

Investigating the respiratory microbiome in Bronchiectasis through “Integrative Microbiomics”

A Thesis

submitted to

Indian Institute of Science Education and Research Pune

in partial fulfillment of the requirements for the

BS-MS Dual Degree Programme

by

Jayanth Kumar Narayana



Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,

Pashan, Pune 411008, INDIA.

June, 2019

Supervisor: Asst. Prof. Sanjay Haresh Chotirmall

Co-supervisor: Prof. Krasimira Tsaneva-Atanasova

© Jayanth Kumar Narayana 2019

All rights reserved

Certificate

This is to certify that this dissertation entitled Investigating the respiratory microbiome in Bronchiectasis through “Integrative Microbiomics” towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Jayanth Kumar Narayana at Lee Kong Chian School of Medicine under the supervision of Asst. Prof. Sanjay Haresh Chotirmall, MD, PhD, Department of Medicine and Prof. Krasimira Tsaneva-Atanasova, PhD, Department of Mathematics, during the academic year 2018-2019.



Prof. Krasimira Tsaneva-Atanasova



Asst. Prof. Sanjay Haresh Chotirmall

Committee:

Asst. Prof. Sanjay Haresh Chotirmall

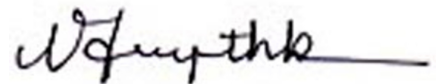
Asoc. Prof. Pranay Goel

Prof. Krasimira Tsaneva-Atanasova

This thesis is dedicated to my parents, who made me who I am

Declaration

I hereby declare that the matter embodied in the report entitled Investigating the respiratory microbiome in Bronchiectasis through “Integrative Microbiomics” are the results of the work carried out by me at the Department of Medicine, Lee Kong Chian School of Medicine, under the supervision of Asst. Prof. Sanjay Haresh Chotirmall and co-supervision of Prof. Krasimira Tsaneva-Atanasova, and the same has not been submitted elsewhere for any other degree.



Jayanth Kumar Narayana

Abstract

Studies of the human microbiome have brought paradigm-shifting implications for translational research and clinical care, and, is now recognized as significant across a range of human organ systems. Despite significant progress in the field over the last decade, a holistic analysis of bacteria, fungi and viruses (the “multi-biome”) is rarely performed despite this most closely representing the true in-vivo state. Integration of these high-dimensional datasets brings challenges in terms of complexity and their translation into clinically actionable outputs. To address this “analytical bottleneck”, we sought to build a computational pipeline for integration of bacterial, fungal and viral datasets from a single well characterised patient population (a process we coin “integrative microbiomics”) as a proof of principle in work described below. Having successfully integrated bacterial, fungal and viral datasets, we characterise the integrated microbial components by identifying a statistically significant super-consensus network representing possible mathematical microbial interactions (which we term the “interactome”). Further, we show that cross-talk between microbes is as significant as the isolated microbes not if higher, in driving specific disease states.

Contents

Abstract	ix
1 Introduction	7
1.1 Microbes and Microbiomes	7
1.2 Chronic Respiratory Diseases	9
1.3 Case study: The Microbiome in Bronchiectasis	11
1.4 Summary	13
2 Methods	15
2.1 Introduction	15
2.2 Network inference using Similarities	16
2.3 Similarity Network Fusion	17
2.4 Weighted Similarity Network Fusion	19
2.5 Cluster Analysis	20
2.6 Cluster Characterisation	21
2.7 Co-occurrence Analysis	23
2.8 Network Construction	28
2.9 Network Analysis using Cytoscape	31

3	Results	33
3.1	Bacteriome clusters identify high risk patients	33
3.2	Integrative Microbiomics using Similarity Network Fusion (Unweighted) . . .	34
3.3	Integrative Microbiomics using weighted Similarity Network Fusion indentifies a high-risk cluster with increased precision	36
3.4	Co-occurrence analysis reveals difference in number of negative interaction between clusters	38
3.5	Busy, influential and critical microbes among indentified clusters	38
3.6	<i>Pseudomonas</i> specific interaction in the clusters	42
4	Discussion	43
4.1	Introduction	43
4.2	Integration of Microbiomes	43
4.3	Microbes and Microbial association network	45
4.4	Clinical Relevance	46
4.5	Conclusion	48

List of Figures

2.1	Work-flow of integrating biomes: Data matrices of microbiomes are converted into Similarity matrices using Bray-Cutis similarity as given in definition 2.2.1. Network of patients for each biome is created based on these similarity matrices. Further, these matrices are merged using SNF and spectral clustering is implemented on this merged network to find subgroups of patients. Figure adapted from [Wang et al., 2014]	18
2.2	An ensemble approach of Network Inference: Network of microbes are built on microbes from all the biomes. Edge weights along with their statistical significance, are ascertained using four different similarity measures and GBLM(Gradient Boosting with Component-wise linear models), resulting in five different microbial networks one based on each measure. A merged microbial network is derived from the 5 networks with appropriate weighting(green colored text). P-values are combined using weighted Sime’s test and edge weights are a weighted sum is taken after proper standardisation.	25
2.3	Network construction from similarities: 1)Similarity measure is calculated between every pair of species over all patients. 2)Bootstrap distribution of for each similarity measure is constructed. 3)Reboot[Permutation and re-normalisation] distribution for each similarity measure is built. 4)Distributional difference between bootstrap and reboot is assessed using Mann-Whitney U test and a p-value is assigned for each edge/similarity measure. .	28
2.4	GBLM for network construction: 1)GBLM model was implemented to predict each microbe from all other microbes. The coefficients of the microbes served as edge weights. 2)Bootstrap distribution of these coefficients were created. 3) A reboot(permutation-renormalisation) distribution of these coefficients were created. 4)Statistical significance(p-values) of the coefficients were assessed using Mann-Whitney U test between bootstrap and Reboot distribution.	30

3.1 **Spectral clustering on the bacteriome of CAMEB patients:** A) PCoA(Principal Coordinate Analysis) plot of the bacteriome clusters using bray-curtis dissimilarity with x and y axis representing the first and second principal axis respectively. B) A histogram representing relative abundance of bacteria's in each cluster. C,D,E,F) Box-plots representing differences in various outcomes and indices, Mann-Whitney U test was used to asses statistical significance. "*" represents p-value ≤ 0.05 , "***" p-value ≤ 0.01 , "****" p-value ≤ 0.001 and "n.s" not significant 34

3.2 **Clusters based on the integrated biome(Unweighted):** A) Heat map of the patient similarity matrix Q_c obtained from the merging of bacteriome, fungome and virome plotted in log scale. x and y axis of the matrix represent the patients and each entry of the matrix represents the similarity between them. B) A plot representing the results of LEfSe on the clusters with x-axis and y-axis representing the effect size and species respectively. The plot only represents species that have an LDA score of ≥ 3 . C,D,E,F) Box plot representing the differences between various outcomes, Mann-Whitney U test is applied to calculate the statistical significance for difference in the clusters. "*" represents p-value ≤ 0.05 , "***" p-value ≤ 0.01 , "****" p-value ≤ 0.001 and "n.s" not significant 36

3.3 **Clusters based on the integrated biome(Weighted):** A) Heat map of the patient similarity matrix Q_c described from the merging of bacteriome, fungome and virome in a weighted fashion plotted in log scale. x and y axis of the matrix represent the patients and each entry of the matrix represents the similarity between them. B) A plot representing the results of LEfSe on the clusters with x-axis and y-axis representing the effect size and species respectively. The plot only represents species that have an LDA score of ≥ 3 and the top 20 species. C,D,E,F) Box plot representing the differences between various outcomes, Mann-Whitney U test is applied to calculate the statistical significance for difference in the clusters. "*" represents p-value ≤ 0.05 , "***" p-value ≤ 0.01 , "****" p-value ≤ 0.001 and "n.s" not significant 37

3.4 **Differential interactions:**The above graphs represent the microbial association network in cluster 1 and 2. The 'white' colour nodes represent the microbes, 'green' coloured edges represents positive interactions, 'red' the negative interactions and the depth of the colour represents the strength of the interaction. 39

- 3.5 **Interactome of Cluster 1:** C) Represents the interactome(i.e, interactions) of cluster 1. Highlighed yellow edges represent the interactions of *Haemophilus*, *Rothia* and *Streptococcus*. A,B,D) Represents the interaction of *Rothia*, *Streptococcus* and *Haemophilus* respectively with ‘green’ edges representing positive interactions and ‘red’ edges the negative interactions. The depth of the color represents the strength of the interactions. 40
- 3.6 **Interactome of Cluster 2:** C) The graph represents the interactome(i.e. interactions) of cluster 2. Highlighed yellow edges represent the interactions of *Cryptococcus*, *Veillonea*, *Prevotella* and *Haemophilus*. A,B,D,E) These graphs represent the interaction of *Cryptococcus*, *Veillonea*, *Prevotella* and *Haemophilus* with other microbes, respectively with ‘green’ edges representing positive interactions and ‘red’ edges the negative interactions. The depth of the color represents the strength of the interactions. 41
- 3.7 ***Pseudomonas* specific interaction :***Pseudomonas* specific interactions in cluster 1 and 2 are coloured in ‘red’ and ‘green’ with ‘red’ representing negative interaction and ‘green’ the positive interaction. The depth of the edge colour represents the strength of the interaction. 42

List of Tables

3.1	Clinical outcome comparison on clusters based on bacteriome, fungome and virome. Each value of the clusters column represents the median value of that outcome in that cluster. Medians of variables that were not statistically significant at an α level of 0.05 are not reported.	35
3.2	Clinical outcome comparison of clusters derived by integrating bacteriome, fungome and virome using SNF. Each value in the clusters column represents the median value of that outcome. Medians of variables that were not statistically significant at an α level of 0.05 are not reported.	35
3.3	Clinical outcome Comparison on the merged bacteriome, fungome and virome using weighted SNF. Each value of the column cluster 1 & 2 represents the median value of that outcome. Medians of variables that were not statistically significant at an α level of 0.05 are not reported.	38
3.4	Table showing total degree, in degree, out degree, betweenness centrality and stress centrality of the nodes from the microbial association networks of both the clusters from section3.3.	42

Chapter 1

Introduction

When we think about who we are and what defines us as species, our thoughts generally drift towards the human genome. The Human Genome Project was the first attempt to sequence the whole human genome with the goal of finding the genetic roots of disease and then developing suitable therapies. However the human genome, which is predetermined at the birth doesn't represent the whole genomic diversity present in our human body. Estimates of the gene content of microbes in our body are at 220 million, exceeding the $\sim 20,000$ human genes by at least a factor of 100 [Knight et al., 2017]. Even, microbial cells outnumber the human cells that we have in our body. The most detailed report up to date proposes that on average we are only 47% human by cell count [Sender et al., 2016]. This enormous amount of microbes living in our body can impact human biology in various ways.

1.1 Microbes and Microbiomes

Microbes or Micro-organisms are microscopic organisms which exist in both single cellular and multi-cellular fashion. Micro-organisms can be found everywhere in the environment, and they survive in virtually any condition from the poles, deserts, geysers, deep sea and even inside our body. The term microbiome is used to refer to the collection of genes within a community of microbes (including bacteria, fungi, virus, protists and bacteriophages) [Knight et al., 2017]. With the advent of next-generation sequencing, it is now possible to sequence the microbiome. Due to the tremendous diversity of microbes, new sequences of mi-

crobes are found in almost every new data-set and hence classifying the microbes sometimes exceeds the capacity of the reference database. To address this issue, researchers cluster sequences into operational taxonomic units (OTUs) often using a 97% sequence identity as a proxy for the species [Knight et al., 2017]. This adoption of OTU concept has allowed well-developed ecological theories to be applied in the context of the microbiomes.

In the last few years, microbiome research has helped us gained new insights into how microbes shape our human biology. Human microbiota are crucial for our body to maintain its homeostasis and disruption of this can lead to diseases. This is because the microbes provide a range of services to the host including bio-conversion of nutrients, protection against pathogenic microbes and production of essential resources. Thus the loss of beneficial microbiota and introduction of maladaptive functions by invading microbes can lead to diseases. Some prominent examples include dental caries and bacterial vaginosis. Obesity, inflammatory bowel disease, malnutrition and even disease such as Parkinsons, Autism, Asthma and depression are linked to the microbiome [Knight et al., 2017]. Most of the present microbiome research focuses on a single profile of the human microbiome, i.e. bacterial, fungal or viral profiles. All these studies have focused on each of these biological entities in isolation, even though bacteria, fungi and viruses coexist in the body as a community. Thus, it is essential to look at these biological components together in an integrated fashion. However, one of the primary reasons for the lack of multi-biomic research is the lack of methods to merge microbiome data-sets together. To address this issue, we propose a method in this thesis to integrate these microbiomes with a particular focus on the lung microbiome in the setting of bronchiectasis. We illustrate the advantages of combining different micro-biomic data-sets such as bacteriome, fungome and virome, with bronchiectasis, a chronic respiratory disease as a case study. However, this method is not limited to bronchiectasis and can be applied to merge any microbiome data-sets in general.

1.1.1 Ecological interactions versus random processes

A problem associated with microbiome data-sets obtained using sequencing techniques is that the observed relative abundance of a microbe may be artifactual and not represent the actual ecological interaction. In a paper published on the distribution of bird species across the islands near New Guinea [Martin Cody, 1975], Jared Diamond proposed a set of community assembly rules from the birds presence-absence data. In short, he stated that

competitive exclusion of bird species is the main force structuring the species composition of the island. This paper triggered a long discussion among ecologists about the importance of such rules in community formation or species composition of an area. Connor and Simberloff in their paper *The Assembly of Species Communities: Chance or Competition?* [Connor and Simberloff, 1979] criticized Diamond's rules suggesting that these rules didn't withstand significance testing in simulations. Discussion on this, whether observed presence-absence data/relative abundance data are due to ecological processes or due to random chance went on until Stephen Hubbell proposed the Unified neutral theory of Biodiversity [Hubbell, 2011]. The theory suggests that observed species distribution can be well explained by the random process of birth, death, and migration. Ecologists further used Hubbell's model as a Null model to test for the effect of ecological interactions. The Null has been confirmed in some studies but contradicted in others [Hubbell, 2011] [Faust and Raes, 2012]. This ambiguity suggests that both random processes and ecological interactions contribute to the species abundance distributions. Since microbiome data-sets resemble ecological data-sets where microbes replace species, microbiome data-sets also suffer from the above issue.

