MICROBIOME DATA ANALYSIS USING COMPOSITIONAL DATA APPROACH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ASLI BOYRAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF HEALTH INFORMATICS

NOVEMBER 2022

MICROBIOME DATA ANALYSIS USING COMPOSITIONAL DATA APPROACH

submitted by **ASLI BOYRAZ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Banu Günel KILIÇ
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın SON
Head of Department, **Health Informatics**

Assist. Prof. Dr. Aybar C. ACAR
Supervisor, **Health Informatics, Middle East Technical University**

Assist. Prof. Dr. Özkan Ufuk NALBANTOĞLU
Co-supervisor, **The Department of Computer Engineering, Erciyes University**

Examining Committee Members:

Prof. Dr. Hasan OĞUL
School of Computer Engineering, Çankaya University

Assist. Prof. Dr. Aybar C. ACAR
Health Informatics, Middle East Technical University

Assoc. Prof. Dr. Yeşim Aydın SON
Health Informatics, Middle East Technical University

Assoc. Prof. Dr. Tunca DOĞAN
School of Computer Engineering, Hacettepe University

Assist. Prof. Dr. Burçak OTLU SARITAŞ
Health Informatics, Middle East Technical University

Date: 18.11.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Aslı BOYRAZ

Signature :

ABSTRACT

MICROBIOME DATA ANALYSIS USING COMPOSITIONAL DATA APPROACH

BOYRAZ, Aslı
Ph.D., Department of Health Informatics
Supervisor: Assist. Prof. Dr. Aybar C. ACAR
Co-Supervisor: Assist. Prof. Dr. Özkan Ufuk NALBANTOĞLU

NOVEMBER 2022, 94 pages

The microorganisms present in the human body play a crucial role in maintaining human health, and the environmental microbiome influences the human microbiome. Advanced understanding of the human microbiome and indoor microbiota is the first step towards understanding the potential relationships between health and microbiome. Next Generation Sequencing (NGS) enables identification and study of a large number of microorganisms in a short time. With the identification of a large number of microorganisms, the studies for the understanding of their role in the environment and human health have become important. This thesis examines the production and the properties of microbiome data and statistical challenges of microbiome analysis. First, we give a brief history of the various methods of analysing microbiome data. We are mainly concerned with performing microbiome analysis using compositional approaches. The proposed procedures were illustrated with the data from 16S rRNA amplicon sequencing but those also apply for microbiome shotgun metagenomics. This dissertation describes the basics of compositional data (CoDa) analysis introducing log-ratio methodology. The first part of this thesis deals with the problem of establishing relationship based on the microbial features annotated with taxonomic information, where a compositional alternative to phylogenetic grouping of microbiome data (Principal Microbial Groups - PMGs) is proposed to enable working with low-level microbial features (OTUs or ASVs). The usefulness of the proposed procedure is illustrated on a Cirrhosis dataset to search for biomarker candidates. The second part of the thesis focuses on the microbial transmission and PMGs are aimed to investigate any hint to track microbial transmission. An experiment that was conducted at Erciyes University Hospital for this purpose, and swab samples were gathered from the Intense Care Unit (ICU) to construct microbiome profiles. Microbial transmission is carried out between objects, so it is expected that resulting microbiome profiles of samples should have similar microbial structure. In this case, not taxonomic changes but OTU/ASV abundance changes between samples need to be investigated. PMGs procedure were applied to microbial transmission dataset in order to analyze the contagion. PMGs provide a valid grouping for OTUs alternative to taxon grouping using CoDa approach and it offers the possibility of working with coarse group of OTUs, which are not present in a phylogenetic tree in microbiome analysis.

Keywords: microbiome; compositional data; balance; microbial biomarkers

ÖZ

BİLEŞİMSEL VERİ YAKLAŞIMI KULLANARAK MİKROBİYOM VERİ ANALİZİ

BOYRAZ, Aslı
Doktora, Sağlık Bilişimi Bölümü
Tez Yöneticisi: Dr. Öğr. Üyesi. Aybar C. ACAR
Ortak Tez Yöneticisi: Dr. Öğr. Üyesi. Özkan Ufuk NALBANTOĞLU

Kasım 2022, 94 sayfa

İnsan vücudunda bulunan mikroorganizmalar, insan sağlığının korunmasında çok önemli bir rol oynar ve çevresel mikrobiyom, insan mikrobiyomunu etkiler. insan mikrobiyomunun ve iç mekan mikrobiyotasının ileri düzeyde anlaşılması, mikrobiotanın insan sağlığı ile olası ilişkilerini anlamaya yönelik ilk adımdır. Yeni Nesil Dizileme (YND) teknolojisi kısa sürede çok sayıda mikroorganizmanın tanımlanmasını ve incelenmesini sağlar. Çok sayıda mikroorganizmanın kısa sürede tanımlanmasıyla birlikte, çevre ve insan sağlığındaki rollerinin anlaşılmasına yönelik çalışmalar da önem kazanmıştır. Bu tez, mikrobiyom veri analizi üretimini, mikrobiyom verilerinin özelliklerini, mikrobiyom analizinin istatistiksel zorluklarını incelemektedir. Öncelikle, mikrobiyom verilerini analiz etmenin çeşitli yöntemlerinin kısa bir tarihçesini anlattık. Temel olarak, bileşimsel (compositional) yaklaşımları kullanarak mikrobiyom analizi yapmakla ilgilendik. Oluşturulan prosedürler 16S rRNA amplikon dizilemesinden elde edilen verilerle gösterildi, ancak bu prosedurler aynı zamanda shotgun metagenomik verileri için de geçerlidir. Bu tez, log-oran metodolojisini tanıtan bileşimsel veri analizinin temellerini açıklar. Bu tezin ilk bölümü, mikrobiyal özelliklere dayalı ilişki kurma problemi ile ilgilenir ve düşük seviyeli mikrobiyal özelliklerle (OTUs veya ASVs) çalışmayı sağlamak için mikrobiyom verilerinin filogenetik gruplandırılmasına alternatif olarak bileşimsel (compositional) bir yaklaşım (Temel Mikrobiyal Gruplar - TMG) önerilir. Önerilen prosedürün kullanışlılığı Siroz veri setinde biyobelirteç adaylarını aramak için gösterilmektedir. Tezin ikinci kısmı mikrobiyal bulaşmaya odaklanmaktadır ve TMG mikrobiyal bulaş takibi için herhangi bir ipucunun araştırmasında kullanılması amaçlanmıştır. Bu amaçla Erciyes Üniversitesi Hastanesi'nde bir deney yapılmış ve mikrobiyom profilleri oluşturmak için Yoğun Bakım Ünitesinden (YBÜ) sürüntü örnekleri alınmıştır. Mikrobiyal aktarım nesneler arasında ardarda gerçekleştirilir. Bu nedenle numunelerin ortaya çıkan mikrobiyom profillerinin benzer mikrobiyal yapıya sahip olması beklenir. Bu durumda örnekler arasındaki taksonomik değişikliklerin değil, OTU-/ASV bolluk değişikliklerinin araştırılması gerekir. Bulaşmayı analiz etmek için mikrobiyal iletim veri setine Temel Mikrobiyal Gruplar prosedürü uygulanmıştır. TMG'ler, OTU'lar için CoDa yaklaşımını kullanarak takson gruplamasına alternatif olarak geçerli bir gruplama sağlar ve mikrobiyom analizinde filogenetik bir ağaçta bulunmayan kaba OTU'lar grubuyla çalışma imkanı sunar.

Anahtar Kelimeler: mikrobiyom, bileşimsel veri, balans, mikrobiyal biyobelirteç

Babama...

ACKNOWLEDGMENTS

Doktora sürecimde maddi ve manevi her türlü desteğini esirgemeyen başta babam olmak üzere tüm aileme teşekkür ederim. İspanya'da ikinci ailem haline gelen ve benim CODA topluluğu ile tanışmama vesile olan Vera Pawlowsky-Glahn ve Juan Jose Egozcue'ya minnnettarım. Bu süreçte destekleri ve değerli katkılarından dolayı başta danışmanlarım Aybar C. Acar ve Özkan Ufuk Nalbantoğlu olmak üzere tez izleme komitesinde bulunmuş Hasan Oğul ve Yeşim Aydın Son'a teşekkür ederim. Bu tez TÜBİTAK tarafından desteklenmiştir [1059B141601395].

I would like to thank my family, especially my father, for their financial and moral support throughout my doctoral process. I am grateful to Vera Pawlowsky-Glahn and Juan Jose Egozcue, my second family in Spain, made me meet with the CODA community. I would like to thank my advisors Aybar C. Acar and Özkan Ufuk Nalbantoğlu, and Hasan Oğul and Yeşim Aydın Son, who were in my thesis monitoring committee, for their support and valuable contributions during this process. This thesis funded by The Scientific and Technological Research Council of Turkey [1059B141601395].

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

DEDICATION viii

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS x

LIST OF TABLES xvi

LIST OF FIGURES xvii

CHAPTERS

1 INTRODUCTION 1

- 1.1 Research Questions 2
- 1.2 Contributions of the Study 2
- 1.3 Organization of the Thesis 2

2 MICROBIOME DATA AND COMPOSITIONALITY 3

- 2.1 Microbiome Data 3
  - 2.1.1 OTU Clustering Methods 4
  - 2.1.2 Problems about OTU Clustering Methods 4
  - 2.1.3 Sequence Variants (SV) Methods 5

2.1.3.1 Dada2 5

2.1.3.2 Unoise2 6

2.1.3.3 Deblur 6

2.2 Statistical Challenges of Microbiome Data 6

2.3 Compositional Data Approach for Microbiome Analysis 7

2.3.1 Why Microbiome Data is Compositional 7

2.3.2 Challenges of CODA Approach on Microbiome Data Analysis 8

2.3.2.1 Normalization of Sequence Data 8

2.3.2.2 The problem of Zero Components 9

2.3.2.3 High Dimensionality 9

3 NETWORK INFERENCE FROM METAGENOMICS DATA 11

3.1 Visualization of Metagenomic Data 11

3.2 Microbial Dependency Measures 11

3.2.1 Correlation and Partial Correlation 12

3.3 Methods for Microbial Co-occurrence 14

3.3.1 CCLasso 15

3.3.2 REBACCA 15

3.3.3 CCREPE 15

3.3.4 SPIEC-EASI 15

3.3.5 SPARCC 15

3.3.6 CONET and MIC 16

3.3.7 LSA 16

3.3.8 PROXI 16

3.3.9 COAT 16

3.4 Methods for Microbial Co-exclusion 17

3.4.1 CO-EX 17

3.5 Multidimensional Boolean Patterns in Microbial Communities 17

3.6 SourceTracker 17

4 COMPOSITIONAL DATA ANALYSIS 19

4.1 What is Compositional Data? 19

4.1.1 Principles of Compositional Data 20

4.1.1.1 Principles of Scale Invariance 20

4.1.1.2 Principles of Permutation Invariance 21

4.1.1.3 Principles of Subcomposition coherence 21

4.2 Aitchison Geometry 21

4.2.1 Defining Simplex as a Vector Space. 22

4.2.2 Log Ratio Analysis: A Statistical Methodology for Compositional Data Analysis 24

4.2.2.1 Additive Log Ratio (alr) Transformation 25

4.2.2.2 Centered Log Ratio (clr) Transformation 25

4.2.2.3 Isometric Log Ratio (ilr) Transformation 26

4.2.3 Compositional Distance 26

4.3 Basis and Balances 27

4.3.1 Principle Balances 29

4.4 Exploratory Data Analysis of Compositional Data 30

4.4.1 Center, Variation Matrix and Covariance Structure of Compositional Data 30

4.4.2 Correlation Analysis of Compositional Data 31

4.4.3 Regression Analysis of Compositional Data 32

4.4.3.1 Case 1: Response is Real , Covariates are Compositional 32

4.4.3.2 Case 2: Covariates is real , Response are compositional 33

4.4.3.3 Case 3: Both Covariates and Response are compositional 33

4.4.4 PCA for Compositional Data 33

4.4.5 Biplot for Compositional Data 34

4.4.6 CODA Dendrogram 35

5 PRINCIPAL MICROBIOME GROUPS FOR BIOMARKER INVESTIGATION 37

5.1 Introduction 37

5.1.1 Compositional Data (CODA) Approach for Microbiome Data Analysis 38

5.2 MATERIALS AND METHODS 39

5.3 Overview of Principal Microbial Groups 39

5.3.1 (1) Select an appropriate SBP 40

5.3.2 (2) Choose the optimal number of PMGs 40

5.3.3 (3) Select Compositional Biomarkers 42

5.4 Dataset and Preprocessing 42

5.4.1 Benchmark Evaluation 43

5.4.2 Results 43

5.4.2.1 PMG Balances as Dimensionality Reduction Method 43

5.4.2.2 PMGs as Feature Aggregation Procedure 44

5.4.2.3 PMG Balances as Biomarker Candidates 47

Compositional Biomarker. 49

5.4.2.4 CODA Dendrogram to Discover Discriminatory Power of the Balances . 51
5.5 Discussion and Conclusion 53
5.6 Key Points 55
6 PRINCIPAL MICROBIAL GROUPS FOR MICROBIAL TRANSMISSION 57
6.1 Introduction 57
6.2 Nosocomial Infections and Indoor Microbiome 58
6.3 Microbial Transmission Modelling 59
6.3.0.1 How to Define Microbial Transmission ? 59
6.4 Controlled Experiment 60
6.4.1 Sampling 60
6.4.2 DNA isolation 61
6.4.3 16S rRNA Polymerase Chain Reaction and Sequencing 61
6.4.4 Otu Picking 63
6.4.5 Preprocessing Data 63
6.5 RESULTS 63
6.5.1 Grouping OTUs as PMGs 63
6.5.2 Hierarchical Clustering of Samples after PMG construction 64
6.6 Discussion and Conclusion 65
7 CONCLUSION AND FUTURE WORK 73
REFERENCES 75

APPENDICES

A PRINCIPAL MICROBIAL GROUPS : EXTRA MATERIAL 87

A.1 Alternatives to PMG evaluation 87

A.2 Stability of Principal Microbial Groups 88

A.3 Benchmarking Methods for Dimension Reduction 91

  A.3.1 PCA representation 91

  A.3.2 Principal Balance representation 91

  A.3.3 Distal Balance representation 92

A.4 Supplementary Tables 93

LIST OF TABLES

Table 1 Table for Observed Counts. Observed Library Size: n_j = 10 8

Table 2 Table for Absolute Counts. Absolute Library Size: N_j = 15 8

Table 3 Observed Counts. Library Size: S = 100 8

Table 4 Proportions of Observed Counts. Library Size: S = 1 8

Table 5 Microbial co-association network methods 14

Table 6 Classification performances of PMG balance combinations selected by balance selection methods. 49

Table 7 The explanations of the pvclust clusters....................................... 65

Table 8 Classification performances of reduced datasets processed with four dimensionality reduction procedures for disease prediction on the cirrhosis dataset. 93

Table 9 Selected balances by balance selection methods on different data types and AUC measures for the classification performance. 93

Table 10 Classification performances (AUC^1) of distal balances with different data types for disease prediction on the cirrhosis dataset. 94

LIST OF FIGURES

Figure 1 Underlying Network of Generated Data 12

Figure 2 The importance of choosing threshold fro network construction 12

Figure 3 Precision matrix to find out not correlated nodes. 13

Figure 4 Choosing the optimal threshold on inverse covariance matrix 13

Figure 5 Ternary Diagram. 22

Figure 6 Perturbation and power transformation effect in simplex. 23

Figure 7 Sequential Binary Partition Example 28

Figure 8 Ray and Link in biplot of a 5-part composition for 10 observation 34

Figure 9 CoDa Dendrogram of 5-part composition 35

Figure 10 A general workflow overview 41

Figure 11 Logistic regression classification performances of dimensionality reduction methods 45

Figure 13 The microbial content of balances selected by different methods 48

Figure 14 Form biplot of selected PMGs by balance selection methods 50

Figure 15 CODA dendrogram of PMGs and the association of species in PMGs with cirrhosis 52

Figure 16 The frequency of genera in each PMG 54

Figure 17 Experiment path, list of objects, and their respective locations in the ICU 62

Figure 18 Rarefaction curve 63

Figure 19 Coda Dendrogram of OTUs and constructed PMGs 66

Figure 20 The taxonomic content of PMGs. 67

Figure 21 Heatmap of PMGs abundance table. 68

Figure 22 Heatmap of OTUs abundance table. 69

Figure 23 Form Biplots on OTU and PMG tables 70

Figure 24 Graph representation of PMGs and connected OTUs 71

Figure 25 Hierarchical clustering via scale bootstrap resampling of samples 72

Figure 26 Quantiles of VD Aitchison norm and square-norm over the number of PMGs . . . 89

Figure 27 Quantiles of VD norm 89

Figure 28 Jaccard distances from the original PMGs to those identified in the 50 re-samplings represented as boxplots 90

CHAPTER 1

INTRODUCTION

Microorganisms are an essential part of life on the earth and can exist in association with virtually any living thing. Not only living things, microorganisms have been found in every part of the built environment such as in the air, on surfaces and on building materials [1]. The interaction between the environmental and human microbiome highly influences human health. The microbes present in the human body play a crucial role in maintaining human health, and the environmental microbiome influences the human microbiome [2]. Advanced understanding of the ecology of the indoor microbiota and human microbiome is the first step towards understanding potential relationships with health outcomes. Next Generation Sequencing (NGS) has led to an explosive growth of studies of microbiome at a very large-scale without requiring cultivation in vitro. NGS enables identification and study of a large number of microorganisms in a short time. With the identification of a large number of microorganisms, the studies for the understanding of their role in the environment and human health have become important. Microbiome-wide association studies have established that numerous diseases are associated with changes in the microbiota [3, 4]. Most of the methods proposed for microbiome analysis are intended to address two main issues: first, whether there is a global association between the microbiome and a phenotype of interest; second, which specific taxa are associated with the disease [5]. There has not been a proposed knowledge and techniques to reveal microbial transmission mechanisms. Transmission of microorganisms from reservoirs within the built environment to human occupants has historically focused on pathogens; however, microorganisms can be transferred to and from occupants and environmental reservoirs within buildings. Advanced understanding of the ecology of the indoor microbiota is the first step towards understanding potential relationship with health outcomes. Built environment microbiome analysis may help tracking biothreats and controlling hospital infections, and so developing early warning systems. Such studies generate large-scale high-dimensional count and compositional data, which are the focus of this dissertation.

Microbiome profiles are produced through sequencing specific genes (often the 16S rRNA gene) that provides diversity of bacterial taxa or shotgun metagenomics that provides further insights at the molecular level. Microbiome profiles are typically high-dimensional and very sparse, leading to two main problems in data analysis. The main approach to deal with these problems is to annotate microbial features with taxonomic information. The taxon grouping allows summarizing microbiome abundance with a coarser resolution in a lower dimension. The similarities or relationships between samples are addressed correspondingly. However, bacterial strains in the same taxonomic group have been found to vary in their relationships with the interested parameters, suggesting that each of them may have a distinct impact on the association [6]. Thus, correlating selected taxa with the parameters can often lead to controversial results in microbiome studies. If members in a taxon have opposite associations with the same parameter, lumping them into one taxon variable will produce degradation of the possible associations. On the other hand, microbial transmission is carried out between objects, so it is naturally expected that resulting microbiome profiles of samples should have similar microbial structure. In this case, not taxonomic changes but OTU/ASV abundance changes between samples need to be investigated.

1.1 Research Questions

Can CoDa make use of highest granularity of microbial genome features while overcoming high dimensionality?

Can CoDa use to develop any other grouping of microbial features without using a phylogenetic tree based on relative abundances?

1.2 Contributions of the Study

We propose a procedure that groups microbial features attending the compositional character of the data making use of the highest possible resolution of microbial features (OTUs). This mathematically consistent aggregation procedure collapses microbial features into units as an alternative to taxon grouping, here called Principal Microbial Groups (PMGs). The procedure reduces the need for user-defined aggregation and offers the possibility of working with coarse group of OTUs, which are not present in a phylogenetic tree.

1.3 Organization of the Thesis

The first part of the this thesis presents the background information of how microbiome data is produced and the statistical challenges of the microbiome data on the way of translating research to clinical practice and also presents a review of the network inference techniques from microbiome data.

The second part of the thesis focuses on the review of compositional data (CoDa) analysis.

The third part of this thesis deals with the problem of establishing relationship based on the microbial features annotated with taxonomic information, where a compositional alternative to phylogenetic grouping of microbiome data (Principal Microbial Groups) is proposed to enable working with low-level microbial features (OTUs or ASVs). The usefulness of the proposed procedure is illustrated on Cirrhosis dataset to search for biomarker candidates.

The fourth part of the thesis focuses on the analysis of microbial transmission. An experiment was conducted in the Erciyes University Hospital for this purposes, and swab samples were gathered from an Intense Care Unit (ICU) to construct microbiome profiles. Consecutively contaminated objects were analyzed using Principal Microbial Groups.

CHAPTER 2

MICROBIOME DATA AND COMPOSITIONALITY

In this chapter, the details of the how microbiome data are produced and the statistical challenges of the microbiome data are presented.

2.1 Microbiome Data

Microorganisms are an essential part of life on the earth and can exist in association with virtually any living thing as well as every part of the built environment [1]. The microbiome is defined as a collection of microorganisms.

The first step of producing microbiome data is gathering swap samples and sample sequencing. Next Generation Sequencing (NGS) has led to an explosive growth of studies of microbiome in very large-scale without cultivation in vitro. NGS enables identification and study of a large number of microorganisms in a short time. Two main approaches are there to produce microbiome profiles: amplicon sequencing and shotgun sequencing. Amplicon sequencing relies on sequencing a phylogenetic marker gene after polymerase chain reaction (PCR) amplification [7]. For bacteria and archaea, the marker gene is the 16S ribosomal RNA gene. The 16S rRNA gene contains both highly conserved areas and hypervariable sites, denoted as V1–V9. The conserved regions can be targeted with PCR primers while the hypervariable regions are specific to each microbial species and make possible to distinguish the different microbes [7]. The V1–V3 and V4 regions are most commonly targeted. PCR amplification creates thousands to millions of copies of the DNA target region, called amplicons. PCR amplicons are then sequenced using HTS platforms and multiple nucleotide sequences, also known as reads, are obtained [8]. Shotgun is an untargeted sequencing method that extracts all genomic material for microbial community classifications and gene annotations [9]. The result of 16S and shotgun sequencing is a virtual “library” of many short sequence fragments.

The second step is sequence processing. Bioinformatics pipelines are available for processing microbiome sequence data (i.e mothur[10] and QIIME[11]). The bioinformatics pipeline consists of five main steps: Preprocessing and quality control filtering, operational taxonomic unit (OTU) binning, taxonomy assignment, construction of the abundance table and phylogenetic analysis. Preprocessing and quality control filtering consists on first assign the sequences to samples (demultiplexing) and then sequences are quality filtered to remove too short sequences, too many ambiguous base pairs and chimeras [7]. Then, OTUs are constructed. Currently two different approaches exist for OTU construction: Clustering similar sequence fragments into OTUs and Sequence Variants (SV) methods. OTUs/SVs are the minimum unit for microbiome studies for downstream analysis. Taxonomy assignment is then obtained by comparing OTU sequences to microbial reference databases such as GreenGenes (http:// greengenes.second.genome.com) for 16S and Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/pathway.html) for shotgun sequencing. Then, sequence data can be represented as an abundance table of counts representing the number of sequences per sample for a specific taxon [7, 12].

2.1.1 OTU Clustering Methods

The analysis of microbiome data begins with the construction of OTUs. Picking OTUs is basically "clustering" reads. All the sequences are clustered into OTUs based on a distance matrix at a specified threshold [13, 14]. For each cluster a "representative sequence" is determined and it is called OTU representative. OTU picking procedure outputs a “otu table” which is an abundance table of counts of OTU representatives for each sample.

Three different methods for OTU picking is proposed: denovo, closed and open-reference. Closed reference method compares reads to a sequence in a reference database and recruites into a corresponding OTU. Denovo method clusters reads into OTUs as a function of their pairwise sequence similarities. Open reference method is combined of closed and denovo methods. First it clusters sequences against a database of references sequences, then uses denovo clustering on those sequences which are not similar to any reference sequences. Which method to choose is depend on what is known about the microbiome community prior. If the studied microbial community is well studied, then reference databases have many representatives and closed reference otu picking strategy is suitable. Denovo method is suitable to discover new species.

Distribution-based operational taxonomic unit (dbOTU3) is another OTU picking method which is different than other OTU picking strategies. It takes account distribution of sequences across samples. This allows to distinguish ecologically-distinct but sequence-similar organisms [15]. For example, OTU methods would likely group two sequences in the same OTU if they differ by only one nucleotide. However, if the two sequences never appeared together in the same sample, an observer would probably conclude that that one nucleotide difference corresponds to two distinct groups of organisms, one which lives in one group of samples, the other living in the other [15].

2.1.2 Problems about OTU Clustering Methods

The main problem with OTU base methods are clustering multiple different sequences in the same cluster. However, even a single nucleotide change of with in a gene sequence might lead to a different organism. Moreover, NGS of the 16S rRNA gene on Illumina instruments is commonly used to identify taxa present in a given sample, but suffers from an error rate of 0.1% per nucleotide [16]. In such experiments, sequence errors caused by PCR and sequencing are difficult to distinguish from true biological variation. The classic approach to overcome sequencing errors is to cluster amplicon sequences into OTUs based on an arbitrary sequence identity threshold. This approach reduces problems caused by erroneous sequences but also reduces phylogenetic resolution because sequences below the identity threshold cannot be differentiated [17].

On the other hand, OTU based methods create obstacles discovering new species due to limits of reference databases and comparison of different microbiome studies. Closed reference OTU picking strategy clusters reads using a reference database if the reads sufficiently similar to a reference sequence. Similarity is determined based on a threshold such as 97% so, for example one base change is not considered as biological variation even if the reads were from two different species. Moreover, OTUs from two different dataset can be compared only if the same reference databases are used. However, biological variation that is not represented in the reference database is lost during assignment to closed reference OTUs [18]. Denovo OTU picking strategy clusters reads together that are similar to one another. Obtained denovo OTUs are directly depend on the dataset and this dependency does not allow comparison of denovo OTUs in two different dataset. dbOTU3 method tries to brings a solution to the problem of closed reference otu picking, but resulting OTUs only depend on the dataset and it also does not allow OTU comparison between different datasets [15].

Sequence Variants (SV) method brings a solution to all those problems. Each SVs corresponds a biological variation which is independent from the processed data and obtained SVs can be compared between different samples.

2.1.3 Sequence Variants (SV) Methods

The goal of SV methods is to infer accurate biological template sequences from noisy reads. SV methods infer the biological sequences in the sample from errors on the basis of the number of repeated observations of distinct sequences. SV methods can distinguish SVs differing by as little as one nucleotide.

Algorithms such as Deblur [17], DADA2 [18] and UNOISE2 [19] use error profiles to resolve sequence data into exact sequence features. This task is generally divided into two phases. First; correcting point errors to obtain an accurate set of amplicon sequences (denoising). Second; filtering of chimeric amplicons [19]. The result is a set of predicted biological sequences which three method calls them with different terminology; DADA2 calls as "sequence variants", Deblur calls as "sub-OTUs" and UNOISE calls as "zero-radius OTUs zOTUs)".

Oligotyping [20] is another method to resolve SVs and it improves traditional OTU picking by including position-specific information from 16S rRNA sequencing to identify subtle nucleotide variation and by discriminating between closely related but distinct taxa.

The resulting output from SV methods is a table of DNA sequences rather than OTU groups and counts of these different sequences per sample. SVs are reusable, reproducible, and comprehensive [18]. The most important opportunity of SV methods is that each SVs corresponds a biological organism which is independent from the processed data and obtained SVs can be compared between different samples [18]. Recent literature on microbiome analysis recommend that SV methods should replace OTU-based approaches for all applications [18, 21].

2.1.3.1 Dada2

DADA2 (Divisive Amplicon Denoising Algorithm) is a divisive partitioning algorithm to infer sample sequences by correcting amplicon errors that incorporates quality information without constructing OTUs [18]. DADA2 enables a complete pipeline produces merged, denoised, chimera-free Svs. Reads with the same sequence are grouped into unique sequences with an associated abundance. "The abundance p-value" is calculated for each unique sequence. This p value is used to determine unique sequences that can not be explained by errors in amplicon sequencing. Singletons have an abundance p-value of 1. A low p indicates that there are more reads of the sequence than can be explained by errors introduced during the amplication and sequencing. If the smallest p-value falls below the threshold, a new partition is formed with that unique sequence as its center. Unique sequences are then allowed to join the partition most likely to have produced them. Division continues until all unique sequences are consistent with being produced as errors from the sequence at the center of their partition. In other words, division continues until all abundance p-values are greater than the threshold [18].

2.1.3.2 Unoise2

UNOISE2 clusters the unique sequences in the reads. Input to the UNOISE2 algorithm is the set of unique read sequences with abundance bigger than a threshold γ where $\gamma = 4$ by default. Unique reads with low-abundance are discarded because they are more likely contain errors that are reproduced by chance or bias [19]. A cluster has a centroid sequence with higher abundance and has members that are similar sequences with lower abundances. Members are inferred to be reads of the same centroid sequence containing one or more point errors.

Let C be a cluster centroid sequence with abundance a_C and M be a member sequence of that cluster with abundance a_M. Let d be the Levenshtein distance (number of differences including both substitutions and gaps) between M and C. The abundance skew of M with respect to C is defined to be $skew(M, C) = \frac{a_M}{a_C}$ [22]. Sequences are considered in order of decreasing abundance. A sequence (Q) is assigned to cluster C if $skew(Q, C) \leq \beta(d)$. If no such C exists, Q becomes a new centroid. The final set of centroids are reported as the predicted amplicons [19].

2.1.3.3 Deblur

DEBLUR is a greedy deconvolution algorithm based on Illumina error profiles [17]. DEBLUR produces sub-OTUs (called sOTUs). Deblur algorithm first sorts sequences by abundance. Second, from the most to least abundant sequence, the number of predicted error-derived reads is subtracted from neighboring reads based on their Hamming distance, using an upper bound on the error probability. Finally, sequences whose abundance drops to 0 after subtraction are removed. After applying DEBLUR, only reads likely to have been presented to the sequencer are retained. However, it is possible that the reads would still contain chimeras originating from PCR. Reads are filtered for denovo chimeras using UCHIME [19] as implemented by VSEARCH [23] using modified parameters [17].

Unlike DADA2 and UNOISE2, DEBLUR operates on each sample independently [17]. DADA2 and UNOISE2 algorithm can not be applied only a read, at least a sample data is necessary to infer SVs [18].

2.2 Statistical Challenges of Microbiome Data

The microbiome data is produced as a result of sample sequencing and constructing OTUs or inferring SVs. The abundance of OTUs/SVs is quantified by sorting and counting the DNA fragments in each sample. Resulting count data are typically high-dimensional and very sparse, which are two main problems in microbiome data analysis.

One of the available approaches to deal with the above mentioned problems is to annotating constructed OTUs/SVs with taxonomic information using databases and agglomerating taxa to any rank. Representative sequences are classified taxonomically via alignment against a database of previously characterized reference sequences. Agglomerating taxa to any rank (phylum, genus etc.), microbiome abundance is summarized with a coarser resolution in lower dimension. The similarities or relationships between samples are addressed correspondingly. This approach helps to reduce dimensionality as well as sparsity.

On the other hand, microbiome profiles are usually represented using relative abundances of the observed OTUs/SVs. Relative data harbors the relationships between the features in the dataset [24, 25, 5]. Relative data needs to be analyzed carefully because it is compositional. Widely used statistical methods are not valid on compositional data such as evaluating correlations might not capture the structural relations [26]. Recent awareness of considering microbiome data as compositional data, Compositional Data Analysis (CoDA) approach has been started to be employed on microbiome studies.

2.3 Compositional Data Approach for Microbiome Analysis

2.3.1 Why Microbiome Data is Compositional

In metagenomics, the abundance of genes is quantified by sorting and counting the DNA/RNA fragments. The resulting count data is high-dimensional and affected by high levels of technical and biological noise that make the statistical analysis challenging [27]. After aligning the sequencing reads to the reference microbial genomes, the observed count data usually depend on the true underlying composition of microbiome genomes, amount of genetic material extracted from the community, and the sequencing depth [28]. In order to account for the large variability in the total number of sequencing reads across different subjects, the observed counts are often normalized to a relative measure of abundances rather than absolute counts, which yields the compositional data [29]. The compositional data lives in the simplex, not in Euclidean space, so many data analyses, including distance measures, correlation coefficients, and multivariate statistical models turn invalid in simplex [30].

