META-ANALYSIS OF LUNG MICROBIOME IN CYSTIC FIBROSIS PATIENTS

Submitted by

Srijan Banerjee

M.Tech Bioinformatics

MAKAUT, WB

Under Joint supervision of

Dr.Sudipto Saha and Prof. Raja Banerjee

বসু বিজ্ঞান মন্দির
বসু विज्ञान मंदिर
BOSE INSTITUTE

MAKAUT WB
Utech
In Pursuit Of Knowledge And Excellence

Summary

Cystic Fibrosis is an autosomal recessive disease caused by the presence of mutations in both copies of the gene for the cystic fibrosis transmembrane conductance regulator (CFTR) protein. Cystic Fibrosis is mostly associated with lung infection and severe damages to the pancreas, liver, kidneys, and intestine. Acute increase in Pulmonary exacerbation which occur throughout the patients' conditions in Cystic fibrosis are treated in part with oral, inhaled and intravenous antibiotics. Several studies were performed in identifying gut and lung microbiome in healthy and diseased condition. We have collected and compiled the gut and lung microbiome studies from human and other experimental animal from PubMed and developed a manually curated database named as GLMdb.Currently,the database includes 265 and 150 gut and lung microbiome published literature respectively.The database provides the knowledge of gut and lung microbiome studies performed separated for respiratory for respiratory and metabolic diseases. This database includes information on sample species (eg.human,monkey,rat),hypervariable region of 16S rRNA, links to the raw sequence data (Bioproject ID/ SRA ID from NCBI/ENA etc.) and disease associated with altered lung and/or gut microbiome. The database is freely available at http://bicresources.jcbose.ac.in/ssaha4/glmdb/index.php. From GLMdb we found that there are total 39 Lung microbiome studies available for Cystic Fibrosis. Among those only 4 studies of lung microbiome have sputum sample, V4 region amplified Paired end raw sequence layouts which are selected for 16s rRNA sequencing analysis further. Mothur (Version 1.42.1) was used to analyse the 16s rRna sequencing data of the selected samples and further performed meta-analysis by systemically combining the previous studies.Although there was variation in patient age and sample collection methods,still we observed dysbiosis of lung microbiome profiling to the respiratory diseases. We have also the complex interplay in 10 different known pathogen of Cystic fibrosis in between Cystic fibrosis lung and healthy lung microbiome.we have also shown how those known pathogen altered in different antibiotic conditions and different Cystic fibrosis patients condition from 2 different antibiotic studies.

Introduction

Acute and chronic lung infection with significant morbidity and early mortality is associated with the autosomal recessive disease Cystic fibrosis which causes not only progressive lung infections but also severe damages to the pancreas, liver, kidneys, and intestine [9]. In people with Cystic fibrosis, a mutation in the gene of CFTR causes a thick, sticky buildup of mucus in the lungs, pancreas, and other organs. The CFTR protein is a channel protein that controls the flow of H2O and Cl- ions and bicarbonate ions in and out of cells inside the lungs. When the CFTR protein is working correctly, ions freely flow in and out of the cells [6]. However, when the CFTR protein is malfunctioning, these ions cannot flow out of the cell due to a blocked channel. This results in the formation of dehydrated and viscous mucus and subsequently aberrant mucosa in the lungs and digestive tract, increasing the risk of recurrent and chronic pulmonary infection and inflammation, pancreatic insufficiency (PI), CF-related liver disease (CFRLD) and diabetes (CFRD)[11].

This mucus act as potential source of nutrients which leads to the formation of bacterial microenvironments known as biofilms.Sometimes these mucus are difficult for immune cells and antibiotics to penetrate. This microbial environment is responsible for progressive lung infections whose interaction contributes to respiratory failure.Despite their importance, the complex interplay between the lung microbiota and the host environment is poorly understood. As there has been an association between a loss of bacterial diversity and progression of lung disease the lung and Gut microbiome in CF is of interest [10].

Several cystic fibrosis lung microbiome studies have shown that *Pseudomonas aeruginosa, Staphylococcus aureus ,Haemophilus influenza, Burkholderia cepacia* complex are the most conventional pathogen in Cystic fibrosis lung infection. Improvements in airway clearance and more effective treatment of the conventional CF pathogens has led to the emergence of new airway pathogens such as Methicillin-resistant *Staphylococcus aureus, Stenotrophomonas maltophilia, Mycobacterium abscessus, and Achromobacter xylosoxidans , Streptococcus salivarius, Granulicatella adiacens, Prevotella melaninogenica, Rothia mucilaginosa, Veillonella dispar, Streptococcus mitis group (oralis, dentisani, mitis, infantis*).There are some other genera apart from the conventional pathogens of Cystic fibrosis like *Actinomyces, Gemella, Fusobacterium, Neisseria, Atopobium* sometimes found in several studies in different Cystic fibrosis condition.

There is also an bi-directional relationship between Cystic fibrosis physiology and gut and lung microbiota[2]. Till date,several studies were performed in identifying gut and lung microbiome in healthy and diseased condition of Cystic fibrosis patients.Gut microbiota is also be important in nutrient and energy metabolism in CF. Duytschaever et al., 2011 showed that Predominant genera (from highest to lowest relative abundance) of gut microbiome are *Bacteroides, Bifidobacterium, Veillonella, Clostridium, Blautia, Parabacteroides, Streptococcus, Lachnospira* etc.

Acute increase in Pulmonary exacerbation which occur throughout the patients conditions in Cystic fibrosis are treated in part with oral,inhaled and intravenous antibiotics.Sometimes when antibiotic resistance is presumed or detected antibiotics with a broader range of antimicrobial activity are used. Therefore Often the chronic use of multiple antibiotics increase with possibility of multidrug resistance (resistance to three or more antibiotic categories) which severely limits the options of antibiotic coverage in patients with advanced disease.

Here in this project work we collected and compiled the gut and lung microbiome studies from human and other experimental animal from PubMed and developed a manually curated database named as GLMdb.The database focuses on lung microbiome and the association of gut and lung microbiome, also known as gut-lung axis. This database provides information on sample species (eg.human,monkey,rat),hypervariable region of 16S rRNA, links to the raw sequence data (Bioproject ID/ SRA ID from NCBI/ENA etc.) and disease associated with altered lung and/or gut microbiome. Currently the database includes 265 and 150 gut and lung microbiome published literature.

From GLMdb we found that there are total 39 Lung microbiome 6 gut microbiome studies available for Cystic Fibrosis. Among those only 4 studies of lung microbiome have sputum sample, V4 region amplified Paired end raw sequence layouts which are selected for 16s rRNA sequencing analysis further. Among them there are 2 antibiotics studies.

Mothur (Version 1.42.1)was used to analyse the 16s rRna sequencing data of the selected samples and further performed meta-analysis by systemically combining the previous studies.We have performed lung microbiome 16s rRNA analysis among group of Cystic fibrosis studies and Cystic fibrosis antibiotic studies and also from healthy groups sample of 2 different asthma studies. We have checked the complex interplay in 10 different known pathogen of Cystic fibrosis in between Cystic fibrosis lung and healthy lung microbiome.we

have also showed how those known pathogen altered in different antibiotic conditions and different Cystic fibrosis patients condition from those different antibiotic studies.

As a meta-analysis approach Pearson's correlation coefficient is checked for some of the most abundant Operational Taxonomic Units in Cystic fibrosis Lung and Healthy group's sample.In this research project work we have tried to identify relationship of human lung microbiome to lung function of Cystic fibrosis from those approaches.

Review of literature

Several clinical studies have demonstrated associations between the human Gut and lung microbiome and respiratory disease, yet fundamental questions remain on how we can generalize this knowledge. Results from individual studies can be inconsistent, and comparing published data is further complicated by a lack of standard processing and analysis methods.

Very few databases provide the knowledge of gut, lung oral or other microbiota alteration to its specific disease. Overall, our current understanding of the precise relationships between the human gut and lung microbiome and diseases remains limited.

So here in this research work we made an effort to collect and compile the gut and lung microbiome studies from human and other experimental animal from PubMed and developed a manually curated database named as GLMdb.

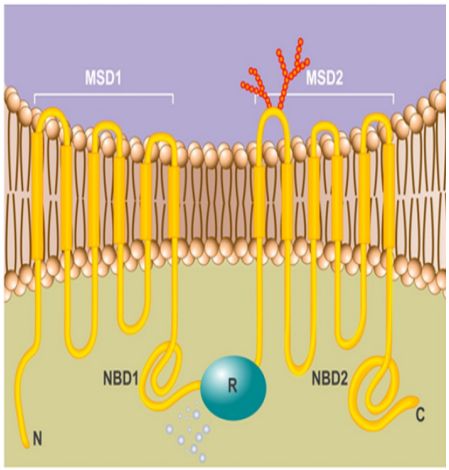
Cystic Fibrosis is an autosomal recessive disease, mostly associated with lung infection and severe damages to the pancreas, liver, kidneys, and intestine[9].In people with Cystic Fibrosis, a defective gene of CFTR causes a thick, sticky build up of mucus in the lungs, pancreas, and other organs[1]

We are aware of the fact that more than 30,000 people are living with Cystic fibrosis (more than 70,000 worldwide). Approximately 1,000 new cases of Cystic fibrosis are diagnosed each year. More than 75 percent of people with Cystic fibrosis are diagnosed by age 2.More than half of the Cystic fibrosis patient population age is 18 or older.It is reported that CF is found to be rare in persons of nonCaucasian origin.[17]

As a historical overview we found the very first reference of this genetic disease Cystic fibrosis from a Eighteenth century German and Swiss literatures warned: “Wehedem Kind, das beim Kuß auf die Stirn salzig schmekt, es ist verhext und muss bald sterben”, which can be translated as: “Woe to that child which when kissed on the forehead tastes salty; he is bewitched and soon must die.” Dorothy Andersen, a pathologist at the New York Babies Hospital made the first clear description of Cystic fibrosis was in 1938 by; her paper entitled “Cystic fibrosis of the pancreas and its relation to celiac disease”[18]. More knowledge on the underlying pathophysiology of the disease was brought about in the 1980s with the description of chloride impermeability of the Cystic fibrosis sweat gland. The CFTR gene was cloned in 1989[20].

The CFTR gene is located in the long arm of chromosome 7 (7q31.2)[21]. The encoded protein functions mainly as an adenosine 3′,5′-cyclic monophosphate (cAMP)-regulated chloride channel in a variety of polarized epithelia. The Cystic Fibrosis gene is large approximately 250 kb, and contains 27 exons. The encoded mRNA is about 6.5 kb long and is translated into a protein product of 1480 amino acids. The amino acid sequence of CFTR protein shows significant homology to the family of ATPbinding cassette (ABC) transporters. The predicted protein structure is composed of two repeated units, each consisting of a membrane spanning domain (MSD) comprising of six hydrophobic transmembrane helices, followed by a nucleotide-binding domain (NBD) that interacts with ATP. Ten of the 12 transmembrane helices contain one or more charged amino acids, and two potential glycosylation sites are found between helices 7 and 8. The two repeated units are linked by a single regulatory (R) domain that contains 9 of the 10 consensus sites for phosphorylation by protein kinase A (PKA) and 7 of the phosphorylation sites for protein kinase C (PKC). The R domain separates the two MSDs and interacts physically with the NBD1. The R domain is unique for CFTR as it is not present in the other members of the ABC superfamily. The protein domains assemble to line the pore of the anion-selective channel through which chloride flows across the plasma membrane. Anion flow through the channel is believed to be gated by cAMP-dependent PKA phosphorylation of the R domain. and by the interaction of ATP to NBD sites that induces conformational changes in the protein, finally resulting in its opening and closing statuses.[17]. As a transepithelial anion channel, CFTR provides a pathway for chloride, gluconate and bicarbonate transport. it has been shown that stimulation of CFTR by cAMP agonists inhibits the amiloride sensitive epithelial Na+ channel,

