**Analysis of the Lung Microbiome in Cystic Fibrosis Patients Using 16S Sequencing**

**Manasvi Pinnaka1, Brianna Chrisman2**

1 Basis Independent Silicon Valley, San Jose, California

2 Department of Bioengineering, Stanford University, Palo Alto, California

**Student Authors**

Manasvi Pinnaka – high school

**SUMMARY**

Cystic fibrosis patients often develop lung infections that range anywhere in severity from mild to life-threatening due to the presence of thick and 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. With a publicly available dataset of DNA sequences from bacterial species in the lung microbiome of cystic fibrosis 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. For our statistical analyses, we distinguished between taxonomies using this referencing and determined the proportions of a certain taxa relative to another. We found that the Fusobacterium, Actinomyces, and Leptotrichia microbial types all had positive correlation with 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 by past literature. Examining the lung microbiology of cystic fibrosis patients can help with the prediction of the current condition of lung function, with the potential to guide doctors when designing personalized treatment plans for patients.

**INTRODUCTION**

More than 160,000 people worldwide suffer from cystic fibrosis (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 damage to organs like the lungs (2). While there have been advances in treatments for CF, babies 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 chance of survival. One setback, however, for this approach is that there are no specific stages doctors have defined to be indicators of disease severity, considering the high variability in the complications and symptoms that each patient faces.

One possible avenue researchers can instead look to for an indication of disease severity is lung microbiome compositions, or the types of different microorganisms and their abundance in the lung, of patients. Nearly every single CF patient is expected to experience lung infections as their disease progresses (4), and approximately 80 to 95% of CF patients experience chronic lung infections and inflammation that eventually lead to respiratory failure (5). Not only are chronic lung infections correlated with the increasing fatality of CF, especially with the presence of the pathogen *Pseudomonas Aeruginosa*, but past studies have also shown that patients who show less microbial diversity in the lung also have decreased lung function (6). This research aims to similarly 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.

The goal of 16S sequencing is to determine the microbiome composition of a specific sample, specifically distinguishing between taxonomies using reference genomes and deciding 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 OTUs, or operational taxonomic units (formed through sequence similarity), are grouped based on how similar they are to the current reference sequences from the reference database, which consists of different rRNA gene sequences that each correspond to specific taxa, allowing for a determination of microbiome composition (7).

We hypothesize that dominant cystic fibrosis pathogens, such as *Pseudomonas Aeruginosa* and *Staphylococcus Aureus*, will have a strong, negative correlation with the FEV1 score variable, which serves as a measure of pulmonary function for cystic fibrosis patients. This study utilizes 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.

**RESULTS**

1. *Overview*

The open-source dataset we used for this research consists of 51 samples of the lung microbiome from cystic fibrosis patients before and after interventions with Trikafta therapy.

Taxonomic classification of the DNA sequences from the QIITA dataset gave us four outputs:

1. Sample metadata: This contained sample information, notably FEV1 and FVC scores for each patient.
2. Representative sequence table: This was a clustered data matrix of the different sequences for each patient and their corresponding representative sequences.
3. Count table: This contained the number of times each initial sequence had corresponded to each patient.
4. 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.

Using multiple hypothesis correction and the Spearman Rank statistical test 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 available in the table of sample information. FEV1, or forced expiratory volume in 1 second, is considered a relevant and effective measure of lung function in the medical field when tracking the progression of cystic fibrosis (8). Before performing Spearman Rank correlations between the bacterial clusters/taxonomies and the FEV1 score variable, we decide the appropriate significance level for the hypothesis tests through the Bonferroni Correction method, under which we divided 0.05 by the number of representative clusters that consisted of at least 300 sequences (simplification to 124 tested bacterial clusters).

1. *Fusobacterium is positively correlated with FEV1.*

Using the procedure described above, we found that Fusobacterium was positively correlated with FEV1 score. As the FEV1 score increases, we can see a positive and moderately strong association between the two variables (p value = 9.863 x 10^-7, r2 = 0.662, Figure 1).