For the compositional data, it is not a requirement for the arbitrary sum to represent complete unity. Microbiome abundance data is lack information about potential true components and hence exist as incomplete compositions [30]. Although, microbiome abundance data have compositional properties, but differ slightly from the formally defined compositional data in that they contain integer values only [30]. Eventually the individual values of the observed counts are irrelevant. The only thing that is accessible is the relationships between points, which is their ratios. As a result, analyzing microbiome data is actually analyzing the "relative abundance" data.

Let consider a $n \times d$ microbiome data matrix where d genes (SVs or OTUs) correspond to the columns and the n (multivariate) samples are displayed in the rows $(n < d)$. Samples are collected from different objects and for each sample the number of gene values sum to a constant that is unrelated with the absolute amount of genes in the object of origin. Each sample, which is the each row in the data matrix, is considered as a composition of genes. A row is denoted by a vector x whose elements x_is are the number of genes extracted from the samples for genes i=1,. . . ,d (count number in OTU table).

Consider the following observed and absolute counts for 3 genes.

	Gene1	Gene2	Gene3
sample 1	5	3	2

Table 1: Table for Observed Counts. Observed Library Size: n_j = 10

	Gene1	Gene2	Gene3
sample 1	8	4	3

Table 2: Table for Absolute Counts. Absolute Library Size: N_j = 15

Let $x_i = a_i/s$ with a_i denoting the absolute gene amount from gene i and s the total gene amount is $s = \sum_{j=i}^{d} a_j$ [31]. We neither know s nor the a_i. But the number of genes ratios, $x_i/x_j = a_i/a_j$ is the only information maintained from the original absolute amounts thus, constant gene ratios can be correctly inferred even on relative data. Taking the log of these ratios makes them symmetric with their reciprocal values [15].

The approach to compositional data analysis originated by John Aitchison uses ratios of parts as the fundamental starting point for description and modeling [32].

2.3.2 Challenges of CODA Approach on Microbiome Data Analysis

2.3.2.1 Normalization of Sequence Data

The simplest normalization would involve rescaling counts by the library size (i.e. the total number of observed reads from a sample)[33]. It basically divide observed count by the library size. This rescaling also solve the problem of uneven sequencing depth. Consider the following normalization example for a sample with 3 genes. Eventually, calculating proportions does not transform compositional counts into absolute counts.

-	Gene1	Gene2	Gene3
sample 1	50	30	10

Table 3: Observed Counts. Library Size: S = 100

-	Gene1	Gene2	Gene3
sample 1	0.5	0.3	0.1

Table 4: Proportions of Observed Counts. Library Size: S = 1

The methods such as EdgeR[34] and DESeq[35] tries to estimate absolute counts considering data distribution, but they were were criticized that if counts were evaluated relative to absolute, then the original absolute count data was recovered after normalization. This means that closed data was converted to "open" data which is not realistic since the microbiome data originally produces closed data by its default [30].

Normalization attempt is the initial step for data analysis and the choice of normalization method eventually impacts the downstream analysis and so the final result. Since current normalization methods are controversial, so avoiding normalization would seem desirable for microbiome data [30, 24]. Thus the relative abundance of OTUs are the main data to analyze. Considering CoDA approach, working with ratios of components led to working with logarithms of ratios since logarithms of ratios are mathematically easier to control than ratios [32]. Compositional data lives in a simplex, but Aitchison presented that compositional data could get mapped into real space by log-ratio transformation so that Euclidean distance become meaningful [32]. Thus, many compositional data analysis begin with transformation instead of normalization that is suitable for microbiome data.

2.3.2.2 The problem of Zero Components

Composition components are non-negative, but zeros in composition makes statistical analysis difficult. Because of compositional data analysis relies on log-transformation, zeros may cause difficulty in downstream data analysis [32] and lead to biased estimate of microbial diversity.

A common technique to handle zeros is to replace zeros with a small value. Another replacement strategies is adding a fixed positive value to all components, so that zeros replaced with the fixed positive value, but of a pseudo-count addition to all components does not preserve the ratios between components [36]. Mathematical models have been employed for replacing zeros. A model that is based on the Dirichlet sampling procedure is developed to replace zeros [37]. Fernandez et al. [38] introduced a bayesian multiplicative model for estimating non-zero compositions from count data. ALDEx2[39] package available on Bioconductor uses a Dirichlet-multinomial model to infer abundance from counts. zCompositions[40] R package is available dealing to estimate zero count values using bayesian multiplicative model.

2.3.2.3 High Dimensionality

It is current practice in microbiome studies to filter out rare OTUs across all samples. The OTU abundance threshold for filtering depends on the user choice. Filtering could help to reduce dimension but it might not be enough for large-scale microbiome dataset.

Another commonly used approach is to use phylogenetic information of OTU. OTUs are assigned to a taxonomic unit using reference databases so that OTUs are grouped by a taxonomic level (i.e phylum, genus) so dimension is reduced. Most of the microbiome studies to investigate the microbial structure of the environment, the relationships between samples, the differential abundance of microorganisms between samples or the microbial difference after a treatment etc. have been performed using the phylogenetic information of the samples.

On the other hand, there are three popular dimension reducing procedures in CoDA: principal component analysis, factor analysis and subcompositional analysis.

Principal components analysis and factor analysis are widely used dimension reduction techniques in Euclidean space. The goal of principal component analysis is to convert possibly correlated original variables from the data into a smaller set of linearly uncorrelated variables called principal components [41]. The goal of factor analysis is to extract a few directions in the data, called the factors or latent variables. Thus factor analysis and principal components reduce the data dimensionality, and therefore aim at summarizing the multivariate information in a compact form[42]. When applying principal component analysis and factor analysis to compositional data, it is crucial to apply an appropriate transformation.

In compositional data, each dimension called as a "part" and groups of parts can be viewed either as a subcomposition or as a group inside the whole composition. Subcompositional analysis is intended to deal with parts within the group and relations with respect to other groups or parts [43]. The goal of grouping parts is to reduce dimension to facilitate interpretation. The summation of parts, called amalgamation, is an easy and apparently intuitive way of grouping parts and a practical way of reducing dimensionality [44]. But, amalgamation introduce a non-linear distortion to the data and amalgamation of parts changes the original problem, thus cannot be considered as a compatible reduction of dimension [43]. The balances are introduced as an alternative to amalgamation and balances have recently become popular for the analysis and classification of microbiome compositions. The main goal of balances is to identify a complete orthonormal basis of the simplex and to make the corresponding coordinates directly interpretable between two groups of parts [43]. The resulting procedures provide tools that improve interpretability and can also be used for an intuitive dimension reduction [45].

CHAPTER 3

NETWORK INFERENCE FROM METAGENOMICS DATA

In this chapter, an overview of the microbial network inference and microbial association methods are summarized.

3.1 Visualization of Metagenomic Data

Due to the compositional nature and the extreme sparsity of the metagenomic data, inferring association relationships is very challenging. One way to begin exploring such large data sets such as microbiome data is to set up a network among data points. Network inference is being applied to studies of microbial ecology to visualize and characterize microbial communities [46]. Graph-based representation of metagenomic data is a promising direction for analyzing microbial interactions [47].

The first step to construct a network is searching for pairs of variables that are closely associated, then calculating some measure of dependence for each pair, rank the pairs by their scores, examine the top-scoring pairs and draw a network considering those pairs. However, the determination of dependency in microbiome data is not an easy. Recent studies have demonstrated that the microbiome composition varies across individuals due to different health and environmental conditions [48].

The compositional nature of the data complicates the investigation of the dependency structure since there are no known multivariate distributions that are flexible enough to model such a dependency [49]. But, understanding the dependence structure among microbial units within a community, including co-occurrence and co-exclusion relationships between microbial units might help to model the behavior of community and may help answer many questions.

3.2 Microbial Dependency Measures

Dependency measures between between data points can be distance, similarity, dissimilarity, correlation and partial correlation.

Distance metrics need to satisfy metric axioms (minimality, symmetry and triangle inequality). Dissimilarity is similar to distance, but it does not need to satisfy triangle inequality. Dissimilarity functions are usually in the range of [0,1]. Similarity is the complement of the dissimilarity measured in the range of [0,1].

Using the formula, $d = \sqrt{1-s}$ where s is similarity and d is Euclidean distance, dissimilarity value can be converted to similarity,s. If necessary, the correlation values also can be transformed into Euclidean distances using the same equation.

3.2.1 Correlation and Partial Correlation

Networks inference algorithms frequently use correlation. If two nodes have higher correlation more than a pre-determined threshold, then an edge is drawn between those nodes and network is constructed. Partial correlation is another measure of the relationship between two continuous variables while controlling for the effect of one or more other variables.

In order to demonstrate the network construction procedure and differences between correlation and partial correlation, let's generate four correlated points such as:

$$x1 \sim N(0,1), x2 \sim N(2 * x1 + 1, 1), x3 \sim N(3 * x2, 1), x4 \sim N(0.5 * x3, 1)$$

It is clear that the underlying network of the generated data is as in the figure 1:

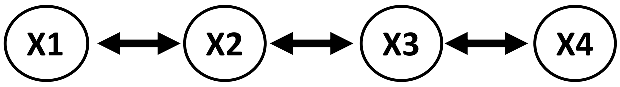
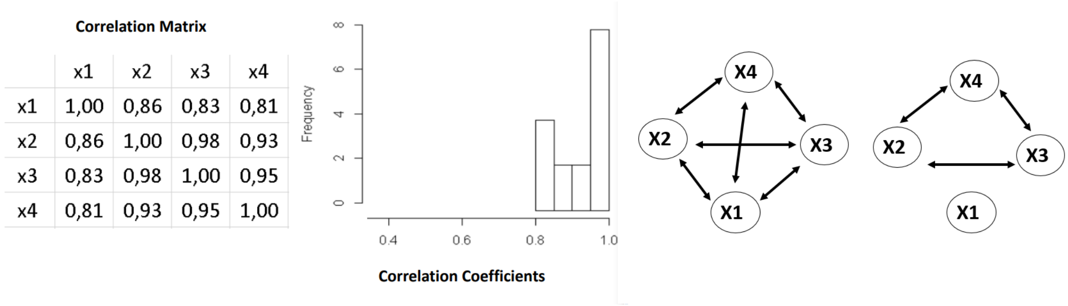


Figure 1: Underlying Network of Generated Data

Calculating the pairwise correlation between points, how to link points can be determined. Choosing threshold is crucial for network construction (see figure 2). If threshold is too high, connections might be lost between points, or if it is too low, then extra connections which do not exist in the real network might be seen in the network.



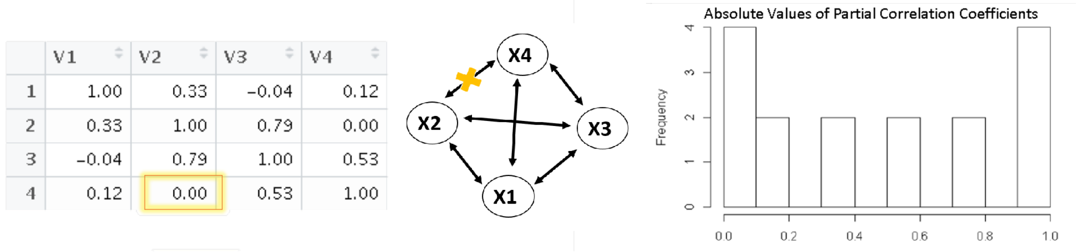
Correlation Matrix

	x1	x2	x3	x4
x1	1,00	0,86	0,83	0,81
x2	0,86	1,00	0,98	0,93
x3	0,83	0,98	1,00	0,95
x4	0,81	0,93	0,95	1,00

Figure 2: Choosing threshold is the crucial for network construction. Different thresholds creates different connected graphs.

Using reverse thinking, determination of "not correlated nodes" can help for network construction. If correlation is zero between two nodes, then nodes are linearly independent, hence there will be no edge in the network graph.

Partial correlation can be more helpful than correlation in finding uncorrelated nodes. Partial correlation measures the relationship between two variables removing other variables' effect. The partial correlation matrix is also called precision matrix. If partial correlation between nodes is zero, then two nodes are conditionally independent. In the network, there will be no edge between those nodes (see figure 3).

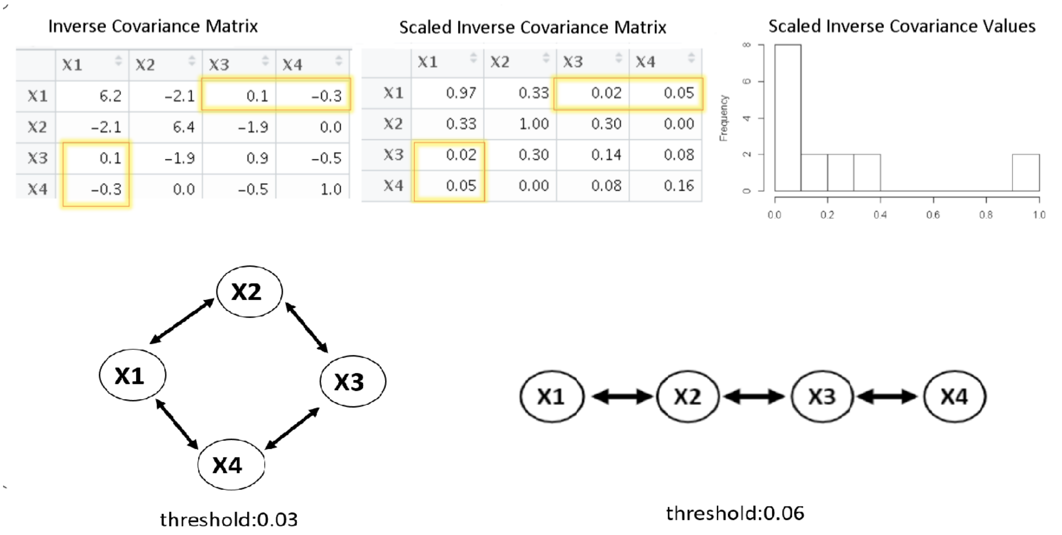


	V1	V2	V3	V4
1	1.00	0.33	-0.04	0.12
2	0.33	1.00	0.79	0.00
3	-0.04	0.79	1.00	0.53
4	0.12	0.00	0.53	1.00

X4
X2
X3
X1
Absolute Values of Partial Correlation Coefficients
Frequency
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3: Precision matrix to find out not correlated nodes.

Mathematically, the inverse covariance matrix is referred to as precision or partial correlation matrix. Inverse covariance matrix is also used for network inference and it is known in the literature as Graphical Gaussian model. Inverse covariance values yields to zero more than correlation and partial correlation and choosing the optimal threshold on inverse covariance matrix might reveal the true underlying network for data points (see figure 4).



Inverse Covariance Matrix

	X1	X2	X3	X4
X1	6.2	-2.1	0.1	-0.3
X2	-2.1	6.4	-1.9	0.0
X3	0.1	-1.9	0.9	-0.5
X4	-0.3	0.0	-0.5	1.0

Scaled Inverse Covariance Matrix

	X1	X2	X3	X4
X1	0.97	0.33	0.02	0.05
X2	0.33	1.00	0.30	0.00
X3	0.02	0.30	0.14	0.08
X4	0.05	0.00	0.08	0.16

Scaled Inverse Covariance Values
Frequency
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
X2
X1
X3
X4
X1
X2
X3
X4

threshold:0.03

threshold:0.06

Figure 4: Inverse covariance values yields to zero more than correlation and partial correlation and Choosing the optimal threshold on inverse covariance matrix might lead the true underlying network.

Methodology	Underlying Metric
Correlation And Regression-Based (CREEPE and CCLasso)	Correlation between pairwise similarities or incidences
REBACCA, COAT	Covariance estimation on compositional data
Graphical Model Inference (Spiec-easi, Proxi)	Conditional independence on compositional data
Local Similarity Analysis (LSA)	Local alignment on time series
Bayesian Networks (SparCC)	Multivariate probability distribution-Conditional independence
Mutual Information (CoNet, MIC)	Mutual dependence between variables-Entropy

Table 5: Microbial co-association network methods

3.3 Methods for Microbial Co-occurrence

Microbial co-occurrence means that two types of microorganism like to live together and usually they are seen together in an environment. If one microorganism has high abundance in an environment, then other one also expected to have high abundance or versa versa.

Presence-absence or abundance data is used for prediction of microbial association networks. These kind of problem is known as network inference in computer science and these techniques are widely used in genomics [50].

Similarity based network inference considers similarity of two species distribution to assess the co-occurrence patterns of two species over multiple samples. Significance of similarity assessed for all pairwise relationships by correlation. However, it is not clear that correlation is the proper measure of association. For example, correlations can arise between OTUs that are indirectly connected in an ecological network [51]. Moreover, in real life, species generally depend on multiple other species, and so pairwise relationship cannot be used model complex systems.

Regression can predict the abundance of one species from the combined abundances of other organisms however this prediction does not always have a biological meaning [50].

Mutual information approaches are also useful for identifying non-random co-association patterns[50, 52].

For more complex relationship inference association rule mining technique can be adopted [50]. Inferred relationships can be represented with network graph as a node refers to a microbial unit (otus, genes, taxa etc.), edge and directed edge refer to relationship and its direction. These graphs can be very complex and visualization of the also another issue. In order to model and visualize complex relationships among microbial communities, more research need to be conducted [50].

Table 5 summaries the microbial co-association network methods and underlying metric on these methods.

3.3.1 CCLasso

Correlation inference for Compositional data through Lasso (CCLasso) [53] uses the log ratio transformation for raw compositional data to infer the correlations among microbes through a latent variable model.

3.3.2 REBACCA

Regularized estimation of the basis covariance based on compositional data (REBACCA) [54] is an algorithm to identify significant co-occurrence patterns by finding sparse solutions to a system using log ratios of count or proportion data.

3.3.3 CCREPE

CCREPE method (Compositionality Corrected by Renormalization and Permutation) [55] determines the significance of association between features in a composition, using any similarity measure (e.g. Pearson correlation, Spearman correlation, etc.). They also implemented the NC-score similarity measure between compositions derived from ecological relative abundance measurements. It calculates Kendall's τ on binned data instead of ranked data. In such cases, features typically represent species abundances, and the NC-score discretized these continuous values into one of N bins before computing a normalized similarity of co-occurrence or co-exclusion.

3.3.4 SPIEC-EASI

Sparse Inverse Covariance Estimation for Ecological Association Inference (SPIEC-EASI) is a graphical model inference method for the inference of microbial ecological networks [51]. SPIEC-EASI method aims to learn a network of pairwise taxon-taxon association from microbiome compositions. SPIEC-EASI leveraged CoDA theory and used the centered log-ratio transform to reconstruct microbial association networks and interactions. SPIEC-EASI addresses interdependence through a centered log ratio (clr) transformation of the relative abundance data and then estimates the sparse inverse covariance matrix, therefore inferring association based on conditional independence. This method is fundamentally distinct from other techniques (CCREPE) which essentially estimate pairwise correlations.

3.3.5 SPARCC

Bayesian networks can capture the conditional interdependence between OTUs and can deal with interactions within complex microbial communities. SparCC uses Bayesian estimate [46]. SparCC estimates the linear Pearson correlations between the log-transformed components [56]. It makes two assumptions: the number of features is large and the correlation network is sparse. The SparCC method uses Bayesian estimates but calculates a mean value of a measure similar to the concordance correlation coefficient [56].

SparCC and SPIEC-EASI algorithms both assume an underlying sparse network and so are less rigorous for estimating correlations in compositional data than is the calculation of phi. However, they both offer the advantage of using a full or partial Bayesian approach, which is generally more powerful than point-estimate based approaches [25].

3.3.6 CONET and MIC

CoNet is an ensemble based network reconstruction method that detects non-random patterns of microbial co-occurrence between OTUs by combining multiple association methodologies (such as Kullback–Leibler divergence, Pearson correlation and Spearman correlation as well as mutual information) simultaneously to identify the highest scoring pairwise relationships and merges the results into a consensus network structure [46].

MIC (Maximal information coefficient) is a measure of the strength of linear or nonlinear associations between variables via mutual information. This is a nonparametric, exploratory statistical approach to identify novel interactions from a large dataset [52]. MIC and CoNet methods, however, lack the ability to discriminate against intuitively difficult to interpret patterns, can miss some important relationships [57].

3.3.7 LSA

the Local Similarity Analysis (LSA) method captures local and potentially time-delayed co-occurrence and association patterns in time series data that cannot otherwise be identified by ordinary correlation analysis. It can be applied to identify shifts in the abundance of a target OTU in response to a change in the composition of another OTU (or set of OTUs) or an environmental condition [58].

3.3.8 PROXI

Proxi [47] constructs a proximity graph from the abundances of microbial operational taxonomic units (OTUs). Proxi learns a proximity graph that each node is an OTU and edges represent proximity relationships between nodes. This tool supports three types of proximity graphs: k-nearest neighbor (k-NN) graphs; radius-nearest neighbor (r-NN) graphs; and perturbed k-nearest neighbor (pk-NN) graphs.

3.3.9 COAT

COmposition-Adjusted Thresholding (COAT) method [49] is to estimate the sparse covariance matrix of the latent log-basis components. The method is based on a decomposition of the variation matrix into a rank-2 component and a sparse component. The resulting procedure can be viewed as thresholding the sample centered log-ratio covariance matrix and hence is scalable to large covariance matrix estimations based on compositional data.

3.4 Methods for Microbial Co-exclusion

Microbial co-exclusion means that two types of bacteria do not like to live together and usually they are not seen together in an environment. In ideal case, one microorganism can be present in any abundance only if the other microorganism is absent. In not ideal cases, if one microorganism has high abundance then the other one expected to have low abundance or versa versa. Or both of them are expected to have low abundance.

Co-exclusion is one of the most important patterns to be identified in microbial communities. Knowing which microorganisms are unable to tolerate each other's presence or can replace one another in the community opens an opportunity to manipulate and control the microbiota, guide microbiota transplantation, and personalize the choice of microorganisms for probiotic treatments [57]. Note that mutual exclusion/avoidance pattern is not anti-correlation (negative Pearson or Spearman correlation).

3.4.1 CO-EX

Co-Ex introduce a quantified definition of the strength and statistical significance of multidimensional co-exclusion patterns between variables describing microbial communities [57].

Co-Ex formulates co-exclusion relationship on ideal case that is one microorganism is present in any abundance and the other microorganism is absent. Co-exclusion represented as the function $X_a X_b = 0$ where X_a and X_b are abundances of two microorganism. The coefficient of determination R^2 is redesigned for co-exclusion function and used to quantify goodness of fit.

3.5 Multidimensional Boolean Patterns in Microbial Communities

The microbial community members are dynamic and they are often simultaneously involved in multiple relations. Such relationships are very hard to detect using traditional correlation, mutual information, principal coordinate analysis, or covariation-based network inference approaches. Recently, a novel pattern-specific method to quantify the strength, and to estimate the statistical significance of two-dimensional co-presence, co-exclusion, and one-way interaction (organism 1 was needs organism 2 to survive and vice versa) patterns between abundance profiles of two organisms are proposed [59]. The basic idea of the proposed approach is to estimate the pattern score by counting the fraction of observations belonging to the pattern under investigation. The approach searches for Boolean patterns in the microbial abundance data and the search is pattern specific. The result is the presence of multidimensional patterns in microbial communities and multilayer networks are used to visualize these multidimensional patterns.

3.6 SourceTracker

Sourcetracker is a technique used to identify the ecological source of microbiomes [60]. It uses a Bayes approach and is adopted from Latent Dirichlet Allocation which was originally used in the natural language processing domain to identify topics of text [61]. Sourcetracker is used to identify the components constituting the habitats (skin, fecal, oral, soil, ocean etc.) of a microbiome. The SourceTracker algorithm needs a OTU table (sample x OTU count matrix), and a metadata file for samples. Samples are marked as “source” and “sink” in the metadata file and algorithm trained with sources and tested with sinks. The approach models contamination as a mixture of entire source communities into a sink community, where the mixing proportions are unknown [60]. The result is the predicted proportions of source samples for each sink sample. SourceTracker considers each sink sample x as a set of n sequences mapped to taxa, in which each sequence can be assigned to any one of the source environments $v_1, \ldots, V_n$, including an unknown source. When part of a sink sample is unlike any of the known sources, it gets assigned to an unknown source.

CHAPTER 4

COMPOSITIONAL DATA ANALYSIS

In this chapter, an overview of the basics of Compositional Data (CoDa) Analysis is summarized.

4.1 What is Compositional Data?

Compositions describe parts of a whole and carry relative information. Examples of compositional data include anything measured as a percent or proportion [43]. The formal definition [62] is as follows:

Definition 1 *Compositional Data. A row vector, x = [x1, x2,..., xD], is defined as a D-part composition when all its components are strictly positive real numbers and they carry only relative information.*

The most common examples of compositional data have a constant sum κ and it is known in literature as closed data [63]. κ might be 1 , 100 or any number. κ is called closure constant. Sum of components of the composition is κ.

Definition 2 *Closed Data. The sample space of compositional data is the simplex, defined as*

$$S^D = x = [x_1, x_2, \ldots, x_D] such as x_i > 0, i = 1, 2, \ldots . D and \sum_{i=1}^{D} x_i = 1$$

The closure is a projection of a point in R^D on S^D.

Definition 3 *Closure of Data. For any vector of D real positive components*

$$z = \{z_1, z_2, \ldots, z_D\} \in R_+^D$$

for all i=1,2,. . . ,D).

The closure of z is defined as

$$C(z) = \{\frac{\kappa.z1}{\sum_{i=1}^{D} z_i}, \frac{\kappa.z2}{\sum_{i=1}^{D} z_i}, \ldots, \frac{\kappa.zD}{\sum_{i=1}^{D} z_i}, \}$$

If only some parts of the composition are needed or available, subcomposition of data needs to be defined.

Definition 4 *Subcomposition of Data. Given a composition x, a subcomposition* x_s *with s parts is obtained applying the closure operation to a subvector* $\{x_i1, x_i2, ..., x_is\}$ *of x.*
Subindexes $i_1, ..., i_s$ *tell us which parts are selected in the subcomposition, not necessarily the first s ones.*

4.1.1 Principles of Compositional Data

Aitchison [32] introduces three important principles of compositional data analysis: scale invariance, subcompositional coherence and permutation invariance.

4.1.1.1 Principles of Scale Invariance

The principle of "scale invariance" states that compositional data only carry "relative information". The analysis should not depend on the closure constant κ. The difference between component values is only meaningful proportionally [64]. For example, the difference between 100 and 200 counts carries the same information as the difference between 1000 and 2000 counts. Thus, proportional vectors with positive components are compositionally equivalent as composition. Nevertheless, when it comes to interpretation of unit, the closure constant will be very important for the correct interpretation of the units.

Definition 5 *Compositionally Equivalent. Let two vectors of D positive real components* $x, y \in R_+^D$. $(x_i, y_i > 0)$ *for all* $i = (1, 2, \dots, D)$. x *and* y *are compositionally equivalent if there exists a positive scalar* $\lambda \in R_+$ *such that x=*λ*.y and, equivalently C(x) = C(y) where C is closure function.*

Only scale invariant functions can be consistently used in CoDA analysis which is defined as $f(\alpha.x) = f(x)$, where $\alpha > 0$

Example: Let define a scale invariant function f as the ratio of elements by the last element of vector. $f(x) = (\frac{x_i}{x_D}, \dots, \frac{x_{D-1}}{x_D})$ where $x \in R_+^D$ is a compositional vector. And let $x = (1.6, 2.4, 4.0)$ and $y = (3.0, 4, 5, 7, 5)$ be two composition vectors.

$f(x) = (\frac{1.6}{4.0}, \frac{2.4}{4.0})$ and $f(y) = (\frac{3.0}{7.5}, \frac{4.5}{7.5})$ and $f(x) = f(y) = (0.4, 0.6)$ so that x and y are compositionally equivalent.

There are many equivalent sets of ratios which may be used for the purpose of creating meaningful functions of compositions. For example the geometric mean of components of a composition $f(x) = \frac{x}{g(x)}$ where $g(x) = (x_1, \dots, x_D)^{\frac{1}{D}}$ would also meet the scale invariant requirement [65].

4.1.1.2 Principles of Permutation Invariance

The principle of ‘permutation invariance’ states that the ordering of components are not matter as long as all compositions are ordered in a consistent manner [32]. Equivalent result should be obtained when ordering of components changed in a composition. For example, calculating distance between two composition vectors x=(1.6, 2.4, 4.0) and y = (3.0, 4,5, 7,5) does not depend on the order of components. x’=(2.4, 4.0,1.6) and y’= (4,5, 7,5 ,3.0) have the same distance. Removing any part from the composition is not permutation invariant since result will depend on the erased component [43].

4.1.1.3 Principles of Subcomposition coherence

The principle of subcompositional coherence states that any method used should produce consistent results between a full composition and a subset obtained by deleting some components. For example, the distance measured between two full compositions must be greater then the distance between them when considering any subcomposition. Moreover, erasing a non-informative part of composition should not change the result.

For example let S be a full composition S= (0.1, 0.2, 0.1, 0.6), (0.2, 0.1, 0.1, 0.6) and s= (0.25, 0.50, 0.25), (0.50, 0.25, 0.25) is a subcomposition of S. The ratio of two components $\frac{s_i}{s_j}$ and $\frac{x_i}{x_j}$ remains unchanged when transfer data from full composition to subcomposition. Thus, as long as working with scale invariant functions, or equivalently to express all statements about composition in terms of ratios, subcompositional coherent will be reserved [43].

4.2 Aitchison Geometry

Compositional data exist in a subspace known as the simplex, so many commonly used metrics in Euclidean space is invalid for relative data [32].

Definition 6 *Simplex. The sample space of compositional data is the simplex, defined as*

$$S^D = \{x = \{x_1, x_2, \dots, x_D\} | xi > 0, i = 1, 2, \dots D; \sum_{i=1}^{D} x_i = \kappa\}$$

The ternary diagram that is shown in the Figure 5 is the standard representation of simplex for D = 3.

Presence or absence of other components affect the distance between two composition [66]. Increasing the abundance of one decreases the proportional abundance of the others so that representing variables as portions of the whole makes them mutually-dependent multivariate objects and multivariate statistics yield erroneous results [66].

Example: Let $C1 = ([5, 65, 30], [10, 60, 30])$ and $C2 = ([50, 20, 30], [55, 15, 30])$

Euclidean distance is the same between compositions C1 and C2 due to 5 unit difference between first and second components. If proportions are consider, first component of C1 is doubled since first component of C2 has a relative increase around 10 %.

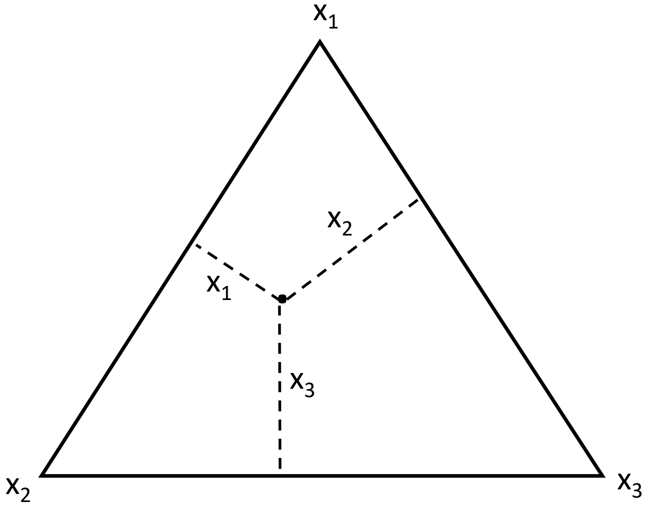


Figure 5: Ternary Diagram.

Hence, another geometry was needed to work with compositional data and Aitchison defined Aitchison Geometry to work with relative data [32].

4.2.1 Defining Simplex as a Vector Space.

In order to calculate distance, length, norm, inner product, orthogonality etc. between compositions, two operations were defined on simplex: perturbation and power transformation.

Definition 7 *Perturbation of a Composition. Let x and $y \in S^D$ are two compositions. Perturbation of x by y is defined as*

$$x \oplus y = C\{x_1 y_1, x_2 y_2, \ldots, x_D y_D\}.$$

where C is closure function.

Definition 8 *Power Transformation of Composition. Let $x \in S^D$ is a composition and $\alpha \in R$ is a scalar, Power transformation of x by α is defined as*

$$\alpha \odot x = C\{x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\}$$

where C is closure function.

Hereby, the simplex, $(S, \oplus, \odot)$ with perturbation operation and power transformation is a vector space.

The Figure 6 shows the perturbation and power transformation effect in simplex. On the left; original composition (*) is perturbed by $p = \{0.1, 0.1, 0.8\}$ and the resulting composition (o) obtained. On the right; original composition (*) is powered by $\alpha = 0.2$ and resulting composition (o) obtained.

