



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

UNIVERSITY OF
EXETER



**DEBUGGING LUNG DISEASES: APPLYING
MATHEMATICAL TECHNIQUES, INVOLVING MODELLING,
DATA INTEGRATION AND MACHINE LEARNING FOR
PRECISION MEDICINE**

QUALIFYING EXAMINATION REPORT

SUBMITTED BY: JAYANTH KUMAR NARAYANA
MATRICULATION NUMBER : G1902804D

Contents

0.1	Introduction	1
1	Title1	2
1.1	Introduction	2
1.2	Methods	3
1.2.1	Integrative-microbiomics, a webtool	3
1.2.2	Longitudinal assessment of Exacerbation	6
1.2.3	Antibiotic action simulation	6
1.2.4	“Time to next exacerbation” prediction	6
1.2.5	Validation of the interactome	7
1.3	Results	7
1.3.1	Math Cheat Sheet	7
1.3.2	Table Cheat Sheet	8
1.4	Discussion	8
1.5	Future works	9
2	Title2	10
2.1	Methods	10
2.1.1	Study population	10
2.1.2	Data-preprocessing	11
2.1.3	Co-occurrence analysis	11
2.1.4	11
2.2	Results	11
2.2.1	Significant overlap of fungal communities of lung and gut contrary to bacteria	11

2.2.2	Co-occurrence analysis reveals lung-gut microbial (bacteria and fungi) interactions suggestive of a potential lung-gut axis.	12
2.2.3	Integrated microbiomes identifies a 'high-risk' patient cluster	13
2.2.4	Dysregulated lung-gut axis in high-risk patients	14
3	Discussion	17
4	Future works	1

List of Figures

1.1	A figure illustrating the different sequencing approaches used to derive the human microbiome, consisting of interacting bacteria, fungi and viruses. Adapted from: [15]	3
1.2	(a) A schematic representing, overview of analysis performed on the CAMEB cohort (n=217). Methodologies: Weighted SNF and Co-occurrence analysis were used for microbiome integration and intreactome construction. (b) A patient similarity matrix with each cell representing the integrated similarity between patients. Two clusters of low (black) and high (red) risk patients identified by wSNF are highlighted by boxes. Visualization of the interactome of low (c) and high (d) risk clusters. Interactions between microbes are classified as negative if the sign of the edge weights between them is negative (coloured red) with positive interactions indicated by green colouration. The strength of the interaction is indicated by the colour depth	4
1.3	A figure describing the workflow of integrative microbiomics. The input microbiome datasets, are converted into patient/sample similarity networks based on the user-specified similarity measures: 1) Bray-Curtis, 2) Gower, 3) Canberra and 4) Jaccard; before merging them using the user-specified algorithm: 1) SNF, 2) wSNF. Further, the tool then implements a spectral clustering algorithm to allow cluster analysis on the merged dataset.	5
1.4	Plot of Exponential Function.	7
1.5	Plot of 3-D Function.	8
1.6	Two functions in one plot.	8
2.1	A histogram illustrating all available publications (including original articles and perspectives) matching the keyword “lung-gut axis” from 1900 to 2020 in the web of science database. . . .	11
2.2	A boxplot illustrating the difference in Shannon-diversity index of the Microbiome (Ochre) and Mycobiome (Blue) between the sputum (Lung) and stool (Gut) samples. Statistical significance of these differences were calculated using ‘wilcoxon test’ and are indicated above as p-values.	12
2.3	Microbial association networks derived using co-occurrence analysis methods 1) GBLM(top) and 2)Spiec-easi(bottom). Nodes represents microbes including bacteria and fungi from both lung(left) and gut(right). Edges illustrate the association/interactions between the microbes derived using the respective methods. Highlighted red circle represents the interactions between the lung and gut microbiome.	13

2.4	Integrative microbiome data analysis using MOFA2: (a) A bar chart representing the cumulative variance explained by each of the individual integrated biomes. (b) A heat-map illustrating the breakdown for variance explained of the individual biomes across the first three factors. Box plots illustrating the factor values of factors 1,2 and 3 across exacerbation category (c) and NTM_ever groups (d). [*] illustrate the statistical significance of kruskal-wallis test in terms p-values; [*] p-value < 0.05, ^{**} p-value < 0.001, ^{***} p-value < 0.0001	14
2.5	Boxplots illustrating the differences in Exacerbation (a), FACED (b) and Reiff score (c) between high-risk cluster 1(yellow) and low-risk cluster 2(blue). Statistical significance of these differences were calculated using wilcoxon test and are indicated above as p-values.	15
2.6	Microbial co-occurrence network across the clusters, derived using GBLM with nodes as microbes (bacteria and fungi) from both lung and gut, and edges representing the significant (p-value<0.0001) interaction between nodes. (a) Overlapping microbes are highlighted as yellow nodes. (b) Inter lung-gut microbial interactions are highlighted as red edges.	16

List of Tables

1.1	Caption	9
-----	---------	---

0.1 Introduction

Introduction of the thesis

Chapter 1

Title1

1.1 Introduction

The term microbiome is used to refer to the collection of genes within a community of microbes (including bacteria, fungi, virus, protists and bacteriophages). In the last few years, microbiome research has helped us gained new insights into how microbes shape our human biology and have brought paradigm-shifting implications for translational research and clinical care. The human microbiota is crucial for our body to maintain its homeostasis. Disruption of this can lead to diseases such as obesity, inflammatory bowel disease, malnutrition, Parkinson's, Autism, Asthma, dental caries, bacterial vaginosis, and depression [10]. Currently, microbiome researchers use culture-independent techniques that involve DNA sequencing to derive the microbiome. Broadly, the community taxonomy/microbiome can be identified using two approaches (see Figure1.1) 1) Targeted and 2) Metagenomic. Targeted sequencing approach uses the PCR amplified, target gene markers (16S rRNA in case of bacteria or ITS in case of Fungi) derived from the samples to reference it against gene-marker databases (Silva, Green Genes, etc.). In contrast, the metagenomic sequencing approach directly sequence the whole community DNA and compares it to reference genomes [15].

Present microbiome studies focus on a single profile of the human microbiome in isolation, even though bacteria, fungi and viruses coexist and interact in the body as a community. Thus, it is essential to look at these biological components together in an integrated fashion to understand more holistically the true underlying *in vivo* state. However, one of the primary reasons for the lack of multi-biomic research is the lack of methods to merge microbiome datasets and integrative analysis. Consequently, I tried addressing some of these challenges in my master's thesis, using microbiome datasets derived from bronchiectasis patients as an example. Bronchiectasis, is a chronic inflammatory respiratory disease associated with progressive, irreversible dilatation of the airway. It is crucial to study bronchiectasis because in most cases it is known to be idiopathic(unknown cause) [3] and it is a significant contributor to lung diseases globally with a substantial four-fold higher predominance in Asian populations [17].

