THESIS FOR THE DEGREE OF DOCTOR

PHD IN EXPERIMENTAL SCIENCES AND TECHNOLOGY

Statistical methods for the analysis of microbiome compositional data in HIV studies

Javier Rivera Pinto

Thesis directors:

M. Luz Calle Rosingana (UVic - UCC)

Marc Noguera Julian (IrsiCaixa)

Vic, September 2018

IrsiCaixa
Institut de Recerca de la Sida

UVIC
UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL
DE CATALUNYA

IrsiCaixa AIDS research institute | UVic-UCC

This thesis has been partially supported by a grant for predoctoral studies from Noel Alimentaria to the University of Vic - Universitat Central de Catalunya (UVic - UCC).

A mi Familia, y a mi familia

After all there's no gene for fate

Gattaca (1997)

Abstract

The human microbiome is involved in many essential functions, such as food digestion and immune system maintenance. Alterations in its composition may have important effects on human health and they have been associated to high impact diseases such as obesity, asthma, cancer or cardiovascular disease among others.

This thesis is focused on the study of the link between the gut microbiome and HIV infection. The interest arises because of the important damages that the virus causes in the gut epithelium, which houses most of our immune system. Because of this damage, HIV patients present systemic and chronic inflammation responsible of an increase in their risk of having non-AIDS related diseases. Thus, understanding how gut microbiome alterations after HIV infection are related to immune dysregulation is of major importance.

The analysis of microbiome data is challenging. Since microbiome abundances are obtained from high-throughput DNA sequencing techniques, the

total number of reads per sample is constrained by the maximum number of sequence reads that the DNA sequencer can provide. This total count constraint induces strong dependencies among the abundances of the different taxa and confers the compositional nature of microbiome data. This means that the abundance values are not informative by themselves and that the relevant information is contained in the ratios of abundances between the different taxa. Ignoring the compositional nature of microbiome data may have important negative effects, such as spurious correlations, subcompositional incoherences, and the increase of type I error. In this context, we have proposed two novel statistical methods for microbiome analysis that preserve the principles of compositional data analysis: *MiRKAT-CoDA* (weighted and unweighted) and *selbal* algorithm.

MiRKAT-CoDA algorithm is a distance-based method for testing the overall association between microbial composition and a response variable of interest. It extends Kernel machine regression to compositional data analysis by considering a subcompositional dominant distance, such as Aitchison distance. The weighted version of *MiRKAT-CoDA* provides a measure of the contribution of each taxon to the global association with the response variable.

selbal algorithm is a new approach for the identification of microbial signatures associated to an outcome. The approach is innovative because, instead of defining the microbial signature as a linear combination of a set of taxa abundances, it is defined as a balance between two groups of taxa, a mathematical notion that preserves the principles of compositional data analysis.

In summary, the major contributions of this thesis are two new methodological strategies: *MiRKAT-CoDA* (weighted and unweighted) and *selbal* algorithm, for microbiome association testing and for the identification of microbiome signatures, respectively. Moreover, the results of this thesis have helped to advance the study of the role of the gut microbiome in HIV infection.

Resumen

El microbioma humano participa en muchas funciones esenciales como la digestión de alimentos y el mantenimiento del sistema inmunitario. Alteraciones en su composición pueden afectar a la salud del individuo, habiendo sido relacionados cambios en el microbioma con enfermedades tales como obesidad, asma, cáncer o enfermedades cardiovasculares entre otras.

Esta tesis está centrada en el estudio de la relación entre el microbioma intestinal y la infección por VIH. Este interés surge debido al importante daño que el VIH produce sobre el epitelio intestinal, el cuál contiene la mayor parte del sistema inmunitario. Debido a este daño, los pacientes infectados por VIH presentan una inflamación sistémica y crónica, responsable del incremento del riesgo de padecer enfermedades no relacionadas directamente con el SIDA. Así pues, resulta importante entender las alteraciones en el microbioma intestinal asociadas a la infección y patogénesis del VIH.

El análisis de los datos de microbioma resulta todo un desafío desde el punto de vista estadístico. Dado que los datos de abundancia del microbioma se obtienen por técnicas de secuenciación del ADN, el número total de *reads* por muestra viene limitado por el número máximo de secuencias que puede proporcionar el secuenciador. Esta limitación en el número de *reads* genera fuertes dependencias entre las abundancias de las diferentes taxas y define la naturaleza composicional de este tipo de datos. Este hecho supone que los valores de abundancia no son informativos en sí mismos, sino que la información la proporcionan realmente los ratios entre distintas componentes. De ignorar la composicionalidad de los datos de abundancia microbiana, los resultados obtenidos pueden ser confusos e incoherentes. Así, pueden aparecer correlaciones espurias, incoherencias subcomposicionales o incluso un incremento de los falsos positivos a la hora de definir las diferencias entre distintos grupos de individuos. En este contexto, presentamos dos nuevas propuestas para el estudio del microbioma que preservan los principios del análisis de datos composicionales: los algoritmos *MiRKAT-CoDA* (ponderada y sin ponderar) y *selbal*.

El algoritmo *MiRKAT-CoDA* es un método basado en distancias que permite evaluar si existe una asociación global entre la composición microbiana y una variable respuesta de interés. Este método es una extensión de la *Kernel machine regression* dentro del ámbito del análisis de datos composicionales, considerando una distancia subcomposicionalmente dominante como es la distancia de Atichison. La versión ponderada de *MiRKAT-CoDA* proporciona para cada variable un valor que mide la contribución de cada una de las taxas en la asociación global con la variable respuesta.

Por otra parte, el algoritmo *selbal* es una nueva propuesta focalizada en la identificación de firmas microbianas asociadas a una variable de interés. El método es novedoso debido a que en lugar de definir la firma microbiana como una combinación lineal de un conjunto de variables, se define como un balance entre dos grupos de taxas, una noción matemática que preserva los principios del análisis de datos composiconales.

En resumen, las mayores aportaciones de esta tesis son dos estrategias metodológicas diferentes: *MiRKAT-CoDA* (ponderada y sin ponderar) y *selbal*. Estas propuestas resultan útiles para evaluar la asociación entre microbioma y variable respuesta así como identificar firmas microbianas, respectivamente. Además, los resultados de esta tesis han contribuido al avance en el estudio del papel que desempeña el microbioma intestinal en la infección por VIH.

Resum

El microbioma humà participa en diverses activitats essencials per a l'hoste, com ara els processos de digestió i el manteniment del sistema immunitari. Alteracions en la seva composició poden tenir efectes importants en la salut humana i fins avui s'han associat a diverses malalties d'alt impacte com ara la obesitat, l'asma, el càncer i les malalties cardiovasculars, entre d'altres.

Aquesta tesis està centrada en l'estudi de la interacció entre el microbioma intestinal i la infecció per VIH. La infecció per VIH causa danys importants en l'epiteli i la mucosa intestinal, la qual acull gran part del nostre sistema immunitari. En relació a la pèrdua de l'homeòstasi intestinal, els individus amb infecció per VIH-1 solen presentar una inflamació sistèmica i crònica, malgrat l'administració de tractament antiretroviral, que sol associar-se a major risc de partir diverses malalties, associades o no, a la SIDA. Per això, és important entendre la relació entre aquestes alteracions del microbioma intestinal associades a la infecció i la patogènesis del VIH.

Malgrat els avenços dels últims anys, l'anàlisi del microbioma segueix essent tot un repte. Donat que les mesures d'abundància del microbioma deriven de dades de seqüenciació massiva, el número total de seqüències per mostra està cenyit al màxim número de seqüències que el seqüenciador pot generar. Així doncs, aquest limitació numèrica condiciona el càlcul de les abundàncies relatives dels diferents taxons detectats i confereix una natura composicional a les dades generades. Això significa que els valors d'abundància no són informatius per ells mateixos, i que la informació important està lligada al rati entre les abundàncies de diferents taxons. Ignorar la natura composicional de les dades de microbioma pot tenir significants efectes negatius importants com ara correlacions espúries, incoherències subcomposicionals i major presència de falsos positius. En aquest context, hem proposat 2 nous mètodes estadístics per a l'anàlisi de dades de microbioma que mantenen els principis de les dades composicionals: *MiRKAT-CoDA* (ponderada i no ponderada) i l'algoritme *selbal*.

L'algoritme *MiRKAT-CoDA* és un mètode basat en distàncies dissenyat per testar l'associació entre la composició del microbioma i alguna variable resposta d'interès. Així doncs, el *MiRKAT-CoDA* aplica la *Kernel machine regression* a l'anàlisi de dades composicionals considerant una distància dominant subcomposicional, com ara és la distància de Aitchison. La versió ponderada de *MiRKAT-CoDA* proporciona, a més a més, una mesura de la contribució de cada taxó a la associació global amb la variable resposta.

Per altra banda, l'algoritme *selbal* és una nova aproximació per a la identificació de signatures microbianes associades a una determinada observació (i.e. variable resposta). Aquesta aproximació és innovadora perquè, enlloc de definir una signatura microbiana com una combinació lineal d'una sèrie de taxons, es defineix com un balanç entre dos grups de taxons, una noció matemàtica que preserva els principis de les dades d'origen composicional.

En resum, les contribucions més significatives d'aquesta tesis són dues noves estratègies metodològiques, el *MiRKAT-CoDA* (ponderada i no ponderada) i *selbal*, útils per a testar associacions i identificar signatures en dades de microbioma, respectivament. A més a més, els resultats d'aquesta tesis han ajudat a avançar en l'estudi del paper del microbioma intestinal en la infecció per VIH.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The study of the microbes that live in our body and their involvement in human health is nowadays one of the most active research topics in biomedicine. The terms human microbiota and human microbiome are commonly used indistinctly to name the collection of all the microorganisms living in association with the human body. This includes eukaryotic microorganisms, archaea, bacteria and viruses. Although a recent study suggests that the number of microorganisms in the human body is approximately equal to the number of human cells (Sender et al., 2016), a ratio 10:1 between microorganisms and host cells is more commonly accepted. Anyhow, the human microbiome accounts for about 1 to 3 percent of total body mass (MacDougall, 2012), a significant proportion of a human.

The human microbiome is involved in a large number of essential functions, like food digestion and immune system maintenance. Alterations in microbiota may have important effects on human health. They have been associated to high impact diseases such as obesity, type 2 diabetes, asthma, irritable bowel syndrome, inflammatory bowel disease, cardiovascular disease, cancer and other disorders.

The human microbiome displays considerable diversity in different body locations and several microbial projects have been launched to characterize the microbiome in different body sites. However, the gut microbiome is by far the most studied one. The great interest generated by the gut microbiota stems from its important role in the host's nutrient metabolism, xenobiotic and drug metabolism, maintenance of structural integrity of the gut mucosal barrier, immunomodulation and protection against pathogens (Thursby and Juge, 2017; Jandhyala et al., 2015). The gut microbiome has been studied in projects focusing on a wide range of illnesses, such as inflammatory bowel disease , obesity, cardiovascular disease or even anxiety (Hold et al., 2014; Davis, 2016; Wilson Tang and Hazen, 2017; Lach et al., 2018).

Although the human microbiome has long been known to influence human health and disease, until recently, its composition and properties were largely unknown, since studies were limited to in-vitro cultivation of some specific microorganisms. Currently, high-throughput DNA sequencing technologies have revolutionized this field, allowing us to study the genomes of all microorganisms of a given environment. Metagenomics is the massive study of genomes of microorganisms and represents a breakthrough in the study of the relationship between the human microbiome and our health.

Microbiome studies are based on microbial DNA sequencing through two main approaches: amplicon sequencing (16S rRNA gene for both bacteria and archaea) and whole metagenomics DNA shotgun sequencing. Both, amplicon and shotgun sequencing allows the study of the microbial diversity and composition but, because of its higher resolution and sensitivity, shotgun sequencing provides more reliable abundance estimations and, in addition, it allows for the functional profiling of the microbial community.

Amplicon sequencing relies on sequencing a phylogenetic marker gene after Polymerase chain reaction (PCR) amplification. For bacteria and archaea, the marker gene is the 16S ribosomal RNA gene that encodes the RNA component of the small ribosomal subunit. The 16S rRNA gene contains both highly conserved areas and hypervariable sites denoted as V1-V9 (Figure 1.1). 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. The V1-V3 and V4 regions are most commonly targeted. PCR amplification creates thousands to millions of copies (amplicons) of the DNA target region. PCR amplicons are then sequenced using high-throughput sequencing platforms and multiple nucleotide sequences, also known as reads, are obtained. After preprocessing and quality control filtering, the sequences are grouped by their similarity into clusters that define Operational Taxonomic Units (OTUs); that is, groups of sequences, usually with at least a 97% of similarity between them. Each OTU is represented by a consensus sequence which is compared against a reference database, such as, *GreenGenes*

(DeSantis et al., 2006), *myRDP* (Cole et al., 2007) or *SILVA* (Pruesse et al., 2007), for taxonomic assignment to different taxonomic ranks. Figure 1.2 shows, as an example, the hierarchy defined for *Escherichia coli.* As we descend in rank, we gain in resolution with a more specific classification. 16S rRNA gene sequencing generally allows you to determine up to the genus rank. As a result of all this process, an OTU abundance table is obtained containing the number of sequences for each sample associated to each OTU. Since OTU tables are usually extremely sparse, they are often merged into abundance tables at higher taxonomic groups or taxa.

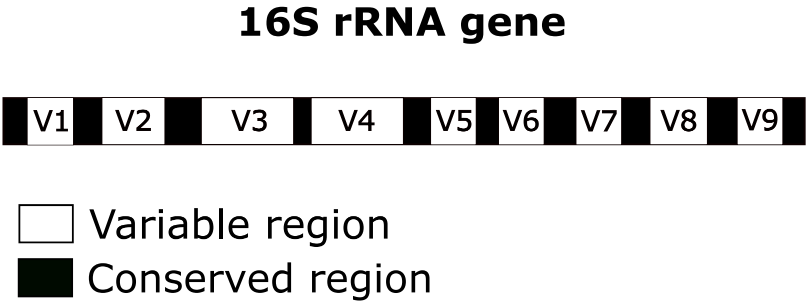


Figure 1.1: *Conserved and variable regions of 16S rRNA gen.*

The second alternative is shotgun metagenomics sequencing which involves sequencing the total microbial DNA of a sample, instead of just a particular marker gene. With this technique, we can infer the relative abundance of every microbial gene and quantify specific metabolic pathways to predict the potential functionality of the entire community. This is achieved by mapping the obtained sequences against a database such as *KEGG* (Kane-

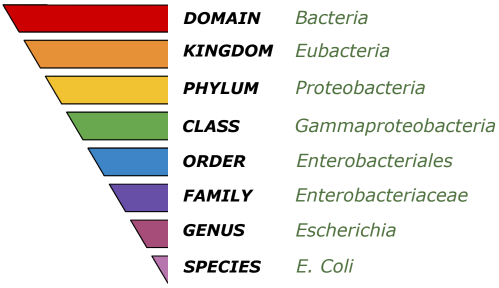


Figure 1.2: *Taxonomic ranks defined for Escherichia coli bacteria.*

hisa, 2002). An abundance matrix resulting from this type of functional study provides the number of sequences associated to a particular function for each sample. Figure 1.3 summarizes the procedure followed by both 16S rRNA sequencing and shotgun metagenome sequencing approaches.

From a statistical perspective, the output of both microbiome approaches, amplicon and shotgun sequencing, is similar: an abundance table of counts representing the number of sequences (reads) per sample for a specific taxon or a particular gene function. Throughout this dissertation, we illustrate the methodologies with data from 16S rRNA amplicon sequencing but most approaches also apply for microbiome shotgun metagenomics.

There are many reasons why the analysis of microbiome data is so challenging. On one hand, we face the usual challenges of count data analysis, i.e., skewed distribution, zero inflation and overdispersion. Because of the experimental process and quality control filtering, the total number of counts per sample is highly variable, which requires some normalization prior to the analysis so that the microbiome abundances among the different samples are comparable. Moreover, the total number of counts per sample is constrained by the maximum number of sequence reads that the sequencer can provide. This total count constraint induces strong dependencies among the abundances of the different taxa characterizing microbiome data as compositional data. How to deal properly with microbiome compositional data is the main focus of this work.

This thesis is the result of my participation in a partnership between the Microbial Genomics group (IrsiCaixa AIDS research institute) and the Research Group in Bioinformatics and Medical Statistics (UVic-UCC, University of Vic, Central University of Catalonia). One of the main areas of research of the Microbial Genomics group is to understand the role of the gut microbiome in human health and disease and, in particular, its involvement in HIV-1 infection. On the other hand, the Research Group in Bioinformatics and Medical Statistics is focused on the development of new statistical methods for the analysis of omics data, in particular, for microbiome data.

Supervised by Dr. Marc Noguera at IrsiCaixa, I have taken part in several microbiome projects in which the Microbial Genomics group is involved. This has given me the opportunity to learn the most relevant questions in HIV-microbiome studies, to understand the difficulties of this research area, and to contribute to the improvement of the field. In turn, Prof. M. Luz

Calle at UVic-UCC has guided me in the mathematical and methodological part of the thesis. This includes the identification of the limitations of the standard methods for microbiome analysis, my immersion in a largely unknown mathematical domain of compositional data analysis, and the development of new approaches for microbiome analysis in this area.

Thus, this thesis is motivated by the need for a rigorous mathematically approach to answer different questions arisen in microbiome studies. More specifically, the main goals of this thesis are, on one hand, to contribute to the understanding of the role of the gut microbiome in HIV-1 infection, and on the other hand, to improve microbiome analysis by proposing novel statistical methods that preserve the principles of compositional data analysis.

The different chapters of this dissertation mainly describe the methodological part of my work, that is, the proposal of two new statistical approaches for microbiome analysis in the context of compositional data analysis, one of which has already been published and is available in the Appendix (Rivera-Pinto et al., 2018). The more applied part of my thesis, as a result of my participation in the HIV-microbiome studies conducted at IrsiCaixa, is not written as specific chapters of this manuscript but it is reflected on three published papers added to the Appendix (Rivera-Pinto et al., 2017; Vesterbacka et al., 2017; Rivera-Pinto et al., 2018).

This thesis is organized as follows. At the end of this first introductory chapter we add two sections, one to explain the importance and interest of the study of the gut microbiome in HIV-AIDS research, and the other section to describe the most common procedures for microbiome analysis and discuss their limitations.

The second chapter explores the mathematical characteristics of microbiome data emphasizing its compositional nature, and introduces the main concepts of Compositional Data Analysis (CoDA). Some recent publications warn of the possible issues that may arise if standard tools are used for compositional datasets (Gloor et al., 2016; Weiss et al., 2017). We illustrate how incoherent results or an increase of type I error can result from the use of statistical tools that are not designed for constrained datasets. We introduce different transformations based on log-ratio analysis, such as *alr*, *clr* and *ilr* transformations, that offer the possibility of using different standard statistical approaches without having to worry about the aforementioned issues. We also discuss different ways of dealing with zeros and evaluate the impact of zero abundance on the most common distances used in microbiology.

In chapter 3 we propose the Kernel machine regression for microbiome compositional data, a distance-based regression method for testing the overall association between microbial composition and a response variable of interest. By applying the subcompositional dominant Aitchison distance we adapt Kernel machine regression to compositional data analysis. Additionally, we develop a weighted version using the recently defined weighted Aitchison distance (Egozcue and Pawlowsky-Glahn, 2016), which provides a measure of the contribution of each taxon to the global association with the response variable. The algorithms for implementing the methodology are publicly available as an R package named `MiRKAT-CoDA` at https://github.com/UVic-omics/MiRKAT-CoDA.

Chapter 4 presents `selbal`, a new algorithm for the identification of microbial signatures associated to an outcome, that is, groups of microbial taxa that are predictive of a phenotype of interest. These microbial signatures can be used for diagnosis, prognosis or prediction of therapeutic response based on an individual's specific microbiota (Knight et al., 2018). This approach is particularly innovative since, instead of defining a microbial signature as a linear combination of a set of taxa abundances, we define a microbial signature as a *balance* between two groups of taxa. The mathematical notion of *balance*, a special kind of log-contrast function, preserves the principles of compositional data analysis. `selbal` performs forward variable selection and identifies two groups of taxa whose balance or relative abundance is most associated with the response variable of interest. The new method is published in a peer-reviewed scientific paper (Rivera-Pinto et al., 2018) and the algorithm is publicly available as an R package named `selbal` at https://github.com/UVic-omics/selbal.

In summary, the major contributions of this thesis are two new methodological strategies: `MiRKAT-CoDA` (weighted and unweighted) and `selbal` algorithm, for microbiome association testing and for the identification of microbiome signatures, respectively. Moreover, the results of this thesis have helped to advance the study of the role of the gut microbiome in HIV infection.

1.1 Microbiome and HIV infection

The acquired immunodeficiency syndrome (AIDS) is the set of symptoms caused by the infection of the human immunodeficiency virus (HIV). HIV infection affects over 36 million people worldwide, mainly in Africa, which accounts for about seventy percent of the global HIV-positive individuals. Since the pandemic began, HIV has been responsible of more than 35 million deaths (UNAIDS, 2017). Figure 1.4 shows the HIV/AIDS epidemic statistics for the last three decades. Concurrently with the increase in social awareness and the intensive research efforts that have given rise to the development and roll-out of new antiretroviral therapies (ART), the number of new HIV infections has steadily decreased since 1990s. These improvements have had a direct impact on the number of HIV-related deaths that has experienced an important decline in the last decade. In 2014, the number of deaths was almost half of those occurred ten years before. However, an estimated 1.8 million individuals worldwide became newly infected with HIV in 2017, about 5000 new infections per day (UNAIDS, 2017). Thus, still additional research is needed to prevent more infections and to help the millions of people living with HIV.

The most important HIV/AIDS research areas are focused on the immune response to the disease, on how to prevent and eradicate the disease and, on the development of new therapies for providing better life conditions. This thesis is focused on one of the main interests of the Microbial Genomics group, to understand the role of the gut microbiome and HIV infection (http://www.irsicaixa.es/en/microbial-genomics). Specifically, the current microbiome-HIV projects at IrsiCaixa are mainly devoted to:

1. Analyse how the gut microbiome influences the ability of HIV-1 infected individuals to achieve adequate immune reconstitution, control HIV-1 replication and limit chronic inflammation.
2. Characterize the co-evolution of the gut microbiome and inflammatory response after acute HIV-1 infection.
3. Understand how the human microbiome can influence the AIDS vaccine response and vice versa, that is, how vaccines and other strategies for eliminating HIV-1 affect the human microbiome.

The interest on studying the link between the gut microbiome and HIV infection arises because of the important changes and damages in the gut epithelium caused since the beginning of the infection by the virus. The gut houses most of our immune cells responsible of the defense against pathogens and blocking the pass of harmful bacteria into the blood. On the other hand, the gut microbiota plays an important role in regulating immune homoeostasis, that is, the balance between immune response to pathogens and self-tolerance to avoid autoimmunity. Since the beginning of the infection, HIV damages the gut associated lymphoid tissue (GALT) causing a depletion of immune cells and greatly reducing the intestinal barrier function. This results in bacterial translocation which triggers inflammation processes. Inflammation persists even under HIV treatment and, hence, HIV patients present systemic and chronic inflammation responsible of an increase in their risk of having non-AIDS related diseases, such as, cardiovascular disease, cancer, kidney disease, liver disease, neurologic disease, and bone diseases (Phillips et al., 2008). Thus, understanding how gut microbiome alterations after HIV infection are related to immune dysregulation is of major importance. In this framework, IrsiCaixa is investigating how we can act on the microbiome to help people living with HIV recover immunity, and to strengthen the immune response of a therapeutic or preventive vaccine (http://www.irsicaixa.es/en/microbial-genomics). We highlight two different projects led by Dr. Roger Paredes in the Microbial Genomics group.

The first project is a transversal study in collaboration with the Karolinska Institutet (Sweden) that includes two independent cohorts of HIV-1-infected subjects and HIV-1-negative controls in Barcelona ($n = 156$) and Stockholm ($n = 84$). The objective of this project is to study the association between microbiome and HIV infection and to elucidate some contradictory associations reported in different studies. In recent years shifts from *Bacteroides* to *Prevotella* predominance have been described in HIV-1 infection (Lozupone et al., 2013; Vázquez-Castellanos et al., 2014), however, these shifts have not been found in previous works including studies with animal models. These results may be easily confounded by other factors such as environmental factors, long term dietary patterns or exercise (Modi et al., 2014; Sommer and Bäckhed, 2013). This project takes into account these important variables in order to evaluate their impact on the microbial composition.

Here we briefly describe the Barcelona cohort since we will use it to illustrate the methods proposed in this thesis. We will refer to it as the IrsiCaixa HIV study. The Barcelona cohort is composed of 156 individuals, most of them males (79.5%), enrolled in Barcelona, Catalonia, Spain, between January and December 2014. The cohort consists of 127 HIV-1 positive individuals (HIV+) and 29 HIV-negative controls (HIV-). Three HIV-1 risk groups were defined: heterosexual (*hts*), men who have sex with men (*msm*), and people who inject drugs (*pwid*). More details can be found in (Noguera-Julian et al., 2016).

The second project is a longitudinal study focused on the changes in gut microbiome composition in the very early stages of HIV-1 infection. For this propose, a group of 49 Mozambican subjects diagnosed with recent HIV-1 infection and 54 HIV-1-negative controls were followed for 9-18 months. Comparing them with 98 chronically HIV-1-infected subjects, the study tries to elucidate which are the microbial changes due to the infection.

Several papers have been published from these two projects, and some others are still ongoing or under revision. I add in the appendix the three papers where my participation was more relevant (Rivera-Pinto et al., 2017;

Vesterbacka et al., 2017; Rivera-Pinto et al., 2018).

1.2 Standard microbiome statistical analysis

From a statistical and mathematical point of view the main element of a microbiome study is the microbiome abundance table that is usually expressed as a matrix of counts, denoted by $\mathbf{X}$, with k columns (taxa) and n rows (samples). Each entry x_{ij} of $\mathbf{X}$ is the number of sequences (reads) corresponding to taxon j in sample i.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

$$x_{ij} \in \mathbb{N} \cup \{0\}, i \in \{1, \dots, n\},\ j \in \{1, \dots, k\}$$

The sum of the counts in each row of $\mathbf{X}$ provides the total number of counts per sample, a concept that in the context of DNA sequencing technologies is known as the *sampling depth*, the *library size* or *sequencing depth.*

Though there is no standardized protocol for microbiome analysis, there are some procedures that are commonly performed and that can be divided into four different steps: normalization, diversity analysis, ordination and differential abundance testing. Without trying to be comprehensive, we briefly describe the most important procedures in the following subsections.

1.2.1 Normalization

The large variability of the total counts per sample, prevents meaningful comparisons of raw abundances between individuals with very different total counts. This is usually addressed through normalization of raw counts before the analysis. The literature offers a wide range of normalization methods, among which the most frequently used are the transformation to relative abundances, dividing each cell by the total number of counts per sample; rarefaction, which consists on subsampling the same number of reads for all the samples from their initial compositions, and some additional percentile transformations or more complex techniques implemented in R packages developed for RNA-seq analysis, such as, {`DESeq`} (Anders and Huber, 2010) or {`edgeR`} (Robinson et al., 2009). See Weiss et al. (2017) and McMurdie et al. (2014) for a comparison and discussion on the performance of different normalization methods for microbiome analysis (McMurdie and Holmes, 2014; Weiss et al., 2017).

1.2.2 Diversity analysis

As in any ecological system, the diversity of the microbiome community is an important indicator of the good or bad conditions of the environment, with larger microbiome diversity being usually associated to better health status. Microbiome diversity can be assessed through multiple ecological indices that can be separated into two kind of measures, alpha and beta diversity. Alpha diversity measures the variability of species within a sample while beta diversity accounts for the differences in composition between samples.

Alpha diversity. Richness and evenness

The most important measure of alpha diversity is *richness*, the number of different species present in an environment. Richness is estimated by the observed richness, R_{obs}, that is the number of different species observed in the sample. Because of the sequencing is limited to the sample being analyzed and to the sequencing depth, the observed richness tends to underestimate the real richness in the environment, where the less frequent species are likely to be undetected. There are different indices that adjust for this and try to estimate the hidden part that has not been detected. One of the most extended is *Chao1* index defined as

$$R_{Chao1} = R_{obs} + \frac{f_1(f_1 - 1)}{2(f_2 + 1)}$$

where f_1 is the number of species observed only once, and f_2 is the number of species observed only twice.

Another important indicator of alpha diversity is *evenness*, which measures the homogeneity in abundance of the different species in a sample. In a composition of k different taxa, the maximum evenness is achieved when the abundances are uniformly distributed, that is, each of taxa has a relative abundance of $\frac{1}{k}$. On the other hand, the evenness is very low when only few taxa accounts for most of the relative abundance in the sample. A commonly used measure of evennes is the *Shannon* index defined as

$$R_{Shannon} = -\sum_{i=1}^{k} p_i \log(p_i)$$

where p_i represents the relative abundances of the i-th taxon.

Beta diversity. Distance

Beta diversity measures the differences in microbiome composition between samples. The usual Euclidean distance is not a good measure of beta diversity since it does not behave as a good ecological distance is expected, i.e., for samples that share most of their species, the distance should be small and when samples have few species in common, the distance should be large. There is a wide range of ecological distances or dissimilarities for measuring how close are two microbial compositions. While some of them only consider the presence or absence of each species in the sample, some others take the relative abundance into account. There is also a family of measures including in their definition the phylogenetic relationship between taxa. Here, we define the main measures that can be found in the literature for microbiome analysis, namely Bray-Curtis, UniFrac and weighted UniFrac distances.

Let $\mathbf{p_1} = (p_{11}, \ldots, p_{1k})$ and $\mathbf{p_2} = (p_{21}, \ldots, p_{2k})$ denote the microbiome relative abundance of two different samples.

Bray-Curtis distance

Bray-Curtis is probably the most extended measure in microbiome studies and is defined as follows:

$$d_{BC}(\mathbf{p_1}, \mathbf{p_2}) = \frac{\sum\limits_{i=1}^{k} \mid p_{1i} - p_{2i} \mid}{\sum\limits_{i=1}^{k} (p_{1i} + p_{2i})} \tag{1.1}$$

UniFrac family of distances

UniFrac family of distances, unlike Bray-Curtis distance, consider the phylogenetic tree that represents the evolutionary relationships among the different taxa. For a tree with r branches, let $\mathbf{b} = (b_1, \ldots, b_r)$ represent the length of the different branches in the phylogenetic tree, and $\mathbf{q}_1 = (q_{11}, \ldots, q_{1r})$, and $\mathbf{q}_2 = (q_{21}, \ldots, q_{2r})$ the relative abundances associated to each branch for the first and the second sample, respectively. According to these values, we can compute the *weighted*, *unweighted* and *generalized* versions of the UniFrac distance.

The unweighted UniFrac distance (Lozupone and Knight, 2005) defined in Equation 1.2 with $I(\cdot)$ denoting the indicator function, measures the relative length of those branches that lead exclusively to species present in only one of the two samples with respect to the total length of all branches in the tree.

$$d_U(\mathbf{b}, \mathbf{q}_1, \mathbf{q}_2) = \frac{\sum\limits_{i=1}^{r} b_i \mid I(q_{1i} > 0) - I(q_{2i} > 0) \mid}{\sum\limits_{i=1}^{r} b_i \left(I(q_{1i} + q_{2i} > 0\right)} \tag{1.2}$$

The unweighted UniFrac distance only takes into account the presence or absence of the taxa but Lozupone et al. also introduced the *weighted UniFrac distance* that includes information on the relative abundance of each taxa and is defined as follows

$$d_W(\mathbf{b}, \mathbf{q}_1, \mathbf{q}_2) = \frac{\sum\limits_{i=1}^{r} b_i \mid q_{1i} - q_{2i} \mid}{\sum\limits_{i=1}^{r} b_i \, (q_{1i} + q_{2i})} \tag{1.3}$$

1.2.3 Ordination

After computing pairwise distances, it is common to represent samples into a plot to visualize the beta diversity and to identify possible data structures.

The goal of ordination plots is to represent graphically the multidimensional data into two or three orthogonal axes, preserving the main trends of the data as well as possible. A distance matrix $\mathbf{D}$ usually defines the information to represent in these graphical representations. While Principal Components Analysis (PCA) preserves Euclidean distance, in microbiology other alternatives are required since Euclidean distance is not appropriate for microbiome abundance tables. We highlight two different approaches for the graphical representation: the Principal Coordinates Analysis (PCoA), also known as Multidimensional Scaling (MDS), and the Non-Metric Multidimensional Scaling (NMDS).