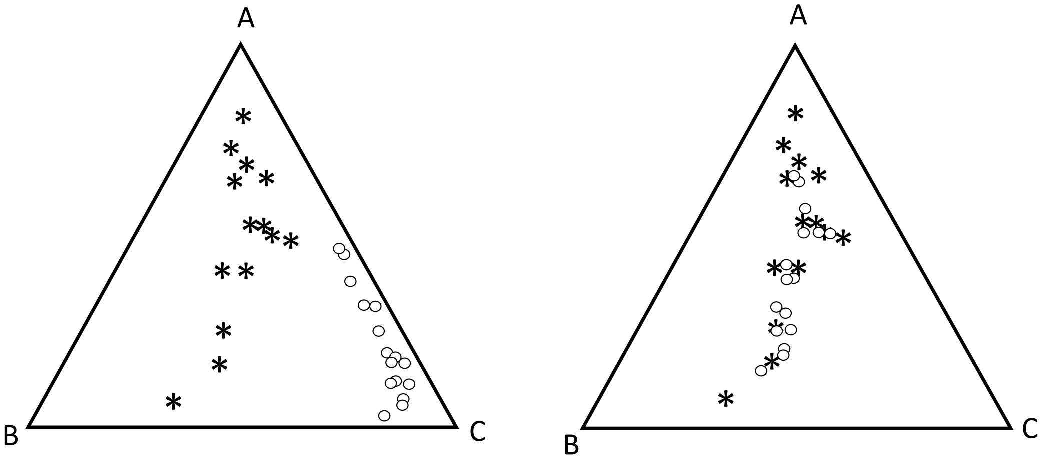


Figure 6: Perturbation and power transformation effect in simplex.

Since, S^D is a vector space then it holds commutative and associative property, neural and inverse element properties.

$(S^D, \oplus)$ has a commutative group structure, for $x, y, z \in S^D$ it holds

1. Commutative property: $x \oplus y = y \oplus x$
2. Associative property: $(x \oplus y) \oplus z = x \oplus (y \oplus x)$
3. Neural Element: $n = C[1, 1, \dots, 1] = [\frac{1}{D}, \frac{1}{D}, \dots, \frac{1}{D}]$
4. inverse of x : x^-1 = $C\{x_1^-1, x_2^-1, \dots, x_D^-1\}$ and $x \oplus x^-1 = n$

The power transformation satisfies the properties of an external product, for $x, y, z \in S^D$, $\alpha, \beta \in R$ it holds:

1. Associative property: $\alpha \odot (\beta \odot x) = (\alpha.\beta) \odot x$
2. Distributive property: $\alpha \odot (x \oplus y) = (\alpha \odot x) \oplus (\beta \odot y)$ and $(\alpha + \beta) \odot x = \alpha.\beta \odot x$
3. Neural Element: $1 \odot x = x$

Definition 9 *Inner Product. Let $< ., . >_a$ stands for the Aitchison inner product and $x, y \in S^D$ are two compositions. Inner product of x and y is defined as*

$$< x, y >_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}$$

Definition 10 *Norm Let* $||.||_a$ *stands for the Aitchison norm and* $x, y \in S^D$ *are two compositions. Norm of* x *is defined as*

$$||x||_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} (\ln \frac{x_i}{x_j})^2}$$

Definition 11 *Distance Let* $d_a(.,.)$ *stands for the Aitchison distance and* $x, y \in S^D$ *are two compositions. Distance between* x *and* y *is defined as*

$$d_a(x, y) = ||x \ominus y|| = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} (\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j})^2}$$

4.2.2 Log Ratio Analysis: A Statistical Methodology for Compositional Data Analysis

Subcompositional coherence principle of compositional data puts conditions on how to conduct a CoDA analysis [67]. Scale invariant functions of the composition guarantee the subcompositional coherence [43]. Any meaningful scale-invariant function of a composition can be expressed in terms of ratios of the components of the composition or only of log-ratios of the components [66]. Thus, it led to working with logarithms of ratios since logarithms of ratios are mathematically easier to control than ratios [32].

Log-contrasts are the most frequent used scale-invariant functions in CoDa analysis. A log-contrast is a simple way of expressing a set of log-ratios in a linear form which is symmetric in the components.

Definition 12 *Linear functions: log-constrasts Let* $x \in S^D$ *is a composition. A log-constrast on x is defined as*

$$f(x) = \sum_{i=1}^{D} \alpha_i . ln(x_i), \sum_{i=1}^{D} x_i = 0$$

Simple log ratios $ln(x_i/x_j)$ is a log-contrast function. The balance function; $ln(\frac{g(x_1,x_2,...x_r)}{g(x_r+1,x_r+2,...x_s)})$, where g(.) is the geometric mean of the arguments is also a log-contrast function.

Aitchison projected the sample space of compositional data, the D-part simplex S^D, to real space $R^D - 1$ or R^D, using log-ratio transformation. The philosophy of logratio analysis can be stated simply [68]

1. Formulate the compositional problem in terms of the components of the composition.
2. Translate this formulation into terms of the logratio vector of the composition.
3. Transform the compositional data into log ratio vectors.

4. Analise the log ratio data by an appropriate standard multivariate statistical method.
5. Translate back into terms of the compositions the inference.

After conceiving log-ratios as coordinates in a real Euclidean space [69, 43], the idea of transforming CoDa to the real space is irrelevant, as the coordinates fully represent compositions and this is equivalent to analyzing compositions using the Aitchison geometry or to analyzing their representation in coordinates using the ordinary Euclidean geometry [65, 67]

4.2.2.1 Additive Log Ratio (alr) Transformation

Many compositional data analysis begin with conversion of absolute data set into relative space by dividing each element of the sample vector by the total sum. It could be possible that the two groups of compositional data appear clearly linearly separable in absolute space, but after transformation, the boundaries between groups might become unclear in relative space. In order to reveal group separation, dividing all or some of the features by a reference feature, one might discover that the resultant ratios can separate the groups and any separation revealed by such ratios can be analyzed by standard statistical techniques.

Alr transformation is achieved by taking the logarithm of each measurement within a composition as divided by a reference feature.

Definition 13 *Alr Transformation.* $alr : S^D \rightarrow R^{D-1}$ *transformation defined by:*

$$y = alr(x) = [\ln \frac{x_i}{x_D}; \ldots ; \ln \frac{x_{D-1}}{x_D}]$$

Alr is a bijective transformation so the inverse of transformation $alr^{-1} : R^{D-1} \rightarrow S^D$ is

$$x = alr^{-1}(y) = C(exp(y1), exp(y2), \ldots , exp(y_{D-1}))$$

The result of alr transformation does not belong to the choice of the final component as reference feature. If components were permuted, the result would not change [70]. One drawback of alr transformation is being a asymmetric in the parts, it depends of the chosen reference feature.

4.2.2.2 Centered Log Ratio (clr) Transformation

In order to overcome drawback of alr and treat parts symmetrically instead of choosing one component as reference feature, an abstract reference ; the geometric mean of the composition g(x) is used and this transformation called centered log ratio (clr) transformation

Definition 14 *Clr Transformation.* $clr : S^D \rightarrow R^{D-1}$ *transformation defined by:*

$$z = clr(x) = \{\ln \frac{x_i}{g(x)}; \ldots ; \ln \frac{x_D}{g(x)}\}$$

where g(x) is the geometric mean of the composition

Clr is also a bijective transformation and the inverse of transformation $clr^{-}1 : U^D \rightarrow S^D$ takes the form

$$x = clr^{-1}(z) = C[exp(z1)...exp(z_D)]$$

where $U^D = [u_1...u_D] : u1 + ...u_D = 0$ a hyperplane of R^D.

One drawback with clr transformation is that the sum of transformed values is 0. Clr transformation creates a constrained vector in R^D ,thus it yields a coordinate system featuring a singular covariance matrix which is unsuitable for many common statistical models [69].

4.2.2.3 Isometric Log Ratio (ilr) Transformation

Unlike alr and clr, the isometric log-ratio bypasses of choosing any feature as divisor; instead of using orthonormal basis for transformation [69]. The isometric log-ratio transformation (ilr) is an isometric linear mapping between the simplex and R^D which preserves distances and angles between points.

The transformation is performed by first finding an orthonormal basis for S^D and transforming it into a "contrast matrix" ψ and use it to define the ilr transformation.

Definition 15 *Ilr Transformation.* $ilr : S^D \rightarrow R^{D-1}$ *transformation defined by:*

$$ilr(X) = clr(X)\psi^t$$

Contrast is a linear combination of variables whose coefficients add up to zero, allowing comparison of different parts. Those coefficients can be used to construct a contrast matrix.

Isometric log ratio (ilr) transformation overcomes drawbacks of alr and clr and assign coordinates with respect to orthonormal basis.

The ILR transform can be built from a sequential binary partition (SBP) of the original variable space. The SBP is a hierarchy of the parts of a composition: in each step it is split into two groups so it ensures the orthogonality. However, a known obstacle of ILR transform is the choice of partition such that the resulting coordinates are meaningful. Details of SBP explained in Basis and Balances section.

4.2.3 Compositional Distance

Compositional data lives in a simplex so Euclidean distance between samples does not make sense [66]. Distance for compositional data is Aitchison distance. It provides a measure of distance between two d-dimensional compositions. Unlike Euclidean distance, Aitchison distance has scale invariance, perturbation invariance, permutation invariance, and sub-compositional dominance properties. There is another distance that accomplish these four properties is Mahalanobis (clr) distance.

Definition 16 *Aitchison Distance. Let x and $y \in S^D$ are two compositions. Aitchison distance is defined as*

$$d_a(x,y) = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2}$$

Aitchison distance is simply the Euclidean distance between clr-transformed compositions. In other words, when data is clr-transformed, then Euclidean distance makes sense.

4.3 Basis and Balances

Principal component analysis (PCA) is a well-known method in statistics to transform variables into a new set of uncorrelated variables called principal components (PC). The first PC is the linear combination of the original variables which acquire the largest sample variance. Geometrically, each PC is associated with a direction represented by a vector (also called Principal Direction (PD)) [71]. PDs constitute an orthonormal basis of the space. The sample values of PCs, called scores, are expressed as coordinates with respect to the PDs [71].

Remember that analyzing compositional data with traditional statistical methods, it needs to be transformed to real space by alr, clr or ilr transformation functions. To calculate the basis in a simplex, basis of R^D must be transformed back to simplex.

Let $(u_1, u_2, \ldots, u_{D-1})$ is a basis of R^{D-1} where

$$u_1 = [1;0;0;\ldots;0;0]; u_2 = [0;1;0;\ldots;0;0]; \ldots; u_{D-1} = [0;0;0;\ldots;0;1]$$

then the compositions $(e_1, e_2, \ldots, e_{D-1})$ is called compositional basis (C-basis) of S^D where

$$e_1 = alr^{-1}(u1) = C[e;1;\ldots;1]; e_2 = alr^{-1}(u2) = C[1;e;1;:\ldots;1]; \ldots e_{D-1} = alr^{-1}(u_{D-1}) = C[1;\ldots;1;e;1]$$

Thus the vector of C-coordinates of a composition x with respect to this basis is

$$alr(x) = [ln(x_1/x_D);\ldots;ln(x_{D-1}/x_D)]$$

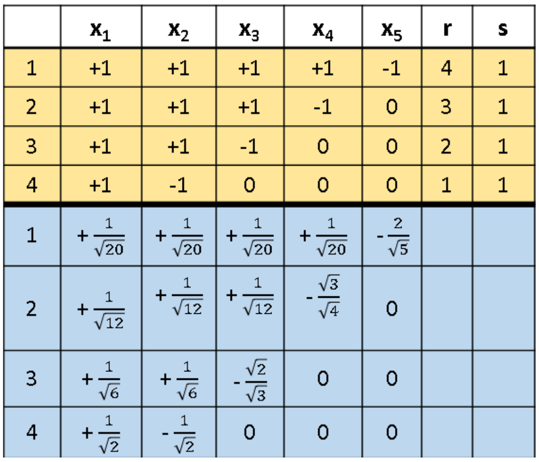
But C-basis are not orthonormal since $||ei||_c^2 = 1 - (1/D) \neq 1$ and $< ei, ej >_c = (-1/D) \neq 0$.

The best manner of defining an orthonormal basis in S^D is by the clr-coefficients of the compositions: it suffices to ensure that the squared coefficients add up to 1; and the ordinary inner product of the clr-coefficients, as real vectors, is 0.

The practical way of defining orthonormal basis is Sequential Binary Partition (SBP). The SBP is a hierarchy of the parts of a composition: in each step it is split into two groups so it ensures the orthogonality. A sign table is constructed dividing composition between groups. The idea is that the first row of the sign table is created by splitting the composition into two groups; each of these groups is then split into two more groups. This process continues until each group has a single part. 1 is used to indicate an inclusion in the first group, -1 indicates an inclusion in the second group and 0 indicates no inclusion [72].

Sequential binary partition process and an example of sign matrix are shown in Figure 7 for $D = 5$.

The contrast matrix ψ is then constructed by the formula: $\psi_{ij} = \frac{1}{r}\sqrt{\frac{r.s}{r+s}}$, when the corresponding value in the sign table is positive and $\psi_{ij} = \frac{1}{s}\sqrt{\frac{rs}{r+s}}$, when the corresponding value in the sign table is negative in which separating parts composed of r and s parts. If the corresponding sign table value is 0, then the value in the contrast matrix is also 0. The lower part of the Figure 7 shows the ψ matrix of the basis.



	x_1	x_2	x_3	x_4	x_5	r	s
1	+1	+1	+1	+1	-1	4	1
2	+1	+1	+1	-1	0	3	1
3	+1	+1	-1	0	0	2	1
4	+1	-1	0	0	0	1	1
1	$+\frac{1}{\sqrt{20}}$	$+\frac{1}{\sqrt{20}}$	$+\frac{1}{\sqrt{20}}$	$+\frac{1}{\sqrt{20}}$	$-\frac{2}{\sqrt{5}}$		
2	$+\frac{1}{\sqrt{12}}$	$+\frac{1}{\sqrt{12}}$	$+\frac{1}{\sqrt{12}}$	$-\frac{\sqrt{3}}{\sqrt{4}}$	0		
3	$+\frac{1}{\sqrt{6}}$	$+\frac{1}{\sqrt{6}}$	$-\frac{\sqrt{2}}{\sqrt{3}}$	0	0		
4	$+\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	0	0	0		

Figure 7: Sequential Binary Partition Example

Once ψ has been found, it is easy to perform the ilr transformation. Positive components $(x_j's)$ are taken as numerator, and negative components $(x_l's)$ are taken as denominator and ilr coordinates are calculated using the formula:

$$X_i^* = \sqrt{\frac{r_i.s_i}{r_i + s_i}} \ln \frac{(\prod_{j \in R_i} X_j)^{1/r_i}}{(\prod_{l \in S_i} X_l)^{1/s_i}}$$

These ilr coordinates are called as balances [43].

Definition 17 *Balances. Balances are the coefficients corresponding orthogonal bases. Balances are log-contrasts which are log-ratios of geometric means of two non-overlapping groups of parts.*

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{g(x_+)}{g(x_-)}$$

where x_+, x_- are two non-overlapping groups of parts of a composition and g(.) is the geometric mean of the arguments.

A positive balance means that the group of parts in the numerator has more weight in the composition than the group in the denominator (and vise verse for negative balances). Moreover, balances can be very useful to project compositions onto special subspaces just by retaining some balances and making other ones null [62].

In summary, a set of orthonormal balances is easily defined using a SBP, resulting in ilr-coordinates.

4.3.1 Principle Balances

Principal balances (PBs) refer to principle components of simplex in real space. PBs are defined as a sequence of orthonormal balances which maximize successively the explained variance in a compositional data set. Given a compositional centered sample, we define the first PB as the balance which maximizes the explained sample variance. Subsequent principal balances, being orthogonal to the preceding ones, also maximize the explained remaining variance [71]. Similar to PCs, PBs maximize the explained variance of a data set in decreasing order.

Computing PBs requires an exhaustive search along all possible sets of orthogonal balances. PBs can be approximated by hierarchical clustering of compositional parts using Ward's method and those hierarchical clusters are used as a SBP for PBs [45]. Hierarchical clustering of components yields, by construction, a series of balances with increasing variance. The Ward clustering method [73] merges clusters with the most similar centroids to form a cluster. The algorithm needs a distance or similarity measure in order to merge parts. The variation matrix can be used as a distance matrix, which is actually a square distance matrix between parts [45, 74]. The variation matrix elements are all positive and the variance of the log-ratio of two parts is zero if the two components are perfectly proportional. Note that proportionality is presented by Lovell et. al. [26, 74] as a valid alternative to correlation for relative data. When two parts x_i and x_j are exactly proportional, $\mathrm{var}(\ln(x_i/x_j)) = 0$, so x_i and x_j are linearly associated. If the variance is large, then the proportionality of the two variables is unreliable, and they likely belong to different groups. The Ward algorithm starts detecting the smallest entry in the variation matrix and the corresponding parts are merged to form a group [45]. Then, the geometric mean of both columns -the group centroid- is calculated and the variation matrix is updated. The algorithm iteratively continues merging groups of parts according to the smallest variance of the corresponding balance. The first principal balance, using the clustering method, includes all parts and defines the largest variance in the data.

Definition 18 *Principal Balances. Let $x = x_1, ..., x_D$ be a D-part composition. Principal Balances are log-linear functions $\sum_{i=1}^{D} a_{ki} \ln X_i$, k=1,2,...D-1 such that the vectors $a_k = (a_{k1}, a_{k2}, ..., a_{kD})$ are constant and they maximize the variance*

$$var\left[\sum_{i=1}^{D} a_{ki} \ln X_i\right]$$

subject to

1. *(balance condition) for k=1,2,...D-1, the coefficients a_{ki} take one of the three values (-c1,0,c2) for some strictly positive c1 and c2.*

2. *(zero sum and unit norm conditions) for k=1,2,...D-1, the coefficients* a_k *satisfies* $\sum_{i=1}^{D} a_{ki} = 0$.
3. *(orthogonality condition) for k=2,...D-1, the coefficients* a_k *is orthogonal to the previous* $a_{k_1}, a_{k_2}, ..., a_1$*, that is* $\sum_{i=1}^{D} a_{ki}a_(k-l)_i = 0$*, l=1,2,...,k-1.*

4.4 Exploratory Data Analysis of Compositional Data

4.4.1 Center, Variation Matrix and Covariance Structure of Compositional Data

To compute variance, the center of data needs to be computed first. Let $X = x_i = (x_{i1}, \dots, x_{iD}), i = 1, \dots, n$ be a dataset of observations of the simplex S^D of size n.

Definition 19 *Center. The center is defined as the geometric mean of the parts.*

$$G = C(g_1, g_2, \dots, g_n) \textit{ with } g_j = \prod_{i=1}^{n} X_ij^{1/n}$$

The center g also can be calculated from the arithmetic mean of the clr -transformed data or alr-transformed data.

Dispersion in a compositional data set can be described either by the variation matrix or by the normalized variation matrix originally defined by Aitchison (1986) using the log-ratio variance.

Definition 20 *Variation Matrix.*

$$T = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1D} \\ t_{21} & t_{22} & \dots & t_{2D} \\ \dots & & & \\ t_{D1} & t_{D2} & \dots & t_{DD} \end{bmatrix} \textit{ where , } t_{ij} = var\big(ln(x_i/x_i)\big)$$

or by the normalized variation matrix

$$T^* = \begin{bmatrix} t^*_{11} & t^*_{12} & \dots & t^*_{1D} \\ t^*_{21} & t^*_{22} & \dots & t^*_{2D} \\ \dots & & & \\ t*_{D1} & t*_{D2} & \dots & t^*_{DD} \end{bmatrix}$$

where, $t^*_{ij} = var\big(\frac{1}{\sqrt{2}} ln(x_i/x_i)\big)$

t_{ij} stands for the log-ratio of parts i and j while t^*_{ij} stands for the normalized log-ratio of parts i and j. Note than $t^*_{ij} = \frac{1}{2}t_{ij}$ and thus $T^* = \frac{1}{2}T$ [62]. Variation matrix is symmetric with 0 diagonal. When the variance (t_{ij}) is null, x_i and x_j are strictly proportional; when it is large, proportionality is lost or it is too noisy to be considered [74].

The total variance in a compositional data set is measured by total log-ratio variance. The total variation summarises the variation matrix T in a single quantity. Total variance is defined as $\frac{1}{2D}$ x (sum of all elements of matrix T).

Definition 21 *Total Variance. A measure of global dispersion is the total variance.*

$$TotalVar[X] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} var\big(\ln\frac{x_i}{x_j}\big) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} t_{ij} = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{D} t^*_{ij}$$

The classic covariance calculation has negative bias to the unit sum constraint. This implies that at least one of the covariances between x_i and another component is negative. Aitchison [32] describes covariance structure of D-part composition as set of all covariances as following:

Definition 22 *Covariance Structure.*

$$\sigma_{ij.kl} = cov(log(x_i/x_j), log(x_k/x_l))$$

Two matrices can be used to describe the covariance structure of X.

1. $Cov(clr(X)) = \Gamma = (\tau_{ij}) = (cov(log(\frac{x_i}{g(x)}), log(\frac{x_j}{g(x)})) : i, j = 1, \ldots, D)$ where g(x) is the geometric mean of the components of X is a covariance matrix and treats parts symmetrically, but it is singular because the sum of the each row of the matrix is 0.
2. $Cov(alr(X)) = \sum = (\sigma_{ij}) == (cov(log(\frac{x_i}{X_D}), log(\frac{x_j}{X_D})) : i, j = 1, \ldots, D$ is a covariance matrix and it is non-singular but asymmetric in its treatment of parts.

4.4.2 Correlation Analysis of Compositional Data

In linear space, perfectly correlated data will follow the formula y = m.x+b, where y and x are variables, m is the slope of the line and b is the intercept. However, in log transformed data, m becomes the slope of the line, and b becomes a non-linear parameter [25]. When the intercept is 0, then the data is linearly related in the normal and log space. The only difference is the changes of slopes; all lines with the intercept 0 has the same slope 1 in log space. When the intercept is not 0, then the data in log space lines are curved with changing intercept which means rations are changing and the lines are not associated in log space. Moreover, correlation is not corrected by using non-parametric correlation measures [25].

The covariance relationship has been exploited to develop algorithms for inferring correlation networks from compositional data [51, 56, 54, 53, 28].

Sparse Correlations for Compositional data (SparCC) [56] captures the conditional interdependence between parts and estimates the linear Pearson correlations between the log-transformed components.

Correlation inference for Compositional data through Lasso (CCLasso) [53] algorithm infers the correlations among parts through a latent variable model after log ratio transformation for raw compositional data.

Regularized estimation of the basis covariance based on compositional data (REBACCA) [54] algorithm is to identify significant co-occurrence patterns by finding sparse solutions to a system with a deficient rank.

SParse InversE Covariance Estimation for Ecological Association Inference (SPIEC-EASI) [51] addresses interdependence through a clr transformation of the relative abundance data and then estimates the sparse inverse covariance matrix, therefore inferring association based on conditional independence.

COmposition-Adjusted Thresholding (COAT) method [28] is to estimate the sparse covariance matrix of the latent log-basis components. The method is based on a decomposition of the variation matrix into a rank-2 component and a sparse component. The resulting procedure can be viewed as thresholding the sample centered log-ratio covariance matrix and hence is scalable to large covariance matrice estimations based on compositional data.

4.4.3 Regression Analysis of Compositional Data

In a regression analysis, compositions can serve as covariates (that is, predictors), as response variables or both. The first step in a regression analysis involving compositions, the components of the composition are transformed to R_D. Once all the variables are in R_D, then a multivariate or univariate regression analysis is performed. After a satisfactory model has been found the model can be converted back into S_D if a suitable transformation can be performed [75].

Due to the drawbacks of clr and alr explained in the section 4.2.2, ilr transformation is preferred for compositional regression models. After transforming classical statistical models can be used on the data. For example, after log-ratio transformation, the estimation can be made with the OLS method and expressed in coordinates. Then, the estimated model can be expressed in the simplex using the inverse transformation.

4.4.3.1 Case 1: Response is Real , Covariates are Compositional

A composition X can be used as predictor of a non-compositional variable W.

W_i represent some real response variable and X_i represent some compositional covariate.

$$W_i = \alpha + \beta(ilr(X_i)) + \epsilon_i$$

β values can be obtained solving an ordinary regression problem in ilr coordinates. Ilr transformed components can be transformed back into the simplex which results in

$$W_i = \alpha + < b, X_i >_A + \epsilon_i$$

where B is the composition created by performing a reverse ilr-transformation on the coefficients β. The intercept is a and the composition parameter is b. ϵ_i is the error term with $\epsilon_i \in N(0, \sum)$

As an example, Washburn et. al., Pinto et. al. and McDonald et. al[76, 77, 5] used otu table as the predictor and a real variable as the response variable in order to identify otus that are strongly associated with a given environment.

4.4.3.2 Case 2: Covariates is real , Response are compositional

A composition X can be predicted using of a non-compositional variable U.

In the model, a and b constant compositions, $X_i \in S^D$ compositional response variable and U_i is the real covariate. ϵ_i is the error term with $\epsilon_i \in \mathcal{L}(0, \sum)$.

$$X_i = a \oplus (U_i \odot b) \oplus \epsilon_i.$$

First the model must be transformed to R^D.

$$ilr(X_i) = ilr(a) + U_i.ilr(b) + \epsilon_i$$

with $\epsilon_i \in \mathcal{N}(0, \sum)$. The parameters a and b can now be estimated via standard multivariate regression.

As an example, Morton et. al. [78] used pH values to predict OTU proportions in the environment using ordinary least-squares linear regression on balances.

4.4.3.3 Case 3: Both Covariates and Response are compositional

A composition X can be predicted using of a compositional variable Y.

The model in S_D space will be

$$X = \alpha \oplus (\beta_c \odot Y) \oplus \epsilon$$

The model in R^D is

$$ilr(X) = \alpha_i r + \beta_i rilr(Y)$$

using ilr transformation.

Once the data have been transformed to unconstrained space then normal multivariate linear modeling techniques can be used to create a model of the data.

4.4.4 PCA for Compositional Data

Principal Component Analysis (PCA) should not get applied directly to compositional data. Instead, PCA could be applied to clr-transformed data (resulting in an additional centering of the rows after log transformation) [70]. When interpreting the resultant PCA, it should be considered that covariances and correlations between features exist with respect to the geometric mean reference [79]. Relative variation biplot reveals associations between samples and features, and can also be used to infer power law relationships between features in an exploratory analysis [70].

4.4.5 Biplot for Compositional Data

When analyzing and interpreting compositional data, it is important to remember that the variance in the ratios of the underlying data is examined (not directly examining abundance). A biplot represents simultaneously the rows (observations) and columns (parts) of the matrix X by means of a rank-2 approximation. In other words, a multidimensional dataset is projected onto two dimensions. Biplots are based on the variance of the ratios of the parts.

Compositional biplots are generated after zero-replacement and clr transformation of the data. Singular value decomposition (SVD) is conducted on clr-transformed data and PCs are visualized. Two types of biplot are possible (form and covariance biplots), depending on the assignment of the singular values to the left or right singular values of the decomposition [80]. In both the projections of one set of points on the other approximate the centered data. The form biplot, where singular values are assigned to the left vectors corresponding to the observations, displays approximate Euclidean distances between the observations. The covariance biplot, where singular values are assigned to the right vectors corresponding to the parts, displays approximate standard deviations and correlations of the parts.

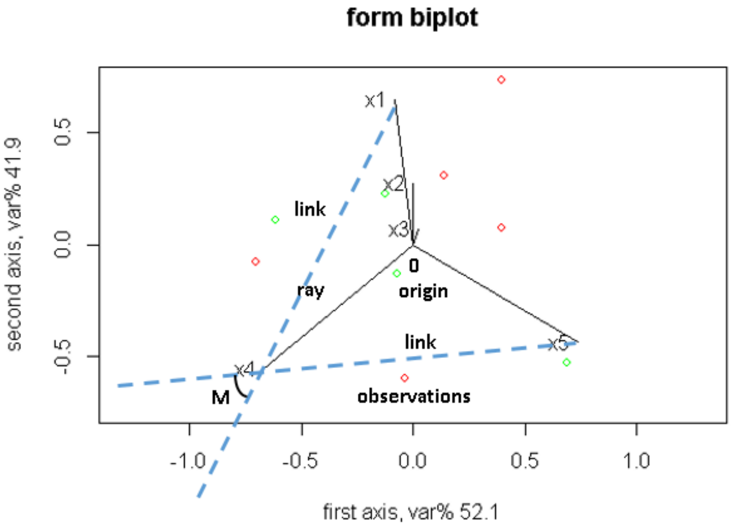


Figure 8: Ray and Link in biplot of a 5-part composition for 10 observation

Figure 8 shows a form biplot for 5-part composition for 10 observations. The length of links and rays provide information on the relative variability in a compositional data set. The length of the links indicates variation of log ratios. If the length is high then variation is high. Short links indicate a constant or near constant ratio between the two linked parts. Each ray represents the variance of a log ratio; for example the $ray|0x_i|^2 \sim var(log(x_i/g(x)))$ and the link $|x_ix_j|^2 \sim var(log(x_i/x_j))$. The angle between links provide information on the correlation of subcompositions. If two links intersect at M and the $cos(M) \sim 0$ then, zero correlation of the two log ratios can be expected. In order words, orthogonal links indicates independent parts in the composition.

The distance between observations is related to their multivariate similarity of the parts as ratios. If all components are relatively the same (ie, the ratios between all parts are identical), then two samples are in the same location [74].

Biplots are a useful tool that provides a visual detection of not only the relationships between the sample compositions, but also which parts are influencing the differences between them.

4.4.6 CODA Dendrogram

The CODA dendrogram is a powerful tool in order to explore a compositional dataset. It can be used as a descriptive tool for visualizing some univariate statistics of the ilr coordinates derived from an SBP [45]. The SBP table in the Figure 7 can be represented by dendrogram-type links between parts, as shown in Figure 9. The leaves of the dendrogram, represented by dotted lines, correspond to the groups of parts formed by a unique element. The vertical bars describe the groups of parts formed at each order of partition. Vertical bars are scaled in the interval (-c,c), where c is used defined. Each branching corresponds a ilr coordinates (balances). The location of the mean of an ilr coordinate is determined by the intersection of the horizontal segment with the vertical segment (variance). The sum of all vertical bars represents the total variance of the sample. A short vertical bar means that the balance has a small variability in the sample, thus explaining only a little bit of the total variance. Conversely, a long vertical bar implies a balance explaining a good deal of the total variance [81].

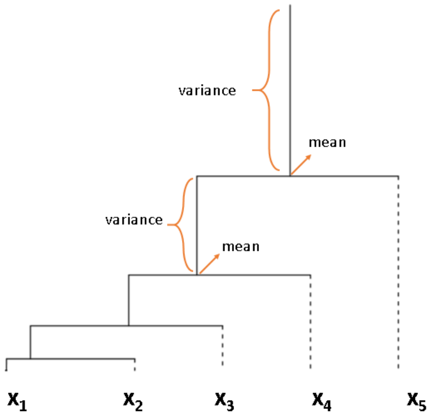


Figure 9: CoDa Dendrogram of 5-part composition

CHAPTER 5

PRINCIPAL MICROBIOME GROUPS FOR BIOMARKER INVESTIGATION

In this chapter, we address the problem of establishing relationship based on the microbial features annotated with taxonomic information, where a compositional alternative to phylogenetic grouping of microbiome data, Principal Microbial Groups (PMGs), is proposed to enable working with low-level microbial features (OTUs or ASVs). The grouping is based only on relative abundances and it is based on Principal Balances [71]. The usefulness of the proposed procedure in order to search for biomarker candidates is illustrated on Cirrhosis dataset.

5.1 Introduction

High-throughput sequencing has led to an explosive growth of studies on the associations between human microbiome and human disease. Many chronic diseases, including obesity, type 2 diabetes, liver diseases, cancer and allergies have linked alteration in the human gut microbiome [82, 83, 84, 85, 77, 86, 87]. Microbiome profiles are typically high-dimensional and very sparse, leading to two main problems in data analysis. The main approach to deal with these problems is to annotate constructed microbial features with taxonomy. The majority of microbiome studies (96.9%) used the OTU approach to cluster reads and assign taxonomy to the clusters [83]. Agglomerating taxa by rank of interest (phylum, genus etc.) allows summarizing microbiome abundance with a coarser resolution in lower dimension. Genus was the most frequently used level (75.7%), followed by phylum (55.3%), and only 16.0% of the studies focused on species level [83]. The similarities or relationships between samples are addressed correspondingly. The higher the taxon level bacteria are collapsed into, the lower dimensionality and sparsity one can achieve [6]. However, bacterial strains in the same taxonomic group have been found to vary in their relationships with the host bio-clinical parameters, suggesting that each of them may have a distinct impact on host health [6]. Thus, correlating selected taxa with disease can often lead to controversial results in biomarker studies. If members in a taxon have opposite associations with the same disease, lumping them into one taxon variable will produce degradation of the possible associations with the disease.