1.2 Chronic Respiratory Diseases

According to the World Health Organization (WHO), Chronic respiratory diseases (CRDs) are diseases of the airway and other structures of the lung [Organization, 2007]. Some of the most common CRDs include chronic pulmonary obstructive disease (COPD), asthma, bronchiectasis and pulmonary hypertension. Hundreds of millions of people from all ages suffer every day from preventable CRDs. According to the latest WHO estimates about 236 million people have asthma, a common lung disease among children, 64 million people have COPD and over 3 million people die each year from COPD which accounts for about 6% of all deaths worldwide [Organization, 2007]. On the other hand, many CRDs such as Bronchiectasis have no licensed treatments worldwide, and most treatments presently used are based on very little evidence [Polverino et al., 2017]. CRDs have significant adverse effects on the quality of life, morbidity and mortality of the affected individuals. Hence, more study is required to address these issues.

1.2.1 Microbiome in Chronic Respiratory Diseases

CRDs have various causative factors and mechanisms associated with disease progression which are presently not fully understood. One such factor is the lung/pulmonary microbiome, traditionally healthy lungs were thought to be sterile [Faner et al., 2017] but with the emergence of culture-independent sequencing techniques, it has been demonstrated that the healthy lung in fact contains an associated microbiome. There are a vast number of microbes including bacteria, fungi and viruses that inhabit the lung. These microbes co-exist and live as communities in the lung of healthy individuals as well patients. However, the composition and diversity of lung microbiota vary significantly between individuals, and they are mainly dictated by both environmental and genetic factors of the host [Rothschild et al., 2018]. On the contrary, in the absence of a major lifestyle change such as diet, disease onset or environment, the lung microbiome is relatively stable. The study of the pulmonary microbiome in the healthy has revealed that *Firmicutes*, *Bacteroidetes* *Proteobacteria* at phylum level and *Prevotella*, *Veillonella*, *Streptococcus* at the genus level are the most predominant microorganisms, with a minimal contribution from common pathogenic *Proteobacteria* including *Haemophilus* [Faner et al., 2017]. In patients with COPD *Proteobacteria*, *Bacteroidetes*, *Actinobacteria* and *Firmicutes*, with *Pseudomonas*, *Streptococcus*, *Prevotella* and *Haemophilus* are common [Faner et al., 2017]. In Cystic Fibrosis(CF) and Bronchiectasis (non-CF), culture-based studies have revealed *H. influenzae*, *P. aeruginosa*, *Moraxella catarrhalis*, *Staphylococcus aureus* and *Burkholderia cepacia* are more prevalent and predominant [Faner et al., 2017]. During exacerbation (the acute episode of progressive worsening of symptoms including shortness of breath and cough) of the patients with CRDs, the relative abundance of some genera increases whereas others don't change significantly. Also, exacerbation seems to be related not only to isolated microbes but also with the changes in microbiome composition as a whole [Faner et al., 2017]. Notwithstanding their significant contribution to the field, none of these studies had dissected relationships between microbes and studied them in detail. It is essential to study these interactions because clinical outcomes may not alone depend on individual micro-organisms but also the interactions between them. In this thesis, we also try to address this issue to a certain extent by using methods such as co-occurrence analysis.

1.3 Case study: The Microbiome in Bronchiectasis

Bronchiectasis is a chronic inflammatory respiratory disease associated with progressive, irreversible dilatation of the airway. This increasingly prevalent disease has the potential to cause a devastating illness which includes frequent respiratory infections, breathlessness, productive cough (a cough that produces mucus or phlegm) and occasional hemoptysis (a cough that involves blood or blood-stained mucus). Records of bronchiectasis can be traced back to the early 19th century to the writings of René Théophile Hyacinthe Laennec, which includes descriptions of patients with suppurative phlegm (Sputum) [Barker, 2002]. Phlegm/Sputum is the liquid secreted by the mucosal membrane of the respiratory system which is expelled by coughing. It is crucial to study bronchiectasis because most bronchiectasis is reported to be idiopathic (unknown cause) [Chalmers and Chotirmall, 2018] and it is a significant contributor to lung diseases globally with a substantial four-fold higher predominance in Asian populations [Seitz et al., 2012].

The definition of Bronchiectasis has remained morphological for over 50 years owing to the work of Reid. Reid in his paper states that bronchiectasis is a name given to any condition where dilatation of one or more bronchi is observed. Since this condition may result from various causes, the term “bronchiectasis” describes an anatomical abnormality rather than a single disease [REID, 1950]. Presently, the disease is diagnosed based on the pathological or radiographic appearance of airways. Hence study of this disease is difficult as it is often an end point of many CRDs. Lungs are the primary organ affected in Bronchiectasis patients since it damages the airways which makes it difficult for the mucus to leave the lungs. This accumulation of mucus attracts various microbes and leads to microbial infection. Hence, lung microbiota can act as a good proxy for the severity of disease and there have been many studies that focus on microorganisms, such as bacteria and fungi, that colonize the lung as the cause of exacerbation in bronchiectasis [Chandrasekaran et al., 2018][Mac Aogáin et al., 2017]. One such study “THE CAMEB STUDY” [Mac Aogáin et al., 2018] published by the Translational Research Laboratory, Singapore suggests that the mycobiome/fungome (fungal microbiome) is clinically important in Bronchiectasis. It is this study and its characterised clinical cohort that provides the source of data for the analysis performed on this thesis.

1.3.1 The CAMEB Study

The CAMEB study is an international multi-center cross-sectional study of the Cohort of Asian and Matched European Bronchiectasis (CAMEB) patients recruited in Asia (Singapore & Malaysia) and the United Kingdom (Scotland) (n=238) matched on age, gender and bronchiectasis severity. Bronchiectasis severity is measured using Bronchiectasis severity index(BSI) which is a composite measure of severity that takes into account factors such as Age, Body Mass Index, Lung function, Exacerbation, *Pseudomonas* colonisation, MMRC dyspnea score, etc.[Chalmers et al., 2014]. BSI is severity score used to predict patients with a future risk of mortality, hospitalisation and exacerbation. MMRC dyspnea score or the Modified Medical Research Council scale (MMRC score) is an assessment score of dyspnea(shortness of breath) widely used in CRDs. MMRC Score is often filled by the patient based on the degree of breathlessness. The MMRC and all components of the BSI were recorded for all CAMEB participants as previously described [Mac Aogáin et al., 2018].

The sputum mycobiome (Fungal Profile from sputum) was determined in these 238 patients by targeted amplicon shotgun sequencing of the 18S-28S rRNA internally-transcribed spacer regions ITS1 and ITS2 [Mac Aogáin et al., 2018]. In addition to this, the translational research laboratory has performed 16s rRNA sequencing to determine the bacterial composition of patient sputum, using the same samples [Mac Aogáin et al., 2017]. The above studies of bacteriome and fungome of the CAMEB cohort have revealed that specific fungal genera such as *Cryptococcus*, *Clavispora* and *Aspergillus* characterise the bronchiectasis mycobiome. Further, *Streptococcus*, *Haemophilus* and *Pseudomonas* species dominated the Bacterial profile of the Asian patients. To characterise the virome, quantitative Polymerase Chain Reaction(qPCR) on panel of 17 viruses was ran on 217 patients of the CAMEB cohort.

In summary, 134 bacteria, 405 fungi and 17 viruses were characterised across all the 217 patients of the CAMEB Cohort. Most of the published studies focus on a single profile of the lung microbiome, i.e. bacterial or fungal profiles while only very few have characterised the viral profile (the virome). All such studies have focused on each of these biological entities in isolation, even though bacteria, fungi and viruses coexist in the lung as a community. One of the primary reasons for this is the lack of methods to merge microbiome data-sets, and if there exists one, no one has shown if merging microbiome data-sets are useful. In this thesis, we use the data from CAMEB cohort to demonstrate the principle of Integrative Micro-Biomics and its advantages.

1.4 Summary

Microbes play a significant role in maintaining lung homeostasis. Changes in the composition are increasingly recognised in CRDs. Hence, a better understanding of the interaction between various microbes is important to predict the disease pathogenesis for future therapeutic intervention. However, most contemporary studies focus on the microbiome in a singular fashion, even though bacteria, fungi and virus co-exist in the lung. In this thesis, we propose a new method to integrate these microbiomes and demonstrate its advantages. There is a vast amount of evidence supporting the concept that abnormal regulation of cross-talk between microbes in different organs may play a critical role in disease. We also in this thesis, look into interactions between the microbes and attempt to find evidence for the above issue in the context of bronchiectasis.

Chapter 2

Methods

2.1 Introduction

In light of this issue, the lack of methods to merge microbiome data-sets we propose a method to integrate these microbiomes. Bacterial, Fungal and Viral similarity graphs with microbes as nodes and Bray- Curtis similarity as edge weights are constructed from the individual microbiome data-sets. These complete graphs are then fused using Similarity Network Fusion [Wang et al., 2014]. Cluster analysis was implemented on the fused network using spectral clustering. Resulting clusters were next assessed for significant differences in clinical outcomes. Further, an ensemble based approach using multiple similarity measures and regression techniques were used to create microbial association networks on these clusters to identify microbes that co-occur.

Definition 2.1.1. *A Microbiome data-set $D = [d_{i,j}]$ is a $m \times n$ matrix defined for a set of microbes $M = \{m_1, \dots, m_n\}$ on a set of patients $P = \{p_1, \dots, p_m\}$ with $d_{i,j}$ representing the relative abundance of microbe m_j in patient p_i .*

Property :

1. $\sum_{k=1}^n d_{i,k} = 100, \forall i \in (1, \dots, m)$, i.e. Sum of all microbes of M for any patient in P is equal to 100.

Example 2.1.1. *Consider the bacteria and fungi that were identified through 16s and ITS*

amplicon sequencing on the CAMEB Cohort then let M_1, M_2 denote the Bacteria and Fungi that were identified on patients P of the CAMEB Cohort then D_1, D_2 represent bacteriome and fungome of P i.e. the bacterial and fungal microbiome data-sets of patients P .

1. $D_1 \in \mathbb{R}^{217 \times 134}$

2. $D_2 \in \mathbb{R}^{217 \times 405}$

Example 2.1.2. Consider the Viruses that were identified through qPCR on the CAMEB cohort then let M_3 denote the viruses that were identified on the patients P of the CAMEB Cohort then $V = [v_{i,j}] \in \mathbb{R}^{217 \times 17}$ where $v_{i,j}$ is the number of genome copies per gram of sputum. This V represents the virome of P .

The virome V of Example 2.1.2 does not satisfy the property 1 of Microbiome data-set (definition 2.1.1).

2.2 Network inference using Similarities

Network inference is the processes of creating graphs or networks from data-sets. Bray-Curtis similarity, a statistic widely used in ecology to characterise compositional similarity between two sites, is employed for this purpose.

Most ecological data-sets are represented as occurrences or relative abundances. A typical ecological data-set $H = [h_{i,j}]$ is a $m \times n$ matrix defined for a set of species $S = \{s_1, \dots, s_n\}$ on different ecological sites $T = \{t_1, \dots, t_m\}$ with $h_{i,j}$ representing the relative abundance of species s_j in site t_i or number of occurrences/counts of species s_j at site t_i . Bray-Curtis similarity on the above data-set H is defined as:

Definition 2.2.1. *Bray-Curtis Similarity between two ecological sites $t_i \in T$ and $t_j \in T$ is defined as*

$$BC_{T,S}(i, j) = \frac{2C_{ij}}{S_i + S_j}$$

where C_{ij} is sum of lesser values for only those species common between both sites i.e., $C_{i,j} = \sum_k \min(h_{i,k}, h_{j,k}) \forall k \in (1, \dots, n)$ and $S_i = \sum_k h_{i,k} \forall k \in (1, \dots, n)$ the total number of species at site t_i .

Property :

1. $BC_{T,S}(i, j) : T \times T \rightarrow [0, 1]$ where i, j are index of ecological sites and $t_i, t_j \in T$.

Microbial data-set D is similar to ecological data-set H . This is because microbes M can be replaced by species S and patients P can be compared to sites T . Therefore, due to this resemblance of our microbiome data-sets D_1, D_2 and V to ecological data-sets, bray-curtis similarity was chosen as a similarity measure for microbiomes. Bray-curtis similarity $BC_{T,S}(i, j)$ in the context of microbiome datasets will measure how similar a patient p_i is to any other patient p_j based on the microbial composition of the patients. If both patients p_i and p_j share the same microbes with same counts/abundance then $BC_{P,M}(i, j) = 1$.

Let D_1, D_2 and V be as defined in Example 2.1.1 & 2.1.2 we need to construct similarity graphs $\mathcal{G}_1, \mathcal{G}_2$ and \mathcal{G}_3 for each of the above datasets. To achieve this, we set P as the vertex set for all the graphs and compute the edge weights between patients using bray-curtis similarity. Let W_1, W_2 and W_3 denote the weighted adjacency matrix of $\mathcal{G}_1, \mathcal{G}_2$ and \mathcal{G}_3 respectively, then for each $W_k \in \{W_1, W_2, W_3\}$ we define $W_k = [w_{i,j}]_{217 \times 217}$, where $w_{i,j} = BC_{P,M_k}(i, j)$ and $(i, j) \in [1, 217] \times [1, 217]$.

This results in three complete weighted graphs $\mathcal{G}_1, \mathcal{G}_2$ and \mathcal{G}_3 representing the similarity network between the patients P based on their bacterial, fungal and virome composition. This was implemented using the “vegan package” [Oksanen et al., 2018] in R.

2.3 Similarity Network Fusion

Bo Wang in his paper introduces Similarity Network Fusion(SNF) [Wang et al., 2014] as a novel method to merge two or more similarity networks. This method is described in detail below:

Let $\{\mathcal{G}_n\}_m$ denote a set of similarity graphs defined on the same vertex set V and $\{W_n\}_m$ a set of its weighted adjacency matrix as described in section2.2. We describe below how we merge these similarity graphs into one single fused graph G_f . For each W_n in $\{W_n\}_m$ we decompose it into Q_n and S_n using

The Q_n matrix represents the similarity of a vertex to all other vertices and it satisfies $\sum_j [Q_n](i, j) = 1 \forall i \in [1, p]$ where p is the cardinality of V . The S_i matrix represents the similarity of a patient to its “ k ” most similar patients. These matrices are then fused in an iterative fashion for each n over t iterations using

$$Q_n^{(\nu)} = S_n \times \frac{\sum_{i \neq n} Q_i^{\nu-1}}{m-1} \times (S_n)^T, \quad n \in [1, m] \quad (2.3)$$

where, m is the cardinality of $\{\mathcal{G}_n\}_m$, ν is the iter and t is a user specified hyper-parameter. The overall fused Q_c matrix is calculated from each n by taking the average of these $\{Q_n\}_m$ matrices

$$Q_c = \frac{\sum_{n \in [1, m]} Q_n^{\nu=t}}{m} \quad (2.4)$$

This Q_c is the weighted adjacency matrix of the fused graph \mathcal{G}_f and V is its vertex set. This graph represents a holistic view of all $\{\mathcal{G}_n\}_m$ i.e., higher weight of an edge in the fused network \mathcal{G}_f implies that the edge has higher weights in many of these graphs $\{\mathcal{G}_n\}_m$.