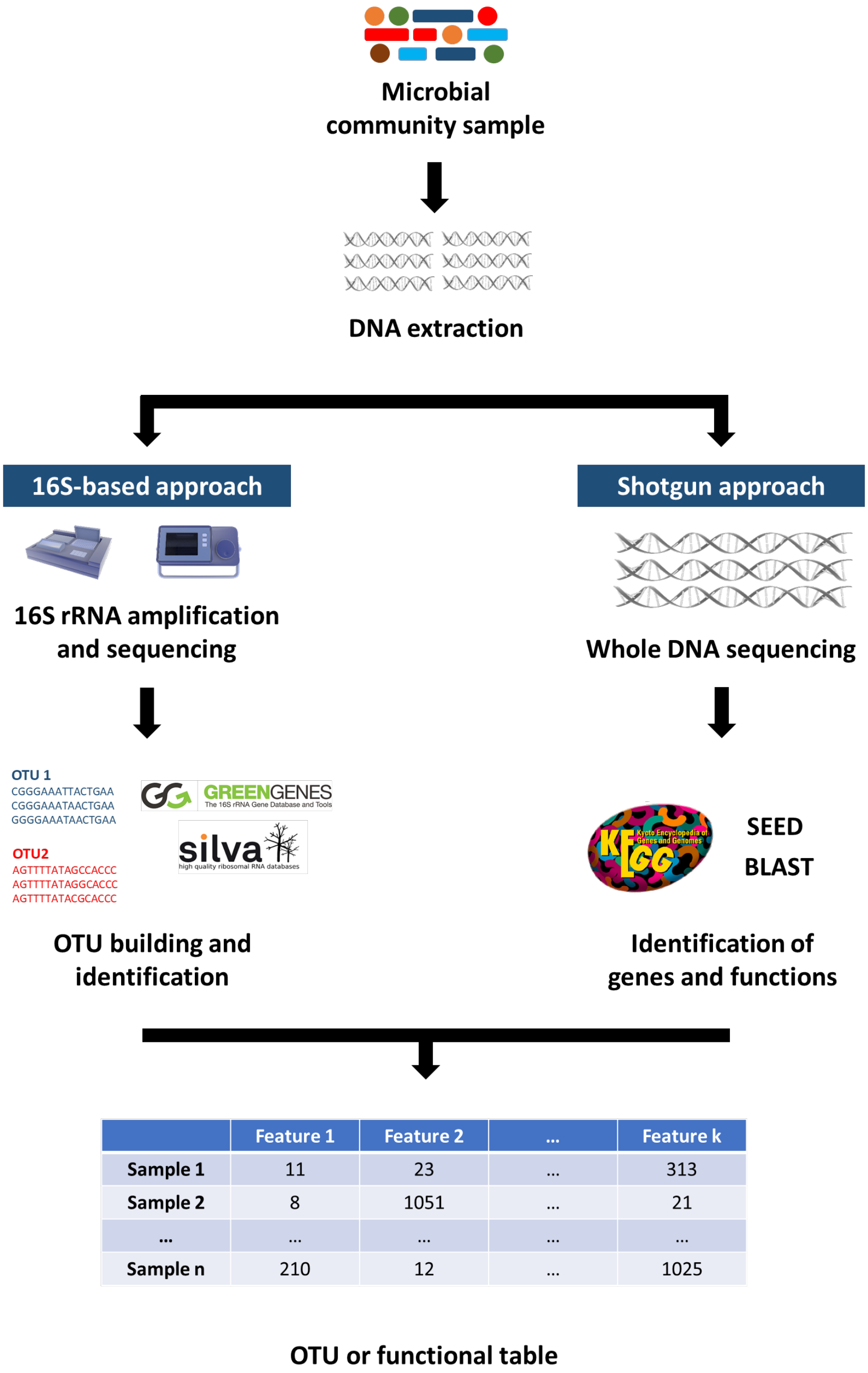
PCoA is based on eigenvalue decomposition of $\mathbf{D_c}'\mathbf{D_c}$ where $D_c = \mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{11'})$ is the centred distance matrix. When $\mathbf{D}$ is defined by the Euclidean distance, PCoA results exactly the same as PCA. Nevertheless, as ecological distances are more informative for microbiome data, other alternatives are often used. Care must be taken with PCoA if the selected distance is not metric, because some eigenvalues may be negative and then, the graphical representation will not be exactly a perfect representation of $\mathbf{D}$.

In order to avoid this problem NMDS is more commonly used. Also based on the distance matrix $\mathbf{D}$, NMDS maximizes rank-based correlation between original distances and those shown in the new ordination space, meaning that for a particular sample, its closest neighbour in the graphic is also the closest sample in $\mathbf{D}$. The positions of samples in the graphic are computed in an iterative procedure where, starting with a random configuration, they are modified so that the discrepancy between the rank defined from $\mathbf{D}$ and the resulted for the plot is minimized. This discrepancy is measured through a parameter known as *stress*.

1.2.4 Differential abundance testing

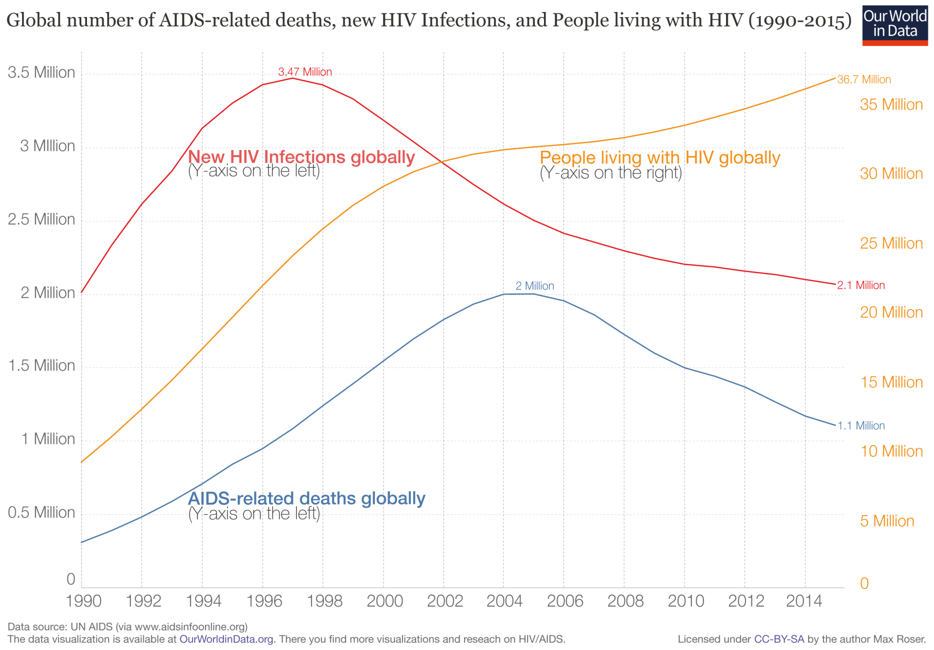
Graphical representations of the microbiome abundance matrix may suggest differences between groups of samples. Sometimes these groups are established before starting the analysis (for example HIV-positive and HIV-negative individuals), and other times they are defined by the user from ordination plots. Differences in composition can be evaluated globally (multivariate) or analyzing each particular taxon (univariate). Each approach answers a different question, the multivariate approach tests the global association between the microbiome and the output of interest while the univariate approach tests which specific taxon are associated with the outcome.

Different proposals are available to test for global differences between two groups of samples. On the one hand, there are distance-based methods such as *PERMANOVA* (Anderson, 2001), *Analysis of Similarity* (ANOSIM) (Clarke, 1993) and Kernel machine regression (Zhao et al., 2015) that use a distance matrix $\mathbf{D}$ to measure the significance of differences in composition. On the other hand, there are model-based methods, like LaRosa et al. that consider the Dirichlet-Multinomial distribution for modelling the whole composition and test for global associations (la Rosa et al., 2012). If significant global differences are detected between two groups, it results interesting to deepen in the analysis and specify which particular taxa are mainly making the difference. While Poisson is the main distribution for modelling counts, the mean-variance equality it assumes differs from the characteristics present in microbiome abundance tables. So, the Negative Binomial (NB) distribution, as an extension of the Poisson distribution allowing differences between mean and variance, is considered to model the counts by some authors (Robinson et al., 2009; Anders and Huber, 2010; Love et al., 2014). Based on NB distribution we highlight {`edgeR`} (Robinson et al., 2009) and {`DESeq2`} (Love et al., 2014) R packages, which by a Generalized Linear Model they test if taxon's distribution can be considered significantly different. Although the NB distribution is the most used for the analysis of individual taxa, there are other alternatives in the literature. For example, for datasets with many zeros, the Zero Inflated family of distributions is proposed to model better the shape of each feature (Paulson et al., 2013; Xu et al., 2015; Zhang et al., 2016).


Microbial community sample
DNA extraction
16S-based approach
Shotgun approach
16S rRNA amplification and sequencing
Whole DNA sequencing
OTU 1
CGGGAAATTACTGAA
CGGGAAATAACTGAA
GGGGAAATAACTGAA
OTU2
AGTTTTATAGCCACCC
AGTTTTATAGGCACCC
AGTTTTATACGCACCC
GREENGENES
The 16S rRNA Gene Database and Tools
silva
high quality ribosomal RNA databases
KEGG
Kyoto Encyclopedia of Genes and Genomes
SEED
BLAST
OTU building and identification
Identification of genes and functions
OTU or functional table

	Feature 1	Feature 2	...	Feature k
Sample 1	11	23	...	313
Sample 2	8	1051	...	21
...	...	...	...	...
Sample n	210	12	...	1025

Figure 1.3: *Scheme of 16S rRNA sequencing and shotgun sequencing for extracting the microbiological content of a sample.*



Global number of AIDS-related deaths, new HIV Infections, and People living with HIV (1990-2015)

Data source: UN AIDS (via www.aidsinfoonline.org)
The data visualization is available at OurWorldinData.org. There you find more visualizations and reseach on HIV/AIDS.

Figure 1.4: *Global number of AIDS-related deaths, new HIV infections, and people living with HIV in 1990-2014 period.*

CHAPTER 2

Compositional data and microbiome analysis

As introduced before, we denote the microbiome abundance table as a matrix $\mathbf{X}$ with n rows, each of them associated to a particular sample, and k columns referred to the different taxa or bacteria included in the study. Each entry x_{ij} represents the number of sequences (reads) corresponding to taxon j in sample i. The characteristics of such microbiome abundance matrices give rise to different problems for their analysis. We highlight three of them:

1. The total number of counts per sample is highly variable along individuals.
2. The total number of counts per sample is constrained by the maximum number of sequence reads of the DNA sequencer.
3. Abundance tables typically contain a large proportion of zeros.

The first issue, i.e., the large variability of the total counts per sample, is usually addressed through normalization of raw counts before the analysis and was already discussed in the introduction.

The second issue arises because the total number of counts per sample is constrained by the maximum number of sequence reads that the DNA sequencer can provide. This total count constraint induces strong dependencies among the abundances of the different taxa: an increase in abundance of one taxon requires the decrease of the observed number of counts for some others so that the total number of counts does not exceed the specified sequencing depth. This constraint induces the compositional nature of microbiome data.

Finally, the third issue is related to the sparsity of abundance tables, that is, the large proportion of zeros in them.

In this chapter, we focus on the second and third issues, how to properly analyze microbiome compositional data and how to deal with zeros in the context of compositional data analysis (CoDA). As we will see, in a proper compositional data analysis the first issue is no longer relevant and no normalization of the data is needed before the analysis.

In the following sections we first introduce the notion of compositional data and describe three important effects of ignoring the compositional nature of microbiome abundance matrices: *spurious correlations*, *subcompositional incoherences*, and the *increase of type I error*. Then, we present the three most common transformations for compositional data analysis: *alr*, *clr* and *ilr*-transformations. Finally, we discuss some issues on the treatment of zeros for CoDA.

2.1 Principles of Compositional Data Analysis

The challenges of dealing with compositional data were already advised by Karl Pearson in 1897, who remarked the problem of spurious correlations. In 1986, J. Aitchison introduced the log-ratio approach and laid the foundations of Compositional Data Analysis (CoDA).

A *compositional vector*, or simply a *composition*, is a vector of k strictly positive components or parts

$$\mathbf{x} = (x_1, \ldots, x_k)\,;\ x_i > 0,\ i \in \{1, \ldots, k\} \tag{2.1}$$

with a constrained or noninformative total sum $\sum x_i$.

In a composition the value of each component is not informative by itself and the relevant information is contained in the ratios between the components or parts. Mathematically, the assertion that the relevant information is contained in the ratios between the components implies that two proportional compositions are equally informative in terms of CoDA. Accordingly, an equivalence relation is defined in order to agglomerate vectors carrying the same information.

Two vectors $\mathbf{x_1} = (x_{11}, \ldots, x_{1k})$ and $\mathbf{x_2} = (x_{21}, \ldots, x_{2k})$ are *composition-*

ally equivalent (denoted by $=_a$), if they are proportional, that is:

$$\mathbf{x}_1 =_a \mathbf{x}_2 \iff \exists p > 0,\ \mathbf{x}_1 = p\mathbf{x}_2 \tag{2.2}$$

Each equivalence class has a representative in the *unit simplex* defined as:

$$\mathcal{S}^k = \{\mathbf{x} = (x_1, \ldots, x_k)\,,\ x_i > 0,\ \sum_{i=1}^{k} x_i = 1\} \tag{2.3}$$

We adopt the unit simplex for simplicity though any other constant constraint for the total sum can be considered. Representatives in the unit simplex are obtained by means of the so-called *closure operation* which consists on dividing each component by the sum of the components, thus scaling the vector to the constant sum 1. Formally, given a composition $\mathbf{x} = (x_1, \ldots, x_k)$ the closure of $\mathbf{x}$ is defined as

$$\mathcal{C}\left(\mathbf{x}\right) = \left({}^{x_1}/\sum_{i=1}^{k} x_i, \ldots, {}^{x_k}/\sum_{i=1}^{k} x_i \right).$$

In microbiome analysis, for example, both the raw counts and their transformation into relative abundances or proportions belong to the same equivalence class and they carry the same relative information.

The simplex is thus the sample space of compositional data. Furthermore, the simplex has a Euclidean structure when the so-called Aitchison geometry in the simplex is considered (Section 2.3).

Three important conditions should be fulfilled for a proper analysis of compositions (Aitchison, 1986): *permutation invariance*, *scale invariance* and *subcompositional coherence.*

- *Permutation invariance*: a change in the order of the parts in the composition should not affect the results.
- *Scale invariance*: any function f used for the analysis of compositional data must be invariant for any element of the same compositionally equivalent class; that is,

  $$f(\mathbf{x_1}) = f(\mathbf{x_2}),\ \forall \mathbf{x_1}, \mathbf{x_2},\ \mathbf{x_1} =_a \mathbf{x_2}$$

- *Subcompositional coherence*: the results that are obtained when a subset of components is analyzed should not contradict those obtained when analyzing the whole composition. In the context of microbiome analysis this principle is important because we usually work with subcompositions, obtained after filtering out the most low-abundant taxa.

2.2 Issues when the compositionality is ignored

As already discussed in the introduction of this chapter, microbiome data is compositional because the information that abundance tables contain (number of DNA sequences) is relative: the total counts per sample is constrained to the sequencing depth and thus they do not represent the true microbiome abundance in the sample. However, many microbiome analyses are performed using standard statistical methods. So, a natural question to be asked is: what are the implications of ignoring the compositional nature of the data in microbiome studies?

In this section we describe three main problems that may arise when ignoring the compositionality of microbiome data: *spurious correlations*, *subcompositional incoherences*, and *increase of type I error.*

2.2.1 Spurious correlations

A consequence of the total sum constraint that characterizes compositional data is that, necessarily, the correlation between some of the components will be negative, regardless of the underlying relationship of the components. This kind of artificial or spurious correlation was already highlighted by Pearson in 1896 (Pearson, 1896). For simplicity we illustrate the problem for a composition $\mathbf{X} = (X_1, \ldots, X_k)$ constrained to a constant sum equal to 1. On one hand we have:

$$cov(X_1, \sum_{i=1}^{k} X_i) = cov(X_1, 1) = 0$$

On the other hand:

$$cov(X_1, \sum_{i=1}^{k} X_i) = cov(X_1, X_1) + \cdots + cov(X_1, X_k)$$

Thus,

$$cov(X_1, X_1) + \cdots + cov(X_1, X_k) = 0$$

which implies that

$$-var(X_1) = cov(X_1, X_2) + \cdots + cov(X_1, X_k)$$

The left hand side of the last equation is negative (except in the particular case that the first component is constant). Thus, at least one of the covariances on the right side is enforced to be negative.

The presence of spurious correlations have important implications since correlations are the basis of many statistical analyses, as important as, linear regression or principal components analysis. Thus, if we know for sure that artificial correlations arise, how could we trust the results of analysis based on correlations?

2.2.2 Subcompositional incoherences

Given a composition **x**, a subcomposition is a vector obtained from a subset of the parts of **x** or, equivalently, by removing some of its components.

One of the principles for a proper analysis of compositions is subcompositional coherence, defined as the agreement between the results obtained for the whole composition and those obtained for a subcomposition. Standard tools cannot ensure this agreement and incoherences arise specially when working with correlations and distances, as we describe below. The problem in microbiome analysis is that we usually work with subcompositions since the whole composition is rarely available.

Subcompositional incoherences of correlations

When standard Pearson correlations are computed on a subcomposition, not only size differences, but also changes in the sign of the association can be found with respect to the correlations obtained with the whole composition. This is illustrated in Table 2.1 with a small example involving four taxa.

The table on the left in the first row represents the whole composition, while the table on the right in the same row is the subcomposition when the fourth component has not been considered and the proportions have been recalculated. The second row contains the correlation matrices for the whole composition (left) and for the subcomposition (right). Pearson correlation between T1 and T2 is equal to 1 for the whole composition while is almost zero (-0.03) for the subcomposition, which is a large difference in correlation value and a change in the association sign. Similarly, we can observe important differences in the correlations between T1 and T3, being initially -0.5, and -0.91 when the fourth component is omitted from the study. This toy example illustrates the erratic results that may be obtained when computing correlations in the whole composition and in a subcomposition. This makes the results based on correlation analysis highly unreliable.

Subcompositional incoherences of distances

Many methods for microbiome analysis involve the computation of distances. However, most commonly used distances in microbiome analysis are not subcompositionally coherent because do not fulfill the condition of *subcompositional dominance*, an important property for the reliability of the results in any distance-based analysis.

Let d be a distance or dissimilarity measure, $\mathbf{x_1}$ and $\mathbf{x_2}$ two compositions and $\mathbf{s_{x_1}}$ and $\mathbf{s_{x_2}}$ two subcompositions with the same number of components.

	T1	**T2**	**T3**	**T4**
Sample 1	0.1	0.2	0.1	0.6
Sample 2	0.1	0.2	0.2	0.5
Sample 3	0.3	0.3	0.1	0.3

	T1	**T2**	**T3**
Sample 1	0.25	0.5	0.25
Sample 2	0.2	0.4	0.4
Sample 3	0.43	0.43	0.14

	T1	**T2**	**T3**	**T4**
T1	1	1	**-0.5**	-0.945
T2		1	-0.5	-0.945
T3			1	0.189
T4				1

	T1	**T2**	**T3**
T1	1	**-0.03**	**-0.91**
T2		1	-0.37
T3			1

Table 2.1: *On the top: the whole composition relative abundances (left) and the subcomposition relative abundances calculated after removing the fourth component (right). Below, the corresponding correlation matrices.*

We say d is *subcompositionally dominant* if $d(\mathbf{s}_{\mathbf{x_1}}, \mathbf{s}_{\mathbf{x_2}}) \leq d(\mathbf{x_1}, \mathbf{x_2})$, for any possible subcomposition $\mathbf{s}_{\mathbf{x_1}}$ and $\mathbf{s}_{\mathbf{x_2}}$ of $\mathbf{x_1}$ and $\mathbf{x_2}$, respectively. Intuitively, this property can be explained as the condition that the distance between two points in a multi-dimensional space should always be larger than their distance when projected in a lower dimensional space (subcomposition).

Distances such as Bray-Curtis or weighted UniFrac do not comply this property. This is a problem because the distribution of the points in the projected space do not represent the distribution of points in the original space.

In Table 2.2 we present a small example for illustrating this. The two tables of counts in the first row show the whole composition (left) and the subcomposition obtained by removing the fourth component (right) for two different samples. The two tables in the second row represent their respective proportions (down). If we compute the Bray-Curtis distance between the two samples, we obtain a value of 0.42 for the whole composition and a value of 0.44 for the subcomposition. Thus, subcompositional dominance is not fulfilled which may affect the interpretation of the analysis.

	T1	**T2**	**T3**	**T4**
Sample 1	20	15	1	2
Sample 2	3	17	6	3

	T1	**T2**	**T3**
Sample 1	20	15	1
Sample 2	3	17	6

	T1	**T2**	**T3**	**T4**
Sample 1	0.53	0.39	0.03	0.05
Sample 2	0.10	0.59	0.20	0.11

	T1	**T2**	**T3**
Sample 1	0.56	0.42	0.02
Sample 2	0.12	0.65	0.23

Table 2.2: *On the top: raw counts for the whole composition (left) and for the subcomposition obtained by removing the fourth component (right). Below, the corresponding proportion matrices.*

2.2.3 Increase of type I error

Differential abundance testing, i.e., the search of microbial taxa presenting different abundances between groups of samples, is highly affected when the compositional nature of microbiome datasets is not acknowledged. More concretely, ignoring the compositionality of the data results in an increase of false-positive findings.

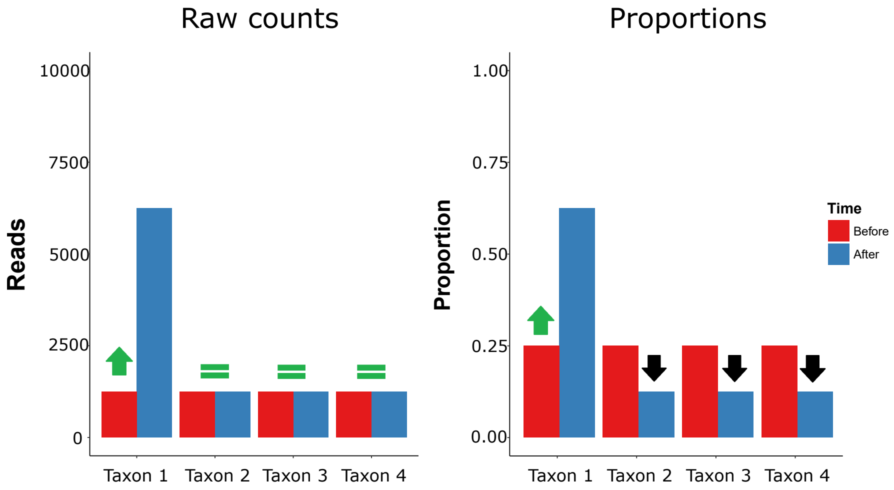


Figure 2.1: *On the left, changes in raw counts only for the first taxon. On the right, how it affects proportions.*

We illustrate this fact with a small toy example (Figure 2.1) involving a microbiome composition with four parts and two conditions, before and after a disease that causes the increase in the abundance of Taxon 1. This example is inspired in what happens in the gastric microbiota when it is invaded by Helicobacter pylori. In this example we assume for simplicity that before the disease all taxa are equally abundant, presenting around 1250 reads each one. Then, we assume that the disease modifies the microbiome in such a way that the first taxon increases its abundance to 6250 reads while the other taxa remain with 1250 reads. The barplot on the left of Figure 2.1, shows that Taxon 1 is the unique taxon presenting a change in raw reads before and after the disease. However, when reads are transformed into proportions, changes all over the four taxa are observed before and after the disease (right side of Figure 2.1) So, though the increment in Taxon 1 has been the unique change (emulating the behaviour of Helicobacter Pylori), the normalization of raw values induces changes in the relative abundance of the other components and all 4 taxa will be identified as differentially abundant.

In general, the constraint that characterizes compositional data causes that changes in one or several taxa induce changes into some others, thus incrementing the type I error in many differential abundance analyses.

2.3 The Aitchison geometry and the representation in coordinates

2.3.1 The log-ratio approach

In 1986, J. Aitchison put the basis of Compositional Data Analysis (CoDA) by introducing what is now called the Aitchison's log-ratio approach. The log-ratio analysis was introduced in order to meet the principle of scale invariance; as stated by Aitchison *"any meaningful (scale-invariant) function of a composition can be expressed in terms of ratios of its components"* (Aitchison, 1986). Because the logarithmic transformation makes ratios mathematically more tractable, the simplest invariant function is given by the *log-ratio* between two components, that is:

$$f(\mathbf{x}) = \log\left(\frac{x_i}{x_j}\right), \ i, j \in \{1, \ldots, k\},$$

whose generalization is named a *logcontrast* function, that is, a linear combination of logarithms of the components with the restriction that the sum of the coefficients is equal to 0:

$$f(\mathbf{x}) = \sum_{i=1}^{k} a_i \log(x_i); \qquad \sum_{i=1}^{k} a_i = 0.$$

There are two operations in the k-dimensional simplex which give it a vector space structure. These two operations are the *perturbation* and *powering.*

Given two compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^k$, the *perturbation* of $\mathbf{x}$ by $\mathbf{y}$ is given as

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left(x_1 y_1, x_2 y_2, \ldots, x_k y_k\right)$$

and the *power transformation* or powering of the composition $\mathbf{x}$ by a constant $\alpha \in \mathbb{R}$ is defined as

$$\alpha \odot \mathbf{x} = \mathcal{C}\left(x_1^{\alpha}, x_2^{\alpha}, \ldots, x_k^{\alpha}\right).$$

Moreover, we can talk about an Euclidean vector space, taking the following inner product and the associated norm and distance.

The *inner product* of $\mathbf{x}$ and $\mathbf{y}$ is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2k} \sum_{i=1}^{k} \sum_{j=1}^{k} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}$$

Thus, the *norm* of $\mathbf{x}$ is given by

$$\parallel \mathbf{x} \parallel_a := \langle \mathbf{x}, \mathbf{x} \rangle_a = \sqrt{\frac{1}{2k} \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\log \frac{x_i}{x_j} \right)}.$$

And the *distance* between $\mathbf{x}$ and $\mathbf{y}$ compositions, known as *Aitchison distance* is the norm of the difference between them, that is,

$$d_a \left(\mathbf{x}, \mathbf{y} \right) = \parallel \mathbf{x} \ominus \mathbf{y} \parallel_a = \sqrt{\frac{1}{2k} \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2}.$$

2.3.2 Coordinate representation

An alternative to perform the analysis of compositional data within the simplex is to transform the compositions so that the transformed observations belong to the real space and classical statistical analysis could be applied. Several data transformations have been proposed that fulfil the principles of compositional data analysis. All them are based on log-ratios between components. We highlight the *alr*, *clr* and *ilr* transformations described below.

The *additive log-ratio* transformation (*alr*) is the first proposal introduced by Aitchison (Aitchison, 1986). Taking one part as the reference, for instance x_k, the *alr* transformation is defined as:

$$\begin{aligned} alr \colon \mathcal{S}^k &\rightarrow \mathbb{R}^{k-1} \\ \mathbf{x} = (x_1, \ldots, x_k) &\rightarrow \left(log \left(\frac{x_1}{x_k} \right), \ldots, log \left(\frac{x_{k-1}}{x_k} \right) \right) \end{aligned} \tag{2.4}$$

alr-transformation returns a new vector with $(k-1)$-components, one less than the initial composition without any restriction for the sum of the new components and with each coordinate taking values all over the real line. Unconstrained multivariate analyses can be applied on the transformed values.

Aitchison also defined the *centered log-ratio* transformation (*clr*) to treat the parts symmetrically. The *clr* transformation is given by:

$$\begin{aligned} clr\colon \mathcal{S}^k &\to \mathcal{U} \subset \mathbb{R}^k \\ \mathbf{x} = (x_1, \ldots, x_k) &\to \left(log\left(\frac{x_1}{g(\mathbf{x})}\right), \ldots, log\left(\frac{x_k}{g(\mathbf{x})}\right)\right) \end{aligned} \tag{2.5}$$

where $\mathcal{U} = \{\mathbf{z} \in \mathbb{R}^k : \sum\limits_{i=1}^{k} z_i = 0\}$ defines a subspace of the k-dimensional real space and $g(\mathbf{x}) = (\prod\limits_{i=1}^{k} x_i)^{\frac{1}{k}}$ represents the geometric mean of the composition $\mathbf{x}$.

One characteristic of the *clr* transformation is that the transformed components are restricted to have a sum equal to zero:

$$\begin{aligned} \sum_{i=1}^{k} clr(\mathbf{x})_i = \sum_{i=1}^{k} log\left(\frac{x_i}{g(\mathbf{x})}\right) = log\left(\prod_{i=1}^{k} \frac{x_i}{g(\mathbf{x})}\right) = \\ = log\left(\frac{1}{g(\mathbf{x})^k}\prod_{i=1}^{k} x_i\right) = log(1) = 0. \end{aligned}$$

This constant sum of zero restriction of the *clr*-transformed coordinates implies that some common statistical analyses can not be applied after the *clr* transformation because of a singular covariance matrix $(det(clr(\mathbf{X})) = 0)$. Aitchison distance, introduced in the previous section, can be expressed in terms of the clr-transformation. Given two compositions $\mathbf{x_1} = (x_{11}, \dots, x_{1k})$ and $\mathbf{x_2} = (x_{21}, \dots, x_{2k})$, Aitchison distance is the Euclidean distance between the *clr*-transformed vectors of $\mathbf{x_1}$ and $\mathbf{x_2}$:

$$d_a(\mathbf{x_1}, \mathbf{x_2}) = \sqrt{\sum_{i=1}^{k} (clr(\mathbf{x_1}) - clr(\mathbf{x_2}))^2} \tag{2.6}$$

The third alternative is named *isometric log-ratio transformation* (*ilr*) and consists on the the representation of a composition given a particular orthonormal basis in $\mathcal{S}^k$ (Egozcue et al., 2003). It overcomes the problem of the singular covariance matrix present in the *clr*-transformation. The *ilr*-transformation can be defined as:

$$\begin{aligned} ilr\colon \mathcal{S}^k &\rightarrow \mathbb{R}^{k-1} \\ \mathbf{x} = (x_1, \dots, x_k) &\rightarrow (ilr_1, \dots, ilr_{k-1}) \end{aligned} \tag{2.7}$$

The formula given in Equation 2.7 is not explicit because multiple orthonormal basis can be defined in $\mathcal{S}^k$. One way to define one of them is to link the basis to a sequential binary partition (SBP), that is, a hierarchy selection of the parts of a composition (Egozcue and Pawlowsky-Glahn, 2006). The process consists on $(k-1)$ steps where first, the whole composition is divided into two groups of features. The membership of each part is expressed with a $+1$ or a -1 depending on the group. In the remaining steps, each group of parts previously defined is subdivided into two sets until no additional subdivision is possible because the subset only contains one component. As a result, a $((k-1) \times k)$-dimensional matrix $\mathbf{S}$ defines the binary partition, where each row characterizes the groups defined at that step through a codification where $+1$, -1 and 0 specify if the feature is included in one group, the other one, or none of them, respectively. $\mathbf{S}$ matrix is used to define $\boldsymbol{\Psi} \in \mathcal{M}_{(k-1,k)}$, a matrix including the vectors for the orthonormal basis associated to the SBP. Each cell in $\boldsymbol{\Psi}$ is defined from $\mathbf{S}$ as:

$$\begin{aligned} \psi_{ij} &= 0 && \text{if } s_{ij} = 0 \\ \psi_{ij} &= \frac{1}{k_+^i}\sqrt{\frac{k_+^i k_-^i}{k_+^i + k_-^i}} && \text{if } s_{ij} = 1 \\ \psi_{ij} &= -\frac{1}{k_-^i}\sqrt{\frac{k_+^i k_-^i}{k_+^i + k_-^i}} && \text{if } s_{ij} = -1 \end{aligned} \tag{2.8}$$

where k_+^i and k_-^i represent the number of $+1$ and -1 values in the i-th row of $\mathbf{S}$, respectively. Hence, each *ilr* - coordinate is defined from the corresponding row of $\boldsymbol{\Psi}$ as:

$$ilr_i = \sum_{j=1}^{k} \psi_{ij} \log x_j \tag{2.9}$$

Rewriting Equation 2.9 we obtain:

$$ilr_i = \sqrt{\frac{k_+^i \cdot k_-^i}{k_+^i + k_-^i}} \log \left(\frac{\left(\prod_{l \in I_+^i} x_l\right)^{1/k_+^i}}{\left(\prod_{j \in I_-^i} x_j\right)^{1/k_-^i}} \right) \tag{2.10}$$

where I_+^i and I_-^i are the sets of indices of those components codified with a $+1$ and a -1 in the i-th row of $\mathbf{S}$, respectively. Each *ilr*-coordinate is also known as a *normalised balance* or simply a *balance.* Expanding Equation 2.10, a simpler expression is obtained as given in Equation 2.11, resulting on a value proportional to the difference between the means of two log-transformed set of variables.

$$ilr_i = \sqrt{\frac{k_+^i \cdot k_-^i}{k_+^i + k_-^i}} \left(\frac{1}{k_+^i} \sum_{l \in I_+^i} \log x_l - \frac{1}{k_-^i} \sum_{j \in I_-^i} \log x_j \right) \tag{2.11}$$

This concept of *balance* will be important in the fourth chapter of this thesis where we define microbial signatures as balances between two groups of taxa and propose a method for obtaining the balance that is most associated with the response variable of interest.

These three transformations of compositional data, especially the *centered log-ratio* and the *isometric log-ratio*, are the most extended in CoDA following a common procedure that involves these steps: first, the compositional problem is formulated in terms of its components; then, the corresponding data transformation is implemented and the appropriate standard multivariate statistical method is applied for finally, translate back the results into terms of initial compositions. (Pawlowsky-Glahn and Buccianti, 2011).

2.4 Zeros in microbiome data analysis

The presence of zeros in microbiome abundance tables makes difficult the application of compositional data analysis which, based on log-ratios, is not defined when zeros are included in the denominator. In this section we first define the different kind of zeros that can be found in a microbiome study and introduce two different ways of modifying abundance matrices in order to avoid the zeros and make possible the use of log-ratio analysis.

2.4.1 Different kinds of zeros

When a microbiome count, x_{ij}, in a microbiome abundance table is equal to zero, this means that no sequences of the j-th taxon have been found for the i-th sample. Depending on the particular study and the method used for measuring the information, we can distinguish between three different types of zeros: *count zeros*, *rounded zeros*, or *essential zeros* (Pawlowsky-Glahn and Buccianti, 2011).

Count zeros appear in sampling studies involving counts. They arise in processes that may be compared with a multinomial experiment. Thus, we consider the presence of all the categories (taxa) with a positive probability of appearing in a sample, but some of them in such a low proportion that depending on the sample size, they are more or less probable to appear. When count zeros are present in a dataset, they are usually replaced by small positive values.

Rounded zeros, also known as values below detection limit, are those null values resulted because its real abundance is below the maximum round-off error or detection limit. Similarly to count zeros, null values considered as rounded zeros are usually replaced by small quantities.

Essential zeros are null values representing the total absence of the taxon in sample's environment. For these type of zeros, the sampling process and sequencing depth are not important, because although increasing the total number of reads, we would not find any read associated to that taxon. So, in this case the observed zeros represent the true content and it does not make sense to replace them by a low positive value. This situation requires a special analysis (see (Martín-Ferńandez et al., 2011) for more details).

Since, in general, we do not have enough information for ensuring the complete absence of a bacteria in the environment of a sample, in what follows we consider all null values as count zeros. That is, we assume that any microbial taxa can be present in any sample, though, some of them in such a small proportion that they might not be detected during the sequencing process.

2.4.2 Dealing with count zeros

As already emphasized in previous chapters, microbiome abundance tables are sparse, that is, they contain many zeros. This should be properly addressed before compositional data methods are applied.

In microbiome analysis this problem has not been treated in depth and the most common way to tackle this issue is by adding a pseudo-count to the whole matrix. We discuss in the following paragraphs some of the approaches that have been proposed in the last decade about the proper treatment of rounded or count zeros in compositional data sets.

Substitution by a pseudo-count

This method consists on overcoming the problem of zeros by adding a constant value to the abundance matrix. Here, we describe two different approaches used in the most recently CoDA literature.