Researchers are interested in identifying a single taxonomic unit that may serve as a biomarker of diseases using classical statistical and machine learning techniques [88, 89, 83, 86]. However, considering the preventive or risk effects of each bacteria separately does not adequately account for the variation in the human microbiome and it is rare for a single bacterial species to be associated with a disease [90, 91]. Indeed, it is suggested that dysbiosis (imbalance in microbial communities) [92, 93] is likely to contribute to diseases [94, 95, 96]. Thus, detection of bacterial species that are out of balance has become important in developing promising diagnostics.

Recent awareness of the compositional nature of microbiome data has led to employ the compositional approach in microbiome studies [97, 24, 5, 98, 99, 76, 100, 101, 30, 102, 25, 103]. Relative abundances are compositional and the relative data contains the relationships between the features of the dataset [97, 24]. If the relative abundance of one microbial feature increases, the relative abundances of some other microbial features must decrease, and vice versa. Pearson correlations of relative abundances are spurious and cannot be relied upon to make coherent inferences about the relationships between pairs of features [26, 53, 51]. There is an increasing number of publications motivating and using the log-ratio methodology for statistical processing of microbiome [5, 76, 104, 105, 106, 107, 29, 108]. Log-ratio methodology brings a new perspective to the biomarker concept since it deals with ratios of microbial features. So, focusing on a single species is not suitable for log-ratio methodology. Then, the ratio of microbial features can be interpreted as positively or negatively correlated with the disease. However, the most of the current biomarker discovery methodologies do not consider the compositional nature of the microbiome data [109], as they assume implicitly the sample space to be the real space endowed with the usual Euclidean geometry. On the contrary, the compositional approach, which assumes the sample space to be the simplex endowed with the Aitchison geometry [110], could reveal relevant microbiome markers among microbiome samples or groups of samples (e.g., sick vs healthy) [111, 112]. Some recent biomarker discovery methodologies that consider the compositional nature of the data usually work with agglomerated taxa by the rank of interests (phylum, genus etc.) that might lead spurious results in biomarker studies [6].

5.1.1 Compositional Data (CODA) Approach for Microbiome Data Analysis

The main idea of CODA is to represent the original microbiome data in coordinates [110, 69] of the simplex corresponding to the Aitchison geometry. These coordinates are, by construction, real, and their support space is the real space endowed with the usual Euclidean geometry. These new variables are formed by interpretable log-ratios or their linear aggregates (log-contrasts), and then one can continue with standard statistical or machine learning processing [68, 113].

The components of a composition are called parts. Linear functions of a composition onto the real numbers are scale invariant additive combinations of the logarithms of parts, called log-contrasts. Log-contrasts are characterized by the fact that the weighting coefficients sum up to zero. Log-contrasts are obvious candidates for describing the characteristics of a composition and are then used as statistics from a sample. Examples of such log-contrasts are compositional principal coordinates [68] and balances, the latter understood as log-ratios of geometric means of groups of parts [114, 43]. The interpretability of log-contrasts depends on the characteristics of the combination of logs. Compositional principal components generally involve all parts of the composition in a non-homogeneous way, thus making its interpretation difficult. That is, they are neither simple nor sparse. Balances are a simple class of log-contrasts, as the combination coefficients have only two values different from zero (simplicity), and when the involved groups of parts are small, balances are also sparse [45]. Additionally, orthonormal cartesian coordinates are sets of log-contrasts called isometric log-ratio (ilr/olr) coordinates [69, 67]. Compositional principal components are ilr-coordinates. Alternatively, ilr coordinates can be obtained by a sequential binary partitions (SBP) of the composition [114]. The SBP procedure produces orthonormal coordinates which are balances. A positive balance means that the group of parts in the numerator has (in geometric mean) more weight in the sample than the group in the denominator (and vice versa for negative balances). Note that amalgamations of parts do, in general, not lead to log-contrasts, unless the composition is extended by a new component or components made up of amalgamated parts.

Several microbiome studies have chosen ilr coordinates using an ad hoc SBP [91, 115]. Interpreting the results of general, blind partitions might not be trivial. In order to choose meaningful parts for balances, an expert opinion is necessary [43], but this is not practical for high-dimensional data.

Recently, some effort has been made to choose balances for classification or prediction purposes, defining the significant balances as those that are associated with the outcome of interest [78, 5, 116, 107]. Rivera-Pinto et. al. [5] introduced the “selbal” algorithm, which identifies the smallest number of microbial features with the highest prediction or classification accuracy of a given response variable. Quinn et. al. [107] introduced “discriminative balance analysis” (DBA-distal), which offers a computationally efficient way to select important 2 and 3-part balances. Distal Balance based disease prediction and biomarker discovery platforms have also been introduced: GutBalance [105] and DisBalance [104]. The most recent balance-based feature selection approach is “codacore” [116] and it finds the sparse subset of balances that are maximally associated with the response variable. “Philr”[106] is a different method from the above-mentioned balance selection methods with respect to SBP construction. It does not use data labels for SBP construction, but the phylogenetic tree.

Rather than building balances with geometric means of parts, amalgamation is proposed as an alternative to balances [44, 117], but amalgams need special care as their meaning changes under perturbation, for instance, when centering data. In fact, amalgamation is a non-linear operation in the simplex endowed with the Aitchison geometry. Moreover, the amalgamation ignores the existence of two different parts in the group. Ratios of involving parts are lost after amalgamation [43]. This is not desirable for biomarker studies, because the association role of microbial features in the groups with disease status is separately important.

We propose a procedure that groups microbial features attending the compositional character of the data making use of the highest possible resolution of microbial features (OTUs). This mathematically consistent aggregation procedure collapses microbial features into units as an alternative to taxon grouping, here called Principal Microbial Groups (PMGs), providing a coherent data analysis for the search of biomarkers in human microbiota.

5.2 MATERIALS AND METHODS

5.3 Overview of Principal Microbial Groups

Principal Microbial Groups (PMGs) procedure creates non-overlapping OTU groups without using taxonomy. The grouping is based only on relative abundances. Thus, microbial features in the same group might have different taxonomy. PMGs offer the possibility of working with coarse groups of OTUs, groups which are not present in a phylogenetic tree. PMGs can be used for facilitating the use of high resolution microbial data in the search of biomarkers.

While grouping, the procedure assigns each OTU to a group (PMG) such that OTUs in the group are highly linearly associated in the Aitchison geometry [74]. Grouping of OTUs in PMGs is an unsupervised R-mode cluster analysis and acknowledges coda methodology. The selection of non-overlapping groups of OTUs is obtained through hierarchical clustering as used to approach principal balances (PBs) [71]. The association between OTUs is determined using the variation matrix that has been proven to be proportional to the square Aitchison distance between parts [74, 45]. These associated OTUs form a PMG and each PMG is represented by the geometric mean of the relative abundances of OTUs, thus reducing dimension of the dataset for further analysis. Note that only the grouped OTUs play a role in each PMG dimension. Thus, it contributes to the better understanding of dimension reduction procedure.

The overview of the PMG procedure is illustrated in Figure 10. The proposed procedure consists of three steps: (1) Select an appropriate SBP for grouping, (2) Choose the optimal number of PMGs and (3) Select compositional biomarkers.

5.3.1 (1) Select an appropriate SBP

Let $\mathbf{x} = (\mathrm{otu}_1, \mathrm{otu}_2, \mathrm{otu}_3, ...\mathrm{otu}_D)$ be a D-part compositional observation, possibly normalized to $\sum_{k=1}^{D} \mathrm{otu}_k = 1$. The data set is then arranged in an (n, D) data matrix $\mathbf{X}$. A hierarchical cluster analysis of the columns of $\mathbf{X}$ (OTUs) is carried out using Ward's method. The variation matrix can be used to define association between OTUs. Variation of parts can also be expressed in terms of the Aitchison distance between parts, because the square root of the variation matrix is actually proportional to the Aitchison distance between parts [45, 74]. The entries of the variation matrix are $\mathrm{var}(\log(\mathrm{otu}_i/\mathrm{otu}_j))$ and they are all positive. The variance of the logratio of two OTUs is 0 if they are equal or if they are perfectly proportional [26], i.e. if otu_i and otu_j are exactly linearly associated [74]. A small variance indicates approximate linear association. The larger the variance is, the more unreliable the proportionality of the two OTUs is, and they likely belong to different groups. Thus, the variation matrix is a natural choice of a distance for merging OTUs.

Each branching of the hierarchical clustering tree is a binary partition that divides the OTUs under the branch into two groups. The procedure is iterated until all groups contain only one single OTU. The number of binary divisions of a group comprising D OTUs to attain the end of the process is $D - 1$. This procedure defines a SBP.

5.3.2 (2) Choose the optimal number of PMGs

Choosing the number of PMGs is critical for the future analysis because the interpretability of constructed balances depends on it. The aim is to get as many groups as possible with a manageable number of OTUs in it to make use of PMG balances in the search of biomarkers. Those groups also should explain the most of the total variance in the dataset. The explained variance of PBs can be used to choose the optimal number of groups. PBs which explain the higher variance than the mean of total variances are chosen to construct PMGs. The minimum number of PMGs is decided by the user. Assume that the SBP process is stopped when the number of groups is z. Denote these groups PMG_j, $j = 1, 2, \dots, z$, so that all otu_i are in one, and only one, PMG. The value assigned to each PMG_j is the geometric mean of the OTUs included. Alternative possibilities are discussed in supplementary data.

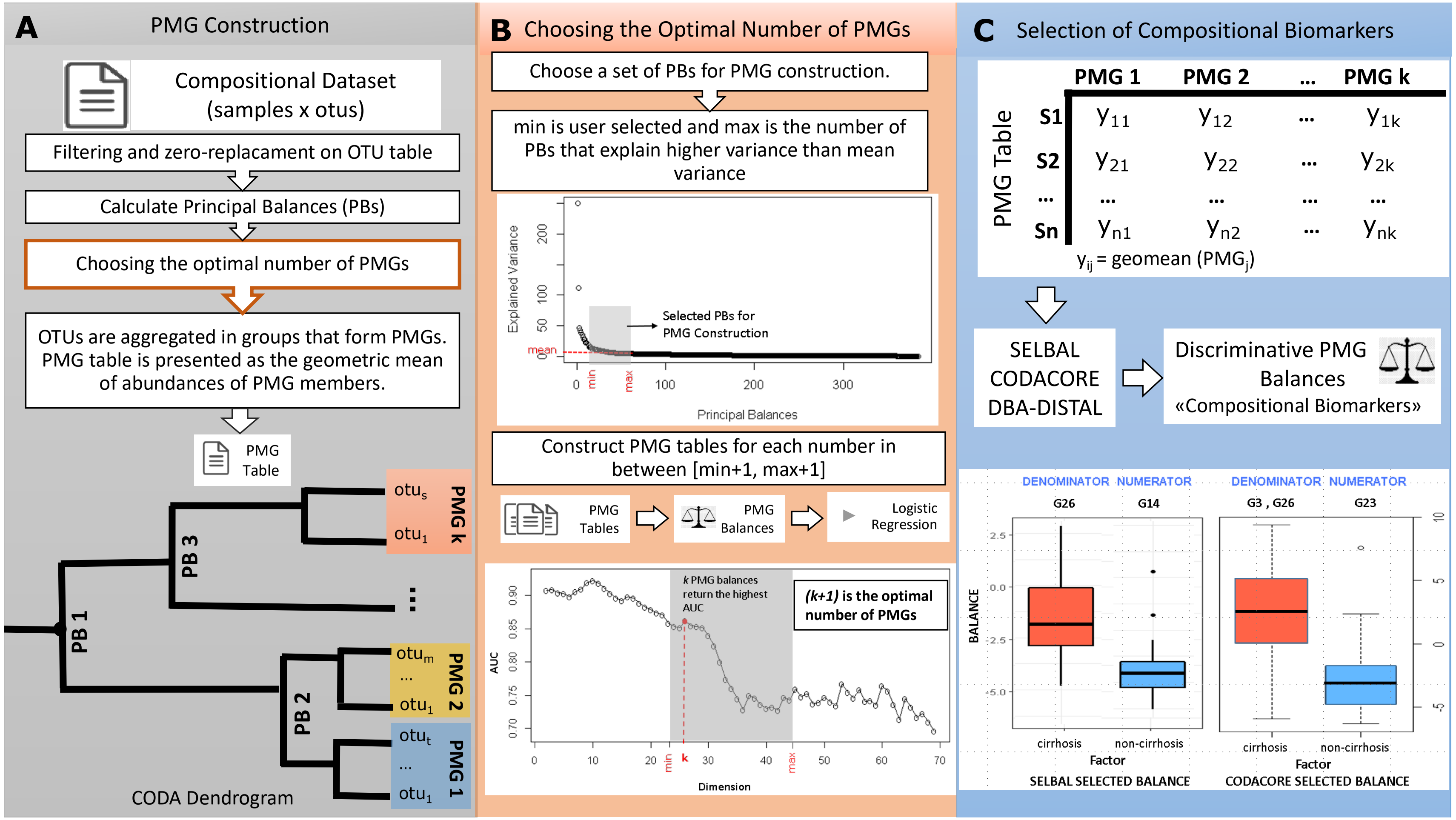

A
PMG Construction
Compositional Dataset (samples x otus)
Filtering and zero-replacament on OTU table
Calculate Principal Balances (PBs)
Choosing the optimal number of PMGs
OTUs are aggregated in groups that form PMGs. PMG table is presented as the geometric mean of abundances of PMG members.
PMG Table
PB 1
PB 2
PB 3
otu$_s$
otu$_1$
PMG k
...
otu$_m$
...
otu$_1$
PMG 2
otu$_t$
...
otu$_1$
PMG 1
CODA Dendrogram
B
Choosing the Optimal Number of PMGs
Choose a set of PBs for PMG construction.
min is user selected and max is the number of PBs that explain higher variance than mean variance
Explained Variance
200
100
50
0
mean
min
max
Selected PBs for PMG Construction
0
100
200
300
Principal Balances
Construct PMG tables for each number in between [min+1, max+1]
PMG Tables
PMG Balances
Logistic Regression
AUC
0.90
0.85
0.80
0.75
0.70
k PMG balances return the highest AUC
(k+1) is the optimal number of PMGs
0
10
20
min
k
30
40
max
50
60
70
Dimension
C
Selection of Compositional Biomarkers
PMG Table
PMG 1
PMG 2
...
PMG k
S1
y$_{11}$
y$_{12}$
...
y$_{1k}$
S2
y$_{21}$
y$_{22}$
...
y$_{2k}$
...
Sn
y$_{n1}$
y$_{n2}$
...
y$_{nk}$
y$_{ij}$ = geomean (PMG$_j$)
SELBAL
CODACORE
DBA-DISTAL
Discriminative PMG Balances
«Compositional Biomarkers»
DENOMINATOR
NUMERATOR
G26
G14
BALANCE
2.5
0.0
-2.5
-5.0
cirrhosis
non-cirrhosis
Factor
SELBAL SELECTED BALANCE
DENOMINATOR
NUMERATOR
G3 , G26
G23
10
5
0
-5
cirrhosis
non-cirrhosis
Factor
CODACORE SELECTED BALANCE

Figure 10: (A) PMG construction procedure starts with a compositional OTU table and ends with a PMG table. In the CODA dendrogram, SBP for principal balances (PBs) is visualized. Choosing n PBs results in $n+1$ PMGs. (B) PBs are utilized to choose the optimal number of PMGs. The min and max number of PBs to include in the PMG construction process is determined. PMG tables are constructed for each number in between [min+1,max+1]. Then, PMG balances (ilr transformed PMGs) are obtained for each PMG table. Logistic regression classifier runs on all PMG balance tables. k PMG balances return the highest AUC value. Thus, the optimal number for PMGs is k+1. (C) Dataset projected on PMGs; the PMG table is calculated so that each group is represented by the geometric mean of the relative microbial abundances in the group ($y_{ij} = \text{geomean}(Si_{\text{otu}_1}, \ldots, Si_{\text{otu}_t})$ is the corresponding PMG$_j$ value for sample i where t is the total number of otus in PMG$_j$). Discriminative PMG balances are chosen by selbal, codacore and DBA-distal and compositional biomarkers are obtained. The box plots show the distribution of selbal selected PMG balance (G26/G14) scores and codacore selected PMG balance (G3,G26/G23) scores for cirrhosis and non-cirrhosis individuals. The PMGs that form the balances are specified at the top of the box plots.

The PMG_j, $j = 1, 2, \dots, z$ form a composition $\mathbf{y}$ with z parts

$$\mathbf{y} = (\mathrm{g_m}(\mathrm{PMG}_1), \mathrm{g_m}(\mathrm{PMG}_2), \dots, \mathrm{g_m}(\mathrm{PMG}_z)),$$

where $z < D$.

The new composition y can be represented by some arbitrary set of $z - 1$ ilr coordinates denoted y_k^*, $k = 1, 2, \dots, z - 1$. That is, for each number of groups z, an $(n, (z - 1))$ matrix of ilr coordinates (PMG Balances) is obtained, and they are used in a logistic regression to predict the presence/absence of the disease. Once the minimum (a) and maximum (b) number of PMGs are chosen, the best accuracy measure (e.g. area under the ROC curve) of these logistic regressions for $z = a, a + 1, \dots, b$ corresponds to the optimal number of PMGs-1.

5.3.3 (3) Select Compositional Biomarkers

PMG balances can be used to construct a dataset to search for microbial groups that are out of balance depending on a factor. The reduced z-part composition $\mathbf{y}$ can be represented by ilr coordinates to obtain PMG balances. If they are defined by means of an SBP, the coordinates will be balances of the form

$$\mathbf{y}^* = K \ln \frac{(\mathrm{g_m}(\mathrm{PMG}_1^+), \mathrm{g_m}(\mathrm{PMG}_2^+), \dots, \mathrm{g_m}(\mathrm{PMG}_{m_+}^+))^{1/m_+}}{(\mathrm{g_m}(\mathrm{PMG}_1^-), \mathrm{g_m}(\mathrm{PMG}_2^-), \dots, \mathrm{g_m}(\mathrm{PMG}_{m_-}^-))^{1/m_-}},$$

$$K = \sqrt{\frac{m_+ m_-}{m_+ + m_-}}, \quad m_+ + m_- \leq z,$$

corresponding to a partition separating the parts $\mathrm{g_m}(\mathrm{PMG}_{m_+}^+)$ from those $\mathrm{g_m}(\mathrm{PMG}_{m_-}^-)$, composed of m_- and m_+ parts, respectively. The $\mathbf{y}^*$'s are the PMG balances.

Selbal [5], codacore [116] and DBA-distal [107] are balance selection methods to find balances associated with response. The input data for those methods can be at any level of the microbial features i.e. phylum, genus, species or OTUs. Genus-level aggregation of microbiome data is the commonly used procedure before starting any analysis, thus obtained balances are basically a ratio of genera. Alternatively, PMGs could provide a set of OTUs whose aggregated ratios are discriminative. The PMG balances are called “compositional biomarkers.” Compositional biomarkers are in line with the recent understanding that diseases are generally associated with a balance of discrete groups of microbial species, as opposed to individual microbes [92, 90, 94, 107].

5.4 Dataset and Preprocessing

To illustrate the proposed procedure and to reveal the biological meaning of PMGs, we choose a cirrhosis dataset [118] because the disease state is considered to be highly predictable by machine learning methodologies [119]. The dataset is available in the Knights Lab GitHub repository https://github.com/knights-lab/MLRepo [120]. There are 130 cirrhosis and non-cirrhosis samples with 2145 features (OTUs). We filtered features that had less than 20 counts in at least 30% of samples, resulting in 130 samples and 385 features. Cirrhosis samples (n=68) and non-cirrhosis samples (n=62) were used for further analysis.

In a microbiome dataset, each observed sample is a composition of microbial features (OTUs). Zero values in the compositional dataset must be handled prior to any analysis, as CODA methods rely on logarithms. The geometric bayesian multiplicative (GBM) method, implemented in the cmultRepl function from the zCompositions package in R, is used for zero replacement [40]. As a result, a closed dataset with no zeros was obtained. Before PMG construction, the optimal number of group has to be determined. As explained in the section “Choose the optimal number of PMGs", 25 was chosen as the minimum and the optimal number was determined as 27. Eventually, cirrhosis dataset was represented with 27 PMGs.

5.4.1 Benchmark Evaluation

We evaluated PMGs with respect to two aspects: (1) PMG balances as a dimensionality reduction method for compositional data, (2) PMG as a feature aggregation procedure that provides an alternative to taxon grouping for construction of microbial balances afterwards used for disease prediction.

First, we bench-marked PMG balances (ilr transformed PMGs) against competing dimension reduction methods designed for compositional data. This includes (i) PCA, (ii) Principal Balances [71] and (iii) Distal balances (DBA-distal)[107]. The OTU tables and genus-level tables (OTU tables agglomerated into genus level) were dimensionally reduced by each of the three methods and reduced datasets were fed to logistic regression (LogReg). Note that PMGs are constructed only on the OTU table, thus PMG balances were not calculated on the genus-level table. For fair comparison between methods, the dataset was reduced to the same number of dimensions. The classification performance of the model was assessed by AUC, the area under the receiver operating characteristic (ROC) curve with ten-fold cross validation. Classification was performed in R using the caret package [121]. More detailed information about benchmarking methods and how to implement them in R are available in supplementary data.

Secondly, whether grouping OTUs as PMGs as an alternative to taxon grouping adds value in terms of creating better balances for classification was assessed. The classification performance of balance selection methods (selbal, codacore and DBA-distal) and selected balances on OTU table, genus-level table and PMG table were examined.

5.4.2 Results

The cirrhosis dataset was preprocessed and a total of 27 PMGs (G1, ..., G27) were identified, as explained in the Material and Methods section. We show that PMG construction is an alternative technique to taxon grouping that enables working with coarse groups of OTUs. PMGs have some interpretation advantages in reducing dimensionality and provide balances of microbial groups that can be used for disease prediction.

5.4.2.1 PMG Balances as Dimensionality Reduction Method

PMG balances were used as a dimensionality reduction method on OTU table for cirrhosis dataset. PMG table with 27 groups was ilr-transformed and 26 PMG balances were obtained. Thus, the cirrhosis dataset was reduced to optimal dimension (26) by using different dimension reduction procedures.

We benchmarked 26 PMG balances against competing dimensionality reduction methods designed for compositional data: PCA, PBA, DBA-distal. LogReg was used for disease prediction. Reduced datasets were constructed on OTU table and on genus-level table separately to compare with PMG balances. Note that PMG balances were not calculated on the genus-level table, they are constructed only on the OTU table for Figure 11. Because, PMGs are designed for grouping OTUs as an alternative to taxon aggregation. The classification performance of the reduced tables (26 dimensions) was reported (supplementary Table 8 and Figure 11-A). Figure 11-B and 2-C show the LogReg classification performances change with the dimension on the reduced datasets obtained by different methods. PMG balances exhibited performance rivaling common dimension reduction methods for compositional data.

PMG balances will not outperform other dimension reduction methods. However, PMG balances have an interpretation advantage compared to other dimension reduction methods. Reducing a dataset by PMG balances creates a ratio of non-overlapping groups and only the grouped OTUs play a role in each PMG dimension. Compared to PCA, each principal component generally involve all parts of the composition in a non-homogeneous way, thus making its interpretation difficult. Moreover, PMG balances offer the possibility of working with coarse groups of OTUs, groups which are not present in a phylogenetic tree. It is assumed that the microbial OTUs related to a given phenotype can be mixed up within coarser units like phylum or genus, leading to degradation of possible associations [6]. Alternatively, representing data by PMGs, one can obtain balances with richer high resolution microbial features that could prevent mixed up associations resulting of taxon aggregation. As a result, PMGs contribute a better understanding of dimension reduction procedures.

5.4.2.2 PMGs as Feature Aggregation Procedure

PMG is an alternative way of grouping OTUs for microbiome research. Whether grouping OTUs as PMGs adds value in terms of creating better balances compared to OTUs and genus level data was assessed. The dataset was redesigned by different data types (OTU table, genus-level table and PMG table) and they were fed to balance selection methods (selbal, codacore and DBA-distal). The classification performance of methods and discriminatory power of the selected balances on different data types were examined. Selbal and codacore have a cross-validation procedure in their model and they return an AUC value for discriminatory power of the selected balances. PMG table exhibited performance rivaling OTU and genus-level tables on selbal and codacore algorithms (supplementary Table 10 and Figure 12-A). Selbal selects a global balance and the performance of the global balance was reported in the Figure 12-A. PMGs provided a small performance boost for selbal. The reason could be that PMGs combined OTUs that have similar discriminative role. The OTU content of selbal selected PMG balance examined in the section “PMG Balances as Biomarker Candidates". The OTU content of PMGs are consistent with the literature findings in term of association with disease (Figure 13-D). A further study is needed for detailed examination.

Unlike selbal and codacore, DBA-distal method returns a dataset that consists of many balances with 2 or 3 parts (distal balances) on the inputted dataset. OTU table, genus-level table and PMG table were fed to DBA-distal method. LogReg performances (AUC) of the distal balances on three different data types were compared. DBA-distal method selects 15 PMG balances, 56 genera balances and 199 OTU balances on the cirrhosis dataset, for fair comparison, the most discriminative 15 distal balances are included in LogReg classification (Figure 12-B). Distal OTU balances have higher AUC values

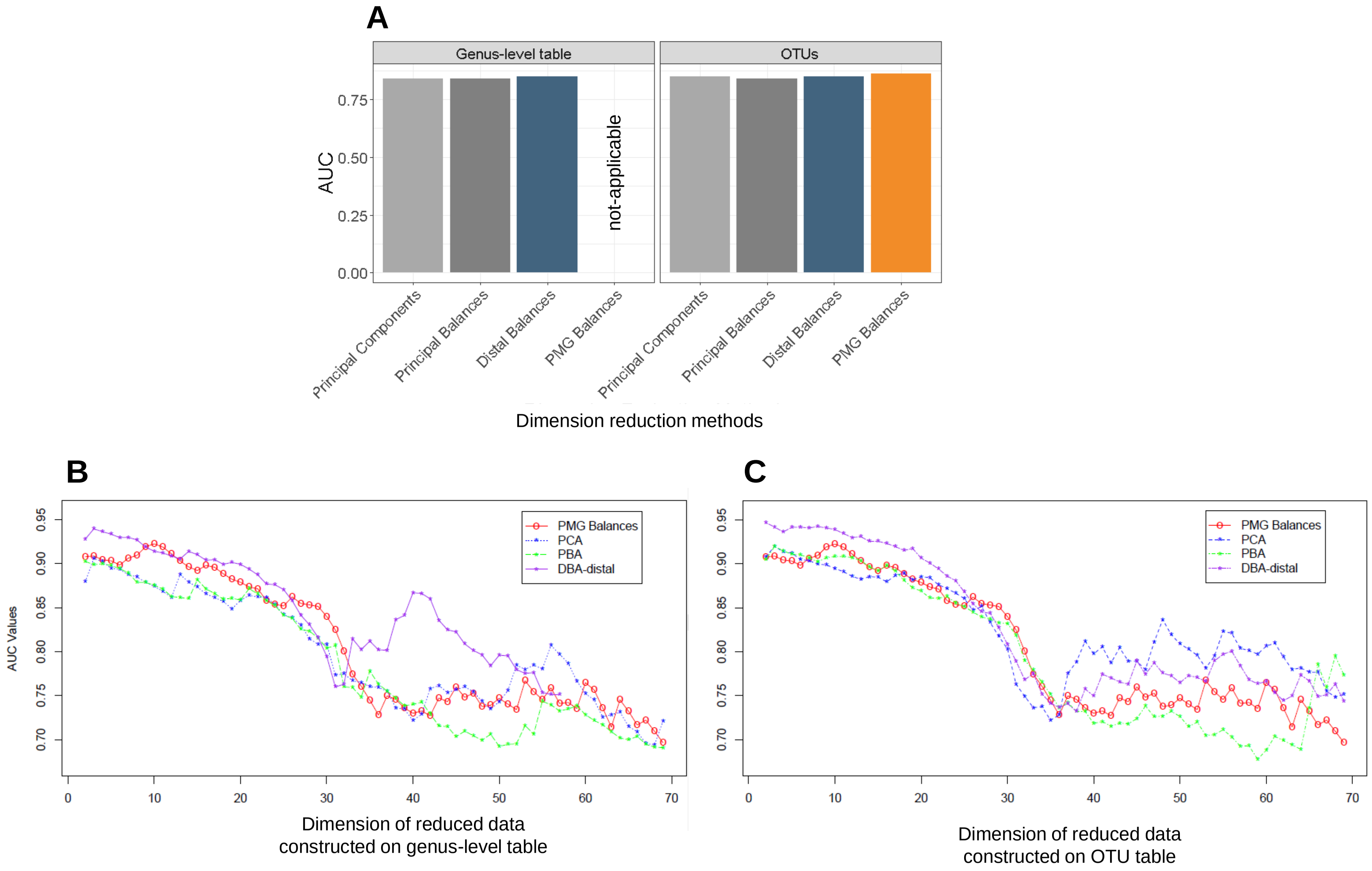


Figure 11: Logistic regression classification performances of dimensionality reduction methods (PCA, PBA and DBA-distal and PMG balances) on OTU table and genus-level table were assessed by AUC. (A) The cirrhosis dataset was reduced to optimal dimension (26) by using different dimension reduction procedures. Optimal dimension is decided based on the number of PMG balances. PMG balances are applied only on OTU table as the procedure is designed to group OTUs as an alternative to taxon grouping. Classification performances of reduced datasets were compared. Figure supplements are available in supplementary table 8.(B) AUC values that change with the dimension of the reduced datasets on genus-level table and (C) OTU table were plotted separately to compare with PMG balances. Note that the reduced dataset with PMG balances are the same in both plots.

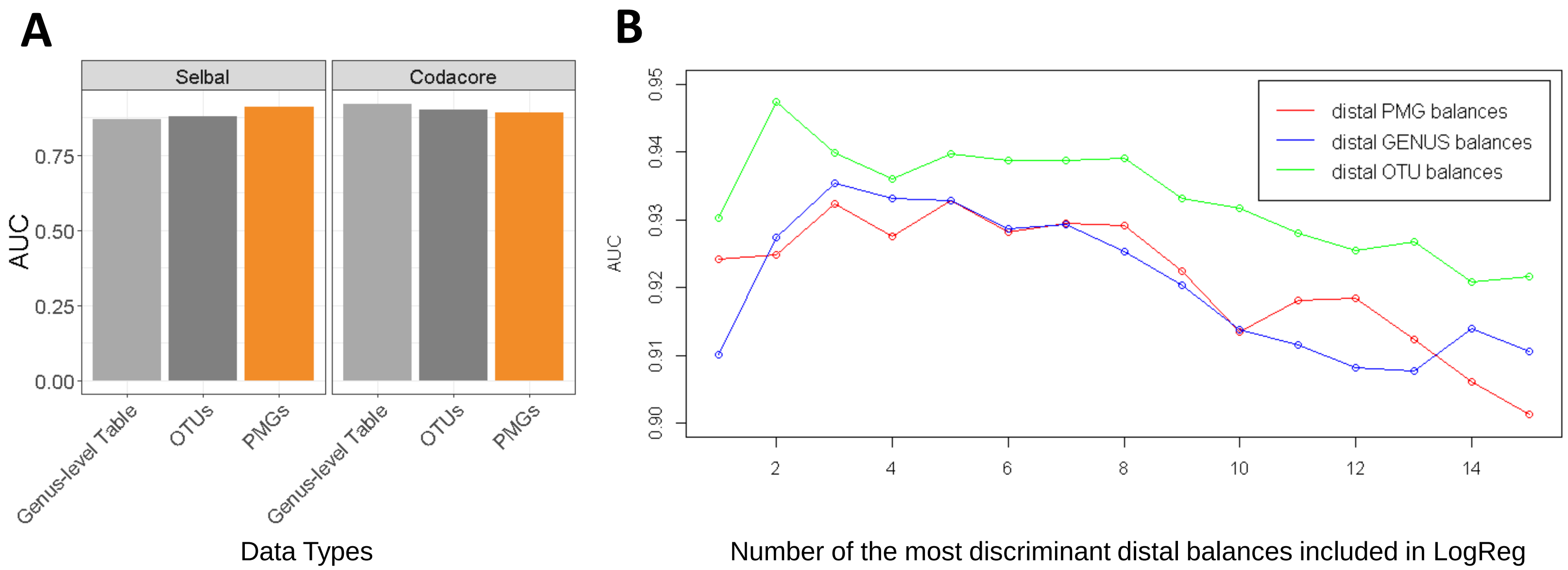


Figure 12: OTU table, genus-level table and PMG table were fed to selbal, codacore and DBA-distal method and the classification performances of the selected balances were compared (AUC). (A) Selbal and codacore have a cross-validation procedure in their model and they return an AUC value for discriminatory power of the selected balances. Global balance performance were reported. Note that the dimension of input PMG table is 27 for those methods. Figure supplements are available in supplementary table 9. (B) OTU table, genus-level table and PMG table were fed to DBA-distal and the classification performances of the most discriminative 15 distal balances on different data types were compared. Figure supplements are available in supplementary table 10. Note that DBA-distal method selects 15 PMG balances, 56 genera balances and 199 OTU balances on the cirrhosis dataset, for fair comparison, the most discriminative 15 distal balances are included in LogReg classification.

than distal genus and PMG balances. On the other hand, distal PMG balances exhibited performance rivaling distal genus balances, noting that genus level is commonly used grouping procedure for high dimensional microbiome data. OTUs are the highest possible resolution of microbial features and the possible associations are more clear on OTU-level data. Grouped OTUs (as genus or PMGs) will be less sparse, thus grouping could cause to degradation of possible associations. PMGs result in a different grouping of OTUs than phylogenetic grouping, and new latent functional features related with disease could be inferred. On the contrary of genus-level table, working with PMGs, one can obtain balances with richer high-level microbial features as biomarker candidates. Figure 13 shows the microbial content of selected balances on different data types. They have overlapping features. It is noticeable that most of the selected taxa by any balance selection methods are already included in PMGs (specified in italic in white boxes). Moreover, the microbial content of PMGs is consistent with the literature in terms of association with disease such as microbes enriched or diminished in cirrhosis patient. PMG balances can be used to enhance existing microbiota analysis pipelines as well as they can be used as a new source in the search of biomarkers.