ENaC.That ENaC activity is increased in CF respiratory epithelia. Moreover, It is reported that CFTR was shown to be expressed in intracellular vesicles where it may play a role in intracellular and intravesicular pH regulations. CFTR also seems to control exocytosis/endocytosis processes.CFTR is mainly located at the apical membrane of polarized epithelial tissue but it may be found in it is also found in a number of non-epithelial tissues such as cardiac myocyte, smooth muscle, erythrocytes, and immune cells such as macrophages.[22]



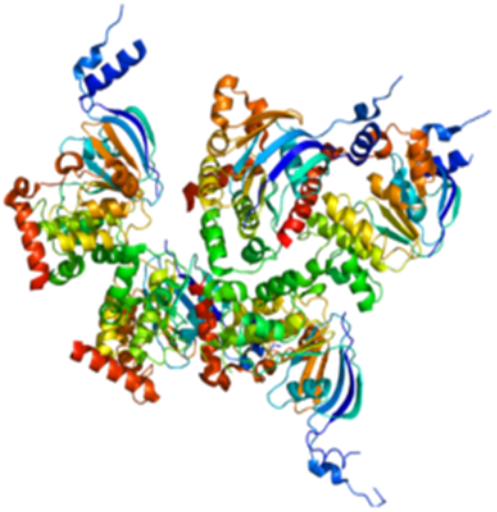


Fig1: Predicted topology and protein structure of CFTR(adapted from Lubamba et al[17])

Fig:PDB id 1XMI(human CFTR)

Under normal circumstances, the CFTR gene undergoes transcription and is translated into a CFTR protein that traffics to the cell membrane where it fully functions as a chloride channel. In Cystic fibrosis, the majority of CFTR mutations involve changes in three or fewer nucleotides and result in amino acid substitutions, frame shifts, splice site, or nonsense mutations. The most common and first identified mutation, the F508del, corresponds to a three base pair deletion that codes for phenylalanine at position 508 of the CFTR protein. But the relative frequency of the F508del mutation in families carrying the CF gene varies between groups. The presence of the F508del mutation increases the frequency of CF in

Caucasian population comparative to other races. The classification of CFTR mutation is summarized in table 1 and table 2(adapted from Lubamba et al[17]).

Table1:Severe CF phenotype

Class	Mutation prototypes	Consequences
I	G542X, W1282X, R553X, 3950delT	CFTR is not synthesized because of stop codons or splicing defects
II	F508del, N1303K	CFTR is synthesized but in an immature form (only partly glycosylated, misfolded, not released from the endoplasmic reticulum) and is mostly degraded by the ubiquitin–proteasomal pathway.
III	G551D	CFTR is synthesized and transported to the plasma membrane, but its activation and regulation by ATP or cAMP are disrupted
VI	1811+1.6 kb A>G	CFTR is synthesized, but membrane stability or conductance of ions other than chloride is reduced

Table2:Milder CF phenotype:

Class	Mutation prototypes	Consequences
IV	R334W, G314E, R347P, D1152H	CFTR is synthesized and expressed at the plasma

		membrane, but chloride conductance is reduced
V	3849+10 kb C>T, 3272-26 A>G	CFTR synthesis or processing is partly defective

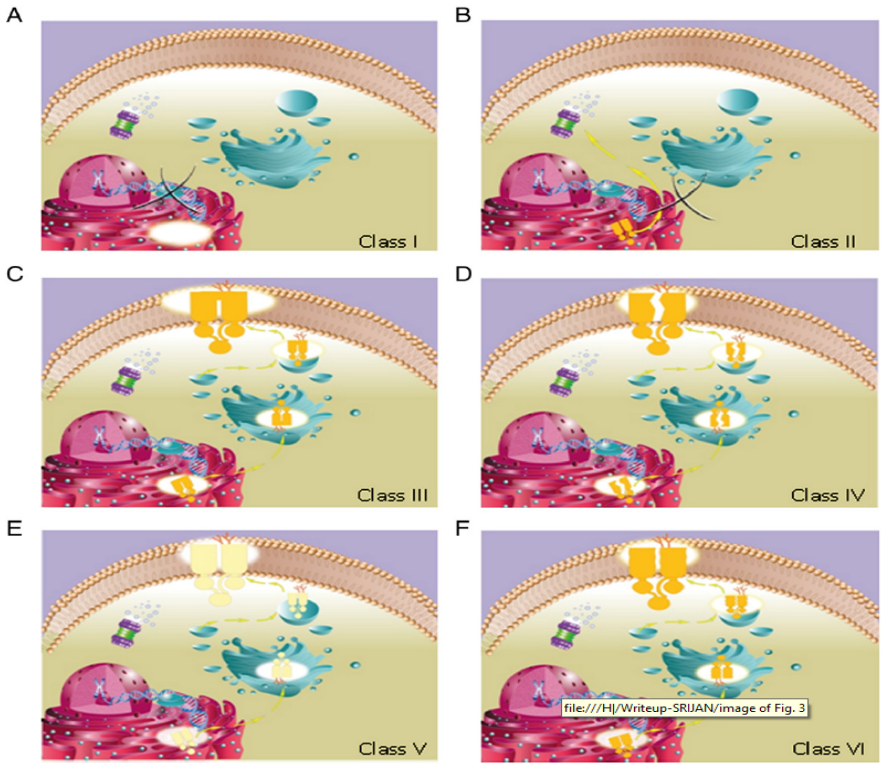


Fig:3The six classes of CFTR mutations(adapted from Lubamba et al[17])

The signs and symptoms may vary from patients to patients, depending on the severity of the disease. It has been seen that People with cystic fibrosis have a higher than normal level of salt in their sweat. Most of the other signs and symptoms of cystic fibrosis affect the respiratory system and digestive system. It has been reported that adults diagnosed with cystic fibrosis are more likely to have atypical symptoms, such as (pancreatitis), infertility and recurring pneumonia.[7]. CF respiratory phenotype is characterized by wheezing, bronchiectasis, Chronic infections, nasal polyps, hemoptysis, pneumothorax leading to respiratory failure and death. Digestive signs and symptoms can be charecterized by

foul-smelling, greasy stools, poor weight gain and growth, severe constipation, rectal prolapse in children. Early gastrointestinal manifestations include meconium ileus that occurs in 10–17% of patients within the first days of life. Cystic fibrosis males are infertile due to absence of vasa deferentia[18].

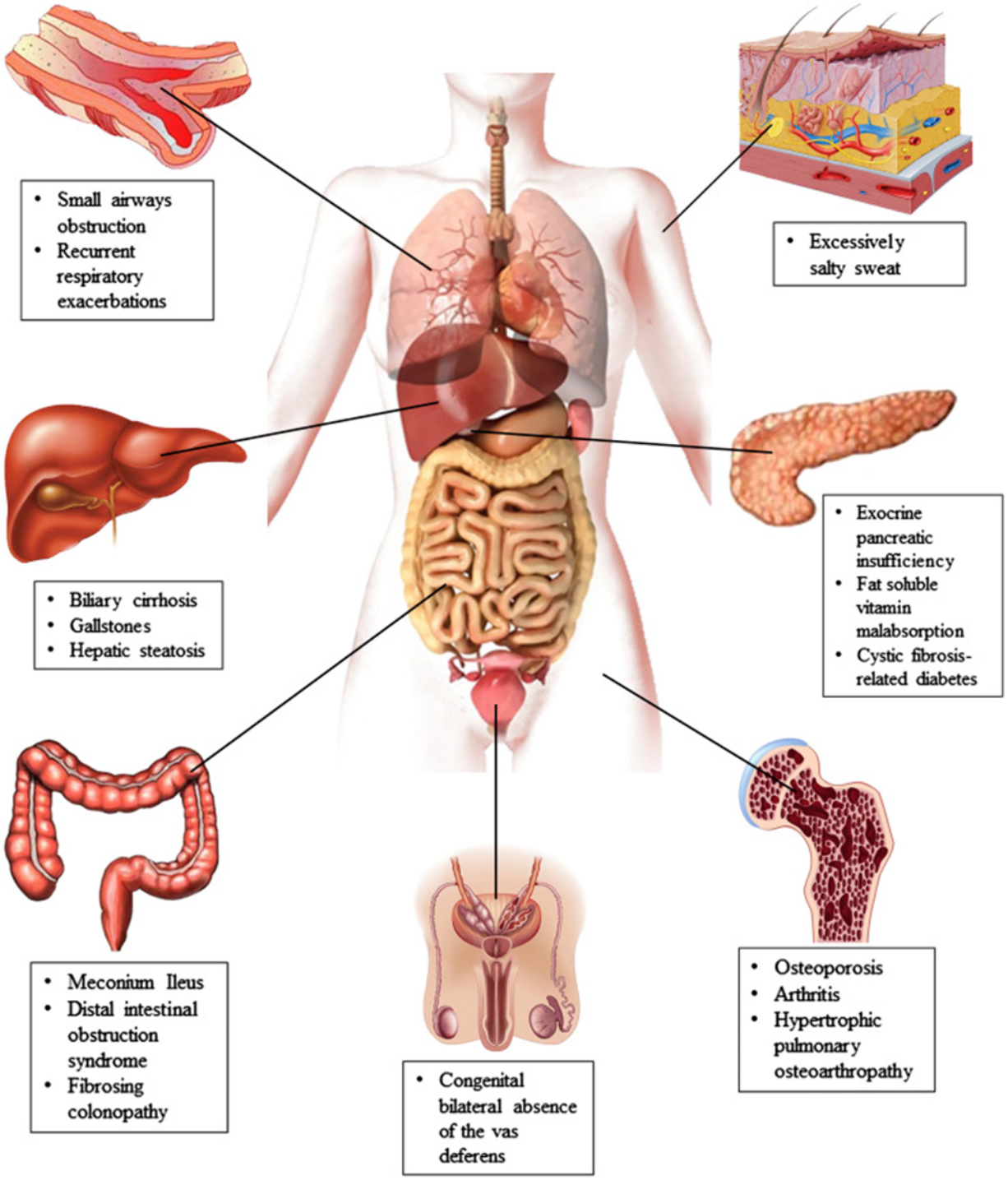


Fig 4:sign and symptoms of Cystic fibrosis.(adapted from Molina et al[18]).

The dehydrated and viscous mucus in Cystic fibrosis patients act as potential source of nutrients which leads to the formation of bacterial microenvironments known as biofilms.Sometimes these mucus are difficult for immune cells and antibiotics to penetrate.This microbial environment is responsible for progressive lung infections whose interaction contributes to respiratory failure.

Despite their importance, the complex interplay between the lung microbiota and the host environment is poorly understood. As there has been an association between a loss of bacterial diversity and progression of lung disease the lung and Gut microbiome in Cystic fibrosis is of interest.

Several cystic fibrosis lung microbiome studies have shown that Pseudomonas aeruginosa, *Staphylococcus aureus,Haemophilus influenza,Burkholderiacepacia* complex are the most conventional pathogen in Cystic fibrosis lung infection.Improvements inairway clearance and more effective treatment of the conventional Cystic fibrosis pathogens has led to the emergence of new airway pathogens such as Methicillin-resistant *Staphylococcus aureus,Stenotrophomonas maltophilia,Mycobacterium abscessus, and Achromobacter xylosoxidans,Streptococcus,salivarius,Granulicatella adiacens,Prevotella melaninogenica,Rothia mucilaginosa,Veillonella dispar;Streptococcus mitis group (oralis, dentisani, mitis, infantis)*.There are some other genera apart from the conventional pathogens of Cystic fibrosis like *Actinomyces,Gemella,Fusobacterium,Neisseria,Atopobium*sometimes found in several studies in different Cystic fibrosis condition.[24]

However there is an bi-directional relationship between Cystic fibrosis physiology and gut and lung microbiota[2].Till date,several studies were performed in identifying gut and lung microbiome in healthy and diseased condition of Cystic fibrosis patients.Gut microbiota is also be important in nutrient and energy metabolism in CF.Duytschaever et al., 2011 showed that Predominant genera (from highest to lowest relative abundance) of gut microbiome are *Bacteroides, Bifidobacterium, Veillonella, Clostridium, Blautia, Parabacteroides,*

Streptococcus, Lachnospira etc where as for lung microbiome are Streptococcus, Veillonella, *Prevotella, Bacteroides, Neisseria, Haemophilus, Gemella, Staphylococcus* etc[12].