The most extended way of avoiding zeros is by adding one to all the cells in the abundance matrix, independently of their initial value. For instance, Silverman et al. (2017) and Morton et al. (2017) include this replacement before their compositional analysis (Silverman et al., 2017; Morton et al., 2017).

The second alternative is to replace zeros by the 65% of the detection limit. Since a microbiome abundance table is defined by counts, the lowest non-zero value that can be found is 1. Thus, in this case the method consists on replacing zeros by a pseudo-count equal to 0.65. This idea is supported by Martín-Fernández et al. (2015) when zeros do not represent more than the 10% of the total number of cells in the abundance matrix. Washburne et al. (2017) apply this technique in their CoDA proposal for identifying the phylogenetic factors driving patterns in the composition of microbial community (Washburne et al., 2017).

It is important to highlight that even though these two alternatives are similar, they affect log-ratio analysis in a different way. While replacing zeros by the 65% of the detection limit preserves constant the log-ratio between non-zero parts, by adding one count to all the cells in the table, these log-ratios are modified.

Bayesian-Multiplicative treatment

Instead of replacing zeros by a pseudo-count or adding a constant value to the matrix, Martín-Fernández et al. (2015) propose the *Bayesian-Multiplicative (BM) treatment*, a replacement involving a Bayesian inference on the zero values, and a multiplicative modification of non-zero values in the composition (Martín-Fernández et al., 2015).

BM-replacement modifies both zero and non-zero values of each composition $\mathbf{x_i}$ by another composition $\mathbf{r}_i = (r_{i1}, \ldots, r_{ik})$, so that the original ratios between parts without zeros are preserved, that is:

$$\frac{r_{ij}}{r_{ik}} = \frac{x_{ij}}{x_{ik}} \ , \ x_{ij} \neq 0 \ , \ x_{ik} \neq 0. \tag{2.12}$$

Additional information about the BM-replacement is detailed in (Martín-Fernández et al., 2015). There, they consider several prior distributions for the Bayesian inference in the BM-replacement, concluding that the Geometric Bayesian Multiplicative (GBM) replacement is the one providing best results. GBM replaces $\mathbf{x_i}$ by $\mathbf{r_i}$, whose components are given by:

$$r_{ij} = \begin{cases} \dfrac{\hat{m}_{ij}}{g_i \sum x_i + 1} & \text{if } x_{ij} = 0 \\ x_{ij}\left(1 - \dfrac{\sum\limits_{k|x_{ik}=0} \hat{m}_{ik}}{g_i \sum x_i + 1}\right) & \text{if } x_{ij} > 0 \end{cases} \tag{2.13}$$

where

$$
\begin{aligned}
\alpha_{ij} &= \sum_{\substack{r=1\\ r\neq i}}^{n} x_{rj} && \text{is the sum of counts for a particular taxa without considering the } i\text{-th sample}\\
\hat{m}_{ij} &= \frac{\alpha_{ij}}{\sum\limits_{l=1}^{k} \alpha_{il}} && \text{is a particular estimation for the prior}\\
g_i &= \left(\prod_{j=1}^{k} \hat{m}_{ij}\right)^{1/k} && \text{is the geometric mean of the vector } \hat{\mathbf{m}}_i
\end{aligned}
\tag{2.14}
$$

Based on the study of Martín Fernández et al. (2015), we set in our proposals the GBM replacement as the default zero replacement procedure prior to the analysis. GBM replacement is appropriate since it preserves the log-ratio values between non-zero components allowing the use of prior information for a better zero replacement. However, we also include in our algorithms the option of adding one count to the whole abundance matrix due to its extended use.

Depending of the degree of sparsity of the microbiome abundance matrix, zero-replacement conducted for the use of compositional tools may have an important effect on the results. If the proportion of zeros is high, the matrix requires the replacement of many cells which may represent an important change of the initial dataset. We have implemented some small simulations in order to explore the robustness of results when zero replacement is performed. The simulations evaluated how consistent was Aitchison distance depending on the zero replacement method, the addition of a constant equal to 1 to all cells in the abundance table and the Bayesian-multiplicative zero replacement. We compare Aitchison distance between samples for the initial dataset with Aitchison distance as sparsity of the matrix increased. This increase in the proportion of zeros was introduced through the rarefaction (subsampling) of the initial dataset. The results suggest that the Aitchison is very robust to zero replacements and that both zero replacement methods have similar performance.

CHAPTER 3

Kernel machine regression for microbial compositional data analysis

One of the first and principal questions to tackle in a microbiome study is whether there is any relationship between the microbiome and a response variable of interest such as disease status or another phenotype. This global question is addressed with multivariate methods that can broadly be divided into two groups: model-based methods and distance-based methods. In this chapter, we briefly describe the most common multivariate proposals for microbiome analysis, both, model-based and distance-based, but we focus mainly on a specially interesting approach: Kernel machine regression (KMR) and a particular variant for microbiome data analysis proposed by Zhao et al. (Zhao et al., 2015). Since this approach does not take into consideration the compositional nature of microbiome data, their results may be questioned. In this work we suggest how to adapt Kernel machine regression to compositional data analysis through the use of a subcompositionally dominant distance. Moreover, we define a method that, using a weighted version of KMR, provides a measure of the contribution of each taxon in the global association between the microbiome and the response variable.

3.1 Model-based multivariate methods for microbiome analysis

Most model-based multivariate methods for microbiome analysis assume that microbiome abundances follow a Dirichlet-Multinomial (DM) distribution. Raw counts for each sample in abundance tables can be interpreted as a particular essay from a multinomial experiment. However, the variability present in raw counts is usually larger than the determined by the multinomial model. This is due to the fixed underlying proportions assumed by the multinomial distribution, which does not fit with the heterogeneity present in microbiome samples. This overdispersion in microbiome abundance tables is accounted considering each proportion as a random variable and assuming a Dirichlet distribution.

Thus, given a set of k taxa $\mathbf{X} = (X_1, \ldots, X_k)$ and a particular vector of counts $\mathbf{x} = (x_1, \ldots, x_k)$, the Dirichlet-Multinomial distribution (DM) is defined as in Equation 3.1. $\Gamma(\cdot)$ denotes the Gamma function in Equation 3.1 and $\gamma = (\gamma_1, \ldots, \gamma_k)$ is a vector of positive parameters proportional to the expected mean for each taxon. x_+ and γ_+ denote the sum of $\mathbf{x}$ and γ, respectively, where γ_+ controls the dispersion. Higher values of γ_+ indicate a lower overdispersion and probability values closer to the expected mean.

$$P\left(\mathbf{X} = \mathbf{x}; \gamma\right) = \frac{x_+!}{x_1! \cdots x_k!} \frac{\Gamma(x_+ + 1)\Gamma(\gamma_+)}{\Gamma(x_+ + \gamma_+)} \prod_{j=1}^{k} \frac{\Gamma(x_j + \gamma_j)}{\Gamma(\gamma_j)\Gamma(x_j + 1)} \tag{3.1}$$

Based on DM distribution, we highlight two different approaches. The first one is introduced as an alternative to non-parametric models (la Rosa et al., 2012). After a proper estimation of the parameters of the distribution, they propose different tests for: determining if the overdispersion is high enough to justify the DM model instead of the multinomial distribution, if the expected frequencies for each taxon are equal to a given vector, or if several groups of samples share a common vector of expected frequencies.

The second approach, proposed by Chen et al. (2013), considers that the parameters γ_j in the DM-model (Equation 3.1) depend on a set of covariates $\mathbf{Z} = (Z_1, \ldots, Z_r)$ via the regression model in Equation 3.2 (Chen and Li, 2013). $\mathbf{z^i}$ represents the covariates for the i-th sample and β_{jl} the coefficient referent to the j-th taxon with respect to l-th covariate, whose sign and magnitude measure the effect of the l-th covariate on the j-th taxon.

$$\gamma_j(\mathbf{z^i}) = exp\left(\sum_{l=1}^{r} \beta_{jl} z_{il}\right) \tag{3.2}$$

The method includes a penalized likelihood approach to estimate the regression parameters and to select only the relevant associations, thus overcoming the possible loss of power that the high-dimensionality of the problem may induce.

We implemented this methodology in the context of the HIV-microbiome studies with the aim of characterizing dietary and gut microbiome association in HIV-1 positive individuals (Rivera-Pinto et al., 2017).

3.2 Distance-based multivariate methods for microbiome analysis

Distance-based methods are more extended than model-based methods in microbiome studies. This type of analyses are focused on a distance or dissimilarity matrix **D** that summarizes differences between samples with respect to the abundance of multiple microorganisms. The good performance of distance-based approaches strongly relies on the selected distance or dissimilarity measure. The Euclidean distance is not a good measure to describe the dissimilarity of two ecosystems (Beals, 1984). Though there are many different measures derived from the ecology field, the most extended distances in microbiology are Bray-Curtis and UniFrac family of distances.

As already introduced in previous chapters, distances are computed at first steps of the analysis and are used for graphical representation of the individuals in a non-metric multidimensional scaling plot (NMDS). This type of representations may be useful to identify possible data structures or groups of samples with a similar composition. As an example, Figure 3.1 represents Bray-Curtis distances between samples included in the IrsiCaixa

HIV study conducted at IrsiCaixa (Noguera-Julian et al., 2016) that, as described in previous chapters, consists of 156 samples (127 HIV-1 positive individuals (HIV+) and 29 HIV-negative controls (HIV-) divided into three risk groups: *hts* (heterosexual males and females), *msm* (men who have sex with men) and *pwid* (people who inject drugs). A clear difference in the distribution of individuals in the MSM group is manifest in Figure 3.1, suggesting that this group presents a different microbial composition than the other risk groups. In other words, MSM is a confounder that should be adjusted for in microbiome studies. This was one of the main findings in Noguera et al. (2016) since it has important implications for future HIV-microbiome studies and for the validity of previous ones. Some tests are available in the literature for measuring the significance of the differences between groups observed in a NMDS plot. Most of them are permutational tests, which evaluate if the centroids and dispersion of each group of samples is the same. Figure 3.2 shows an example where two pre-defined groups of samples are represented in a NMDS and the samples are linked to the centroid of the group through a line. The null hypothesis of distance-based tests specifies that there are no differences in composition among groups of samples, what in graphical terms means that all the groups have the same center of mass. In this framework, two popular tests are the *Analysis of similarity* (Clarke, 1993) and the *Permutational analysis of variance (PERMANOVA)* (Anderson, 2001), implemented in the `anosim()` and `adonis()` functions of the {`vegan`} R library, respectively. While the analysis of similarity takes a rank-based statistic to evaluate differences in the center of gravity of the different groups, PERMANOVA evaluates an score similar to the Firsher's F-ratio that, in the same way as ANOVA, compares the variability within groups (SSW) against the variability between groups (SSA), but partition of sums-of-squares (SS) is applied directly to dissimilarities. Equation 3.3 defines the F-statistic, where g is the number of groups considered in the study and n the total number of samples.

$$F = \frac{SSA/(g-1)}{SSW/(n-g)} \tag{3.3}$$

The total sum of squares (SS) is computed as expressed in Equation 3.4, where d_{ij} represents the distance between i-th and j-th samples, and n is the total number of samples.

$$SS = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^2 \tag{3.4}$$

The variability within groups (SSW) is computed as defined in Equation 3.5, where ϵ_{ij} is defined as a 1 if i-th and j-th samples belong to the same group and 0 otherwise.

$$SSW = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^2\epsilon_{ij} \tag{3.5}$$

Finally, because $SS = SSA + SSW$, the variability between groups can be easily computed as the difference between SS and SSW.

In the same way that the linear regression model is an extension of the ANOVA that allows the adjustment for covariates, there is also a methodology that extends PERMANOVA to a regression model approach: *Kernel* *machine regression* (KMR). In the next section we will describe in more detail this powerful distance-based regression model.

3.3 Kernel machine regression for microbiome analysis

Kernel machine regression is a semi-parametric regression model that includes a non-parametric component to associate a set of covariates $\mathbf{X}$, for instance microbiome abundances, with a response variable of interest $\mathbf{Y}$. It compares pairwise similarity in the outcome variable to pairwise similarity in the microbiome profile. The model allows the adjustment for additional covariates $\mathbf{Z}$. Kernel machine regression is expressed according to Equation 3.6a when the response variable is continuous, and to Equation 3.6b for a dichotomous outcome.

$$Y_i = \beta_0 + \beta'\mathbf{Z_i} + h(\mathbf{X_i}) + \epsilon_i \tag{3.6a}$$

$$logit(Y_i) = \beta_0 + \beta'\mathbf{Z_i} + h(\mathbf{X_i}) \tag{3.6b}$$

Equation 3.6a corresponds to a linear regression model with a non-parametric component given by $h(\cdot)$, and Equation 3.6b to a logistic regression model also with a non-parametric part. β_0 denotes the intercept, β is the vector of the regression coefficients for the covariate adjustment, and ϵ_i is an error term with mean 0 and variance σ^2. The non-parametric component $h(\mathbf{X})$ measures the relationship between the microbiome composition and the outcome. This association can be tested according to the following hypothesis:

$$\begin{aligned} H_0\colon &\quad h(\mathbf{X}) = 0 \\ H_1\colon &\quad h(\mathbf{X}) \neq 0 \end{aligned} \tag{3.7}$$

where the null hypothesis represents no association between microbiome composition and $\mathbf{Y}$. The non-parametric component $h(\mathbf{X})$ is related to a positive definite Kernel function $K(\cdot,\cdot)$ that measures the dissimilarity between the composition of two individuals. Depending on the complexity of the selected Kernel function, the user can obtain a measure that includes information about non-linear relationships or taxa interactions, among others. There are many ways of defining a Kernel matrix $\mathbf{K}$, but the most natural is to define it from a given distance matrix $\mathbf{D}$, as follows:

$$\mathbf{K} = -\frac{1}{2}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\mathbf{D}^2\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right) \tag{3.8}$$

where $\mathbf{I}$ denotes the identity matrix and $\mathbf{1}$ corresponds to a vector of ones (Pan, 2011).

Kernel machine regression is a special kind of mixed model (Liu et al., 2007, 2008; Gianola and van Kaam, 2008) where $h(\mathbf{X})$ is a subject-specific random effect that follows a normal distribution with mean 0 and variance $\tau\mathbf{K}$, that is, $h(\mathbf{X}) \sim N(0, \tau\mathbf{K})$. Thus, the association test defined in Equation 3.7 can be rewritten as

$$\begin{aligned} H_0\colon &\quad \tau = 0 \\ H_1\colon &\quad \tau \neq 0 \end{aligned} \tag{3.9}$$

Making use of the methodology developed for mixed models, the test in

Equation 3.9 can be evaluated through the statistic Q

$$Q = \frac{1}{2\phi}(\mathbf{Y} - \hat{\mathbf{Y}}_0)'\mathbf{K}(\mathbf{Y} - \hat{\mathbf{Y}}_0) \tag{3.10}$$

where $\hat{\mathbf{Y}}_\mathbf{0}$ is the predicted mean of $\mathbf{Y}$ under the null hypothesis, that is, when we remove $h(\mathbf{X_i})$ from equation 3.6a or equation 3.6b. ϕ is the dispersion parameter which is defined as 1 for the logistic regression, and the variance in the sample for linear regression. Under the null hypothesis, the statistic Q asymptotically follows a mixture of chi-squared distributions and the p-value can be obtained in different ways (Liu et al., 2008; Duchesne and Lafaye de Micheaux, 2010). Zhao et al. (2015) consider this test to be very conservative and, based on a previous work (Chen et al., 2016), they propose and implement some adjustments in the so called MiRKAT algorithm, a microbiome regression-based kernel association test available as an R package with the same name {`MiRKAT`} (Zhao et al., 2015).

Zhao et al. (2015) discuss how the power of the test may be affected by the selection of a good Kernel function. In order to choose the best distance-based kernel, they evaluate a collection of dissimilarities including the weighted UniFrac distance, unweighted UniFrac distance, Bray-Curtis dissimilarity or the generalized UniFrac distance. Since the subcompositional dominance of these measures cannot be guaranteed, none of them prevents from incoherences between distances based on the global composition and those based on a subset of the features. In the next section, we introduce a methodology which adapts the Kernel machine regression to compositional data analysis with the aim of overcoming subcompositional problems.

3.4 Kernel machine regression for microbiome compositional data

We name MiRKAT-CoDA the extension of Kernel machine regression for compositional data analysis through the use of the subcompositional dominant Aitchison distance (Aitchison, 1986). Given two compositions denoted by $\mathbf{x_1} = (x_{11}, \ldots, x_{1k})$ and $\mathbf{x_2} = (x_{21}, \ldots, x_{2k})$, Aitchison distance between $\mathbf{x_1}$ and $\mathbf{x_2}$ is defined as

$$d_A(\mathbf{x_1}, \mathbf{x_2}) = \sqrt{\sum_{i=1}^{k} \big(clr(\mathbf{x_1})_i - clr(\mathbf{x_2})_i\big)^2} \tag{3.11}$$

where $clr(\mathbf{x_1})$ and $clr(\mathbf{x_2})$ are the *clr*-transformed values of $\mathbf{x_1}$ and $\mathbf{x_2}$, respectively, that is, $clr(\mathbf{x_i}) = \Big(log(\frac{x_{i1}}{g(\mathbf{x_i})}), \ldots, log(\frac{x_{i2}}{g(\mathbf{x_i})})\Big)$, for $i = 1, 2$, being $g(\cdot)$ operator the geometric mean of a vector. Aitchison distance is just the Euclidean distance between the *clr*-transformed coordinates of the initial abundances. It can also be computed as the Euclidean distance of any *ilr*-transformed coordinates.

We propose to adapt Kernel machine regression for compositional data analysis by using Aitchison distance. Thus, the model can be expressed as in Equation 3.12a for a continuous response variable or as in Equation 3.12b for a dichotomous outcome, both including $h(clr(\mathbf{X}))$, which denotes the non-parametric component applied to the *clr*-transformed values of $\mathbf{X}$.

$$Y_i = \beta_0 + \beta'\mathbf{Z} + h(clr(\mathbf{X_i})) + \epsilon_i \tag{3.12a}$$

$$logit(Y_i) = \beta_0 + \beta'\mathbf{Z} + h(clr(\mathbf{X_i})) \tag{3.12b}$$

The association between the microbiome and the response variable is tested through the following hypothesis

$$\begin{aligned} H_0\colon &\quad h(clr(\mathbf{X})) = 0 \\ H_1\colon &\quad h(clr(\mathbf{X})) \neq 0 \end{aligned} \tag{3.13}$$

The non-parametric component $h(\cdot)$ is assumed to follow a normal distribution $N(0, \tau\mathbf{K}_A)$, with τ a real value and $\mathbf{K}_A$ the Kernel matrix that, similarly to Equation 3.8, is given by

$$\mathbf{K_A} = -\frac{1}{2}\left(\mathbf{I} - \frac{\mathbf{11}^T}{n}\right)\mathbf{D_A}^2\left(\mathbf{I} - \frac{\mathbf{11}^T}{n}\right) \tag{3.14}$$

where $\mathbf{D}_A$ denotes Aitchison distance matrix.

3.4.1 Global association between microbiome and HIV infection with MiRKAT-CoDA

We can implement MiRKAT-CoDA by making use of the available functions in MiRKAT package with some previous transformations of the microbiome data. The process is described in the following steps:

1. **Replacement of zeros**: we use the Geometric Bayesian Multiplicative (GBM) replacement (Martín-Fernández et al., 2015) to avoid zeros in the microbiome abundance matrix $\mathbf{X}$, replacing them by a small value obtained with a proper Bayesian imputation.

2. **Compute Aitchison distance matrix**: once the count matrix does not contain zeros, the `clr()` function of {`compositions`}

package is used to compute the centered log-ratio transformed values. Then, Euclidean distance is calculated for this *clr*-transformed scores, thus getting Aitchison distance matrix.

3. **Obtain the Kernel matrix**: we use `D2K()` function available in the {`MiRKAT`} package. It transforms the distance matrix $\mathbf{D_A}$ into a Kernel matrix $\mathbf{K_A}$ according to Equation 3.14.

4. **Implement Kernel machine regression**: we run `MiRKAT()` function, available in {`MiRKAT`} package where we specify the response variable, the Kernel matrix and the covariate adjustment.

The output of this process is a p-value for the null hypothesis of no association between $\mathbf{X}$ and $\mathbf{Y}$ adjusted by $\mathbf{Z}$. Here we illustrate the use of MiRKAT-CoDA for exploring the global association between microbiome abundance and HIV infection for the IrsiCaixa HIV study (Noguera-Julian et al., 2016). We evaluate the association between microbiome composition and HIV-Status (HIV+ or HIV-) using the logistic Kernel machine regression given in Equation 3.12b, where the response variable $\mathbf{Y}$ is HIV-Status, $\mathbf{X}$ contains microbiome abundances at genus level, and $\mathbf{Z}$ are the adjustment covariates. As concluded by Noguera et al., sexual practice is a possible confounding variable when studying the microbiome composition in an HIV study. We implement the analysis with and without adjustment by including sexual practice in $\mathbf{Z}$ as a dichotomous variable defined as `MSM` for those men who have sex with men, and `non-MSM` for the rest of individuals. The result of running this algorithm to the HIV-study, is a p-value of 0.00215 if we do not adjust by MSM, and a p-value of 0.084 after adjusting for sexual practice. Though near the usual significance level of 0.05, the result is inconclusive and does not allow to confirm a significant association between microbiome composition and HIV-Status.

3.5 Weighted Kernel machine regression for compositional data

Kernel machine regression as implemented in the previous section provides a global answer about the relationship between the response variable and microbiome composition as a whole. If this global test results significant, it is reasonable to think that not all taxa but only few of them are responsible of the association. In this sense, we present a new proposal in order to rank each variable according to its contribution to the global association with the phenotype. The method uses Kernel machine regression together with the recently defined *weighted Aitchison distance* (Egozcue and Pawlowsky-Glahn, 2016) and provides a score for each variable measuring its contribution to the microbiome-outcome association. Furthermore, the method can also be applied to a particular group of variables instead of to just one component.

The weighted Aitchison distance is the generalization of Aitchison distance where each component is weighted according to a vector $\mathbf{w}$. Given two compositions denoted by $\mathbf{x_1} = (x_{11}, \ldots, x_{1k})$ and $\mathbf{x_2} = (x_{21}, \ldots, x_{2k})$, and a vector of weights $\mathbf{w} = (w_1, \ldots, w_k)$, the weighted Aitchison distance

(Egozcue and Pawlowsky-Glahn, 2016) between $\mathbf{x_1}$ and $\mathbf{x_2}$ is defined as follows

$$d_w(\mathbf{x_1}, \mathbf{x_2}) = \sqrt{\sum_{i=1}^{k} w_i \left(\log \frac{y_{1i}}{g_{\mathbf{w}}(\mathbf{y_1})} - \log \frac{y_{2i}}{g_{\mathbf{w}}(\mathbf{y_2})} \right)^2} \tag{3.15}$$

where $\mathbf{y_1}$ and $\mathbf{y_2}$ are the initial compositions $\mathbf{x_1}$ and $\mathbf{x_2}$ divided by $\mathbf{w}$

$$\mathbf{y_i} = \frac{\mathbf{x_i}}{\mathbf{w}} = \left(\frac{x_{i1}}{w_1}, \ldots, \frac{x_{ik}}{w_k} \right), \tag{3.16}$$

and $g_{\mathbf{w}}(\cdot)$ denotes the weighted geometric mean as defined in Equation 3.17, where the $s_{\mathbf{w}} = \sum\limits_{i=1}^{k} w_i$ is the total sum of the weights

$$g_{\mathbf{w}}(\mathbf{y}) = \exp\left(\frac{1}{s_{\mathbf{w}}} \sum_{i=1}^{k} w_i \log(y_i) \right). \tag{3.17}$$

Each vector of weights $\mathbf{w}$ provides a different weighted Aitchison distance. In order to evaluate the contribution of a particular taxon (or a group of taxa) we explore how the global association between the microbiome and the response variable changes as we modify the weight of the taxon of interest. If a particular taxon is important in the association, as we increase its weight we expect the global association to also improve and get a more significant result of the Kernel machine regression, that is, a smaller p-value. Based on this idea, we consider a sequence of weights and we analyse how the global p-value changes as different weights are considered. As described below, we summarize the contribution of a taxon to the global association as the slope of the linear regression model between the different weights and $-\log(p)$. This proposal named *weighted MiRKAT-CoDA* is implemented as an algorithm defined by the following steps:

1. **Zero replacement**: a replacement of zeros is required before compositional data analysis can be performed. Although we consider the GBM replacement as the default option, there are other alternatives like adding one count to all the cells in the abundance matrix.

2. **MiRKAT-CoDA with weights**: given a composition $(X_1, \ldots, X_k)$, the analysis of the contribution of component X_i in the global association is implemented by considering a sequence of weights $S = \{s_1, \ldots, s_q\}$. Thus, for each component X_i , $i \in \{1, \ldots, k\}$ and for each particular value $s_r \in S$, the components of $\mathbf{w} = (w_1, \ldots, w_k)$ are defined as:
$$\begin{cases} w_j = 1 & \forall j \neq i \\ w_i = s_r \ , s_r \in S \end{cases}$$
This idea can also be extended for measuring the contribution of a set of features indexed by I, defining the vector $\mathbf{w}$ as
$$\begin{cases} w_j = 1 & \forall j \notin I \\ w_l = s_r \ , \forall l \in I, s_r \in S \end{cases}$$
Using the response vector $\mathbf{Y}$ and the weighted Aitchison distance matrix $\mathbf{D}_w$ for a particular vector of weights $\mathbf{w}$, we implement `MiRKAT()` function, which returns a p-value measuring the significance of the association between the response and the microbiome. Once the kernel regression model has been run for each combination of taxon and weight value, a table similar to Table 3.1 summarizes the results:

3. **Linear regression**: the contribution of each taxon is summarized by the slope of the linear regression model between the weights S

Weight	**Taxon 1**	**Taxon 2**	...	**Taxon k**
$w = s_1$	p_{11}	p_{12}	$\cdots$	p_{1k}
$w = s_2$	p_{21}	p_{22}	$\cdots$	p_{2k}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$w = s_q$	p_{q1}	p_{q2}	$\cdots$	p_{qk}

Table 3.1: *Structure of p-values obtained after the evaluation of the contribution of each taxon for each weight.*

and minus the logarithm of the p-values obtained for each different weight. The larger the slope, the larger is the contribution of the variable to the global association.

4. **Slope ranking**: once the contribution of each taxon has been estimated they can be ranked in a decreasing order so that the most important features appear on the top of the list.

3.5.1 Weighted MiRKAT-CoDA for microbiome-HIV association

We use weighted MiRKAT-CoDA algorithm to the IrsiCaixa HIV study in order to measure the individual contribution of each feature to the global microbiome-HIV association. Though this global association is not significant, we use this data to illustrate weighted MiRKAT-CoDA algorithm. Using the default set of weights $S = (.1, .4, .7, 1, 2, 3, 4, 5)$, we obtain the Table 3.2, where each p-value measures the association between microbiome and HIV-Status when the weight in the corresponding row is assigned to the taxon in the corresponding column. A linear regression model is implemented for each taxon being S the explanatory variable, and minus log-transformed p-values the response. Figure 3.3 presents the result for two different taxa: *g_Prevotella* presenting a decrease of the response variable as the weight is increased, and *g_Bacteroides*, with a positive relationship between the weight and the significance. From these results we can infer that *g_Prevotella* is not relevant to the global microbiome-HIV association since when we increase its weight the significance of the association decreases (negative slope). Instead, *g_Bacteroides* is shown to be important in the global microbiome-HIV association since the significance increases as we increase its weight (positive slope).

Finally, we represent in Figure 3.4 the top 10 taxa that most contribute to the joint microbiome-HIV association.

3.6 Discussion

There are different methods to answer if there is an association between microbial composition and an outcome of interest. In this chapter we describe the Kernel machine regression (KMR) as a distance-based method to answer this question. Defined for microbiome studies, `MiRKAT` R package implements KMR methodology for different distances. As none of them is

Weight	**g_Prevotella**	**g_Bacteroides**	...	**g_Solobacterium**
$w = 0.1$	0.076	0.120	...	0.077
...	...	...	...	...
$w = 1$	0.084	0.084	...	0.084
...	...	...	...	...
$w = 5$	0.146	0.027	...	0.125

Table 3.2: *Matrix with the p-values for the corresponding taxon (column) obtained after using the weight indicated in the Weight column.*

subcompositionally dominant, we introduce an adaptation of the method including Aitchison distance as the measure used for building the Kernel matrix in KMR. Because of its subcompositional dominance, we use this distance in order to avoid the possible incoherences resulted with some others.

Moreover, we define another new method based on KMR which uses the weighted Aitchison distance. Modifying the weight of the different taxa they can be ranked according to their importance in the global association with the outcome. This approach, although useful for pointing the most important microorganisms, presents some limitations. The first limitation is that the set of weights S considered may result uninformative. Variability in the p-values is required to compute the contribution of each taxon, so weights should be defined to ensure this fluctuation. The second limitation is that if the reference global p-value (when no weights are considered) is very small or zero, the weighting method is not very informative since changing the weight of just one taxon can hardly modify the global p-value, remaining equal to zero in most cases.

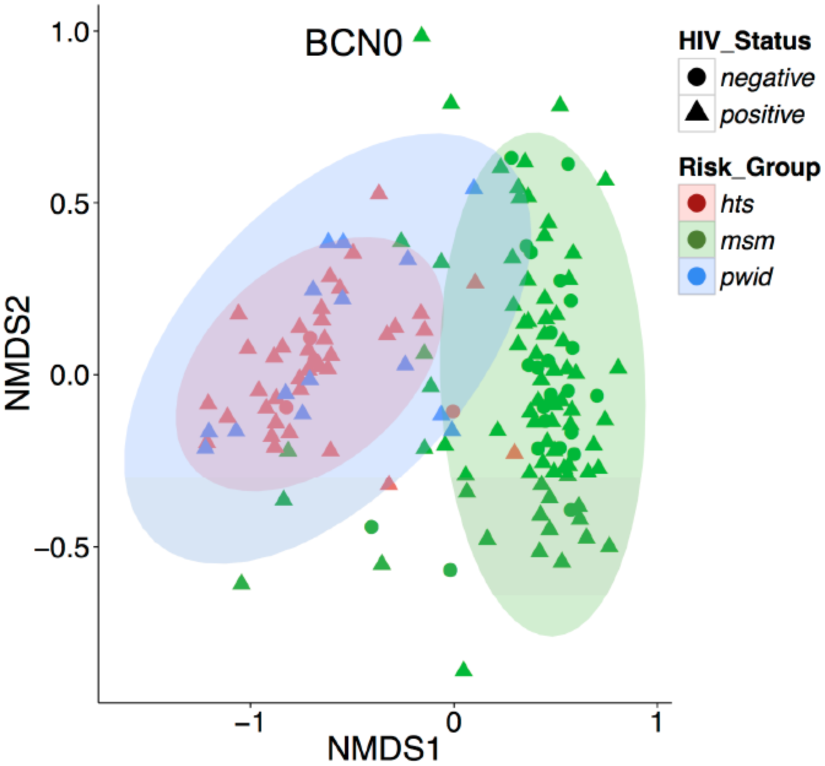


Figure 3.1: *Non-metric multidimensional scalling (NMDS) representation of an HIV cohort labelling samples according to their HIV-Status and Risk group.*

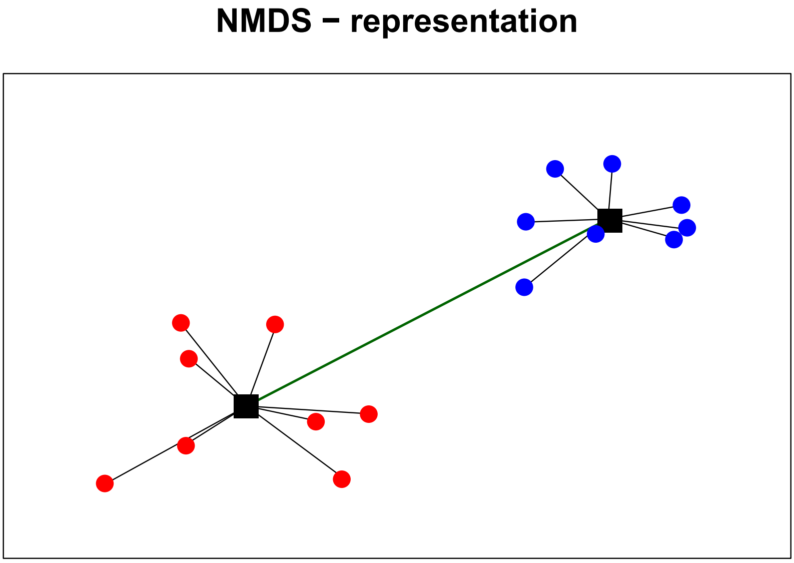


Figure 3.2: *NMDS-representation of samples of two pre-defined groups. Squares denote their centroids, with which they are linked.*

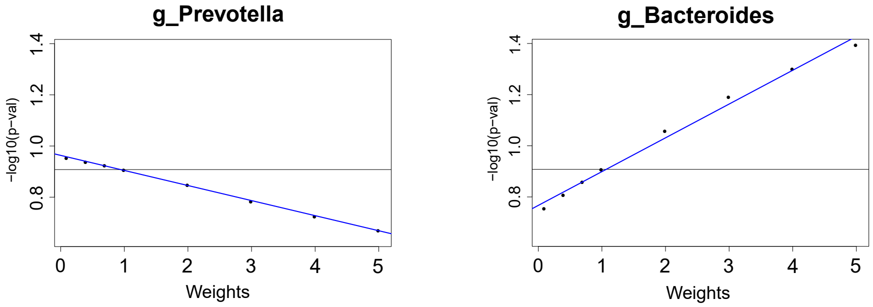


Figure 3.3: *Slopes for two different bacteria. g_Prevotella with a negative slope and g_Bacteroides with a positive slope.*

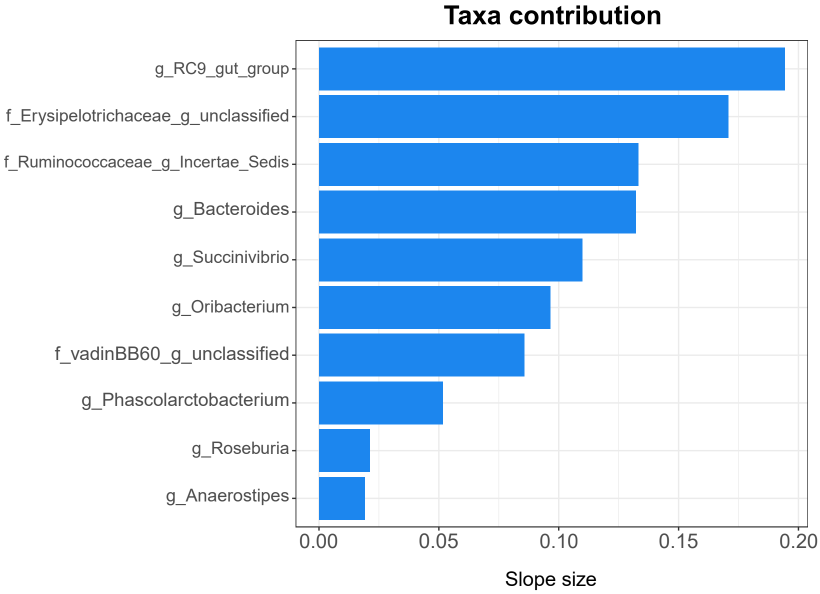


Figure 3.4: *Top 10 taxa contribution in the microbiome - HIV association.*

CHAPTER 4

Identification of microbial signatures with selbal

Most of the methods proposed for microbiome analysis are intended to answer two main questions: first, whether there is a global association between the microbiome and a phenotype of interest, and second, which specific taxa are associated with the outcome. Multivariate methods like PERMANOVA (Anderson, 2001), implemented in the `adonis()` function of the R {`vegan`} package (Oksanen, 2015), or MiRKAT (Zhao et al., 2015) answer the first question. The second one is approached with univariate methods where each taxa is tested for association with the outcome. When the response variable is dichotomous, this is known as differential abundance testing and methods specifically developed for RNA-Seq data, such as `DESeq2` (Love et al., 2014) and `edgeR` (Robinson et al., 2009), are commonly used. Other methods have also been proposed, such as `ANCOM` (Mandal et al., 2015) and `ALDEX` (Fernandes et al., 2013), that acknowledge the compositionality of microbiome data.

