JOURNAL OF
EMERGING INVESTIGATORS

Article

Analysis of the lung microbiome in cystic fibrosis patients using 16S sequencing

Manasvi Pinnaka[1], Brianna Chrisman[2]

[1] Basis Independent Silicon Valley, San Jose, California

[2] Department of Bioengineering, Stanford University, Palo Alta, California

SUMMARY

Cystic fibrosis (CF) patients often develop lung infections that range anywhere in severity from mild to life-threatening due to the presence of thick, sticky mucus that fills their airways. Since many of these infections are chronic, they not only affect a patient's ability to breathe but also increase chances of mortality by respiratory failure. Examining the lung microbiology of CF patients can help with the prediction of the current condition of lung function, with the potential to guide doctors when designing personalized and thereby more effective treatment plans for patients. With a publicly available dataset of DNA sequences from bacterial species in the lung microbiome of CF patients, we investigated the correlations between different microbial species in the lung and the extent of deterioration of lung function. 16S sequencing technologies allowed us to determine the microbiome composition of the samples in the dataset. We found that the *Fusobacterium, Actinomyces,* and *Leptotrichia* microbial types all had positive correlations with forced expiratory volume in 1 second (FEV1) score, indicating potential displacement of these species by pathogens as the disease progresses. However, the dominant pathogens themselves, including *Pseudomonas aeruginosa* and *Staphylococcus aureus,* did not have statistically significant negative correlations with FEV1 score as described in past literature.

INTRODUCTION

More than 160,000 people worldwide suffer from cystic fibrosis (CF) (1). Patients with this disease often develop thick and sticky mucus caused by an abnormality in the cystic fibrosis transmembrane conductance regulator (CFTR) gene that is known to cause organ damage in the lungs (2). While there have been advances in treatments for CF, people born with this disease still, on average, live just until their 50s (3). Due to the fatal nature of CF, understanding and tracking disease progression for each CF patient is crucial in determining what treatment interventions would most effectively increase the patient's survival. However, one setback for this approach is that there are no specific indicators of disease severity due to the high variability in the complications and symptoms that each patient faces.

One possible avenue researchers can instead look to as an indicator of disease severity is the patient's lung microbiome composition, or the types of different microorganisms and their abundance in the lung. Nearly every single CF patient is expected to experience lung infections as their disease progresses, and approximately 80 to 95% of CF patients experience chronic lung infections and inflammation that eventually lead to respiratory failure (4,5). Not only are chronic lung infections correlated with CF fatality, especially when the pathogen *Pseudomonas aeruginosa* is present, but past studies have also shown that patients who show less microbial diversity in the lung also have decreased lung function (6). This research thus aims to examine the relationship between the composition of the lung microbiome in CF patients and the severity of the disease, as indicated by a measure of lung function, through 16S rRNA sequencing methodologies.

Particularly for bacterial and archaeal isolates, the goal of 16S sequencing is to determine the microbiome composition of a certain sample, distinguishing between taxonomies using reference genomes and determining proportions of a certain taxa relative to another. Since the 16S gene is present in the DNA of every bacterial type, the presence of variable regions of the gene can be used for taxonomic classification. In 16S sequencing, the operational taxonomic units (OTUs), formed through sequence similarity, are grouped based on how similar they are to the sequences from the reference database, which consists of different rRNA gene sequences that each correspond to specific taxa. This then allows for a determination of microbiome composition (7).

We hypothesized that dominant CF pathogens, such as *P. aeruginosa* and *Staphylococcus aureus*, would have a strong, negative correlation with the forced expiratory volume in 1 second (FEV1) score variable, which measures what maximum air volume can be expelled from the lungs after maximum inhalation; it was chosen for this study because it is considered a relevant and effective measure of lung function in the medical field when tracking the progression of CF (8). This study utilized multiple hypothesis correction, after the taxonomic classification steps, to identify the top (statistically significant) positive and negative correlations between different bacterial species and FEV1 score. From these analyses, we were able to identify three bacterial types that were positively correlated with FEV1 score in patients with CF: *Fusobacterium*, *Actinomyces*, and *Leptotrichia*.