Previously in my master's thesis, I developed weighted similarity network fusion (wSNF) to allow weightage of input datasets during integration, otherwise unaccounted by conventional SNF [18]. Ensemble-based co-occurrence analysis strategy developed by Faust et al. [5] was extended to allow weightage of individual methods in the ensemble along with other modification to better infer microbial association networks. Microbiome and Mycobiome derived using targeted amplicon sequencing of the 16S and ITS regions from the sputum samples of the CAMEB cohort [13]; virome from qPCR on an extensive panel of 17 respiratory viruses, were used as the example dataset to integrate the microbiomes (Figure1.2a). Multi-biome (Microbiome, Mycobiome and Virome) integration by wSNF identifies a high-risk exacerbation cluster

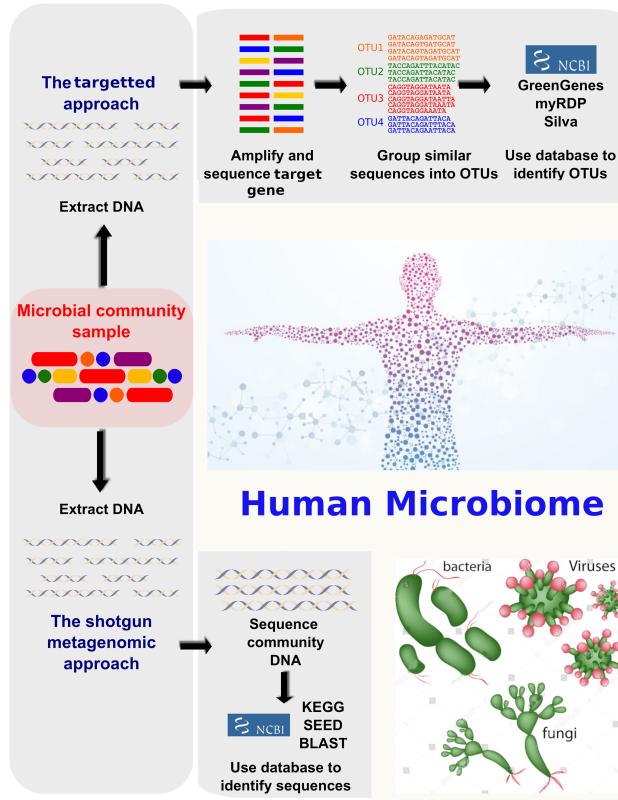


Figure 1.1: A figure illustrating the different sequencing approaches used to derive the human microbiome, consisting of interacting bacteria, fungi and viruses. Adapted from: [15]

with increased precision (Figure 1.2b). Co-occurrence network analysis of this high-risk cluster revealed an elevated antagonistic interactome with reduced alpha-diversity (Figure 1.2c) [16].

Having developed the wSNF and shown its increased precision to identify exacerbators (clinical outcomes); here in this chapter of my PhD thesis, I attempt to extend my results further. I aim to develop a web tool to enable users to integrate their microbiome datasets and to illustrate its advantages using publicly available microbiome datasets. The tool would motivate clinicians and microbiome researchers to explore multi-biome strategies for their problem and aid them in integrating their datasets. Secondly, I aim to study exacerbation events, antimicrobial perturbations and “Time to next exacerbation” using the developed “Interactome” framework. Thirdly, I aim to validate the “high-risk” exacerbation cluster of Bronchiectasis patients and its “interactome” as derived in my previous work [16] using an alternate sequencing approach: metagenomics. Further, we also pick an interaction from the interactome of the high and low-risk clusters and validate it experimentally.

1.2 Methods

1.2.1 Integrative-microbiomics, a webtool

Given the input microbiome datasets, the tool converts them into patient/sample similarity networks for each view based on the user-specified similarity measure before merging them using the user-specified algorithm. Further, the tool then implements a spectral clustering algorithm to allow cluster analysis on the merged dataset outputting the cluster assignments for each sample/patient. The optimum default number of clusters is computed using ensemble-based voting of three differing methodologies: Best Eigen Gap, Rotation cost and average silhouette method (Figure 1.3). For a given value of ‘k’ (the number of clusters), we calculate a

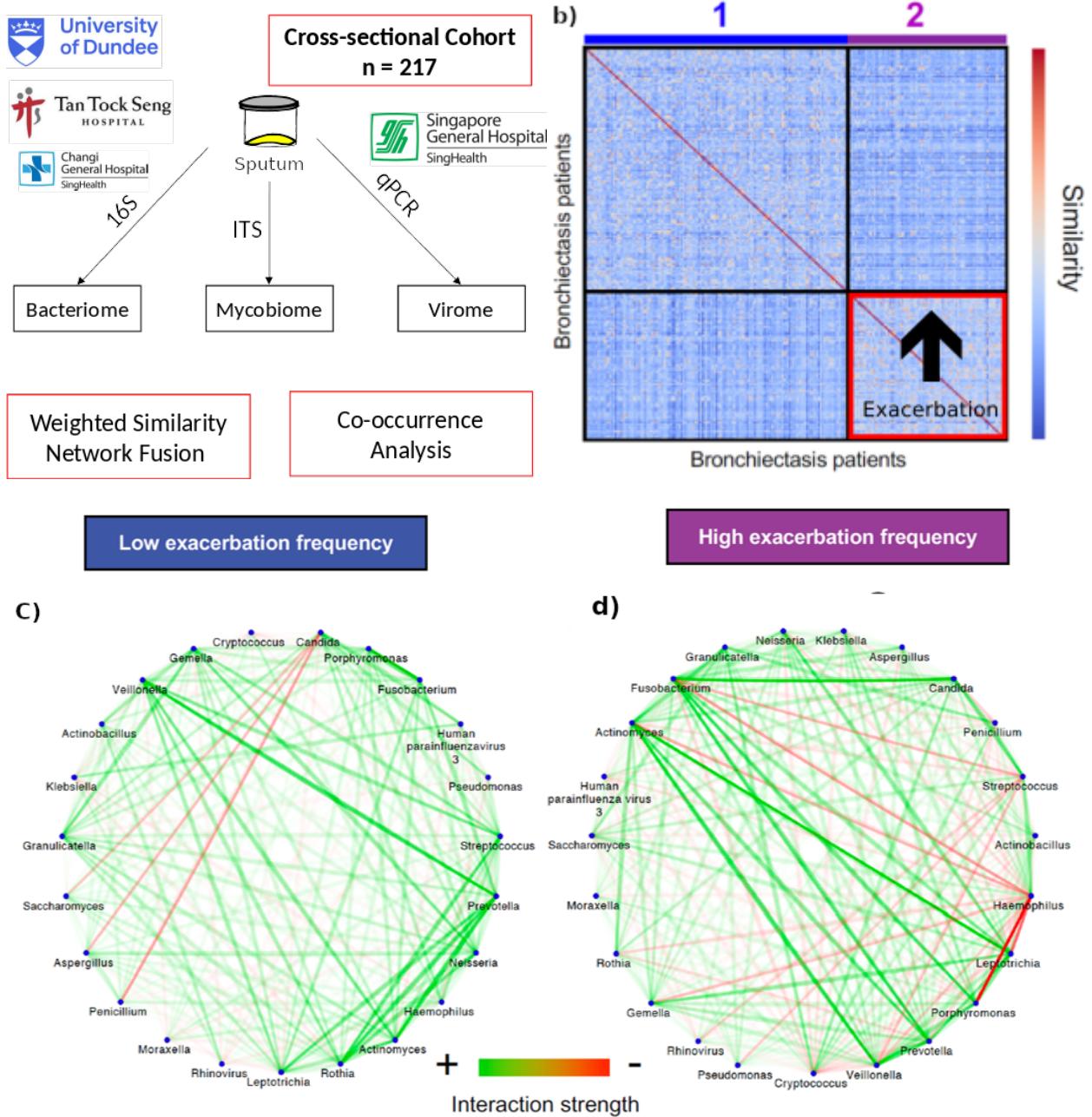


Figure 1.2: (a) A schematic representing, overview of analysis performed on the CAMEB cohort ($n=217$). Methodologies: Weighted SNF and Co-occurrence analysis were used for microbiome integration and interactome construction. (b) A patient similarity matrix with each cell representing the integrated similarity between patients. Two clusters of low (black) and high (red) risk patients identified by wSNF are highlighted by boxes. Visualization of the interactome of low (c) and high (d) risk clusters. Interactions between microbes are classified as negative if the sign of the edge weights between them is negative (coloured red) with positive interactions indicated by green colouration. The strength of the interaction is indicated by the colour depth

score/vote using the below rules

1. If the average silhouette score $\geq 0.7 \rightarrow \text{Score} = \text{Score} + 3$
2. If $0.5 \leq \text{average silhouette score} < 0.7 \rightarrow \text{Score} = \text{Score} + 2$
3. If $0.3 \leq \text{average silhouette score} < 0.5 \rightarrow \text{Score} = \text{Score} + 1$
4. If k equals the first best value as derived from eigen gap method $\rightarrow \text{Score} = \text{Score} + 3$
5. If k equals the second-best value as derived from eigen gap method $\rightarrow \text{Score} = \text{Score} + 2$