Acute increase in Pulmonary exacerbation which occur throughout the patients conditions in Cystic fibrosis are treated in part with oral,inhaled and intravenous antibiotics.Sometimes when antibiotic resistance is presumed or detected antibiotics with a broader range of antimicrobial activity are used. Therefore Often the chronic use of multiple antibiotics increase with possibility of multidrug resistance (resistance to three or more antibiotic categories) which severely limits the options of antibiotic coverage in patients with advanced disease.

So we sought to determine to perform meta-analysis for identifying the heterogeneity of lung microbiome between the Cystic fibrosis phenotype and healthy lungs.We have also tried to find out the antibiotic susceptibility among some conventional and emerging pathogen of Cystic fibrosis.

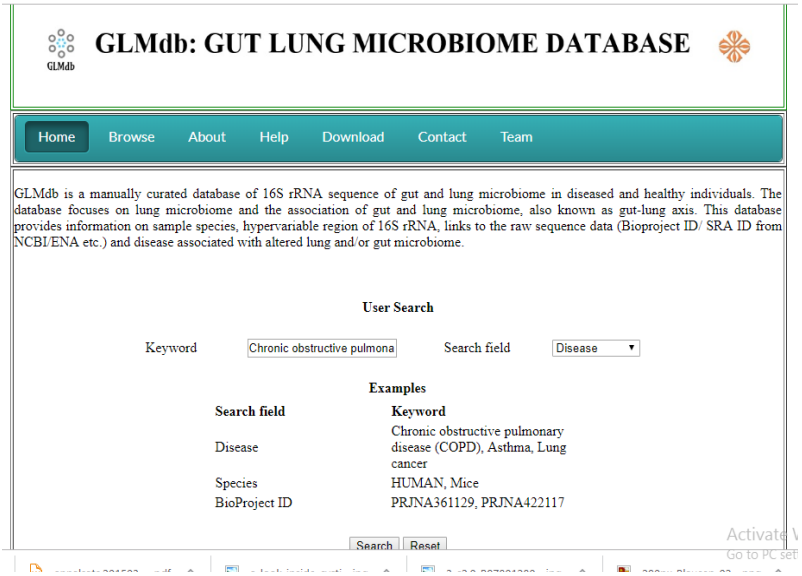
Methodology

Dataset selection:

Literature data was collected by using search engine PubMed, search query we have used [((16S) AND lung) AND microbiome] and [(Pyrosequencing) AND gut microbiota human]. All the information related to the sequencing experiments such as variable region, type of sample, disease information, model species, Bioproject ID, SRA ID etc. along with author and paper information were collected, manually curated and submitted to the database GLMdb.

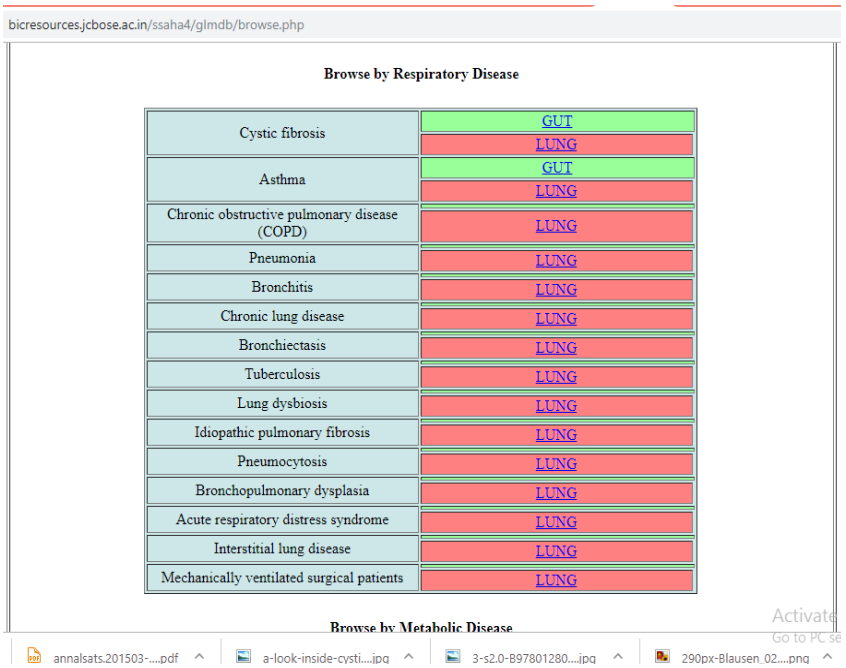
GLMdb is a manually curated database of 16S rRNA sequence of gut and lung microbiome in diseased and healthy individuals. The database focuses on lung microbiome and the association of gut and lung microbiome, also known as gut-lung axis. The database aims to bridge the knowledge of gut and lung microbiome studies performed separated for respiratory for respiratory and metabolic diseases. This database provides information on sample species (eg.human,monkey,rat),hypervariable region of 16S rRNA, links to the raw sequence data (Bioproject ID/ SRA ID from NCBI/ENA etc.) and disease associated with altered lung and/or gut microbiome. The database is freely avialable at

http://bicresources.jcbose.ac.in/ssaha4/glmdb/index.php . Currently the database includes 265 and 150 gut and lung microbiome published literature.

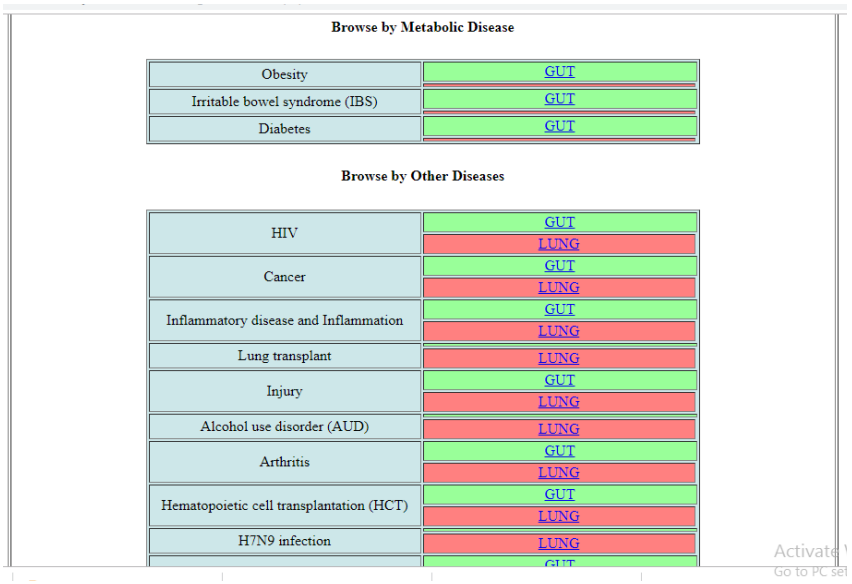


Screenshot1: GLMdb home page.

In GLMDB we can browse by either 'Respiratory diseases' or by 'Metabolic diseases'.There are an another category called 'Other disease'.In Respiratory disease group 15 disease are included.In Metabolic diseases till date only Obesity, Irritable bowel syndrome (IBS) and Diabetes is added.Several other metabolic disease will be included in future. HIV, Cancer, Inflammatory disease and Inflammation, Lung transplant, Injury, Alcohol use disorder (AUD), Arthritis, Hematopoietic cell transplantation (HCT), H7N9 infection, Healthy categories are included in Other disease.



Screenshot2: GLMdb browse page of Respiratory diseases



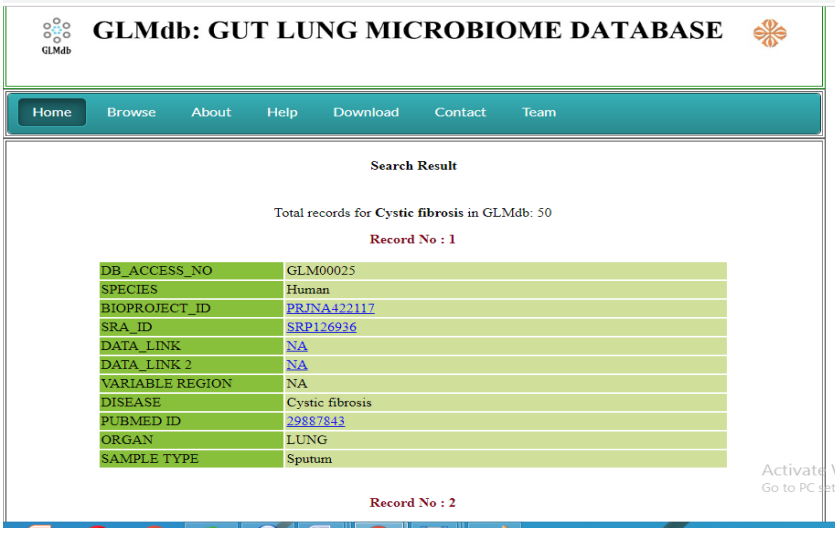
Browse by Metabolic Disease

Obesity	GUT
Irritable bowel syndrome (IBS)	GUT
Diabetes	GUT

Browse by Other Diseases

HIV	GUT
	LUNG
Cancer	GUT
	LUNG
Inflammatory disease and Inflammation	GUT
	LUNG
Lung transplant	LUNG
Injury	GUT
	LUNG
Alcohol use disorder (AUD)	LUNG
Arthritis	GUT
	LUNG
Hematopoietic cell transplantation (HCT)	GUT
	LUNG
H7N9 infection	LUNG

Screenshot3: GLMdb browse page of Metabolic and Other disease



GLMdb: GUT LUNG MICROBIOME DATABASE

Home | Browse | About | Help | Download | Contact | Team

Search Result

Total records for **Cystic fibrosis** in GLMdb: 50

Record No : 1

DB_ACCESS_NO	GLM00025
SPECIES	Human
BIOPROJECT_ID	PRJNA422117
SRA_ID	SRP126936
DATA_LINK	NA
DATA_LINK 2	NA
VARIABLE REGION	NA
DISEASE	Cystic fibrosis
PUBMED ID	29887843
ORGAN	LUNG
SAMPLE TYPE	Sputum

Record No : 2

Screenshot4:search result of Cystic fibrosis

From GLMdb we found that there are total 39 Lung microbiome studies available for Cystic Fibrosis. Among those only 5 studies have sputum sample, V4 region amplified paired end raw sequence layouts. From these studies we have continued our 16s rRNA sequence analysis further for this study..

For Cystic Fibrosis phenotype, 4 datasets were selected for studies having publicly available 16S sequencing data. The selection criteria for dataset was to have V4 region paired end sequences. For lung microbiome studies in Cystic Fibrosis phenotype, we selected sputum samples .

Among those 4 datasets two of them are antibiotic cohort studies.

Due to lack of publicly available datasets for adult Healthy lung microbiome studies, we have collected healthy lungs samples from two asthmatic patients' lung microbiome studies with V4 paired end dataset in which one has sputum sample and other has (BAL/bronchial brushing) sample .

Details of datasets taken have been summarized in Table 4

Table4:Summary of datasets taken for analysis.