RESULTS

Overview

The open-source dataset we used for this research (Qiita) consists of 51 samples of the lung microbiome from CF patients before and after interventions with Trikafta therapy (standard therapy for CF patients with the most common disease mutation — p.F580del).

Taxonomic classification through 16s rRNA sequencing using the Galaxy pipeline of the raw DNA sequences from the Qiita dataset gave us four outputs: (a) sample metadata: this contained sample information, notably the FEV1 scores for each patient, (b) representative sequence table: this was a clustered data matrix of the different sequences for each patient and their corresponding representative sequences, (c) count table: this contained the number of times each initial sequence had corresponded to each patient, and (d) aggregated sequence table: this combined the representative sequence table and the count table to display the total counts of each of the representative sequences for each of the patients (9). One especially notable value for the sample metadata ouput was the FEV1 scores for each patient. For a healthy individual, an FEV1 score above 80% is expected as this is considered normal. On the other hand, for those with obstructive lung diseases like CF, the lower the FEV1 score, the greater the severity of the disease in the patient's current condition. For example, an FEV1 score greater than 70% indicates a mild condition while an FEV1 score between 35% and 49% indicates a severe condition (10).

Using this data and Spearman's rank correlation coefficient, we were able to correlate each of the species to the FEV1 scores for each patient available in the table of sample information. For each bacterial type, correlations with both the bacterial cluster and the genus were performed. Due to the uncertainty of sequencing during taxonomic classifications, we can evaluate statistical significance in terms of whether a certain microbial type correlates with FEV1 score. This can be done using either the different bacterial clusters or the different taxonomic classifications (genus, family, order etc.). Bacterial clusters group 16S sequences together by genetic similarity. On the other hand, taxonomies rely on human-created categories, many of which were created before sequencing was available and were simply based on the visual or chemical properties of a bacteria.

Fusobacterium is positively correlated with FEV1 score

We found that considering the bacterial cluster classification, *Fusobacterium* was positively correlated with FEV1 score. As the FEV1 score increased, we saw a positive and moderately strong association between the two variables ($p = 9.863 \times 10\text{-}7$, $r2 = 0.662$, **Figure 1**). Meanwhile, the correlation using the genus cluster classification was not found to be statistically significant.

Actinomyces is positively correlated with FEV1 score

We also found that considering the bacterial cluster classification, *Actinomyces* was positively correlated with FEV1 score. As the FEV1 score increased, we once again saw a positive and moderately strong association between the two variables ($p = 1.074 \times 10\text{-}4$, $r2 = 0.551$, **Figure 2**). Meanwhile, the correlation using the genus cluster classification was once again not found to be statistically significant.

Leptotrichia is positively correlated with FEV1 score

Finally, there was a positive correlation between *Leptotrichia* and FEV1 score. Considering the *Leptotrichia* bacterial cluster classification first, as the FEV1 score increases, there is a positive and moderately strong association between the two variables ($p = 3.315 \times 10^{-4}$, $r2 = 0.517$, **Figure 3**).

Considering the *Leptotrichia* genus classification rather than the bacterial cluster classification next, this positive correlation persists between the *Leptotrichia* genus classification and FEV1 score. As the FEV1 score increases, there is a similar positive and moderately strong association between the two variables as seen in Figure 3 between the Leptotrichia bacteria cluster and FEV1 score ($p = 1.436 \times 10^{-4}$, $r2 = 0.542$, **Figure 4**). With this correlation, specifically, it should be noted that the correlated label was 'unclassified Leptotrichiaceae,' indicating that the genus classification is *Leptotrichia*. However, with the taxonomic classifications steps performed through 16s rRNA sequencing, the confidence level on the Galaxy platform was not high enough for the classification system to fully define it as such. Nevertheless, the correlation between this microbial type and FEV1 score is still evident, further supporting the relationship between the Leptotrichia bacteria and FEV1 score as a moderately strong