INTEGRATIVE MICROBIOMICS

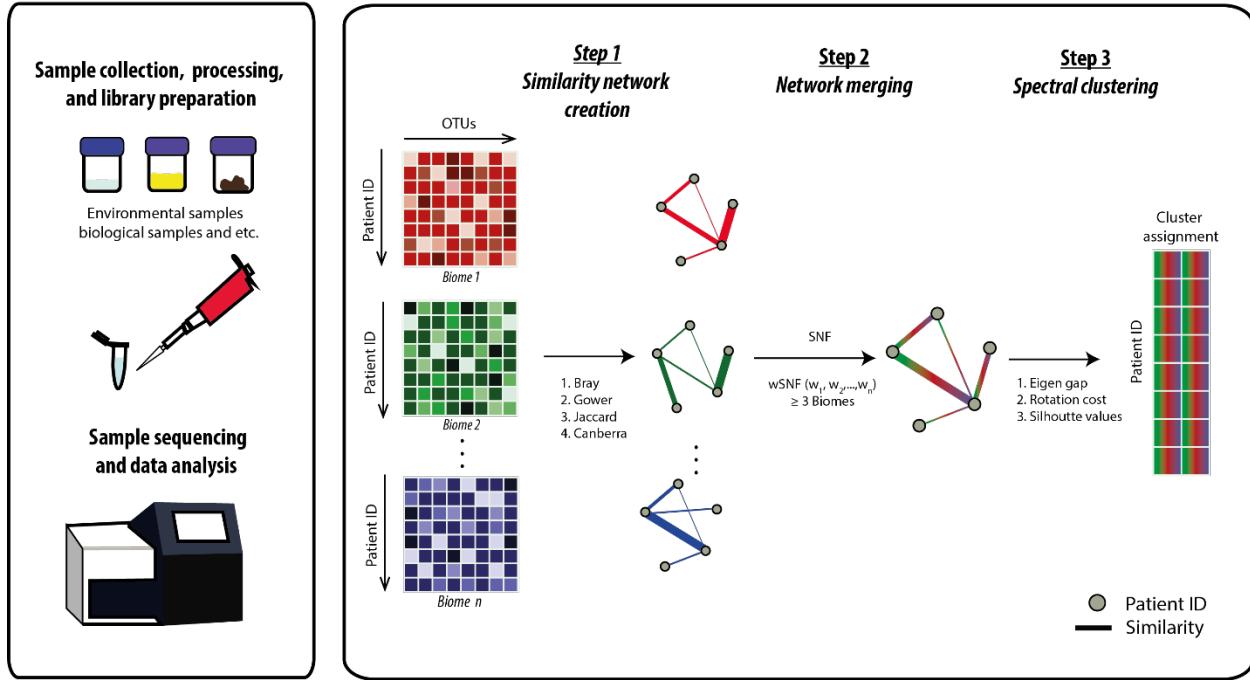


Figure 1.3: A figure describing the workflow of integrative microbiomics. The input microbiome datasets, are converted into patient/sample similarity networks based on the user-specified similarity measures: 1) Bray-Curtis, 2) Gower, 3) Canberra and 4) Jaccard; before merging them using the user-specified algorithm: 1) SNF, 2) wSNF. Further, the tool then implements a spectral clustering algorithm to allow cluster analysis on the merged dataset.

6. If k equals the first best value as derived from rotation cost method → Score = Score + 3
7. If k equals the second-best value as derived from rotation cost method → Score = Score + 2

The value of k for which the Score is the highest is chosen as the default optimum number of clusters. In addition, the tool also outputs the integrated similarity matrix which can be used for downstream analysis such as for label propagation and survival analysis [18].

The tool presently provides four similarity measures 1) Bray-Curtis, 2) Gower, 3) Canberra and 4) Jaccard, appropriate for microbiome datasets which is used to construct patient/sample similarity network and two approaches 1) SNF, 2) wSNF to integrate these networks. For the implementation of wSNF the following formula in SNF

$$P^{(v)} = S^{(v)} \times \frac{\sum_{k=v} p^k}{(m-1)} \times (S^{(v)})^T, v = 1, 2, 3, \dots, m$$

was modified into

$$P^{(v)} = S^{(v)} \times \frac{\sum_{k=v} \omega_k \times p^{(k)}}{\sum_{k \neq v} \omega_k} \times (S^{(v)})$$

$v = 1, 2, 3, \dots, m$ where ω_k is the weight of the k^{th} dataset, m the total number of views, P the status matrix and S the kernel matrix as defined by Wang et.al [18].

This webtool allows the users to integrate multiple microbiome datasets obtained from different sites in a patient/biological entity or from various methods (targeted sequencing, metagenomics and qPCR) from the same site. For example, the lung microbiome (bacteria) with the gut microbiome (bacteria) or the lung microbiome (bacteria) with lung mycobiome. The tool assumes each input microbiome datasets represent a view of an underlying biological mechanism or a disease. Reliable estimation of each view is assumed when using SNF [9]. However, it may not be always practical to reliably estimate each view, although they

play an equal role in the underlying biological process. This is due to the limitations and differing rates of development, in the present technologies and reference databases. In such cases, a weighted SNF approach is preferred, which still assumes the input datasets share an underlying biological mechanism but accounts for inconsistency of the microbiome data based on the user specified weights. The default weights are assigned based on the taxonomical richness (i.e. the number of microbes present) of the datasets.

The interface of the webtool was developed using Rshiny and is available through Shiny Server (Open Source) in confluence with nginx-1.19.1. The tool is powered by custom scripts written in python2.7 and R; and containerized using Docker for ease of offline implementation. The developed webtool can be accessed at <https://integrative-microbiomics.ntu.edu.sg>.

1.2.2 Longitudinal assessment of Exacerbation

A longitudinal cohort of n=17 patients were recruited from two hospitals in the east of Scotland (2016-2017) to study changes in the microbiome during exacerbation and following antibiotic treatment. DNA and RNA extraction were performed on sputum samples obtained from each patient and on a blank sterile PBS (Phosphate buffer solution). The extracted DNA was subjected to targeted amplicon sequencing of the 16S rRNA and ITS2 regions of the genome to derive the Microbiome and Mycobiome, by mapping them to green genes and UNITE databases, respectively. Blank samples contained read counts many orders of magnitude lower than test samples and hence unlikely to have any influence on the observed microbiome. RT-qPCR (real time quantitative polymerase chain reaction) was performed on the cDNA derived from the extracted RNA to quantify the viral burden of the 17 viruses investigated in each patient. α and β diversity of the multi-biome was calculated from the concatenated microbiome and the integrated patient similarity matrix using the vegan package in R.

1.2.3 Antibiotic action simulation

To predict the impact of antibiotics on the interactome, β -lactam antibiotic action was simulated by a 75% reduction in the relative abundance of the microbes targeted by this antibiotic in the baseline (pre-antibiotic) state including the following genera: *Streptococcus*, *Staphylococcus*, *Haemophilus*, *Moraxella*, *Actinomyces*, *Arachnia*, *Bacteroides*, *Bifidobacterium*, *Eubacterium*, *Fusobacterium*, *Lactobacillus*, *Leptotrichia*, *Peptococcus*, *Peptostreptococcus*, *Propionibacterium*, *Selenomonas*, *Treponema* and *Veillonella*. In order to remove interactions resulting from random noise at the expense of sensitivity to weak signals and to allow comparison between the derived interactomes, the following abundance and prevalence filters were applied followed by co-occurrence analysis; retention of microbes present at greater than 1% abundance in at least three subjects; in the pre OR post OR modelled antibiotic state.