In this chapter, we focus on a different question. Here, we propose a methodology whose goal is the identification of microbial signatures, that is, groups of microbial taxa that are predictive of a phenotype of interest. These microbial signatures can be used for diagnosis, prognosis or prediction of therapeutic response based on an individual's specific microbiota (Knight et al., 2018). The identification of microbial signatures involves both, modeling and variable selection: modeling the response variable and identifying the smallest number of taxa with the highest prediction or classification accuracy. In this context, we propose `selbal`, a model selection procedure that searches a sparse model that adequately explains the response variable of interest. Similarly to forward stepwise linear regression, `selbal` performs multiple regressions a number of times, adding a new taxon to the model at every step. Unlike linear regression, the raw variables in `selbal` are not included in a linear equation, but as part of what is called a *balance* in compositional data analysis literature. The method proposed in this chapter has already been published and is available in Appendix D (Rivera-Pinto et al., 2018).

Introduced in Chapter 2 with the general name of an *ilr*-coordinate, we redefine the concept of balance. Let $\mathbf{X} = (X_1, \cdots, X_k)$ be a composition with k parts. Given two disjoint subsets of components in $\mathbf{X}$, denoted by $\mathbf{X}_+$ and $\mathbf{X}_-$, indexed by I_+ and I_-, and composed by k_+ and k_- parts, respectively; the balance between $\mathbf{X}_+$ and $\mathbf{X}_-$ is defined as the normalized log-ratio of the geometric mean of the two groups of components:

$$\mathbf{B}(\mathbf{X}_+, \mathbf{X}_-) = \sqrt{\frac{k_+ \cdot k_-}{k_+ + k_-}} \log \frac{\left(\prod_{i \in I_+} X_i\right)^{1/k_+}}{\left(\prod_{j \in I_-} X_j\right)^{1/k_-}} \tag{4.1}$$

Equation 4.1 can be expanded as in Equation 4.2 for easier interpretation. This formulation expresses a balance as a value proportional to the difference between the means of the log-transformed variables of the two groups of components.

$$\mathbf{B}(\mathbf{X}_+, \mathbf{X}_-) \propto \frac{1}{k_+} \sum_{i \in I_+} \log X_i - \frac{1}{k_-} \sum_{j \in I_-} \log X_j \tag{4.2}$$

With the expression in Equation 4.2, it is clear that a compositional balance is a particular case of a log-contrast, defined as a linear combination of the log-transformed components of a composition with the restriction that the coefficients of the linear function add-up to zero (Pawlowsky-Glahn et al., 2015). The importance of working with balances or in general with log-contrasts, when analyzing compositional data is that this kind of functions preserve scale invariance, one of the principles that should be fulfilled in CoDA (Aitchison, 1986; Pawlowsky-Glahn et al., 2015).

`selbal`, the algorithm we propose for balance selection, starts with a thorough search of the two taxa whose balance, or log-ratio, is most associated with the response. Once the first two-taxa balance is selected, the algorithm performs a forward selection process where, at each step, a new taxon is added to the existing balance, so that the specified optimization criterion is improved (*Area Under the ROC Curve (AUC)* for dichotomous responses or *Mean Squared Error (MSE)* for continuous outcomes). The algorithm stops when there is no additional variable that improves the current optimization parameter, or when the maximum number of components to include in the balance is achieved. This number is established with a cross-validation procedure, which is also used to explore the robustness of the identified balance.

The idea of model selection for microbial signature identification can also be performed in two separate steps: first, variable selection, and next, model building with the selected variables. When the outcome variable is dichotomous, variable selection can be obtained with differential abundance testing methods such as `DESeq2` (Love et al., 2014), `edgeR` (Robinson et al., 2009), or, in the context of compositional data analysis, `ANCOM` (Mandal et al., 2015) or `ALDEx2` (Fernandes et al., 2013). However, it is not clear how to combine the selected variables to obtain the best joint sparse model. This is specially challenging for microbiome analysis, where the compositional nature of microbiome data induces spurious correlations among the variables (Gloor et al., 2016). We think that a joint procedure that involves both modelling and variable selection, as performed in `selbal`, is more appropriate in this context.

Other authors (Silverman et al., 2017; Washburne et al., 2017; Morton et al., 2017) have previously proposed the use of balances for microbiome analysis regarding the construction of an isometric log-ratio (*ilr*) transformation (Egozcue et al., 2003), that allows compositional data to be represented in a real Euclidean space where standard statistical methods can be applied. Silvermann et al. (2017) and Washburne et al. (2017) propose methods that use microbial phylogenetic information to guide the sequential binary partition in the construction of a particular *ilr*-transformation. This phylogenetically driven *ilr*-transformation would help to detect relevant evolutionary factors or phylogenetically associated bacterial groups related to host-microbiome interactions (Silverman et al., 2017; Washburne et al., 2017). In the method proposed by Morton et al. (2016), instead of using phylogenetic information, they use the response variable to define the binary sequential partitions of the *ilr*-transformation (Morton et al., 2017). `selbal` is different from these methods: first, in `selbal` only one balance is considered and not a sequence of balances, and second, the purpose of the selected balance is classification or prediction and not a new representation of the data.

4.1 SELBAL ALGORITHM

The main goal of `selbal` algorithm is to find the best balance for prediction of a variable of interest. So, given a numeric or dichotomous response variable $\mathbf{Y}$, a composition $\mathbf{X} = (X_1, \dots, X_k)$, and additional covariates $\mathbf{Z} = (Z_1, \dots, Z_r)$, the goal of the algorithm is to determine two subcompositions of $\mathbf{X}$, $\mathbf{X}_+$ and $\mathbf{X}_-$, whose balance $\mathbf{B}(\mathbf{X}_+, \mathbf{X}_-)$ is highly associated with $\mathbf{Y}$ after adjustment for covariates $\mathbf{Z}$. Depending on the nature of the dependent variable, the association can be defined in several ways. For a continuous response, the optimization criterion is the minimization of the mean squared error (MSE) of the linear regression model defined in Equation 4.3.

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{B}(\mathbf{X}_+, \mathbf{X}_-) + \gamma' \mathbf{Z} \tag{4.3}$$

Similarly, for a dichotomous outcome $\mathbf{Y}$, we consider the logistic regression model in Equation 4.4, and three possible optimization criteria corresponding to the maximization of the area under the ROC curve, the maximization of the explained variance (Mittlböck and Schemper, 1996), or the discrimination coefficient (Tjur, 2009).

$$logit(\mathbf{Y}) = \beta_0 + \beta_1 \mathbf{B}(\mathbf{X}_+, \mathbf{X}_-) + \gamma' \mathbf{Z} \tag{4.4}$$

The search of the optimal balance is hard because there are multiple possible candidates. As we increase the number of components k, the number of possibilities increases exponentially. For instance, for $k = 10$, we need to consider 57002 possible balances, whereas working in a common range of between 50 and 100 features, to evaluate all the balances implies the analysis of between $7.17 \cdot 10^{23}$ and $5.15 \cdot 10^{47}$ alternatives. The evaluation of such a huge quantity of balances is infeasible in terms of computational time. As an alternative, we propose a greedy-forward selection algorithm defined by the following steps:

1. **Zero replacement**: compositional techniques are defined for vectors with strictly positive values. As microbiome datasets present zeros, they require a modification in order to use CoDA. There are several alternatives to solve the zero problem. The default option we consider

uses the Geometric Bayesian Multiplicative (GBM) replacement (Martín-Fernández et al., 2015) implemented in the `cmuultRepl()` function of the {`zCompositions`} R package (Palarea-Albaladejo and Martín-Fernández, 2015). We also provide the option of adding one count to all values in the dataset in order to avoid zeros, which is a very extended practice.

2. **Optimal balance between two components**: once zeros have been replaced in $\mathbf{X}$, in the next step the algorithm evaluates exhaustively all the possible balances composed by only two components; that is, all balances of the form:

$$\mathbf{B_{ij}} = \mathbf{B}(X_i, X_j) = \sqrt{\frac{1}{2}}\big(log(X_i) - log(X_j)\big) \quad \text{for } i, j \in \{1, \cdots, k\},\ i \neq j \tag{4.5}$$

Each two component balance $\mathbf{B_{ij}}$ is tested for association with the response variable $\mathbf{Y}$ with the linear regression model in Equation 4.6a if $\mathbf{Y}$ is continuous, and the logistic regression model in Equation 4.6b if $\mathbf{Y}$ is dichotomous.

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{B_{ij}} + \gamma' \mathbf{Z} \tag{4.6a}$$

$$logit(\mathbf{Y}) = \beta_0 + \beta_1 \mathbf{B_{ij}} + \gamma' \mathbf{Z} \tag{4.6b}$$

The balance that maximizes the optimization criteria (MSE or AUC) is selected and denoted by $\mathbf{B^{(1)}}$.

It is important to remark that $\mathbf{B}(X_i, X_j)$ and $\mathbf{B}(X_j, X_i)$ only differ in their sign and they present the same association value with the response variable. The algorithm returns the one whose regression coefficient β_1 is positive.

3. **Optimal balance adding a new component**: for $s > 1$ and until the stop criterion is fulfilled, let $\mathbf{B^{(s-1)}}$ be the balance defined in the previous step $(s-1)$ given by:

$$\mathbf{B^{(s-1)}} \propto \frac{1}{k_+^{(s-1)}} \sum_{i \in I_+^{(s-1)}} \log(X_i) - \frac{1}{k_-^{(s-1)}} \sum_{j \in I_-^{(s-1)}} \log(X_j) \tag{4.7}$$

Equation 4.7 is defined by $I_+^{(s-1)}$ and $I_-^{(s-1)}$, which are two disjoint subsets of indices in $\{1, \cdots, k\}$ with $k_+^{(s-1)}$ and $k_-^{(s-1)}$ elements, respectively.

For each of the remaining variables, X_p not yet included in the balance, $p \notin \left(I_+^{(s-1)} \cup I_-^{(s-1)}\right)$, the algorithm considers the balance that is obtained by adding $\log(X_p)$ to the positive part of $\mathbf{B^{(s-1)}}$ (Equation 4.8a), or to its negative part (Equation 4.8b):

$$\mathbf{B_p^{(s+)}} \propto \frac{1}{k_+^{(s-1)}+1} \Big(\sum_{i \in I_+^{(s-1)}} log(X_i) + \log(X_p) \Big) - \frac{1}{k_-^{(s-1)}} \sum_{j \in I_-^{(s-1)}} \log(X_j) \tag{4.8a}$$

$$\mathbf{B_p^{(s-)}} \propto \frac{1}{k_+^{(s-1)}} \sum_{i \in I_+^{(s-1)}} log(X_i) - \frac{1}{k_-^{(s-1)}+1} \Big(\sum_{j \in I_-^{(s-1)}} \log(X_j) + \log(X_p) \Big) \tag{4.8b}$$

Each of these pairs of balances, $\mathbf{B_p^{(s+)}}$ and $\mathbf{B_p^{(s-)}}$, for each of the remaining variables, X_p, is tested for association with the response variable through one of the regression models in Equation 4.9a and Equation 4.9b , where **B** represents the tested balance.

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{B} + \gamma' \mathbf{Z} \tag{4.9a}$$

$$logit(\mathbf{Y}) = \beta_0 + \beta_1 \mathbf{B} + \gamma' \mathbf{Z} \tag{4.9b}$$

The balance that maximizes the optimization criterion, defines the new balance $\mathbf{B^{(s)}}$ for the s-th step.

4. **STOP criterion**: there are two stopping rules: the iterative algorithm stops when the improvement of the optimization parameter is lower than a specified threshold (default equal to 0) or, when the specified maximum number of components has been included in the balance (default equal to 20).

These previous steps define the main function for the search of a microbial signature. In addition, we implement an iterative cross-validation (CV) procedure with two goals: (a) to identify the optimal number of components to include in the balance, and (b) to explore the robustness of the global balance identified with the whole dataset.

Cross-validation

Let T be the number of iterations (default $T = 10$), F the number of folds in the cross-validation (default $F = 5$), and C the maximum number of variables or components included in a balance (default $C = 20$). At each iteration $t \in \{1, \cdots, T\}$, the data is divided into F folds $\{D_1^t, \cdots, D_F^t\}$.

Then, for each $f \in \{1, \cdots, F\}$ the main algorithm of `selbal` is applied to the training dataset $\bigcup_{j \neq f} D_j^t$, and the optimal balance with C variables is obtained, $B_f^t(C)$. Actually, since algorithm is a forward selection process where variables are included sequentially at each step, we have a sequence of balances including from 2 to C variables: $B_f^t(2),\ B_f^t(3), \cdots,\ B_f^t(C)$.

The classification accuracy (MSE or AUC) of these balances is measured on the test dataset, D_f^t, giving a sequence of accuracy measures for each number of variables included in the balance. For each iteration-fold pair of values, if the response variable is continuous, a set of values as in Equation 4.10a is obtained, and a sequence like in Equation 4.10b when $\mathbf{Y}$ is dichotomous. We have denoted it as AUC_f^t, but analogous scores are obtained for other accuracy measures.

$$MSE_f^t(2),\ MSE_f^t(3),\ \cdots,\ MSE_f^t(C) \tag{4.10a}$$

$$AUC_f^t(2),\ AUC_f^t(3),\ \cdots,\ AUC_f^t(C) \tag{4.10b}$$

For each number of components $c \in \{2, \cdots, C\}$ we have $F \times T$ measures of accuracy. The mean and the standard error are computed an represented in a plot, as shown in Figure 4.1 or Figure 4.4.

Similarly to the cross-validation process in LASSO for finding the optimal penalization parameter *lambda* (Hastie and Tibshirani..., 2009), we follow the *1se strategy* and define the optimal number of variables (k_{opt}) to be included in the balance as the lowest number whose mean MSE is within 1 standard error of the minimum mean MSE (or whose mean AUC is within 1 standard error of the maximum mean AUC). Usually, the *1se strategy* provides sparser models than taking the minimum mean MSE (or maximum AUC), with very similar accuracy. This *1se strategy* is the default option in `selbal`, but there is also the possibility of determining the optimal number of variables as the value reaching the optimum (minimum mean MSE or maximum mean AUC).

Once the optimal number of variables k_{opt} has been determined, we obtain the *global balance*, that is, we apply the main algorithm to the whole dataset $\mathbf{X}$ with the specification that the maximum number of components in the balance is k_{opt}. We use the cross-validation results to explore the robustness of the global balance. We retrieve all the balances with k_{opt} components obtained in the cross-validation process $B^t_f(k_{opt}),\ f \in \{1, \cdots, F\}, t \in \{1, \cdots, T\}$ and compare them with the global balance. We summarize these cross-validation balances in two different ways, per balance and per variable. We provide the relative frequency of the different balances obtained in the CV process and the proportion of times that each taxon has been included into a balance. This information is summarized in a table as shown in Figure 4.3 or Figure 4.6.

The cross-validation process is also used to obtain the cross-validation accuracy, defined as the mean MSE or mean AUC of the balances obtained in the CV process that have the same number of variables as the global balance, that is, $\text{mean}_{t,f}\left(MSE_f^t(k_{opt})\right)$ or $\text{mean}_{t,f}\left(AUC_f^t(k_{opt})\right)$.

All this methodology including the search of the optimal balance and the cross-validation process for measuring the robustness of the result is included in an R package named {selbal} which can be found in https://github.com/UVic-omics/selbal.

4.2 Applications

We illustrate how {selbal} package can be useful for obtaining microbial signatures associated either with continuous or dichotomous variables. We first consider a Crohn's disease study where the goal is to to find a microbiome biomarker able to differentiate patients with and without the disease. The second analysis is focused on the search of a balance linked with a continuous inflammation marker of interest in HIV infection.

4.2.1 Microbiome and Crohn's disease

Crohn's disease (CD) is an inflammatory bowel disease that has been linked to microbial alterations in the gut (Ren et al., 2015; Øyri et al., 2015) . We use data from a large pediatric CD cohort study (Ren et al., 2015) to illustrate the proposed methodology for the identification of a microbial signature. Microbiome data from 16S rRNA gene sequencing and QIIME 1.7.0 bioinformatics processing are downloaded from Qiita (https://qiita.ucsd.edu, study identifier [ID]: 1939). Only patients with Crohn's disease ($n = 662$) and those without any symptom ($n = 313$) are included in the analysis. Agglomerating OTUs to genus level, it results a matrix with 48 genera and 975 samples.

The goal of the analysis with `selbal` is to identify a microbial signature for Crohn's disease which discriminates between CD and non CD individuals.

We first run a cross-validation process (through `selbal.cv()` function included in {`selbal`}) in order to determine the optimal number of taxa to consider in the balance. Figure 4.1 provides the mean AUC and standard error of the balances obtained in the CV process as a function of the number of taxa. In this case, and following the 1se rule, the optimal number of components is twelve. Once the number of taxa is determined, the main function `selbal()` is applied to the whole dataset specifying that the number of desired taxa to include is twelve. Thus, the global balance is defined by

$$\begin{aligned}
\mathbf{X}_{+} = \{&g\_Roseburia,\ o\_Clostridiales\_g\_,\ g\_Bacteroides,\\
&f\_Peptostreptococcaceae\_g\_\}\\
\mathbf{X}_{-} = \{&g\_Dialister,\ g\_Dorea,\ o\_Lactobacillales\_g\_,\ g\_Eggerthella,\\
&g\_Adlercreutzia,\ g\_Streptococcus,\ g\_Oscillospira\}.
\end{aligned}$$

Figure 4.2 describes the distribution of the values for the microbial signature both for CD and non CD individuals. Cases with Crohn's disease present lower balance scores than controls, which means lower relative abundances of taxa in group $\mathbf{X}_+$ with respect to taxa in group $\mathbf{X}_-$. Despite differences in the interpretation of the results, we highlight that *Bacteroides* and *Clostridiales* have been previously identified as less abundant in cases than in controls (Ren et al., 2015).

The discrimination value of the identified balance is important, with an apparent AUC of 0.838. However, this apparent AUC is known to overestimate the discrimination value of the microbial signature, since it has been measured on the same dataset that was used to build the model. A more accurate estimation is obtained from the CV process that provides a cv-AUC of 0.819, which is also a very good discrimination value.

Cross-validation can also be helpful to assess the robustness of the proposed global balance. In Figure 4.3 we summarize the different balances with twelve taxa obtained in the CV process. On one hand, we have the frequency of the different CV balances and, on the other, the frequency of selection of each taxon. Rows represent the most frequent taxa with their percentage of selection given in the second column; the third column represents the global balance, and the last three columns show the three most frequent balances selected in the CV procedure. Colored rectangles indicate whether the taxon is in the numerator of the balance (*red*), in the denominator (*blue*) or not included (*white*). The last row provides the proportion of times each balance has been selected as optimal in the CV procedure. From Figure 4.3 it follows that the identified global balance for Crohn's disease is very robust: it coincides with the most frequently selected one in the CV process, which turns out to be the optimal a 36% of times. Moreover, the taxa which form the global balance are also those most frequently selected in the CV procedure.

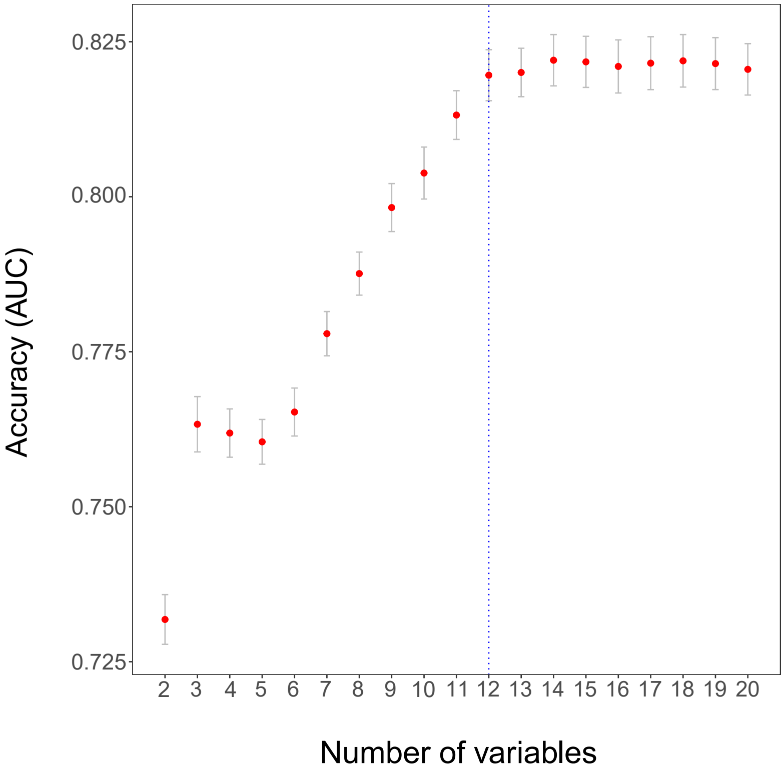


Figure 4.1: *AUC values for different number of components in the balance for the association of Crohn's disease status with proposed balances.*

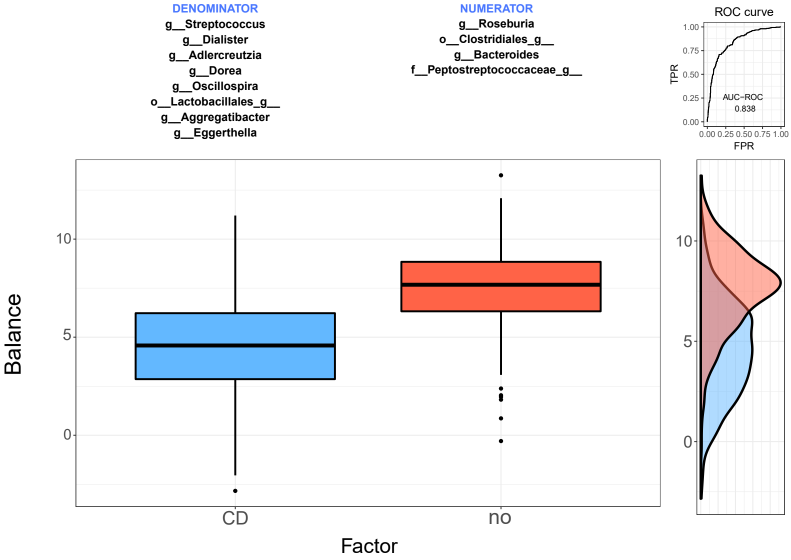
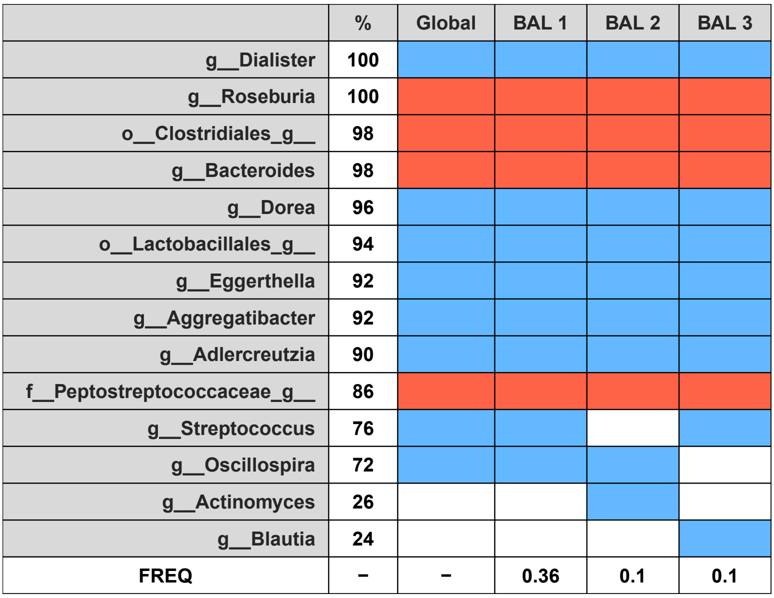


Figure 4.2: *Boxplot for the scores of the proposed balance in order to differentiate samples from patients with Crohn's disease (CD) and patients without the disease (no).*

4.2.2 Microbiome and sCD14 inflammation marker

Acute and chronic inflammation typically occur after HIV infection. Even patients under antiretroviral medications and undetectable viral load present chronic inflammation, which may cause tissue damage and is associated with many chronic diseases (Brenchley et al., 2006). In this context, there is a great interest in defining possible interventions involving modifications of the gut bacterial environment, which may reduce inflam-



	%	Global	BAL 1	BAL 2	BAL 3
g__Dialister	100	🟦	🟦	🟦	🟦
g__Roseburia	100	🟥	🟥	🟥	🟥
o__Clostridiales_g__	98	🟥	🟥	🟥	🟥
g__Bacteroides	98	🟥	🟥	🟥	🟥
g__Dorea	96	🟦	🟦	🟦	🟦
o__Lactobacillales_g__	94	🟦	🟦	🟦	🟦
g__Eggerthella	92	🟦	🟦	🟦	🟦
g__Aggregatibacter	92	🟦	🟦	🟦	🟦
g__Adlercreutzia	90	🟦	🟦	🟦	🟦
f__Peptostreptococcaceae_g__	86	🟥	🟥	🟥	🟥
g__Streptococcus	76	🟦	🟦		🟦
g__Oscillospira	72	🟦	🟦	🟦	
g__Actinomyces	26			🟦	
g__Blautia	24				🟦
FREQ	–	–	0.36	0.1	0.1

Figure 4.3: *Table with the global balance and those most frequent in the CV-procedure for Crohn's disease study.*

mation in HIV patients (Klatt et al., 2013; D'Ettorre et al., 2015). This requires a good understanding of the association between gut microbial composition and several inflammation markers. In this case, we focus on an inmune-marker related to the chronic inflamation, the levels of soluble CD14 (sCD14), which is measured for a subset of samples ($n = 151$) of the IrsiCaixa HIV study. We apply `selbal` to search for a microbial signature that is predictive of sCD14 inflammation marker. According to the cross-validated mean squared error (cv-MSE), the optimal number of components to include in the model is four (Figure 4.4). The balance identified as the most associated with sCD14 is composed by

$$\mathbf{X}_+ = \{g\_Subdoligranulum\ ,\ f\_Lachnospiraceae\_g\_Incertae\_Sedis\}$$
$$\mathbf{X}_- = \{\ f\_Lachnospiraceae\_g\_unclassified\ ,\ g\_Collinsella\}$$

The association is moderate, with a correlation coefficient $R = 0.53$. Since sCD14 is continuous, we represent the result with a scatter plot of balance scores and sCD14 values. We observe in Figure 4.5 that higher values are associated with higher amounts of sCD14. The robustness of the selected balance can be evaluated through the results of the CV procedure. Thus, in Figure 4.6 we appreciate that the proposed global balance is also the one that has been the most selected in the CV, a 34% of the times. The four taxa defining the global balance correspond to the top 4 most frequently selected in the cross-validation, so that these results emphasize the robustness of the global balance.

Despite `selbal` does not explore the whole balance space, as it is shown in these examples, it can be a useful tool in CoDA framework for defining biomarkers in order to differentiate groups of samples and to associate the microbiome with a continuous response.

4.3 SELBAL AGAINST OTHER METHODS

Although differential abundance tests are not designed for the identification of microbial signatures, they can be adapted in order to predict a response variable. Thus, using the Crohn's disease dataset, we compare the classification accuracy of some of the most extended differential abundances tests against `selbal`. We follow a two-steps strategy: first, a variable selection, and next, model building. For the variable selection step we consider `DESeq2`, `edgeR`, `ANCOM` and `ALDEx2`. Then, a model or microbial signature is built with the selected variables. For `DESeq`, `edgeR` and `ANCOM` the model is a linear combination of the selected variables while for `ALDEx2` the model is defined as a linear combination of the selected clr-transformed variables (Fernandes et al., 2013).

`selbal` cannot be compared with these methods in terms of *false discovery rate* (FDR) and power because its goal is not to identify all taxa that are associated with the response, but to obtain the best sparse model to predict the response. So, in a cross-validation process implemented for the Crohn's disease dataset, we measure the test prediction accuracy and sparsity of the models (microbial signatures) obtained with each method. The results are given in Table 4.1, and in Figure 4.7 we can see the variability of cv-AUC for the different methods.

`selbal` and `ALDEx2` are the methods with the best classification accuracy, but `selbal` is more parsimonious, which is also a desirable feature of microbial signatures. With only 12 taxa, `selbal` obtains similar discrimination accuracy than `ALDEx2` with 31 taxa. `DESeq2` and `edgeR` provide similar results: large number of selected taxa but lower classification accuracy. This suggests that among the selected variables by `DESeq2` and `edgeR`

METHOD	Median number of taxa	Mean cv-AUC
selbal	12	0.8196
DESeq2	33	0.7752
edgeR	34	0.7721
ANCOM	5	0.7125
ALDEx2	31	0.8156

Table 4.1: *Mean cv-AUC and median of the number of taxa considered both for* selbal *and each of the differential abundance methods included in the study for classification.*

there are some false positives. ANCOM is the best in terms of parsimony, it selects the smallest number of variables with a classification accuracy comparable to DESeq2 and edgeR. This is in accordance to previous simulation studies (Weiss et al., 2017) that reports that ANCOM has very low FDR and comparable power to other methods. These results cannot be generalized since they only reflect the behaviour of the methods on one specific dataset. A more general conclusion would require a comprehensive simulation study.

4.4 Discussion

As an alternative for the differential abundance tests available in the literature and far from the increase of false positives that many of them may present working with normalized counts, we introduce `selbal`, an algorithm framed in the compositional data theory. In this context, `selbal` looks for the best balance in terms of association with a response variable of interest. The association is measured through a regression model, allowing covariate adjustment, which is interesting since some confounders may affect the analysis.

`selbal` is a useful tool for defining biomarkers both to differentiate groups of samples or to associate the microbiome with a continuous response. Comparing it with other alternatives based on the most extended differential abundance tests used in microbiology, `selbal` offers the best results including higher accuracy and lower number of taxa included in the balance. However, as a forward stepwise algorithm, it does not cover all the possible balances defined from a set of k taxa and the result could be suboptimal. Because the evaluation of all the possibilities is extremely computationally demanding, future research should be focused on the search of optimal balances through alternative approaches such as penalized regression conveniently adapted to fulfill the restrictions imposed by on the coefficients of a balance.