Author	PubMed ID	Bioproject ID	Sample	16S rRNA gene hypervariable region	Library layout	Sequencing Platform	Type of Study
Deborah A. Hogan et al[3]	27548479	PRJNA288589	Sputum	V4-V5	Paired-end	Illumina MiSeq	Cystic fibrosis patient
Flynn JM et al[14]	27548479	PRJNA305156	Sputum	V4	Paired-end	Illumina MiSeq	Cystic fibrosis patient
Durack et al[15]	27838347	PRJEB15534	Bronchial brushing	V4	Paired-end	Illumina MiSeq	Healthy lung
Durack et al.[16]	29885665	PRJEB22676	Induced Sputum	V4	Paired-end	Illumina MiSeq	Healthy lung
Andrea Hahn et al[1]	30238064	PRJNA437613	Sputum	V4	Paired-end	Illumina MiSeq	Antibiotic cohort study
Lenka Kramná et al[13]	29127619	PRJNA339813	sputum	V4	Paired-end	Illumina MiSeq	Antibiotic cohort study

In our first antibiotic study (Andrea Hahn et al)[1]we have got 16s rRNA sequence sample from 6 patients with several groups of antibiotic received for 4 conditions (i.e. Baseline, Exacerbation, Treatment and recovery). We have got 5 baseline (B), 6 exacerbation (E), 2 treatment (T), and 6 recovery (R) sample. This is summarized in table 5. In this study the baseline visit (B) occurred at a routine evaluation when patients were had not been on IV antibiotics for at least 30 days. The exacerbation visit (E) occurred when they were being hospitalized to begin treatment with IV antibiotics. The treatment visit (T) occurred at the end of their antibiotic treatment course. The recovery visit (R) occurred more than 30 days after completing their IV antibiotic therapy.

Our second antibiotic cohort study(Lenka Kramná et al)[13] we have got total 32 samples from 9 patients with total 16 antibiotic course. Three patients provided a single pair of samples covering one course of antibiotic therapy, five patients underwent two courses, and one patient had three courses of the therapy. Expectorated sputum was collected on the day when antibiotic therapy started referred as 'before' and on the last day of antibiotic treatment referred as 'after'.This is summarized in table 6.

Table5: Summary of datasets taken for analysis for antibiotic study1 (Andrea Hahn et al)[1]

Patient and Course	Antibiotic received	BERT Condition
A1	Piperacillin-tazobactam, imipenem-cilastatin, linezolid	E
A2	Meropenem, amikacin	E, T
A3	Meropenem,amikacin, ertapenem	E
A4	Meropenem,amikacin, vancomycin, linezolid	R
B	Ceftazidime, tobramycin	B, E, R
C1	Ceftazidime, tobramycin	B, R
C2	Meropenem, tobramycin	E, R
D	Meropenem, tobramycin	B, R
E	Ceftazidime, tobramycin	B, E, T
F	Meropenem, tobramycin	B, R

Table6: Summary of datasets taken for analysis for antibiotic study2 (Lenka Kramná et al)[13]

Patient	Course of antibiotic therapy	Antibiotic received
P1	course 1	piperacillin-tazobactam, amikacin, co-trimoxazole / withdrawn azithromycin p.o., tobramycin inh.
P2	course 1	piperacillin-tazobactam, ciprofloxacin
P2	course 2	meropenem, ciprofloxacin / withdrawn azithromycin p.o.
P3	course 1	piperacillin-tazobactam, tobramycin
P4	course 1	meropenem, tobramycin / withdrawn azithromycin p.o
P4	course 2	meropenem, colistin / withdrawn azithromycin p.o
P5	course 1	tobramycin, ceftazidime, co-trimoxazole
P5	course 2	colistin, cefoperazone-sulbactam / withdrawn tobramycin
P5	course 3	meropenem, tobramycin / withdrawn tobramycin inh.

P6	course 1	colistin, cefepime, co-trimoxazole / withdrawn tobramycin
P6	course 2	meropenem, tobramycin / withdrawn tobramycin inh.
P7	course 1	cefepime, amikacin / withdrawn azithromycin p.o.
P7	course 2	colistin, cefepime / azithromycin p.o.
P8	course 1	piperacillin-tazobactam, ofloxacin
P8	course 2	piperacillin-tazobactam, ofloxacin / withdrawn tobramycin inh
P9	course 1	meropenem, colistin / withdrawn azithromycin p.o.

Data processing:

At first all the datasets were downloaded from SRA using their Run accession IDs. Raw FASTQ files were run through FastQC and none of the sequences had adapter linked with it. So adapter trimming was not required and we further processed our data in Mothur (version 1.41.1) pipeline.

Some of our sequences contained certain elements in identifier line in the fastq files that Mothur was unable to read. So we modified our sequence files in the identifier line so that forward file identifier and reverse file should match

Some of the sequence files contained different number of lines in their forward and reverse read files. This was creating problem to make contigs during make.contigs command in Mothur. To solve this problem, we used list.seqs and get.seqs command in Mothur which created forward and reverse files with same number of lines.

```
mothur>list.seqs(fastq=A113.1.fastq)

Output File Names:

A113.1.accnos

mothur>get.seqs(accnos=A113.1.accnos, fastq=A113.2.fastq)

Selected 784263 sequences from your fastq file.

Output File Names:

A113.2.pick.fastq

mothur>list.seqs(fastq=A113.2.pick.fastq)

Output File Names:

A113.2.pick.accnos

mothur>get.seqs(accnos=A113.2.pick.accnos, fastq=A113.1.fastq)

Selected 784263 sequences from your fastq file.

Output File Names:

A113.1.pick.fastq
```

After preparing all our FASTQ files, we processed them in Mothur for each datasets. The workflow for Mothur has been followed from the standard protocol in https://www.mothur.org/wiki/MiSeq_SOP.[8]

Files needed for sequence processing:

- Silva reference file for alignment: Bacterial reference file for 16SrRNA gene was downloaded from https://www.mothur.org/wiki/Silva_reference_files and processed the file using **pcr.seqs** command to customise the region of our choice, that is V4 hypervariable region of 16S rRNA gene.

  pcr.seqs : A customised reference database has been made for V4 region using **pcr.seqs** from the downloaded SILVA reference file by using the following command .

```
mothur> pcr.seqs(fasta=silva.bacteria.fasta,
start=11894, end=25319, keepdots=F)

mothur>rename.file(input=silva.bacteria.pcr.fasta,
new=silva.v4.fasta)
```

rename.file(input=silva.bacteria.pcr.fasta, new=silva.v4.fasta)

- sequences. The mothur formatted version of RDP reference files can be downloaded from the following link:https://www.mothur.org/wiki/RDP_reference_files.

We moved all these files to mothur executable folder.

Creating files easy to handle:

make.file : A stability file was created to handle huge number of samples,shows the sample number and its corresponding forward and reverse reads (Paired-end data) by the command **make.file** in mothur.

Input: Fastq files

```
mothur>make.file(inputdir=Control_samples,type=fastq,,prefix=stabi
lity)
```

Inputdir= mothur executable folder

Output: .file

```
SRR2083518      SRR2083518_1.fastq      SRR2083518_2.fastq
SRR2083524      SRR2083524_1.fastq      SRR2083524_2.fastq
SRR2083531      SRR2083531_1.fastq      SRR2083531_2.fastq
SRR2083538      SRR2083538_1.fastq      SRR2083538_2.fastq
SRR2083551      SRR2083551_1.fastq      SRR2083551_2.fastq
SRR2083557      SRR2083557_1.fastq      SRR2083557_2.fastq
```

Reducing sequencing and PCR errors

make.contigs : **make.contigs** command was used for combining forward and reverse sets of reads for each sample and then to combine the data from all of the samples.This command extracted the sequence and quality score data from fastq files and created the reverse complement of the reverse read and then joined the reads into contigs.

Input: Output **.file** from **make.file** step

```
mothur>make.contigs(file=stability.files, processors=4)
```

Output:.**fasta** and **.qual** files for scraped and trimmed sequences. One **.groups** file and one **.report** file.

Stability.trim.contigs.fasta file have been used in downstream processes.

. groups file gave information about which read originated from which sample.

```
Group count:
SRR2083518      375518
SRR2083524      313285
SRR2083531      255486
SRR2083538      1326998
SRR2083551      92670
SRR2083557      101899

Total of all groups is 2465856

It took 2935 secs to process 2465856 sequences.

Output File Names:
frstprj/stability.trim.contigs.fasta
frstprj/stability.trim.contigs.qual
frstprj/stability.scrap.contigs.fasta
frstprj/stability.scrap.contigs.qual
frstprj/stability.contigs.report
frstprj/stability.contigs.groups
```

summary.seqs : To have an idea about the sequences, we checked the **.fasta** file by running in **summary.seqs** command.

Input: Output trimmed **.fasta** file from **make.contigs** step.

```
mothur>summary.seqs(fasta=stability.trim.contigs.fasta)
```

Output: **.summary** file

According to the output of the **summary.seqs** command, the parameters has been set to screen our sequences.

```
mothur > summary.seqs(inputdir=5thprj, fasta=stability.trim.contigs.fasta)
Setting input directory to: 5thprj/

Using 4 processors.

                Start   End     NBases  Ambigs  Polymer NumSeqs
Minimum:        1       247     247     0       3       1
2.5%-tile:      1       252     252     0       3       15072
25%-tile:       1       253     253     0       4       150713
Median:         1       253     253     0       6       301425
75%-tile:       1       253     253     0       6       452137
97.5%-tile:     1       253     253     7       6       587778
Maximum:        1       501     501     88      250     602849
Mean:    1      253     253     0       4
# of Seqs:      602849

It took 10 secs to summarize 602849 sequences.

Output File Names:
 5thprj/stability.trim.contigs.summary
```

Data cleaning:

screen.seqs :The quality of our sequences has been improved by using **screen.seqs** command with maxambig and maxlength parameter set according to our data.

Input: **.fasta** file created in **make.contigs** step and **.groups** file created in **make.contigs** step

```
mothur>screen.seqs(fasta=stability.trim.contigs.fasta,group=st
ability.contigs.groups, maxambig=0, maxlength=275)
```

Output: **good.groups** file, **bad.accnos** file and **good.fasta** file.

Reads that were removed during this step were listed in **bad.accnos** file.

unique.seqs : We anticipated that many of our sequences are duplicates of each other.So our sequences are unique by running **unique.seqs** command.

Input: **good.fasta** output file created in last **screen.seqs** step.

```
mothur>unique.seqs(fasta=stability.trim.contigs.good.fasta)
```

Output: **.names** file and **unique.fasta** file.

.names file consists of unique sequence names and their identical sequence.

unique.fasta file contains only unique sequences.

count.seqs : Then we combined our **.names** file and **.groups** file into a single count table by running **count.seqs** command.

Input: **good.groups** file from **screen.seqs** output, **.names** file from the **unique.seqs** output

```
mothur>count.seqs(name=stability.trim.contigs.good.names,
group=stability.contigs.good.groups)
```

Output: **.count_table**

Sequence alignment:

align.seqs :align.seqs command was used to align our sequences to reference.This command will use kmer searching with 8mers and will use the Needleman-Wunsch pairwise alignment method with a reward of +1 for a match and penalties of -1 and -2 for a mismatch and gap.

Input: **unique.fasta** output from **unique.seqs** step, customized reference file from SILVA

```
mothur>align.seqs(fasta=stability.trim.contigs.good.unique.fast
a, reference=silva.v4.fasta)
```

Output: **.align** file, **.report** file and **.accnos** file.

.accnos file gave a list of discarded sequence after alignment.

.report file gave all information related to the alignment.

summary.seqs : The positions of alignment has been checked after alignment by summary.seqs then.

Input: **.align**fastafile from **align.seqs** step, **.count_table** output from **count.seqs** step.

```
mothur>summary.seqs(fasta=stability.trim.contigs.good.unique.alig
n, count=stability.trim.contigs.good.count_table)
```

Output: **.summary** file

More data cleaning:

screen.seqs : After aligning the sequences and running **summary.seqs**, we checked the alignment positions from the last **.summary** output and re-run **screen.seqs** with start, end and maxhomopolymer parameters given.

Input: **.align**fasta file from **align.seqs** step, **.count_table** from **count.seqs** step, **.summary** file from last **summary.seqs** step.

```
mothur>screen.seqs(fasta=stability.trim.contigs.good.unique.ali
gn, count=stability.trim.contigs.good.count_table,
summary=stability.trim.contigs.good.unique.summary, start=1968,
end=11550,maxhomop=8)
```

Output: **good.summary**, **good.align**, **bad.accnos**, **good.count_table**

filter.seqs : To make sure that our sequences overlap only the region specified in the screen.seqs, filter.seqs command is used to remove any overhangs on either end of that region using 'trump' parameter. **filter.seqs** tool additionally cleaned up our alignment file by removing any columns that have a gap character in that position for every aligned sequence by giving 'vertical' parameter.