5.4.2.3 PMG Balances as Biomarker Candidates

The selected PMG balances by balance selection methods are called compositional biomarkers. Compositional biomarkers are in line with the recent understanding that diseases are generally associated with a ratio of discrete groups of microbial species, as opposed to individual microbes [92, 90, 94, 107]. The microbial content of the selected balances on different data types was examined with respect to association with disease mentioned in the literature.

Selected balance by selbal on OTU table was *(Veil.parvula, Mega.micro.)/Bac.uni.* After PMG construction, selbal selected the balance of G26/G14 (global balance). G26 has seven unique species that are *Veil.parvula* and *Mega.micro.* as well as *Fus.nucleatum*, *Fus.periodonticum*, *Camp.-concisus, St.mutans, St.anginosus*. Of these, *St.anginosus, Camp.concisus* and *Veil.parvula* are specifically mentioned in the literature as cirrhosis related species [119, 118]. *Fusobacterium* is considered to be associated with cirrhosis at genus level, however the aforementioned *Fusobacterium* species have not been specifically mentioned in the literature. In contrast with G26, G14 has nineteen unique species belonging to *Bacteroides, Odoribacter, Parabacteroides, Coprobacter* and *Barnesiella* genera. Of these, *Bacteroides* is specifically mentioned in the literature as the dominant genus in both cirrhosis and non-cirrhosis groups, but significantly diminished in the liver cirrhosis group [119, 118, 87]. *Odoribacter* and *Parabacteroides* are also mentioned as diminished genera in cirrhosis individuals [118]. To the best of our knowledge, *Coprobacter* and *Barnesiella* genera in G14 have not been specifically mentioned in the cirrhosis biomarker literature, but the *Porphyromonadaceaa* that is the family of *Coprobacter* and *Barnesiella* genera is mentioned as diminished in patients [122]. Overall, G26 includes species mostly related to, i.e. enriched, in patients with cirrhosis, whereas G14 includes species that are diminished in cirrhosis patients. Since the balance of G26/G14 is discriminant between diseased and non-diseased samples, balance of those species should be one of the priorities for future therapies to prevent and treat cirrhosis. On the other hand, selbal selects the balance of *Megasphaera/Unc.Erysip* genera on genus-level table. *Megasphaera* is mentioned as cirrhosis enriched taxa [118, 119], but *Unc.Erysip* has not been mentioned in the literature.

Selected balance by codacore on OTU table was (*Lac.saliva., Megas.-micro.*)/(*Adler.equ., Alis.indis*). After PMG construction, codacore selects the balance of (G26,G3)/G23 on PMG table. G3 has five

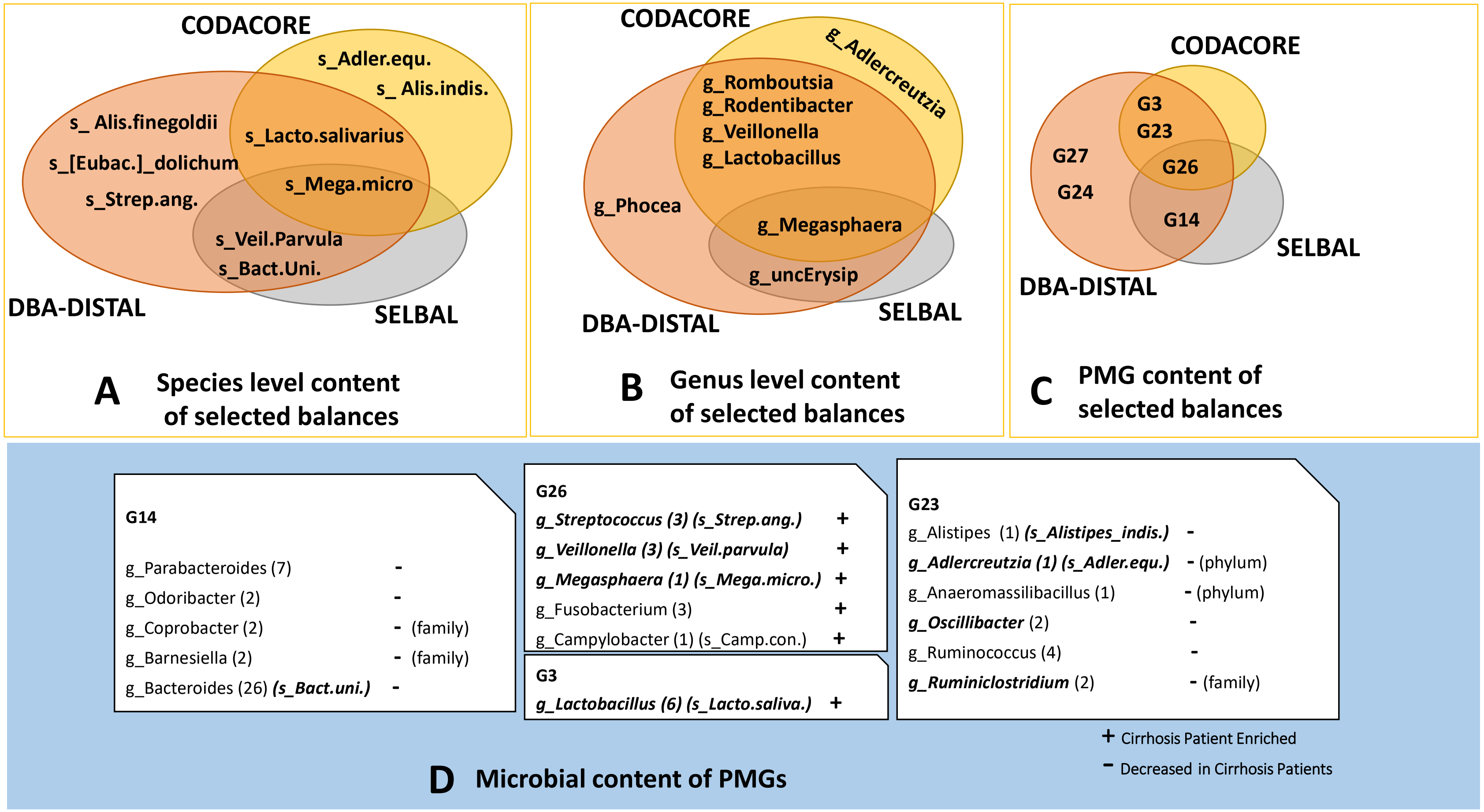

CODACORE
s_Adler.equ.
s_ Alis.indis.
s_ Alis.finegoldii
s_Lacto.salivarius
s_[Eubac.]_dolichum
s_Mega.micro
s_Strep.ang.
s_Veil.Parvula
s_Bact.Uni.
DBA-DISTAL
SELBAL
A
Species level content of selected balances
CODACORE
g_Adlercreutzia
g_Romboutsia
g_Rodentibacter
g_Veillonella
g_Lactobacillus
g_Phocea
g_Megasphaera
g_uncErysip
DBA-DISTAL
SELBAL
B
Genus level content of selected balances
CODACORE
G3
G23
G27
G26
G24
G14
SELBAL
DBA-DISTAL
C
PMG content of selected balances
G14
g_Parabacteroides (7) -
g_Odoribacter (2) -
g_Coprobacter (2) - (family)
g_Barnesiella (2) - (family)
g_Bacteroides (26) (s_Bact.uni.) -
G26
g_Streptococcus (3) (s_Strep.ang.) +
g_Veillonella (3) (s_Veil.parvula) +
g_Megasphaera (1) (s_Mega.micro.) +
g_Fusobacterium (3) +
g_Campylobacter (1) (s_Camp.con.) +
G3
g_Lactobacillus (6) (s_Lacto.saliva.) +
G23
g_Alistipes (1) (s_Alistipes_indis.) -
g_Adlercreutzia (1) (s_Adler.equ.) - (phylum)
g_Anaeromassilibacillus (1) - (phylum)
g_Oscillibacter (2) -
g_Ruminococcus (4) -
g_Ruminiclostridium (2) - (family)
+ Cirrhosis Patient Enriched
- Decreased in Cirrhosis Patients
D
Microbial content of PMGs

Figure 13: The microbial content of balances selected by different methods are presented at (A) species, (B) genus-level and (C) PMG level. Different colors indicate different balance selection methods. The selected PMGs (G14, G26, G23) and the microbial content of PMGs is presented at genus level in default in the blue box (D). In each PMG (white boxes), the number of OTUs belonging to each genus is specified in parenthesis. If a genus is in a PMG selected by any balance selection method, then it is specified in italic. Selected species by any balance selection methods are specified in the parenthesis next to each genus. Enriched and diminished genera in cirrhosis (according to literature) are specified with + and − in the PMGs, respectively. If the association is reported on another phylogeny, it is specified in the parenthesis next to + and − signs. Note that DBA-distal method selects 15 PMG balances, 56 genera balances and 199 OTU balances on the cirrhosis dataset. For the simplicity of the figure, only a set of the most discriminative distal balances are presented in panel A, B and C and only the microbial content of the most two discriminative distal PMG balances are presented in panel D.

unique species from *Lactobacillus* genus. *Lactobacillus* is mentioned in the literature as the genus that increases in cirrhosis patients, specifically *Lac.salivarius* species [118]. In contrast with G3 and G26, G23 has seven unique species from *Ruminococcus, Ruminiclostridium, Oscillibacter, Alistipes, Adlercreutzia, Anaeromassilibacillus* genera. Of these, *Ruminococcus, Oscillibacter* and *Alistipes* are mentioned as the phyla diminished in cirrhosis patients [118]. Other listed genera are associated at higher level such as phylum and family. Similar to the selbal selected balance, G26 and G3 include species that are mostly enriched in cirrhosis patients, whereas G23 has species that are mostly diminished in cirrhosis patients. The balance of (G26,G3)/G23 might be another important biomarker candidate for future therapies to prevent and treat cirrhosis. On the other hand, codacore selects a balance of (*Lactobacillus, Megasphaera, Veillonella, Rodentibacter*)/(*Adlercreutzia, Romboutsia*) on genus-level table. Among them, *Rodentibacter* and *Romboutsia* have not been mentioned in the literature and *Adlercreutzia* has only been mentioned at phylum level.

The selected genera balances are not well supported by literature, whereas the microbial content of the selected PMG balances are consistent with the literature on species level or on higher taxonomic level. Thus, it can be concluded that the reliability of the selected genera balances is controversial [6].

Compositional Biomarker. Selected PMG balances by balance selection methods be defined as compositional biomarkers. Selbal and codacore methods select a single balance, whereas DBA-distal selects many balances with 2 or 3 parts. The box plots at the bottom of Figure 10-C show the PMGs that constitute the balances selected by selbal and codacore on cirrhosis dataset. They can be directly interpreted as an important ratio of groups of microbial features that are highly discriminatory between cirrhosis and non-cirrhosis individuals. DBA-distal method selected 15 distal PMG balances. It is noticeable that PMG balances together selected by selbal and codacore were the most two discriminative distal PMG balances on cirrhosis dataset. Prediction power of balance combinations was tested by LogReg. The classification performance of two PMG balances together selected by selbal and codacore is higher than separately tested balances. The classification performance of the most three discriminative distal PMG balances is the best as can be seen in Table 6. Moreover, the PMGs included in the compositional biomarker explain the most of the total variance covered by all PMGs. Figure 14 shows the compositional form biplot of PMGs included in compositional biomarkers on the cirrhosis dataset. Explained variance is %89.5 of the total variance retained by the 6 selected PMGs.

Table 6: Classification performances of PMG balance combinations selected by balance selection methods.

Method	Compositional Biomarkers	AUC
Selbal	(G26/G14)	0.91
Codacore	(G3,G26)/G23	0.90
DBA-Distal (2 balance)	(G26/G14) (G3/G26)/G23	0.92
DBA-Distal (3 balance)	(G26/G14) (G3/G23) (G27/G24)	0.93

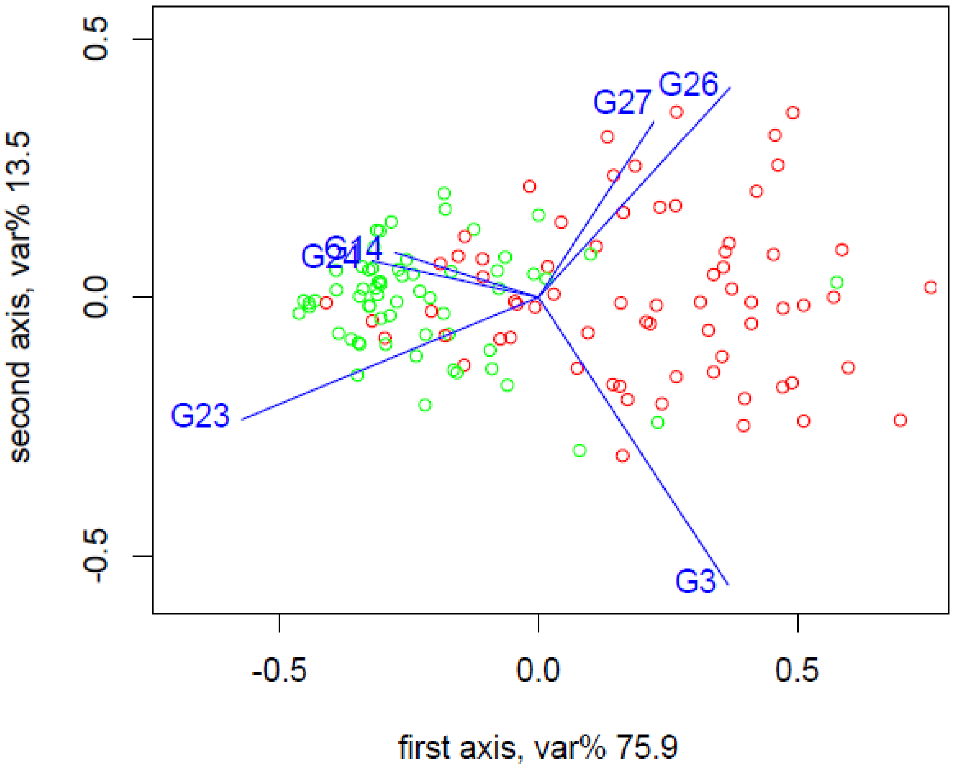


Figure 14: Form biplot of selected PMGs by balance selection methods (G26, G14, G3, G23, G24, G27). Red points indicate cirrhosis samples whereas green points indicate non-cirrhosis samples in the dataset. Explain variance is %89.5 of the total variance retained by the 6 selected PMGs.

5.4.2.4 CODA Dendrogram to Discover Discriminatory Power of the Balances

The CODA dendrogram is a powerful tool to explore a compositional dataset. The aim of the CODA dendrogram is to represent most of the information contained in the SBP in a comprehensive plot [81, 123, 124]. Figure 15-A shows a list of species mentioned in the literature as associated with cirrhosis and non-cirrhosis samples and in which PMGs those species are located. In Figure 15-B, the CODA dendrogram of PMGs is presented. Red and green horizontal bars represent the cirrhosis and non-cirrhosis samples, respectively. The length of each colored horizontal bar is proportional to the balance contribution to the total variance of the sample. The discriminatory power of each balance can be visually seen in the dendrogram. The first balance has almost double variance in cirrhosis patients than in non-cirrhosis patients. Reviewing literature about taxa associated with cirrhosis, cirrhosis related genera are all placed in PMGs located at the bottom of the coda dendrogram, whereas non-cirrhosis related genera are all placed in PMGs located on the upper side of the coda dendrogram.

The location of PMGs included in compositional biomarkers on the CODA dendrogram reveals the discriminative property of selected balances. The numerator groups (G3, G26, G27) include bacteria that have been mostly associated with cirrhosis and lie on the upper part of the dendrogram, whereas the denominator (G14, G23, G24) groups include bacteria that are enriched in non-cirrhosis (diminished in cirrhosis) and lie at the bottom of the dendrogram.

The members of PMGs are consistent with the literature in terms of association with cirrhosis disease. However, we can make inferences using the CODA dendrogram on association of species that are not previously reported with cirrhosis in the literature. For example, *Campylobacter* and *Veillonella* are two cirrhosis related genera mentioned in the literature. The cirrhosis dataset has two species that belong to *Campylobacter*: *Camp.concisus* and *Camp.coli*. *Camp.concisus* is a cirrhosis associated species specifically mentioned in the literature [118], and is located in G26. There is not any finding about *Cam.coli* specifically in the literature that we are aware of and it is located in G24. Since G24 lies at the bottom of the CODA dendrogram, there is a high probability that *Camp.coli* should not be strongly enriched, relative to the other species, in cirrhosis. Similarly, the cirrhosis dataset has two species that belong to *Veillonella*: *Veil.parvula* and *Veil.seminalis*. *Veil.parvula* is a cirrhosis associated species specifically mentioned in the literature [119] and is located in G26, whereas *Veil.seminalis*, which is not mentioned specifically in the cirrhosis literature, lies in G17 and each group is laying on different sides of the CODA dendrogram. Since G17 lies at the bottom of the CODA dendrogram, there is a high probability that *Veil.seminalis* should not be strongly enriched with respect to *Veil.parvula* in cirrhosis. *Streptococcus* is another genus strongly associated with cirrhosis in the literature. In the CODA dendrogram, all species belonging to the *Streptococcus* genus were located in three PMGs: G11, G26 and G27. Two of 22 species belong to G11, which lies at the bottom of the CODA dendrogram. The remaining 20 out of the 22 species belong to G26 and G27 laying on the upper side of the CODA dendrogram. *S. anginosus, S. parasanguinis* and *S. salivarius* species are specifically mentioned in the literature as enriched in cirrhosis patient [119, 118]. *S. anginosus* is located in G26 and *S. parasanguinis* and *S. salivarius* are located in G27. *Streptococcus* species in G26 and G27 have high potential to be relatively enriched in cirrhosis, whereas species in G11 have a high potential of being relatively diminished in cirrhosis samples.

It is important to note that the balance of species is the key aspect to take into account, not the individual species. It is because focusing on a single bacteria and trying to find association with the classical

A

	Genera	Species	PMGs	Literature Support
Cirrhosis Related Phylotype in Literature	Veillonella	Veillonella parvula	G26	yes
		Veillonella seminalis	G17	yes
		Veillonella dispar	-	yes
		Veillonella atypica	-	yes
	Campylobacter	Campylobacter-concisus	G26	yes
		Campylobacter_coli	G24	no
	Streptococcus	Streptococcus anginosus	G26	yes
		Streptococcus parasanguinis	G27	yes
		Streptococcus salivarius	G27	yes
		Streptococcus_orisratti	G11	no
		Streptococcus_gallolyticus	G11	no
		other Streptococcus spp.	G26, G27	no
	Haemophilus	Haemophilus parainfluenzae	G13	yes
	Prevotella	-	G2, G11, G20, G21	yes (as genus)
	Clostridium	Clostridium_perfringens	G12	yes
		other Clostridium spp.	G16, G17, G24	no
Non-Cirrhosis Related Phylotype in Literature	Eubacterium	Eubacterium eligens	G25	yes
		Eubacterium rectale	G25	yes
		other Eubacterium spp.	G17, G24	no
	Alistipes	Alistipes putredinis	G15	yes
		Alistipes finegoldii	G15	yes
		other Alistipes spp.	G15, G23, G25	no
	Bacteroides	Bacteroides_uniformis	G14	yes
		other Bacteroides spp.	G11, G14, G17	no
	Odoribacter	Coprobacter_fastidiosus	G14, G15	yes (as genus)
	Parabacteroides	Parabacteroides_distasonis Parabacteroides_johnsonii	G14	yes (as genus)

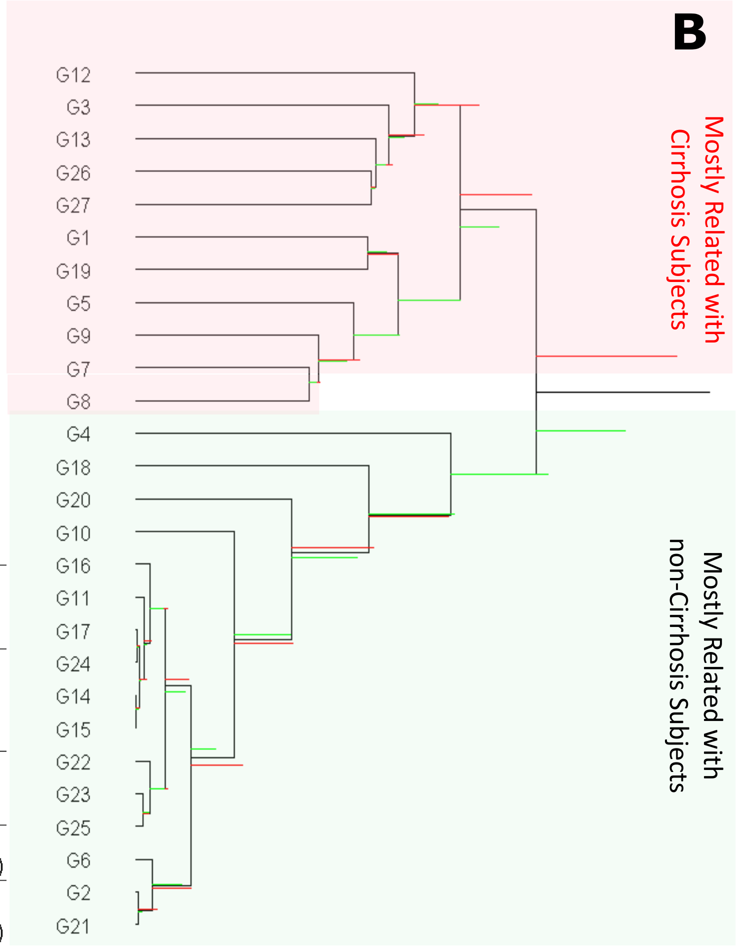


Figure 15: Species mentioned in the literature as associated with cirrhosis and non-cirrhosis samples take place on the different side of the first balance of the CODA dendrogram. (A) Cirrhosis and non-cirrhosis related genera mentioned in the literature are listed. Then, species exist in cirrhosis dataset belong to each genus are listed. In which PMGs those species are located is determined. If there is a research that mentions a species specifically associated with cirrhosis, then literature support is marked as "yes". Some phylotypes are only associated with cirrhosis as genus, so they are marked 'yes (as genus)' under the literature support column. (B) Coda dendrogram of PMGs for the cirrhosis dataset. Red and green horizontal bars on the coda dendrogram represent the cirrhosis and non-cirrhosis samples, respectively. The length of each colored horizontal bar is proportional to the balance contribution to the total variance of the sample. Cirrhosis related genera mentioned in the literature are mostly placed at the bottom of the CODA dendrogram, whereas non-cirrhosis related genera mentioned in the literature are mostly placed in the upper side of the CODA dendrogram.

statistical and machine learning techniques might detect some species that might be delusively associated with disease since the correlation of relative abundances is unreliable.

5.5 Discussion and Conclusion

Recognising microbiome datasets as compositional data leads researchers to utilize log-ratio methodology that brings a new perspective to biomarker research. Instead of focusing on a single or a group of microbial features assumed to be associated with a disease, focus is placed on ratios of microbial features - balances of species - and shall be one of the priorities for future therapies to prevent and treat diseases.

However, a known obstacle in the construction of balances is the choice of partition such that the resulting balances are meaningful. Some effort has been made to choose balances for classification or prediction purposes, defining the significant balances that are associated with the outcome of interest [107, 116, 5]. Philr [106] is another study that define balances utilizing phylogenetic tree. Phylogenetic agglomerated data might not be suitable for biomarker research since the microbial OTUs related to a given phenotype can be mixed up within coarser units like phylum or genus [6].

In this study, we introduce a novel SBP methodology utilizing principal balances that naturally groups microbial features based only on relative abundances making use of the highest possible resolution of microbial features. It offers the possibility of working with coarse group of OTUs, which are not present in a phylogenetic tree. Each PMG could contain species from different genera so that the constructed balances based on PMGs have a unique microbial characterization. Figure 16 shows the frequencies of genera represented in each PMG. There are cases in which many genera participate in one PMG and conversely there are PMGs that contain exclusively OTUs coming from a single genus.

Filtering options and the number of PMGs constructed on OTU table could change the members of PMGs. Even though the number of OTUs change in each PMG with the total number of PMGs, core OTUs usually lie in the selected discriminative PMG balances. The construction of a different number of PMGs will not affect the core species composition in the groups dramatically, but will change their density. The stability of PMGs has been assessed under re-sampling and changing the sample size (similar to subsampling). The discussion about the stability of PMGs is available in supplementary data. The conclusion is that construction of PMGs are quite stable.

PMG balances have an interpretation advantage compared to other dimension reduction methods. Reducing a dataset by PMG balances creates a ratio of non-overlapping OTU groups. Comparing to PCA dimension reduction, only the grouped OTUs play a role in each PMG dimension, but each principal component does not create a discrete groups of OTU, thus making its interpretation difficult.

The important PMG balances are determined using discriminative balance selection methods: selbal, codacore and DBA-distal. The results show that the PMG balances selected by those methods (compositional biomarkers) are highly discriminatory for the cirrhosis dataset. The content of balances are reliable since they include microbial features supported by literature. Combination of selected balances increases the classification performance for predicting cirrhosis. A set of the most informative distal PMG balances has a potential to be a set of strong biomarker candidates.

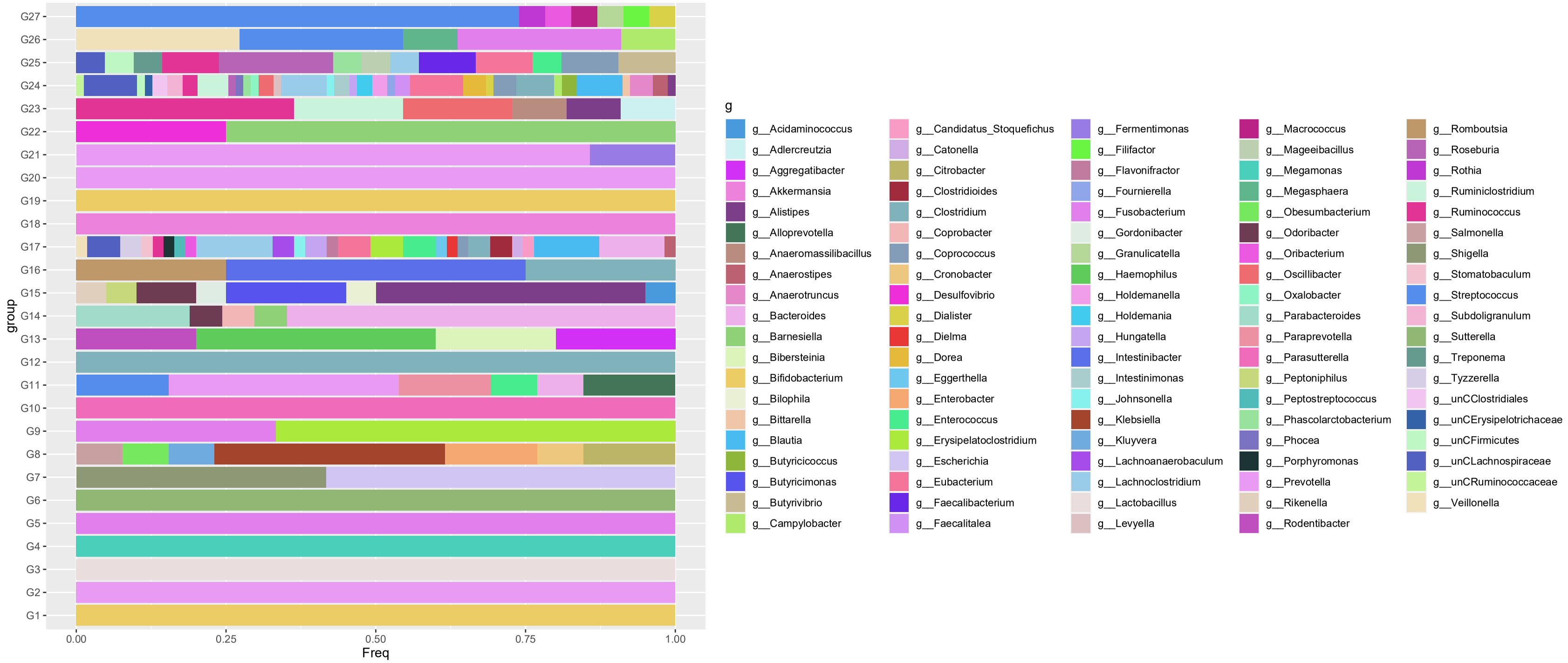


Figure 16: The frequency of genera (g) in each PMG. The members of PMGs are approximately proportional across samples. There are cases in which many genera participate in one PMG and conversely there are PMGs that contain exclusively OTUs coming from a single genus.

The proposed mathematically consistent aggregation procedure collapses OTUs into PMGs as a new alternative to taxon grouping and provides a possibility of working with high resolution microbial features. PMGs overcome the high dimensionality problem of analyzing microbiome data. PMG balances provide a coherent data analysis in the search of biomarkers and have a potential to identify biomarkers candidates. Extra Materials are available in Appendix A.

5.6 Key Points

- High dimensionality, sparsity, and the compositional character of microbiome data present statistical challenges on the way of translating research to clinical practice. Recently, the use of the log-ratio methodology developed for compositional data to process statistically the microbiome has been shown to be a successful option for biomarker research.
- Taxon grouping of microbiome data and inferences based on genera level, or the attempt to identify a single bacterial species associated with a disease, are up to now the main techniques for biomarker studies. As an alternative to taxon grouping, PMGs offer the possibility of working with coarse groups of OTUs, groups which are not present in a phylogenetic tree. PMGs can contain species from different genera so that constructed PMG balances have a unique microbial characterization other than phylogenetic agglomeration.
- Reducing dimensionality of the data by PMGs contributes to the better understanding of dimension reduction procedures. PMG balances creates a ratio of non-overlapping OTU groups and only the grouped OTUs play a role in each PMG dimension. Representing data by PMGs, one can obtain balances with richer high resolution microbial features. Discriminative balance selection methods can be used to determine important PMG balances, termed “compositional biomarkers." Compositional biomarker can be directly interpreted as an important ratio of two groups of microbial features that are discriminatory between health status. PMG balances can be used to enhance existing microbiota analysis pipelines as well as they can be used as a new source in the search of biomarkers.
- A cirrhosis dataset has been analyzed as a demo to illustrate how PMG balances work. Most PMG members of the compositional biomarkers selected by balance selection methods are individually consistent with the literature in terms of association with disease. We strongly emphasize that researchers should focus on compositional biomarkers, preferably represented by balances, to develop promising therapies.

CHAPTER 6

PRINCIPAL MICROBIAL GROUPS FOR MICROBIAL TRANSMISSION

Gaining insights on the ecology of indoor microbiota is the first step towards understanding potential relationships with health outcomes. Built environment microbiome analysis may help tracking in biothreats and diseases, and so developing early warning systems. Recently, there has been ongoing research, developing knowledge and techniques to reveal environmental microbial transmission mechanisms and microbial transmission networks [125, 126, 127, 128]. This chapter focuses on the microbial transmission mechanisms in a hospital environment. An experiment was conducted in the Erciyes University Hospital for this purposes, and swab samples were gathered from the Intensive Care Unit (ICU) to construct microbiome profiles. Microbial transmission is carried out between objects, so it is naturally expected that resulting microbiome profiles of samples should have similar microbial structure. To track the contamination between samples, not taxonomic changes but rather OTU/ASV abundance changes between samples need to be investigated. Principal Microbial Groups procedure was applied to microbial transmission dataset in order to analyze the contagion between samples.