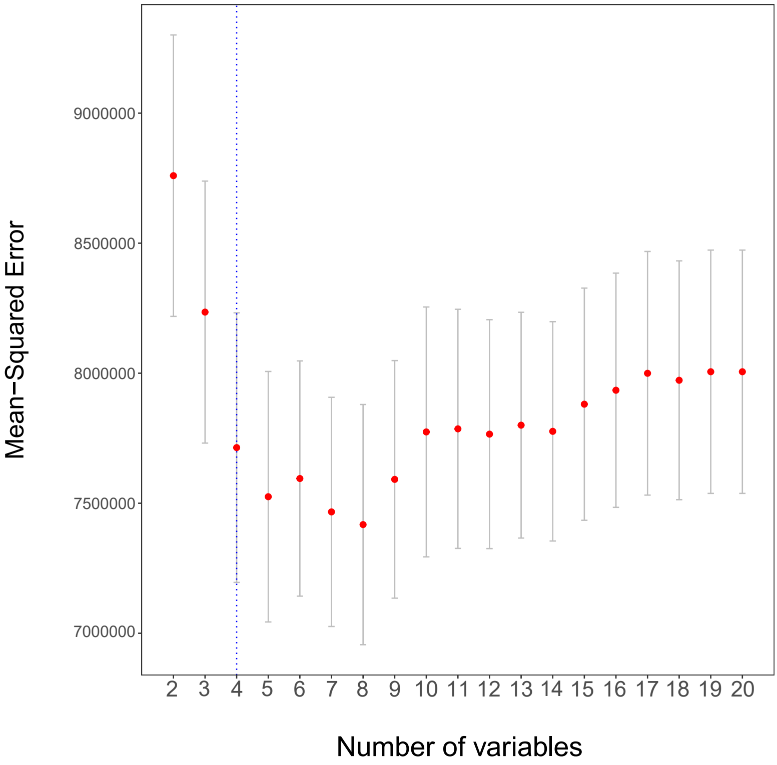


Figure 4.4: *Mean squared error for different number of components in the balance for the association of sCD14 with proposed balances*

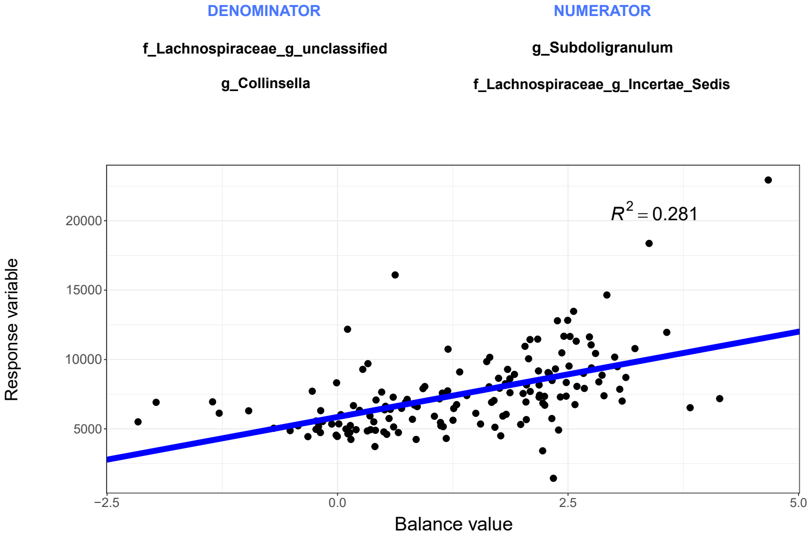
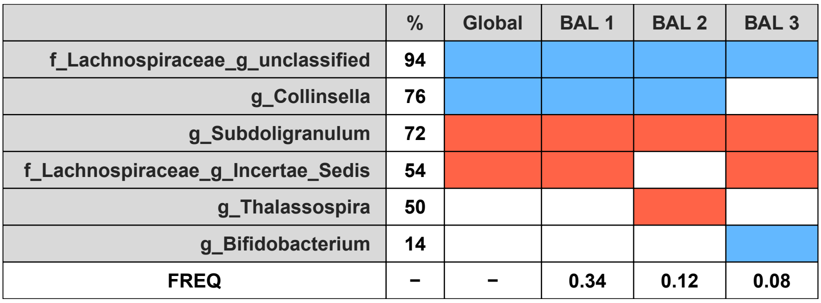


Figure 4.5: *Representation of the score for the proposed balance (X-axis) with the sCD14 values (Y-axis). Additionally, the regression line and the squared correlation coefficient.*



	%	Global	BAL 1	BAL 2	BAL 3
f_Lachnospiraceae_g_unclassified	94	🟦	🟦	🟦	🟦
g_Collinsella	76	🟦	🟦	🟦	
g_Subdoligranulum	72	🟥	🟥	🟥	🟥
f_Lachnospiraceae_g_Incertae_Sedis	54	🟥	🟥		🟥
g_Thalassospira	50			🟥	
g_Bifidobacterium	14				🟦
FREQ	–	–	0.34	0.12	0.08

Figure 4.6: *Table with the global balance and those most frequent in the CV-procedure for sCD14 inflammation marker.*

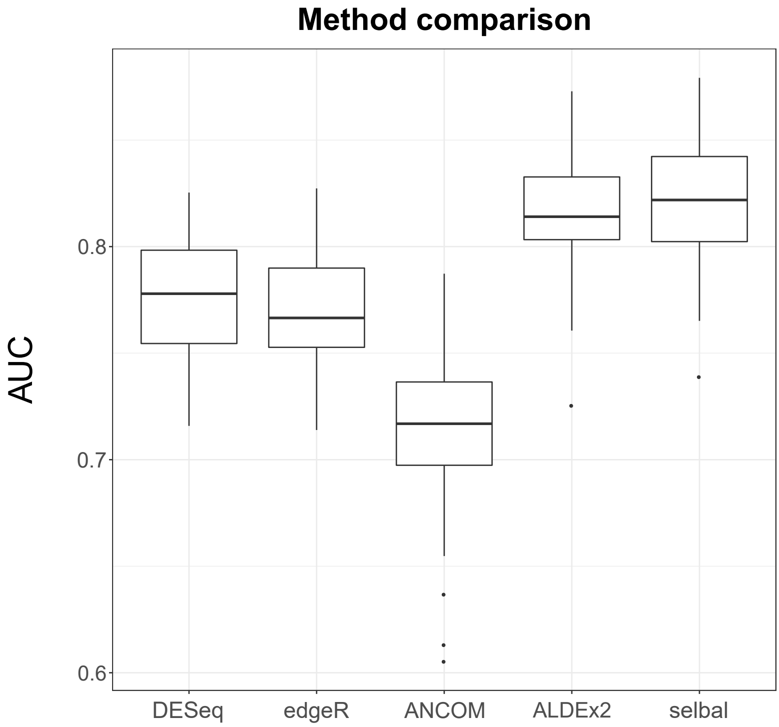


Figure 4.7: *Boxplots of the AUC resulted from the classification of patients for different methods.*

CHAPTER 5

Conclusions

What we have learned about microbiome and HIV infection

One of the main clinical problems of people living with HIV in areas with adequate healthcare standards and continued antiretroviral therapy (ART) supply is chronic inflammation related to structural or metabolic perturbations of the gut microbiota. This chronic inflammation is responsible of an increased risk of presenting non-AIDS related diseases and premature aging.

HIV-1 infection causes severe gut and systemic immune damage but its effects on the gut microbiome remain unclear. Previous results indicating a clear shift from *Bacteroides* to *Prevotella* in HIV-1 infection should be revised accounting for possible confounders, such as HIV risk factors, exercise or diet. The most evident hallmark of HIV infection on the gut microbiome is a reduction in bacterial richness. However, this is not specific of HIV infection but it is characteristic of other intestinal inflammatory diseases. The lowest bacterial richness was observed in subjects with a poor response to antiretroviral therapy.

Patients who spontaneously maintain sustained control of HIV, elite controllers (EC), have different microbiota from individuals with progressive infection and more similar to HIV negative individuals. EC have richer gut microbiota than untreated HIV patients, with unique bacterial signatures and a distinct metabolic profile. Composition and functional capacity of gut microbiota in EC may be one of the factors contributing to control of HIV-infection in absence treatment. This microbiota related control of HIV infection in EC, if confirmed, supports the search for new microbiota intervention strategies for HIV patients.

Though diet is known to have an important effect on gut microbiome composition in healthy individuals, measuring its effects on HIV infection is difficult because of the lack of extensive and reliable information at this level.

New approaches for the analysis of microbiome compositional data

Microbiome abundance data is compositional since the total counts per sample is constrained by the maximum number of sequence reads that the DNA sequencer can provide. This constraint induces strong dependencies among the abundances of the different taxa.

The use of standard statistical methods that ignore the compositional nature of microbiome data can lead to important adverse implications, such as, spurious correlations, subcompositional incoherences and the increase of type I error.

Distance-based multivariate methods, such as Kernel machine regression, are convenient for exploring patterns in microbiome data. However, most distances used in microbiome research do not fulfil the principles of compositional data analysis.

Kernel machine regression applied to a sub-compositional dominant distance, for example, Aitchison distance, provides a powerful framework for testing global associations between microbiome and a response variable of interest, while preserving from possible incoherences that may arise if non-subcompositional dominant distances are used.

Weighted Kernel machine regression, that is, Kernel machine regression applied to a weighted Aitchison distance, provides a measure of the contribution of each taxon to the joint microbiome association with the outcome. Further research is needed to define how to select the most informative set of weights in each application.

The identification of microbial signatures for diagnosis prognosis or prediction of therapeutic response is of primar interest for translating microbiome research to clinical practice. Hoewever, the decision of which taxa have to be included in the microbial signature is challenging because of the compositional nature of microbiome data.

Defining a microbiome signature as a compositional balance between two groups of taxa is innovative and preserves the principles of compositional data analysis. The search of microbial signatures with `selbal` is a powerful approach for defining biomarkers that could be used to differentiate groups of samples or to identify associations between the microbiome and a continuous response. `selbal` performs forward variable selection and, since not all possible balances are explored, the result could be suboptimal. Future research may be focused on alternative approaches to find the optimal balance.

Bibliography

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society*, 44(2):139–177.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

Beals, E. W. (1984). *Advances in Ecological Research Volume 14*, volume 14.

Brenchley, J. M., Price, D. A., Schacker, T. W., Asher, T. E., Silvestri, G., Rao, S., Kazzaz, Z., Bornstein, E., Lambotte, O., Altmann, D., Blazar, B. R., Rodriguez, B., Teixeira-Johnson, L., Landay, A., Martin, J. N., Hecht, F. M., Picker, L. J., Lederman, M. M., Deeks, S. G., and Douek, D. C. (2006). Microbial translocation is a cause of systemic immune activation in chronic HIV infection. *Nature Medicine*, 12:1365.

Chen, J., Chen, W., Zhao, N., Wu, M. C., and Schaid, D. J. (2016). Small

Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. *Genetic Epidemiology*, 40(1):5–19.

Chen, J. and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Annals of Applied Statistics*, 7(1):418–442.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1988):117–143.

Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Bandela, A. M., Cardenas, E., Garrity, G. M., and Tiedje, J. M. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Research*, 35(Database issue):D169–D172.

Davis, C. D. (2016). The gut microbiome and its role in obesity. *Nutrition Today*, 51(4):167–174.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072.

D'Ettorre, G., Ceccarelli, G., Giustini, N., Serafino, S., Calantone, N., De Girolamo, G., Bianchi, L., Bellelli, V., Ascoli-Bartoli, T., Marcellini, S., Turriziani, O., Brenchley, J. M., and Vullo, V. (2015). Probiotics

Reduce Inflammation in Antiretroviral Treated, HIV-Infected Individuals: Results of the “Probio-HIV” Clinical Trial. *PLoS ONE*, 10(9):e0137200.

Duchesne, P. and Lafaye de Micheaux, P. (2010). *Computing the Distribution of Quadratic Forms: Further Comparisons between the Liu–Tang–Zhang Approximation and Exact Methods*, volume 54.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2006). Simplicial geometry for compositional data. *Geological Society, London, Special Publications*, 264(1):145–159.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2016). Changing the reference measure in the simplex and its weighting effects. *Austrian Journal of Statistics*, 45(4):25–44.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35(3):279–300.

Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., and Gloor, G. B. (2013). ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS ONE*, 8(7).

Gianola, D. and van Kaam, J. B. C. H. M. (2008). Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. *Genetics*, 178(4):2289–2303.

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016).

It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology*, 26(5):322–329.

Hastie, T. and Tibshirani…, R. (2009). The elements of statistical learning: data mining, inference, and prediction. *books.google.com*.

Hold, G., Mukhopadhya, I., and Hansen, R. (2014). Diet, the microbiome, and IBD. *Inflammatory Bowel Disease Monitor*, 14(2):39–44.

Jandhyala, S. M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., and Reddy, D. N. (2015). Role of the normal gut microbiota. *World Journal of Gastroenterology*, 21(29):8836–8847.

Kanehisa, M. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30(1):42–46.

Klatt, N. R., Chomont, N., Douek, D. C., and Deeks, S. G. (2013). Immune activation and HIV persistence: Implications for curative approaches to HIV infection. *Immunological reviews*, 254(1):326–342.

Knight, R., Navas, J., Quinn, R. A., Sanders, J. G., and Zhu, Q. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*.

la Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W. D. (2012). Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. *PLoS ONE*, 7(12):1–13.

Lach, G., Schellekens, H., Dinan, T. G., and Cryan, J. F. (2018). Anxiety, Depression, and the Microbiome: A Role for Gut Peptides.

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9:292.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, 63(4):1079–1088.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21.

Lozupone, C. and Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12):8228–8235.

Lozupone, C., Li, M., Campbell, T., Flores, S., Linderman, D., Gebert, M., Knight, R., Fontenot, A., and Palmer, B. (2013). Alterations in the Gut Microbiota Associated with HIV-1 Infection. *Cell Host & Microbe*, 14(3):329–339.

MacDougall, R. (2012). NIH Human Microbiome Project defines normal bacterial makeup of the body.

Mandal, S., Treuren, W. V., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. 1:1–7.

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158.

Martín-Ferńandez, J. A., Palarea-Albaladejo, J., and Olea, R. A. (2011). Dealing with Zeros. In *Compositional Data Analysis: Theory and Applications*, pages 43–58.

McMurdie, P. J. and Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology*, 10(4).

Mittlböck, M. and Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, 15(19):1987–1997.

Modi, S. R., Collins, J. J., and Relman, D. A. (2014). Antibiotics and the gut microbiota. *The Journal of Clinical Investigation*, 124(10):4212–4218.

Morton, J. T., Sanders, J., Quinn, R. A., McDonald, D., Gonzalez, A., Vázquez-Baeza, Y., Navas-Molina, J. A., Song, S. J., Metcalf, J. L., Hyde, E. R., Lladser, M., Dorrestein, P. C., and Knight, R. (2017). Balance Trees Reveal Microbial Niche Differentiation. *mSystems*, 2(1):e00162–16.

Noguera-Julian, M., Rocafort, M., Guillén, Y., Rivera, J., Casadellà, M., Nowak, P., Hildebrand, F., Zeller, G., Parera, M., Bellido, R., Rodríguez, C., Carrillo, J., Mothe, B., Coll, J., Bravo, I., Estany, C., Herrero, C., Saz, J., Sirera, G., Torrela, A., Navarro, J., Crespo, M., Brander, C., Negredo, E., Blanco, J., Guarner, F., Calle, M. L., Bork, P., Sönnerborg, A., Clotet, B., and Paredes, R. (2016). Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine*, 5:135–146.

Oksanen, J. (2015). Multivariate Analysis of Ecological Communities in R.

Øyri, S. F., Muzes, G., and Sipos, F. (2015). Dysbiotic gut microbiome: A key element of Crohn's disease.

Palarea-Albaladejo, J. and Martín-Fernández, J. A. (2015). ZCompositions - R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96.

Pan, W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic Epidemiology*, 35(4):211–216.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200–2.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications.*

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data.*

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution.–On a Form of Spurious Correlation Which May Arise When

Indices Are Used in the Measurement of Organs. *Proceedings of the Royal Society of London (1854-1905)*, 60(1):489–498.

Phillips, A. N., Neaton, J., and Lundgren, J. D. (2008). The Role of HIV in Serious Diseases Other than AIDS. *AIDS (London, England)*, 22(18):2409–2418.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B., and Ludwig, W. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data .... *Nucleic Acids Research*, 35(21):7188–7196.

Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., Haberman, Y., Walters, T., Baker, S., and Rosh, J. (2015). The treatment-naïve microbiome in new-onset Crohn ' s disease. 15(3):382–392.

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. (2018). Balances: a New Perspective for Microbiome Analysis. *mSystems*, 3(4).

Rivera-Pinto, J., Estany, C., Paredes, R., Calle, M. L., and Noguera-Julián, M. (2017). Statistical Challenges for Human Microbiome Analysis. In Ainsbury, E. A., Calle, M., Cardis, E., Einbeck, J., Gómez, G., and Puig, P., editors, *Extended Abstracts Fall 2015*, volume 7, pages 47–51, Cham. Springer International Publishing.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A

Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14(8):1–14.

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:1–20.

Sommer, F. and Bäckhed, F. (2013). The gut microbiota — masters of host development and physiology. *Nature Reviews Microbiology*, 11:227.

Thursby, E. and Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal*, 474(11):1823–1836.

Tjur, T. (2009). Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. *American Statistician*, 63(4):366–372.

UNAIDS, G. (2017). Global AIDS monitoring 2017: indicators for monitoring the 2016 United Nations Political Declaration on HIV and AIDS.

Vázquez-Castellanos, J. F., Serrano-Villar, S., Latorre, A., Artacho, A., Ferrús, M. L., Madrid, N., Vallejo, A., Sainz, T., Martínez-Botas, J., Ferrando-Martínez, S., Vera, M., Dronda, F., Leal, M., Del Romero, J., Moreno, S., Estrada, V., Gosalbes, M. J., and Moya, A. (2014). Altered metabolism of gut microbiota contributes to chronic immune activation in HIV-infected individuals. *Mucosal Immunology*, 8:760.

Vesterbacka, J., Rivera, J., Noyan, K., Parera, M., Neogi, U., Calle, M., Paredes, R., Sönnerborg, A., Noguera-Julian, M., and Nowak, P. (2017). Richer gut microbiota with distinct metabolic profile in HIV infected Elite Controllers. *Scientific Reports*, 7(1):1–13.

Washburne, A. D., Silverman, J. D., Leff, J. W., Bennett, D. J., Darcy, J. L., Mukherjee, S., Fierer, N., and David, L. A. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5:e2969.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27.

Wilson Tang, W. H. and Hazen, S. L. (2017). The Gut Microbiome and Its Role in Cardiovascular Diseases. *Circulation*, 135(11):1008–1010.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE*, 10(7):1–30.

Zhang, X., Mallick, H., and Yi, N. (2016). Zero-Inflated Negative Binomial

Regression for Differential Abundance Testing in Microbiome Studies. *Journal of Bioinformatics and Genomics*, (December):1–9.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *American Journal of Human Genetics*, 96(5):797–807.

CHAPTER A

Gut Microbiota Linked to Sexual Preference and HIV Infection

EBioMedicine 5 (2016) 135–146

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Gut Microbiota Linked to Sexual Preference and HIV Infection

Marc Noguera-Julian [a,b,c,1], Muntsa Rocafort [a,c,1], Yolanda Guillén [a,c], Javier Rivera [a,b], Maria Casadellà [a,c], Piotr Nowak [d], Falk Hildebrand [e], Georg Zeller [e], Mariona Parera [a], Rocío Bellido [a], Cristina Rodríguez [a], Jorge Carrillo [a,c,g], Beatriz Mothe [a,b,c,f], Josep Coll [a,f], Isabel Bravo [f], Carla Estany [f], Cristina Herrero [f], Jorge Saz [h], Guillem Sirera [f], Ariadna Torrela [i], Jordi Navarro [i], Manel Crespo [i], Christian Brander [a,b,c,j], Eugènia Negredo [b,c,f], Julià Blanco [a,b,c], Francisco Guarner [k], Maria Luz Calle [b], Peer Bork [e,l,m], Anders Sönnerborg [d], Bonaventura Clotet [a,b,c,f], Roger Paredes [a,b,c,f,*]

[a] IrsiCaixa AIDS Research Institute, Ctra de Canyet s/n, 08916 Badalona, Catalonia, Spain
[b] Universitat de Vic-Universitat Central de Catalunya, C. Sagrada Família 7, 08500 Vic, Catalonia, Spain
[c] Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain
[d] Department of Medicine, Unit of Infectious Diseases, Karolinska University Hospital, Karolinska Institutet, Huddinge 141, 86, Stockholm, Sweden
[e] Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany
[f] HIV Unit & Lluita Contra la SIDA Foundation, Hospital Universitari Germans Trias i Pujol, Ctra de Canyet s/n, 08916 Badalona, Catalonia, Spain
[g] ISGLOBAL, Carrer Rosselló, 132, 08036 Barcelona, Catalonia, Spain
[h] BCN Checkpoint, Carrer del Comte Borrell, 164, 08015 Barcelona, Catalonia, Spain
[i] Infectious Diseases Unit, Hospital Universitari Vall d'Hebrón, Passeig de la Vall d'Hebrón, 119–129, 08035 Barcelona, Catalonia, Spain
[j] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
[k] Digestive Diseases Department, Vall d'Hebrón Institute of Research, Hospital Universitari Vall d'Hebrón, Passeig de la Vall d'Hebrón, 119–129, 08035 Barcelona, Catalonia, Spain
[l] Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Str. 10, 13092 Berlin, Germany
[m] Molecular Medicine Partnership Unit, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form 12 January 2016
Accepted 27 January 2016
Available online 28 January 2016

Keywords:
HIV-1
Microbiome
Microbiota
16S rDNA
Prevotella
Bacteroides

ABSTRACT

The precise effects of HIV-1 on the gut microbiome are unclear. Initial cross-sectional studies provided contradictory associations between microbial richness and HIV serostatus and suggested shifts from *Bacteroides* to *Prevotella* predominance following HIV-1 infection, which have not been found in animal models or in studies matched for HIV-1 transmission groups. In two independent cohorts of HIV-1-infected subjects and HIV-1-negative controls in Barcelona (n = 156) and Stockholm (n = 84), men who have sex with men (MSM) predominantly belonged to the *Prevotella*-rich enterotype whereas most non-MSM subjects were enriched in *Bacteroides*, independently of HIV-1 status, and with only a limited contribution of diet effects. Moreover, MSM had a significantly richer and more diverse fecal microbiota than non-MSM individuals. After stratifying for sexual orientation, there was no solid evidence of an HIV-specific dysbiosis. However, HIV-1 infection remained consistently associated with reduced bacterial richness, the lowest bacterial richness being observed in subjects with a virological-immune discordant response to antiretroviral therapy. Our findings indicate that HIV gut microbiome studies must control for HIV risk factors and suggest interventions on gut bacterial richness as possible novel avenues to improve HIV-1-associated immune dysfunction.

1. Introduction

The main clinical problems of people living with HIV (PLWH) in areas with adequate healthcare standards and continued antiretroviral therapy (ART) supply are increasingly related to premature aging (Paiardini and Müller-Trutwin, 2013). That is, a precocious development of type 2 diabetes, dislipidemia, cardiovascular diseases, osteoporosis and frailty syndrome. Such diseases have been related to structural or metabolic perturbations in the gut microbiota of non-HIV-infected subjects (Claesson et al., 2012; Koeth et al., 2013; Le Chatelier et al., 2013; Tang et al., 2013) whereas, in PLWH, have been linked to chronic inflammation, immune activation and endotoxemia (Brenchley et al., 2006; Douek, 2003; Sandler and Douek, 2012). Thus there is considerable interest in understanding the role of the human gut microbiome in HIV pathogenesis and, in particular, its ability to perpetuate chronic inflammation and foster immune senescence. This has immediate clinical implications because, in theory, it might be

[*] Corresponding author at: HIV Unit and IrsiCaixa AIDS Research Institute, Ctra de Canyet s/n, 08916 Badalona, Catalonia, Spain.
E-mail address: rparedes@irsicaixa.es (R. Paredes).

[1] Contributed equally to this work.

http://dx.doi.org/10.1016/j.ebiom.2016.01.032

Table 1
Baseline chacteristics of subjects in the Barcelona test dataset (BCN0).

		Full dataset	HIV-1 positive	HIV-1 negative	p-Value	
No. of subjects		156	129	27		
Age (years)[a]		43 (35, 51)	44 (36, 52)	37 (34, 44)	0.021	
Gender	Male	124 (79.5%)	101 (78.3%)	23 (85.2%)	0.076	0.600
	Female	31 (19.9%)	28 (21.7%)	3 (11.1%)		0.291
	Transgender	1 (0.6%)	0	1 (3.7%)		0.173
Ethnicity	Asiatic	1 (0.6%)	1 (0.8%)	0	0.900	1
	Caucasian	124 (79.5%)	101 (78.3%)	23 (85.2%)		0.600
	Hispanic–Latino	28 (18%)	24 (18.6%)	4 (14.8%)		0.786
	Others	3 (1.9%)	3 (2.3%)	0		1
Risk group	HTS	41 (26.3%)	37 (28.7%)	4 (14.8%)	0.027	0.156
	MSM	100 (64.1%)	77 (59.7%)	23 (85.2%)		0.014
	PWID	15 (9.6%)	15 (11.6%)	0		0.075
Residency	Barcelona	51 (32.7%)	36 (27.9%)	15 (55.6%)	0.058	0.007
	BCN Met	56 (35.8%)	50 (38.8%)	6 (22.2%)		0.125
	Outside BCN Met	38 (24.4%)	33 (25.6%)	5 (18.5%)		0.622
	na	11 (7.1%)	10 (7.7%)	1 (3.7%)		0.691
Profile	Late presenter	11 (7.1%)	11 (8.5%)	0	–	–
	Discordant	18 (11.5%)	18 (14%)	0		–
	Concordant	53 (34%)	53 (41.1%)	0		–
	Early-treated	13 (8.3%)	13 (10.1%)	0		–
	Naïve	15 (9.6%)	15 (11.6%)	0		–
	Viremic control	11 (7.1%)	11 (8.5%)	0		–
	Elite control	8 (5.1%)	8 (6.2%)	0		–
	HIV-1 negative	27 (17.3%)	0	27 (100%)		–
BMI (kg/m^2)[a]		23.8 (22, 26)	23.8 (22, 26)	24.9 (22, 27)	0.469	
Allergy	No	122 (78.2%)	101 (78.3%)	21 (77.8%)	0.205	1
	Yes	30 (19.2%)	26 (20.2%)	4 (14.8%)		0.603
	na	4 (2.6%)	2 (1.5%)	2 (7.4%)		0.138
ATB during the previous 3–6 months		35 (22.4%)	32 (24%)	4 (14.8%)	0.446	
Fecal consistency	Hard	56 (35.9%)	44 (34.1%)	12 (44.4%)	0.535	0.378
	Soft	91 (58.3%)	77 (59.7%)	14 (51.9%)		0.521
	Liquid	5 (3.2%)	5 (3.9%)	0		0.588
	na	4 (2.6%)	3 (2.3%)	1 (3.7%)		0.536
Abdominal transit alterations	Yes	23 (14.7%)	22 (17.1%)	1 (3.7%)	0.089	0.134
	No	127 (81.4%)	103 (79.8%)	24 (88.9%)		0.414
	na	6 (3.9%)	4 (3.1%)	2 (7.4%)		0.277
Defecation frequency (per day)	1	88 (56.4%)	70 (54.3%)	18 (66.7%)	0.669	0.288
	2	47 (30.1%)	40 (31%)	7 (25.9%)		0.653
	3	12 (7.7%)	11 (8.5%)	1 (3.7%)		0.692
	4	5 (3.2%)	5 (3.9%)	0		0.588
	na	4 (2.6%)	3 (2.3%)	1 (3.7%)		0.536
HBV co-infection	Positive	19 (12.2%)	19 (14.7%)	0	0.054	0.045
	Negative	112 (71.8%)	91 (70.6%)	21 (77.8%)		0.638
	na	25 (16%)	19 (14.7%)	6 (22.2%)		0.386
HCV co-infection	Positive	24 (15.4%)	24 (18.6%)	0	0.013	0.015
	Negative	120 (76.9%)	94 (72.9%)	26 (96.3%)		0.005
	na	12 (7.7%)	11 (8.5%)	1 (3.7%)		0.692
Syphilis serology	Positive	21 (13.5%)	20 (15.5%)	1 (3.7%)	0.262	0.128
	Negative	116 (74.3%)	93 (72.1%)	23 (85.2%)		0.225
	na	19 (12.2%)	16 (12.4%)	3 (11.1%)		1
PCR *Chlamydia trachomatis*	Positive	9 (5.8%)	9 (7.0%)	0	0.161	0.360
	Negative	115 (73.7%)	91 (70.5%)	24 (88.9%)		0.055
	na	32 (20.5%)	29 (22.5%)	3 (11.1%)		0.293
PCR *Neisseria gonorrhoeae*	Positive	0	0	0	0.109	
	Negative	125 (80.1%)	100 (77.5%)	25 (92.6%)		
	na	31 (19.9%)	29 (22.5%)	2 (7.4%)		
PCR human papilloma virus	Yes	72 (46.2%)	61 (47.3%)	11 (40.7%)	0.613	0.671
	No	83 (53.2%)	67 (51.9%)	16 (59.3%)		0.530
	na	1 (0.6%)	1 (0.8%)	0		1
Anal cytology	ASCUS	22 (14.1%)	17 (13.2%)	5 (18.5%)	0.664	0.542
	HSIL	7 (4.5%)	7 (5.4%)	0		0.605
	LSIL	30 (19.2%)	26 (20.2%)	4 (14.8%)		0.603
	Normal	80 (51.3%)	64 (49.6%)	16 (59.3%)		0.402
	na	17 (10.9%)	15 (11.6%)	2 (7.4%)		0.738
CD4+ T-cell count (cells/mm^3)[a]	All	–	700 (462, 860)	–	–	–
	Late presenters	–	100 (33, 189)	–		–
	Discordant	–	263 (223, 287)	–		–
	Concordant	–	761 (640, 932)	–		–
	Early-treated	–	785 (506, 930)	–		–
	ART naive	–	701 (564, 813)	–		–
	Viremic control	–	783 (525, 920)	–		–
	Elite control	–	940 (821, 1009)	–		–
Lymphocytes (×10×9/L)[a]		2 (1.7, 2.5)	2 (1.6, 2.5)	2.1 (1.8, 2.3)		0.438
Leukocytes (×10×9/L)[a]		5.8 (4.8, 7.2)	5.6 (4.8, 6.7)	7.1 (5.2, 8.4)		0.011

Table 1 (*continued*)

		Full dataset	HIV-1 positive	HIV-1 negative	p-Value
No. of subjects		156	129	27	
HIV-1 RNA (copies/mL)[a]	Late presenters	–	178,500 (61,880, 340,300)	–	–
	Discordant	–	<40 (<40, <40)	–	–
	Concordant	–	<40 (<40, <40)	–	–
	Early-treated	–	<40 (<40, <40)	–	–
	ART naive	–	13,900 (6867, 43,410)	–	–
	Viremic control	–	794 (243, 1360)	–	–
	Elite control	–	<40 (<40, <40)	–	–

HTS, heterosexual; MSM, men who have sex with men; PWID, people who inject drugs; ATB, antibiotic; BCN Met, Barcelona Metropolitan Area; na, not available.

[a] Median (IQR), p-values for continuous and discrete variables were calculated with the Wilcoxon rank sum and Fisher's tests, respectively.

possible to gear the gut microbiota towards "healthier" equilibrium states with the host, which might allow, for example, to achieve faster immune reconstitution, improve vaccine responses or reduce HIV reservoirs.

However, although expectations are high, the HIV microbiome science is still at its early stages, and much remains to be known. Simple questions such as whether there is a consistent HIV-specific dysbiosis pattern, or which factors are relevant in shaping the microbiome in PLWH remain unanswered. Initial cross-sectional studies in humans have provided contradictory associations between microbial richness and HIV serostatus, and suggested shifts from *Bacteroides* to *Prevotella* predominance following HIV-1 infection (Lozupone et al., 2013; Vázquez-Castellano et al., 2015). Such shifts, however, have neither been found in animal models (Handley et al., 2012) nor in studies matching for HIV-1 risk groups (Yu et al., 2013). Conversely, large international studies in healthy populations have shown that at least in resource-rich countries, the gut microbiome forms a composition landscape with density peaks that can stratify the human population into enterotypes dominated by *Bacteroides*, *Prevotella* and *Ruminococcus*, respectively (Arumugam et al., 2011; Koren et al., 2013). The origin and clinical significance of such enterotypes is uncertain, but they have been linked to genetic (Goodrich et al., 2014), as well as to lifestyle (Clarke et al., 2014; David et al., 2013; Wu et al., 2011) and environmental factors (Modi et al., 2014; Sommer and Bäckhed, 2013), including long-term dietary patterns and exercise. Thereby, associations between *Prevotella* or *Bacteroides* and HIV infection might be easily confounded by other factors. Obtaining reliable information at this level is critical to advance our understanding of HIV pathogenesis, as well as to define the specific targets of novel therapeutic interventions on the human gut microbiome.

2. Methods

2.1. Study Design

This was a cross-sectional study in two independent European cohorts of HIV-1-infected subjects and HIV-negative controls. The study included one test cohort, one internal validation cohort and one external validation cohort (Supplementary Fig. 1).

The test cohort (BCN0) was enrolled in Barcelona, Catalonia, Spain, between January and December 2014. HIV-1 infected patients were recruited from HIV Clinics at the University Hospitals Germans Trias i Pujol and Vall d'Hebrón. HIV-1-negative controls were mainly recruited from an ongoing prospective cohort of HIV-negative MSM at risk of becoming infected by HIV-1 (Coll et al., 2015), who attend quarterly medical and counseling visits including HIV-1 testing (Alere Determine™ HIV-1/2 Ag/Ab Combo, Orlando, FL) at a community-based center for MSM in Barcelona (Meulbroek et al., 2013). Additional controls were HIV-1-negative partners from HIV-1-infected subjects attending the HIV clinics.

The inclusion criteria were: age within 18 and 60 years and body mass index (BMI) within 18.5 and 30. Exclusion criteria were: (a) any gross dietary deviation from a regular diet, or any specific regular diet, i.e., vegetarian, low-carb, etc.; (b) antibiotic use during the previous 3 months (with the exception of late presenters, who could receive antibiotics to treat opportunistic infections); (c) pregnancy or willingness to become pregnant; (d) current drug consumption or alcohol abuse; (e) any chronic digestive disease such as peptic ulcer, Crohn's disease, ulcerative colitis or coeliac disease; (f) any surgical resection of the intestines except for appendectomy; (g) any autoimmune disease; and (h) any symptomatic chronic liver disease or presence of hepatic insufficiency defined as a Child–Pugh C score. In addition, HIV-infected subjects were classified as elite controllers, viremic controllers, ART-naïve, early treated, late presenters, immune concordant or immune discordant (Supplementary methods).

The internal validation cohort (cohort BCN1) included individuals from BCN0 who provided a second fecal sample one month later.

Observations in Barcelona were externally validated in an independent observational cohort recruited at the HIV outpatient clinic, Karolinska University Hospital, Stockholm, Sweden (cohort STK). All HIV-1-infected patients in cohort STK were at least 18 years old, had been diagnosed with HIV-1 between one and 25 years earlier and were ART-naïve at the time of fecal sampling. Controls were healthy HIV-1-negative individuals matched by sex and age. Neither patients nor controls had been prescribed antibiotics or probiotics, or had had infectious diarrhea during the preceding two months.

2.2. Data Collection

Clinical and laboratory data from BCN0 and BCN1 were collected in a centralized database specifically designed for this study (OpenClinica™, © 2015 OpenClinica, LLC). The clinical evaluation was performed following a standardized questionnaire including: a checklist for fulfillment of inclusion and exclusion criteria, anthropometric data, age at study entry, age at HIV diagnosis, gender, ethnicity, city of residence, HIV risk group, history of allergies, antibiotic intake between 3 and 6 months before inclusion, frequency and consistency of feces, history of medical or surgical problems or interventions, present and previous ART, history of AIDS- and non-AIDS-related diseases, nadir and most recent CD4+ T-cell counts, HIV-1 RNA levels, history of sexually transmitted diseases and infection by the human papillomavirus (HPV), hepatitis B (HBV) or hepatitis C (HCV).

HIV-1 risk categories in our study were mutually excluding: male study participants who reported being MSM or referred insertive or receptive anal intercourse with other men were included in the MSM category, even if they also reported intravenous drug use or sex with women. Females and males not included in the MSM category reporting past intravenous drug use were classified as PWID. Heterosexual males or females not included in any of the previous 2 categories were classified as HTS. None of our study participants belonged to any other HIV-1 transmission category.