Input: good.alignfasta file from last screen.seqs

```
mothur>filter.seqs(fasta=stability.trim.contigs.good.unique.go
od.align, vertical=T, trump=.)
```

Output: filter.fasta file, .filter file

Filter.fasta file contained the aligned filtered sequences.

```
Length of filtered alignment: 543
Number of columns removed: 12881
Length of the original alignment: 13424
Number of sequences used to construct filter: 1374199

Output File Names:
frstprj/stability.filter
frstprj/stability.trim.contigs.good.unique.good.filter.fasta
```

unique.seqs : Any filtering step that was performed might lead to sequences that are redundant, so **unique.seqs** has been reran again to deduplicate our data.

Input: **filter.fasta** output from **filter.seqs** step, **.count_table** output from last **screen.seqs** step.

```
mothur>unique.seqs(fasta=stability.trim.contigs.good.unique.good.fi
lter.fasta, count=stability.trim.contigs.good.good.count_table)
```

Output: **.count**_table, **unique.fasta**

PreClustering:

pre.cluster : we have then pre-clustered our sequences to merge our nearly identical sequences together,. This is because, sequences that only differ by around 1 in every 100 bases are likely to represent sequencing errors and not true biological variation. So we pre-clustered our sequences given threshold to 2 mismatches. This command splits the

sequences by group and then sorts them by abundance and goes from most abundant to least. Then it identifies sequences that are within 2 nucleotide variation from each other. If it finds such sequences, it merges them.

Input: **unique.fasta** and **.count_table** from last **unique.seqs** step

```
mothur>pre.cluster(fasta=stability.trim.contigs.good.unique.good.fi
lter.unique.fasta,
count=stability.trim.contigs.good.unique.good.filter.count_table,
diffs=2)
```

Output: **precluster.fasta**, **precluster.count_table**, **precluster.map**.

Chimera Removal

chimera.vsearch : It is possible that two unrelated templates are combined to form a sort of hybrid sequence ,formed a chimera during PCR amplification,. To remove these sequencing artefacts **chimera.vsearch** tool was used.It uses VSEARCH algorithm.VSEARCH stands for vectorized search, VSEARCH uses an optimal global aligner by Needleman-Wunsch algorithm, in contrast to USEARCH which by default uses a heuristic seed and extend aligner. This usually results in more accurate alignments and overall improved sensitivity (recall) with VSEARCH, especially for alignments with gaps.

This command split the data by sample and check for chimeras. In addition, if a sequence is flagged as chimeric in one sample, the default (dereplicate=f) removes it from all samples. Because of this, rare sequences get flagged as chimeric when they are the most abundant sequence in another sample. To solve this, the parameter dereplicate=t is given as input.

Input: **precluster.fasta** and **precluster.count_table** from **pre.cluster** step.

```
mothur>chimera.vsearch(fasta=stability.trim.contigs.good.unique.goo
d.filter.unique.precluster.fasta,count=stability.trim.contigs.good.
unique.good.filter.unique.precluster.count_table, dereplicate=t)
```

Output: **vsearch.chimeras**, **vsearch.accnos** and **vsearch.pick.count_table**

remove.seqs : chimera.vsearch removed the chimeric sequences from the count_table. In order to remove the chimeric sequences from our fasta files also,**remove.seqs** command was used then.

Input: **precluster.fasta** from **pre.cluster** step, **vsearch.accnos** from **chimera.vsearch** step

```
mothur>remove.seqs(fasta=stability.trim.contigs.good.unique.good.f
ilter.unique.precluster.fasta,accnos=stability.trim.contigs.good.
unique.good.filter.unique.precluster.denovo.vsearch.accnos)
```

Output: **pick.fasta**

Removal of non-bacterial sequences

classify.seqs : Next we have to classify our sequences to remove 16S rRNA gene fragments from other sources than of bacteria,. We used **classify.seqs** tools which classified our sequences with reference of mothur formatted version of RDP classifier files. Wang method was used as default of **classify.seqs** to classify all our sequences. We provided a cut-off value of 80 which corresponds to the 80% confidence to the bootstrap values.

Input: **pick.fasta** from **remove.seqs**, **pick.count_table** from **chimera.vsearch** output, reference and taxonomy from mothur formatted version of RDP trainset.

```
mothur>classify.seqs(fasta=stability.trim.contigs.good.unique.good.fi
lter.unique.precluster.pick.fasta,count=stability.trim.contigs.good.
unique.good.filter.unique.precluster.denovo.vsearch.pick.count_table
,reference=trainset9_032012.pds.fasta,taxonomy=trainset9_032012.pds.
tax, cutoff=80)
```

```
Output File Names:
frstprj/stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pds.wang.taxonomy
frstprj/stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pds.wang.tax.summary
frstprj/stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pds.wang.flip.accnos
```

remove.lineage : ** Next undesired non-bacterial sequences was removed using **remove.lineage command.

Input: **pick.fasta** from **remove.seqs**, **pick.count_table** from **chimera.vsearch** output, **wang.taxonomy** from **classify.seqs** step

```
mothur>remove.lineage(fasta=stability.trim.contigs.good.unique.g
ood.filter.unique.precluster.pick.fasta,count=stability.trim.con
tigs.good.unique.good.filter.unique.precluster.denovo.vsearch.pi
ck.count_table,taxonomy=stability.trim.contigs.good.unique.good.
filter.unique.precluster.pick.pds.wang.taxonomy,
taxon=Chloroplast-Mitochondria-unknown-Archaea-Eukaryota)
```

Output: **.taxonomy**, **.fasta** and **.count_table**

```
Output File Names:
frstprj/stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pds.wang.pick.taxonomy
frstprj/stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.fasta
frstprj/stability.trim.contigs.good.unique.good.filter.unique.precluster.denovo.vsearch.pick.pick.count_table
```

Clustering sequences into OTU:

cluster.split : The sequences are then clusterd into OTUs (Operational Taxonomic Unit) using **cluster.split** command. This command uses the taxonomic information to split the sequences into bins and then cluster within each bin. **cluster.split** by default uses opticlust algorithm to cluster sequences. We specified the split method and taxonomy level to split by and the cut-off level to 0.03 as we wanted a 97% similarity.

Input: **.fasta**, **.taxonomy** and **.count_table** from **remove.lineage** step

```
mothur>cluster.split(fasta=stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.fasta,count=stability.trim.contigs.good.unique.good.filter.unique.precluster.denovo.vsearch.pick.pick.count_table,taxonomy=stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pds.wang.pick.taxonomy, splitmethod=classify, taxlevel=4, cutoff=0.03)
```

Output: **.dist**, **.list**, **.sensspec**

make.shared :Now we have checked how many sequences are in each OTU from each group, by using **make.shared** command.

Input: **.list** file from **cluster.split**, **.count_table** from **remove.lineage**

```
mothur>make.shared(list=stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.opti_mcc.list,count=stability.trim.contigs.good.unique.good.filter.unique.precluster.denovo.vsearch.pick.pick.count_table, label=0.03)
```

Output: **.shared** file

classify.otu : To classify the consensus taxonomy for each OTU,**classify.otu** command has been used.

Input: **.list** file from **cluster.split**, **.count_table** and **.taxonomy** from **remove.lineage**

```
mothur>classify.otu(list=stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.opti_mcc.list,count=stability.trim.contigs.good.unique.good.filter.unique.precluster.denovo.vsearch.pick.pick.count_table,taxonomy=stability.trim.contigs.good.unique.good.filter.unique.precluster.pick.pds.wang.pick.taxonomy, label=0.03)
```

Output: **cons.taxonomy**, **cons.tax.summary**

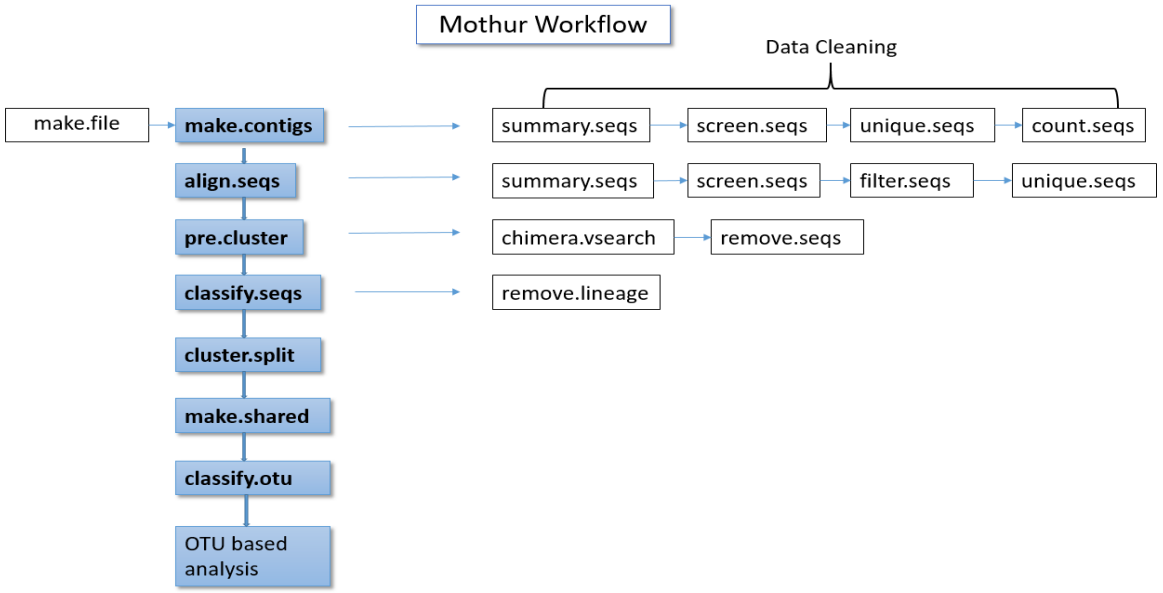


Fig5:Flowchart of mothur workflow

OTU Based Analysis :

The classify.otu output file **cons.taxonomy**and **cons.tax.summary** gave the OTU information and its corresponding taxonomic information. For each study, we normalized the OTU counts by taking the average from all samples in that study. Samples showed a tendency of clustering by studies due to different sampling size and sequencing depth in different studies. So we normalized the OTU counts between studies by taking relative proportion of OTU counts in percentage for each study (from now on we will call this relative proportion of OTU counts in percentage as 'OTU score'). We picked OTU score from each study keeping the threshold of minimum 1 OTU. We also picked OTU score from each study with threshold of 0.1 relative proportion in percent. We constructed two separate long format matrix with the two different threshold including all the studies. Then we converted the long format matrix to wide format matrix in R-Studio using 'reshape' package. The R code has been given below.

```
install.packages("reshape")

library(reshape)

library(readxl)

setwd("C:/Users/suniket/Desktop")

Long_format<-data.frame(read_excel("C:/Users/Srijan/Desktop/long
bp_CF_H.xlsx"))

View(Long_format)

wide_format<-reshape(Long_format, idvar = "taxon", timevar =
"bioproject", direction = "wide")

write.csv(wide_format, "wide_CF_H.csv")
```

write.csv(wide_format, "wide_CF_H.csv")

In the wide format matrix, we got the OTU score for each taxa in each bioproject/study in a concise table. From this wide format matrix table data, we performed Pearson correlation between Cystic Fibrosis patients and control healthy patients, and log2 Fold Change (FC) in Microsoft Excel. We also made Venn diagram using Venny 2.1.0 to find out exclusive microbes present in Cystic Fibrosis lung and Healthy control lung.