6.1 Introduction

In the developed world, people's natural ecosystem has been restricted to the built environment, an average of 90 % of our lives takes place indoors. We live in a highly interconnected world, not only with living things, but also everything surrounding us. The last 10 years, built environments have been considered not only habitats for humans; but also for diverse microbes and we live with microorganisms that can have direct or indirect effects on the quality of our living spaces, health, and well-being in the buildings [129]. Modern buildings are equipped with surfaces and environmental systems designed to reduce the potential for microbial life to flourish. This fundamental shift in our lifestyle is likely impacting on the development and function of our immune systems in ways that we are only beginning to understand [130] .

We have witnessed an explosion of technologies and informatics pipelines including DNA sequencing-based approaches to study and explore these microbial communities in-depth living in buildings and the overall built environment. These investigations have facilitated understanding microbes that surround us in our daily lives and their impacts on our lives. The emergence of the "microbiology of the built environment" field has required bridging disciplines, including microbiology, ecology, building science, architecture, and engineering [129]. The main aim of this emerging field is to understand the sources of microbes in built environment and estimating the structure the distributions and abundances of microbes within buildings.

Inanimate environments can be considered as a potential utility for identifying and tracking bacterial diversity. Research results indicate that people, animals, plants, outdoor air communities are the major microbial source for indoor microbiome. Moreover, ventilation strategies are also one of the important factors for understanding result of research on microbial community dynamics in the built environment [131, 132]. Another study reveals that working with crops or animals influence the microbiota inside homes [133].

Various microbiome studies revealed that microbial communities not only have relationship among them but also very strong relationship with the environmental characteristic and geographical locations. Studies in different counties revealed that every environment and surfaces have their own characteristic microbial communities and it is not randomly colonized [134]. Analysis of metagenomics data obtained from different biogeographic sites in different time periods show that similar microbiome communities are grouped with high probability in the same biogeography [135]. There are efforts on public health using built environment studies such as sampling public subways in different cities all over the world. Researchers sequence DNA from surfaces in the subways, determine the microbial diversity and microbial sources. Their aim is to develop a "pathogen map" of a city [136].

Microbiome studies also change the definition of sterilized surfaces. Research investigating surfaces for biological diversity revealed that many surfaces known as sterilized actually are not sterilized, some characteristic microbiome colonization had identified [137, 138, 139, 140, 141]. In an ICU biodiversity study, the presence of the species belonging 15 different phyla of bacteria were reported using 16s rRNA sequencing [142]. Next generation sequencing analysis of samples from ICU showed that nosocomial agents are found together with commensal bacteria in the environment on inanimate surfaces [143]. Surface to surface transmission of microbes depends on the characteristics of the microbes and the surface itself [127]. Commonly touched objects are potential hotspots which can facilitate the exchange of microbes during direct hand contact [127].

Machine learning and pattern recognition techniques were applied to microbiome studies to determine the hidden rules underlying microbial structure. For example, pattern reorganization techniques were successfully utilized for the studies such as comparing finger microbiomes and computer keyboard microbiomes to identify who touched which keyboard [144]; analyzing swabs taken from different shoes and room floors to identify who was in which room [145].

Built environment analysis may help in tracking biothreats and diseases and so developing early warning systems. Understanding the determinants of the indoor microbiota is the first step toward understanding potential relationships with health outcomes.

6.2 Nosocomial Infections and Indoor Microbiome

The significance of the indoor microbiome has been critical during the COVID pandemic. How long a virus can survive in the air or on surfaces is a key question. It is well known that contaminated surfaces are the significant vectors in the transmission of the infection both in hospitals and in the community [146, 147, 148].

Nosocomial infections are a worldwide health problem and one of the major sources of morbidity and mortality. Especially, outbreaks of Multi Drug Resistant (MDR) pathogens within Intensive Care Units (ICUs) constitute a grand healthcare threat on immune supressed patients. An important strategy for the control of hospital infections is to prevent pathogen transmission in the hospital environment. Transmission of microorganisms from reservoirs within the built environment to human occupants has been historically studied focusing on pathogens; however, communities of microorganisms can spread through the interaction of their carriers (e. g., air and surfaces) [128].

the ICU is a closed environment and has its own microbiome community. Drifting microbiome habitats from one point to another may cause infection. Hewit et al. [143] sampled two Neonatal Intensive Care Units (NICU) and analyzed the bacterial diversity. Their findings provided evidence that NICU inanimated hospital environments harbor a high diversity of human-associated bacteria. So, inanimated hospital environments can be considered as the potential utility for identifying and tracking bacterial diversity. It was reported that patient follow-up folders were contaminated with around 63% - 83% percent pathogens [149]. Another study reported important resistant pathogens were isolated from sampled objects such as automated censored sinks, stethoscope, computer keyboard, pen, folders and watch etc. [150, 151, 152]. Another study shows that microbes in hospital water can also cause nosocomial infections [131]. These studies support that the source of infections in ICU are sourced in objects in the ICU unit and infection might be caused by environmental transmission.

6.3 Microbial Transmission Modelling

Microbiome studies in the literature have been focused on two main aims: (1) to divulge any global association between microbiome data and phenotype of interest; (2) to specify the microbial feature in the data that are related with outcome. [5]. Recently, research has started developing knowledge and techniques to reveal environmental microbial transmission mechanisms and microbial transmission networks [125, 126, 127, 128].

6.3.0.1 How to Define Microbial Transmission ?

It is plausible to consider the microbial transmission problem to be similar to the problem of determining the differentially abundant OTUs between samples. It could be hypothetically assumed that if there is not a transmission event between a pair of sampled environments, then their compositions should differ at relatively significant extend. However, in practise, this assumption might not hold for microbial transmission in built environment problems. This is due to the fact that, as revealed by various microbiome studies, microbial communities not only have transmissional relationships among them but also are shaped dominantly by the environmental characteristics [1]. Each physical surface in built environments has its own characteristic microbial community, and it is not randomly colonized but shaped by certain ecological drivers [134]. In accordance with the literature, it could be assumed that the objects in the same indoor environment should have similar microbial composition structure. For any microbial transmission experiment objects in an indoor environment, objects are supposed to have similar structure before any contamination, since they were located in the same environment and it makes it possible to track contagion among objects.

The first step of microbiome data analysis is sample sequencing and the construction of OTUs. Microbiome data consists of the relative abundances of the observed OTU counts. Each OTU vector is an abundance vector and it is the only data used to decide microbial transfer.

There are two possible ways of describing “contagion” or “microbial transfer” between two objects:

1. Source object has the same OTUs as the target object before contagion. Contagion occurs. Then both source and target objects have that OTU, but only abundance levels of OTUs change; source sample might have lower or higher OTU abundance than before.
2. Source object has at least one different OTU from the target object before contagion. Contagion occurs. It is supposed that the OTU is transferred to target object and target object has lower OTU abundance than source object. Eventually both source and target object have that OTU.

In a realistic scenario, it is not possible to know the microbiome of each object before contagion. If the previous microbiome of object before contagion is not known, it is not possible to be sure whether an OTU was present there before contagion or it had transferred from source object. But, we could infer that if there is a contagion between two objects, they both must have shared OTUs.

The general approach taken for most of the microbiome studies is summarizing microbial abundance by agglomerating taxa to any rank however, taxonomic changes do not help in understanding the transmission patterns. For this, changes in OTU abundance, rather than taxonomic changes, between samples need to be investigated. The total number of OTUs in any microbiome study are generally vast and the investigation of abundance changes on OTUs individually is intractable. This is not only due to the large number of OTUs, but also because it is only possible to measure relative abundances, and an apparent increase of the relative abundance can be due to an increase in the abundance of the OTU in consideration, or to a decrease in the abundances of the other OTUs in the sample. On the other hand, if microbial transmission is carried out between objects, it is expected that resulting samples ought to have similar microbial structures. In this case, not taxonomic changes but OTU abundance difference between samples need to be investigated. OTUs that play a role in the transmission, therefore, should be determined.

Grouping correlated OTUs can help to reduce dimensionality and make it possible to investigate abundance differences on groups of OTUs. Principal Microbial Groups (PMGs) can be utilized for this purposes.

6.4 Controlled Experiment

6.4.1 Sampling

A controlled experiment conducted in Erciyes University Hospital and swab samples gathered from an Intense Care Unit (ICU) to construct microbiome profiles. 25 objects that were used in daily routine by doctors and ICU personnel were determined in the ICU. One person conducted the contagion experiment and she touched each object in a predetermined order one by one at least for 45 seconds. After the touches, objects were sampled. Swab samples were gathered from the contamination points of objects. Some objects contaminated couple times in the experiment and sampled more than once.

So, a total of 29 swab samples were collected out of 25 objects. Figure 17 shows the experiment path, list of objects, and their respective locations. Object 1 was sampled three times as S1, S2 and S13. Object 10 was sampled twice as S10 and S16. Object 4 was sampled twice as S4 and S17.

6.4.2 DNA isolation

DNA isolation and 16S rRNA gene amplification analysis performed for gathered samples after contamination. For Total microbial DNA isolation, MoBio PowerSoil DNA Isolation Kits (Mobio Laboratory) was used and was carried out as specified in the manufacturer's procedure (http://www.mobio.com-/files/protocol/12888.pdf). In preparation for DNA isolation, cotton samples taken from swab samples were vortexed for 30-40 seconds in 5mL MoBio lysis buffer to be removed from the strip with a sterile scalpel. After mixing and after centrifugation at 1500 rcf for 5 minutes, supernatant were removed and the cotton which had collapsed to the bottom was taken to MoBio Garnet Bead tubes containing 750ul MoBio buffer. After these tubes were treated at 65 °C and 10 °C for 10 minutes, the horizontal blended MoBio vortex adapter was treated with DNA extractions with the kit for 2 minutes at the highest speed [153]. If extraction was not carried out immediately, the mixture was stored at $-80\,^{\circ}\mathrm{C}$. Quantification of double stranded DNA (dsDNA) quantities of DNA samples between 1.8-2.0 purity isolated from the kit were performed using the Qubit dsDNA BR assay kit in the Qubit fluorimetric system (Qubit fluorimetric, Life Technology). Total bacterial DNA obtained after extraction was stored at $-20\,^{\circ}\mathrm{C}$ until use.

6.4.3 16S rRNA Polymerase Chain Reaction and Sequencing

16S Ribosomal RNA sequencing studies were performed on the Illumina MiSeq system and all pre-sequencing had been performed in accordance with the protocols of the device (16S Metagenomic Sequencing Library Preparation Preparing 16S Ribosomal RNA Gene Amplicons for the Illumina MiSeq System). In this study, 16S r RNA gene amplification were performed targeting V3-V4 regions. The Hotstar Master Mix (5Prime) was used for the polymerase chain reaction (PCR). Primers that were expected to form single amplicons of about 460 bp in the PCR and to which overhang adapters appropriate to the Illumina MiSeq system was used. Primers that were expected to form single amplicons of about 460 bp in the PCR and to which overhang adapters appropriate to the Illumina MiSeq system was added used. Forward 5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG3', Reverse 5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC3' primers were used.

The reaction was preincubated in a 96-well thermocycler, with final primer concentrations of 0.2 µ M, for 3 minutes at 95 °C in a total of 50 L reaction mixture; 25 cycles of denaturation at 95 °C for 30 seconds, 30 seconds at 55 °C and 30 seconds at 72 °C, and final extension for 5 minutes at 72 °C. The PCR products obtained after the amplification purified using AMPure XP beads (Beckman Coulter). The purified amplicons were used as templates in the Index PCR. The index PCR was be performed in accordance with the company protocol using the Nextera XT Index Kit. Index PCR products were cleaned using AMPure XP beads and the clean amplicons obtained were used to create libraries. The PhiX Control V3 Kit were used as a positive control when creating the Amplicon library.

Rarefaction is a technique to assess species richness from the results of sampling. A rarefaction curve which shows the change in the alpha diversity measure as the number of sample increases. If the curve

S1: phone before contagion
S2: phone after contagion
S3 : patient folder
S4: coughmeter
S5: mouse
S6: scissors
S7: blood sugar meter
S8: blue patient folder
S9: stethoscope
S10: common desk
S11: cabinet drawer
S12: trash can
S13: phone - again
S14: chair
S15: cabinet
S16: common desk - again
S17: coughmeter -again
S18: metal staple
S19: drug set bag surface
S20: alcohol sprey
S21: door handle
S22: glass door
S23: computer keyboard
S24: drug box
S25: patient desk
S26: patient box
S27: emergency car handles
S28: bed head
S29: subject hand

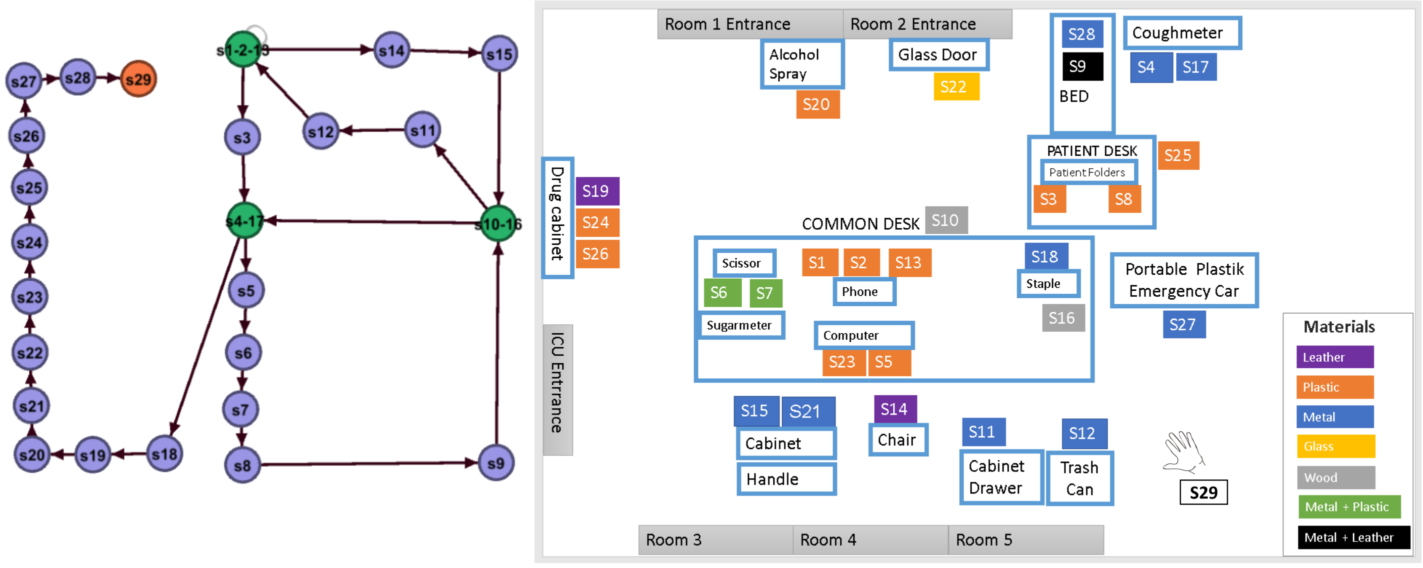

s1-2-13
s14
s15
s3
s12
s11
s4-17
s10-16
s5
s6
s7
s8
s9
s18
s19
s20
s21
s22
s23
s24
s25
s26
s27
s28
s29
Room 1 Entrance
Room 2 Entrance
Alcohol Spray
Glass Door
S20
S22
S28
S9
BED
Coughmeter
S4
S17
PATIENT DESK
Patient Folders
S3
S8
S25
Drug cabinet
S19
S24
S26
ICU Entrrance
COMMON DESK
S10
Scissor
S6
S7
Sugarmeter
S1
S2
S13
Phone
Computer
S23
S5
S18
Staple
S16
Portable Plastik Emergency Car
S27
S15
S21
Cabinet
Handle
S14
Chair
S11
Cabinet Drawer
S12
Trash Can
S29
Room 3
Room 4
Room 5
Materials
Leather
Plastic
Metal
Glass
Wood
Metal + Plastic
Metal + Leather

Figure 17: Experiment path, list of objects, and their respective locations in the ICU. Green circles in the experiment path represent the objects touched multiple times during experiment.

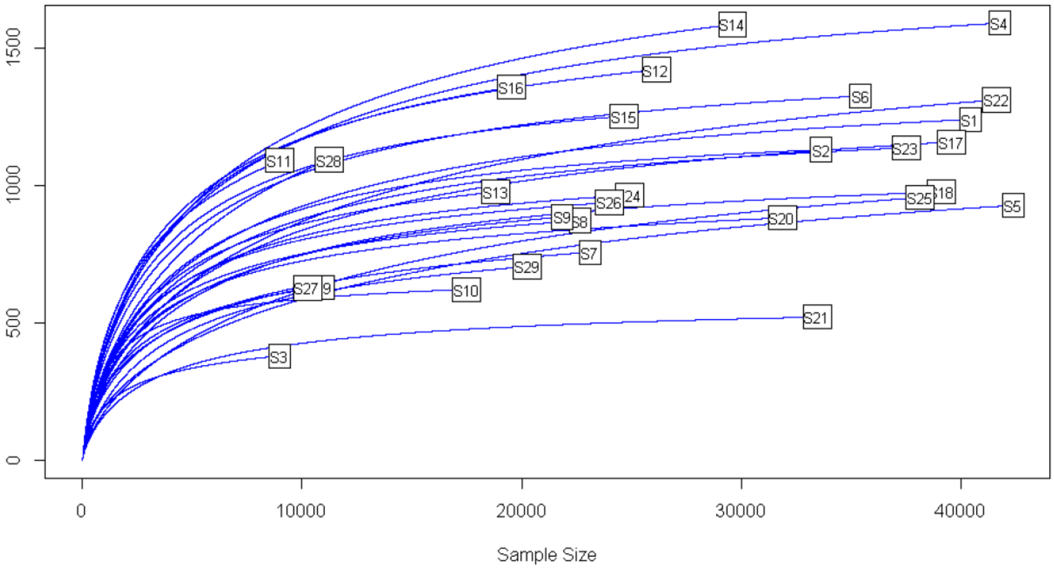


Figure 18: Rarefaction curve

converges to a horizontal asymptote, this indicates that further more reads will have little or no effect on the diversity [22]. The figure 18 shows the rarefaction curve for the 29 samples in the experiment.

6.4.4 Otu Picking

DADA2 (Divisive Amplicon Denoising Algorithm) was used for OTU table construction [18]. DADA2 enables a complete pipeline that produces merged, denoised, chimera-free sequence variants and OTU abundance matrix. A total of 2950 OTUs were constructed.

6.4.5 Preprocessing Data

It is assumed that if there is a microbial transmission (contagion) between two objects, they both should have shared OTUs. To model transmission in the experiment data, the OTUs that are only present in all samples were kept for further analysis. The original OTU table has 2950 OTUs; after preprocessing, an OTU table with 76 OTU were obtained. This filtering helped reduce dimensionality as well as solving sparsity problems.

6.5 RESULTS

6.5.1 Grouping OTUs as PMGs

PMGs is performed on OTU table obtained from the microbiome transmission experiment. After OTU table construction and data preprocessing steps 76 OTUs for 29 samples were grouped in 9 PMGs. The Figure 19 shows the 9 PMGs on CODA dendrogram. Each node corresponds to a principal balance. The first 8 most informative principle balances were chosen to construct 9 PMGs. OTUs were classified using Greengenes reference database (http://greengenes.lbl.gov) and the bar plot in the figure 20 shows the taxonomy information of OTU groups on genus level. As seen, OTUs in each group may have different taxonomies. Note that OTUs in the same group are approximately proportional. Figure 21 and Figure 22 shows a heatmap of PMG and OTU abundance tables. OTUs and PMGs which are abundant on each sample can be seen.

Biplots might help to interpret the microbiome abundance data by showing which OTUs dominate on which samples. Figure 23 shows the biplot on the OTUs (before grouping) and PMGs (after grouping). On the left biplot, the first 3 principal components explain the %58 of total variance. The plot is clutter with 76 OTUs and the explained total variance is not informative, so that it is not interpretable. On the right, the plot is clearer and the explained total variance for the first 3 principal components is %90 of the total variance retained by the 9 selected .

The graph in the figure 24 shows the 9 PMGs (G_i) as a hub node and and OTUs included in each PMG as connected nodes. Samples (S_i) are connected to OTUs which are the most abundant in that sample. The heatmap in the figure 21 also shows that G1 and G9 together are abundant on most of the samples. Those microbial features in the G1 and G9 might refer to the colonizing base community in the environment, in other words, the common characteristic of samples. On the other hand, it may refer to a transmission between the samples.

6.5.2 Hierarchical Clustering of Samples after PMG construction

Hierarchical clustering analysis of samples might be used for detection of densely interconnected objects. Branches of the hierarchical clustering dendrogram might correspond consecutively contaminated objects. The first step of hierarchical clustering is deciding the distance metric between object. It could be hypothetically assumed that if there is a transmission event between a pair of objects, then their OTU compositions should be similar, i.e the distance between two composition should be small. The variation matrix can be used as a distance matrix between objects. If the variation of two compositions is 0, then they are proportional.

One drawback of hierarchical clustering is the determination of a cutting point to detect clusters in the data set. Pvclust[154] is a R package for assessing the uncertainty in hierarchical clustering. Pvclust calculates p-values for hierarchical clustering via multiscale bootstrap resampling. Pvclust provides three p-values: SI (selective inference) p-value, AU (approximately unbiased) p-value and BP (bootstrap probability) value for each cluster in a dendrogram. Clusters with high AU or SI values are strongly supported by data.

Pvclust was applied on OTU table and PMGs table with 10000 bootstrap replications. The Ward method was used as the agglomerative method in hierarchical clustering and the square root of the variation matrix was used as a distance measure.

Before grouping data, pvclust could detect two clusters with %98 confidence (S4-S6-S16 and S11-S12). After grouping four clusters were detected with %98 confidence (S10-S11-S12, S19-S20-S29, S2-S13, S8-S16-S4-S6, S22-S5-S1-S14, S25-S26) as seen in the figure 25. In the table 7, detailed comments on why the samples go together in a cluster is explained. As a result, grouped data revealed more clusters in which consecutive samples in.

Table 7: The explanations of the pvclust clusters

Cluster	**Description**
S10-S11-S12	consecutively contaminated objects
S19-S20-S29	S19 and S20 is consecutively contaminated objects and S29 is hand. The reason S29 goes with this cluster might be that the material of S19 is leather.
S2-S13	the same object that have sampled two times
S22-S5-S1-S14	S1 and S14 could be considered consecutively contaminated objects, because S1 and S13 are the same object. S1, S5 and S14 are also located close to each other. They are objects on the common desk and their microbiome could have high variability. S5 and S23 are located closely and S22-S23 are consecutively contaminated objects. The material of S22 is glass so microbial features after a touch could grow easily on the material. That could be the reasons these samples go together in a cluster. On the other hand, those samples have high microbial materials after sequencing (See the rarefaction curve in the figure 18).
S25-S26	consecutively contaminated objects
S8-S16-S4-S6	S4-S5-S6-S7-S8 are consecutively contaminated objects, but S5 and S7 do not show up in this cluster. The reason could be the high variability of S5. S4 and S17 are the same object. This cluster might reflect the circle effect in the experiment path (See the figure 17. The square at the bottom reflecting the circle contagion order of the objects between S4 and S17.

6.6 Discussion and Conclusion

In this research, the aim was to group OTUs and to investigate any hint of microbial transmission. Grouping OTUs as PMGs improves the revelation of some consecutive samples in the experiment data. However, the microbial transmission experimental data is limited and drawing a conclusion is difficult with only one experiment. More structured experiments are needed for investigation of microbial transmission problem.

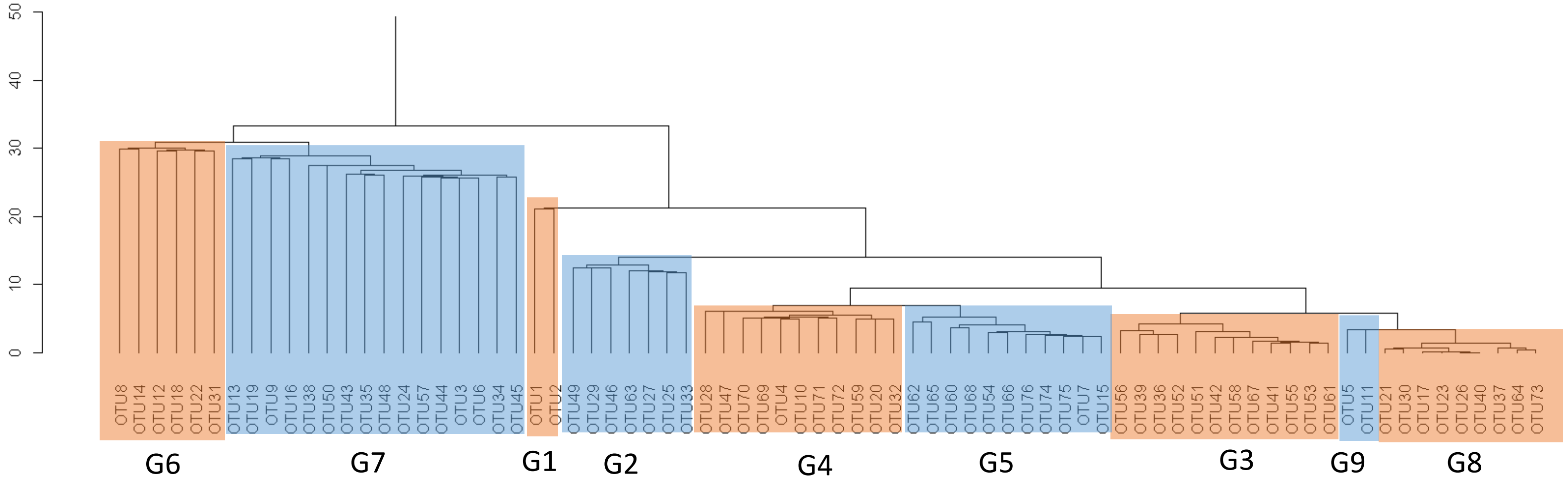


Figure 19: Coda Dendrogram of OTUs and constructed PMGs

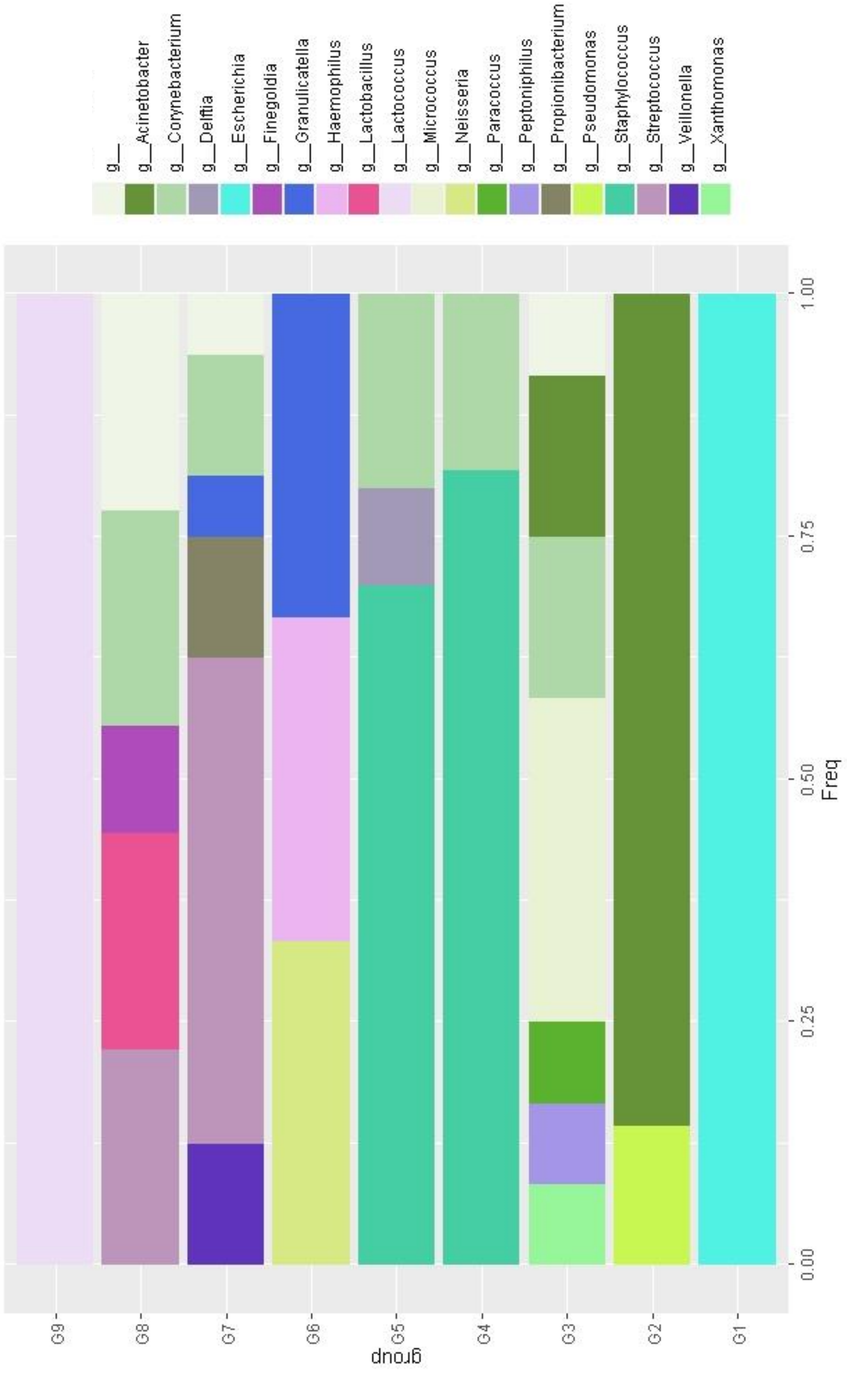


Figure 20: The taxonomic content of PMGs.

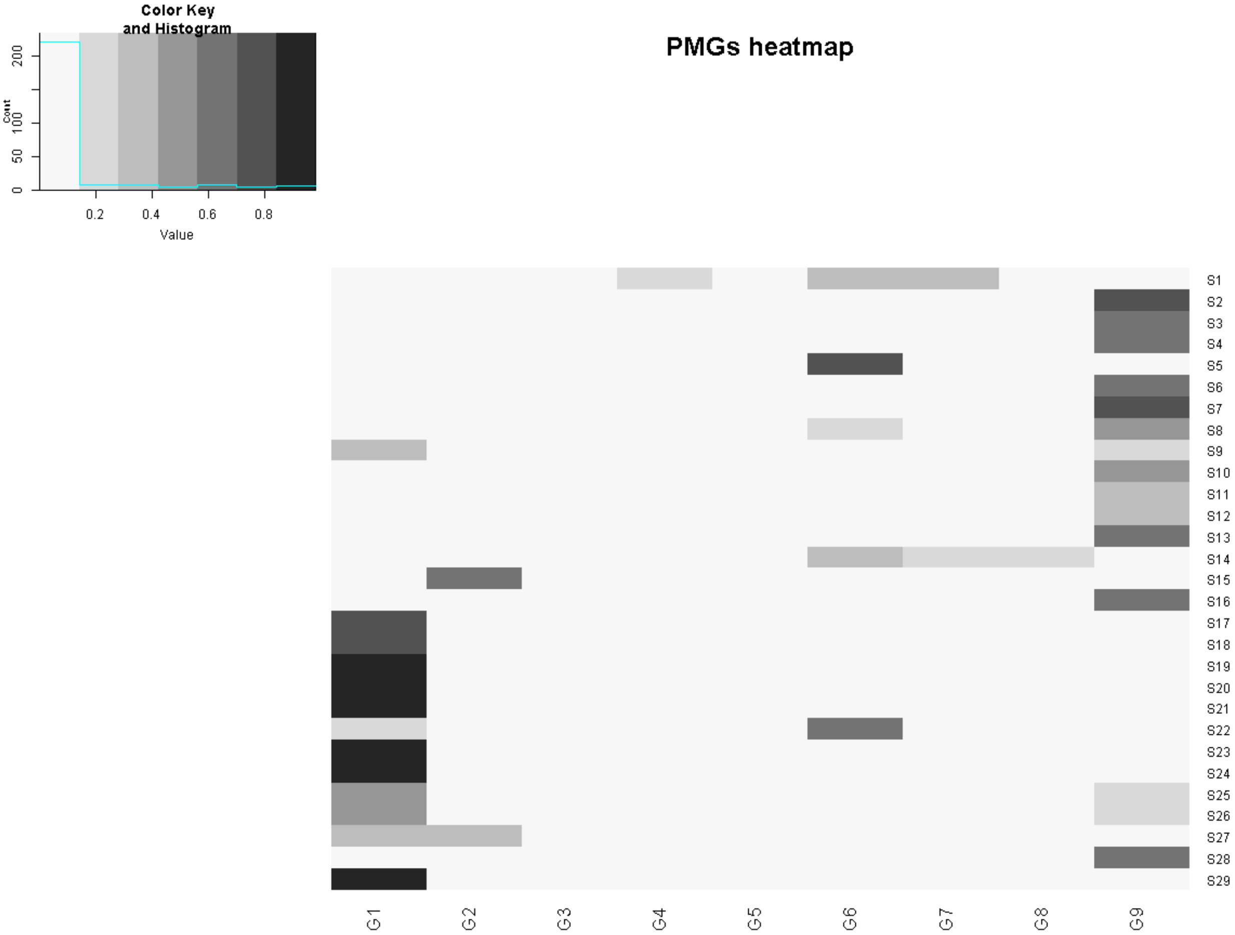


Figure 21: Heatmap of PMGs abundance table.