The above described method was implemented using the “SNFtool” package [Wang et al., 2018] in R.

2.4 Weighted Similarity Network Fusion

SNF described in section treats all the similarity networks with equal weights/importance during merging. Consider $\{\mathcal{G}_n\}_m, V, \{\mathcal{W}_n\}_m$ as defined in section 2.4 and let $\Omega = \{\omega_1, \dots, \omega_m\}$ be the non-zero weights or importance of each of $\{\mathcal{G}_n\}_m$ these similarity graphs. In order to introduce weights in SNF we modify equation 2.3

$$Q_n^{(\nu)} = S_n \times \frac{\sum_{i \neq n} \omega_i \times Q_i^{\nu-1}}{\sum_{i \neq n} \omega_i} \times (S_n)^T, \quad n \in [1, m] \quad (2.5)$$

Q_c is then computed and treated as weighted adjacency matrix of the fused graph \mathcal{G}_f . The codes for weighted SNF was written in R and is available at <https://github.com/Jayanth-kumar5566/Integrative-Microbiomics>

We use the both SNF and Weighted SNF to merge $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$ where $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ are the

similarity graphs of bacteriome, fungome and virome of the CAMEB patients respectively, as described in section 2.2.

2.5 Cluster Analysis

The final merged network \mathcal{G}_f is a network of similarities between vertices $v_i \in V$. In order to identify subgroups or subsets of these vertices, we apply clustering methods and find set of vertices $\{v_i\}$ such that vertices in the same set are more similar to each other than to those in the other set. Similarity-based clustering is a clustering technique that uses similarity measure between data points to identify clusters in the data. Spectral-Clustering is similarity based clustering algorithm that can be implemented on any similarity graph \mathcal{G} to identify clusters in the vertex set V of the graph \mathcal{G} . It uses eigenvalues of the Laplacian matrix computed using the weighted adjacency matrix(similarity matrix) of \mathcal{G} to perform dimensionality reduction before clustering them in lower dimensions using k-means [Macqueen, 1967]. The Laplacian matrix L of a similarity matrix W is defined as $L := D - W$ where D is a diagonal matrix with $D_{i,i} = \sum_j W_{ij}$. Let $\lambda_1 \geq \lambda_2 \geq \dots \lambda_{\mathcal{K}}$ be the eigen first \mathcal{K} eigen values of L and $U = [u_i]_{i \in [1, \mathcal{K}]}$ be the corresponding eigen vectors. This U corresponds to W in the reduced dimension. k-means [Macqueen, 1967], a clustering algorithm is applied on U to cluster these vertices(given as rows of U) into \mathcal{K} clusters. The optimal number of clusters \mathcal{K} was calculated using the Eigen gap method, i.e. \mathcal{K} is given by the value that maximises the eigengap of the laplacian matrix L (difference between consecutive eigenvalues) after ordering them in an ascending fashion. This method was implemented using the ‘‘SNFtool’’ package [Wang et al., 2018] in R.

Further, to assess the quality of these clusters we calculate a silhouette width/score [Rousseeuw, 1987] for each cluster. Silhouette width of a cluster is the average of all silhouette values of its objects. The silhouette value measures how similar is an object to its own cluster (cohesion) compared to other clusters (separation). This measure ranges from -1 to $+1$, with a high value indicating that the object lie well within their own cluster compared to other clusters. We define the silhouette value as below:

Definition 2.5.1. *The Silhouette value of an object ‘‘i’’ is defined as:*

$$S(i) = \frac{\text{inter Cluster affinity}(i) - \text{intra cluster affinity}(i)}{\max\{\text{inter cluster affinity}(i), \text{intra cluster affinity}(i)\}} \quad (2.6)$$

where,

- *inter cluster affinity*(i) : is the mean of similarity between i and all other data points within the same cluster
- *intra cluster affinity*(i) : is the lowest mean similarity between i and all other data points of which i is not a member of.

The value of the hyper-parameter k which controls the number of nearest neighbours as described in section 2.4 was tuned to an optimal value using the average silhouette width/score of all the resulting clusters i.e., average silhouette width was calculated for all possible values of k and the k yielding the maximum average silhouette width was chosen. This is implemented in Python, codes are available at https://github.com/Jayanth-kumar5566/Integrative-Microbiomics/blob/master/Weighted_SNF/All_biomes/snf.R

2.6 Cluster Characterisation

The resulting clusters obtained from the similarity graphs of CAMEB patients, on different microbiomes and their combination were mapped back to the clinical outcomes of the CAMEB patients. For continuous variables such as BMI, BSI, Number of exacerbation, Age etc. the Kruskal Wallis test, a non-parametric test for assessing whether two or more independent samples come from the same distribution was implemented to check if the distribution of continuous clinical variables from the two or more clusters came from the same population distribution at an α level of 0.05. Further, if the number of clusters is more than two, dunn's test, a post-hoc test for multiple comparison was implemented with Benjamini and Hochberg, False Discovery Rate(FDR) correction [Benjamini and Hochberg, 1997] to account for multiple testing.

In the case of categorical variables such as Inhaled corticosteroids, positive sputum culture, gender etc., a contingency table is computed for each variable between the clusters. Further, Pearsons Chi-Squared test, a test for independence (i.e., H_0 : The joint distribution of the cell counts in a 2-dimensional contingency table is the product of the row and column marginals. H_a : contrary to null) was implemented on this contingency table. The p-values

for this test was computed by a Monte Carlo method [Hope, 1968] with 2000 replicates i.e., by simulating random sampling from the set of all contingency tables with given marginals. The above described test was applied to check if there was any significant relationship between the clusters and categorical variables at an α level of 0.05.

Also, Shannon diversity index H , an ecological diversity measure was calculated for each patient. This measure reflects the number of different types of microbes present in that patient and simultaneously accounts for the evenness in distribution among those microbes. Shannon diversity index H is defined as

Definition 2.6.1. $H := \sum_1^s p_i \ln(p_i)$

where, p_i is the proportion ($\frac{n}{N}$) of individuals of one particular species found (n) divided by the total number of individuals found (N). A distribution $\{H\}$ of H was calculated on each cluster using the values of individual patients. Further a Mann-Whitney U test with Dunn's test as post-hoc was implemented to check for statistical difference of $\{H\}$ between the clusters.

The methods described in this section was implemented in R and can be found at <https://github.com/Jayanth-kumar5566/Integrative-Microbiomics>

2.6.1 Linear Discriminant Analysis Effect Size (LEfSe)

We sought to find the microbes that drive this clustering, in order to do this we used Linear Discriminant Analysis(LDA) Effect Size (LEfSe)[Segata et al., 2011] available as a web-tool on Galaxy [Afgan et al., 2018]. The web tool developed by the Huttenhower group attempts to find species that are most influential in differentiating the given clusters. Firstly, the tool selects microbes by implementing a Kruskal-Wallis(KW) test to check for significant differences in microbial abundances between the clusters at a p-value of < 0.05 . The factorial KW rank sum test is a non-parametric test that checks whether the samples are drawn from the same population distribution. Secondly, an Linear Discriminant Analysis(LDA) model is built with clusters as the dependent variable and the selected microbes from the first part as independent variables. LDA is a linear classifier that tries to find a linear combination of features (i.e. Microbes in our case) that best separates two or more classes(i.e. Clusters in our case). Thirdly, this LDA model is used to estimate the effect sizes of microbes. These

effect sizes of microbes would represent species that are most influential in differentiating between the clusters.

2.7 Co-occurrence Analysis

To address this issue of microbes being not independent and to identify these inter-plays. We need to construct microbial association networks (i.e., graphs with nodes as microbes and edges representing a measure of interaction between them) which is a classical problem of network inference in computer science, constructing association networks from datasets. There are two main groups of approaches to address this issue, with one group capturing **Pairwise relationships** and the other capturing **Complex relationships** [Faust and Raes, 2012]. Here we propose an ensemble-based approach encompassing both these approaches with significance testing. Karoline Faust in his paper [Faust et al., 2012] introduces “Reboot”, a novel bootstrap and re-normalisation approach to assess the degree of association present purely due to ecological interactions alone, addressing the problem stated in section 1.1.1. This paper also introduces an ensemble approach with multiple similarity measures and generalised boosted linear models, for network inference. We adopt this framework with some improvements, which is stated in section 4.3.3.

2.7.1 Similarity-based Network Analysis

Similarity-based Network analysis belongs to the group that captures pairwise relationships. The Network inference technique described in section 2.2 belongs to this method of similarity based network analysis and this technique can be generalised to any arbitrary similarity measure \mathcal{S} and any microbiome data-set.

Consider an arbitrary similarity measure \mathcal{S} and an arbitrary microbiome data-set D as defined by definition 2.1.1 then, $\mathcal{S} : M \times M \rightarrow \mathbb{R}^+$ where \mathbb{R}^+ denotes the set of all positive rationals, the codomain of \mathcal{S} . For microbe m_i and m_j that are dissimilar, $\mathcal{S}(m_i, m_j) = 0$

We construct a complete weighted graph \mathcal{N} on D with M as the vertex set and the edge set E computed using the similarity measure \mathcal{S} i.e., $\forall e_{m_i, m_j} \in E, e_{m_i, m_j} = \mathcal{S}(m_i, m_j)$, calculated over multiple patients. Since the network/graph is constructed using the similarity

measure \mathcal{S} that considers a pair of microbes (m_i, m_j) at a time to compute the edge weights e_{m_i, m_j} this approach falls under the class that captures pairwise relationships.

The difference between the above described method and the method described in section 2.2 is that one method considers microbes M as vertex set and the other considers patients P as vertex set to create similarity graphs \mathcal{N} and \mathcal{G} respectively, where \mathcal{G} is as described in section 2.2.

Being one of the simpler method to implement, it does not capture more complex form of interactions in which one microbe depends on (or is influenced by) multiple other microbes and it also suffers from the problem of spurious edges. For instance, let m_1, m_2 depend on each other and m_3 depend on m_2 then $\mathcal{S}(m_1, m_3) > 0$ even though m_1 and m_3 don't interact with each other directly but rather through m_2 , where \mathcal{S} is a similarity measure such as correlation.

2.7.2 Regression-based Network Analysis

Regression-based network analysis in contrast captures complex relationships. Consider an arbitrary regression model $Y \approx f(X, \beta)$ on data-set D_s as defined in section 2.2, where $Y \in M_s$, $X \subseteq M_s \setminus Y$ and β is the coefficient matrix that quantifies the strength of relation between Y and $m_i \in X$. f is the function that relates X and Y . If f is linear, then the regression model is called a linear model.

We construct a weighted graph \mathcal{N}_s on D_s with M_s as the vertex set and the edge set E_s computed using the regression model. $\forall e_{m_i, m_j} \in E_s$

$$e_{m_i, m_j} = \begin{cases} [\beta_{m_i}]_{m_j} & \text{if } m_j \in X \\ 0 & \text{otherwise} \end{cases}$$

where, $[\beta_{m_i}]_{m_j}$ is an element of the coefficient matrix $[\beta_{m_i}]$ corresponding to the m_j^{th} microbe and $[\beta_{m_i}]$ is the coefficient matrix obtained from the regression model, $m_i \approx f(X, \beta)$

Unlike the Similarity-based analysis, this method captures both pairwise and complex interactions. However, implementation of this method is difficult as it's computationally intensive.

2.7.3 Ensemble approach for network inference

In similarity based network inference, a single similarity measure such as Pearson's or Spearman's correlation is employed to identify relationships/edges. The issue with this is, the microbial association networks are dependent on the choice of similarity measure. Similarity measures are only capable of capturing certain relationships such as linearity, monotonicity etc. Consequently, there is a need for multiple similarity measures to identify relationships more reliably. Further, due to the disadvantage of the similarity based network inference, a regression-based network inference is also required.

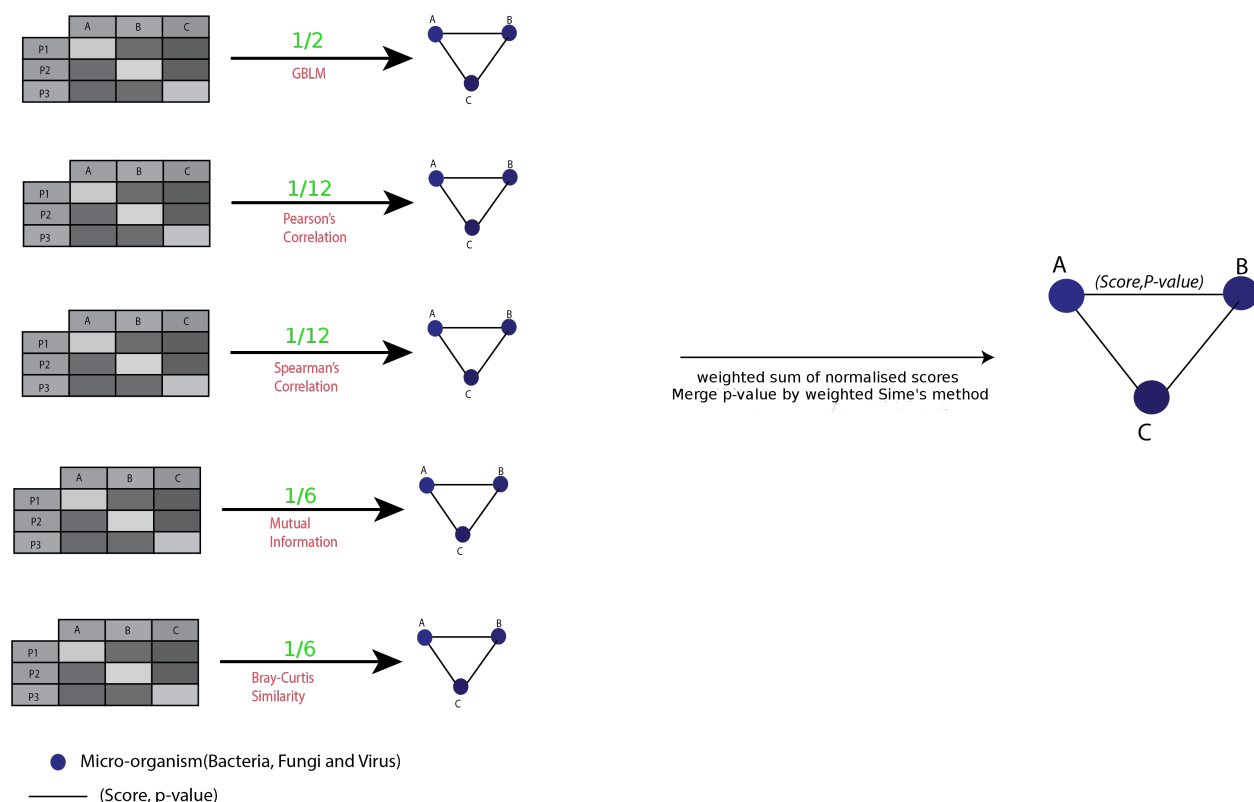


Figure 2.2: **An ensemble approach of Network Inference:** Network of microbes are built on microbes from all the biomes. Edge weights along with their statistical significance, are ascertained using four different similarity measures and GBLM(Gradient Boosting with Component-wise linear models), resulting in five different microbial networks one based on each measure. A merged microbial network is derived from the 5 networks with appropriate weighting(green colored text). P-values are combined using weighted Sime's test and edge weights are a weighted sum is taken after proper standardisation.