For antibiotic studies, we have also generated wide format matrix from long format matrix with the relative scores of 10 known pathogen *(Alcaligenaceae family, Haemophilus, Streptococcus, Granulicatella, Stenotrophomonas, Prevotella, Rothia, Veillonella, Pseudomonadaceae family, Staphylococcus)*of Cystic fibrosis along with patient condition(Baseline,Exacerbation,Treatmment,Recovery) for first antibiotic study(Andrea Hahn et al). For Second antibiotic study (Lenka Kramná et al) we have 32 smples for 16 antibiotic course of 9 Cystic fibrosis patients.Each antibiotic course have two samples named “after “ and “before”.We have constructed heatmap for both studies by higherarchial clustering using R studio by using hclust and heatmap.2 function.We have also create a separate heatmap of those 10 pathogen from previous CF patient and healthy studies.The R code has given below:

```
install.packages("gplots")

install.packages("heatmap.plus")

install.packages("RColorBrewer")

library("gplots")

library("heatmap.plus")

library("RColorBrewer")

test <-
read.csv("C:/Users/Srijan/Desktop/wide_4th_Pseudo_change2.csv",row.
names = 1)

input <- as.matrix(t(test))heatmap.2(input, Colv="FALSE",
Rowv="FALSE", trace="none", density="none", col=bluered(20),
cexRow=0.7, cexCol=1, margins = c(4,15),scale="row",hclustfun =
hclust)
```

below:

Results

From the wide format matrix of microbe with OTU score, we observed some general trends in between Cystic fibrosis patients' lung and healthy lung. *Firmicutes* are highly colonized in both cases. Comparative to the healthy lung, Cystic fibrosis patients' lungs harbours more *Firmicutes* in overall community structure. Significantly large number of microbes under the class bacilli has been seen in CF lung than healthy lung. Microbes of *Pseudomonadaceae* and *Streptococcaceae* family were also found in large numbers in Cystic fibrosis patients' lung than healthy lung.Members of *Gammaproteobacteria* are more in Cystic fibrosis patients' lung. *Fusobacteria, Clostridia, Negativicutes, Neisseria, Leptotrichia* have relatively higher colonization in healthy lung microbiome rather than Cystic fibrosis patients' lung microbiome.

In genus level *Rothia, Gemella, Neisseriaceae_unclassified, Alcaligenaceae_unclassified, Achromobacter, Stenotrophomonas* have relatively higher abundance in Cystic fibrosis patients' lung microbiome.Suprisingly we have observed that genus *Prevotella has* relatively higher abundance in healthy lung microbiome.

Log2 foldchange:

We have calculated log2 foldchange in between Cystic fibrosis patients' lung and healthy lung microbiome.We have tried to found out some specific taxa associated with the disease Cystic fibrosis.We have observed that *Pseudomonadaceae, Streptococcaceae, Alcaligenacea* have large increased abundance in Cystic fibrosis lung. While In genus level *Rothia, Gemella*,

Neisseriaceae_unclassified, Alcaligenaceae_unclassified, Achromobacter, Stenotrophomonas have slightly increased abundance in Cystic fibrosis lung microbiome compared to healthy lung muicrobiome. Comparatively decresed abundance of Fusobacteria, Clostridia, Negativicutes, Neisseria, Leptotrichia , *Prevotella* has been observed. In healthy lung microbiome we got slightly less abundance streptococcus than Cystic fibrosis lung microbiome.

Table7: Important bacterial genus found in altered microbiome

Taxon	Change in abundance
Pseudomonadaceae *Streptococcaceae* *Alcaligenacea* *Rothia* *Gemella* *Neisseriaceae_unclassified* *Stenotrophomonas* Achromobacter	**Increased abundance in Cystic fibrosis patients** 
Fusobacteria *Negativicutes* *Neisseria* *Clostridia* *Negativicutes*	**Decreased abundance in Cystic fibrosis patients.** 

Correlation:

To find out correlation between Cystic fibrosis lung microbiome and normal heathy lung microbiome, we calculated the Pearson correlation coefficient between these groups by using **PEARSON** function in Microsoft Excel. For easier calculation, we took the average for each group and then calculated the correlation coefficient. We found positive correlation of **0.78** in

between two groups. Some taxa specific changes may be present which could not be assessed using correlation.

Venn diagram:

To observed what OTUs found solely and what are OTUs are common in between two groups of Cystic fibrosis lung microbiota and healthy lung microbiome we have created to venn diagrams.

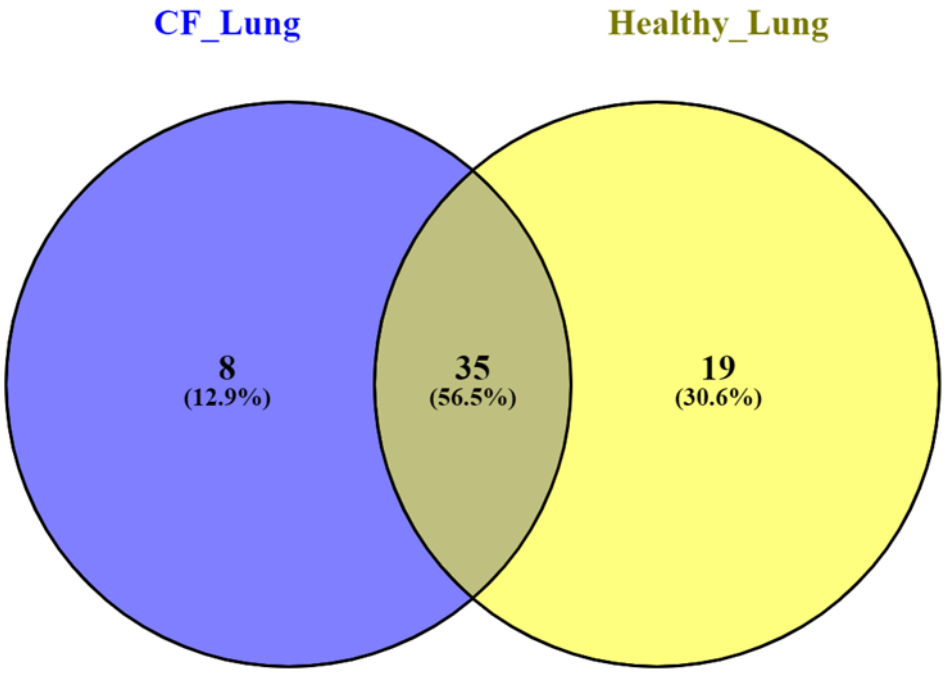


Fig6: Venn diagram 1(Created using minimum threshold of 1 OTU)

At minimum threshold of 1 OTU 8 taxon included exclusively in Cystic fibrosis patients' lung(CF_LUNG):Pseudomonadales,*Pseudomonadaceae,Pseudomonadaceae_unclassified,Staphylococcaceae,Bacillales_unclassified,Staphylococcaceae_unclassified,Staphylococcus,Burkholderiales*. 19 taxon included exclusively in Healthylung microbiome.These are *Pasteurellales,Pasteurellaceae,Pasteurellaceae_unclassified,Leptotrichiaceae,Neisseriales,Neisseriaceae,Actinomycetales_unclassified,Leptotrichia,Lachnospiraceae_unclassified,Prevotellaceae_unclassified,Veillonellaceae_unclassified,Bacteria_unclassified,Neisseria,Bacteroidales_unclassified,Porphyromonadaceae,Bacillales_Incertae_Sedis_XI,Gemella,Neisseriaceae_unclassified,Oribacterium.*35 common taxon in Cystic fibrosis and Healthylung microbiome.*Firmicutes,Bacilli,Lactobacillales,Proteobacteria,Gammaproteobacteria,Bacilla*

les,Streptococcaceae,Streptococcus,Bacteroidetes,Bacteroidia,Bacteroidales,Bacilli_unclassified,Prevotellaceae,Prevotella,Actinobacteria,Lactobacillales_unclassified,Actinomycetales,Veillonellaceae,Negativicutes,Selenomonadales,Veillonella,Clostridia,Micrococcaceae,Clostridiales,Rothia,Fusobacteria,FusobacterialesBetaproteobacteria,Lachnospiraceae,Fusobacteriaceae,Actinomycetaceae,Fusobacterium,Actinomyces,Firmicutes_unclassified,Carnobacteriaceae are those taxons.

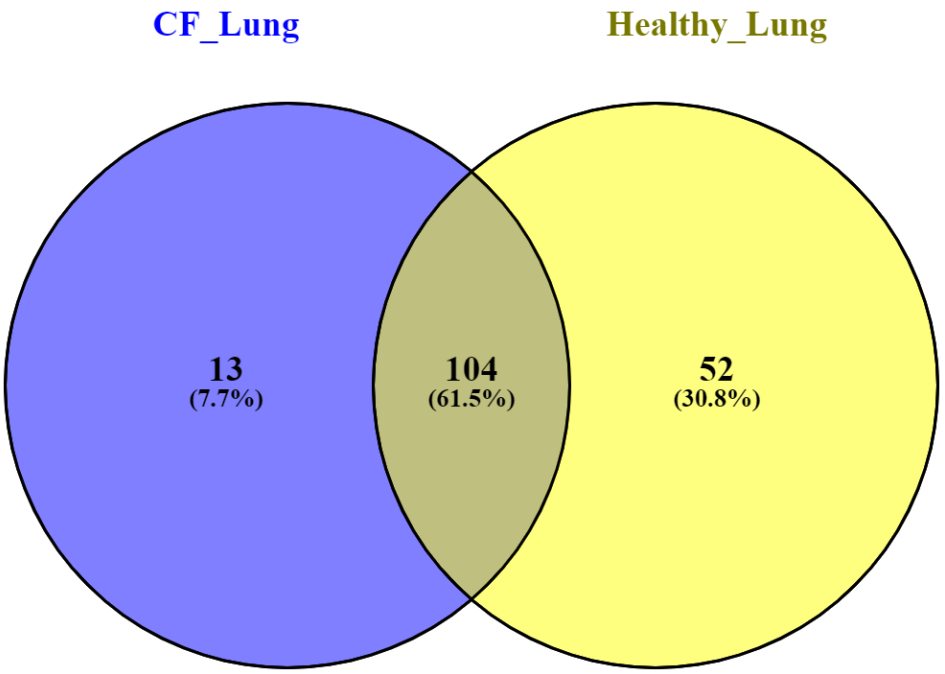


Fig7: Venn diagram 1(Created using minimum threshold of 1 OTU)

At minimum threshold of 0.1 OTU 13 taxons included exclusively in Cystic fibrosis patients'lung.*Staphylococcaceae,Staphylococcaceae_unclassified,Staphylococcus,Alcaligenaceae,Alcaligenaceae_unclassified,Gammaproteobacteria_unclassified,Achromobacter,Xanthomonadaceae,Streptobacillus,Deltaproteobacteria,Verrucomicrobia,Bifidobacterium,Stenotrophomonas* are these. 52 taxon included exclusively in Healthy_Lung.These are *Neisseriaceae_unclassified,Parvimonas,Catonella,Erysipelotrichia,Erysipelotrichales,Erysipelotrichaceae,Clostridiales_Incertae_Sedis_XIII,Mogibacterium,SR1,SR1_class_incertae_sedis,SR1_order_incertae_sedis,SR1_family_incertae_sedis,SR1_genus_incertae_sedis,Eubacteriaceae,Eubacterium,Fusobacteriaceae_unclassified,Selenomonas,Escherichia_Shigella,Solobacterium,Bacteroidaceae,Bacteroides,Corynebacteriaceae,Corynebacterium,Enterobacteri*

*aceae_unclassified,Sphingomonadales,Micrococcaceae_unclassified,Sphingomonadaceae,Ruminococcaceae,Moraxellaceae,Acidobacteria,Comamonadaceae,Haemophilus,Actinomycetaceae_unclassified,Alphaproteobacteria_unclassified,Centipeda,Sphingobacteria,Sphingobacteriales,Tenericutes,Mollicutes,Mycoplasmatales,Mycoplasmataceae,Mycoplasma,Clostridia_unclassified,Tannerella,Ruminococcaceae_unclassified,Pseudomonas,Aerococcaceae,Filifactor,Rhodospirillales,Deinococcus-Thermus,Deinococci,Hallella.*We found there are total 104 taxon are common in between Cystic fibrosis lung microbiome and healthy lung microbiome.