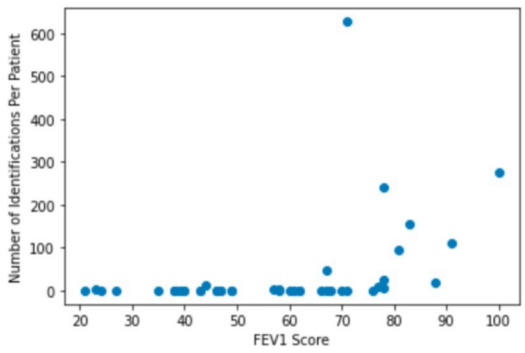

Figure 1. Positive correlation between Fusobacterium and FEV1 score. FEV1 score, an effective measure of lung function in the medical field, was provided for each CF patient (n = 51) as a measure of the maximum volume of air that can be expelled from the lungs after maximum inhalation. Spearman rank correlation, ***p < 0.0004.

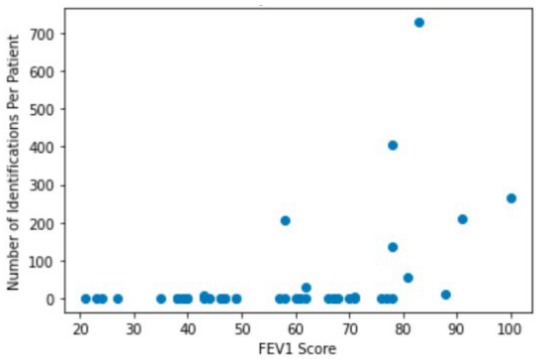

Figure 2. Positive correlation between Actinomyces and FEV1 score. FEV1 score, an effective measure of lung function in the medical field, was provided for each CF patient (n= 51) as a measure of the maximum volume of air that can be expelled from the lungs after maximum inhalation. Spearman rank correlation, ***p < 0.0004.

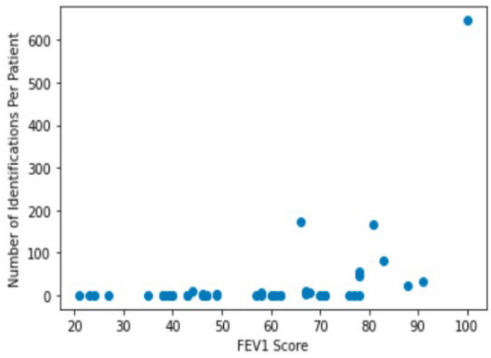

Figure 3. Positive correlation between Leptotrichia (bacterial cluster) and FEV1 score. FEV1 score, an effective measure of lung function in the medical field, was provided for each CF patient (n = 51) as a measure of the maximum volume of air that can be expelled from the lungs after maximum inhalation. Spearman rank correlation, ***p < 0.0004.

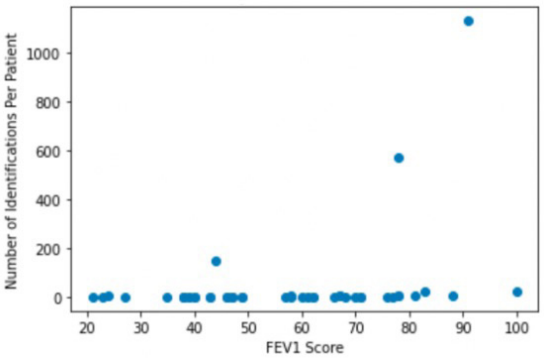

Figure 4. Positive correlation between Leptotrichia (genus) and FEV1 score. FEV1 score, an effective measure of lung function in the medical field, was provided for each CF patient (n = 51) as a measure of the maximum volume of air that can be expelled from the lungs after maximum inhalation. Spearman rank correlation, ***p < 0.0004.

and statistically significant correlation as seen with both analysis of the bacterial cluster and analysis of the genus.