Therefore, we use the following measures,

Pearson's Correlation

Measure of linearity between two random variables X and Y

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where cov , is the co-variance and σ_X is the standard deviation of X . This was implemented in R using the “cor” function.

Spearman’s Correlation

Measure of Monotonicity between two random variables X and Y

$$\rho_{r(X),r(Y)} = \frac{\text{cov}(r(X), r(Y))}{\sigma_{r(X)} \sigma_{r(Y)}}$$

where $r(X)$ and $r(Y)$ denote the ranked random variables of X and Y . This was implemented in R using the “cor” function.

Mutual Information

Measures stochastic dependence or mutual dependence between two random variables X and Y

$$I(X, Y) = E \left(\log \frac{p(X, Y)}{p(X)p(Y)} \right)$$

where, $E(Z)$ denotes the expectation of Z , $p(X, Y)$ denotes the joint density and $P(X)$ and $P(Y)$ denote their respective marginal densities of X and Y . This was implemented using the “minet” package [Meyer et al., 2008] in R using Miller-Madow corrected estimator to estimate Mutual information and the ARCANe algorithm to build the network.

Bray-Curtis Similarity

Measures Compositional similarity between two sites S_i and S_j as defined in section 2.2

$$BC_{i,j} = \frac{2C_{ij}}{S_i + S_j}$$

This was implemented in R using the “vegan” package [Oksanen et al., 2018]

Gradient Boosting with Component-wise Linear Models -(GBLM)

This method uses a regression-based network inference approach to identify complex relationships. A Linear Model(LM), captures the relationship between outcome variable y and predictor variable $\mathbf{x} := (x_1, x_2, \dots, x_p)$ to get an “optimal” prediction of y

given x . This is accomplished by minimising a loss function $\rho(y, f) \in \mathbb{R}$ over prediction function $f(\mathbf{x})$ which is linear. In the framework of gradient boosting, the aim is to estimate the optimal prediction function f^* which is defined as,

$$f^* := \operatorname{argmin}_f \mathbf{E}_{Y, \mathbf{X}}[\rho(y, f(\mathbf{x}))]$$

“Component-wise gradient boosting” [Hofner et al., 2014] algorithm is implemented to find the minimum of the above expectation over f . Component-wise linear models of the predictors \mathbf{x} are used, in this algorithm as base-learners. Hence, named Gradient Boosting with Component-wise Linear Models (GBLM). This method was implemented in R using the “glmboost” function of “mboost” package [Hothorn et al., 2018] with $\nu = 0.05$ and `mstop` was tuned for the optimal value using 10-fold cross validation.

All the above five approaches are used to create five different microbial association networks as described in section 2.8. Further, these networks are then merged together by a weighted sum of the edge weights after appropriate standardisation. Standardisation is achieved by scaling the edge weights for each approach to a percentage with respect to the maximum edge weight of that network $score = \frac{score}{\max(score)} \times 100$ where $score$ is the edge weights. $\frac{1}{12}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{2}$ were used as weights for Pearson’s, Spearman’s, Mutual Information, Bray-Curtis similarity and GBLM respectively. These weights were assigned because GBLM is the only measure from the ensemble that captures complex interaction and spearman, pearson’s both capture correlation. Additionally, FDR corrected p-values of the edges were merged using weighted Sime’s method [Benjamini and Hochberg, 1997] using the above weights.

Let $H_{(0,i)}$ with $i \in (1, n)$ denote n hypothesis, p_i denoting their p-values and w_i denote their weights. Under $H_{(0,i)}$, $p_i \sim U(0, 1)$, we want to test for the global null $H_0 = \cap_i H_{(0,i)}$. Weighted Sime’s test [Benjamini and Hochberg, 1997] does this by reordering p-values $p_{(1)} \leq p_{(2)} \leq \dots p_{(n)}$ to calculate the Sime’s statistic $T_n = \min_i \{p_{(i)} \frac{n}{\sum_{k=1}^i w_{(k)}}\}$, where $w_{(i)}$ is the weight corresponds to the p-value $p_{(i)}$. Under H_0 and independence of p_i Sime’s test rejects H_0 if $T_n < \alpha$.

2.8 Network Construction

Consider the bacteriome(D_1), fungome(D_2) and virome(V) from example 2.1.1 & 2.1.2 these datasets were concatenated on patients P i.e. Let p, q and r be the number of microbes in M_1, M_2 and M_3 respectively. We define the concatenated microbiome data-set $\mathcal{D} = [d_{ij}] \in \mathbb{R}^{217 \times (p+q+r)}$ with patients $P = \{p_1, p_2, \dots, p_{217}\}$ and microbes $M = \cup(M_1, M_2, M_3)$. Where, $d_{i,j}$ represents the relative abundance of microbe m_j in patient p_i . Virome(V) was renormalised to satisfy property described in section 2.2.1, before it was concatenated with others and renormalised to calculate relative abundance.

2.8.1 Similarity-based Network construction

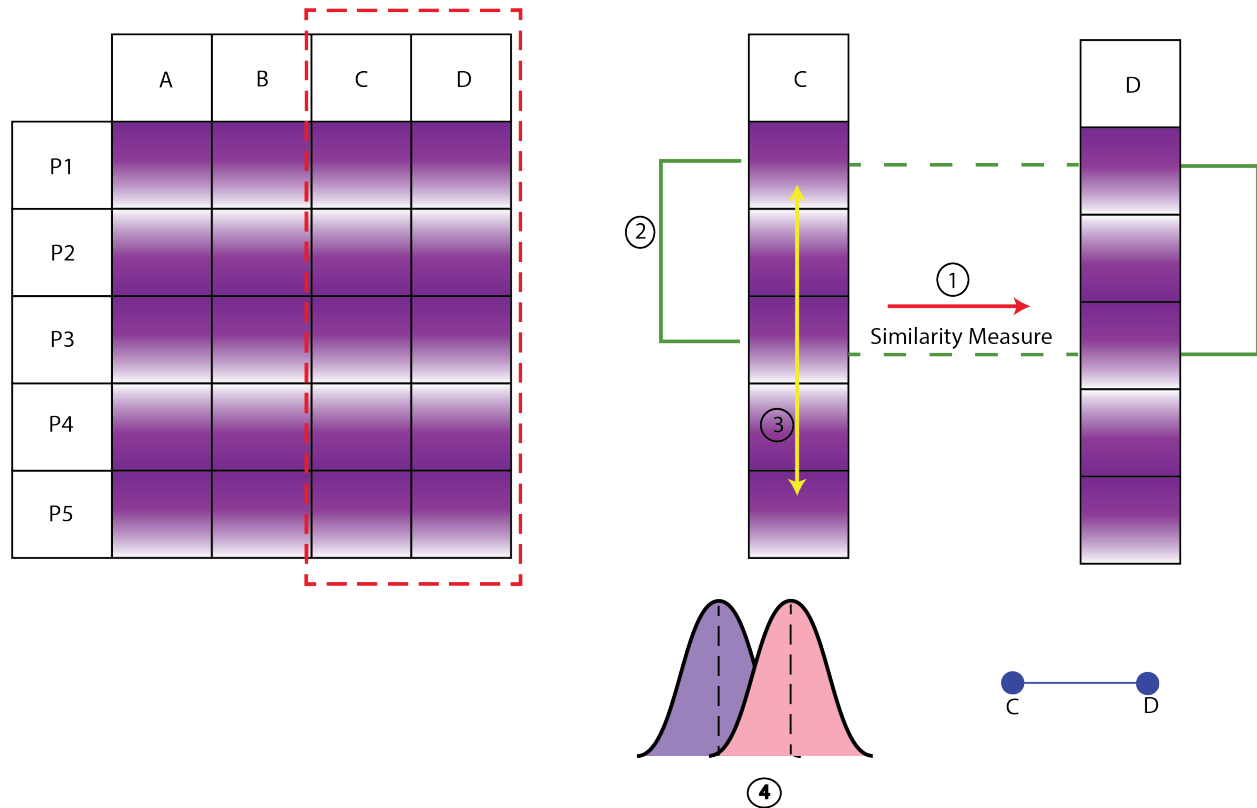


Figure 2.3: **Network construction from similarities:** 1) Similarity measure is calculated between every pair of species over all patients. 2) Bootstrap distribution of for each similarity measure is constructed. 3) Reboot [Permutation and re-normalisation] distribution for each similarity measure is built. 4) Distributional difference between bootstrap and reboot is assessed using Mann-Whitney U test and a p-value is assigned for each edge/similarity measure.

We construct the microbial association network for each similarity measure $S \in (\text{Pear-$

son's, Spearman's, Mutual Information, Bray-Curtis) as follows:-

1. A similarity graph \mathcal{N} was constructed on \mathcal{D} using the method previously described in section 2.7.1.
2. A bootstrap distribution of similarities for each edge i.e., $\forall e_{m_i, m_j} \in E$ was calculated over 100 iterations, where E is the edge set of \mathcal{N} . Bootstrapping is a type of re-sampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample. $S(m_i, m_j)$ was assessed on these bootstrap samples to get the bootstrap distribution.
3. A 'reboot' distribution of similarities for each edge is calculated over 100 permutations i.e., $\forall e_{m_i, m_j} \in E$, abundance values of microbe m_i is permuted across all fixed patients P followed by re-normalisation of the rows and computation of $S(m_i, m_j)$. This process is repeated over 100 iterations, resulting in the 'reboot' distribution of edge weights between each m_i and m_j
4. Mann-Whitney U test, a non-parametric test to assess difference in distribution, is applied between the 'reboot' and the bootstrap distribution for each edge weight $e_{m_i, m_j} \in E$ and a p-value is calculated for each edge.
5. P-values of all edges were FDR(False Discovery Rate) corrected to account for multiple testing and the resulting p-value of the test is appended with the edge weight representing the statistical significance of the edge.

2.8.2 Network construction using GBLM

1. We create a Network \mathcal{N} on \mathcal{D} using the method previously described in section 2.7.2 with $M_s = M$ and $X = \hat{M} \setminus Y$, where $\hat{M} = \{m_i \in M \mid \text{abs}(S(m_i, Y)) > 0.05 \ \& \ \text{abs}(S(m_i, Y)) \neq 1\}$ where, S is spearman correlation.
2. The models for which the R^2 , coefficient of determination < 0.5 was dropped from further analysis and edge weights were set to 0.
3. A bootstrap distribution of the edge weights for each edge is constructed by repeated fitting of a GBLM model over 100 bootstraps samples.

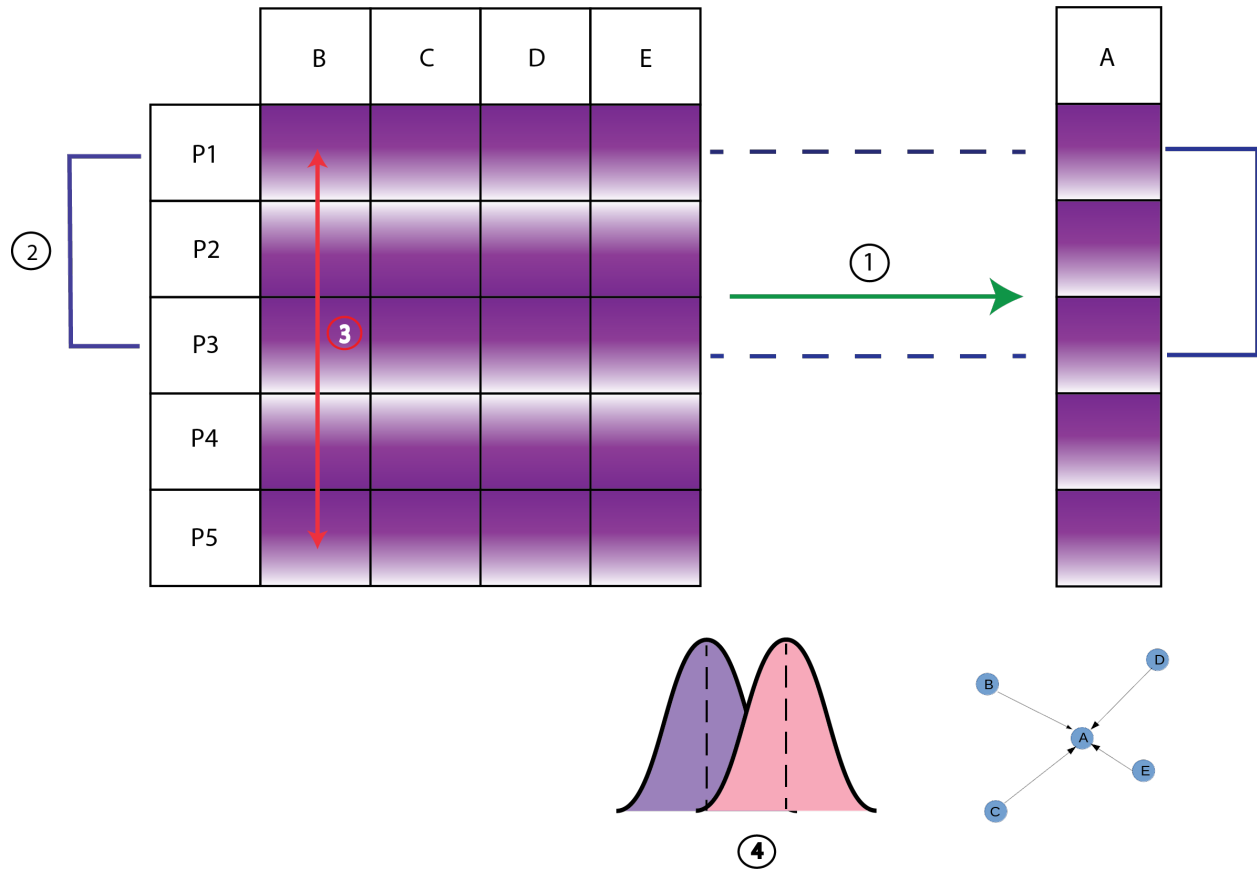


Figure 2.4: **GBLM for network construction:** 1)GBLM model was implemented to predict each microbe from all other microbes. The coefficients of the microbes served as edge weights. 2)Bootstrap distribution of these coefficients were created. 3) A reboot(permutation-renormalisation) distribution of these coefficients were created. 4)Statistical significance(p-values) of the coefficients were assessed using Mann-Whitney U test between bootstrap and Reboot distribution.