1.2.4 “Time to next exacerbation” prediction

To predict Time to next exacerbation, Microbiome datasets were CLR (Centred log ratio) transformed before concatenation and microbes that are present in at least 4 patients at an abundance of 1% were considered for further analysis. To derive pairwise microbial interactions for each patient, LIONESS [11], a single patient network inference framework was implemented with General Boosted Linear model (GBLM) as the network inference algorithm. Correlation between the abundance of each microbes and interaction strength with time to next exacerbation was assessed using Spearmans rank correlation with statistical testing. Multivariate adaptive regression spline (MARS) [6], a non-linear regression model was implemented with microbes or

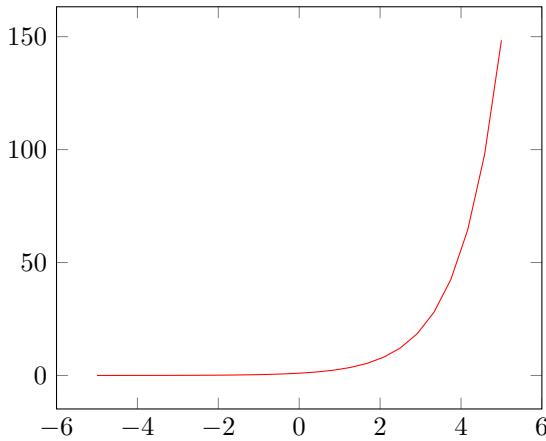


Figure 1.4: Plot of Exponential Function.

interaction strength as the predictor variable to predict time to next exacerbation groups; defined as (Time to exacerbation: <12 weeks and >12weeks). The goodness of the fit of the model was evaluated by computing the R-squared (RSq) and the Generalized R-squared metric (GRsq). A feature importance plot based on Generalized Cross validation score (gcv) was also computed on the feature selected (microbes) by the model. All the above analysis was implemented in R using the following packages 1)Hmisc 2) earth 3)vegan 4)compositions 5)lionessR.

1.2.5 Validation of the interactome

Experimental microbiological validation of *Pseudomonas aeruginosa* and *Aspergillus fumigatus* interaction was performed using one strain *Aspergillus fumigatus* (Af293) and three strains of *Pseudomonas aeruginosa*: (1) lab strain (PAO1) as control and (2,3) two clinical isolates of *Pseudomonas aeruginosa* derived from low-risk and high-risk patient clusters. The interaction was investigated using the disk inhibition method as described by Homa et al. [8]. An independent cohort of 166 patients was recruited from 4 sites (3 in Singapore and 1 in Dundee, Scotland) to validate the high-risk cluster and its interactome. DNA extraction was performed on the collected sputum samples of each patient. A shotgun metagenomic sequencing was performed at the NTU core sequencing facility on these samples according to the methods described by Gusareva et al. [7]. Kaiju [14] with default parameters was implemented on the raw sequences after human read removal to estimate the taxonomic composition by referencing against NCBI BLAST nr+euk database. Estimation of the viruses that include prokaryotic phages and eukaryotic viruses was implemented using a custom pipeline that uses Demovir (<https://github.com/feargalr/Demovir>).

1.3 Results

1.3.1 Math Cheat Sheet

Figure 1.4, Figure 1.5, and Figure 1.6 are generated from function in L^AT_EX.

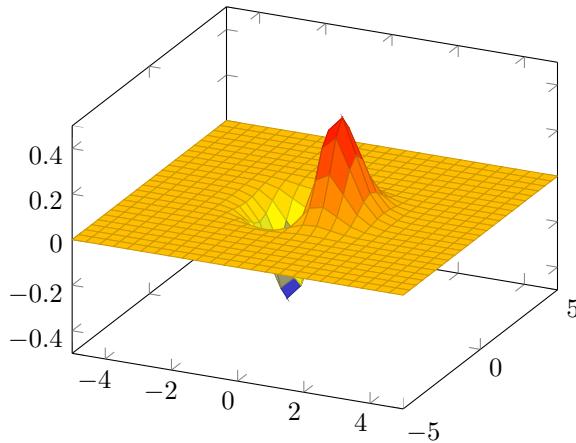


Figure 1.5: Plot of 3-D Function.

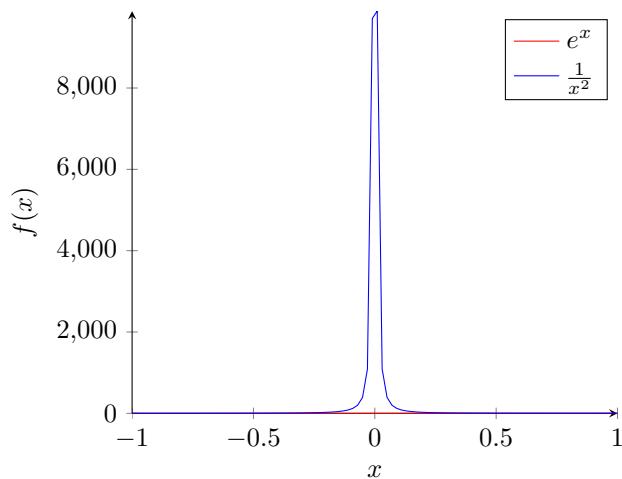


Figure 1.6: Two functions in one plot.

1.3.2 Table Cheat Sheet

1.4 Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc, quis gravida magna mi a libero. Fusce vulputate eleifend sapien. Vestibulum purus quam, scelerisque ut, mollis sed, nonummy id, metus. Nullam accumsan lorem in dui. Cras ultricies mi eu turpis hendrerit fringilla.

	Apple	Oranges	Strawberries
A	1	2	3
B	1	2	3
C	1	2	3
D	1	2	3

Table 1.1: Caption

Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; In ac dui quis mi consectetur lacinia. Nam pretium turpis et arcu. Duis arcu tortor, suscipit eget, imperdiet nec, imperdiet iaculis, ipsum. Sed aliquam ultrices mauris. Integer ante arcu, accumsan a, consectetur eget, posuere ut, mauris. Praesent adipiscing. Phasellus ullamcorper ipsum rutrum nunc. Nunc nonummy metus. Vestibulum volutpat pretium libero. Cras id dui. Aenean ut

1.5 Future works

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc, quis gravida magna mi a libero. Fusce vulputate eleifend sapien. Vestibulum purus quam, scelerisque ut, mollis sed, nonummy id, metus. Nullam accumsan lorem in dui. Cras ultricies mi eu turpis hendrerit fringilla. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; In ac dui quis mi consectetur lacinia. Nam pretium turpis et arcu. Duis arcu tortor, suscipit eget, imperdiet nec, imperdiet iaculis, ipsum. Sed aliquam ultrices mauris. Integer ante arcu, accumsan a, consectetur eget, posuere ut, mauris. Praesent adipiscing. Phasellus ullamcorper ipsum rutrum nunc. Nunc nonummy metus. Vestibulum volutpat pretium libero. Cras id dui. Aenean ut [?]

Chapter 2

Title2

Introduction

The Gut-microbiome is by far the best and widely studied microbial ecosystem of the human anatomy, partly due to the rich microbial environment and partly due to the ease of sample collection (non-invasive) through faeces [2]. On the other hand, healthy lung was long considered to be sterile but with advent high-throughput sequencing techniques this has been proven otherwise [cite]. Extensive research on the gut-microbiota has shown that gut microbiota is capable of influencing other organs, such as the brain, liver or lungs [1]. This has led to the coining of terms such as the ‘gutbrain axis’ and the ‘lung-gut axis’.

The epithelial surfaces of the gut and lung are exposed to diverse microbes; ingested microorganisms can access both sites and the microbiota from the gut can enter the lungs through processes such as micro-aspiration [2]. Furthermore, the lung and gut can interact thorough the systemic cytokines released by host immune cells in response to microbes or microbes from one-site may secrete metabolites which are absorbed into the blood stream and thus regulate the organs [4]. A study used germ free mice, which lack an appropriately developed immune system and showed mucosal alterations, both of which is restored through colonization with gut microbiota. Thus, supplementing the concept of ‘lung-gut’ axis [2].