Hierarchical clustering and heatmaps:

A heatmap is created based on hierarchial clustering from the wide format matrix which generated based on relative abundance scores of 10 wellknown OTU associated with Cystic fibrosis (*Alcaligenaceae family, Haemophilus, Streptococcus, Granulicatella, Stenotrophomonas, Prevotella, Rothia, Veillonella, Pseudomonadaceae family, Staphylococcus*) In between studies of Cystic fibrosis lung microbiome and healthy lung microbiota.

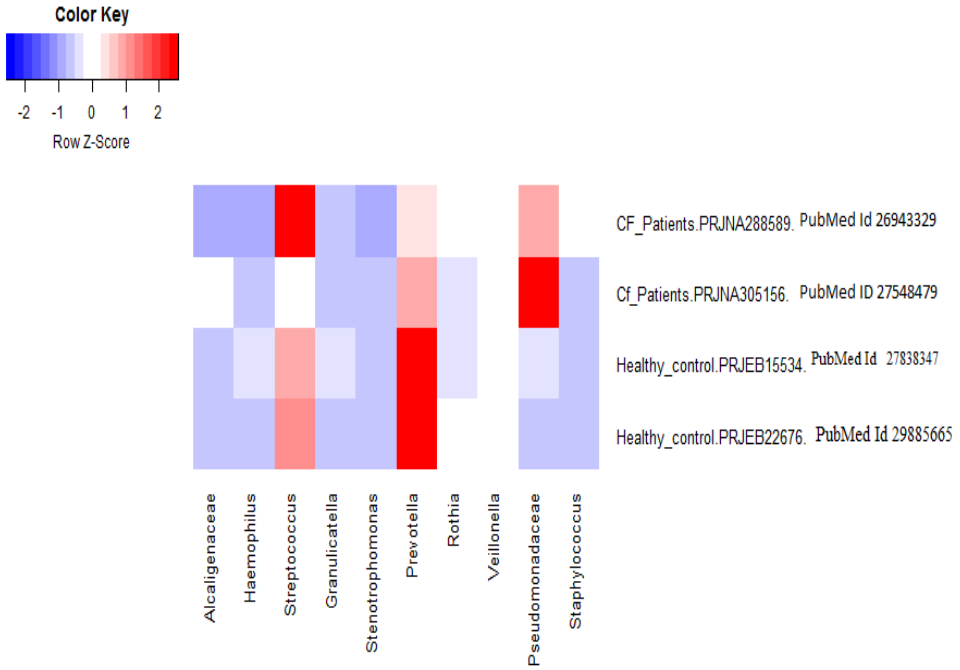


Fig8:heatmap generated based on 10 known pathogen of Cystic fibrosis from studies of Cystic fibrosis lung microbiome and healthy lung microbiome.

We have observed there is an increased abundance of *Streptococcu*s and the members of *Pseudomonadaceae* family in Cystic fibrosis patients' lung microbiome studies comparative to healthy lung microbiota.Suprisingly members of the *Prevotella* have large population in healthy microbiome studies.

In between two Cystic fibrosis study *Alcaligenaceae family, Haemophilus, Stenotrophomonas* have more relative abundance in Deborah A. Hogan et al (PubMed Id 26943329) than other. We found slightly lower abundance of those taxon in healthy lung microbiome studies.

Antibiotic Cohort studies:

We have created an another heatmap from based on hierarchical clustering from the wide format matrix, generated based on relative abundance scores of those taxon in our first antibiotic study (**Andrea Hahn et al, PubMed Id 30238064**)[1] to see how those taxon altered in different Cystic fibrosis patients' condition.

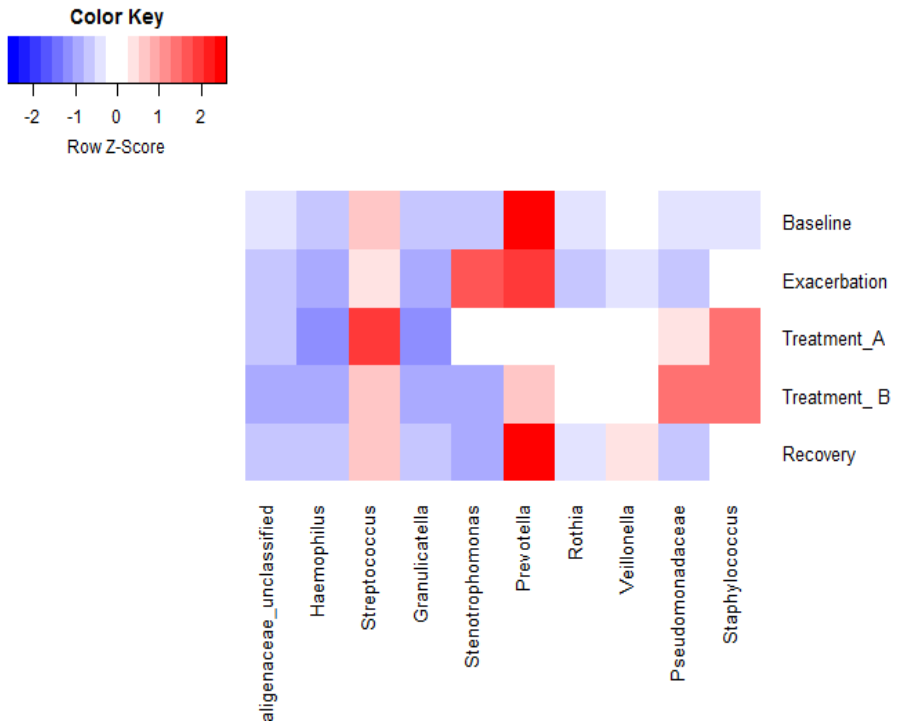


Fig9:heatmap generated based on 10 known pathogen of Cystic fibrosis from antibiotic study 1(Andrea Hahn et al, PubMed Id 30238064)[1]

As our aim was to observe how those known pathogen change in different patients' condition at antibiotic course. Although different patient have given different antibiotics we have seen a correlation. We observed that *Prevotella* has more abundance in baseline and exacerbation than both treatment condition sample. *Stenotrophomonas* also have more relative abundance in exacerbation condition. At recovery stage it again have higher abundance. *Staphylococcus* has relatively higher abundance in treatment_A and treatment_B where as for *streptococcus and Pseudomonadacea* it only have relative higher abundance in Treatment_B.

For our second antibiotic study (**Lenka Kramná et al, PubMed Id 29127619)** [13]we have created another heatmap like this for all 32 samples with those taxa to observe how the taxon are reduced in abundance in different antibiotic course of 9 patients.

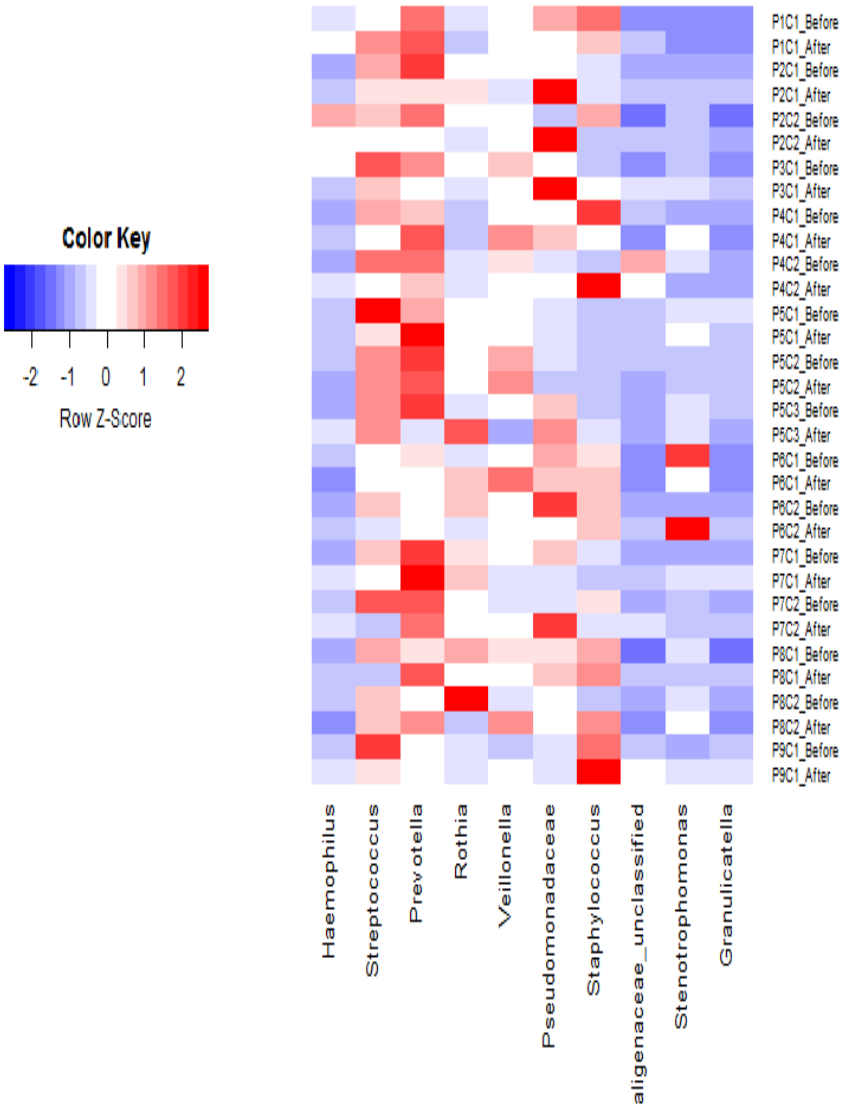


Fig10:heatmap generated based on 10 known pathogen of Cystic fibrosis from antibiotic study 2(Lenka Kramná et al, PubMed Id 29127619)[13]

Here We observed that abundance of *Prevotella is* significatantly decreased in courses in P5C3,P4C2,P2C2,P2C1 (after samples) where Meropenem antibiotic is given.In P6C2 after sample we can see *Pseudomonadacea* family has been higly reduced in abundance than other sample.In piperacillin-tazobactum,oflloxacin antibiotic course *rothia* got decreased in P8C2 after sample. Abundance of streptococcus is highly reduced in large type of antibiotic course in P5C1,P9C1,P3C1.In P4C1 *staphylococcus* abundance is highly reduced in meropenem antibiotic.Abundance of *haemophilus* is also got reduced in abundance in presence of

Meropenem antibiotic in P2C2,P2C1,P4C1,P4C2 after samples. Relative abundance of *Granulicatella* also reduced in *P*2C2,P8C1 in piperacillin-tazobactum.Relative abundance of *Alcaligenaceae_unclassified* was reduced in P2C2 and P3C1 after sample where meropenem-ciprofloxacin and piperacillin-tazobactum antibiotic is given.This is summarized in table 8

Table8: Decreased abundance of 10 known pathogen of Cystic fibrosis in different antibiotic courses from antibiotic study 2(Lenka Kramná et al, PubMed Id 29127619)[13]

Taxon decreased	Antibiotioc courses	Antibiotics
Prevotella	**P5C3**	**meropenem, tobramycin / withdrawn tobramycin inh.**
	P4C2	**meropenem, colistin / withdrawn azithromycin p.o**
	P2C1	**piperacillin-tazobactam, ciprofloxacin**
	P2C2	**meropenem, ciprofloxacin / withdrawn azithromycin p.o**
Pseudomonadacea	**P6C2**	**meropenem, tobramycin / withdrawn tobramycin inh**
Rothia	**P8C2**	**piperacillin-tazobactam, ofloxacin / withdrawn tobramycin inh**
Streptococcus	**P3C1**	**piperacillin-tazobactam, tobramycin**
	P5C1	**tobramycin, ceftazidime, co-trimoxazole**
	P9C1	**meropenem, colistin / withdrawn azithromycin p.o.**

Staphylococcus	**P4C1**	**meropenem, tobramycin / withdrawn azithromycin p.o**
Haemophilus	**P2C2** **P2C1** **P4C1** **P4C2**	**meropenem, ciprofloxacin / withdrawn azithromycin p.o.** **meropenem, colistin / withdrawn azithromycin p.o** **meropenem, tobramycin / withdrawn azithromycin p.o** **meropenem, colistin / withdrawn azithromycin p.o**
Granulicatella	**P2C2** **P8C1**	**meropenem, ciprofloxacin / withdrawn azithromycin p.o** **piperacillin-tazobactam, ofloxacin**
Alcaligenaceae_unclassified	**P2C2** **P3C1**	**meropenem, ciprofloxacin / withdrawn azithromycin p.o** **piperacillin-tazobactam, tobramycin**
Stenotrophomonas	**P6C1**	**colistin, cefepime, co-trimoxazole / withdrawn tobramycin**
Veillonella	**P3C1**	**piperacillin-tazobactam, tobramycin**

Discussion

Despite the fact that the bacterial composition of the lung and gut microbiota is particular to each individual depending on their food habits, age and other factors, we tried to find trends throughout all studies of lung microbiome during Cystic fibrosis patients' condition in adults. As sample size and depth of sequencing varies across studies and they tend to cluster by each study, we normalized them by taking the relative proportion of taxa from each study rather than taking absolute values. Also there are some studies like Flynn JM et al would seem presenting less taxa than other studies.Our antibiotic studies also have very less samle size rather than Cystic fibrosis patients' or healthy lung microbiome. Stille there were no significant difference in relative proportion of the taxa in Cystic fibrosis patients' studies.

