcome of cystic fibrosis patients colonized by Scedosporium species following lung transplantation: A single-center 15-year experience.** *Transpl Infect Dis* 2017, **19**.
11. Zhou YH, Gallins P: **A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction.** *Frontiers in Genetics* 2019, **10**.

12. Topcuoglu BD, Lesniak NA, Ruffin MIV, Wiens J, Schlossa PD: **A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems.** *Mbio* 2020, **11**.
13. Qi Y: *Random forest for bioinformatics. In Ensemble machine learning.* Springer, Boston, MA.; 2012.
14. Szymczak S, Holzinger E, Dasgupta A, Malley JD, Molloy AM, Mills JL, Brody LC, Stambolian D, Bailey-Wilson JE: **r2VIM: A new variable selection method for random forests in genome-wide association studies.** *Biodata Mining* 2016, **9**.
15. Ram M, Najafi A, Shakeri MT: **Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest.** *Iran J Pathol* 2017, **12:**339-347.
16. Roguet A, Eren AM, Newton RJ, McLellan SL: **Fecal source identification using random forest.** *Microbiome* 2018, **6**.
17. Thompson J, Johansen R, Dunbar J, Munsky B: **Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition.** *Plos One* 2019, **14**.
18. Nagpal R, Neth BJ, Wang SH, Mishra SP, Craft S, Yadav H: **Gut mycobiome and its interaction with diet, gut bacteria and alzheimer's disease markers in subjects with mild cognitive impairment: A pilot study.** *Ebiomedicine* 2020, **59**.
19. Das A, O'Herlihy E, Shanahan F, O'Toole PW, Jeffery IB: **The fecal mycobiome in patients with Irritable Bowel Syndrome.** *Scientific Reports* 2021, **11**.
20. Cuthbertson L, Craven V, Bingle L, Cookson W, Everard ML, Moffatt MF: **The impact of persistent bacterial bronchitis on the pulmonary microbiome of children.** *PLoS One* 2017, **12:**e0190075.
21. McMurdie PJ, Holmes S: **phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.** *PLoS One* 2013, **8:**e61217.
22. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ: **Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data.** *Microbiome* 2018, **6:**226.
23. Jari Oksanen FGB, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner: **vegan: Community Ecology Package. R package version 2.5-6.** 2019.
24. Kuhn M: **Building Predictive Models in R Using the caret Package.** *Journal of Statistical Software* 2008, **28:**1-26.
25. Wright MN, Ziegler A: **ranger: A Fast Implementation of Random Forests for High Dimensional Data in C plus plus and R.** *Journal of Statistical Software* 2017, **77:**1-17.
26. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB: **ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq.** *Plos One* 2013, **8**.
27. Morgan M: **DirichletMultinomial: Dirichlet-Multinomial Mixture Model Machine Learning for Microbiome Data.** *R package version 1340* 2021.