Literature survey using the keyword ‘lung-gut axis’ shows that this concept was first introduced in 2004 and increasing work is being done [2]. This increasing evidence also suggests, a potential existence of lung-gut axis and its effectual role in lung diseases. Although the gut-lung axis is only beginning to be understood, emerging evidence indicates that there is potential for manipulation of the gut microbiota in the treatment of lung diseases. Despite this, the influence of microbial gut health in Bronchiectasis lung is poorly studied. Hence, in the second chapter of my PhD thesis

Want to check if microbes interact

2.1 Methods

2.1.1 Study population

57 patients with stable bronchiectasis were recruited as a part of this study by our collaborators in Milan, Italy. Recruitment of patients was performed cross-sectionally as a part of the Bronchiectasis Program of

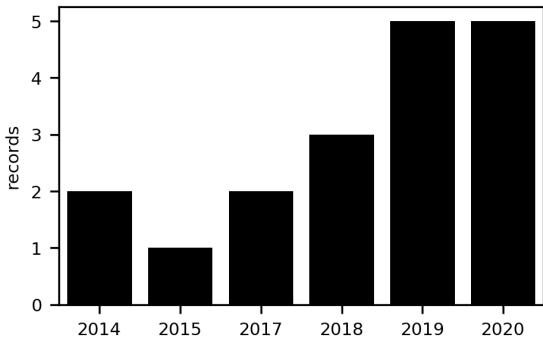


Figure 2.1: A histogram illustrating all available publications (including original articles and perspectives) matching the keyword “lung-gut axis” from 1900 to 2020 in the web of science database.

Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, Milan, Italy. Patients were enrolled during their clinical stability (at least one month apart from the last exacerbation and antibiotic course) and underwent clinical, radiological and microbiological evaluation. Patients were asked to provide a sputum and stool sample with a maximum gap of 12hrs. between the sputum and stool sample. DNA was extracted by our lab members from the sputum and stool samples as described previously [13]. The extracted DNA was subjected to targeted amplicon sequencing of the 16S rRNA and ITS2 regions of the genome to derive the Microbiome and Mycobiome, by mapping them to green genes and UNITE databases, respectively.

2.1.2 Data-preprocessing

Read counts of Microbiome and Mycobiome datasets from the sputum and stool samples of the 57 samples were converted into relative abundances. Only microbes present $\geq 1\%$ in atleast 5 patients were considered for further analysis.

2.1.3 Co-occurrence analysis

2.1.4

We report the first integration of lung-gut microbiome (bacteria and fungi) in bronchiectasis (n=57) perfomed using MOFA2 and a weighted similarity network fusion approach following spectral clustering. ALDEX2 was implemented to identify discriminant taxa. Co-occurrence analysis was employed using GBLM to dervie microbial association networks.

2.2 Results

2.2.1 Significant overlap of fungal communities of lung and gut contrary to bacteria

Intersection analysis of bacterial and fungal communities between sputum and stool samples reveals increased overlap of fungal communities between the lung and gut, contrary to bacteria. Three bacterial genera including *Lactobacillus*, *Prevotella* and *Streptococcus* compared to six fungal genera including *Candida*, *Cryptococcus*, *Curvularia*, *Debaryomyces*, *Lodderomyces* and *Saccharomyces* were present in both sputum

and stool samples. Interestingly, upon assessment of diversity between the sputum and stool samples, a similar pattern is observed. Overall, mycobiome exhibits a decreased diversity compared to microbiome. Further, an increased diversity of bacteria is found in the gut compared to the lung whereas the fungal diversity doesn't change.

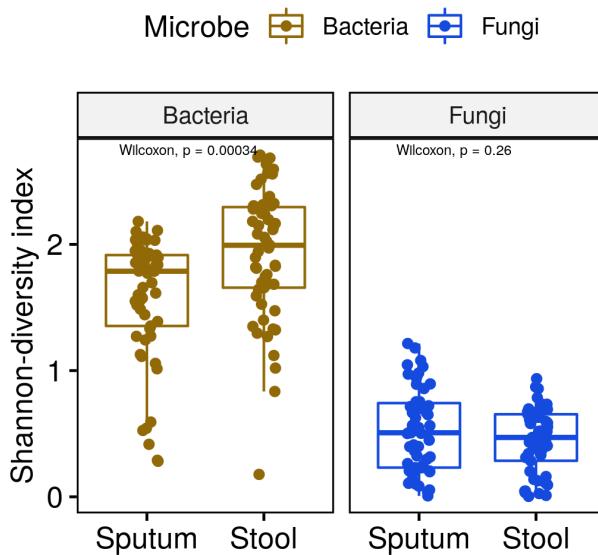


Figure 2.2: A boxplot illustrating the difference in Shannon-diversity index of the Microbiome (Ochre) and Mycobiome (Blue) between the sputum (Lung) and stool (Gut) samples. Statistical significance of these differences were calculated using ‘wilcoxon test’ and are indicated above as p-values.

2.2.2 Co-occurrence analysis reveals lung gut microbial (bacteria and fungi) interactions suggestive of a potential lung-gut axis.

Assessment of interactions of microbes (bacteria and fungi) between the lung and gut was construed using co-occurrence analysis. Co-occurrence analysis was performed using two methods 1)GBLM: captures complex linear interactions and 2) Spiec-easi: captures non-linear interactions but assumes sparsity, to derive microbial association networks. Microbial associations networks from both these methods shows cross-talk between the microbes of the lung and the gut, indicative of potential existence of the lung-gut axis [Figure2.3]. However, a greater number of inter-axis interactions is observed through the use of GBLM methods as compared to spiec-easi and this is partially due to the assumption of sparsity by spiec-easi. Integrative assessment of the microbiomes was performed using MOFA2. Roughly, MOFA2 works like a PCA to create factors which maximises the explained variance of the integrated microbiome. MOFA2 analysis on the microbiomes (bacteria lung, bacteria gut, fungi lung and fungi gut) reveals a factor (Factor1) associated with exacerbations and Non-tuberculosis mycobacterial(NTM) infections [Figure2.4(c,d)]. Following, assessment of Factor1 in terms of contribution from individual microbiomes reveal gut microbiome as the highest contributor [Figure2.4(b)], possibly explainable by lung-gut axis which further supplements the existence of lung-gut axis. Moreover, upon assessment of the loadings of Factor1 we find four of the six overlapping microbes including *Streptococcus*, *Saccharomyces*, *Candida* and *Curvularia* have loadings $\geq 0.5\%$.