Top CF pathogens were not statistically significant

Although these bacterial types were expected to be top hits in our analysis, the correlations between the counts of these bacteria for all the patients and FEV1 score were not statistically significant. This was the case for both *P. aeruginosa* (p = 0.368, r2 = 0.019, **Figure 5**) and *S. aureus* (p = 0.238, r2 = 0.032, **Figure 6**).

DISCUSSION

From our study, we found that *Fusobacterium, Actinomyces*, and *Leptotrichia* were all positively correlated with FEV1 score in CF patients. The fact that a greater FEV1 score is associated with less disease severity explains why in each of the reported correlations, the number of identifications for patients with an FEV1 score under 70% were largely zero. Since FEV1 scores of 70% and higher are indicative of these milder conditions, the identifications of these bacterial types were concentrated there rather than for FEV1 scores below 70%.

In a review by Kiedrowski and Bomberger, the authors identified the following bacterial genera as core members of the general airway microbiome: *Streptococcus*, *Prevotella*, *Veillonella*, *Rothia*, *Granulicatella*, *Gemella*, and *Fusobacterium* (6). Furthermore, the levels of these microbial types in the microbiome decrease as pathogens take over the lungs of CF patients. In our results, we found that *Fusobacterium* was positively correlated with FEV1 score, in agreement with the previous study described that higher levels of that microbial type indicate better lung function in CF patients. Similarly, in a study by Cuthbertson et al., the frequency of *Fusobacterium* species was seen to increase with improvements in FEV1 score when considering all lung disease categories (11). Although they were found in our patient samples, the top CF pathogens, such as *P. aeruginosa* and *S. aureus*, which were identified in both studies as the driving deleterious forces for CF disease progression, did not show statistically significant correlations with FEV1 score in our analysis. Looking at the scatterplots, the lack of statistical significance for the correlation seemed to come from the variability of the data (without as clear of a positive or negative trend when compared to the other scatterplots) rather than the patients not having these bacterial types in sufficiently high counts. *Actinomyces* and *Leptotrichia* microbial types are not well-known as core bacterial species nor CF pathogens, and our study is one of the first to show that they may play a role in the CF lung microbiome, with the possibility of being pushed out like *Fusobacterium* as the pathogens come along.

One notable limitation of our research was the problem of sparse data. For many of the bacterial types, the number of identifications was a non-zero value for only a couple patients, which might simply be due to sample collection using the sputum method (i.e. contamination and identification of bacteria mainly from the mouth). While this is not an issue during the data cleaning and classification processes, the abundance of zeroes hinders many of the analyses completed on the data after taxonomic classification, leading to poor performance from these analysis algorithms. Another notable limitation of our research was the reliability of the FEV1 score. The FEV1 score is considered an accurate measure of lung function in the medical field for diseases like CF, but biometric information, such as body-mass index has been shown to influence these scores when used as an indicator for lung function (12). It is important to consider adjusted scores for these measurements, whether adjusting to age, weight, or height, to eliminate any bias these factors may have on the score, but the dataset use did not indicate whether the FEV1 scores had been adjusted. Researchers have also considered the reliability of FEV1 scores in comparison to the FEV1/FVC ratio, where finding that the ratio, being a very sensitive measure of lung function for CF patients, is likely to miss several cases of obstructed pulmonary function if used alone (13). During the FEV test, FEV2 (for the second second) and FEV3 (for the third second) are also measured. FVC is the total amount of air that is forcefully expelled during exhalation throughout the entire FEV test. When we tried to apply the correlations that had been done with FEV1 to the FEV1/FVC ratio as well, there were no statistically significant hits for microbial types, further supporting the use of FEV1 score over the ratio alone when evaluating CF disease progression.