4. A ‘reboot’ distribution of edge weights for each edge is constructed by permuting the abundance values of m_i across the fixed patients P . Followed by, re-normalisation across rows. This processes is repeated over 100 iterations to get the ‘reboot’ distribution of the edge weights between each microbes.
5. Mann-Whitney U test, a non-parametric test to assess difference in distribution, is applied between the ‘reboot’ and the bootstrap distribution for each edge weight $e_{m_i, m_j} \in E$ and a p-value is calculated for each edge.
6. P-values of all edges were FDR(False Discovery Rate) corrected to account for multiple testing and the resulting p-value of the test is appended with the edge weight representing the statistical significance of the edge.

All the described methods in this section are implemented in R and Python and is

2.9 Network Analysis using Cytoscape

Let \mathcal{N}_f denote the final merged network from section 2.7.3 with microbes as vertex and edges $e_{i,j} = (w_{(i,j)}, p_{(i,j)})$ where $w_{(i,j)}$ represents the merged scores or edge weights between (m_i, m_j) and $p_{(i,j)}$ represents the merged p-value of that edge. Edge weights $w_{(i,j)}$ of edges with $p_{(i,j)} < 0.001$ were set to zero. We reconstructed \mathcal{N}_f using only statistically significant edge weights and graph theoretical measures such as number of nodes, number of edges, average number of neighbours and characteristic path length were computed. All the above analysis was done in Cytoscape[Shannon et al., 2003] and python. Cytoscape was also used to calculate the following:

Degree: Degree of a vertex v_i is the total number of edges incident on that vertex.

Stress Centrality: Stress centrality $C_s(v)$ of a vertex v is defined as

$$C_s(v) = \sum_{p \neq q \neq v \in V} \rho_{p,q}(v)$$

where $\rho_{p,q}(v)$ is the number of shortest paths from p to q passing through v .

Betweenness Centrality: Betweenness centrality $C_b(v)$ of a vertex v is defined as

$$C_b(v) = \sum_{p \neq q \neq v \in V} \frac{\rho_{p,q}(v)}{\rho_{p,q}}$$

where p and q are vertices that lie in a different network from v , $\rho_{p,q}$ is the total number of shortest paths from p to q and $\rho_{p,q}(v)$ is the number of shortest paths from p to q that pass through v

The cytoscape and python codes can be found at <https://github.com/Jayanth-kumar5566/Integrative-Microbiomics/tree/master/Co-occurrence>

Chapter 3

Results

3.1 Bacteriome clusters identify high risk patients

Spectral clustering on the bacterial microbiome with bacteria that were present in at least 5% of the CAMEB patients(n=10) revealed 3 clusters. The average silhouette value of these clusters is 0.386 indicating moderate clustering [Table 3.1]. Figure 3.1A further validates the presence of 3 clusters visually using a PCoA(Principal coordinates analysis) plot. Figure 3.1B reveals that cluster 1 is a *Pseudomonas* dominant, cluster 2 is *Streptococcus* dominant and Cluster 3 as *Haemophilus* dominant. The *Haemophilus* dominant cluster has a relatively low BSI (median = 9) compared to the *Pseudomonas* dominant cluster (median = 12) but have a relatively higher MMRC score and *Aspergillus terreus* conidial burden with a relatively low number of hospitalisation (median = 0) [Table 3.1]. Even though the *Pseudomonas* dominant cluster have relatively low MMRC score (symptomatic score) than the *Haemophilus* dominant cluster the *Pseudomonas* dominant cluster seem to associate with greater number of hospitalisations. We also observe that the median Body Mass Index(BMI) is relatively higher in *Haemophilus* dominated cluster than *Pseudomonas* dominant cluster. On the other hand, the *Streptococcus* dominant cluster seem to be clinically more favourable than the other two clusters with less hospitalisations and exacerbations.

The fungome and virome clusters did not show any difference in clinical measures other than *Aspergillus terreus* conidial burden in fungome and “Asian or European” in both fungome and virome [Table 3.1].

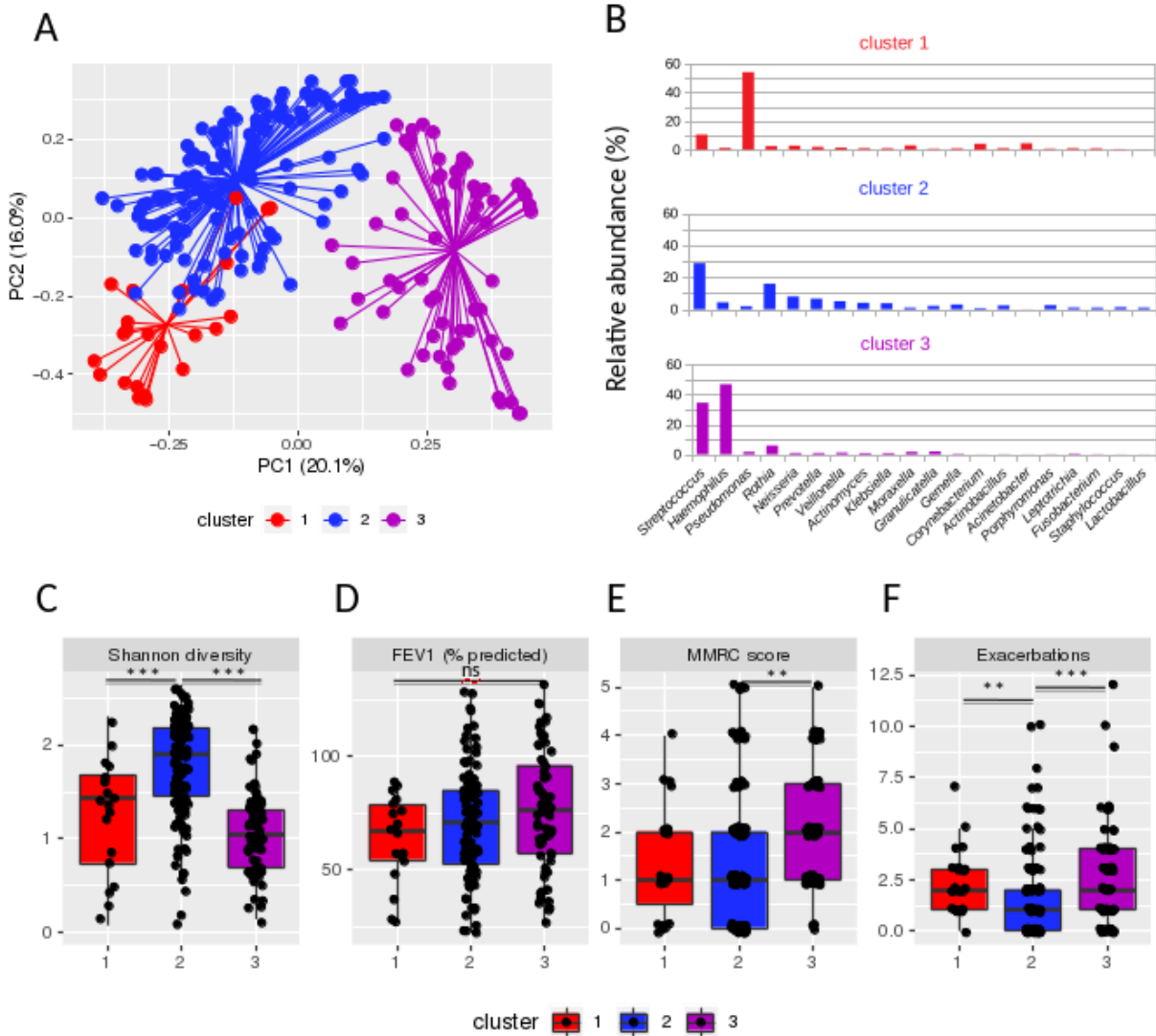


Figure 3.1: **Spectral clustering on the bacteriome of CAMEB patients:** A) PCoA(Principal Coordinate Analysis) plot of the bacteriome clusters using bray-curtis dissimilarity with x and y axis representing the first and second principal axis respectively. B) A histogram representing relative abundance of bacteria's in each cluster. C,D,E,F) Box-plots representing differences in various outcomes and indices, Mann-Whitney U test was used to assess statistical significance. “*” represents p-value ≤ 0.05 , “**” p-value ≤ 0.01 , “***” p-value ≤ 0.001 and “n.s” not significant

3.2 Integrative Microbiomics using Similarity Network Fusion (Unweighted)

Spectral clustering on the integrated bacterial, fungal and viral datasets using SNF with $t = 20$ and $k = 9$ (tuned) was implemented using the microbes that were present in at least

Slno	Outcomes	Bacteria (Silhouette: 0.386)				Fungi (Silhouette: 0.808)			Virus (Silhouette: 0.998)				
		Cluster 1	Cluster 2	Cluster 3	P-value	Cluster 1	Cluster 2	P-value	Cluster 1	Cluster 2	Cluster 3	Cluster 4	P-value
	Number of patients	23	125	69		190	17		121	80	14	2	
1	BSI	12	9	9	0.0039								
2	A.terreus conodial burden	231.03	1499.1	3096.2	0.0326	2205.55	810.4	0.035					
3	MMRC Score	1	1	2	0.0036								
4	No of Hospitalisation bf study	1	0	0	0.0144								
5	No of Exacerbation bf study	2	1	2	0.00021								
6	BMI	20.69	20	25.76	4.39E-05								
7	Asian/European(Asian,European)	(73.91%,26.09%)	(72%,28%)	(18.84%,81.16%)	0.00049	(52.63%,47.37%)	(25.92%,74.07%)	0.0274	(69.42%,30.58%)	(35%,65%)	(57.14%,42.85%)		0.00049

Table 3.1: Clinical outcome comparison on clusters based on bacteriome, fungome and virome. Each value of the clusters column represents the median value of that outcome in that cluster. Medians of variables that were not statistically significant at an α level of 0.05 are not reported.

5% of the CAMEB patients(n=10). This revealed 3 clusters with an average silhouette score of 0.799. Figure 3.2A illustrates these three clusters by a heatmap of the merged pairwise similarity matrix of the patients. Figure 3.2B represents the species that are most influential in differentiating the clusters. *Fuscoporia* in cluster 3, *Candida* in cluster 2 and *Para Influenza Virus 3 (PIV3)*, *Aspergillus* and *Trechispora* in cluster 1 are the microbes that most differentiate the clusters. Upon assessing for clinical outcome for these clusters we find a cluster that is differentiated by *Candida* of relatively high-risk patients with a higher median number of exacerbation(median=2) and MMRC score compared to the rest [Table 3.2]. However, we find that the p-value to identify the median number of exacerbation is increased (i.e. decrease in precision) from 0.00021, bacteriome alone to 0.03964 merged biome.

Slno	Outcomes	Cluster 1	Cluster 2	Cluster 3	P-value
	Number of patients	115	88	14	
1	BSI				
2	A.terreus conodial burden	1254.4	2713.3	1409.3	0.01599
3	MMRC Score	1	2	1.5	0.0030
4	No of Hospitalisation bf study				
5	No of Exacerbation bf study	1	2	1	0.03964
6	BMI				
7	Asian/European(Asian,European)	(75.65%,24.34%)	(28.41%,71.59%)	(57.14%,42.86%)	0.00049

Table 3.2: Clinical outcome comparison of clusters derived by integrating bacteriome, fungome and virome using SNF. Each value in the clusters column represents the median value of that outcome. Medians of variables that were not statistically significant at an α level of 0.05 are not reported.

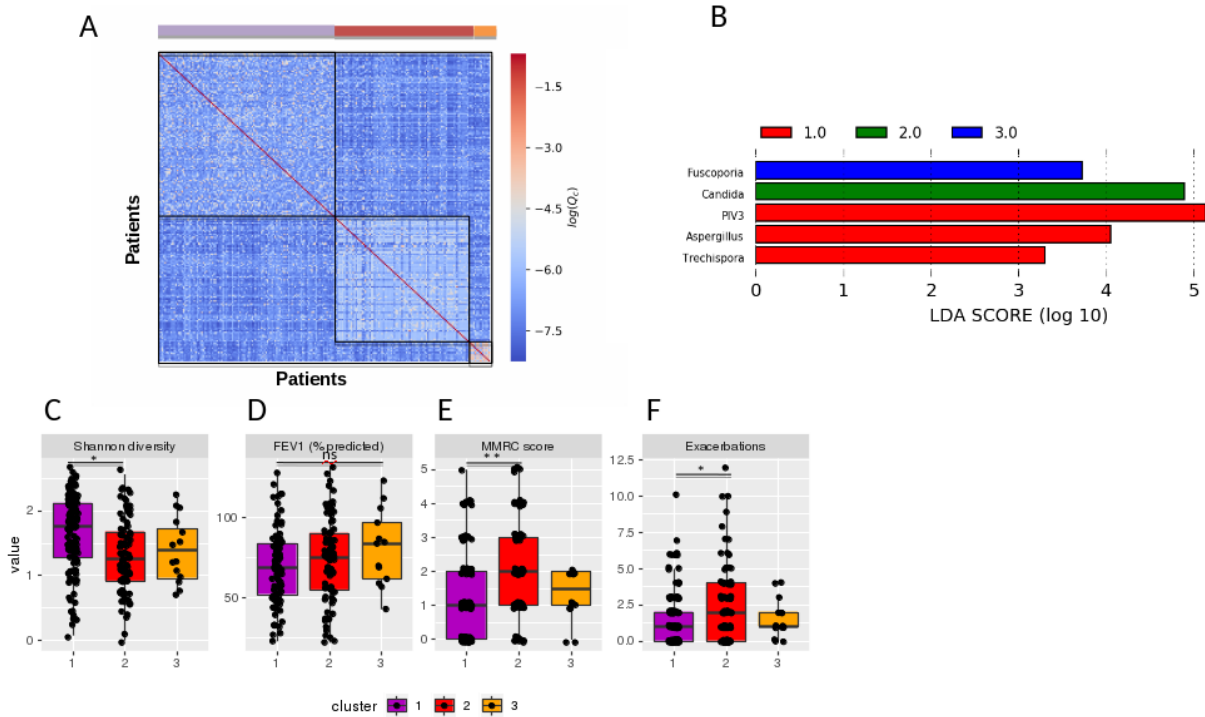


Figure 3.2: **Clusters based on the integrated biome(Unweighted):** A) Heat map of the patient similarity matrix Q_c obtained from the merging of bacteriome, fungome and virome plotted in log scale. x and y axis of the matrix represent the patients and each entry of the matrix represents the similarity between them. B) A plot representing the results of LEfSe on the clusters with x-axis and y-axis representing the effect size and species respectively. The plot only represents species that have an LDA score of ≥ 3 . C,D,E,F) Box plot representing the differences between various outcomes, Mann-Whitney U test is applied to calculate the statistical significance for difference in the clusters. “*” represents p-value ≤ 0.05 , “**” p-value ≤ 0.01 , “***” p-value ≤ 0.001 and “n.s” not significant

3.3 Integrative Microbiomics using weighted Similarity Network Fusion identifies a high-risk cluster with increased precision

Spectral clustering on the integrated microbiome (bacteriome, fungome and virome) using weighted SNF with $t = 20$ and $k = 5$ (tuned) was implemented using the microbes that were present in at least 5% of the patients ($n=10$) revealed two clusters [Table 3.3] with an average silhouette score of 0.796. The weights for each biome were set to the number of microbes that were present in at least 5% of the CAMEB patients ($n=10$) i.e. Bacteriome: 62, Fungome: 52, Virome: 4, because the number of microbes in each dataset affects the quality of the individual similarity network since similarity is assessed using bray-curtis. Figure 3.3A illustrates these 2 clusters by a heat map of the merged pair-wise similarity

matrix of the patients. Figure 3.3B represents the top 20 species that are most influential in differentiating the clusters. Analysing the clinical difference between clusters we find a cluster of high risk patients that have relatively high number of exacerbation (median = 2) and MMRC Score (median=2). Further, we also observe that using a weighted strategy has increased the statistical significance of “number of exacerbations” between the two clusters, which is reflected by the drop of p-value from 0.03964 [Table 3.2] to 2.46×10^{-5} [Table 3.3].