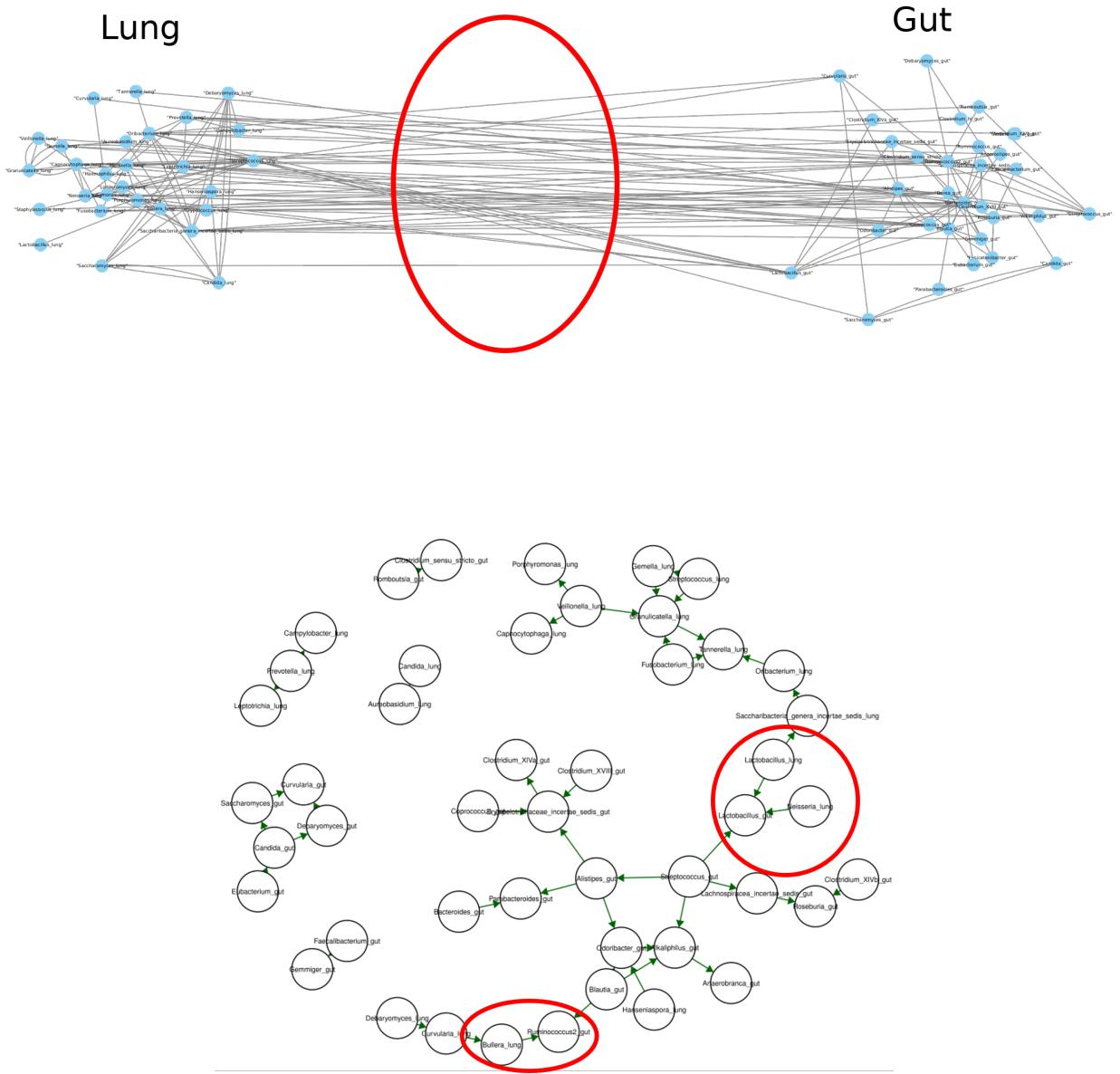


Figure 2.3: Microbial association networks derived using co-occurrence analysis methods 1) GBLM(top) and 2)Spieci-easi(bottom). Nodes represents microbes including bacteria and fungi from both lung(left) and gut(right). Edges illustrate the association/interactions between the microbes derived using the respective methods. Highlighted red circle represents the interactions between the lung and gut microbiome.

2.2.3 Integrated microbiomes identifies a 'high-risk' patient cluster

Integration of microbiomes and mycobiomes from lung and gut using weighted SNF with $k=9$ was performed, following which spectral clustering was implemented to cluster the patients into two groups. Li *et.al.* in their paper showed that integrating multiple views of the same patient/sample increases the power and compensates for smaller sample sizes[12]. Hence, clustering of the integrated microbiomes of these ($n=57$) patients is admissible. Cluster robustness was evaluated using a bootstrap approach and was found to be 79.15%. Evaluation of the derived clusters across clinical attributes reveals patients belonging to cluster1 have a higher median risk of exacerbation, FACED score and Reiff score [Figure2.5] compared to cluster2. Differential analysis illustrates significantly increased *Candida* in Gut and *Fusobacterium* in Lung, of high-risk patients (cluster1) compared to that of low-risk patients(cluster2).

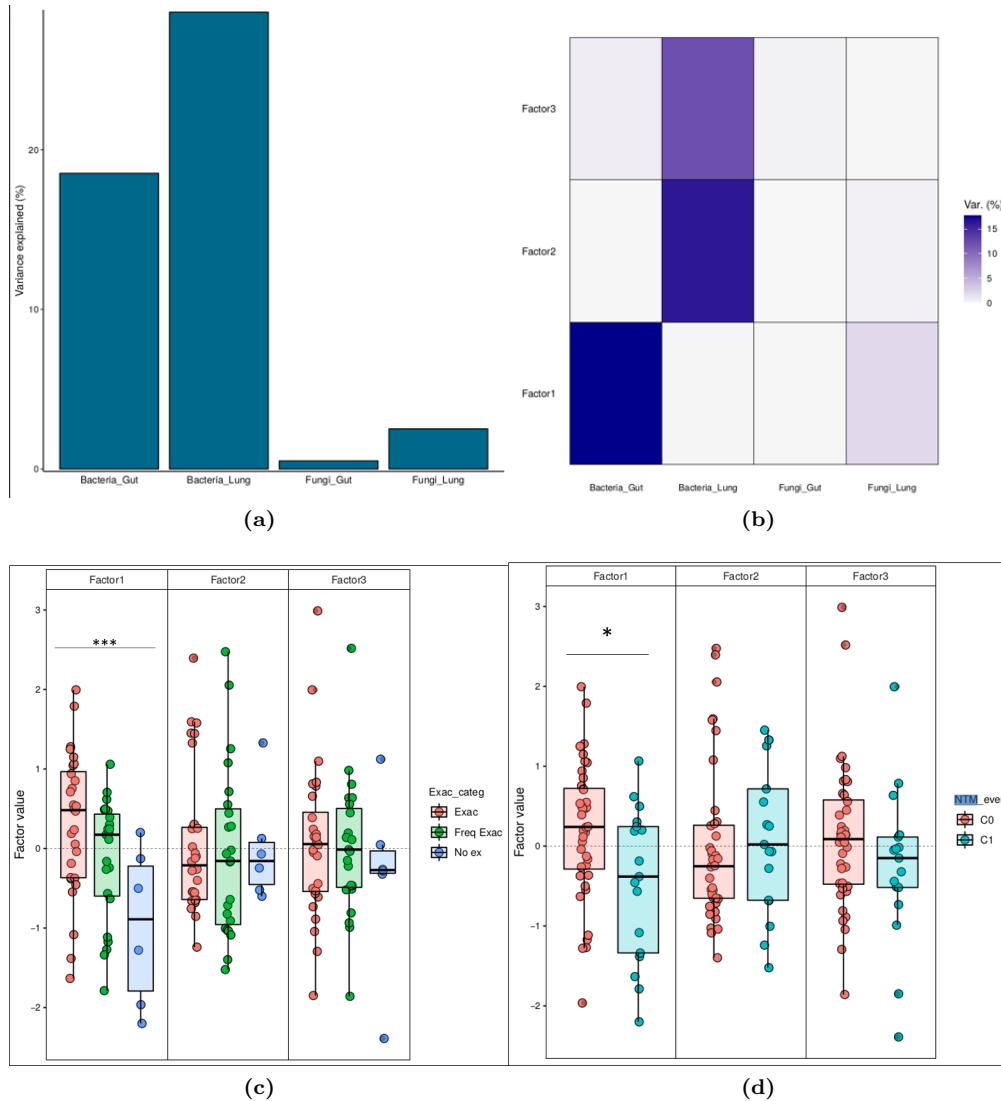


Figure 2.4: Integrative microbiome data analysis using MOFA2: (a) A bar chart representing the cumulative variance explained by each of the individual integrated biomes. (b) A heat-map illustrating the breakdown for variance explained of the individual biomes across the first three factors. Box plots illustrating the factor values of factors 1,2 and 3 across exacerbation category (c) and NTM_{ever} groups (d).^{*} illustrate the statistical significance of kruskal-wallis test in terms p-values; ^{*} p-value < 0.05, ^{**} p-value < 0.001, ^{***} p-value < 0.0001