Study participants in Barcelona received a thorough dietary and nutritional assessment by a specialized dietitian/nutritionist using two standardized and validated questionnaires, i.e.: (a) a prospective dietary nutrient survey aimed at recording, as precisely as possible, any food, supplement or liquid intake during 3 to 5 consecutive days, including at least one weekend day, and (b) a recall of food portions taken per week, on average, during the last year.

Participants also went through a proctology evaluation by a specialized HIV physician/proctologist. In addition to visual inspection for anal or perianal lesions, HPV-related or not, the physician performed a rectal swab to rule out *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infection using real-time PCR and an anal cytology. If the anal cytology reported an abnormal result, such as ASCUS (atypical squamous cells of undetermined significance), LSIL (low-grade squamous intraepithelial lesion) or HSIL (high-grade squamous intraepithelial lesion), the subject was properly treated and PCR typing of HPV was performed. No cases of anal cancer were detected.

In all study participants, we produced MiSeq™ 16S rRNA sequence data on fecal microbiomes and measured soluble plasma markers of enterocyte damage (intestinal fatty acid-binding protein, IFABP), microbial translocation [soluble CD14 (sCD14) and lipopolysaccharide binding protein (LBP)] and systemic inflammation [interleukin-6 (IL-6), C-reactive protein (CRP) and interferon-gamma-inducible protein-10 (IP-10)].

Study participants collected fecal samples in sterile fecal collection tubes the same day or the day before their clinical appointment, before the proctology exam, and following instructions pre-specified on standard operating procedures. If required, samples were stored at 4 °C overnight until DNA extraction. All samples collected in Barcelona were immediately extracted upon arrival to the laboratory. Additional aliquots were cryopreserved at −80 °C for future studies. Samples collected in Stockholm were cryopreserved at −80 °C and shipped on dry ice in batch to the IrsiCaixa AIDS Research Institute, where they were extracted, amplified, sequenced and analyzed using the exact same procedures applied to the Barcelona samples. The lag times to freezing were always <36 h and no particular chemical stabilizers were added to samples used for the analyses presented here. Fecal sample collection procedures were the same for cases and controls.

Detailed descriptions of the wet-lab procedures and the ecological and statistical analyses of the microbiome, soluble plasma markers and the nutritional assessment are available in the Supplementary methods section.

2.3. Ethics & Community Involvement

The study was reviewed and approved by the Institutional Review Boards of the Hospital Universitari Germans Trias i Pujol (reference PI-13-046) and the Hospital Vall d'Hebrón (reference PR(AG)109/2014). The Stockholm study cohort was approved by the Regional Ethical Committee (Stockholm, Sweden, Dnr 2009-1485-31-3). All participants provided written informed consent in accordance with the World Medical Association Declaration of Helsinki. The study concept, design, patient information and results were discussed with the IrsiCaixa's Community Advisory Committee, who also provided input on the presentation and dissemination of study results (Supplementary methods).

2.4. Sequence and Data Availability

Raw Illumina MiSeq sequences and study metadata were deposited in the National Center for Biotechnology Information — NCBI repository (Bioproject accession number: PRJNA307231, SRA accession number: SRP068240).

2.5. Financial Support and Role of the Funding Sources

This study was mainly funded through philanthropy and private donations, which had no influence on its contents. Funds were obtained from a personal donation from Mr. Rafael Punter, the *Gala contra la SIDA* 2013 and 2014 editions, and the *Nit per la Recerca a la Catalunya Central* 2015 edition. M.R. is funded through a FI-DGR grant (FI-B00184) from *Agència de Gestió d'Ajuts Universitaris i de Recerca* (AGAUR) at the *Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya*. Y.G. is supported through a post-doctoral grant from the *Fundación Paideia Galiza*. M.C. is funded through the *Red de Investigación en SIDA*, RD12/0017/0002 as part of the Plan Nacional R + D + I and cofinanced by the *Instituto de Salud Carlos III* (ISCIII)-*Subdirección General de Evaluación y el Fondo Europeo de Desarrollo Regional* (FEDER). J.R. is supported through a grant for doctoral studies from Noel Alimentaria to the University of Vic (UVic-UCC). B.M. is a Joan Rodés investigator from the ISCIII

Table 2
Baseline chacteristics of subjects in the Stockholm validation dataset (STK).

		Full dataset	HIV-1 positive	HIV-1 negative	p-Value
No. of subjects		84	77	7	
Age (years)		40 (32, 48)	38 (32, 49)	44 (38, 47)	0.615
Gender	Male	51 (60.7%)	46 (59.7%)	5 (71.4%)	0.699
	Female	33 (39.3%)	31 (40.3%)	2 (28.6%)	
Risk group	HTS	55 (66.5%)	48 (62.3%)	7 (100%)	0.214
	MSM	19 (22.6%)	19 (24.7%)	0	
	PWID	10 (11.9%)	10 (13.0%)	0	
Ethnicity	Asian	2 (2.4%)	2 (2.6%)	0	1
	Black	28 (33.3%)	26 (33.8%)	2 (28.6%)	
	Caucasian	52 (61.9%)	47 (61.0%)	5 (71.4%)	
	Hispanic–Latino	2 (2.4%)	2 (2.6%)	0	
Country of origin	Sweden	39 (46.4%)	34 (43.4%)	5 (71.4%)	0.069
	Kenya	5 (5.9%)	5 (6.5%)	0	
	Finland	4 (4.8%)	4 (5.2%)	0	
	Ethiopia	3 (3.6%)	1 (1.3%)	2 (28.6%)	
	Eritrea	3 (3.6%)	3 (3.9%)	0	
	Nigeria	3 (3.6%)	3 (3.9%)	0	
	Uganda	3 (3.6%)	3 (3.9%)	0	
	Other	24 (28.5%)	24 (31.2%)	0	
CD4 + T-cell count (cells/mm^3)[a]		–	480 (380, 630)	–	–
CD4 + T-cell count (%)[a]		–	26 (20, 32)	–	–
CD8 + T-cell count (cells/mm^3)[a]		–	970 (660, 1290)	–	–
CD8 + T-cell count (%)[a]		–	51 (44, 59)	–	–
HIV-1 RNA copies/mL[a]		–	19,100 (1590, 69,900)	–	–

[a] Median (IQR), p-values for continuous and discrete variables were calculated with the Wilcoxon rank sum and Fisher's tests, respectively.

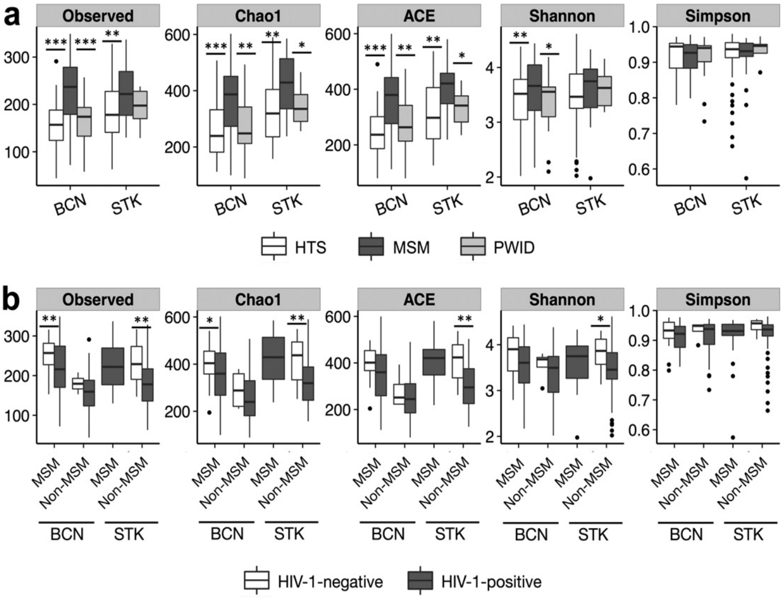


Fig. 1. Both HIV transmission group and HIV-1 infection are linked to the human fecal microbiome richness and diversity. a) The highest richness and diversity in human fecal microbiota were observed in men who have sex with men (MSM). There were no differences between heterosexual subjects (HTS) and people who acquired HIV-1 infection through intravenous drug use (PWID). Kruskal–Wallis p-values in a were adjusted for multiple comparisons using the Benjamini–Hochberg method. b) HIV-1 infection was associated with significant reductions in fecal human microbiome richness after stratifying for sexual preference. Comparisons were made with the Wilcoxon rank sum test with continuity correction. All alpha diversity findings were consistent in Barcelona (test cohort, month 0) (BCN0) and Stockhom (STK). Identical results were found in BCN at month 1 (Supplementary Fig. 2) and when using an independent sequence analysis pipeline (Supplementary Fig. 14). Note: "Simpson" refers to 1-Simpson index. The remaining ecological index names are self-explanatory. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

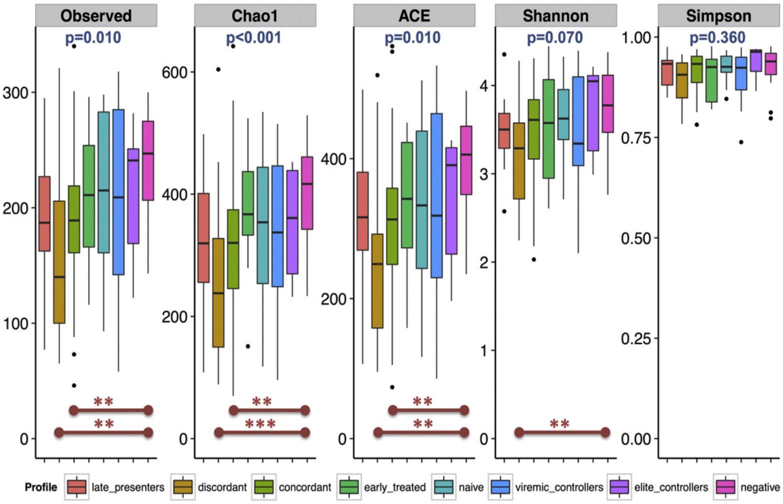


Fig. 2. Alpha diversity by HIV-1 phenotype. HIV-1-infected subjects with an immune discordant phenotype (i.e. those who do not recover CD4+ counts >300 cells/mm^3 despite at least 2 years of effective antiretroviral therapy) had the lowest microbiome richness of all HIV-1 phenotypes. Individuals with an immune concordant phenotype (i.e., those achieving CD4+ count reconstitution >500 cells/mm^3 on antiretroviral therapy) also had lower microbiome richness than HIV-1-negative individuals, but not as low as immune discordant subjects. "Simpson" refers to 1-Simpson index. The remaining ecological index names are self-explanatory. Comparisons were done using a Kruskal–Wallis test including post-hoc pairwise analyses. Benjamini–Hochberg-adjusted p-values are shown at the top of each index; for post-hoc pairwise comparisons: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

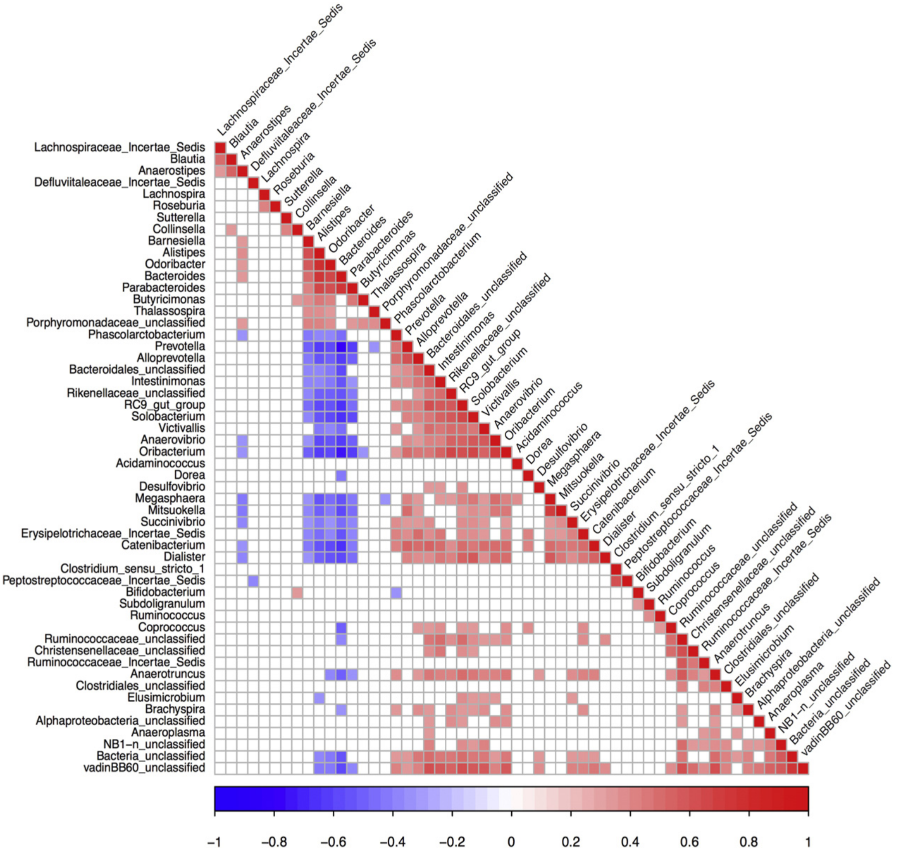


Fig. 3. Spearman correlation by genus abundance. Only significant values (Holm's-corrected $p < 0.05$) are shown. The plot confirms previous observations, i.e.: a) strong positive correlations between *Bacteroides*, *Parabacteroides*, *Barnesiella*, *Alistipes* and *Odoribacter*, b) strong positive correlations between *Prevotella*, *Alloprevotella*, *Mitsuokella* and *Intestinimonas*, among others, and c, strong inverse correlations between the groups including *Prevotella* and *Bacteroides*.

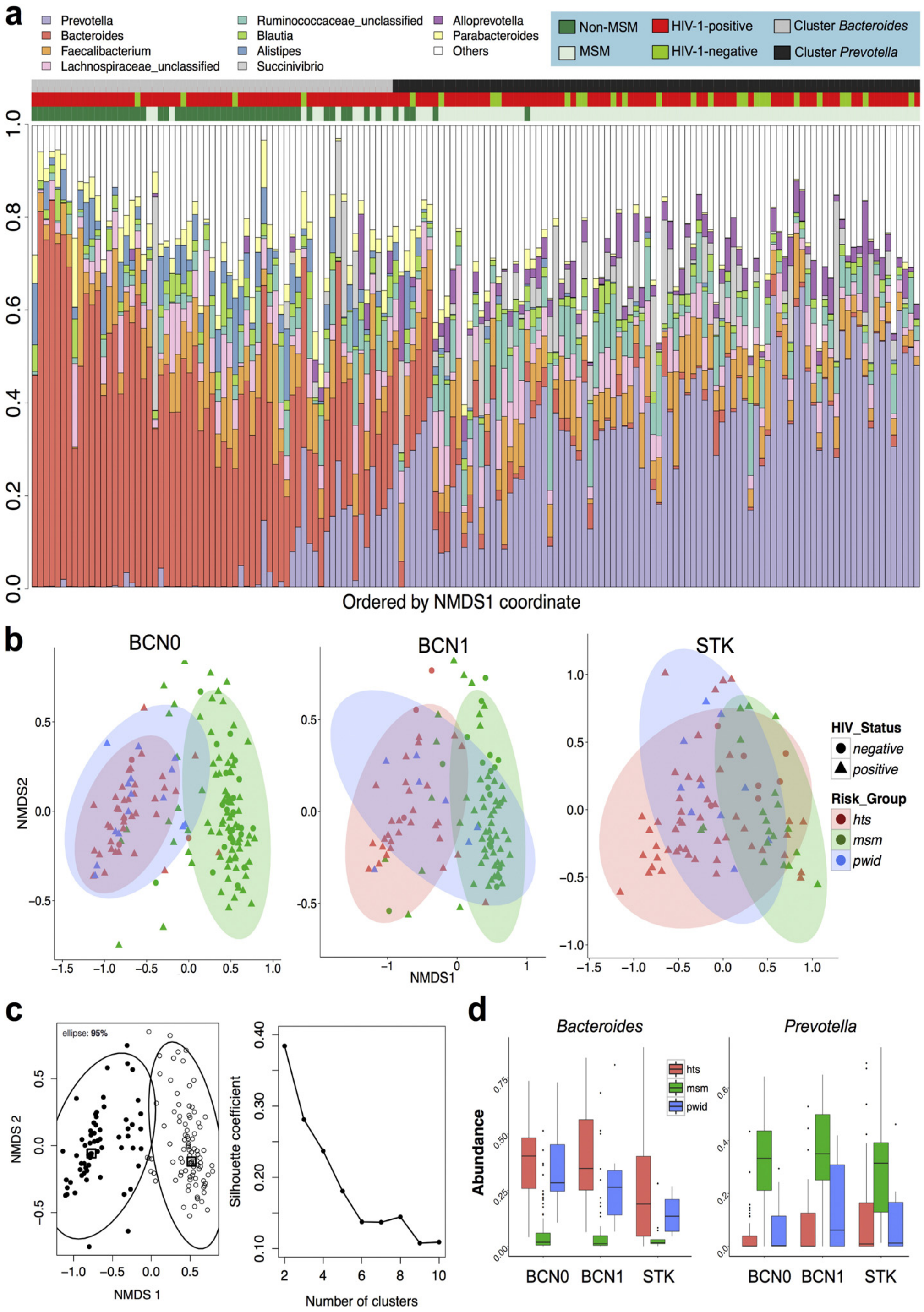
(JR13/00024), Madrid, Spain. M.L.C. is funded through the grant MTM2012-38067-C02-02, Spanish Ministry of Economy and Competitiveness, Spain. PB, FH and GZ are supported by the European Molecular Biology Laboratory. F.H. was funded from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 600375. P.B. acknowledges the European Research Council grant Cancerbiome, reference 268985. The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all study data, and had final responsibility for the decision to submit for publication.

3. Results

3.1. Study Subjects

The study included 240 individuals, 156 in Barcelona (Table 1) and 84 in Stockholm (Table 2). The test cohort BCN0 comprised 129 (82.7%) HIV-1-infected and 27 (17.3%) HIV-negative subjects. The internal validation cohort BCN1 included 110 individuals, 87 HIV-1-infected (79.1%) and 23 non-HIV-infected (20.9%). The external validation cohort STK had 77 HIV-1-infected (91.6%) and 7 non-HIV-infected individuals (8.4%). In Barcelona, the median age of study participants

Fig. 4. The bacterial genus composition of the human fecal microbiome is mainly linked to HIV transmission group. a) The bacterial genus composition of the fecal microbiota in the Barcelona test dataset (BCN0) was largely determined by HIV transmission group, with MSM being enriched in the *Prevotella* cluster and non-MSM in the *Bacteroides* cluster. Genera with mean abundance of at least 2% across all samples are represented in colors; those with <2% abundance are grouped into the category "Others". Each column represents one individual. A similar plot for the Stockholm cohort is shown in Supplementary Fig. 12. b) Non-metric multidimensional scaling (NMDS) ordination plots of Bray–Curtis distances showing that microbiomes in the BCN0, BCN1 and STK datasets mainly cluster by HIV transmission group (MSM vs. non-MSM) rather than by HIV serostatus. Ellipses include 95% of samples. Similar plots using other distances are shown in Supplementary Figs. 8 to 10. c) Partitioning around medoids (PAM) analysis of the BCN0 dataset showing this population structure in this dataset is better explained by 2 rather than more clusters, with reasonable Silhouette support. This information was used to define the *Bacteroides* and *Prevotella* clusters in our study. d) Abundance box plots showing that MSM were enriched in *Prevotella* and non-MSM (HTS or PWID) were enriched in *Bacteroides*. Comparisons between MSM and the non-MSM categories were always highly significant ($p < 0.001$) after adjusting for multiple comparisons using the Benjamini–Hochberg method. Plots of all genera showing significant differences between MSM and non-MSM categories are shown in the Supplementary Figs. 13 to 15.


a
Prevotella
Bacteroides
Faecalibacterium
Lachnospiraceae_unclassified
Ruminococcaceae_unclassified
Blautia
Alistipes
Succinivibrio
Alloprevotella
Parabacteroides
Others
Non-MSM
MSM
HIV-1-positive
HIV-1-negative
Cluster Bacteroides
Cluster Prevotella
0.0
0.2
0.4
0.6
0.8
1.0
Ordered by NMDS1 coordinate
b
BCN0
BCN1
STK
NMDS2
NMDS1
HIV_Status
negative
positive
Risk_Group
hts
msm
pwid
c
ellipse: 95%
NMDS 2
NMDS 1
Silhouette coefficient
Number of clusters
d
Bacteroides
Prevotella
Abundance
hts
msm
pwid
BCN0
BCN1
STK

was 43 years and their median body mass index was 23.8 kg/m^2. Eighty percent of subjects were men, mostly from Caucasian ethnicity. Sixty-four percent of all subjects were MSM, 26% HTS and 10% PWID. There were 8 (5.1%) elite controllers, 11 (7.1%) viremic controllers, 15 (9.6%) ART-naïve, 13 (8.3%) early-treated, 53 (34.1%) immune concordant, 18 (11.5%) immune discordant, and 11 (7.1%) late presenters. HIV-1-infected subjects were slightly older and were more likely to be HBV and HCV positive than HIV-negative controls. Groups were well balanced in all other factors. In Stockholm, 60% of subjects were men; 23% were MSM, 66% HTS and 11% PWID. Only half were nationals from Scandinavian countries; 62% individuals were Caucasian and 33% were Black.

3.2. Richness and Diversity of the Fecal Microbiota

The fecal microbiota was significantly richer and more diverse in MSM than non-MSM individuals in both cities, also after correcting for multiple comparisons (Fig. 1, Supplementary Figs. 2 and 14). This indicated that the measurement of the effect of HIV-1 on gut microbial richness and diversity had to take HIV transmission group into account. After stratifying for MSM vs. non-MSM, HIV-1 infection remained consistently associated with reduced bacterial richness (15% to 30% reduction relative to HIV-negative individuals) in both groups and both cities (Fig. 1, Supplementary Figs. 2 and 14). In the Barcelona cohort, the lowest microbial richness and diversity was observed among HIV-1-infected individuals with an immune-virological discordant phenotype (Fig. 2). Subjects with an immune-virological concordant phenotype had higher microbial richness than immune discordant individuals, but, nevertheless, still showed reduced microbial richness relative to HIV-negative controls, suggesting that despite adequate immune recovery [median (IQR) CD4 + T-cell counts: 761 (640, 932) cells/mm^3] at the time of testing, ART had not been able to fully normalize microbial richness.

3.3. Bacterial Composition of the Fecal Microbiota

Clustering of the fecal microbiomes in BCN0 and STK using a partitioning around medoids (PAM) algorithm suggested the

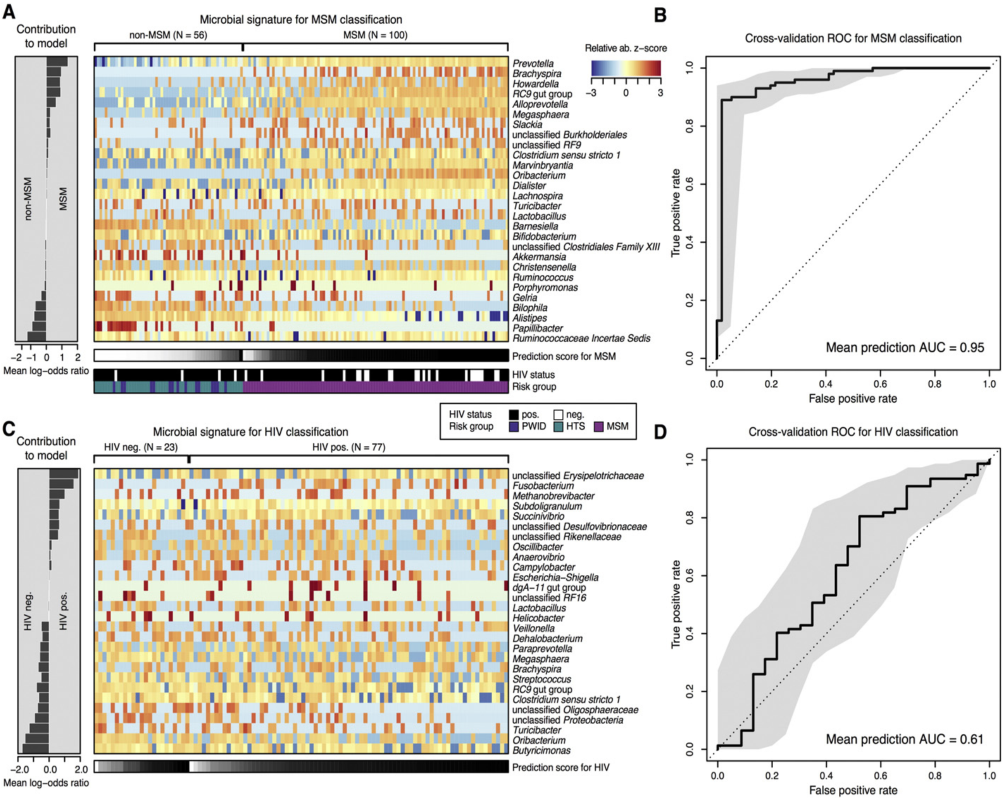


Fig. 5. Global microbiota classifier by sexual preference group and HIV-1 status. A and C) Relative abundances of 28 gut microbial genera collectively associated with MSM and HIV-1 infection, respectively, are displayed as heatmap of log-abundance z-scores with the direction of association indicated to the left. To avoid confounding by sexual preference, the HIV-1 classifier only includes MSM subjects. The mean contribution of each marker species to the classification is shown to the left (bars correspond to log-odds ratio in logistic regression). Below each heatmap the classification score of the microbial signature from cross-validation is shown as gray scale. HIV-1 status and HIV-1 risk group are color-coded below the first heatmap (see color key). B and D) Cross-validation accuracy of the microbiota classifier is depicted as receiver–operator-characteristic (ROC) curve summarizing mean test predictions made in ten times resampled tenfold cross-validation with the area under the curve (AUC) indicated inside each plot. As shown, there was a strong association between the global microbiome genus composition and sexual orientation, whereas the association with HIV-1 infection was much weaker and of uncertain significance.

presence of at least 2 clusters of fecal microbiomes in both cities (Fig. 4c). Such clusters were enriched either in *Bacteroides* or *Prevotella*, and had a similar bacterial composition to the corresponding previously described enterotypes (Arumugam et al., 2011; Koren et al., 2013) (Supplementary Fig. 3). As expected from previous work on gut enterotypes, there were strong positive correlations between the genus *Bacteroides* and *Parabacteroides*, *Barnesiella*, *Alistipes* and *Odoribacter*, as well as between *Prevotella* and *Alloprevotella*, *Catenibacterium*, *Mitsuokella* and *Intestinimonas*, among others (Fig. 3), highlighting that differences between the groups extended beyond a single genus. The genera correlating with *Prevotella* were negatively correlated with *Bacteroides* and vice versa. Moreover, the microbiomes of the *Bacteroides* and *Prevotella* clusters showed remarkably different functional profiles (Supplementary Figs. 4 and 5), also in agreement with previous enterotype descriptions (Arumugam et al., 2011).

3.4. Factors Associated With the Fecal Microbiota Composition

We explored variables potentially influencing the composition of the fecal microbiomes, according to a univariate ADONIS test of ecological

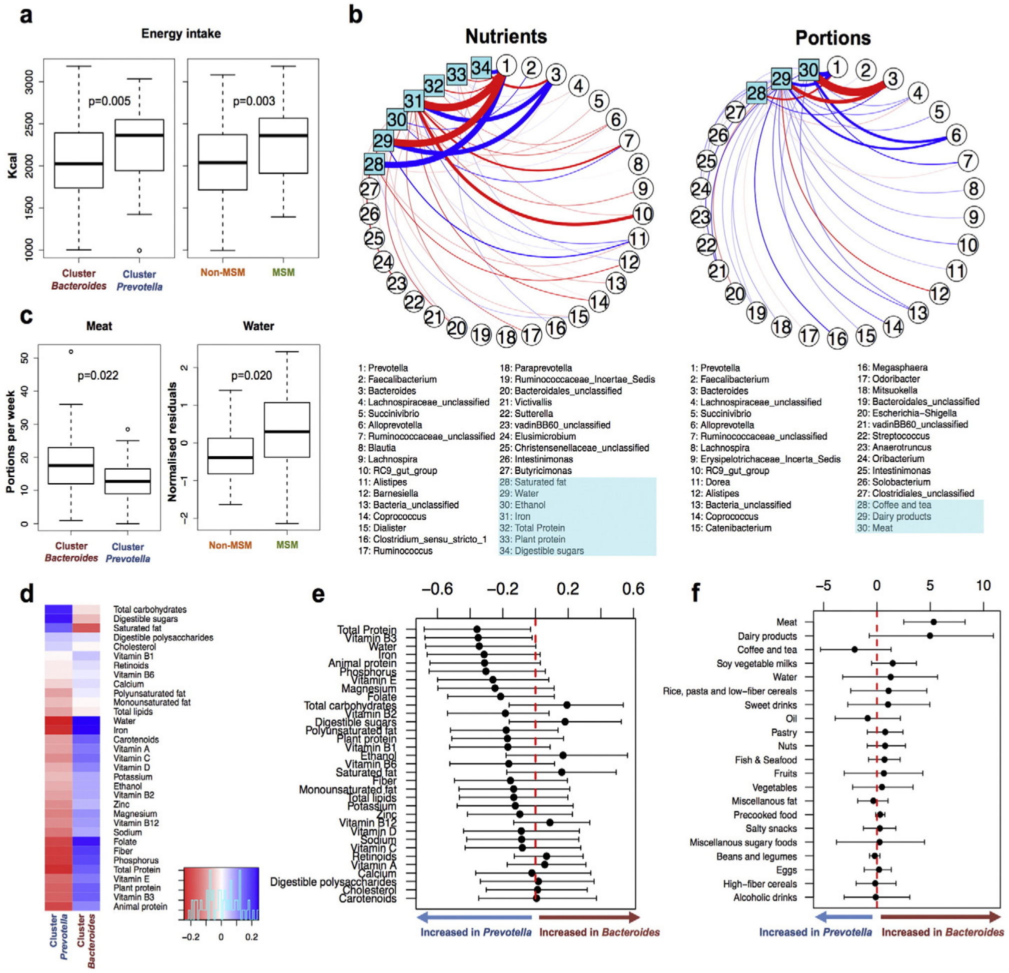


Fig. 6. Limited effect of diet on the composition of the microbiome. a) Subjects belonging to the *Prevotella* cluster and men who had sex with men (MSM) had significantly higher total energy intake. Therefore, all subsequent nutritional analyses were normalized for this factor. b) Main associations between bacterial genera, normalized amounts of nutrients (left) and food portions (right), according to a Dirichlet multinomial regression model. Positive and negative associations are shown in red and blue, respectively. Line thickness is proportional to the strength of the association. c) Of all links identified by the Dirichlet approach, the only significant differences between groups after adjusting for multiple comparisons (Benjamini–Hochberg FDR < 0.1) were increased consumption of meat in the cluster *Bacteroides* and increased intake of dietary water in MSM. d) Spearman correlations between normalized amounts of nutrients and Bray–Curtis distance to the furthest subject in the opposite cluster. Negative correlations imply increased amounts of nutrient with shorter distance to each cluster. Therefore, values in red and blue represent increased and decreased amounts of nutrients within each cluster, respectively. Although, in general, the direction of the correlations was concordant with previous publications, note the small effect sizes (R^2 below the color key). None of the comparisons were statistically significant after correction for multiple comparisons (Benjamini–Hochberg FDR < 0.1); Permanova p = 0.20 for overall differences between clusters. e, f) Mean and 95% confidence intervals for the differences between clusters in consumption of nutrients (e) and portions of food (f). Comparisons were significant if the 95 confidence interval did not cross 0 (dashed red line).

distance and found possible effects of HIV-1 risk group, gender, feces consistency, place of residency, ethnicity, HIV-1 serostatus and altered abdominal transit (Supplementary Table 1). However, only the HIV-1 risk group retained statistical significance in a multivariate ADONIS analysis with terms added sequentially (R^2: 0.373, $p < 0.001$).

Fecal microbiomes in BCN0, BCN1 and STK clustered by HIV transmission group rather than by HIV-1 serostatus, using either Bray–Curtis (Fig. 4b) or other ecological distances (Supplementary Figs. 6 to 8). Although a few individuals showed marked differences between the two time points, fecal microbiota ordination was highly concordant between BCN0 and BCN1 (Procrustes $m2 = 0.3475$, PROTEST $p = 0.001$) (Supplementary Fig. 9), indicating that differences in microbial ordination were not due to random variation. The fecal microbiota composition in both BCN0 and STK significantly differed by HIV transmission group, with MSM and non-MSM subjects mostly belonging to the *Prevotella* and *Bacteroides* clusters, respectively (Fig. 4a and 4d and Supplementary Figs. 10 to 14). Alpha and beta diversity and genus abundance analyses were reproducible using a different analysis pipeline (Hildebrand et al., 2014) (Supplementary Fig. 14 and Supplementary methods).

In an analysis accounting for the potential interdependency of sexual preference and HIV-1 serostatus (LEfSe) (Segata et al., 2011), there were consistent differences in both cities only by sexual preference group, with enrichment of *Prevotella*, *Alloprevotella*, *Succinvibrio*, *Dorea*, RC 9 gut group, *Desulfovibrio*, *Phascolarctobacterium* and unclassified *Bacteroidales* in MSM, and enrichment in *Bacteroides*, *Odoribacter* and *Barnesiella* in non-MSM individuals (Supplementary Fig. 15).

3.5. Strength of the Associations

To quantify the strength of the association between HIV transmission group, HIV serostatus and global fecal microbiota composition, we applied a previously validated global microbiota classification concept based on LASSO regression (Zeller et al., 2014) to our BCN0 dataset. Cross-validation accuracy was extraordinarily high for sexual preference group (mean AUC = 95%), confirming a different fecal microbiota composition in MSM and non-MSM individuals (Fig. 5). In contrast, HIV-1 status was not associated with consistent changes in the global fecal microbiota composition at the genus level, suggesting that the reduction in microbial richness observed in HIV-infected individuals was not genus-specific.