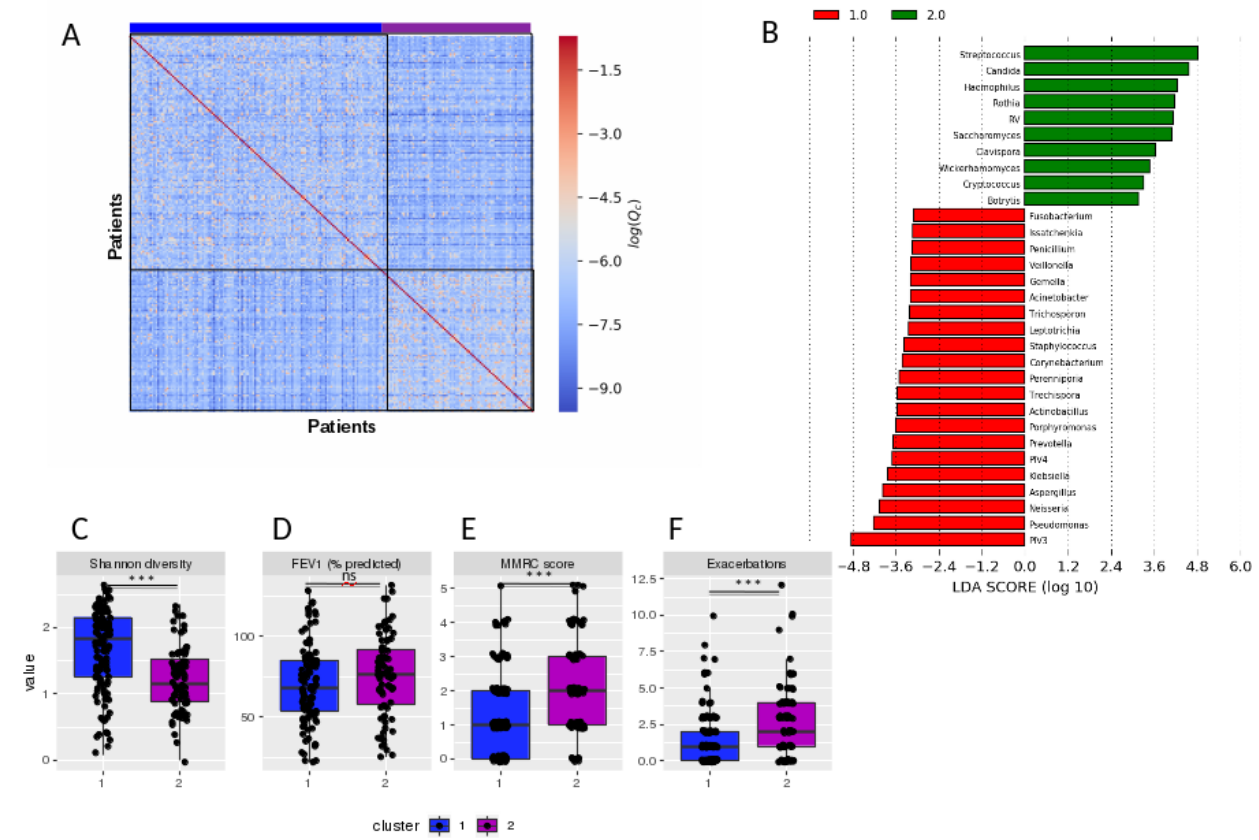


Figure 3.3: **Clusters based on the integrated biome(Weighted):** A) Heat map of the patient similarity matrix Q_c described from the merging of bacteriome, fungome and virome in a weighted fashion plotted in log scale. x and y axis of the matrix represent the patients and each entry of the matrix represents the similarity between them. B) A plot representing the results of LefSe on the clusters with x-axis and y-axis representing the effect size and species respectively. The plot only represents species that have an LDA score of ≥ 3 and the top 20 species. C,D,E,F) Box plot representing the differences between various outcomes, Mann-Whitney U test is applied to calculate the statistical significance for difference in the clusters. “**” represents p-value ≤ 0.05 , “***” p-value ≤ 0.01 , “****” p-value ≤ 0.001 and “n.s” not significant

Slno	Outcomes	Cluster 1	Cluster 2	P-value
	Number of patients	134	83	
1	BSI			
2	A.terreus conidial burden	1186.5	3617	0.00027
3	MMRC Score	1	2	6.25E-06
4	No of Hospitalisation bf study			
5	No of Exacerbation bf study	1	2	2.46E-05
6	BMI	20.25	26.24	
7	Asian/European(Asian,European)	(78.35%,21.64%)	(18.07%,81.92%)	0.00049

Table 3.3: Clinical outcome Comparison on the merged bacteriome, fungome and virome using weighted SNF. Each value of the column cluster 1 & 2 represents the median value of that outcome. Medians of variables that were not statistically significant at an α level of 0.05 are not reported.

3.4 Co-occurrence analysis reveals difference in number of negative interaction between clusters

A co-occurrence analysis with microbes that were present in at least 5% of the patients at an abundance of $\geq 1\%$ was implemented on the two clusters obtained in section 3.3. Interactions between microbes were classified as negative if the sign of the edge weights between them is negative. If an interaction is negative then increase in the abundance of source microbe leads to decrease in abundance of the target microbe. 26.05% of the interactions were negative in cluster 1 whereas 32.51% were negative in cluster 2 as shown in Figure 3.4

3.5 Busy, influential and critical microbes among identified clusters

The degree, betweenness centrality and stress centrality of the nodes/microbes were assessed in the microbial association networks of the two clusters described in section 3.3 [Table 3.4]. From a biological standpoint, highest degree nodes are regarded as “busy” microbes. Nodes that have high-stress centrality are considered as the most “critical” microbes, and the nodes that have high betweenness centrality are equated to the most “influential” microbe.

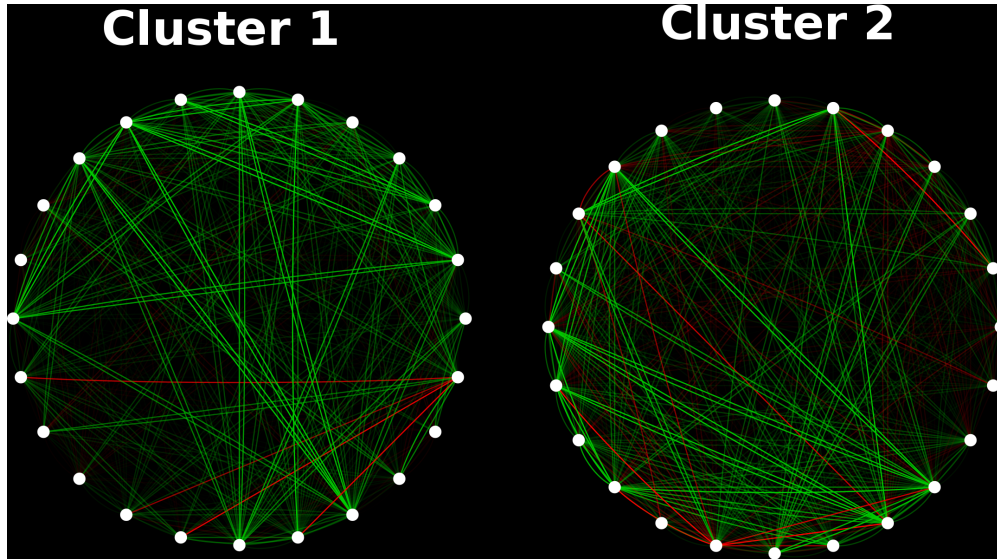


Figure 3.4: **Differential interactions:**The above graphs represent the microbial association network in cluster 1 and 2. The 'white' colour nodes represent the microbes, 'green' coloured edges represents positive interactions, 'red' the negative interactions and the depth of the colour represents the strength of the interaction.

Upon assessing for microbes that are busy, influential and critical in the microbial association network of cluster 1, we find *Rothia*, *Streptococcus* and *Haemophilus* as the top 3 microbes. Figure 3.5 illustrates the interactome [i.e. interaction between microbes] in cluster 1 and highlights the interactions of these 3 microbes. These 3 microbes together, interact with all other microbes of the network [Figure 3.5C]. Also, Figure 3.5D shows that *Haemophilus* interacts negatively with some microbes.

Haemophilus, *Leptotrichia*, *Porphyromonas*, *Prevotella*, *Veillonella* and *Cryptococcus* have the highest betweenness centrality, degree and stress centrality in the microbial association networks of cluster 2 [Table 3.4]. Hence, these microbes are busy, influential and critical in cluster 2. However, *Leptotrichia* and *Porphyromonas* are bacteria that are mainly present in the oral cavity [Eribe and Olsen, 2008][Darveau et al., 2012]. Figure 3.6 illustrates the interactome of cluster 2 highlighting the interactions of *Haemophilus*, *Prevotella*, *Veillonella* and *Cryptococcus*. These microbes together, connect all other microbes in the network.

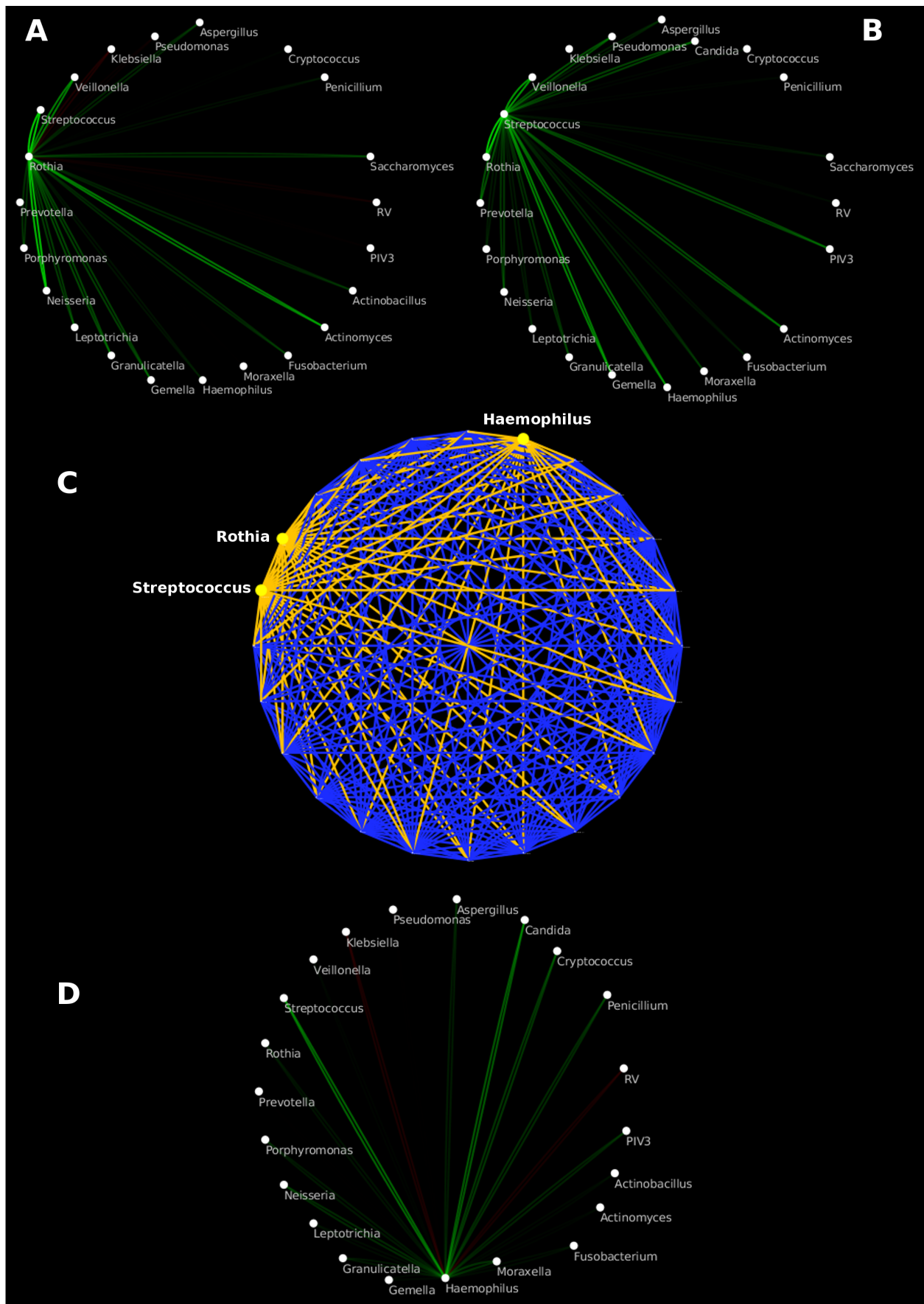


Figure 3.5: **Interactome of Cluster 1:** C) Represents the interactome(i.e, interactions) of cluster 1. Highlighted yellow edges represent the interactions of *Haemophilus*, *Rothia* and *Streptococcus*. A,B,D) Represents the interaction of *Rothia*, *Streptococcus* and *Haemophilus* respectively with 'green' edges representing positive interactions and 'red' edges the negative interactions. The depth of the color represents the strength of the interactions.