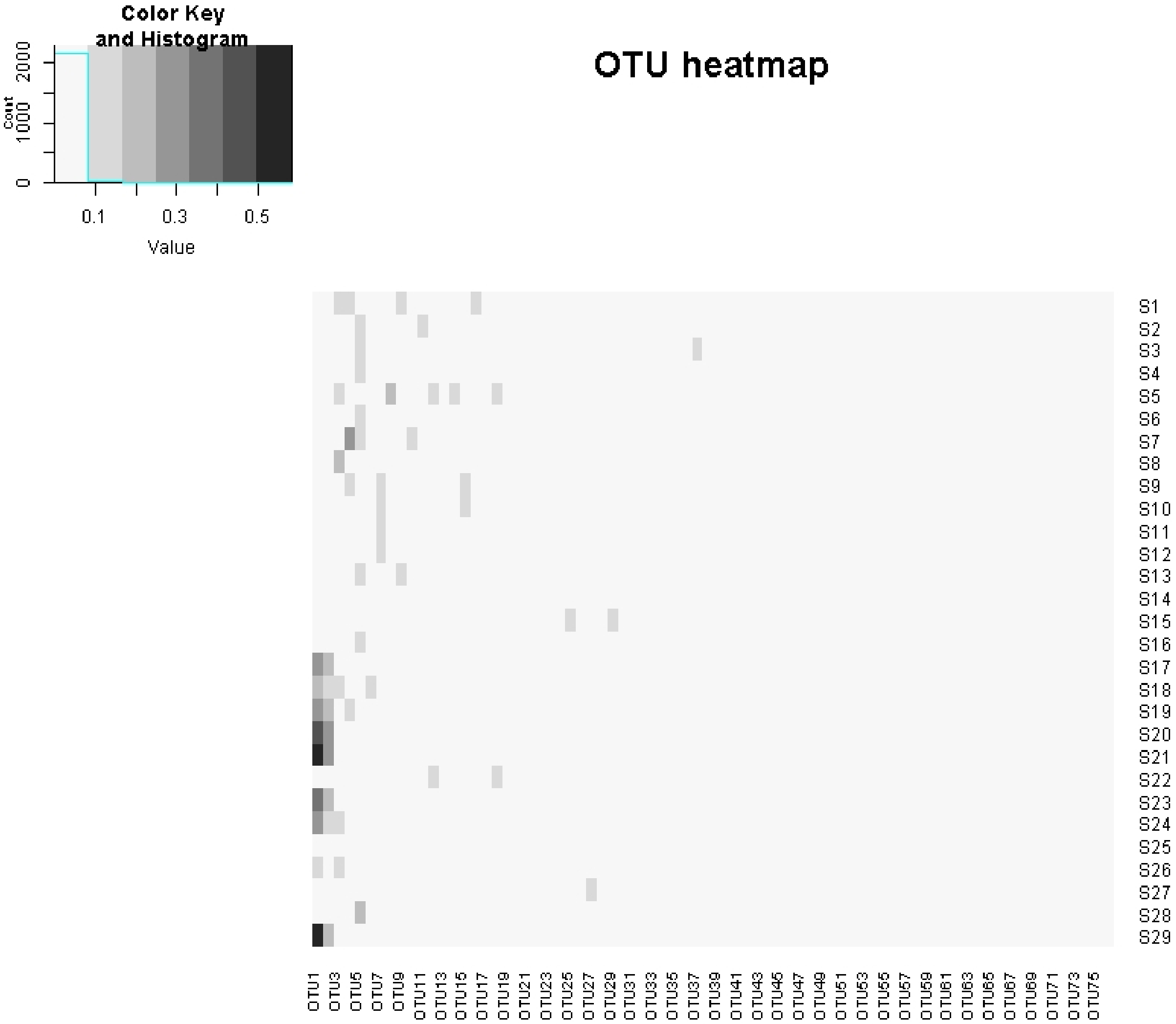
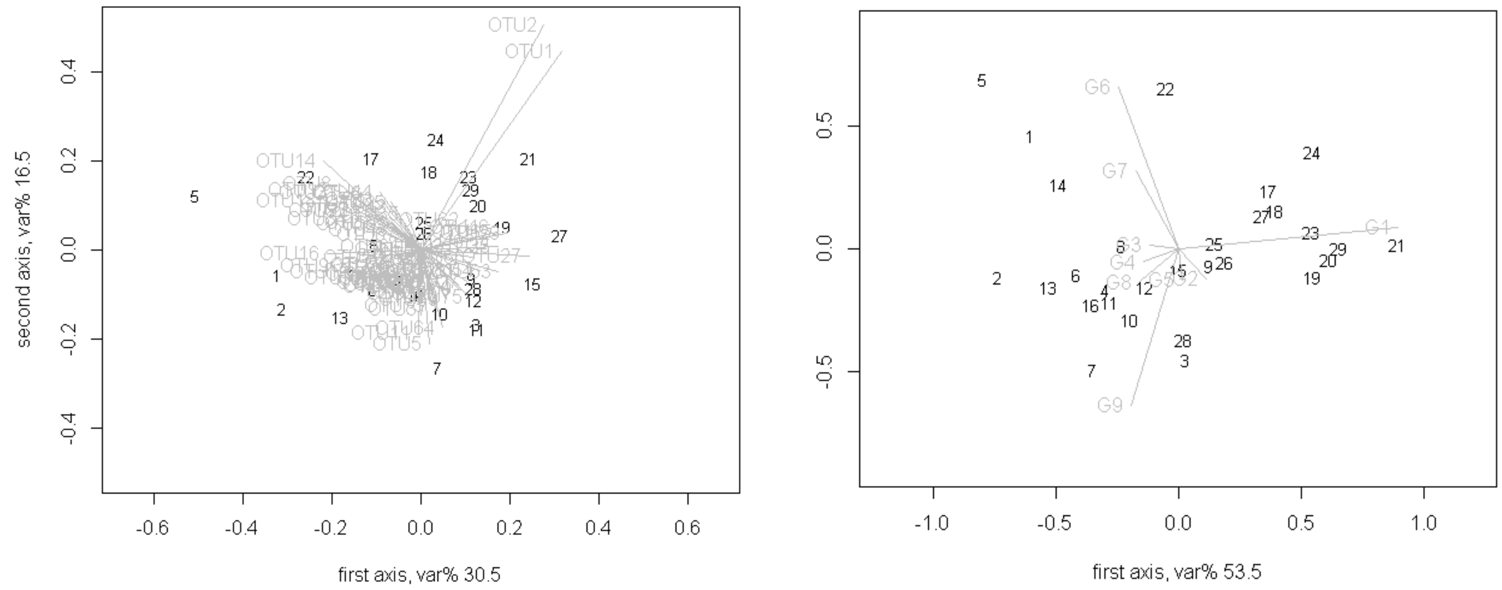


Figure 22: Heatmap of OTUs abundance table.



Form Biplot on OTU table

Form Biplot on PMG table

Figure 23: Form Biplots on OTU and PMG tables

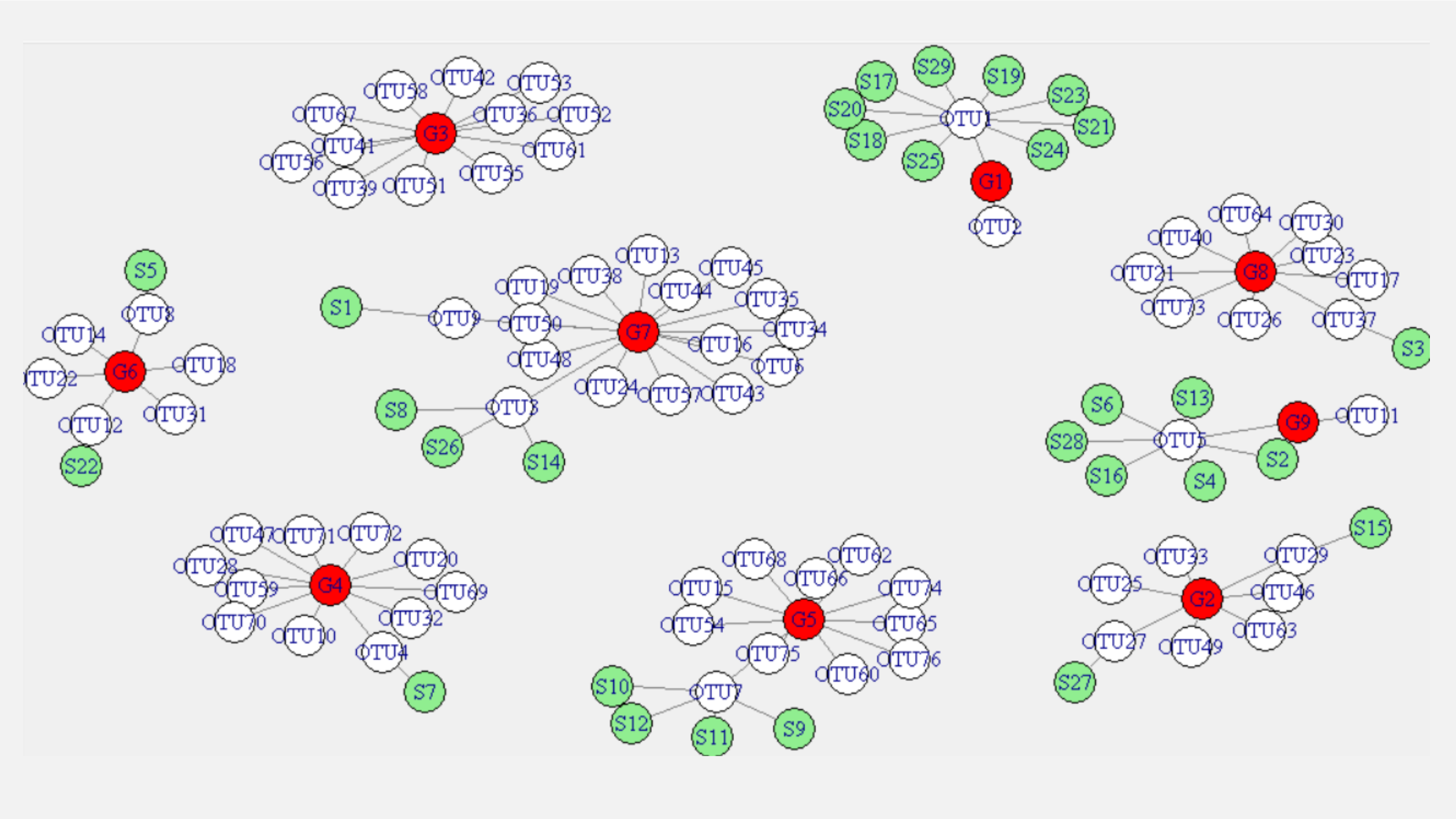
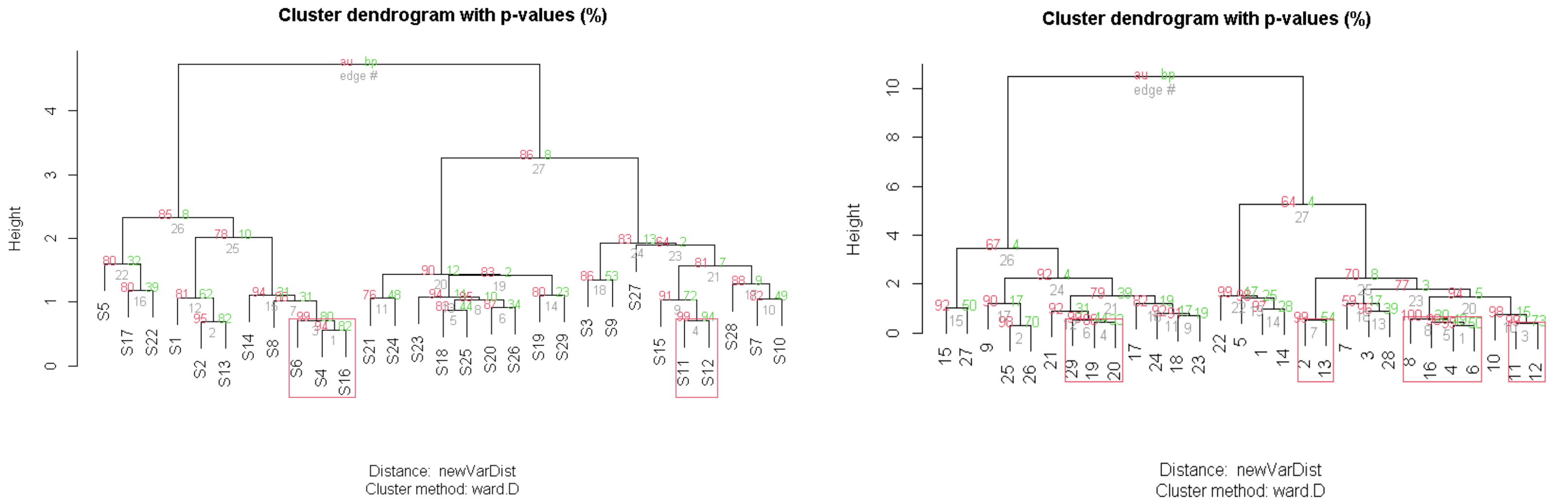

G1
G2
G3
G4
G5
G6
G7
G8
G9
OTU1
OTU2
OTU3
OTU4
OTU5
OTU6
OTU7
OTU8
OTU9
OTU10
OTU11
OTU12
OTU13
OTU14
OTU15
OTU16
OTU17
OTU18
OTU19
OTU20
OTU21
OTU22
OTU23
OTU24
OTU25
OTU26
OTU27
OTU28
OTU29
OTU30
OTU31
OTU32
OTU33
OTU34
OTU35
OTU36
OTU37
OTU38
OTU39
OTU40
OTU41
OTU42
OTU43
OTU44
OTU45
OTU46
OTU47
OTU48
OTU49
OTU50
OTU51
OTU52
OTU53
OTU54
OTU55
OTU56
OTU57
OTU58
OTU59
OTU60
OTU61
OTU62
OTU63
OTU64
OTU65
OTU66
OTU67
OTU68
OTU69
OTU70
OTU71
OTU72
OTU73
OTU74
OTU75
OTU76
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
S18
S19
S20
S21
S22
S23
S24
S25
S26
S27
S28
S29

Figure 24: OTUs (white nodes) are connected PMGs (red nodes) which they belong to and samples (Si) (green nodes) are connected to OTUs which are the most abundant in that sample.



A Hierarchical clustering of samples based on OTU table

B Hierarchical clustering of samples based on PMGs table

Figure 25: Hierarchical clustering via scale bootstrap resampling of samples. Samples in the red rectangle are the clusters that are strongly supported by the data with %98 confidence.

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this dissertation I have discussed several certain aspects of the statistical analysis of microbiome data and proposed a novel grouping procedure for microbiome data using a compositional data approach. The proposed procedure is used for two problems for demonstration: (1) in the search of biomarkers and (2) to track microbial transmission.

I started by giving a brief introduction how microbiome data is produced and the compositional nature of the microbiome data. Then, the principals and statistical methodologies of compositional data were reviewed. Next, a novel procedure called "Principal Microbial Groups (PMGs)" was proposed as an alternative to phylogenetic grouping of microbial features in Chapter 5. PMGs were used in the search of biomarkers, and had promising results for the Cirrhosis dataset. PMG construction enables working with coarse groups of OTUs in the dataset. PMGs also have some interpretation advantages in reducing dimensionality and provide balances of microbial groups that can be used for disease prediction. PMG table of the Cirrhosis dataset exhibited performance rivaling OTU and genus-level tables on balance selection methods selbal and codacore algorithms. The OTU content of PMGs are consistent with the literature findings in term of association with disease. As a future work, more datasets will be analyzed in order to understand the biological meaning of PMGs and their role in the search of biomarker.

We also used, PMGs are aimed to investigate to tracking microbial transmission and aimed to eventually to infer a microbial network. In Chapter 3, an overview of the microbial network inference methodologies and microbial association methods are summarized. Then, the microbial transmission experiment is explained and PMGs are used to investigate microbial transmission in Chapter 6. Grouping OTUs with the proposed methodology improves the revelation of some consecutive samples in the experiment data. However, the microbial transmission experimental data was limited and constructing a network was not possible with only this experiment. More structured experiments are needed in order to investigate microbial transmission procedure and to infer a microbial network for contagion.

While the proposed methodology provides a valid grouping of OTUs using the CoDa approach, determination of the number of groups and interpreting the biological meaning of those groups are open questions for future work.

REFERENCES

[1] J. A. Gilbert and B. Stephens, “Microbiology of the built environment,” *Nature Reviews Microbiology*, vol. 16, no. 11, pp. 661–670, 2018.

[2] B. Panthee, S. Gyawali, P. Panthee, and K. Techato, “Environmental and human microbiome for health,” *Life*, vol. 12, no. 3, p. 456, 2022.

[3] S. V. Lynch and O. Pedersen, “The human intestinal microbiome in health and disease,” *New England Journal of Medicine*, vol. 375, no. 24, pp. 2369–2379, 2016.

[4] N. K. Surana, D. L. Kasper, *et al.*, “Deciphering the tête-à-tête between the microbiota and the immune system,” *The Journal of clinical investigation*, vol. 124, no. 10, pp. 4197–4203, 2014.

[5] J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, and et al., “Balances: a new perspective for microbiome analysis,” *MSystems*, vol. 3, no. 4, pp. e00053–18, 2018.

[6] G. Wu, N. Zhao, C. Zhang, and et al., “Guild-based analysis for understanding gut microbiome in human health and diseases,” *Genome medicine*, vol. 13, no. 1, pp. 1–12, 2021.

[7] M. L. Calle, “Statistical analysis of metagenomics data,” *Genomics & informatics*, vol. 17, no. 1, 2019.

[8] K. R. Amato, “An introduction to microbiome analysis for human biology applications,” *American Journal of Human Biology*, vol. 29, no. 1, p. e22931, 2017.

[9] S. Kodikara, S. Ellul, and K.-A. Lê Cao, “Statistical challenges in longitudinal microbiome data analysis,” *Briefings in Bioinformatics*, vol. 23, no. 4, p. bbac273, 2022.

[10] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, “Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities,” *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[11] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, *et al.*, “Qiime allows analysis of high-throughput community sequencing data,” *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.

[12] M. Griffith, J. R. Walker, N. C. Spies, B. J. Ainscough, and O. L. Griffith, “Informatics for rna sequencing: a web resource for analysis on the cloud,” *PLoS computational biology*, vol. 11, no. 8, p. e1004393, 2015.

[13] W. Chen, C. K. Zhang, Y. Cheng, S. Zhang, and H. Zhao, “A comparison of methods for clustering 16s rrna sequences into otus,” *PloS one*, vol. 8, no. 8, p. e70837, 2013.

[14] W. SL and S. PD., “De novo clustering methods outperform reference-based methods for assigning 16s rrna gene sequences to operational taxonomic units,” *PeerJ*, vol. 3, 2015.

[15] S. W. Olesen, C. Duvallet, and E. J. Alm, “dbotu3: A new implementation of distribution-based otu calling,” *PLoS One*, vol. 12, no. 5, p. e0176335, 2017.

[16] T. C. Glenn, “Field guide to next-generation dna sequencers,” *Molecular ecology resources*, vol. 11, no. 5, pp. 759–769, 2011.

[17] A. Amir, D. McDonald, J. A. Navas-Molina, E. Kopylova, J. T. Morton, Z. Zech Xu, E. P. Kightley, L. R. Thompson, E. R. Hyde, A. Gonzalez, *et al.*, “Deblur rapidly resolves single-nucleotide community sequence patterns,” *MSystems*, vol. 2, no. 2, pp. e00191–16, 2017.

[18] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. Johnson, and S. P. Holmes, “Dada2: High resolution sample inference from amplicon data,” *BioRxiv*, p. 024034, 2015.

[19] R. C. Edgar, “Unoise2: improved error-correction for illumina 16s and its amplicon sequencing,” *BioRxiv*, p. 081257, 2016.

[20] A. M. Eren, L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison, and M. L. Sogin, “Oligotyping: differentiating between closely related microbial taxa using 16s rrna gene data,” *Methods in ecology and evolution*, vol. 4, no. 12, pp. 1111–1119, 2013.

[21] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, *et al.*, “Best practices for analysing microbiomes,” *Nature Reviews Microbiology*, vol. 16, no. 7, pp. 410–422, 2018.

[22] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight, “Uchime improves sensitivity and speed of chimera detection,” *Bioinformatics*, vol. 27, no. 16, pp. 2194–2200, 2011.

[23] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, “Vsearch: a versatile open source tool for metagenomics,” *PeerJ*, vol. 4, p. e2584, 2016.

[24] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, “Microbiome datasets are compositional: and this is not optional,” *Frontiers in microbiology*, vol. 8, p. 2224, 2017.

[25] G. B. Gloor and G. Reid, “Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data,” *Canadian journal of microbiology*, vol. 62, no. 8, pp. 692–703, 2016.

[26] D. Lovell, V. Pawlowsky-Glahn, J. J. Egozcue, and et al., “Proportionality: a valid alternative to correlation for relative data,” *PLoS computational biology*, vol. 11, no. 3, p. e1004075, 2015.

[27] V. Jonsson, T. Österlund, O. Nerman, and E. Kristiansson, “Variability in metagenomic count data and its influence on the identification of differentially abundant genes,” *Journal of Computational Biology*, vol. 24, no. 4, pp. 311–326, 2017.

[28] Y. Cao, A. Zhang, and H. Li, “Microbial composition estimation from sparse count data,” *Preprint. Available at*, 2017.

[29] H. Li, “Microbiome, metagenomics, and high-dimensional compositional data analysis,” *Annual Review of Statistics and Its Application*, vol. 2, pp. 73–94, 2015.

[30] T. P. Quinn, I. Erb, M. F. Richardson, and et al., “Understanding sequencing data as compositions: an outlook and review,” *Bioinformatics*, vol. 34, no. 16, pp. 2870–2878, 2018.

[31] I. Erb and C. Notredame, “How should we measure proportionality on relative gene expression data?,” *Theory in Biosciences*, vol. 135, no. 1, pp. 21–36, 2016.

[32] J. Aitchison, “The statistical analysis of compositional data,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[33] C. Soneson and M. Delorenzi, “A comparison of methods for differential expression analysis of rna-seq data,” *BMC bioinformatics*, vol. 14, no. 1, pp. 1–18, 2013.

[34] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edger: a bioconductor package for differential expression analysis of digital gene expression data,” *bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[35] S. Anders, “Analysing rna-seq data with the deseq package,” *Mol Biol*, vol. 43, no. 4, pp. 1–17, 2010.

[36] J. A. Martín-Fernández and S. Thió-Henestrosa, “Rounded zeros: some practical aspects for compositional data,” *Geological Society, London, Special Publications*, vol. 264, no. 1, pp. 191–201, 2006.

[37] A. D. Fernandes, J. M. Macklaim, T. G. Linn, G. Reid, and G. B. Gloor, “Anova-like differential expression (aldex) analysis for mixed population rna-seq,” *PloS one*, vol. 8, no. 7, p. e67019, 2013.

[38] J.-A. Martín-Fernández, K. Hron, M. Templ, and et al., “Bayesian-multiplicative treatment of count zeros in compositional data sets,” *Statistical Modelling*, vol. 15, no. 2, pp. 134–158, 2015.

[39] G. Gloor, “Aldex2: Anova-like differential expression tool for compositional data,” *ALDEX manual modular*, vol. 20, pp. 1–11, 2015.

[40] J. Palarea-Albaladejo and J. A. Martín-Fernández, “zcompositions—r package for multivariate imputation of left-censored data under a compositional approach,” *Chemometrics and Intelligent Laboratory Systems*, vol. 143, pp. 85–96, 2015.

[41] P. Filzmoser, K. Hron, C. Reimann, and R. Garrett, “Robust factor analysis for compositional data,” *Computers & Geosciences*, vol. 35, no. 9, pp. 1854–1861, 2009.

[42] J. de Sousa, K. Hron, K. Fačevicová, and P. Filzmoser, “Robust principal component analysis for compositional tables,” *Journal of Applied Statistics*, vol. 48, no. 2, pp. 214–233, 2021.

[43] J. J. Egozcue and V. Pawlowsky-Glahn, “Groups of parts and their balances in compositional data analysis,” *Mathematical Geology*, vol. 37, no. 7, pp. 795–828, 2005.

[44] M. Greenacre, “Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation,” *Applied Computing and Geosciences*, vol. 5, p. 100017, 2020.

[45] J. A. Martín-Fernández, V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosona-Delgado, “Advances in principal balances for compositional data,” *Mathematical Geosciences*, vol. 50, no. 3, pp. 273–298, 2018.

[46] C. Cardona, P. Weisenhorn, C. Henry, and J. A. Gilbert, “Network-based metabolic analysis and microbial community modeling,” *Current Opinion in Microbiology*, vol. 31, pp. 124–131, 2016.

[47] Y. EL-Manzalawy, "Proxi: a python package for proximity network inference from metagenomic data," *Systems biology*, 2018.

[48] V. K. Gupta, S. Paul, and C. Dutta, "Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity," *Frontiers in microbiology*, vol. 8, p. 1162, 2017.

[49] Y. Cao, *Statistical methods for high dimensional count and compositional data with applications to microbiome studies*. University of Pennsylvania, 2016.

[50] s. R. J. Faust K., "Microbial interactions: from networks to models," *Nature Reviews Microbiology*, vol. 10, p. 538–550, 2012.

[51] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, and et al., "Sparse and compositionally robust inference of microbial ecological networks," *PLoS computational biology*, vol. 11, no. 5, p. e1004226, 2015.

[52] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[53] H. Fang, C. Huang, H. Zhao, and M. Deng, "Cclasso: correlation inference for compositional data through lasso," *Bioinformatics*, vol. 31, no. 19, pp. 3172–3180, 2015.

[54] Y. Ban, L. An, and H. Jiang, "Investigating microbial co-occurrence patterns based on metagenomic compositional data," *Bioinformatics*, vol. 31, no. 20, pp. 3322–3329, 2015.

[55] E. Schwager, G. Weingart, C. Bielski, and C. Huttenhower, "Ccrepe: Compositionality corrected by permutation and renormalization," *R/Bioconductor https://doi. org/10.18129 B*, vol. 9, 2014.

[56] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," 2012.

[57] L. Albayrak, K. Khanipov, G. Golovko, and Y. Fofanov, "Detection of multi-dimensional co-exclusion patterns in microbial communities," *Bioinformatics*, vol. 34, no. 21, pp. 3695–3701, 2018.

[58] L. C. Xia, J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun, "Extended local similarity analysis (elsa) of microbial community and other time series data with replicates," in *BMC systems biology*, vol. 5, pp. 1–12, Springer, 2011.

[59] G. Golovko, K. Kamil, L. Albayrak, A. M. Nia, R. S. A. Duarte, S. Chumakov, and Y. Fofanov, "Identification of multidimensional boolean patterns in microbial communities," *Microbiome*, vol. 8, no. 1, pp. 1–10, 2020.

[60] D. Knights, J. Kuczynski, E. S. Charlson, J. Zaneveld, M. C. Mozer, R. G. Collman, F. D. Bushman, R. Knight, and S. T. Kelley, "Bayesian community-wide culture-independent microbial source tracking," *Nature methods*, vol. 8, no. 9, pp. 761–763, 2011.

[61] A. Chaney and D. Blei, "Visualizing topic models," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, pp. 419–422, 2012.

[62] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana Delgado, "Lecture notes on compositional data analysis," 2007.

[63] F. Chayes, “On correlation between variables of constant sum,” *Journal of Geophysical research*, vol. 65, no. 12, pp. 4185–4193, 1960.

[64] K. G. Van den Boogaart and R. Tolosana-Delgado, ““compositions”: a unified r package to analyze compositional data,” *Computers & Geosciences*, vol. 34, no. 4, pp. 320–338, 2008.

[65] J. Aitchison and J. J Egozcue, “Compositional data analysis: where are we and where should we be heading?,” *Mathematical Geology*, vol. 37, no. 7, pp. 829–850, 2005.

[66] J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn, “Logratio analysis and compositional distance,” *Mathematical Geology*, vol. 32, no. 3, pp. 271–275, 2000.

[67] J. J. Egozcue and V. Pawlowsky-Glahn, “Compositional data: the sample space and its structure,” *TEST*, vol. 28, no. 3, pp. 599–638, 2019.

[68] J. Aitchison, “Principal component analysis of compositional data,” *Biometrika*, vol. 70, no. 1, pp. 57–65, 1983.

[69] J. J. Egozcue and V. Pawlowsky-Glahn, “Isometric logratio transformations for compositional data analysis,” *Mathematical geology*, vol. 35, no. 3, pp. 279–300, 2003.

[70] J. Aitchison and M. Greenacre, “Biplots of compositional data,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 51, pp. 375–392, 2002.

[71] V. Pawlowsky-Glahn, J. J. Egozcue, R. Tolosana Delgado, and et al., “Principal balances,” in *Proceedings of the 4th International Workshop on CODA(2011)* (J. J. Egozcue, R. Tolosana-Delgado, and M. I. Ortego, eds.), CIMNE, Barcelona, Spain ISBN 978-84-87867-76-7, 2011.

[72] P. Babington, *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society of London, 2014.

[73] F. Murtagh and P. Legendre, “Ward’s hierarchical clustering method: clustering criterion and agglomerative algorithm,” *arXiv preprint arXiv:1111.6285*, 2011.

[74] J. J. Egozcue, V. Pawlowsky-Glahn, and G. B. Gloor, “Linear association in compositional data analysis,” *Austrian Journal of Statistics*, vol. 47, no. 1, pp. 3–31, 2018.

[75] K. G. Van den Boogaart and R. Tolosana-Delgado, *Analyzing compositional data with R*, vol. 122. Springer, 2013.

[76] A. D. Washburne, J. D. Silverman, J. W. Leff, and et al., “Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets,” *PeerJ*, vol. 5, p. e2969, 2017.

[77] D. McDonald, E. Hyde, J. W. Debelius, and et al., “American gut: an open platform for citizen science microbiome research,” *Msystems*, vol. 3, no. 3, pp. e00031–18, 2018.

[78] J. T. Morton, J. Sanders, R. A. Quinn, and et al., “Balance trees reveal microbial niche differentiation,” *MSystems*, vol. 2, no. 1, pp. e00162–16, 2017.

[79] T. P. Quinn, “Visualizing balances of compositional data: a new alternative to balance dendrograms,” *F1000Research*, vol. 7, 2018.

[80] M. Greenacre, "Contribution biplots," *Journal of Computational and Graphical Statistics*, vol. 22, no. 1, pp. 107–122, 2013.

[81] V. Pawlowsky-Glahn and J. J. Egozcue, "Exploring Compositional Data with the Coda-Dendrogram," *Austrian Journal of Statistics*, vol. 40, no. 1 & 2, pp. 103–113, 2011.

[82] J. C. Clemente, L. K. Ursell, L. W. Parfrey, and et al., "The impact of the gut microbiota on human health: an integrative view," *Cell*, vol. 148, no. 6, pp. 1258–1270, 2012.

[83] S. Kleine Bardenhorst, T. Berger, F. Klawonn, and et al., "Data analysis strategies for microbiome studies in human populations—a systematic review of current practice," *Msystems*, vol. 6, no. 1, pp. e01154–20, 2021.

[84] S. Roy and G. Trinchieri, "Microbiota: a key orchestrator of cancer therapy," *Nature Reviews Cancer*, vol. 17, no. 5, pp. 271–285, 2017.

[85] J. Wang and H. Jia, "Metagenome-wide association studies: fine-mining the microbiome," *Nature Reviews Microbiology*, vol. 14, no. 8, pp. 508–522, 2016.