683 28. LeDell E, Petersen M, van der Laan M: **Computationally efficient**
684 **confidence intervals for cross-validated area under the ROC curve**
685 **estimates.** *Electronic Journal of Statistics* 2015, **9:**1583-1607.
686 29. Nembrini S, Konig IR, Wright MN: **The revival of the Gini importance ?**
687 *Bioinformatics* 2018, **34:**3711-3718.
688 30. Strobl C, Boulesteix AL, Zeileis A, Hothorn T: **Bias in random forest**
689 **variable importance measures: Illustrations, sources and a solution.**
690 *Bmc Bioinformatics* 2007, **8**.
691 31. Wang H, Yang F, Luo Z: **An experimental study of the intrinsic stability**
692 **of random forest variable importance measures.** *BMC Bioinformatics*
693 2016, **17:**60.
694 32. Behnamian A, Millard K, Banks SN, White L, Richardson M, Pasher J: **A**
695 **Systematic Approach for Variable Selection With Random Forests:**
696 **Achieving Stable Variable Importance Values.** *Ieee Geoscience and*
697 *Remote Sensing Letters* 2017, **14:**1988-1992.
698 33. Altmann A, Tolosi L, Sander O, Lengauer T: **Permutation importance: a**
699 **corrected feature importance measure.** *Bioinformatics* 2010, **26:**1340-
700 1347.
701 34. Friedman JH: **Greedy function approximation: A gradient boosting**
702 **machine.** *The Annals of Statistics* 2001, **29:**1189:1232.
703 35. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM: **A**
704 **field guide for the compositional analysis of any-omics data.**
705 *Gigascience* 2019, **8**.
706 36. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison
707 KR, Consortium D, Kellis M, Collins JJ, Stolovitzky G: **Wisdom of crowds**
708 **for robust gene network inference.** *Nat Methods* 2012, **9:**796-804.
709 37. Aghaeepour N, Finak G, Flow CAPC, Consortium D, Hoos H, Mosmann TR,
710 Brinkman R, Gottardo R, Scheuermann RH: **Critical assessment of**
711 **automated flow cytometry data analysis techniques.** *Nat Methods*
712 2013, **10:**228-238.
713 38. Kunst H, Wickremasinghe M, Wells A, Wilson R: **Nontuberculous**
714 **mycobacterial disease and Aspergillus-related lung disease in**
715 **bronchiectasis.** *Eur Respir J* 2006, **28:**352-357.
716 39. Joao I, Bujdakova H, Jordao L: **Opportunist Coinfections by**
717 **Nontuberculous Mycobacteria and Fungi in Immunocompromised**
718 **Patients.** *Antibiotics (Basel)* 2020, **9**.
719 40. Parker BJ, Guenter S, Bedo J: **Stratification bias in low signal**
720 **microarray studies.** *Bmc Bioinformatics* 2007, **8**.
721 41. Holmes I, Harris K, Quince C: **Dirichlet Multinomial Mixtures:**
722 **Generative Models for Microbial Metagenomics.** *Plos One* 2012, **7**.
723 42. !!! INVALID CITATION !!! [39].
724 43. Weyand NJ: **Neisseria models of infection and persistence in the**
725 **upper respiratory tract.** *Pathog Dis* 2017, **75**.
726 44. Davies G, Wells AU, Doffman S, Watanabe S, Wilson R: **The effect of**
727 **Pseudomonas aeruginosa on pulmonary function in patients with**
728 **bronchiectasis.** *Eur Respir J* 2006, **28:**974-979.
729 45. Lyratzopoulos G, Ellis M, Nerringer R, Denning DW: **Invasive infection**
730 **due to penicillium species other than P. marneffei.** *J Infect* 2002,
731 **45:**184-195.

46. Pendleton KM, Huffnagle GB, Dickson RP: **The significance of Candida in the human respiratory tract: our evolving understanding.** *Pathog Dis* 2017, **75**.
47. Chotirmall SH, Greene CM, McElvaney NG: **Candida species in cystic fibrosis: A road less travelled.** *Med Mycol* 2010, **48 Suppl 1:**S114-124.
48. de Boer K, Vandemheen KL, Tullis E, Doucette S, Fergusson D, Freitag A, Paterson N, Jackson M, Lougheed MD, Kumar V, Aaron SD: **Exacerbation frequency and clinical outcomes in adult patients with cystic fibrosis.** *Thorax* 2011, **66:**680-685.
49. Nguyen LDN, Deschaght P, Merlin S, Loywick A, Audebert C, Van Daele S, Viscogliosi E, Vaneechoutte M, Delhaes L: **Effects of Propidium Monoazide (PMA) Treatment on Mycobiome and Bacteriome Analysis of Cystic Fibrosis Airways during Exacerbation.** *Plos One* 2016, **11**.
50. Willger SD, Grim SL, Dolben EL, Shipunova A, Hampton TH, Morrison HG, Filkins LM, Toole GAO, Moulton LA, Ashare A, et al: **Characterization and quantification of the fungal microbiome in serial samples from individuals with cystic fibrosis.** *Microbiome* 2014, **2**.
51. Costa-Silva J, Domingues D, Lopes FM: **RNA-Seq differential expression analysis: An extended review and a software tool.** *Plos One* 2017, **12**.
52. Hughes DA, Rosenthal M, Cuthbertson L, Ramadan N, Felton I, Simmonds NJ, Loebinger MR, Price H, Armstrong-James D, Elborn JS, et al: **P130 Pseudomonas aeruginosa impairs growth of aspergillus from CF airway samples.** *Thorax* 2021, **76:**A159.