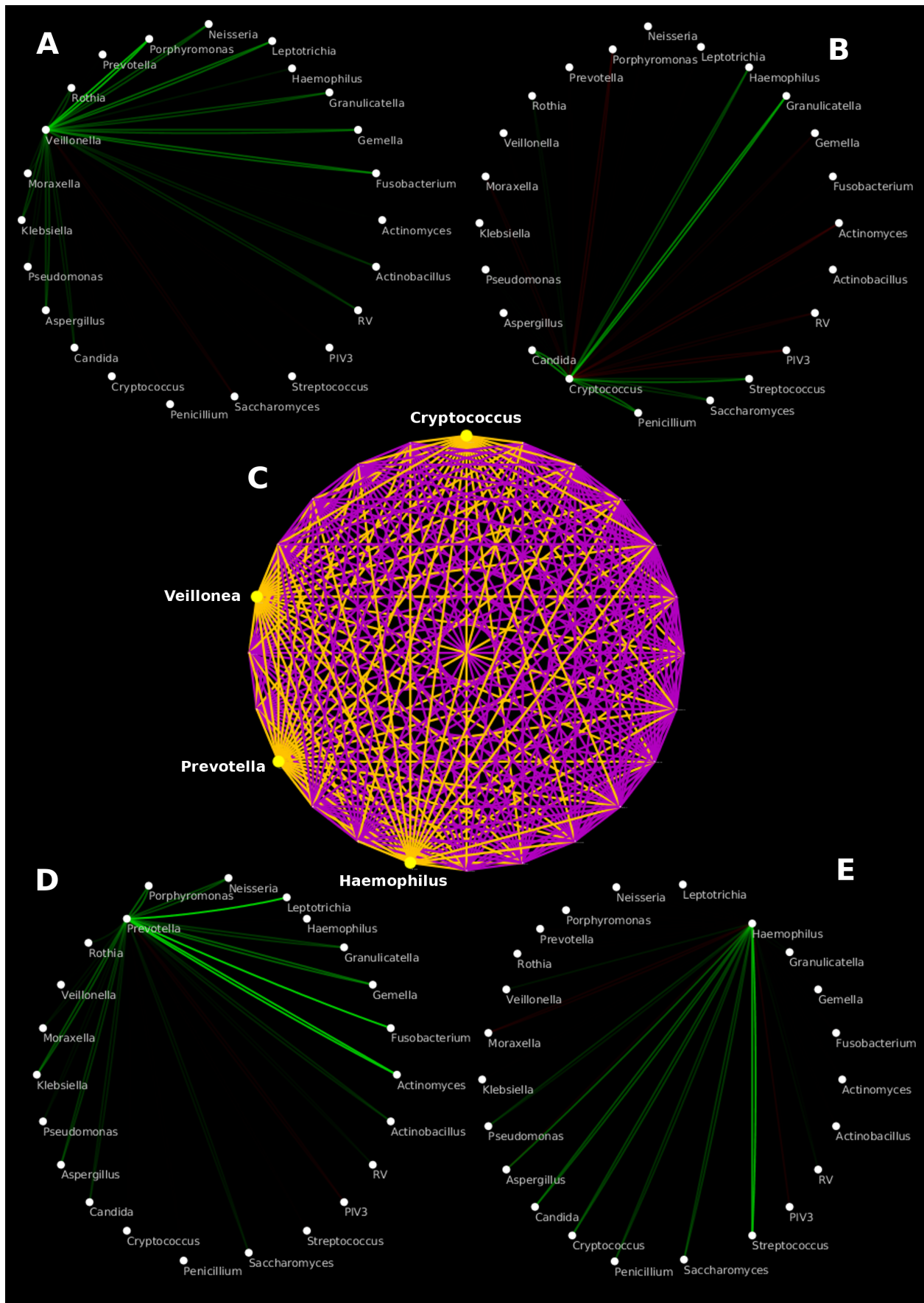


Figure 3.6: **Interactome of Cluster 2:** C) The graph represents the interactome (i.e. interactions) of cluster 2. Highlighted yellow edges represent the interactions of *Cryptococcus*, *Veillonella*, *Prevotella* and *Haemophilus*. A,B,D,E) These graphs represent the interaction of *Cryptococcus*, *Veillonella*, *Prevotella* and *Haemophilus* with other microbes, respectively with 'green' edges representing positive interactions and 'red' edges the negative interactions. The depth of the color represents the strength of the interactions.

Cluster 1

Name	BetweennessCentrality	Degree	Indegree	Outdegree	Stress
Rothia	0.00908866	44	22	22	68
Streptococcus	0.00848975	44	22	22	64
Haemophilus	0.00827524	44	22	22	62
Actinomyces	0.00786141	44	22	22	60
PIV3	0.00735932	44	22	22	58
Leptotrichia	0.00667178	44	22	22	54
Neisseria	0.00667178	44	22	22	54
Pseudomonas	0.00667178	44	22	22	54
Candida	0.0085593	42	21	21	64
Fusobacterium	0.00751064	42	21	21	56
Prevotella	0.00738574	42	21	21	56
Porphyromonas	0.00667775	42	21	21	50
Cryptococcus	0.00507429	42	21	21	42
Gemella	0.0057534	40	20	20	44
Veillonella	0.00416799	40	20	20	34
Actinobacillus	0.00550245	38	19	19	42
Klebsiella	0.00499935	38	19	19	38
Saccharomyces	0.00625511	36	18	18	46
Granulicatella	0.00223116	36	18	18	20
Aspergillus	0.00406501	34	17	17	32
Penicillium	0.00272523	34	17	17	24
Moraxella	0.00130456	32	16	16	12
Rhino Virus	0.00276366	28	14	14	24

Cluster 2

Name	BetweennessCentrality	Degree	Indegree	Outdegree	Stress
Haemophilus	0.00295824	46	23	23	29
Leptotrichia	0.00295824	46	23	23	29
Porphyromonas	0.00295824	46	23	23	29
Prevotella	0.00295824	46	23	23	29
Veillonella	0.00295824	46	23	23	29
Cryptococcus	0.00295824	46	23	23	29
Actinobacillus	0.00277858	44	22	22	27
Actinomyces	0.00207556	44	22	22	21
Fusobacterium	0.0025724	44	22	22	25
Granulicatella	0.00207556	44	22	22	21
Neisseria	0.00258181	44	22	22	25
Streptococcus	0.00277858	44	22	22	27
Klebsiella	0.00258181	44	22	22	25
Aspergillus	0.00244046	44	22	22	24
Candida	0.00205755	44	22	22	21
Penicillium	0.00255044	44	22	22	25
Gemella	0.00168971	42	21	21	17
Moraxella	0.00186933	42	21	21	19
Saccharomyces	0.00258181	42	21	21	25
PIV3	0.00235281	42	21	21	23
Rothia	0.00176581	41	20	21	18
Pseudomonas	0.00094194	38	19	19	10
Rhino Virus	0.00122427	38	19	19	13

Table 3.4: Table showing total degree, in degree, out degree, betweenness centrality and stress centrality of the nodes from the microbial association networks of both the clusters from section3.3.

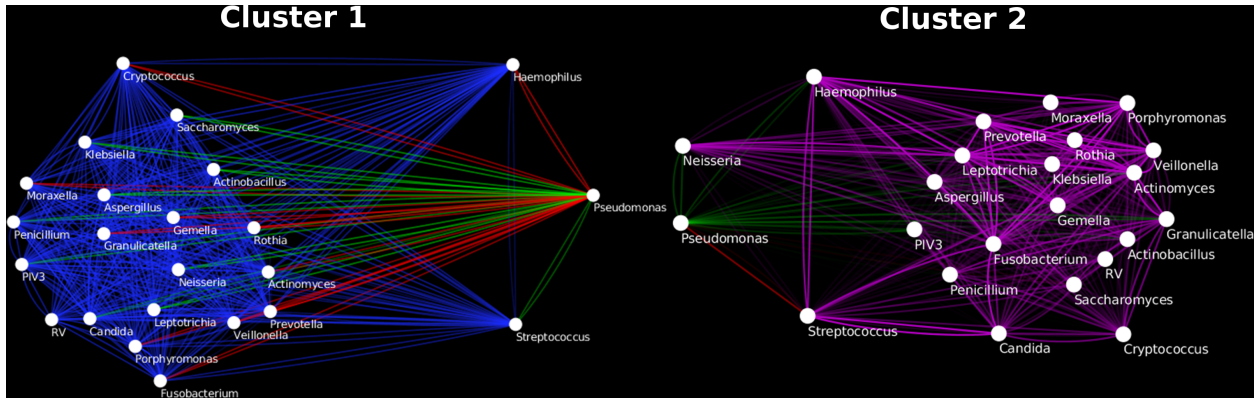


Figure 3.7: *Pseudomonas* specific interaction : *Pseudomonas* specific interactions in cluster 1 and 2 are coloured in 'red' and 'green' with 'red' representing negative interaction and 'green' the positive interaction. The depth of the edge colour represents the strength of the interaction.

3.6 *Pseudomonas* specific interaction in the clusters

Pseudomonas is a pathogenic bacteria and well known for its role in bronchiectasis [Evans et al., 1996]. *Pseudomonas* specific interaction was assessed in microbial association networks from both the clusters described in section3.3. Figure 3.7 shows that *Pseudomonas* interacts negatively with *Haemophilus* and positively with *Sterptococcus* in Cluster 1. Whereas *Pseudomonas* interacts negatively with *Streptococcus* and positively with *Haemophilus* and *Neisseria* in Cluster 2. Further, a loss of *Pseudomonas* specific negative interactions is observed, from 22 in cluster 1 to 4 in cluster 2.

Chapter 4

Discussion

4.1 Introduction

Clustering of Bray-Curtis similarity graphs on CAMEB patients constructed using their bacteriome, by spectral clustering, identifies clusters of potential high-risk patients, i.e. patients that have a relatively high median number of exacerbation whereas other individual biomes such as fungome and virome fail to do so. Integrating these microbiomes using weighted SNF and further clustering them using spectral clustering identifies clusters of potential high-risk patients with increased precision (i.e., decrease in p-value from 2.1×10^{-4} to 2.4×10^{-5}). Interestingly, analysis of the microbial association network between these two clusters revealed an increase in the number of negative interactions in the potential high-risk cluster and dissection of *Pseudomonas* specific interaction revealed that *Pseudomonas* interacts differently with the same microbes in different clusters.

4.2 Integration of Microbiomes

In this thesis, we have integrated microbiomes using SNF. However, SNF[Wang et al., 2014] suffers from limitations which include assigning equal weights to all datasets; not all biological datasets are collected with the same precision/quality. In the context of the microbiomes, there is an inherent problem in the quality of different biomes; this is because the quality/precision of the microbiomes depends on the reference database that the sequences are

matched with. Presently, the 16s bacterial database is better characterised than ITS1 and ITS2 fungal databases, and viral reference databases are the least well characterised. Hence there is an inherent need to weight datasets in the application of any integration strategy, particularly in the context of the microbiome. Therefore, we modified the method of SNF to incorporate weights and showed that weighting increases the precision of identifying high-risk patients based on clinical outcomes [section 3.3]. There are several modifications of SNF including Affinity Network Fusion(ANF)[Ma and Zhang, 2018] and Robust Similarity Network Fusion(RSNF)[Zhang et al., 2017] which are as good as SNF or even better. ANF is computationally less expensive than SNF and also supports weighting on the other hand RSNF is robust to noise and uses random forest for similarity matrix construction. The weighted SNF proposed in this thesis is well suited in the microbiome and ecological setting as, unlike SNF, ANF and RSNF which uses the gaussian kernel and random forest to create similarity matrices which is often not suitable due to sparse and relative abundance structure of the microbiome datasets. We use biologically relevant, context-dependent, bray-curtis similarity measure to merge microbiome datasets. However, the proposed weighted SNF can only be applied in the integration of 3 or more datasets, and further analysis is needed to benchmark the weighted SNF algorithm with other modified SNF algorithms. Data integration can be achieved through different methods, Rappoport in his paper [Rappoport and Shamir, 2018] describes and benchmarks different type of data integration algorithms including SNF(an intermediate integration technique), LRAcluster(an early integration technique), PINS(a late integration technique) and Deep learning based integration methods. Performance of SNF in terms of identifying significant clinical parameters with precision was not the best of all algorithms considered.

Even though better data integration algorithms other than weighted SNF can be applied to integrate microbiomes, in this thesis, we show for the first time that integration of microbiomes is possible, advantageous and a weighting strategy is necessary. As proof of concept, we showed this using SNF and weighted SNF with bray-curtis similarity as a similarity index in the context of bronchiectasis. Singular analysis of fungal and viral datasets separately is not capable of identifying high-risk patients and traditionally would have been considered not useful and dropped. However, integration of these fungal and viral with bacterial data-set refines clustering by increasing or decreasing the statistical significance of clinical outcomes as it makes use of even small signals that are common between the microbes and amplifies it. This increase in statistical significance can be observed by a decrease in a p-value of “number of exacerbation” from 2.1×10^{-4} , bacteriome alone to 2.4×10^{-5} , merged biome

using weighted SNF. Further, we observe that integrating microbiomes using Unweighted SNF does not increase statistical significance in identifying the potential high-risk patients which is reflected by the increase in a p-value of “number of exacerbation” from 0.00021, bacteriome alone to 0.03964, merged biome using SNF.

4.3 Microbes and Microbial association network

4.3.1 LEfSe and isolated microbes

LEfSe is a bio-marker discovery tool that uses relative abundance data between two or more groups. In the context of microbiomes, this tool is capable of identifying species that are most influential in differentiating the clusters. Some microbes such as *Aspergillus*, *Pseudomonas*, *Streptococcus* and *Candida* are known to be associated with disease severity in bronchiectasis. LEfSe successfully captures these organisms in the clusters of weighted SNF [Figure 3.3B]. *Fuscoporia* and *Trechispora* are one of the many microbes that differentiate the clusters from Unweighted SNF [Figure 3.2B]. However, these fungi are previously not known to be pathogenic and associated with bronchiectasis. This may be due to the drawbacks of LEfSe which includes assumption of independence between microbes, this is not true as microbes interact with each other and co-exist in communities. These interactions might be more important in disease progression rather than isolated organisms. Hence we sought resolve this by implementing a co-occurrence approach.

4.3.2 Co-occurrence analysis

Microbes interact in various ways with each other, some common types of interactions are Parasitism or Predation (+, -), Amensalism (0, -), Commensalism (+, 0), Mutualism (+, +) and Competition (-, -). Also, microbial datasets suffer from the problem described in section 1.1.1 i.e. observed relative abundance values might be due to both random processes and actual ecological interactions. To account for this, we implemented a co-occurrence analysis with statistical significance testing using “Reboot” as described in [Faust et al., 2012] with some modifications [described in section 4.3.3]. Although this method tries to capture actual ecological interactions between the microbes, it does not stratify the type of interaction

which is crucial in identifying possible mechanisms and pathways. Hence further experimental validation of these interactions is needed. Our lung microbiome is dynamic and changes as time progresses. The co-occurrence analysis presented in this thesis does not consider dynamic changes in the lung microbiome due to the cross-sectional nature of the CAMEB cohort [Mac Aogáin et al., 2018]. An interaction from the microbial association network of a time slice may not necessarily be constant over time. However, these methods serve as good starting point to make hypothesis which can be further validated experimentally.