Relative to non-MSM subjects, MSM were younger, were more likely to live in Barcelona City, reported softer fecal consistency, and were less likely to be infected with HBV and HCV (Supplementary Table 2). However, none of these factors among others were likely to confound the previous LASSO models (Supplementary Fig. 16). Although long-term dietary patterns have been linked to alternative enterotype states (Wu et al., 2011), the effect of diet on microbiota composition was limited in our setting (Fig. 6 and Supplementary Fig. 17) and none of the diet components was selected by multivariate LASSO regression as a consistent predictor of microbiota clustering.

3.6. Consequences on Enterocyte Damage, Microbial Translocation and Systemic Inflammation

Markers of enterocyte damage, microbial translocation and systemic inflammation followed an overall predictable response across different HIV phenotypes (Brenchley and Douek, 2012), being generally higher in immune discordant and late presenters (Supplementary Figs. 18 and 19). However, they did not differ between the *Bacteroides* or *Prevotella* clusters or between MSM and non-MSM individuals.

4. Discussion

In two independent European cohorts with different ethnic and cultural background, the fecal microbiota of MSM was consistently richer and more diverse than that of non-MSM subjects, and was systematically enriched in genera from the *Prevotella* enterotype. The strength of such association was unusually high, reaching 95% accuracy in a microbial composition-based classifier. These findings have important implications for HIV microbiome science. To our knowledge, this is the first evidence that, in addition to genetic (Goodrich et al., 2014), lifestyle (Clarke et al., 2014; David et al., 2013; Wu et al., 2011) and environmental factors (Modi et al., 2014; Sommer and Bäckhed, 2013), factors related with sexual preference might also affect the gut microbiota composition.

Based on our findings, previous associations between HIV infection and *Prevotella* might be explained by enrichment of HIV-infected groups by MSM relative to HIV-negative controls selected from hospital or research staff, gut biopsy donors, or college students (Lozupone et al., 2013; Mutlu et al., 2014; Vázquez-Castellano et al., 2015). Contradictory associations between HIV infection and microbial richness could also be affected by unbalances in the proportion of MSM between groups. Of note, a selection bias as such could also affect the interpretation of in silico inferences on bacterial metabolism, or even direct metabolomic or metatranscriptomic measurements, which also rely on bacterial composition.

In concordance with data from animal models (Handley et al., 2012) and studies matching for HIV risk factors (Yu et al., 2013), we were unable to identify a consistent HIV-specific fecal dysbiosis pattern after stratifying for HIV transmission group. Yet, HIV-1 infection remained associated with reduced bacterial richness independently of sexual orientation, indicating that the most evident hallmark of HIV infection on the gut microbiome is, like in other intestinal inflammatory diseases (Manichanh et al., 2012), a reduction in bacterial richness. In line with previous observations linking bacterial richness with immune dysfunction (Nowak et al., 2015), the lowest bacterial richness was found in immune discordant subjects, followed by immune concordant individuals with adequate immune recovery on ART. Conversely, bacterial richness was conserved in subjects initiating ART during the first 6 months of HIV infection, as well as in ART-naïve individuals with >500 CD4+ counts/mm^3, suggesting that early ART initiation might help to preserve gut microbial richness.

The strong epidemiological association of fecal microbiota composition with sexual orientation in two independent cities is yet to be translated into specific mechanisms. We ruled out multiple confounders and only found a limited effect of diet in our setting. We did not collect information on exercise, but exercise has been linked to fecal microbiota composition in athletes (Clarke et al., 2014) and even in them diet plays an important role. A formal assessment of the socioeconomic status of our patients was out of the scope of this work, although based on our findings, rigorous studies assessing the role of socioeconomic status in the fecal microbiota composition are needed. Non-MSM subjects in our study were older and more likely to be co-infected with HBV and HCV than MSM, reflecting current trends of the HIV epidemic in Europe, i.e.: most new HIV-1 infections occur in young MSM who rarely use intravenous drugs. Fecal consistency was also softer in MSM than in non-MSM subjects, which, indirectly, might reflect better overall health habits, including a healthier diet, higher water consumption and physical activity. However, none of these factors, nor ethnicity, achieved a significant weight in LASSO models.

Further studies are needed to evaluate the existence of ecological adaptations of commensal bacteria to changes in gut mucosa induced by sexual practices. Populations of commensal bacteria are controlled by substrate competition and glycan availability (Koropatkin et al., 2012) and several factors might affect distal colorectal mucosa, including hyperosmolar substances like semen or certain lubricants (Fuchs et al., 2007; McGowan, 2012), colorectal cleansing or use of sexual toys. Longitudinal studies should also clarify if the observed association is stable over time, and if it varies according to the number of sexual partners (i.e., long-term single relationships versus frequent partner exchange) or by insertive versus receptive anal sex. It is also important to clarify if the observed association remains in heterosexual women who engage in receptive anal sex and if increased microbiota richness can be related to person-to-person transmission of commensal bacteria. Future studies should also investigate if the observed association has implications for transmission of infectious agents, including HIV-1. We did not find an association between fecal microbiome and HBV, HCV, syphillis or rectal HPV, *C. trachomatis* or *N. gonorrhoeae* infections, but did not evaluate HSV-2 infection. In our study, the observed association between sexual orientation and microbiota composition did not translate into gross differences in terms of systemic inflammation or microbial translocation. Shotgun metagenomic analyses of bacterial species and richness, as well as the virome and perhaps the mycobiome, in clinical trials balanced by HIV risk factors might provide novel clues as to the impact of HIV infection on the gut microbiome.

In conclusion, the fecal microbiota of gay men in Europe is richer and has a distinct composition. However, HIV-1 infection remains independently associated with reduced bacterial richness. This offers new avenues for therapeutic interventions on the gut microbiome which might improve HIV-associated immune dysfunction.

Author Contributions

R.P., M.N., P.N., A.S., J.B. and B.C. conceived and designed the study. R.P., I.B., B.M., E.N., J.Co., J.S., A.T., J.N., C.B. and B.C in Barcelona and P.N., and A.S., in Stockholm, recruited the study participants and performed their clinical evaluations. C.E. performed the dietary assessment. G.S. and J.C. performed the proctology studies. C.H. coordinated the study logistics including the fulfillment of all ethical and legal requirements of the study as member of the Contract Research Organization overseeing the study, in coordination with R.P. Fecal 16S rDNA was extracted, amplified and sequenced by M.P, M.C., M.R. and R.B. under the supervision of M.N. and R.P. M.R., Y.G, J.R., C.R., F.H. and G.Z. performed the bioinformatic and statistical analyses of the 16S rDNA data, with the supervision of M.N., M.L.C., P.B., F.G. and R.P. M.C. performed the inflammation analyses under the supervision of J.Ca., J.B and R.P. J.R., did the statistical analyses of the relationship between the microbiota and inflammation and diet, under the supervision of M.N., M.L.C., F.G. and R.P. F.H and G.Z. performed the multivariate analysis of factors determining microbiota clusters and ran the confirmatory analyses with the independent sequence analysis pipeline LotuS, under supervision of P.B. R.P. wrote the paper, which was reviewed, edited and approved by all authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ebiom.2016.01.032.

References

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H.B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E.G., Wang, J., Guarner, F., Pedersen, O., de Vos, W.M., Brunak, S., Doré, J., Antolín, M., Artiguenave, F., Blottiere, H.M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K.U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S.D., Bork, P., 2011. Enterotypes of the human gut microbiome. Nature 473, 174–180. http://dx.doi.org/10.1038/nature09944.

Brenchley, J.M., Douek, D.C., 2012. Microbial translocation across the GI tract. Annu. Rev. Immunol. 30, 149–173. http://dx.doi.org/10.1146/annurev-immunol-020711-075001.

Brenchley, J.M., Price, D.A., Schacker, T.W., Asher, T.E., Silvestri, G., Rao, S., Kazzaz, Z., Bornstein, E., Lambotte, O., Altmann, D., Blazar, B.R., Rodriguez, B., Teixeira-Johnson, L., Landay, A., Martin, J.N., Hecht, F.M., Picker, L.J., Lederman, M.M., Deeks, S.G., Douek, D.C., 2006. Microbial translocation is a cause of systemic immune activation in chronic HIV infection. Nat. Med. 12, 1365–1371. http://dx.doi.org/10.1038/nm1511.

Claesson, M.J., Jeffery, I.B., Conde, S., Power, S.E., O'Connor, E.M., Cusack, S., Harris, H.M.B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., Fitzgerald, G.F., Deane, J., O'Connor, M., Harnedy, N., O'Connor, K., O'Mahony, D., van Sinderen, D., Wallace, M., Brennan, L., Stanton, C., Marchesi, J.R., Fitzgerald, A.P., Shanahan, F., Hill, C., Ross, R.P., O'Toole, P.W., 2012. Gut microbiota composition correlates with diet and health in the elderly. Nature 488, 178–184. http://dx.doi.org/10.1038/nature11319.

Clarke, S.F., Murphy, E.F., O'Sullivan, O., Lucey, A.J., Humphreys, M., Hogan, A., Hayes, P., O'Reilly, M., Jeffery, I.B., Wood-Martin, R., Kerins, D.M., Quigley, E., Ross, R.P., O'Toole, P.W., Molloy, M.G., Falvey, E., Shanahan, F., Cotter, P.D., 2014. Exercise and associated dietary extremes impact on gut microbial diversity. Gut 63, 1913–1919. http://dx.doi.org/10.1136/gutjnl-2013-306541.

Coll, J., Leon, A., Carrillo, A., Fernandez, E., Bravo, I., Saz, J., Meulbroek, M., Pujol, F., Gonzalez, V., Casabona, J., Ferrer, L., Blanco, J.L., Piñol, M., Garcia-Cuyas, F., Sirera, G., Chamorro, A., Revollo, B., Gatell, J.M., Clotet, B., Brander, C., 2015. Early diagnosis of HIV infections and detection of asymptomatic STI in a community-based organization addressed to MSM. 8th IAS Conference on HIV Pathogenesis Treatment and Prevention, 19–22 July 2015 International AIDS Society, Vancouver.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., Biddinger, S.B., Dutton, R.J., Turnbaugh, P.J., 2013. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505, 559–563. http://dx.doi.org/10.1038/nature12820.

Douek, D.C., 2003. Disrupting T-cell homeostasis: how HIV-1 infection causes disease. AIDS Rev. 5, 172–177.

Fuchs, E.J., Lee, L.A., Torbenson, M.S., Parsons, T.L., Bakshi, R.P., Guidos, A.M., Wahl, R.L., Hendrix, C.W., 2007. Hyperosmolar sexual lubricant causes epithelial damage in the distal colon: potential implication for HIV transmission. J. Infect. Dis. 195, 703–710. http://dx.doi.org/10.1086/511279.

Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T., Spector, T.D., Clark, A.G., Ley, R.E., 2014. Human genetics shape the gut microbiome. Cell 159, 789–799. http://dx.doi.org/10.1016/j.cell.2014.09.053.

Handley, S.A., Thackray, L.B., Zhao, G., Presti, R., Miller, A.D., Droit, L., Abbink, P., Maxfield, L.F., Kambal, A., Duan, E., Stanley, K., Kramer, J., Macri, S.C., Permar, S.R., Schmitz, J.E., Mansfield, K., Brenchley, J.M., Veazey, R.S., Stappenbeck, T.S., Wang, D., Barouch, D.H., Virgin, H.W., 2012. Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. Cell 151, 253–266. http://dx.doi.org/10.1016/j.cell.2012.09.024.

Hildebrand, F., Tadeo, R., Voigt, A.Y., Bork, P., Raes, J., 2014. LotuS: an efficient and user-friendly OTU processing pipeline. Microbiome 2, 1–7. http://dx.doi.org/10.1186/2049-2618-2-30.

Koeth, R.A., Wang, Z., Levison, B.S., Buffa, J.A., Org, E., Sheehy, B.T., Britt, E.B., Fu, X., Wu, Y., Li, L., Smith, J.S., DiDonato, J.A., Chen, J., Li, H., Wu, G.D., Lewis, J.D., Warrier, M., Brown, J.M., Krauss, R.M., Tang, W.H.W., Bushman, F.D., Lusis, A.J., Hazen, S.L., 2013. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat. Med. 19, 533–534. http://dx.doi.org/10.1038/nm.3178.

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C., Ley, R.E., 2013. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput. Biol. 9, e1002863. http://dx.doi.org/10.1371/journal.pcbi.1002863.

Koropatkin, N.M., Cameron, E.A., Martens, E.C., 2012. How glycan metabolism shapes the human gut microbiota. Nat. Rev. Microbiol. 10, 323–335. http://dx.doi.org/10.1038/nrmicro2746.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., Leonard, P., Li, J., Burgdorf, K., Grarup, N., Jørgensen, T., Brandslund, I., Nielsen, H.B., Juncker, A.S., Bertalan, M., Levenez, F., Pons, N., Rasmussen, S., Sunagawa, S., Tap, J., Tims, S., Zoetendal, E.G., Brunak, S., Clément, K., Doré, J., Kleerebezem, M., Kristiansen, K., Renault, P., Sicheritz-Ponten, T., de Vos, W.M., Zucker, J.-D., Raes, J., Hansen, T., Bork, P., Wang, J., Ehrlich, S.D., Pedersen, O., Guedon, E., Delorme, C., Layec, S., Khaci, G., van de Guchte, M., Vandemeulebrouck, G., Jamet, A., Dervyn, R., Sanchez, N., Maguin, E., Haimet, F., Winogradski, Y., Cultrone, A., Leclerc, M., Juste, C., Blottière, H., Pelletier, E., LePaslier, D., Artiguenave, F., Bruls, T., Weissenbach, J., Turner, K., Parkhill, J., Antolin, M., Manichanh, C., Casellas, F., Boruel, N., Varela, E., Torrejon, A., Guarner, F., Denariaz, G., Derrien, M., van Hylckama Vlieg, J.E.T., Veiga, P., Oozeer, R., Knol, J., Rescigno, M., Brechot, C., M'Rini, C., Mérieux, A., Yamada, T., 2013. Richness of human gut microbiome correlates with metabolic markers. Nature 500, 541–546. http://dx.doi.org/10.1038/nature12506.

Lozupone, C.A., Li, M., Campbell, T.B., Flores, S.C., Linderman, D., Gebert, M., Knight, R., Fontenot, A.P., Palmer, B.E., 2013. Alterations in the gut microbiota associated with HIV-1 infection. Cell Host Microbe 14, 329–339. http://dx.doi.org/10.1016/j.chom.2013.08.006.Alterations.

Manichanh, C., Borruel, N., Casellas, F., Guarner, F., 2012. The gut microbiota in IBD. Nat. Rev. Gastroenterol. Hepatol. 9, 599–608. http://dx.doi.org/10.1038/nrgastro.2012.152.

McGowan, I., 2012. Rectal microbicide development. curr. opin. HIV AIDS 7, 526–533.

Meulbroek, M., Ditzel, E., Saz, J., Taboada, H., Pérez, F., Pérez, A., Carrillo, A., Font, G., Marazzi, G., Uya, J., Cabrero, J., Ingrami, M., Marín, R., Coll, J., Pujol, F., 2013. BCN

Checkpoint, a community-based centre for men who have sex with men in Barcelona, Catalonia, Spain, shows high efficiency in HIV detection and linkage to care. HIV Med. 14, 25–28. http://dx.doi.org/10.1111/hiv.12054.

Modi, S., Collins, J., Relman, D., 2014. Antibiotics and the gut microbiota. J. Clin. Invest. 124, 4212–4218. http://dx.doi.org/10.1172/JCI72333.The.

Mutlu, E.A., Keshavarzian, A., Losurdo, J., Swanson, G., Siewe, B., Forsyth, C., French, A., Demarais, P., Sun, Y., Koenig, L., Cox, S., Engen, P., Chakradeo, P., Abbasi, R., Gorenz, A., Burns, C., Landay, A., 2014. A compositional look at the human gastrointestinal microbiome and immune activation parameters in HIV infected subjects. PLoS Pathog. 10 http://dx.doi.org/10.1371/journal.ppat.1003829.

Nowak, P., Troseid, M., Avershina, E., Barqasho, B., Neogi, U., Holm, K., Hov, J.R., Noyan, K., Vesterbacka, J., Svärd, J., Rudi, K., Sönnerborg, A., 2015. Gut microbiota diversity predicts immune status in HIV-1 infection. AIDS 29, 2409–2418.

Paiardini, Müller-Trutwin, M., 2013. HIV-associated chronic immune activation. Immunol. Rev. 254, 78–101. http://dx.doi.org/10.1111/imr.12079.

Sandler, N.G., Douek, D.C., 2012. Microbial translocation in HIV infection: causes, consequences and treatment opportunities. Nat. Rev. Microbiol. 10, 655–666. http://dx.doi.org/10.1038/nrmicro2848.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C., 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12, R60. http://dx.doi.org/10.1186/gb-2011-12-6-r60.

Sommer, F., Bäckhed, F., 2013. The gut microbiota — masters of host development and physiology. Nat. Rev. Microbiol. 11, 227–238. http://dx.doi.org/10.1038/nrmicro2974.

Tang, W.H.W., Wang, Z., Levison, B.S., Koeth, R.a., Britt, E.B., Fu, X., Wu, Y., Hazen, S.L., 2013. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. N. Engl. J. Med. 368, 1575–1584. http://dx.doi.org/10.1056/NEJMoa1109400.

Vázquez-Castellano, J.F., Serrano-Villar, S., Latorre, A., Artacho, A., Ferrus, M.L., Madrid, N., Vallejo, A., Sainz, T., Martinez-Botas, J., Ferrando-Martinez, S., Vera, M., Dronda, F., Leal, M., del Romero, J., Moreno, S., Estrada, V., Gosalbes, M.J., Moya, A., 2015. Altered metabolism of gut microbiota contributes to chronic immune activation in HIV-infected individuals. Mucosal. Immunol. 8, 760–762. http://dx.doi.org/10.1038/mi.2014.107.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., 2011. Linking long-term dietary patterns with gut microbial enterotypes. Science 334, 105–108.

Yu, G., Fadrosh, D., Ma, B., Ravel, J., Goedert, J.J., 2013. Anal microbiota profiles in HIV-positive and HIV-negative MSM. AIDS 28, 753–760. http://dx.doi.org/10.1097/QAD.0000000000000154.

Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Paul, I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D.R., Schneider, M.A., Schrotz-king, P., Tournigand, C., Nhieu, J.T. Van, Yamada, T., Zimmermann, J., 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. 10, 1–18.

CHAPTER B

Statistical Challenges for Human Microbiome Analysis

Statistical Challenges for Human Microbiome Analysis

Javier Rivera-Pinto, Carla Estany, Roger Paredes, M.Luz Calle, Marc Noguera-Julián and the MetaHIV-Pheno Study Group

Abstract DNA sequencing technologies have revolutionized microbiome studies. In this work we analyze microbiome data from an HIV study focused on the characterization of microbiome composition in HIV-1 infected patients. A 155 cohort of HIV infected and non-infected individuals is analyzed to characterize dietary and gut microbiome association in this group of patients. A penalized Dirichlet Multinomial regression model has been considered. The assumed underlying Dirichlet distribution in this modelization provides additional flexibility to the multinomial model which results in a better fit of the typically overdispersed microbiome data.

1 Introduction

Until recently, the composition and properties of the human microbiome were largely unknown, since the study was limited to in vitro cultivation of some specific microorganisms. Currently, high-throughput DNA sequencing technologies have revolutionized this field, allowing the study of the genomes of all microorganisms of a given environment. Metagenomics is the massive study of the genomes of the microorganisms and represents a breakthrough in the study of the relationship between the human microbiome and our health. The data from these studies provide valuable information about the composition and functional properties of microbial communities.

However, microbiome data analysis poses important statistical challenges. After DNA sequencing data analysis, microbiome data consists of a count matrix repre-

J. Rivera-Pinto (✉) · R. Paredes · M. Noguera-Julián
IrsiCaixa AIDS Research Institute, Barcelona, Spain
e-mail: jrivera@irsicaixa.es

C. Estany · R. Paredes
HIV Unit & Lluita contra la SIDA Foundation, Hospital Universitari Germans Trias i Pujol, Barcelona, Spain

R. Paredes
Universitat Autònoma de Barcelona, Catalonia, Spain

J. Rivera-Pinto · R. Paredes · M.Luz Calle · M. Noguera-Julián
University of Vic - Central University of Catalonia, Barcelona, Spain

E.A. Ainsbury et al. (eds.), *Extended Abstracts Fall 2015*,
Trends in Mathematics 7, DOI 10.1007/978-3-319-55639-0_8

senting the number of sequences corresponding to a specific bacterial taxa for each individual. Statistical techniques assuming the normal distribution are usually not appropriate. Instead, specific distributions for count data are required. An additional important feature of microbiome data is zero inflation (a large proportion of zero counts corresponding to taxa that are only present in some subjects) and the overdispersion in the rest of values. Since the total number of counts is not equal for every subject, there is the possibility of working with compositional data by dividing each count by the total number of counts giving the proportion that each taxa represents for each individual. In this case, appropriate methods for compositional data analysis are required.

In this work we analyze microbiome data from an HIV study focused on the characterization of microbiome composition along the different inflammatory profiles in healthy individuals and HIV-1 infected patients. HIV-linked chronic inflammation is associated with metabolic disorders, cardiovascular disease, immune senescence, premature aging and other inflammatory diseases. The role of the intestinal microbiome in these inflammatory processes has shown to be relevant. Interestingly, HIV infection clinical course, even when treated, is accompanied by an increase in gut permeability, bacterial translocation and low-level chronic inflammation. However, the precise effects of HIV-1 and related factor on the human gut microbiome are not well understood. It has been shown that diet has an important effect on gut microbiome composition; see [2, 6]. Therefore, it was important to characterize dietary-gut microbiome associations in this cohort. Available information was obtained from IrsiCaixa retrovirology laboratory, where microbiome and dietary information was collected from healthy and HIV infected patients showing different immune and inflammatory profiles and clinical outcomes.

First results of this project have been published in Noguera-Julián et al. [4].

2 Methods

Microbiome information was derived from 16s gene next generation sequencing from fecal samples of 155 subjects. Each one of them fulfilled both a nutrient and food portion independent diet questionnaires whose information was standardized (see Willet–Howe–Kushi [5]) to have total energy intake into account and apply the analysis over energy-relative information and not over raw data which could lead to erroneous conclusions. The standardization was made taking the residuals of a linear regression over total energy intake as new variable values.

The analysis of dietary-gut microbiome associations involves multivariate multiple regression between two matrices: $\mathbf{X}$, of size $n \times p$, and $\mathbf{Y}$, of size $n \times q$. Matrix $\mathbf{X}$ contains dietary information for p different nutrient and $\mathbf{Y}$ the microbiome abundance (*count data*) for q bacterial taxa, being n the total number of individuals.

The previously proposed penalized Dirichlet-Multinomial (DM) regression model (see [3]) was used to analyze the associations. This regression model addresses the overdispersion present in microbiome data by considering the DM distribution, with density function

$$f_{DM}(y_1, y_2, \ldots, y_q; \gamma) = \binom{y_+}{y} \frac{\Gamma(y_+ + 1)\Gamma(y_+)}{\Gamma(y_+ + \gamma_+)} \prod_{j=1}^{q} \frac{\Gamma(y_j + \gamma_j)}{\Gamma(\gamma_j)\Gamma(y_j + 1)}, \tag{1}$$

where $(y_1, \ldots, y_q)$ represents the counts for each genus, $y_+ = \sum_{j=1}^{q} y_j$, $\gamma = (\gamma_1, \ldots, \gamma_q)$ are parameters associated with the mean and variance of each genus, and $\gamma_+ = \sum_{j=1}^{q} \gamma_j$ is controling the degree of overdispersion, with a larger value indicating less overdispersion. In this modelization, the counts of the different taxa are assumed to follow a Dirichlet Multinomial distribution (see [1, 3]), which corresponds to a multinomial distribution

$$f_M(y_1, y_2, \ldots, y_q, \pi) = \binom{y_+}{y} \prod_{j=1}^{q} \pi_j^{y_j}, \tag{2}$$

with random underlying probability vectors following a Dirichlet distribution

$$f_D(\pi_1, \pi_2, \ldots, \pi_q; \gamma) = \frac{\Gamma(\gamma_+)}{\prod_{j=1}^{q} \Gamma(\gamma_j)} \prod_{j=1}^{q} \pi_j^{\gamma_j - 1}, \tag{3}$$

where $\pi = (\pi_1, \ldots, \pi_q)$ are the probabilities for a certain count to belong to the corresponding genus ($\sum_{i=1}^{q} \pi_i = 1$).

Penalized maximum likelihood estimation jointly performs model fitting and variable selection. As a result, the algorithm returns a matrix $\mathbf{C}$ of size $p \times q$, where c_{ij} represents the association between the i-th nutrient and the j-th genus. The penalization used in DM-regression assings zeroes to some coefficients selecting only the strongest associations.

3 Results

DM-regression provides the strongest associations between nutritional and genus composition information as a first step for deeper analysis. In the analyzed cohort, both *Prevotella* and *Bacteroides* are the genus with the strongest associations with nutrition parameters but in an inverse way. *Prevotella* is positively linked specially with *water* and *iron* and negatively associated with *saturated fat*. In the other hand, *Bacteroides* is negatively associated with *water* and *iron* (Fig. 1).

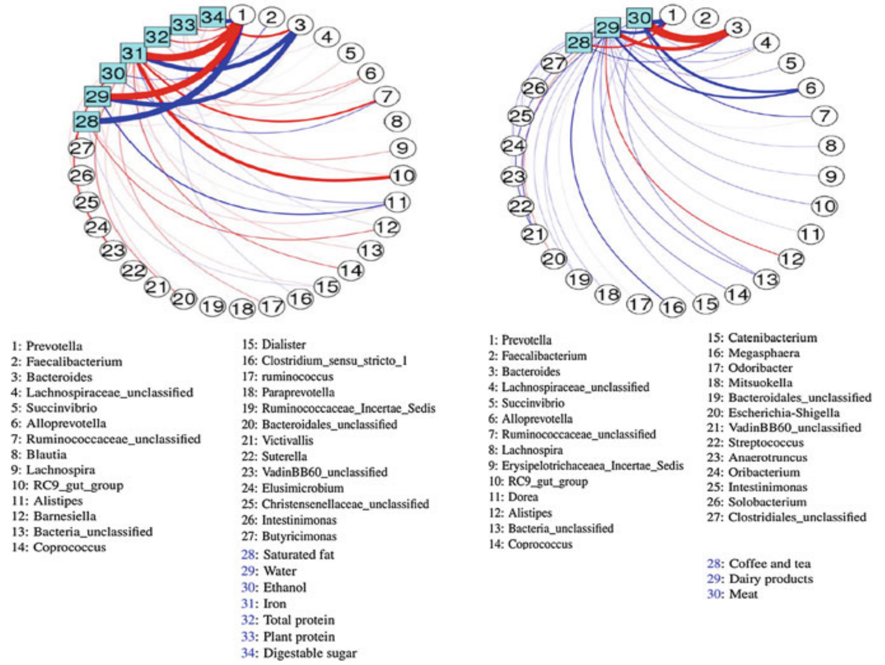


Fig. 1 Results with DM-regression model after penalization both for Nutrients (*left*) and Portions (*right*). *Red* lines represent positive relationship, while *blues* negative associations

4 Conclusions

DM-regression model allows to link two multivariate data matrices, one of them a count matrix. In this analysis those matrices where composed by genus counts after 16s rRNA sequencing and by the nutritional information of the individuals. DM distribution over the counts, has the overdispersion into account and links better with the nature of the data. In the other hand, the penalization included in the regression model selects only the strongest associations between genus and nutrients, allowing to the user to get more interpretable results.

Acknowledgements This study was mainly funded through philanthropy and private donations, which had no influence on its contents. Funds were obtained from a personal donation from Mr. Rafael Punter, the “Gala contra la SIDA” 2013 and 2014 editions, the “Nit per la Recerca a la Catalunya Central” 2015 edition, and by grant MTM2012-38067-C02-02 from the Ministerio de Economía e Innovación (Spain).

The first author is supported through a grant for doctoral studies from Noel Alimentaria to the University of Vic (UVic-UCC).

Other members of MetaHIV-Pheno Study Group were supported trough FI-DGR grant (FI-B00184) from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) at the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya.

References

1. J. Chen and H. Li, "Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis", *Annals of Applied Statistics* **7**(1) (2013), 418–442.
2. L. David *et al.*, "Diet rapidly and reproducibly alters the human gut microbiome", *Nature* **505** (2014), 559–563.
3. P.S. La Rosa, J.P. Brooks, E. Deych, E.L. Boone, D.J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W.D. Shannon, "Hypothesis testing and power calculations for taxonomic-based human microbiome data", *PLoS One* **7**(12) (2012), e52078.
4. Noguera-Julián *et al.*, "Gut microbiota linked to sexual preference and HIV infection", *Ebiomedicine* **5** (2016), 135–146.
5. W.C. Willet, G.R. Howe, and L.H. Kushi, "Adjustment for total energy intake in epidemiologic studies", *Am. J. Clin. Nutr.* **65**, (1997), 1220S–1228S; discussion 1229S–1231S.
6. G.D. Wu *et al.*, "Linking long-term dietary patterns with gut microbial enterotypes", *Science* **334** (2011), 105–108.

CHAPTER C

Richer gut microbiota with distinct metabolic profile in HIV infected Elite Controllers

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Richer gut microbiota with distinct metabolic profile in HIV infected Elite Controllers

Received: 26 January 2017
Accepted: 15 June 2017
Published online: 24 July 2017

Jan Vesterbacka[1], Javier Rivera[2], Kajsa Noyan[3], Mariona Parera[2], Ujjwal Neogi[3], Malu Calle[4], Roger Paredes[2,4,5,6], Anders Sönnerborg[1,3], Marc Noguera-Julian[2,4] & Piotr Nowak[1]

Gut microbiota dysbiosis features progressive HIV infection and is a potential target for intervention. Herein, we explored the microbiome of 16 elite controllers (EC), 32 antiretroviral therapy naive progressors and 16 HIV negative controls. We found that the number of observed genera and richness indices in fecal microbiota were significantly higher in EC versus naive. Genera *Succinivibrio*, *Sutterella*, *Rhizobium*, *Delftia*, *Anaerofilum* and *Oscillospira* were more abundant in EC, whereas *Blautia* and *Anaerostipes* were depleted. Additionally, carbohydrate metabolism and secondary bile acid synthesis pathway related genes were less represented in EC. Conversely, fatty acid metabolism, PPAR-signalling and lipid biosynthesis proteins pathways were enriched in EC vs naive. The kynurenine pathway of tryptophan metabolism was altered during progressive HIV infection, and inversely associated with microbiota richness. In conclusion, EC have richer gut microbiota than untreated HIV patients, with unique bacterial signatures and a distinct metabolic profile which may contribute to control of HIV.

Progressive HIV-1 infection is characterized by depletion of CD4+ T cells in gut-associated lymphoid tissue, followed by immune activation, gut microbiota dysbiosis, and microbial translocation[1–3]. Elite controllers (EC) constitute less than 1% of the HIV-infected population[4], and have sustained viral suppression in absence of antiretroviral therapy (ART). Due to definitional bias, a high rate of heterogeneity is observed among EC cohorts[5]. It appears that host genetic rather than demographic factors contribute to the viral controlling properties, e.g. with an increased rate of HLA B*5701 allele positivity. Also, unique immunological cellular responses against HIV-1 have been proposed as a mechanism for viral control[6]. Despite spontaneous suppressed plasma viremia, microbial translocation and immune activation are present in EC[7].

The gut microbial composition in EC has not been extensively explored, with only three studies (with low numbers of subjects) investigating their gut microbiome[8–11]. In a previous work, differences in the bacterial composition of gut microbiota between ART naive HIV patients and EC were observed at the phylum level, with an enrichment of Bacteroidetes and a reduction of Actinobacteria in EC[9]. It was also found that the EC had lower beta-diversity (i.e. inter-individual variation in the gut microbiota) than the viremic patients, and principal coordinate analysis (PCoA) revealed that EC clustered separately, indicating a different gut microbiome compared to other HIV-infected individuals.

Use of metagenomic techniques has illuminated the complex interactions between the host metabolic activities and gut microbial species in several diseases[12]. Thus, alterations in the catabolism of tryptophan have been linked to progressive HIV-infection, and correlated with a pathological shift in the gut microbiota[10]. In depth, tryptophan degradation products have been linked to loss of Th17/regulatory T cell balance fueling the chronic inflammation in progressive HIV disease[13]. Whether the gut microbiota in EC differently influences the tryptophan metabolism has not been explored, but markers of tryptophan catabolism were not elevated in EC as compared to healthy subjects[11].

[1]Department of Medicine Huddinge, Unit of Infectious Diseases, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden. [2]IrsiCaixa & AIDS Unit, Hospital Universitari Germans Trias i Pujol, Universitat Autònoma de Barcelona, Badalona, Spain. [3]Department of Laboratory Medicine, Division of Clinical Microbiology, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden. [4]Universitat de Vic-Universitat Central de Catalunya, Catalonia, Spain. [5]Universitat Autònoma de Barcelona, 08193, Bellaterra, Catalonia, Spain. [6]HIV Unit & Lluita Contra la SIDA Foundation, Hospital Universitari Germans Trias i Pujol, Ctra de Canyet s/n, 08916, Badalona, Catalonia, Spain. Jan Vesterbacka and Javier Rivera contributed equally to this work. Correspondence and requests for materials should be addressed to J.V. (email: jan.vesterbacka@sll.se)

	EC	Naive	Negative	p-value
Number of individuals	16	32	16	
Age (years, median (IQR))*	47 (40.3–54.3)	43.5 (37.3–50.5)	49 (44–52.8)	ns
Gender (n, male/female)†	9/7	16/16	8/8	ns
Ethnicity (n)				
Black	9	13	0	
Caucasian	6	17	15	
Latin	1	1	0	
Oriental	0	1	1	
Mode of transmission (n)				
Heterosexually	8	21	NA	
MSM	4	8		
IVDU	1	3		
Blood transfusion	2	0		
Unknown	1	0		
Sexual practice†				
Heterosex	12	24	12	ns
MSM	4	8	4	
Time since diagnosis (years, median (IQR))*	8.55 (5.0–18.0)	3 (0.7–6.9)	NA	0.0008
Body Mass Index (BMI) (score (IQR))*	26.4 (24.1–32.2)	25 (23.0–30.0)	24.2 (22.9–25.6)	ns
CD4+ T-cell count (median (IQR)*	806 (676–1049)	390 (298–475)	NA	<0.0001
CD8+ T-cell count (median (IQR)*	705 (541–904)	995 (678–1373)	NA	0.02
CD4/CD8+ T-cell ratio (median (IQR)*	1.41 (0.74–1.55)	0.38 (0.27–0.51)	NA	<0.0001
CD4+ T-reg cells (FoxP3+CD25+) % (median (IQR)*	4.91 (4.13–5.66)	NA	5.88 (4.77–7.28)	ns
CD4+ HLA-DR+ CD38+ T cells % (median (IQR))*	0.44 (0.34–0.75)	7.78 (5.34–14.8)	0.53 (0.39–0.61)	<0.0001
CD8+ HLA-DR+ CD38+ T cells % (median (IQR))*	1.16 (0.77–1.6)	36.9 (23.6–44.9)	0.71 (0.52–1.71)	<0.0001

Table 1. Cohort demographics and cellular immune activation markers at baseline. EC = elite controllers. Naive = viral progressors. Negative = negative controls. *Kruskal-Wallis test was used for comparison between three groups, and Dunn's Multiple Comparison Test was adapted for "post hoc" testing. Mann-Whitney was applied for comparisons between two groups. †Chi-square test was applied. NA (not available). ns (non significant) indicates p-value > 0.05.