In this research, we were able to identify three statistically significant correlations between microbial types in the lung

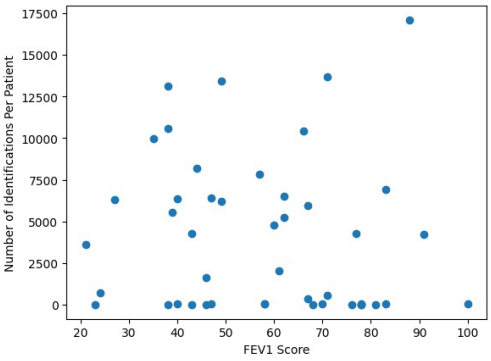

Figure 5. No statistically significant correlation between Pseudomonas and FEV1 score. FEV1 score, an effective measure of lung function in the medical field, was provided for each CF patient (n = 51) as a measure of the maximum volume of air that can be expelled from the lungs after maximum inhalation. Spearman rank correlation, $p > 0.0004$.

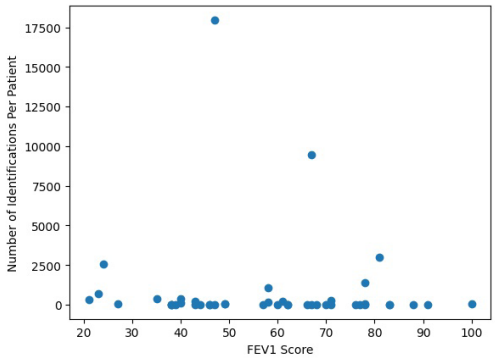

Figure 6. No statistically significant correlation between Staphylococcus and FEV1 score. FEV1 score, an effective measure of lung function in the medical field, was provided for each CF patient (n = 51) as a measure of the maximum volume of air that can be expelled from the lungs after maximum inhalation. Spearman rank correlation, $p > 0.0004$.

microbiome of CF patients and FEV1 score. Considering that each of the correlations for the microbial types showed increasing counts as FEV1 score increased as well, future research could focus not only on identifying other microbial types that were correlated with higher FEV1 scores, such as the ones the research papers above had found with statistical significance, but also on evaluating microbial types that correlated with lower FEV1 scores. Since the bacterial species that negatively correlate with FEV1 score are the ones that could be targeted with treatments, further identifying those species could improve medical insight in fighting lung infections in CF patients. For deeper evaluation of the negative and positive correlations that microbial types have with FEV1 score, or lung function, a key first step would be finding and utilizing numerous datasets with more samples, compared to the 51 samples in the dataset used for this study. Due to these limitations of the data size, we were not able to identify enough statistically significant correlations to consider whether microbial diversity decreases in the lung with CF progression, as previous studies have found. Future research could, therefore, seek to support or qualify this idea, allowing for better determination of which dominant bacteria to target for treatment.

MATERIALS AND METHODS

Data Preparation

The dataset retrieved from the QIITA database was first analyzed using the Galaxy workflow, a platform for computational biology data analysis, to determine the abundance of different bacterial types for each sample (**Figure 7**) (14).

Before the alignment of the DNA sequences from each sample to the reference database, three steps were taken to improve data quality. The sequences were first filtered based on length, identifying and removing any that consisted of more than 275 base pairs. To eliminate classifications based on high levels of uncertainty about the true DNA sequence, any sequences consisting of too many of these unidentified base pairs, as defined by the workflow system, were also removed. The third step included the process of deduplication, where unique DNA sequences were separated from the total number of sequences. By also identifying the number of sequences that are represented by each of the unique DNA sequences, deduplication maximizes efficiency during the classification process, which is seen in this data of 884,047 total sequences, but only 276,868 unique sequences.

During the sequence alignment process, these sequences were then outputted alongside reference sequences that best matched different microbial types in the reference database, a compilation of sequences corresponding to different microbial types. This process identified which sequences from the patient samples aligned with the sequences for any microbial type, which is achieved by keeping only the sequences that entirely overlap the V4 region of interest on the genome, which is one of the main places with the most nucleotide heterogeneity for the sake of classification efforts.