Prevotella strains are classically considered commensal bacteria due to their extensive presence in the healthy human body and their rare involvement in infections.We have seen that although being an known pathogen of Cystic fibrosis *Prevotella* has increased abundance in healthy lung microbiome.This may be due to only a few strains have been reported to give rise to opportunistic endogenous infections, including chronic infections, abscesses and anaerobic pneumonia.There may be other nonharmful strains of Prevotella colonizing in healthy lung microbiome. However, emerging studies have showed that increased Prevotella abundance and specific strains to Cystic fibrosis, suggesting that at least some strains exhibit pathobiontic properties.

Several studies has reported and also in our meta-analysis aproaches we have seen that in most of the cases Cystic fibrosis lung microbiota are dominated by the microbes of the genera like *Streptococcus, pseudomonas, staphylococcus,gamella* etc which are facultatitave anaerobic.This may be due to the buid up thick sticky mucus in Cystic fibrosis patients' airways which leads to almost anaerobic condition in patients' lung.

We have also observed that some strictly aerobic genus like *rothia,stenotrophomonas*,members of *Alcaligenaceae* family also have increased abundance

in Cystic fibrosis patients' lung.These are not conventional pathogen of Cystic fibrosis. .It may be possible that improvements in airway clearance and more effective treatment of the conventional Cystic fibrosis pathogens has led to the emergence of new airway pathogens.

In antibiotic cohort study 1 we have noticed that Staphylococcus has relatively higher abundance in treatment_A and treatment_B where as for *streptococcus* and *Pseudomonadacea* it only have relative higher abundance in Treatment_B.In our approach we have compared the dominant taxons in different antibiotic condition but different patients have given different antibiotics,and we do not have all patients' all antibiotic courses Baseline,Exacerbation,Treatments and Recovery sample.So we could not conclude why those taxon have increased in relative abundance despite of antibiotics given during treatment.It may be possible that in Baseline and Exacerbation those 2 taxons present more than treatment.

From both antibiotic studies we strongly recommend that *Prevotella* species are more susceptible to meropenem antibiotic.*Haemophilus* species are also susceptible to meropenem antibiotic. Studies have been reported that *Alcaligenes* (also called *Achromobacter xylosoxidans*) is intrinsically resistant to aminoglycosides, cephalosporins (except ceftazidime) and aztreonam.It may be the possible cause of observation *alcaligenaceae* have been reduced in courses like P2C2 and P3C1,where betalactum or carbapenem antibiotic is given. Abundance of *Granulicatella* also reduced betalactum and betalactamase inhibitor antibiotics like piperacillin-tazobactam.

Conclusion

In conclusion, our study exemplifies the heterogeneity of changes in the composition of the lung microbiome between Cystic fibrosis patients and Healthy lungs. We have also tried to find out the antibiotic susceptibility of most common conventional and emerging pathogen of this autosomal recessive disease. However, if we could perform species level identification our study would be more accurate. Only then we could distinguish between which strains exhibit pathogenic properties and which strains are not harmful in a specific genus. If we

have more datasets of V4 16srRNA hypervariable region sputum sample from Cystic fibrosis and healthy lung microbiome our study will be more statistically advanced. Among the limitations in antibiotic cohort study, detecting DNA from dead bacterial cells may lead to underestimation of a taxa. Here we believe that most of the abundance come from live organisms. To estimate the sample variability, we would strongly recommend to analyse several replicate for each study.

Future scope

In future a species level identification can be done through the help of taxonomy reference file of **Ezbiocloud** database in this study.[23]. Only then we could distinguish in between which strains exhibit pathogenic properties and which strains are not harmful for a specific genus.When this is totally understood we can design Ecobiotic drugs to re-established the microbial diversity and health state functions.We could even create genetically engineered phage viruses to treat the multi-drug resistant bacterias.

There are total 6 Gut microbiome studies available for Cystic Fibrosis in Glmdb.Only one studies have Raw sequence data availability mentioned in the published article which is single end layout, V4-V6 region amplified sequence.If we got more publicly available gut microbiome datasets for Cystic fibrosis patients' in future,the comparison of lung and gut microbiome can be done to observe the vital crosstalk between the gut and lung microbial communities in Cystic fibrosis patients.

Lung microbiome transplantations can be done to treat the exacerbation condition of a Cystic fibrosis patients.

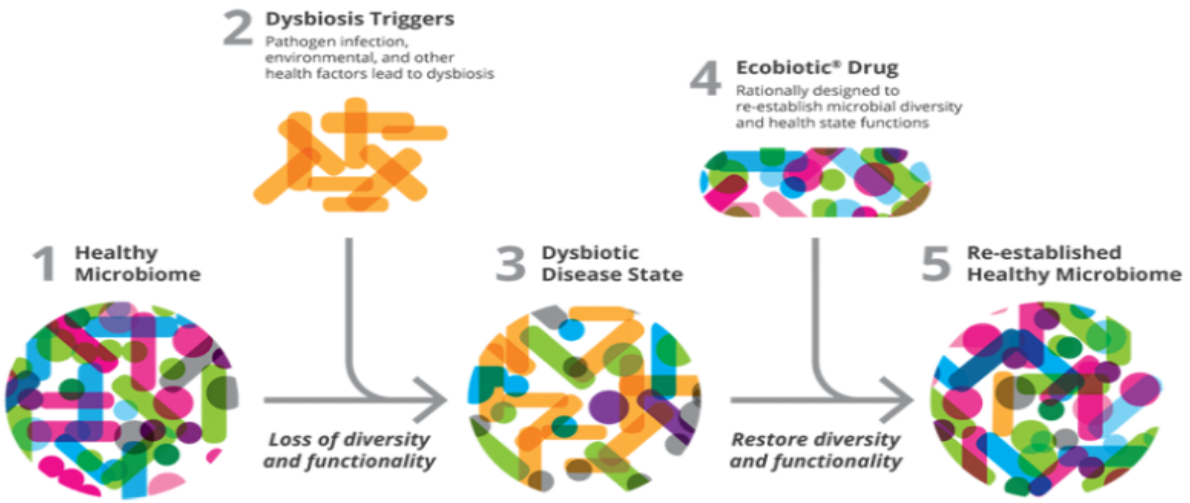


Fig11: Flowchart of steps to re-establish healthy microbiome diversity in Cystic fibrosis patients.

References

1. Andrea Hahn,Amit Sanyal,Geovanny F. Perez,Anamaris M,Colberg-Poley, Joseph Campos,Mary C. Rose,and Marcos Pérez-Losada Different next generation sequencing platforms produce different microbial profiles and diversity in cystic fibrosis sputum.
2. Li Li,Shawn Somerset The clinical significance of the gut microbiota in cystic fibrosis and the potential for dietary therapies.
3. Hogan DA, Willger SD, Dolben EL, Hampton TH, Stanton BA, Morrison HG, Sogin ML, Czum , Ashare A Analysis of Lung Microbiota in Bronchoalveolar Lavage, Protected Brush and Sputum Samples from Subjects with Mild-To-Moderate Cystic Fibrosis Lung Disease
4. Carmody LA1, Zhao J1, Kalikin LM1, LeBar W2, Simon RH3, Venkataraman A4, Schmidt TM4, Abdo Z5, Schloss PD4, LiPuma JJ1. The daily dynamics of cystic fibrosis airway microbiota during clinical stability and at exacerbation
5. https://galaxyproject.github.io/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html
6. https://en.wikipedia.org/wiki/Cystic_fibrosis
7. https://www.mayoclinic.org/diseases-conditions/cystic-fibrosis/symptoms-causes/syc-20353700
8. https://www.mothur.org/wiki/MiSeq_SOP
9. Ramsey BW. Management of pulmonary disease in patients with cystic fibrosis. N Engl J Med. 1996; 335(3):179–188. [PubMed: 8657217]

10. Carmody LA, Zhao J, Schloss PD, Petrosino JF, Murray S, Young VB, et al. Changes in cystic fibrosis airway microbiota at pulmonary exacerbation. Ann Am Thorac Soc. 2013 Jun; 10(3):179–187. [PubMed: 23802813]
11. O'Sullivan BP,Steven Freedman Cystic fibrosis
12. Gwen Duytschaever, Geert Huys, Maarten Bekaert, Linda Boulanger, Kris De Boeck, Peter Vandamme Cross-Sectional and Longitudinal Comparisons of the Predominant Fecal Microbiota Compositions of a Group of Pediatric Patients with Cystic Fibrosis and Their Healthy Siblings.
13. Kramná, Lenka, et al. "Changes in the lung bacteriome in relation to antipseudomonal therapy in children with cystic fibrosis." Folia microbiologica 63.2 (2018): 237-248.
14. Flynn, Jeffrey M., et al. "Evidence and role for bacterial mucin degradation in cystic fibrosis airway disease." *PLoS pathogens* 12.8 (2016): e1005846.
15. Durack, Juliana, et al. "Features of the bronchial bacterial microbiome associated with atopy, asthma, and responsiveness to inhaled corticosteroid treatment." *Journal of Allergy and Clinical Immunology* 140.1 (2017): 63-75.
16. Durack, Juliana, Yvonne J. Huang, Snehal Nariya, Laura S. Christian, K. Mark Ansel, Avraham Beigelman, Mario Castro et al. "Bacterial biogeography of adult airways in atopic asthma." *Microbiome* 6, no. 1 (2018): 104.
17. Lubamba, Bob, et al. "Cystic fibrosis: insight into CFTR pathophysiology and pharmacotherapy." *Clinical biochemistry*45.15 (2012): 1132-1144.
18. Molina, Samuel A., and William R. Hunt. "Cystic Fibrosis: An Overview of the Past, Present, and the Future." Lung Epithelial Biology in the Pathogenesis of Pulmonary Disease. Academic Press, 2017. 219-249.
19. Andersen, Dorothy H. "Cystic fibrosis of the pancreas and its relation to celiac disease: a clinical and pathologic study." American journal of Diseases of Children 56.2 (1938): 344-399.
20. Kerem, B. S., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., ... & Tsui, L. C. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science, 245(4922), 1073-1080.
21. Tsui, L. C., Buchwald, M., Barker, D., Braman, J. C., Knowlton, R., Schumm, J. W., ... & Plavsic, N. (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science*, *230*(4729), 1054-1057.
22. Nagel, Georg, et al. "The protein kinase A-regulated cardiac CI− channel resembles the cystic fibrosis transmembrane conductance regulator." Nature 360.6399 (1992): 81.
23. Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H., & Chun, J. (2017). Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *International journal of systematic and evolutionary microbiology*, *67*(5), 1613.
24. Surette, Michael G. "The cystic fibrosis lung microbiome." Annals of the American Thoracic Society 11.Supplement 1 (2014): S61-S65.