2.2.4 Dysregulated lung-gut axis in high-risk patients

Having shown the existence of lung-gut axis and identifying a sub-group of high-risk bronchiectasis patients using the integrative microbiomics. We next evaluated the changes in lung-gut axis across the two clusters. Microbes between the lung and gut can broadly interact in two ways: 1) Microbes can travel between the sites through mechanisms such as micro-aspiration and 2) Microbes can secrete secondary metabolites and other biomolecules through which they can interact. To assess the first mechanism, a linear correlation analysis between the clr transformed abundance of the overlapping microbes from the lung and gut was performed. *Lactobacillus* lung was found to be significantly correlated with *Lactobacillus* gut in high-risk cluster 1 but not in cluster 2. However, the observed correlation of *Lactobacillus* between the two sites may be due to confounding microbes. Therefore, a GBLM analysis was performed to assess the association of overlapping microbes between the two sites given all the other microbes. Interestingly, the correlation between *Lactobacillus* lung and *Lactobacillus* gut disappears. However, a new association between *Streptococcus* lung and *Streptococcus* gut is found in high-risk cluster 1 but not in low-risk patients; suggestive of movement of *Streptococcus* between lung and gut in high-risk patients (probably due to dysregulation of

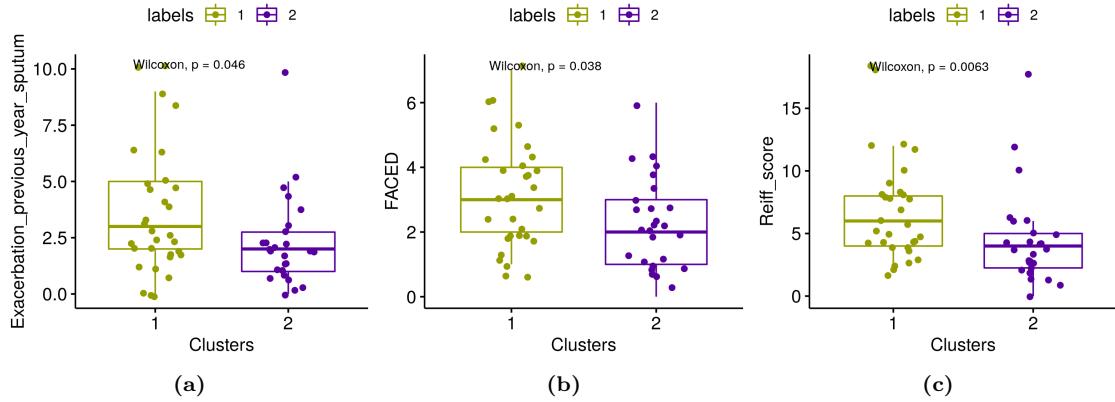
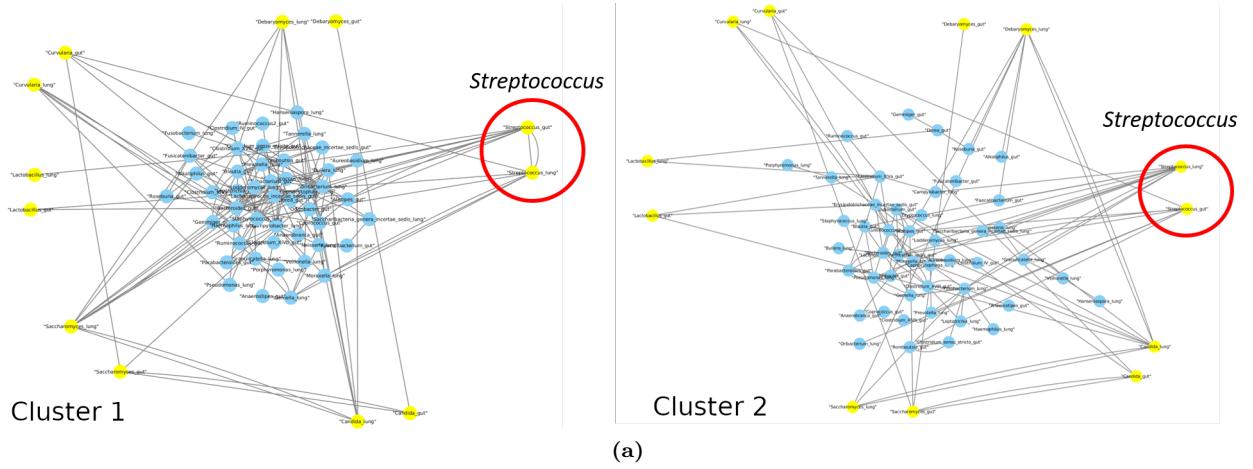
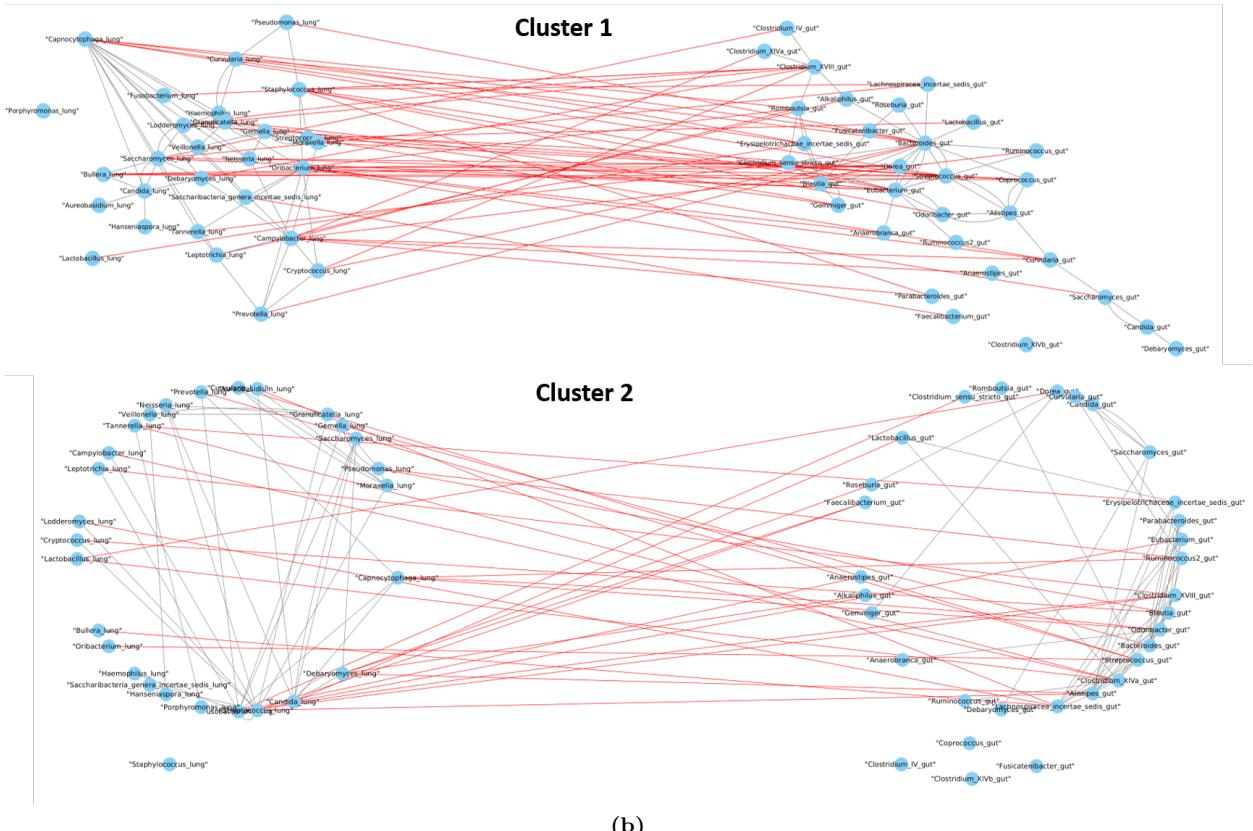


Figure 2.5: Boxplots illustrating the differences in Exacerbation (a), FACED (b) and Reiff score (c) between high-risk cluster 1(yellow) and low-risk cluster 2(blue). Statistical significance of these differences were calculated using wilcoxon test and are indicated above as p-values.

the axis) [Figure2.6(a)]. Assessment of the second mechanism was performed by evaluating the inter-axis microbial interactions between clusters. This reveals an increased lung-gut microbial interaction in high-risk cluster(35%) compared to low-risk cluster(29%) [Figure2.6(b)]; further supplementing the dysregulation of this axis in high-risk patients.