Two more steps were taken to improve data quality: one to directly maximize the efficiency of the classification process and the second to remove any chimeras. The first step consisted of clustering together any sequences that differed only by two or fewer base pairs. This small difference was likely to have been a result of errors during the initial machine sequencing process rather than any genetic variations signifying a completely different microbial identity. With this pre-clustering process, the resulting 276,868 sequences from the deduplication process above dropped down to 228,634 sequences. Next, the chimera removal eliminated any incorrect mixes of two or more biological sequences formed during the PCR amplification step of 16S rRNA sequencing. In general, the removal of chimeras is a crucial component of the data cleaning methodology, ensuring that these hybrid sequences are not interpreted as any novel organisms. No chimeras were found for this data, hence there was still 228,634 sequences after all the data cleaning steps.

Taxonomic Classification

Before completing taxonomic classification on the cleaned data, the remaining aligned sequences were first down sampled to 10,000 sequences (a process done randomly without replacement), taking a representative subset of the sequences to feed into the classification process. To further increase efficiency past the down sampling step, the sequences were also pre-clustered using more general similarity requirements based on the number of base pair

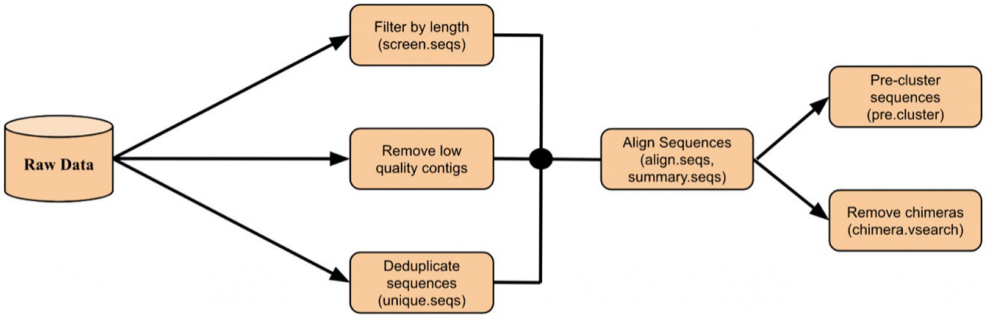

Figure 7. The galaxy workflow. Data preprocessing steps on the galaxy pipeline before taxonomic classification. Sequences that were too long, had too much ambiguity, or were duplicated were removed. Sequences were aligned with a reference database to ensure classification would yield bacterial species. Similar sequences were grouped together and any hybrid sequences joined together during PCR amplification were removed.

differences, creating larger clusters. Finally, the taxonomic classification workflow on the galaxy pipeline, with the Bayesian classifier and the Ribosome Database Project (RDP) reference taxonomy, classified the sequences from domain all the way down to genus (species was primarily left blank or unclassified). It is important to note that the classifier automatically establishes an estimated confidence level of the sequence assignments for each taxon rank. Thus, whenever this estimated level does not meet or exceed the default threshold level for the classifier, the ranks are displayed as an 'unclassified' taxon, which happened in a couple cases such as the 'unclassified Leptotrichiaceae' label. The full pipeline used to analyze our data can be found here: https://github.com/maavistar/CF-microbiome.git.

Statistical Analyses

Multiple hypothesis correction and the Spearman Rank statistical test were used to determine whether different microbial types would be indicative of lung condition, we correlated each of the species to the FEV1 scores for each patient. Before performing Spearman Rank correlations between the bacterial clusters/taxonomies and the FEV1 score variable, we decided the appropriate significance level for the hypothesis tests through the Bonferroni Correction. With the latter method, we divided 0.05 by the number of representative clusters that consisted of at least 300 sequences (simplification to 124 tested bacterial clusters).