1. *Actinomyces is positively correlated with FEV1.*

We also found that Actinomyces was positively correlated with FEV1 score. As the FEV1 score increases, we can once again see a positive and moderately strong association between the two variables (p value = 1.074 x 10^-4, r2 = 0.551, Figure 2).

1. *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 x 10^-4, r2 = 0.517, Figure 3).

This positive correlation was present between the Leptotrichia genus classification and FEV1 score as well. 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 value = 1.436 x 10^-4, r2 = 0.542, Figure 4). With this correlation, it should be noted that the genus came up as unclassified Leptotrichiaceae, indicating that the genus classification is Leptotrichiaceae, but the confidence level was not high enough for the classification system to fully define it as such. Nevertheless, alongside Figure 3, the correlation between this microbial type and FEV1 score is still evident.

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. Thus, the relationship between the Leptotrichia bacteria and FEV1 score is further supported by how a moderately strong and statistically significant correlation is seen with both analysis of the bacterial cluster and analysis of the genus.

**DISCUSSION**

From our study, we found that Fusobacterium, Actinomyces, and Leptotrichia were all positively correlated with FEV1 score in cystic fibrosis patients. In a 2018 literature 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. Furthermore, they found that the levels of these microbial types in the microbiome decrease as pathogens take over the lungs of cystic fibrosis patients (6). In our results, we found that Fusobacterium was positively correlated with FEV1 score, in agreement with the authors that greater levels of that microbial type indicate better lung function in cystic fibrosis patients. Similarly, in a 2020 study by Cuthbertson et al., the frequency of the Fusobacterium species was seen to increase with improvements in FEV1 score when considering all lung disease categories (9). However, although they were found in our patient samples, the top cystic fibrosis pathogens, such as Pseudomonas Aeruginosa and Staphylococcus Aureus, which were identified in both studies as the driving deleterious forces for cystic fibrosis disease progression, did not show statistically significant correlations with FEV1 score in our analyze Actinomyces and Leptotrichia microbial types, are not well-known as a core bacterial species nor a cystic fibrosis pathogen, and our study is the first to show that they may play a role in the CF lung microbiome

One notable limitation of our research was the problem of sparse data. In the count matrix, which showed the counts of each representative sequence for each patient, there were primarily zeroes in nearly every row. 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 cystic fibrosis, but biometric information, such as BMI (10), has been shown to influence these scores when used as an indicator for lung function. 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 there is no specific indication as to what the FEV1 scores from the dataset have been adjusted to. Researchers have also considered the reliability of FEV1 scores in comparison to the FEV1/FVC ratio, finding that the ratio, being a very sensitive measure of lung function for cystic fibrosis patients, is likely to miss several cases of obstructed pulmonary function if used alone (11). 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 cystic fibrosis disease progression.

In this research, we were able to identify three statistically significant correlations between microbial types in the lung microbiome of cystic fibrosis 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 doctors need to target with their treatments, further identifying those species could improve medical insight in fighting lung infections in cystic fibrosis 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. While we were not able to identify sufficient statistically significant correlations to consider whether microbial diversity decreases in the lung with cystic fibrosis progression, as past research states, future research could seek to support or qualify this idea as well, allowing for better determination of which dominant bacteria types to target using treatments.

**MATERIALS AND METHODS**

*A. Data Preparation* (Figure 5)

The dataset we retrieved from the QIITA database was first analyzed using the Galaxy workflow (12) to determine the abundance of different bacterial types for each sample.

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.

When sequencing machines are unable to determine the true nucleotide for a certain segment of the sequence, an ambiguous nucleotide is outputted instead. 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 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 the best matching sequences in the reference database, a compilation of sequences corresponding to different microbial types. This process identified which sequences could even align with the sequences for any microbial type in the first place, which is achieved by keeping only the sequences that entirely overlap the V4 region of interest on the genome.