(a)



(b)

Figure 2.6: Microbial co-occurrence network across the clusters, derived using GBLM with nodes as microbes (bacteria and fungi) from both lung and gut, and edges representing the significant ($p\text{-value} < 0.0001$) interaction between nodes. (a) Overlapping microbes are highlighted as yellow nodes. (b) Inter lung-gut microbial interactions are highlighted as red edges.

Chapter 3

Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc, quis gravida magna mi a libero. Fusce vulputate eleifend sapien. Vestibulum purus quam, scelerisque ut, mollis sed, nonummy id, metus. Nullam accumsan lorem in dui. Cras ultricies mi eu turpis hendrerit fringilla. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; In ac dui quis mi consectetur lacinia. Nam pretium turpis et arcu. Duis arcu tortor, suscipit eget, imperdiet nec, imperdiet iaculis, ipsum. Sed aliquam ultrices mauris. Integer ante arcu, accumsan a, consectetur eget, posuere ut, mauris. Praesent adipiscing. Phasellus ullamcorper ipsum rutrum nunc. Nunc nonummy metus. Vestibulum volutpat pretium libero. Cras id dui. Aenean ut

Chapter 4

Future works

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc, quis gravida magna mi a libero. Fusce vulputate eleifend sapien. Vestibulum purus quam, scelerisque ut, mollis sed, nonummy id, metus. Nullam accumsan lorem in dui. Cras ultricies mi eu turpis hendrerit fringilla. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; In ac dui quis mi consectetur lacinia. Nam pretium turpis et arcu. Duis arcu tortor, suscipit eget, imperdiet nec, imperdiet iaculis, ipsum. Sed aliquam ultrices mauris. Integer ante arcu, accumsan a, consectetur eget, posuere ut, mauris. Praesent adipiscing. Phasellus ullamcorper ipsum rutrum nunc. Nunc nonummy metus. Vestibulum volutpat pretium libero. Cras id dui. Aenean ut [?]

Bibliography

- [1] BELL, J., SPENCER, J., YATES, R., YEE, S., JACOBS, B., AND DELUCA, G. Invited review: From nose to gut—the role of the microbiome in neurological disease. *Neuropathology and applied neurobiology* 45, 3 (2019), 195–215.
- [2] BUDDEN, K. F., GELLATLY, S. L., WOOD, D. L., COOPER, M. A., MORRISON, M., HUGENHOLTZ, P., AND HANSBRO, P. M. Emerging pathogenic links between microbiota and the gut–lung axis. *Nature Reviews Microbiology* 15, 1 (2017), 55–63.
- [3] CHALMERS, J. D., AND CHOTIRMALL, S. H. Bronchiectasis: new therapies and new perspectives. *Lancet Respir Med* 6, 9 (Sep 2018), 715–726.
- [4] DANG, A. T., AND MARSLAND, B. J. Microbes, metabolites, and the gut–lung axis. *Mucosal Immunology* 12, 4 (2019), 843–850.
- [5] FAUST, K., SATHIRAPONGSASUTI, J. F., IZARD, J., SEGATA, N., GEVERS, D., RAES, J., AND HUTTENOWER, C. Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology* 8, 7 (2012), e1002606.
- [6] FRIEDMAN, J. H. Multivariate adaptive regression splines. *The Annals of Statistics* 19, 1 (mar 1991), 1–67.
- [7] GUSAREVA, E. S., ACERBI, E., LAU, K. J. X., LUHUNG, I., PREMKRISHNAN, B. N. V., KOLUNDŽIJA, S., PURBOJATI, R. W., WONG, A., HOUGHTON, J. N. L., MILLER, D., GAULTIER, N. E., HEINLE, C. E., CLARE, M. E., VETTATH, V. K., KEE, C., LIM, S. B. Y., CHÉNARD, C., PHUNG, W. J., KUSHWAHA, K. K., NEE, A. P., PUTRA, A., PANICKER, D., YANQING, K., HWEE, Y. Z., LOHAR, S. R., KUWATA, M., KIM, H. L., YANG, L., UCHIDA, A., DRAUTZ-MOSES, D. I., JUNQUEIRA, A. C. M., AND SCHUSTER, S. C. Microbial communities in the tropical air ecosystem follow a precise diel cycle. *Proceedings of the National Academy of Sciences* 116, 46 (oct 2019), 23299–23308.
- [8] HOMA, M., SÁNDOR, A., TÓTH, E., SZEBENYI, C., NAGY, G., VÁGLGYI, C., AND PAPP, T. In vitro interactions of pseudomonas aeruginosa with scedosporium species frequently associated with cystic fibrosis. *Frontiers in Microbiology* 10 (mar 2019).
- [9] JIANG, D., ARMOUR, C. R., HU, C., MEI, M., TIAN, C., SHARPTON, T. J., AND JIANG, Y. Microbiome multi-omics network analysis: Statistical considerations, limitations, and opportunities. *Frontiers in Genetics* 10 (nov 2019).
- [10] KNIGHT, R., CALLEWAERT, C., MAROTZ, C., HYDE, E. R., DEBELIUS, J. W., McDONALD, D., AND SOGIN, M. L. The microbiome and human biology. *Annual Review of Genomics and Human Genetics* 18, 1 (aug 2017), 65–86.
- [11] KUIJPER, M. L., HSIEH, P.-H., QUACKENBUSH, J., AND GLASS, K. lionessR: single sample network inference in r. *BMC Cancer* 19, 1 (oct 2019).

- [12] LI, C.-X., WHEELOCK, C. E., SKLD, C. M., AND WHEELOCK, Å. M. Integration of multi-omics datasets enables molecular classification of COPD. *European Respiratory Journal* 51, 5 (mar 2018), 1701930.
- [13] MAC AOGÁIN, M., CHANDRASEKARAN, R., LIM YICK HOU, A., TECK BOON, L., LIANG TAN, G., HASSAN, T., THUN HOW, O., HUI QI NG, A., BERTRAND, D., YU KOH, J., LEI PANG, S., YANG LEE, Z., WEI GWEE, X., MARTINUS, C., YIE SIO, Y., ANUSHA MATTA, S., TIM CHEW, F., KEIR, H. R., CONNOLLY, J. E., ARPUTHAN ABISHEGANADEN, J., SIYUE KOH, M., NAGARAJAN, N., CHALMERS, J. D., AND CHOTIRMALL, S. H. Immunological corollary of the pulmonary mycobiome in bronchiectasis: The cameb study. *European Respiratory Journal* (2018).
- [14] MENZEL, P., NG, K. L., AND KROGH, A. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications* 7, 1 (apr 2016).
- [15] MORGAN, X. C., AND HUTTENHOWER, C. Human microbiome analysis. *PLoS Comput Biol* 8, 12 (2012), e1002808.
- [16] NARAYANA, J. K. Investigating the respiratory microbiome in bronchiectasis through "integrative micro- biomics". Master's thesis, Indian Institute of Science Education and Research, Pune, 2019.
- [17] SEITZ, A. E., OLIVIER, K. N., ADJEMIAN, J., HOLLAND, S. M., AND PREVOTS, D. R. Trends in bronchiectasis among medicare beneficiaries in the united states, 2000 to 2007. *CHEST* 142, 2 (Aug 2012), 432–439.
- [18] WANG, B., MEZLINI, A. M., DEMIR, F., FIUME, M., TU, Z., BRUDNO, M., HAIBE-KAINS, B., AND GOLDENBERG, A. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* 11 (Jan 2014), 333 EP –. Article.