ACKNOWLEDGEMENTS

We would like to thank the patients with CF and their families who gave consent to include their data in the QIITA database.

Received: September 27, 2022
Accepted: February 15, 2023
Published: August 29, 2023

REFERENCES

1. Guo, Jonathan, et al. "Worldwide Rates of Diagnosis and Effective Treatment for Cystic Fibrosis." *Journal of Cystic Fibrosis*, 4 Feb. 2022, https://doi.org/10.1016/j.jcf.2022.01.009.
2. "What Is Cystic Fibrosis?" *National Heart Lung and Blood Institute*, U.S. Department of Health and Human Services, 24 Mar. 2022, nhlbi.nih.gov/health/cystic-fibrosis.
3. Keogh, Ruth H, et al. "Up-to-date and Projected Estimates of Survival for People with Cystic Fibrosis Using Baseline Characteristics: A longitudinal study using UK patient registry data." *Journal of Cystic Fibrosis,* vol. 17, 2 Mar. 2018, https://doi.org/10.1016/j.jcf.2017.11.019.
4. "Cystic Fibrosis and Lung Infections: Symptoms, Types & Causes." *Cystic-Fibrosis.com*, 14 Aug. 2019, cystic-fibrosis.com/symptoms/lung-infections.
5. Lyczak, Jeffrey B., et al. "Lung Infections Associated with Cystic Fibrosis." *Clinical Microbiology Reviews* vol. 15, 2 Apr. 2002, https://doi.org/10.1128/CMR.15.2.194-222.2002.
6. Kiedrowski, Megan R., and Bomberger, Jennifer M. "Viral-Bacterial Co-Infections in the Cystic Fibrosis Respiratory Tract." *Frontiers*, 20 Dec. 2018, https://doi.org/10.3389/fimmu.2018.03067.
7. Jo, Jay-Hyun, et al. "Research Techniques Made Simple: Bacterial 16S Ribosomal RNA Gene Sequencing in Cutaneous Research." *Journal of Investigative Dermatology*, 19 Feb. 2016, https://doi.org/10.1016/j.jid.2016.01.005.
8. Szczesniak, Rhonda, et al. "Use of FEV1 in Cystic Fibrosis Epidemiologic Studies and Clinical Trials: A Statistical Perspective for the Clinical Researcher." *Journal of Cystic Fibrosis* vol. 16, 20 Jan. 2017, https://doi.org/10.1016/j.jcf.2017.01.002.
9. "A Multiomics Study on the Effects of TRIKAFTA on CF Patients." *Qiita*, qiita.ucsd.edu/study/description/13507.
10. Ponce, Mario C, et al. *Pulmonary Function Tests*. National Library of Medicine, 3 Sept. 2022, ncbi.nlm.nih.gov/books/NBK482339/.
11. Cuthbertson, Leah, et al. "Lung Function and Microbiota Diversity in Cystic Fibrosis - Microbiome." *BioMed Central*, 2 Apr. 2020, https://doi.org/10.1186/s40168-020-00810-3.
12. Ishikawa, Caren, et al. "Comparison of Body Composition Parameters in the Study of the Association Between Body Composition and Pulmonary Function." *BMC Pulmonary Medicine* vol. 21,1 178. 25 May 2021, https://doi.org/10.1186/s12890-021-01543-1.
13. König, Peter, et al. "Is an FEV1 of 80% Predicted a

Normal Spirometry in Cystic Fibrosis Children and Adults?" *Clinical Respiratory Journal*, Aug. 2018, https://doi.org/10.1111/crj.12920.

14. Hiltemann, Saskia, et al. "Galaxy Training: 16s Microbial Analysis with Mothur (Short)." *Galaxy Training Network*, Galaxy Training Network, 1 Jan. 1970, training.galaxyproject.org/training-material/topics/metagenomics/tutorials/mothur-miseq-sop-short/tutorial.html#data-cleaning.