At this point, two more steps were taken for the sake of 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 only differed by two or fewer base pairs, considering that this trivial 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 came the chimera removal, which essentially 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, but since no chimeras were found for this data, there were still a total number of 228,634 sequences after all the data cleaning steps.

*B. Data Classification*

Before completing taxonomic classification on the cleaned data, the remaining aligned sequences were first down sampled to 10,000 sequences, essentially 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, creating larger clusters. Finally, the taxonomic classification workflow on the galaxy pipeline, with the Bayesian classifier and the Ribosome Database Project (RDP) reference taxonomy, the sequences were classified from domain all the way down to genus (species was primarily left blank or unclassified).

**ACKNOWLEDGEMENTS**

We thank patients with cystic fibrosis and their families for having given consent to include their data in the QIITA database.

**REFERENCES**

1. Guo, Jonathan, et al. “Worldwide Rates of Diagnosis and Effective Treatment for Cystic Fibrosis.” *Journal of Cystic Fibrosis*, 4 Feb. 2022, www.sciencedirect.com/science/article/pii/S1569199322000315.

2. “What Is Cystic Fibrosis?” *National Heart Lung and Blood Institute*, U.S. Department of Health and Human Services, www.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. 2018, doi:10.1016/j.jcf.2017.11.019

4. “Cystic Fibrosis and Lung Infections: Symptoms, Types & Causes.” *Cystic-Fibrosis.com*, cystic-fibrosis.com/symptoms/lung-infections.

5. Lyczak, Jeffrey B., et al. “Lung Infections Associated with Cystic Fibrosis.” *Clinical Microbiology Reviews* vol. 15,2. 2002, doi:10.1128/CMR.15.2.194-222.2002

6. Kiedrowski, Megan R., and Jennifer M. Bomberger. “Viral-Bacterial Co-Infections in the Cystic Fibrosis Respiratory Tract.” *Frontiers*, 20 Dec. 2018, www.frontiersin.org/articles/10.3389/fimmu.2018.03067/full.

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, www.sciencedirect.com/science/article/pii/S0022202X16003481.

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, doi:10.1016/j.jcf.2017.01.002

9. Cuthbertson, Leah, et al. “Lung Function and Microbiota Diversity in Cystic Fibrosis - Microbiome.” *BioMed Central*, 2 Apr. 2020, microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00810-3.

10. 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, doi:10.1186/s12890-021-01543-1

11. König, Peter, et al. “Is an FEV1 of 80% Predicted a Normal Spirometry in Cystic Fibrosis Children and Adults?” *Clinical Respiratory Journal*, Aug. 2018, doi: 10.1111/crj.12920. PMID: 29920961.

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

**Figures and Figure Captions**

**Chart, scatter chart

Description automatically generated**

***Figure 1. Positive correlation between Fusobacterium and FEV1 score.*** *Scatterplot relationship between Fusobacterium and FEV1 score. Spearman rank correlation, \*\*\*p<0.0004*

**Chart, scatter chart

Description automatically generated**

***Figure 2. Positive correlation between Actinomyces and FEV1 score.*** *Scatterplot relationship between Actinomyces and FEV1 score. Spearman rank correlation, \*\*\*p<0.0004*

**Chart, scatter chart

Description automatically generated**

***Figure 3. Positive correlation between Leptotrichia (bacterial cluster) and FEV1 score.*** *Scatterplot relationship between Leptotrichia (bacterial cluster) and FEV1 score. Spearman rank correlation, \*\*\*p<0.0004*

**Chart, scatter chart

Description automatically generated**

***Figure 4. Positive correlation between Leptotrichia (genus) and FEV1 score.*** *Scatterplot relationship between Leptotrichia (Genus) and FEV1 score. Spearman rank correlation, \*\*\*p<0.0004*

Diagram

Description automatically generated

***Figure 5. The galaxy workflow.*** *Flowchart showing the galaxy pipeline steps*

**Appendices**

[*https://github.com/maavistar/CF-microbiome.git*](https://github.com/maavistar/CF-microbiome.git)