4.3.3 Improvements

Improvements from the method of [Faust et al., 2012] includes the implementation of Mutual Information as a similarity measure instead of Kullback-Leibler(KL) divergence. Mutual information is the KL divergence of uni-variate distribution of X from conditional distribution of $X | Y$ and this biologically more relevant than KL divergence between X and Y . Secondly, we implemented the Mann-Whitney U test instead of a Z-test with the pooled variance to compare between the null and bootstrap distribution. As the distributions are not necessarily normal and hence a non-parametric test such as Mann-Whitney is more appropriate. Thirdly, we merge the networks from the ensemble in a weighted fashion using weighted Sime's test. This is important as an imbalance in the ensemble method can suppress actual signals and exemplify the errors. Lastly, we use an abundance-prevalence filter to the microbiome datasets before running co-occurrence analysis. The abundance-prevalence filter keeps only the microbes that are more than 1% abundant in at least 5% of the patients. This filter is applied to remove interactions that result from random noise. However, the filter also removes weak signals.

4.4 Clinical Relevance

Exacerbation is a significant event in bronchiectasis because during exacerbation symptoms such as cough, sputum purulence (sputum that contains white blood cells), breathlessness, fatigue and haemoptysis (cough with blood) get worse abruptly requiring clinical intervention. It is associated with increased hospitalisation and mortality. The ability to predict exacerbation may allow early identification and treatment. Exacerbation in CRDs is known to be associated with microbes such as virus and bacteria [Dickson et al., 2014].

However, it is not known whether it is a cause or effect [Dickson et al., 2014]. The microbiome that is used in this thesis is assessed from the sputum of stable bronchiectasis patients. Dickson showed that in CRDs there is no change in bacterial density or community diversity during exacerbation compared to the stable state [Dickson et al., 2014]. Hence studying microbes at a stable state is useful. We find that the bacteriome clusters identify high-risk patients that have a higher exacerbation [Figure 3.1] which is consistent with the literature [Dickson et al., 2014]. However, our virome data-set does not seem to identify patients that have a higher exacerbation [Table 3.1]; this might be due to small number of viruses in the data-set. Only four viruses were used for further analysis after filtering the data.

Co-occurrence analysis of the two clusters from weighted SNF shows an increase in the number of negative edges in the potential high-risk cluster. Hence there is relatively greater competition among microbes in the high-risk cluster compared to cluster 1. This increase in competition may lead to a decrease in the diversity of the lung microbiome which is associated with increased exacerbation.

Microbes/nodes that have a high degree(i.e. busy) may not always be suitable for targeted therapy. For example, *Pseudomonas* in cluster 1 is busy (highest degree) and it is known that *Pseudomonas* is associated with bronchiectasis [Rogers et al., 2014]. However, it is difficult to eradicate *Pseudomonas* from the lung microbiome of the patients [Rogers et al., 2014]. Hence, it is important to target critical nodes (i.e. nodes that have high stress, number of shortest paths passing through it) and influential nodes (i.e. nodes with high Betweenness centrality). Hence we looked at microbes that are all busy, influential and critical in different clusters. *Rothia*, *Streptococcus* and *Haemophilus* are the busy, influential and critical microbes of the potential low-risk cluster from weighted SNF. These microbes have been reported in the context of bronchiectasis previously [Lee et al., 2018][Mac Aogáin et al., 2017]. Lee in his paper [Lee et al., 2018] finds that *Rothia* and *Haemophilus* are significantly more abundant in a mild bronchiectasis group consistent with our finding. On the other hand, *Haemophilus*, *Leptotrichia*, *Porphyromonas*, *Prevotella*, *Veillonella* and *Cryptococcus* are busy, influential and critical in the potential high-risk cluster of weighted SNF. *Haemophilus*, *Prevotella*, *Veillonella* and *Cryptococcus* are previously known to be associated with bronchiectasis [Mac Aogáin et al., 2018][Faner et al., 2017]. However, *Leptotrichia* and *Porphyromonas* is not well known in bronchiectasis. *Leptotrichia* species is found in the oral cavity and genitourinary tract. It is reported as an emerging pathogen in neutropenic (lack of neutrophils) patients [Eribe and Olsen, 2008]. *Porphyromonas* is an oral bacteria mainly found in dental

plaques. It has been found that *Porphyromonas* has a strong association with periodontitis disease [Darveau et al., 2012]. These two oral taxa's could be possibly due to oral contamination of the sputum from the patients if not this shows that microbes from different sites such as lung and oral cavity interact with each other and probably play an influential and critical role to maintain in the lung microbiome of the potentially high-risk patients.

Pseudomonas is a well-known pathogen in the context of bronchiectasis [Purcell et al., 2014]. *Pseudomonas* in the potential low-risk cluster is one of the busiest and critical but not in the top 5 influential microbes [Table 3.4]. Whereas, *Pseudomonas* in the potential high-risk cluster is the least busy, least critical and least influential. However, *Pseudomonas* is known to be associated with exacerbation and reduced lung function [Purcell et al., 2014], but we observe *Pseudomonas* is the least busy, critical and influential in the high-risk cluster with relatively high exacerbation. Hence, we looked into the *Pseudomonas* specific interactions in both the clusters. Interestingly, we find that how *Pseudomonas* interacts with *Haemophilus* and *Streptococcus* is different between both the clusters [Figure 3.7]. In the low-risk cluster, we find that *Pseudomonas* interacts negatively with *Haemophilus* and positively with *Streptococcus* whereas in the high-risk cluster this interaction is reversed. Interestingly, *Pseudomonas* also interacts positively with *Neisseria*. Hence, it is not the presence of *Pseudomonas* that characterises the high-risk cluster; rather it is what and how *Pseudomonas* is interacting with other organisms which defines it. Hence interactions between organisms are of equal if not greater importance than isolated organisms. However, experimental validation is needed to confirm this interaction. Further, little is known about *Neisseria* in the context of bronchiectasis, and future studies are required to assess its role in bronchiectasis.

4.5 Conclusion

In this thesis, as a proof of concept, we showed that integrating microbiomes is advantageous as it increases precision in identification of high-risk patients based on clinical data in bronchiectasis. However “Integrative Microbiomics” is not disease-specific and can be applied to merge microbiomes in general. Implementation of other data-set merging algorithms such as ANF, Deep learning based methods instead of SNF may further increase the precision of identification of high-risk patients. We also showed that a weighted strategy in integrating microbiomes increases the precision of identifying high-risk patients as not all microbiomes are of equal biological relevance. Future microbiome research should consider

integrating microbiome datasets for their studies rather than focusing on singular microbiome data. Using co-occurrence analysis, we showed that interactions between microbes are more important than isolated microbes in driving various disease states by showing the differential interaction of *Pseudomonas* between low-risk and high-risk patients. However, this is not limited to bronchiectasis and *Pseudomonas*, and can be applied to microbiomes of any disease. Hence, future research should start looking into the interactome (i.e. interactions between microbes) rather than associating clinical outcomes to isolated organisms.

Bibliography

- [Afgan et al., 2018] Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., et al. (2018). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46(W1):W537–W544.
- [Barker, 2002] Barker, A. F. (2002). Bronchiectasis. *New England Journal of Medicine*, 346(18):1383–1393.
- [Benjamini and Hochberg, 1997] Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418.
- [Chalmers and Chotirmall, 2018] Chalmers, J. D. and Chotirmall, S. H. (2018). Bronchiectasis: new therapies and new perspectives. *Lancet Respir Med*, 6(9):715–726.
- [Chalmers et al., 2014] Chalmers, J. D., Goeminne, P., Aliberti, S., McDonnell, M. J., Lonni, S., Davidson, J., Poppelwell, L., Salih, W., Pesci, A., Dupont, L. J., et al. (2014). The bronchiectasis severity index. an international derivation and validation study. *American journal of respiratory and critical care medicine*, 189(5):576–585.
- [Chandrasekaran et al., 2018] Chandrasekaran, R., Aogáin, M. M., Chalmers, J. D., Elborn, S. J., and Chotirmall, S. H. (2018). Geographic variation in the aetiology, epidemiology and microbiology of bronchiectasis. *BMC Pulmonary Medicine*, 18(1).
- [Connor and Simberloff, 1979] Connor, E. F. and Simberloff, D. (1979). The assembly of species communities: Chance or competition? *Ecology*, 60(6):1132.
- [Darveau et al., 2012] Darveau, R., Hajishengallis, G., and Curtis, M. (2012). Porphyromonas gingivalis as a potential community activist for disease. *Journal of dental research*, 91(9):816–820.
- [Dickson et al., 2014] Dickson, R. P., Martinez, F. J., and Huffnagle, G. B. (2014). The role of the microbiome in exacerbations of chronic lung diseases. *The Lancet*, 384(9944):691–702.
- [Eribe and Olsen, 2008] Eribe, E. R. K. and Olsen, I. (2008). Leptotrichia species in human infections. *Anaerobe*, 14(3):131–137.

- [Evans et al., 1996] Evans, S., Turner, S., Bosch, B., Hardy, C., and Woodhead, M. (1996). Lung function in bronchiectasis: the influence of *Pseudomonas aeruginosa*. *European Respiratory Journal*, 9(8):1601–1604.
- [Faner et al., 2017] Faner, R., Sibila, O., Agust, A., Bernasconi, E., Chalmers, J. D., Huffnagle, G. B., Manichanh, C., Molyneaux, P. L., Paredes, R., Prez Brocal, V., Ponomarenko, J., Sethi, S., Dorca, J., and Mons, E. (2017). The microbiome in respiratory medicine: current challenges and future perspectives. *The European respiratory journal*, 49.
- [Faust and Raes, 2012] Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538.
- [Faust et al., 2012] Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology*, 8(7):e1002606.
- [Hofner et al., 2014] Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2014). Model-based boosting in r: a hands-on tutorial using the r package mboost. *Computational statistics*, 29(1-2):3–35.
- [Hope, 1968] Hope, A. C. (1968). A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(3):582–598.
- [Hothorn et al., 2018] Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2018). *mboost: Model-Based Boosting*. R package version 2.9-1.
- [Hubbell, 2011] Hubbell, S. P. (2011). *Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press.
- [Knight et al., 2017] Knight, R., Callewaert, C., Marotz, C., Hyde, E. R., Debelius, J. W., McDonald, D., and Sogin, M. L. (2017). The microbiome and human biology. *Annual Review of Genomics and Human Genetics*, 18(1):65–86.
- [Lee et al., 2018] Lee, S., Lee, Y., Park, J., Cho, Y.-J., Yoon, H., Lee, C.-T., and Lee, J. (2018). Characterization of microbiota in bronchiectasis patients with different disease severities. *Journal of clinical medicine*, 7(11):429.
- [Ma and Zhang, 2018] Ma, T. and Zhang, A. (2018). Affinity network fusion and semi-supervised learning for cancer patient clustering. *Methods*, 145:16–24.
- [Mac Aogáin et al., 2018] 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. (2018). Immunological corollary of the pulmonary mycobioime in bronchiectasis: The cameb study. *European Respiratory Journal*.

- [Mac Aogáin et al., 2017] Mac Aogáin, M., Lim, A. Y. H., Low, T. B., Tan, G. L., Yii, A. C., Chandrasekaran, R., Poh, T. Y., Ng, A. H. Q., Bertrand, D., Koh, J. Y., Leong, C. K.-L., Nagarajan, N., Abisheganaden, J., Koh, M., and Chotirmall, S. H. (2017). The pulmonary microbiome in non-cystic fibrosis bronchiectasis. *European Respiratory Journal*, 50(suppl 61).
- [Macqueen, 1967] Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.
- [Martin Cody, 1975] Martin Cody, J. D. (1975). *Eco and Evol Comm C*. HARVARD UNIV PR.
- [Meyer et al., 2008] Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). Minet: An open source r/bioconductor package for mutual information based network inference. *BMC Bioinformatics*, 9.
- [Oksanen et al., 2018] Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O’Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2018). *vegan: Community Ecology Package*. R package version 2.5-3.
- [Organization, 2007] Organization, W. H. (2007). *Global Surveillance, Prevention and Control of Chronic Respiratory Diseases: A Comprehensive Approach*. World Health Organization.
- [Polverino et al., 2017] Polverino, E., Goeminne, P. C., McDonnell, M. J., Aliberti, S., Marshall, S. E., Loebinger, M. R., Murriss, M., Cantón, R., Torres, A., Dimakou, K., et al. (2017). European respiratory society guidelines for the management of adult bronchiectasis. *European Respiratory Journal*, 50(3):1700629.
- [Purcell et al., 2014] Purcell, P., Jary, H., Perry, A., Perry, J. D., Stewart, C. J., Nelson, A., Lanyon, C., Smith, D. L., Cummings, S. P., and De Soyza, A. (2014). Polymicrobial airway bacterial communities in adult bronchiectasis patients. *BMC microbiology*, 14(1):130.
- [Rappoport and Shamir, 2018] Rappoport, N. and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46(20):10546–10562.
- [REID, 1950] REID, L. M. (1950). Reduction in bronchial subdivision in bronchiectasis. *Thorax*, 5:233–247.
- [Rogers et al., 2014] Rogers, G. B., Bruce, K. D., Martin, M. L., Burr, L. D., and Serisier, D. J. (2014). The effect of long-term macrolide treatment on respiratory microbiota composition in non-cystic fibrosis bronchiectasis: an analysis from the randomised, double-blind, placebo-controlled bless trial. *The Lancet Respiratory Medicine*, 2(12):988–996.

- [Rothschild et al., 2018] Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P. I., Godneva, A., Kalka, I. N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature*, 555(7695):210.
- [Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.
- [Segata et al., 2011] Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6):R60.
- [Seitz et al., 2012] Seitz, A. E., Olivier, K. N., Adjemian, J., Holland, S. M., and Prevots, D. R. (2012). Trends in bronchiectasis among medicare beneficiaries in the united states, 2000 to 2007. *CHEST*, 142(2):432–439.
- [Sender et al., 2016] Sender, R., Fuchs, S., and Milo, R. (2016). Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. *Cell*, 164(3):337–340.
- [Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13:2498–2504.
- [Wang et al., 2018] Wang, B., Mezlini, A., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2018). *SNFtool: Similarity Network Fusion*. R package version 2.3.0.
- [Wang et al., 2014] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11:333 EP –. Article.
- [Zhang et al., 2017] Zhang, Y., Hu, X., and Jiang, X. (2017). Multi-view clustering of microbiome samples by robust similarity network fusion and spectral clustering. *IEEE/ACM transactions on computational biology and bioinformatics*, 14(2):264–271.