[86] G. Yao, W. Zhang, M. Yang, and et al., "Microphenodb associates metagenomic data with pathogenic microbes, microbial core genes, and human disease phenotypes," *Genomics, proteomics & bioinformatics*, vol. 18, no. 6, pp. 760–772, 2020.

[87] Y. Chen, F. Yang, H. Lu, and et al., "Characterization of fecal microbial communities in patients with liver cirrhosis," *Hepatology*, vol. 54, no. 2, pp. 562–572, 2011.

[88] F. Yang and Q. Zou, "maml: an automated machine learning pipeline with a microbiome repository for human disease classification," *Database*, vol. 2020, 2020.

[89] F. Yang and Q. Zou, "mAML: an automated machine learning pipeline with a microbiome repository for human disease classification," *Database*, vol. 2020, 06 2020. baaa050.

[90] B. D. Topçuoğlu, N. A. Lesniak, M. T. Ruffin IV, and et al., "A framework for effective application of machine learning to microbiome-based classification problems," *MBio*, vol. 11, no. 3, pp. e00434–20, 2020.

[91] M. M. Finucane, T. J. Sharpton, T. J. Laurent, and K. S. Pollard, "A taxonomic signature of obesity in the microbiome? getting to the guts of the matter," *PloS one*, vol. 9, no. 1, p. e84689, 2014.

[92] M. Kriss, K. Z. Hazleton, N. M. Nusbacher, and et al., "Low diversity gut microbiota dysbiosis: drivers, functional implications and recovery," *Current opinion in microbiology*, vol. 44, pp. 34–40, 2018.

[93] M. Fassarella, E. E. Blaak, J. Penders, and et al., "Gut microbiome stability and resilience: elucidating the response to perturbations in order to modulate gut health," *Gut*, vol. 70, no. 3, pp. 595–605, 2021.

[94] R. Liu, C. Zhang, Y. Shi, and et al., "Dysbiosis of gut microbiota associated with clinical parameters in polycystic ovary syndrome," *Frontiers in microbiology*, vol. 8, p. 324, 2017.

[95] G. K. Gerber, "The dynamic microbiome," *FEBS letters*, vol. 588, no. 22, pp. 4131–4139, 2014.

[96] E. Rinninella, P. Raoul, M. Cintoni, and et al., “What is the healthy gut microbiota composition? a changing ecosystem across age, environment, diet, and diseases,” *Microorganisms*, vol. 7, no. 1, p. 14, 2019.

[97] G. B. Gloor, J. R. Wu, V. Pawlowsky-Glahn, and J. J. Egozcue, “It’s all relative: analyzing microbiome data as compositions,” *Annals of epidemiology*, vol. 26, no. 5, pp. 322–329, 2016.

[98] T. W. Randolph, S. Zhao, W. Copeland, and et al., “Kernel-penalized regression for analysis of microbiome data,” *The annals of applied statistics*, vol. 12, no. 1, p. 540, 2018.

[99] C. Martino, J. T. Morton, C. A. Marotz, and et al., “A novel sparse compositional technique reveals microbial perturbations,” *MSystems*, vol. 4, no. 1, pp. e00016–19, 2019.

[100] Y. Cao, A. Zhang, and H. Li, “Multisample estimation of bacterial composition matrices in metagenomics data,” *Biometrika*, vol. 107, no. 1, pp. 75–92, 2020.

[101] S. Mandal, W. Van Treuren, R. A. White, and et al., “Analysis of composition of microbiomes: a novel method for studying microbial composition,” *Microbial ecology in health and disease*, vol. 26, no. 1, p. 27663, 2015.

[102] I. Erb, G. B. Gloor, and T. P. Quinn, “Compositional data analysis and related methods applied to genomics—a first special issue from nar genomics and bioinformatics,” *NAR Genomics and Bioinformatics*, vol. 2, no. 4, p. lqaa103, 2020.

[103] M. C. Mert, P. Filzmoser, and K. Hron, “Sparse principal balances,” *Statistical Modelling*, vol. 15, no. 2, pp. 159–174, 2015.

[104] F. Yang and Q. Zou, “Disbalance: a platform to automatically build balance-based disease prediction models and discover microbial biomarkers from microbiome data,” *Briefings in Bioinformatics*, 2021.

[105] F. Yang, Q. Zou, and B. Gao, “Gutbalance: a server for the human gut microbiome-based disease prediction and biomarker discovery with compositionality addressed,” *Briefings in Bioinformatics*, 2021.

[106] J. D. Silverman, A. D. Washburne, S. Mukherjee, and et al., “A phylogenetic transform enhances analysis of compositional microbiota data,” *Elife*, vol. 6, p. e21887, 2017.

[107] T. P. Quinn and I. Erb, “Interpretable log contrasts for the classification of health biomarkers: a new approach to balance selection,” *Msystems*, vol. 5, no. 2, pp. e00230–19, 2020.

[108] T. P. Quinn, I. Erb, G. Gloor, and et al., “A field guide for the compositional analysis of any-omics data,” *GigaScience*, vol. 8, no. 9, p. giz107, 2019.

[109] L. J. Marcos-Zambrano, K. Karaduzovic-Hadziabdic, T. Loncar Turukalo, P. Przymus, V. Trajkovik, O. Aasmets, M. Berland, A. Gruca, J. Hasic, K. Hron, *et al.*, “Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment,” *Frontiers in microbiology*, vol. 12, p. 313, 2021.

[110] V. Pawlowsky-Glahn and J. J. Egozcue, “Geometric approach to statistical analysis on the simplex,” *Stochastic Environmental Research and Risk Assessment*, vol. 15, no. 5, pp. 384–398, 2001.

[111] A. Susin, Y. Wang, K.-A. Lê Cao, and et al., "Variable selection in microbiome compositional data analysis," *NAR genomics and bioinformatics*, vol. 2, no. 2, p. lqaa029, 2020.

[112] I. Moreno-Indias, L. Lahti, M. Nedyalkova, and et al., "Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions," *Frontiers in Microbiology*, vol. 12, p. 277, 2021.

[113] G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. J. Egozcue, "The principle of working on coordinates," in *Compositional Data Analysis: Theory and Applications* (V. Pawlowsky-Glahn and A. Buccianti, eds.), pp. 31–42, John Wiley & Sons, 2011. 378 p.

[114] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado, *Modeling and analysis of compositional data*. John Wiley & Sons, 2015.

[115] K.-A. Lê Cao, M.-E. Costello, V. A. Lakis, and et al., "Mixmc: a multivariate statistical framework to gain insight into microbial communities," *PloS one*, vol. 11, no. 8, p. e0160169, 2016.

[116] E. Gordon-Rodriguez, T. P. Quinn, and J. P. Cunningham, "Learning sparse log-ratios for high-throughput sequencing data," *bioRxiv*, 2021.

[117] T. P. Quinn and I. Erb, "Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data," *NAR genomics and bioinformatics*, vol. 2, no. 4, p. lqaa076, 2020.

[118] N. Qin, F. Yang, A. Li, and et al., "Alterations of the human gut microbiome in liver cirrhosis," *Nature*, vol. 513, no. 7516, pp. 59–64, 2014.

[119] E. Pasolli, D. T. Truong, F. Malik, and et al., "Machine learning meta-analysis of large metagenomic datasets: tools and biological insights," *PLoS computational biology*, vol. 12, no. 7, p. e1004977, 2016.

[120] P. Vangay, B. M. Hillmann, and D. Knights, "Microbiome learning repo (ml repo): A public repository of microbiome regression and classification tasks," *Gigascience*, vol. 8, no. 5, p. giz042, 2019.

[121] M. Kuhn, "The caret package," *Journal of Statistical Software*, vol. 28, no. 5, 2009.

[122] J. S. Bajaj, D. M. Heuman, P. B. Hylemon, and et al., "Altered profile of human gut microbiome is associated with cirrhosis and its complications," *Journal of hepatology*, vol. 60, no. 5, pp. 940–947, 2014.

[123] K. G. van den Boogaart and R. Tolosana-Delgado, *Analyzing Compositional Data with R*. Springer, Berlin, 2013. 258p.

[124] S. Thió-Henestrosa, J. J. Egozcue, V. Pawlowsky-Glahn, and et al., "Balance-dendrogram. a new routine of codapack," *Computers & geosciences*, vol. 34, no. 12, pp. 1682–1696, 2008.

[125] P. Wang, X. Tong, N. Zhang, T. Miao, J. P. Chan, H. Huang, P. K. Lee, and Y. Li, "Fomite transmission follows invasion ecology principles," *Msystems*, pp. e00211–22.

[126] P. Wang, N. Zhang, T. Miao, J. P. Chan, H. Huang, P. K. Lee, and Y. Li, "Surface touch network structure determines bacterial contamination spread on surfaces and occupant exposure," *Journal of Hazardous Materials*, vol. 416, p. 126137, 2021.

[127] T. Lam, D. Chew, H. Zhao, P. Zhu, L. Zhang, Y. Dai, J. Liu, and J. Xu, "Species-resolved metagenomics of kindergarten microbiomes reveal microbial admixture within sites and potential microbial hazards.," *Frontiers in microbiology*, vol. 13, pp. 871017–871017, 2022.

[128] P. Zhao, Q. Wang, P. Wang, S. Xiao, and Y. Li, "Influence of network structure on contaminant spreading efficiency," *Journal of Hazardous Materials*, vol. 424, p. 127511, 2022.

[129] R. I. Adams, S. Bhangar, K. C. Dannemiller, J. A. Eisen, N. Fierer, J. A. Gilbert, J. L. Green, L. C. Marr, S. L. Miller, J. A. Siegel, *et al.*, "Ten questions concerning the microbiomes of buildings," *Building and Environment*, vol. 109, pp. 224–234, 2016.

[130] S. Lax, C. R. Nagler, and J. A. Gilbert, "Our interface with the built environment: immunity and the indoor microbiota," *Trends in immunology*, vol. 36, no. 3, pp. 121–123, 2015.

[131] J. F. Meadow, A. E. Altrichter, S. W. Kembel, J. Kline, G. Mhuireach, M. Moriyama, D. Northcutt, T. K. O'Connor, A. M. Womack, G. Brown, *et al.*, "Indoor airborne bacterial communities are influenced by ventilation, occupancy, and outdoor air source," *Indoor air*, vol. 24, no. 1, pp. 41–48, 2014.

[132] A. Mahnert, C. Moissl-Eichinger, and G. Berg, "Microbiome interplay: plants alter microbial abundance and diversity within the built environment," *Frontiers in microbiology*, vol. 6, p. 887, 2015.

[133] M. K. Lee, M. U. Carnes, N. Butz, M. A. Azcarate-Peril, M. Richards, D. M. Umbach, P. S. Thorne, L. E. Beane Freeman, S. D. Peddada, and S. J. London, "Exposures related to house dust microbiota in a us farming population," *Environmental health perspectives*, vol. 126, no. 6, p. 067001, 2018.

[134] S. T. Kelley and J. A. Gilbert, "Studying the microbiology of the indoor environment," *Genome biology*, vol. 14, no. 2, pp. 1–9, 2013.

[135] K. P. Aßhauer, H. Klingenberg, T. Lingner, and P. Meinicke, "Exploring neighborhoods in the metagenome universe," *International journal of molecular sciences*, vol. 15, no. 7, pp. 12364–12378, 2014.

[136] E. Afshinnekoo, C. Meydan, S. Chowdhury, D. Jaroudi, C. Boyer, N. Bernstein, J. M. Maritz, D. Reeves, J. Gandara, S. Chhangawala, *et al.*, "Geospatial resolution of human and bacterial diversity with city-scale metagenomics," *Cell systems*, vol. 1, no. 1, pp. 72–87, 2015.

[137] S. D. Perkins, J. Mayfield, V. Fraser, and L. T. Angenent, "Potentially pathogenic bacteria in shower water and air of a stem cell transplant unit," *Applied and environmental microbiology*, vol. 75, no. 16, pp. 5363–5372, 2009.

[138] M. T. La Duc, A. Dekas, S. Osman, C. Moissl, D. Newcombe, and K. Venkateswaran, "Isolation and characterization of bacteria capable of tolerating the extreme conditions of clean room environments," *Applied and Environmental Microbiology*, vol. 73, no. 8, pp. 2600–2611, 2007.

[139] M. T. La Duc, P. Vaishampayan, H. R. Nilsson, T. Torok, and K. Venkateswaran, "Pyrosequencing-derived bacterial, archaeal, and fungal diversity of spacecraft hardware destined for mars," *Applied and environmental microbiology*, vol. 78, no. 16, pp. 5912–5922, 2012.

[140] M. K. Lee, M. U. Carnes, N. Butz, M. A. Azcarate-Peril, M. Richards, D. M. Umbach, P. S. Thorne, L. E. Beane Freeman, S. D. Peddada, and S. J. London, "Exposures related to house dust microbiota in a us farming population," *Environmental health perspectives*, vol. 126, no. 6, p. 067001, 2018.

[141] L. Oberauner, C. Zachow, S. Lackner, C. Högenauer, K.-H. Smolle, and G. Berg, "The ignored diversity: complex bacterial communities in intensive care units revealed by 16s pyrosequencing," *Scientific reports*, vol. 3, no. 1, pp. 1–12, 2013.

[142] M. Poza, C. Gayoso, M. J. Gomez, S. Rumbo-Feal, M. Tomas, J. Aranda, A. Fernandez, and G. Bou, "Exploring bacterial diversity in hospital environments by gs-flx titanium pyrosequencing," 2012.

[143] K. M. Hewitt, F. L. Mannino, A. Gonzalez, J. H. Chase, J. G. Caporaso, R. Knight, and S. T. Kelley, "Bacterial diversity in two neonatal intensive care units (nicus)," *PloS one*, vol. 8, no. 1, p. e54703, 2013.

[144] N. Fierer, C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight, "Forensic identification using skin bacterial communities," *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6477–6481, 2010.

[145] S. Lax, D. P. Smith, J. Hampton-Marcell, S. M. Owens, K. M. Handley, N. M. Scott, S. M. Gibbons, P. Larsen, B. D. Shogan, S. Weiss, *et al.*, "Longitudinal analysis of microbial interaction between humans and the indoor environment," *Science*, vol. 345, no. 6200, pp. 1048–1052, 2014.

[146] C. B. Hall, R. G. Douglas Jr, and J. M. Geiman, "Possible transmission by fomites of respiratory syncytial virus," *Journal of Infectious Diseases*, vol. 141, no. 1, pp. 98–102, 1980.

[147] D. Seong, M. Kingsak, Y. Lin, Q. Wang, and S. Hoque, "Fate and transport of enveloped viruses in indoor built spaces-through understanding vaccinia virus and surface interactions," *Biomaterials Translational*, vol. 2, no. 1, p. 50, 2021.

[148] K.-H. Chan, J. M. Peiris, S. Lam, L. Poon, K. Yuen, and W. H. Seto, "The effects of temperature and relative humidity on the viability of the sars coronavirus," *Advances in virology*, vol. 2011, 2011.

[149] K.-H. Chen, L.-R. Chen, and Y.-K. Wang, "Contamination of medical charts: an important source of potential infection in hospitals," *PLoS One*, vol. 9, no. 2, p. e78512, 2014.

[150] C. J. Uneke, A. Ogbonna, P. G. Oyibo, and C. M. Onu, "Bacterial contamination of stethoscopes used by health workers: public health implications," *The Journal of Infection in Developing Countries*, vol. 4, no. 07, pp. 436–441, 2010.

[151] W. Loh, V. Ng, and J. Holton, "Bacterial flora on the white coats of medical students," *Journal of Hospital Infection*, vol. 45, no. 1, pp. 65–68, 2000.

[152] A. Jeans, J. Moore, C. Nicol, C. Bates, and R. Read, "Wristwatch use and hospital-acquired infection," *Journal of Hospital Infection*, vol. 74, no. 1, pp. 16–21, 2010.

[153] K. Aagaard, J. Petrosino, W. Keitel, M. Watson, J. Katancik, N. Garcia, S. Patel, M. Cutting, T. Madden, H. Hamilton, *et al.*, “The human microbiome project strategy for comprehensive sampling of the human microbiome and why it matters,” *The FASEB Journal*, vol. 27, no. 3, pp. 1012–1022, 2013.

[154] R. Suzuki and H. Shimodaira, “Pvclust: an r package for assessing the uncertainty in hierarchical clustering,” *Bioinformatics*, vol. 22, no. 12, pp. 1540–1542, 2006.

[155] B. Efron, E. Halloran, and S. P. Holmes, “Bootstrap confidence levels for phylogenetic trees,” *Proc. Natl. Acad. Sci. USA*, vol. 93, pp. 7085–7090, 1996.

[156] L. J. Billera, S. P. Holmes, and K. Vogtmann, “Geometry of the space of phylogenetic trees,” *Advances in Applied Mathematics*, vol. 27, no. 733–767, 2001.

[157] M. Levandowsky and D. Winter, “Distance between sets,” *Nature*, vol. 234, pp. 34–35, 1971.

APPENDIX A

PRINCIPAL MICROBIAL GROUPS : EXTRA MATERIAL

A.1 Alternatives to PMG evaluation

Let $\mathbf{x} = (\mathrm{otu}_1, \mathrm{otu}_2, ...\mathrm{otu}_D)$ be a compositional observation with D microbial features. The procedure of PMG construction aggregates OTUs in groups that form the PMGs. The observation $\mathbf{x}$ is substituted by the z-part composition of PMGs, $\mathbf{y} = (\mathrm{g_m}(\mathrm{PMG}_1), \mathrm{g_m}(\mathrm{PMG}_2), \dots, \mathrm{g_m}(\mathrm{PMG}_z))$, where $z < D$. In the present approach, the value assigned to PMGs is the geometric mean of the relative frequencies of OTUs included. This option needs a justification and discussion of possible alternatives.

A selection of PMGs is equivalent to an orthogonal projection of the data set onto a subspace in which dimension is the number of selected PMGs minus one. The projection of $\mathbf{x} = (\mathrm{otu}_1, \mathrm{otu}_2, \mathrm{otu}_3, \dots, \mathrm{otu}_D)$ onto a subspace (Aitchison geometry) in which all otu_i included in a specific PMG are replaced by a single value, i.e. information within a PMG is removed. This constitutes an orthogonal projection [43] and the resulting projected D-part composition has the form

$$(\underbrace{\mathrm{g_m}(\mathrm{PMG}_1), \dots, \mathrm{g_m}(\mathrm{PMG}_1)}_{r_1 \text{equal components}}, \dots, \underbrace{\mathrm{g_m}(\mathrm{PMG}_z), \dots, \mathrm{g_m}(\mathrm{PMG}_z)}_{r_z \text{equal components}}) \ ,$$

$$\mathrm{g_m}(\mathrm{PMG}_i) = \prod_{\ell=1}^{r_j} (\mathrm{otu}_\ell)^{1/r_j} \ , \ \mathrm{otu}_\ell \text{ included in } \mathrm{PMG}_i \ ,$$

where $\mathrm{g_m}(\mathrm{PMG}_i)$ appears r_i times, as many as OTUs were grouped in PMG_i. In this projected compositions, any balance between PMGs always depends on the number r_is.

However, the number of OTUs included in each PMG does not seem relevant in this context, since these OTUs are not thought as the primary units for definition of biomarkers. Instead, the selected PMGs are viewed as the parts of a new composition from which the compositional biomarker is selected. It is thus natural to represent each PMG_i by $\mathrm{g_m}(\mathrm{PMG}_i)$, obtaining the $z \leq D$ part composition

$$\mathbf{y} = (\mathrm{g_m}(\mathrm{PMG}_1), \mathrm{g_m}(\mathrm{PMG}_2), \dots, \mathrm{g_m}(\mathrm{PMG}_z)).$$

This technique is seldom used in applications [123]. However, there are alternatives to this dimension reduction. For instance, the geometric mean can be replaced by the arithmetic mean or simply the sum (amalgam) of the included OTUs. Note that in the case of amalgams, the number of initial OTUs is again implicitly re-introduced.

A.2 Stability of Principal Microbial Groups

The selection of PMGs is based on an unsupervised clustering of OTUs, frequently used to approach principal balances [45]. Therefore, stability in the construction of PMGs is of a major concern. There are, at least, two ways for examining how PMGs can vary: (1) a global view of the clustering tree, and (2) focusing on the changes in a particular PMG. In both procedures the main difficulty is the identification of a PMG in the clustering results obtained using different samples when the PMGs are not exactly the same.

The distance between trees with equal leafs (OTUs) has been studied [155, 156]. These techniques can be used for constructing confidence intervals, or in general, studying variability of trees. Here we adopted a different metric between trees based on a compositional perspective.

A principal balance tree generates an ilr basis in the space of OTUs (compositions) and the corresponding decomposition of the total variance in the sample as shown in CoDa-dendrograms [81]. Although not rigorously proven, the composition made of the variances of the ilr-balance-coordinates appears to characterize the clustering tree up to a permutation of the leafs of the tree. Then, the Aitchison-norm of the composition of variances can be used to evaluate the variability of trees. Also, the distance between trees can be defined as the Aitchison distance between the respective Variance Decompositions (VD) when the clusters (PMGs) are well identified.

The following study is based on re-sampling (bootstrap) of the cirrhosis dataset. Individual samples are randomly chosen, with replacement, thus obtaining new data sets, with prescribed sizes.

In a first study, 50 re-samplings where used to generate PMGs. The cluster trees where cut at different levels (number of PMGs). Figure 26-A, shows quantiles (minimum –circles–, 0.25 –black line–, median –blue line–, 0.75 –black line–, maximum –circles–) of the VD norm for different levels of the tree characterized by the number of PMGs generated. The red line is the VD norm obtained for the original data. Figure 26-B shows the square-norms divided by the number of generated PMGs. This parameter was introduced as an inequality index in the VD [67] that can be interpreted as an information index.

The norm of VD is used in order to avoid detailed identification of each generated PMG. The smoothness of these curves shows the stability of groups when deciding an appropriate number of PMGs. The original VD norm (red line) is not in the lower tail of the VD norm distribution, thus suggesting that the original data is not an outlier with respect to the re-sampling distribution. The behaviour of the curves corresponds to what is expected. The VD norm for a large number of groups tends, monotonically increasing, to the square root of the total variance of the data set (Figure26-A). In the Figure26-B, the VD square-norm decreases after the first groupings. This means that the information in the VD associated with the PMGs slowly increases by increasing the number of PMGs after the first groupings. It also means that the variance within the PMGs decreases with the number of groups.

Aitchison distances between re-samplings of different sizes were also studied using 50 re-samplings of the original data set. Figure 27 shows that there is an imperceptible change of the VD norm in median of the re-sampling, while there is a slight decreasing of variability with the sample size. These features were expected since the cluster trees are based on the variation matrix of the data. The increase of the sample size reduces the variance of the variation matrix entries, while it approximately maintains the value of the variance estimator.

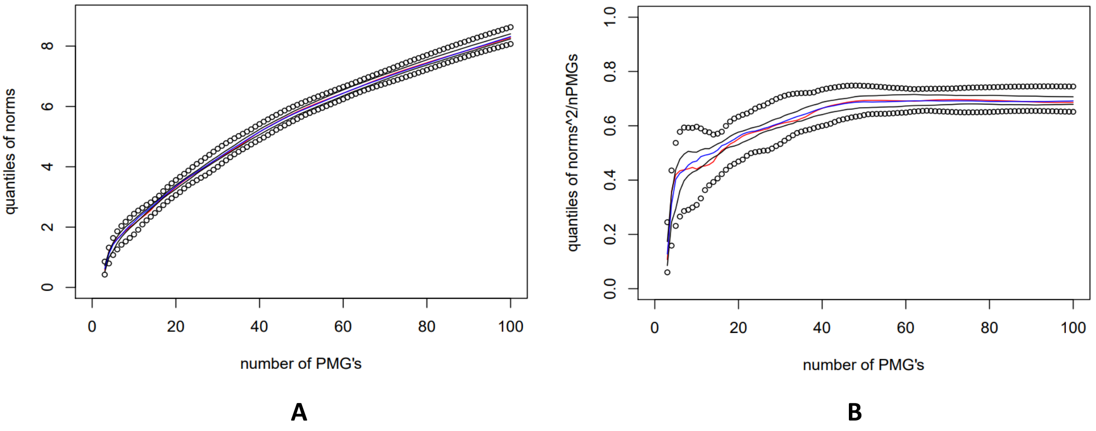


Figure 26: Quantiles of VD Aitchison norm (Panel A) and square-norm over the number of PMGs (Panel B) : minimum –circles–, 0.25 –black line–, median –blue line–, 0.75 –black line–, maximum –circles– for 50 re-samplings of different sample size. The red line indicates the VD norm of the original data set.

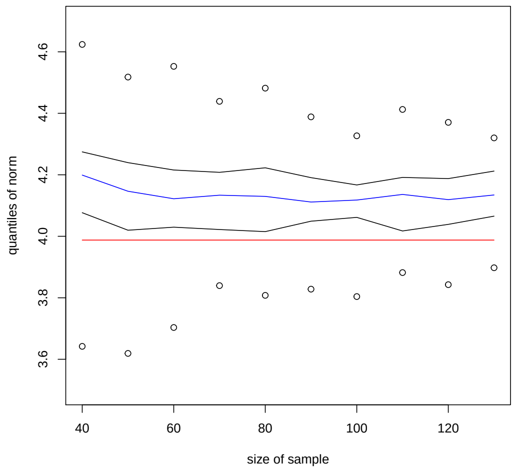


Figure 27: Quantiles of VD norm: minimum –circles–, 0.25 –black line–, median –blue line–, 0.75 –black line–, maximum –circles– for 50 re-samplings of different sample size. The red line indicates the VD norm of the original data set.

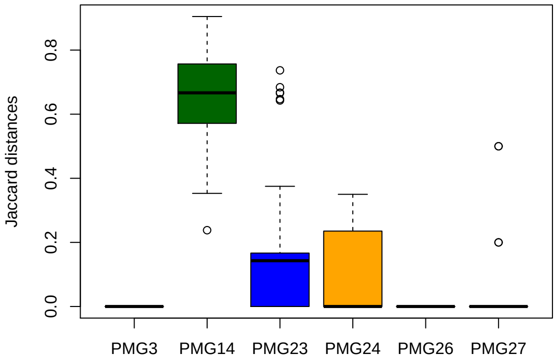


Figure 28: Jaccard distances from the original PMGs to those identified in the 50 re-samplings represented as boxplots. PMGs are indicated in the x-axis.

Individual PMGs can be examined along the 50 re-samplings. The following six (PMG3, PMG14, PMG23, PMG24, PMG26 and PMG27) have been chosen and their OTU composition, based on the original sample, has been considered as a reference. For each re-sampling, 27 groups are built using the clustering techniques. Each reference PMG is then compared with the groups in a re-sampling using the Jaccard distance [157]. The group with the smallest Jaccard distance to the reference is identified as the new group in the re-sampling.

The Jaccard distance [157] between two groups A and B is given by

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \, ,$$

where $| \cdot |$ denotes cardinal of the group and $\cap$ and $\cup$ are the ordinary intersection and union of sets, respectively. The Jaccard distance can be interpreted as the proportion of elements in $|A \cup B|$ which are not in the intersection of the two groups. Therefore, $d_J(A, B) = 0$ corresponds to total coincidence between A and B, whereas $d_J(A, B) = 1$ indicates that A and B are disjoint.

Figure 28 shows the Jaccard distance boxplots for the six mentioned PMGs along the 50 re-samplings. PMG3, PMG26, and PMG27 are almost always exactly identified, with 0 Jaccard distances to original PMG. PMG23 and PMG24 are also well identified, although not exactly. PMG14 is in general not well identified.

We can conclude that PMG construction using cluster of OTUs (Wards method using variation matrix) is reasonably stable when using the cirrhosis dataset.

The R-scripts to reproduce the above studies (up to simulations) are available at https://github.com/asliboyraz/PMGs.

A.3 Benchmarking Methods for Dimension Reduction

PMG balances are benchmarked against competing dimension reduction methods designed for compositional data (PCA, PBA, DBA-distal). The logistic regression classification performances of reduced datasets were compared. The following list summarizes each dimensionality reduction technique applied to the dataset for benchmarking and how to represent the dataset with each technique in R.

- PMG Balances : obtain ilr transformed PMGs that are $z-1$ coordinates and performs logistic regression.
- Principal Components: obtain the log-transformed OTU table and apply PCA. Retain the first $z-1$ principal components and perform logistic regression.
- Principal Balances: retain the first $z-1$ principal balances and perform logistic regression.
- Distal Balances: retain the first $z-1$ distal balances and perform logistic regression.

A.3.1 PCA representation

The preprocessed OTU table (filtered and zeros removed) (see Section Dataset and Preprocessing) is compositional and PCA can not be directly applied to a compositional dataset [68]. The clr-transformation is the commonly used method to apply PCA on compositional data [114]. Selecting the first z principal coordinates reduces the dimension to z.

```
// x must be non zero
pcaRep     function(x,numberOfDim)
  prc     prcomp(x=clr(X), retx=TRUE
                         , rank=numberOfDim
                         , center=TRUE)
  data.pca     as.data.frame(prc x)
  return(data.pca)
```

A.3.2 Principal Balance representation

Principal Balances (PBs) [71] are defined as a sequence of orthonormal balances which maximize successively the explained variance in a compositional dataset. A set of orthonormal balances is defined using a SBP and SBP can be approximated by the hierarchical clustering of parts using Ward's method. In R package *balance* [79], pba function performs a principal balance analysis using the hierarchical clustering of components. Selecting the first z principal balances reduces the dimension to z.

```
// x must be non zero
pbaRep     function(x,numberOfDim)
  library(balance)
  modelPba     pba(x)
  data.pb     as.data.frame(modelPba@pba)
  data.pb      data.pb[,1:(numberOfDim)]
  return(data.pb)
```

A.3.3 Distal Balance representation

DBA-distal method selects only predictive small balances (those involving 2 or 3 parts of the composition) [107]. The algorithm tries to generate a SBP that maximizes the discriminant potential of the distal branches. Unlike other methods, DBA-distal is supervised method and needs data labels (cirrhosis, non-cirrhosis in our case). Methods in the R package *balance* are used to obtain distal balances. Selecting the first z distal balance reduces the dimension to z.

```
// x must be non zero.
distalBalRep      function(x,labels,numberOfDim)
  library(balance)
  sbp      sbp.fromADBA(x, labels)
  sbp      sbp.subset(sbp)

  modelDistal      balance.fromSBP(x=x, y = sbp)
  data.distalBal      as.data.frame
              (modelDistal[,1:numberOfDim])

  return(data.distalBal)
```

A.4 Supplementary Tables

Table 8: Classification performances of reduced datasets processed with four dimensionality reduction procedures for disease prediction on the cirrhosis dataset.

Feature Reduction Methods[*]	AUC (on OTUs)	AUC (on Genus Level Table)
PMG balances	0.86	-
Principal Components	0.85	0.84
Principal Balances	0.84	0.84
Distal balances	0.85	0.85

Each method uses 26 dimension for fair comparison. PMG balances were not calculated on the genus-level table, since PMGs are designed for grouping OTUs as an alternative to taxon grouping.

Table 9: Selected balances by balance selection methods on different data types and AUC measures for the classification performance.

Method	Data types[1]	AUC[2]	Selected Balances[3]
Selbal	OTUs	0.88	(Veil.parvula, Mega.micro.) / Bact.uni.
	Genus Level Table	0.87	(Megasphaera/Unc.Erysip)
	PMGs	0.91	G26/G14
Codacore	OTUs	0.90	(Lac.salivarius, Megas.micro.)/(Adler.equ., Alis.indis.)
	Genus Level Table	0.92	(Lactobacillus, Megasphaera, Veillonella, Rodentibacter) /(Adlercreutzia, R
	PMGs	0.89	(G3,G26)/G23

The dimension of input PMG table is 27.

5-fold CV. n.iter=10.

Selected OTU balances were named with corresponding species. Selected PMG balances were named with group name. Only the global balance were reported for selbal.

Table 10: Classification performances (AUC[1]) of distal balances with different data types for disease prediction on the cirrhosis dataset.

Number of Balances[2]	distal-OTU balances	distal-genus balances	distal-PMG balances
1	0.930	0.910	0.924
2	0.947	0.927	0.925
3	0.940	0.935	0.932
4	0.936	0.933	0.928
5	0.940	0.933	0.933
6	0.939	0.929	0.928
7	0.939	0.929	0.929
8	0.939	0.925	0.929
9	0.933	0.920	0.922
10	0.932	0.915	0.913
11	0.928	0.912	0.918
12	0.925	0.908	0.918
13	0.927	0.908	0.912
14	0.921	0.914	0.906
15	0.922	0.911	0.901

5-fold CV. n.iter=10.

15 distal PMG balances were obtained on PMG table by DBA-distal. Thus, the first 15 distal balances were included in logistic regression for all data types.