In the current work, we investigated if HIV infection differently affects the gut microbiome in patients with progressive HIV infection and EC. We also explored the link between the composition, inferred functionality of gut microbiome and systemic inflammatory, immunological and metabolic markers in these patients.

Material and Methods

Study design. This was a cross-sectional study including both HIV seropositive and seronegative participants.

Patients. Detailed characteristics are presented in Table 1. Totally, 48 study subjects were recruited from the out-patient HIV clinic at Karolinska University Hospital, Stockholm, Sweden. Additionally we included 16 HIV negative controls (negative). Inclusion criteria were age >18 years, HIV positive for at least 6 months and no ongoing HIV-related complications. All viremic progressors had to be ART naive (naive). Exclusion criteria were inflammatory bowel disease or infectious gastroenteritis within the last four weeks. EC were defined by: (I) HIV positive for ≥1 year and with ≥3 consecutive viral loads (VLs) <75 c/ml over one year with all previous VLs < 1000 c/m, or (II) HIV positive for ≥10 years, with ≥2 VLs and ≥90% of all VLs < 400 c/ml. Four female EC had been on short time ART due to pregnancy (three for 3.5 months, one for 14 days), all more than four years before study entry. The study subjects were categorized into three groups (EC: n = 16; naive: n = 32; negative: n = 16) and were matched by Body Mass Index (BMI), age, gender and sexual practice. All participants gave written informed consent. All the work and experiments were performed in accordance with relevant guidelines, regulations and with the Declaration of Helsinki. The study was approved by the Regional Ethics Committee at Karolinska University Hospital, Stockholm (2009/1485-31, 2013/1944-31/4, 2014/920-3).

Blood Sample Collection and Isolation of Peripheral Blood Mononuclear Cells. Plasma, isolated from EDTA-treated peripheral blood, and serum samples were stored at −80 °C until analyses. Peripheral blood mononuclear cells (PBMCs) were isolated from EDTA-treated blood using Hypaque-Ficoll (GE Healthcare) density gradient centrifugation, counted with Nucleocounter® and then cryopreserved at −150 °C in fetal bovine serum (Sigma-Aldrich) containing 10% DMSO (Sigma-Aldrich), at a concentration of 10^6 cells/ml of cryopreservation media. Soluble markers of inflammation and microbial translocation, and metabolites of tryptophan catabolism pathway were analyzed in plasma by ELISA (hs-CRP (Abcam, UK), sCD14 (R&D, Minnesota, USA), IL-6 (R&D), LBP (Hycult Biotech, The Netherlands)) or HPLC (http://bevital.no), respectively, according to manufacturer's instructions.

Flow Cytometry, Immunophenotyping, and Viral Load. Quantification of CD4+ and CD8+ T-cells and plasma HIV-1 RNA were performed as part of the clinical routine with flow cytometry and Cobas Amplicor (Roche Molecular Systems Inc., Branchburg, New Jersey, USA), respectively. At the day of analysis, cryopreserved PBMCs were thawed and stained for HLA-DR and CD38 as markers of immune activation of CD4+ and CD8+ T- cells, and FoxP3 and CD25 as markers of CD4+ T-regulatory cells[14]. HIV negative samples were not analyzed by routine flow cytometry, which is mirrored by the lack of CD4+ and CD8+ T-cell total counts in that group (Table 1).

Fecal Sample Collection. A sterile tube for fecal sampling without preservation media was used when participants were able to donate feces adjacent to their study visit at the clinic. The sample was frozen and stored at −80 °C within 24 hours. PSP® Spin Stool DNA sampling tube (Stratec Biomedical) was used for participants who submitted feces at home. The stool samples were delivered to the out-patient clinic by the participant, or instantly sent by post and stored at −70 °C according to the manufacturer's instructions[15]. All participants were asked to complete a standardized questionnaire, collecting data about recent use of antibiotics (last 3 months) and probiotics, current medication, alcohol use, smoking, chronic diseases, recent infectious gastroenteritis (last 4 weeks), special diet (vegan/vegetarian/gluten-/lactose- free), colectomy, recent travelling abroad (>4 weeks last 12 months) and time since arrival in Sweden for non-natives.

DNA extraction, 16s rRNA gene amplification and Sequencing. DNA extraction was performed using the PowerSoil DNA Extraction Kit (MO BIO Laboratories, Carlsbad, CA, US). To amplify the variable region V3-V4 from the 16S rRNA gene (amplicon size expected ~460 bp), we used the primer pair described in the MiSeq rRNA Amplicon Sequencing protocol which already have the Illumina adapter overhang nucleotide sequences added to the 16S rRNA V3-V4 specific-primers, i.e.: 16S_F 5′-(TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG **CCT ACG GGN GGC WGC AG**)-3′ and 16S_R 5′-(GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G**GA CTA CHV GGG TAT CTA ATC C**)-3′.

Amplifications were performed in triplicate 25 µL reactions, each containing 2.5 µL of non-diluted DNA template, 12.5 µL of KAPA HiFi HotStart Ready Mix (containing KAPA HiFi HotStart DNA Polymerase, buffer, $MgCl_2$, and dNTPs, KAPA Biosystems Inc., Wilmington, MA, USA), and 5 µL of each primer at 1 µM. Thermal cycling conditions consisted of an initial denaturation step (3 min at 95 °C), followed by 30 cycles of denaturation (30 sec at 95 °C), annealing (30 sec at 55 °C) and extension (30 sec at 72 °C). These were followed by a final extension step of 10 min at 72 °C. Once the desired amplicon was confirmed in 1% agarose gel electrophoresis, all three replicates were pooled and stored at −30 °C until sequencing library preparation. Amplified DNA templates were cleaned-up for non-DNA molecules and Illumina sequencing adapters and dual indices were attached using Nextera XT Index Kit (Illumina, Inc.) followed by the corresponding PCR amplification program as described in the MiSeq 16S rRNA Amplicon Sequencing protocol. After a second round of cleanup, amplicons were quantified using Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen, Carlsbad, MA, USA) and diluted in equimolar concentrations (4 nM) for further pooling. Sequencing was performed on an Illumina MiSeq™ platform according to the manufacturer's specifications to generate paired-end reads of 300 base-length in each direction.

Data Analysis. Sequencing data was processed using Mothur[16] phylotype approach. Briefly, paired-end data were merged and quality filtered and all reads not matching the used V3-V4 amplicon design were discarded. Chimeric sequences were filtered using Mothur Uchime[17] implementation. Sequences were classified using RDP algorithm[18] in combination with 16s rRNA Silva database[19]. Obtained sequences from five subjects (one EC, three naive and one negative) were of poor quality and were excluded from further analyses. To assess alpha diversity, richness (Chao1 and ACE) and diversity (Shannon and Simpson) indices were computed using R/vegan library[20,21] selecting a subsample of ten thousand counts for each individual.

Bacterial genera count table were normalized to relative abundance measures. These were used to compute Bray – Curtis[22] dissimilarity between each pair of individuals, which was used as input ordination analysis using non-metric multidimensional scaling (NMDS). Correlation between NMDS plot axis coordinates and inflammation parameters were tested by applying Spearman test. Additionally, a PERMANOVA (adonis) test was performed on this distance matrix to partition different sources of variation using R/vegan package.

Microbiome function was inferred using PICRUSt[23] on GreenGenesDB[24] classified phylotypes. Counts were normalized by considering 16s rRNA gene copy number. To infer the gene content, the normalized phylotype abundances were multiplied by the respective set of gene abundances, represented by Kyoto Encyclopedia of Genes and Genomes (KEGG) identifiers estimated for each taxon.

Statistics. Multiple group differences in diversity indices, inflammation and activation markers and bacterial abundances were analyzed via Kruskal–Wallis rank-based test. Benjamini–Hochberg[25] correction was used to correct for multiple testing. Two-tailed Mann-Whitney U-test was applied for comparisons of inflammation markers between two groups.

Inflammation indices were associated both with genus and functional composition using Spearman correlation. Associations with a Benjamini–Hochberg adjusted p-value lower than 0.01 were considered as relevant and inflammation parameters associated with less than two bacteria were discarded when plotting the heatmap. Bacterial genus and functions were ordered in the heatmap according to a clustering between them using Ward hierarchical clustering.

With the aim of evaluating the power of the classification of individuals according to their microbiome composition profile, a LASSO penalized logistic regression model as proposed in the bibliography[26] was computed for each pair of profiles. LiblineaR and pROC libraries were used to obtain the regression models, represent ROC curves and estimate model accuracy using AUC.

Data Availability. Metagenomics raw sequencing data along with sample level metadata have been deposited using the NCBI/SRA Web service and compliance to MIMARKS standard. Data can be accessed using BioProject accession number PRJNA354863.

Results

This was a cross-sectional study including 64 participants (Table 1). The groups were balanced by age, gender, sexual practice and BMI. The heterosexual transmission route was slightly more common in the naive (65.6 vs 50.0%), whilst the rate of the MSM transmission route was the same in both groups. Two naive patients had chronic hepatitis B infection, whereas two EC and two naive had chronic hepatitis C infection. Use of antibiotics within three months before inclusion was declared from 2 EC, 6 naive and 2 negative. One EC was vegetarian, one EC and one negative were on lactose/gluten-free diet (Supplementary Table 1). The median viral load (copies/mL) of EC was <20 (75% percentile 30.25), of naive 31700 (IQR 4430-100250).

Comparable T-cell activation in Elite Controllers and Negative. As expected, BL CD4+ T-cell count was lower and CD8+ T-cell count significantly higher in naive vs EC. Proportions of CD4+ T-regulatory cells tended to be higher in negative compared to EC (p = 0.07). The level of immune activation of CD4+ and CD8+ T-cells in blood (CD4/8+ T-cell ratio and by expression of HLA-DR+ CD38+) was similar in EC and negative but significantly lower compared to naive group (Table 1).

Richness, diversity and composition of fecal microbiota. Overall, the fecal microbiota was richer and more diverse in EC as compared to naives and similar to negative. Thus, the number of observed taxa in fecal microbiota was higher in EC vs naive (Δ 19.8; p = 0.0001), and not different compared to negative (Δ 8.3; p = 0.14) (Fig. 1a). Similarly, naive patients had decreased estimated richness indices Chao 1 (EC-naive: Δ 19.6; p = 0.0002, EC-negative: Δ 10.4; p = 0.07, naive-negative Δ −9.2; p = 0.007) and ACE (EC-naive: Δ 20.5; p = 0.0001, EC-negative: Δ 9.7; p = 0.09, naive-negative Δ − 10.8; p = 0.03); (Fig. 1b,c). The Shannon index was increased in negative group as compared to naive (Δ − 13.5; p = 0.01) (Fig. 1d) suggesting HIV induced changes in alpha diversity in the latter group. To further characterize the inter-individual differences between groups (beta-diversity) at group level, non-metric multidimensional scaling (NMDS) and LASSO regression analysis with ROC curve and AUC were performed. NMDS analysis revealed separation and clustering of EC along NMDS1 axis, whilst naive tended to cluster along NMDS2 (Fig. 2a). The lowest accuracy of LASSO regression was found when using microbiome composition to classify EC vs negative patients (AUC = 0.77), confirming that the gut microbiota composition was least different among these individuals. Additionally, LASSO classification was more accurate when classifying naive vs either EC (AUC = 0.88) and negative (0.87) (Fig. 2b). Furthermore, PERMANOVA (adonis) test yielded that the bacterial composition varied between the groups ($R^2 = 0.12$; p = 0.001). The groups differed significantly in abundance of 17 bacterial taxa at the genus level (Fig. 2 and Supplementary material Figure S1). We found that genera of *Succinivibrio* and *Sutterella* were enriched in EC only. Additionally, *Rhizobium, Delftia, Anaerofilum* and *Oscillospira* genera were more abundant in EC than in naive, but not significantly different from negative. Moreover, genus *Blautia* and *Anaerostipes* were enriched in naive as compared to EC and negative (Fig. 3). We also found significant differences in abundance of unclassified genera at higher taxonomic levels between the groups (Supplementary material Figure S1).

Inferred gut microbiota functionality. The PICRUSt analysis, predicting the metagenomic functional content of gut microbiota, revealed several significant differences between the groups at both KEGG level II and III. Hence, at the KEGG level II, we found that the predicted pathway of carbohydrate metabolism was significantly reduced in the gut bacterial metagenome of EC as compared to both naive and negative patients. Instead, genes encoding cardiovascular diseases and circulatory system pathways were enriched in EC as compared to naive, but were not significantly different as compared to negative (Fig. 4a). Moreover, several pathways related to the metabolism of carbohydrates were decreased in EC in relation to naive and negative at KEGG level III. Thus, galactose metabolism, pentose-glucoronate interconversions, pyruvate metabolism and pentose-phosphate pathway (PPP) were predicted to have a lower abundance in EC vs naive. PPP was significantly reduced in EC as compared to all other groups, and both galactose and PPP were significantly more abundant in naive vs negative (Fig. 4b). Pathways related to lipid metabolism were differentially distributed in the metagenome of the cohort. Those involved in metabolism of fatty acids and lipid biosynthesis proteins were significantly reduced in naive as compared to the other groups. Conversely, the essential fatty acid linoleic acid metabolism pathway was more represented in naive. The metagenomic proportion of secondary bile acid biosynthesis metabolism pathway, which has a key function in cholesterol homeostasis, was significantly reduced in EC, but present at similar level in naive and negative (Fig. 4c). We also found that the PPAR (peroxisome proliferator-activated receptors)-signaling pathway, which plays an essential role in metabolism of carbohydrates, lipids and proteins, was significantly reduced in naive. Additionally, pathways related to synthesis and degradation of ketone bodies were reduced in naive, whereas significantly enriched in EC, also when compared to negative. Tryptophan metabolism related genes were decreased in naive vs negative. In contrast, proportions of phenylalanine, tyrosine and tryptophan biosynthesis pathway were enriched in naive (Fig. 4d). Additional functional pathways with different distribution in the cohort are presented in supplementary material (Figure S2).

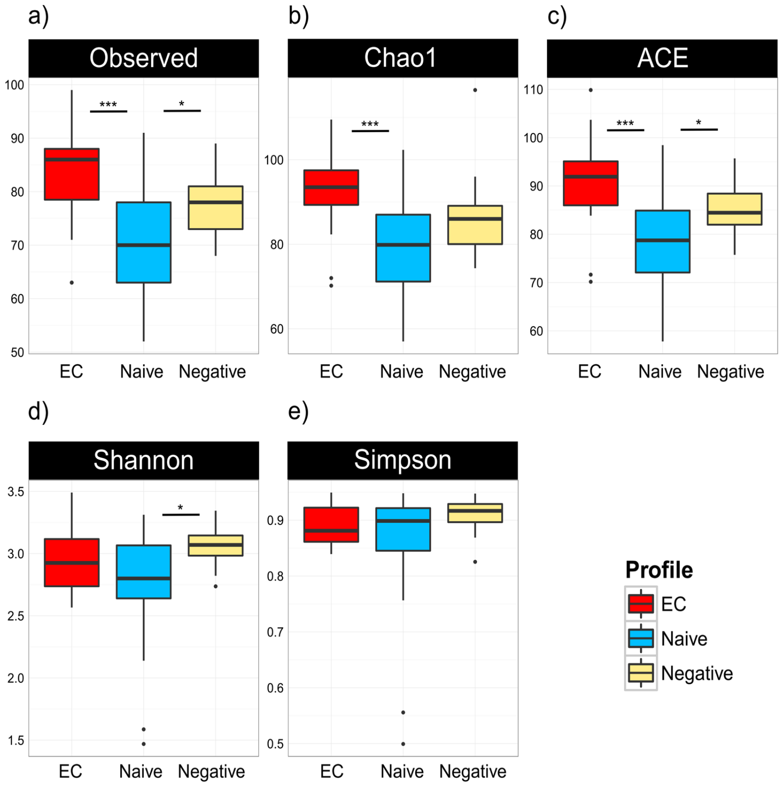


Figure 1. Similar richness and diversity of fecal microbiota in EC and negative controls. Number of observed bacterial genera was significantly lower in naive patients as compared to the other groups (**a**). Richness indices Chao-1 (**b**) and ACE (**c**) were reduced in naive, but no significant differences were observed between EC and negative. Alpha-diversity, assessed by Shannon index was lower in naive as compared to negative (**d**), whereas Simpson index was similar in all groups (**e**). Comparisons between groups were obtained via Kruskal-Wallis rank based test including Dunn's post-hoc pairwise analyses. Benjamini-Hochberg method was used for correction of multiple testing. A p-value < 0.05 was considered significant. Box plots represent median (black horizontal line), 25th and 75th quartiles (edge of boxes), upper and lower extremes (whiskers). Outliers are represented by a single data point.

Plasma levels of soluble markers of inflammation and tryptophan catabolism metabolites. Plasma levels of soluble markers of inflammation, immune activation and metabolites related to the kynurenine pathway of tryptophan degradation are presented in Table 2. We found that EC had higher levels of IL-6 and hs-CRP than negative; however levels of soluble immune activation marker sCD14 were not different among groups. Levels of LBP, commonly used as a marker of microbial translocation, were significantly increased in naive group as compared to others.

Tryptophan levels in plasma were reduced in naive as compared to both EC and negative. Additionally, the naive group had several divergent levels of metabolites. Thus, xanthurenic and kynurenic acid levels were lower in naive as compared to negative; in contrary anthralinic acid levels and kynurenine/tryptophan (K/T)-ratio were increased in naive vs EC/negative. K/T-ratio was correlated to the number of observed genera ($r = -0.47$, $p = 0.0009$), richness indices: Chao-1 ($r = -0.53$, $p = 0.0002$) and ACE ($r = -0.44$, $p = 0.002$), but not to alpha-diversity indices. Significant correlations between levels of tryptophan, xanthurenic acid, K/T-ratio and NMDS2 axis were found (Table 3), mirroring a separation of naive from EC and negative in this axis (Fig. 5).

Factors associated with the composition and functionality of gut microbiota. We observed a distinct pattern of correlations between gut microbial composition, immunological markers and tryptophan catabolism (Fig. 6a). Interestingly, nadir and BL CD4+ T-cell counts, CD4/8+ T-cell ratio and tryptophan levels were strongly correlated to the abundance of genus *Sutterella*, whilst BL CD4+ correlated to *Rhizobium* and *Butyricimonas*. Moreover, CD4/8+ T-cell ratio was positively correlated to *Oscillopira* and *Butyricimonas*.

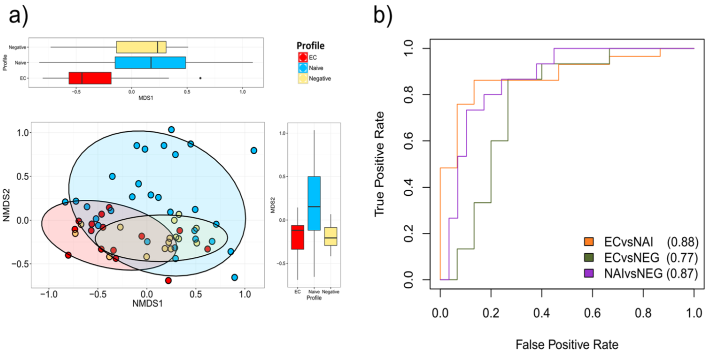


Figure 2. Separation between EC and naive patients in inter-individual (ß-diversity) analyses. Non-metric multidimensional scaling (NMDS) analysis was performed to characterize inter-individual differences between groups, revealing clustering of EC at NMDS axis 1 and naive at axis 2. The separations between groups at each axis are presented in respective box-plot. Box plots represent median (black horizontal line), 25th and 75th quartiles (edge of boxes), upper and lower extremes (whiskers). Outliers are represented by a single data point (**a**). LASSO regression analysis with AUROC (ROC curves; AUC used for estimation of model accuracy) curve was used for classification of gut microbiota composition between groups, and lowest accuracy was found between EC and negative patients (AUC 0.77, suggesting that the similarity of microbiota composition was highest between these groups) (**b**).

Genera of *Sutterella*, *Oscillospira*, *Rhizobium*, *Anaerofilum*, *Alistipes*, *Anaerotruncus* and *Odirobacter* had all at least two inverse correlations with some of the cellular immune activation markers (CD38, HLA-DR). In contrary, abundance of *Blautia* was positively associated with immune activation (CD4+ CD38+, CD8+ CD38+, CD4+ CD38+ HLA-DR+ and VL). Additionally, unclassified genera of Burkholderiales, Bacteriodales, Proteobacteria, Betaproteobacteria and Rhizobiaceae were also positively correlated to BL CD4+ T-cell count. Inversely, there was a strong negative correlation between all of these taxa, unclassified genera of family Porphyromonadaceae, and most of the cellular immune activation markers. Only one of the identified genera, *Rhizobium*, was significantly inversely associated with K/T-ratio (Fig. 6a). There was an inverse correlation between BL CD4+ T-cell count, CD4/8+ T-cell ratio and several pathways related to carbohydrate metabolism, as also the essential omega-6 fatty acid linoleic acid. Conversely, alpha-linoleic acid (an essential n−3 fatty acid) metabolism was negatively associated to these markers. Furthermore, positive correlations were found between BL CD4+ T-cell count and synthesis and degradation of ketone bodies and lipid biosynthesis proteins pathways, both involved in lipid metabolism (Fig. 6b). CD4/8+ T-cell ratio was positively correlated to degradation of amino acids valine, leucine and isoleucine.

Cellular immune activation correlated with several pathways. Proportions of CD4/8+ (CD38+) T-cells were positively associated to carbohydrate metabolism, pentose-phosphate pathway (PPP) and also to overall metabolism of lipids and linoleic acid, whereas both fatty acid and alpha-linoleic acid metabolism were negatively correlated. Further inverse correlations were found between cellular immune activation and pathways involved in PPAR-signaling, steroid biosynthesis, adipocytokine signaling, citrate (TCA) cycle, degradation of amino acids, diabetes mellitus type I and tryptophan metabolism. Most of these associations were both significant for CD4+ and CD8+ (HLA-DR+/CD38+) T-cells. Only a few associations between soluble plasma markers sCD14 and LBP and microbiota function were found at the significance level of 0.01 (Fig. 6b), though additional correlations were observed at level 0.05 (Supplementary material Figure S3).

Discussion

It has been widely accepted that HIV-infection is accompanied by immune activation, microbial translocation[2, 27–31] and gut microbiota dysbiosis[8–10, 32]. Our study provides important observations concerning these pathogenic events in patients who spontaneously maintain sustained control of HIV, the elite controllers (EC). Thus, we present that their microbiota is richer and differs in predicted functionality from treatment naive HIV progressors, resembling the microbiota of HIV negative controls. We also confirm that the level of systemic immune activation and plasma markers of tryptophan catabolism pathway in EC are similar to uninfected individuals. Additionally we show that the microbiota richness is inversely correlated to K/T-ratio, a surrogate marker of IDO-1 activity, the rate limiting enzyme of systemic tryptophan catabolism.

To date, the mechanisms behind the viral control in EC are not fully understood. It has been postulated that more potent HIV-specific CD8+ T-cell responses, expression of restriction factors like APOBEC3 family proteins and enrichment of specific NK-cell receptors contribute to this persistent control of HIV[33]. Even if these individuals can suppress the virus, microbial translocation and chronic immune activation still feature the course of HIV-infection also in EC[7].

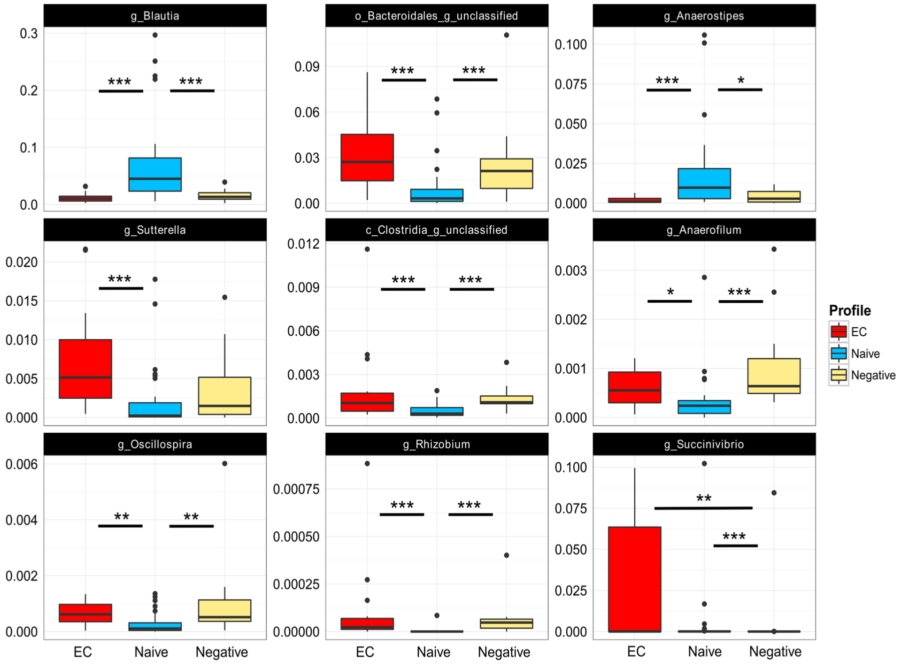


Figure 3. Compositional differences in fecal microbiota between groups. Several differences in bacterial abundance were observed between the groups at genus level. Comparisons of taxa abundances were performed via Kruskal -Wallis rank based test and Benjamini-Hochberg method was used for correction of multiple testing. Adjusted p-value < 0.01 was considered significant for Kruskal-Wallis. Dunn's post-hoc pairwise analyses: $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Box plots represent median (black horizontal line), 25th and 75th quartiles (edge of boxes), upper and lower extremes (whiskers). Outliers are represented by a single data point.

The dysbiosis in progressive HIV infection has been described in several studies[30, 34, 35]. Albeit, even if only handful of EC has been included in these cohorts[8–10], their microbiome diversity and composition have differed from HIV progressors. Our current study, which included the so far highest number of EC, confirms and expands the previous observations. We found that several ecological indices of EC microbiota (including richness and number of observed species) were significantly higher in EC as compared to naive and not different from matched negative controls. Additionally LASSO analysis showed a higher similarity between the microbiota of EC and negatives than that of viremic HIV infected individuals. Furthermore, we found that EC had a unique bacterial signature at genus level with 17 genera that were significantly differently distributed between the groups. Hence, *Succinivibrio*, *Sutterella*, *Rhizobium*, *Delftia*, *Anaerofilum* and *Oscillospira* were more abundant, whereas *Blautia* and *Anaerostipes* were depleted in EC.

In a previous work, initiation of ART was followed by higher abundance of *Succinivibrio*[9]. Interestingly, the metabolic properties of Succinivibrionaceae family members have been associated with ART related immune recovery[36]. The study suggested that bacteria of this family have anti-inflammatory capacity by accumulating molecules involved in reduction of viral infections and inflammation.

Members of *Sutterella* genus are prevalent commensals in the GI-tract with mild pro-inflammatory capacity, except for *Sutterella wadsworthensis* whose pathogenic properties have been described recently[37]. The authors proposed that members of *Sutterella* may have different immunomodulatory roles, as *Sutterella* spp. except from *S. wadsworthensis* may elicit T_H-17 differentiation by adhering to intestinal epithelial cells. Additionally, lower abundance of *Sutterella* has been found in the gut microbiome of patients with multiple sclerosis, and in Hodgkin lymphoma patients after allogenic hematopoetic stem cell transplantation[38, 39]. In our study, we present increased abundance of *Sutterella* in EC with several correlations to immune markers (positive with BL CD4+ T-cell counts and negative to markers of cellular activation). Thus, our findings warrant further characterization of *Sutterella* genus at species level to determine its involvement in the modulation of the immune system.

Similar to us, Mutlu *et al*. found decreased abundance of *Oscillospira* in HIV positive patients with progressive infection[32]. The strong positive correlation between the *Oscillisipira* and CD4/CD8 ratio suggests that this genus was associated with lower systemic inflammation in our cohort, which has also been shown in patients with Crohn disease and obesity[40].

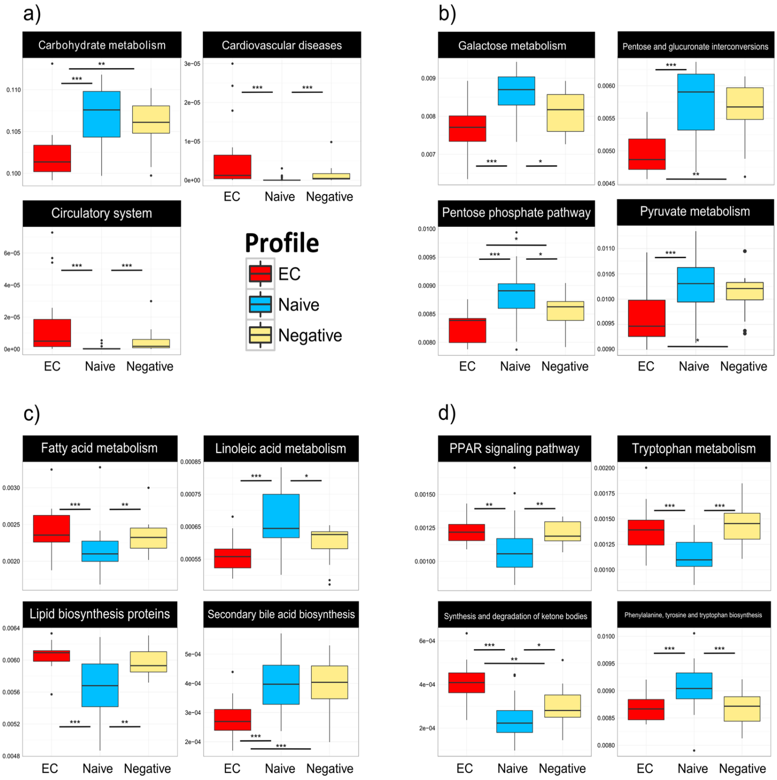


Figure 4. Inferred functional content of gut microbiota. The metagenomic functional content of gut microbiota was predicted by inferred PICRUSt analysis. Abundance of pathways involved in carbohydrate metabolism, cardiovascular diseases and circulatory system at KEGG level II (**a**), or level III (**b**–**d**). Pathways involved in carbohydrate metabolism, galactose metabolism, pentose and glucoronate interconversions, pentose-phosphate pathway and pyrovate metabolism (**b**). Pathways related to metabolism of lipids and fatty acids and biosynthesis of secondary bile acids (**c**). Bacterial tryptophan metabolism, PPAR signaling, phenylalanine, tyrosine and tryptophan biosynthesis and synthesis and degradation of ketone bodies pathways (**d**). Kruskal – Wallis rank-based test was applied, and Benjamini – Hochberg method was used to correct for multiple testing. Adjusted p-value < 0.01 was considered significant for Kruskal-Wallis. Dunn's post-hoc pairwise analyses: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Box plots represent median (black horizontal line), 25th and 75th quartiles (edge of boxes), upper and lower extremes (whiskers). Outliers are represented by a single data point.

Conversely, depletion of *Blautia* and *Anaerostipes* has been described in patients with HIV infection[31, 32], but instead we now report enrichment of *Blautia* and *Anaerostipes* in naive patients, linked to cellular immune activation. Additionally, we found increased abundance of genus *Rhizobium* in EC, with positive correlation to BL-CD4 counts and inverse to viral load, cellular immune activation markers and K/T-ratio. This bacterium, belonging to phylum Proteobacteria, has been attributed to nitrogen fixing properties in plants[41]. Interestingly, the K/T-ratio correlated only with genus *Rhizobium*, suggesting that this particular taxa may play a role in the bacterial metabolism of tryptophan. Moreover, similar to a previous study[42], we found that the proportions of tryptophan metabolism related bacterial genes were depleted in naive as compared to both negative and EC. This probably reflects the loss of intraluminal commensal bacteria involved in tryptophan catabolism, like *Lactobacillus* spp[43]. Based on our results, we speculate that *Rhizobium* genus may be a factor orchestrating tryptophan degradation as *Rhizobium* members are able to convert tryptophan to indole-3-acetic acid[44]. The reduced ability of gut microbiota to produce tryptophan derived indole metabolites related to dysbiosis in progressive HIV-infection is known to affect the production of IL-22 by innate lymphoid cells which together with loss of T_H-17 cells increase the disruption of the epithelial barrier and exacerbate overgrowth of pathogenic bacteria[43, 45, 46]. These events in the gut were mirrored by signs of increased microbial translocation and immune activation in naive group.

The metabolism of tryptophan along the kynurenine pathway in peripheral tissues (including skeletal muscle, liver and white blood cells) is mediated by several enzymes, but the main inducible and rate-limiting enzyme is Indolamine-2,3-Dioxygenase 1 (IDO-1)[47]. During HIV-infection, the IDO-1 activity is induced in dendritic cells

	EC	Naive	Negative	p-value*
Soluble marker: median(IQR)				
LBP (ng/ml)	3727 (2206–15123)	6805 (5984–8031)	2862 (1956–3705)	0.0004
sCD14 (pg/ml)	1.7×10^6 ($1.53–1.95 \times 10^6$)	1.47×10^6 ($1.46–1.71 \times 10^6$)	1.5×10^6 ($1.44–1.65 \times 10^6$)	ns
IL-6 (pg/ml)	1.73 (1.18–3.20)	NA	0.84 (0.67–1.78)	0.035
hs-CRP (pg/ml)	1.37×10^6 ($0.76–2.7 \times 10^6$)	NA	635419 (378718–941463)	0.005
Tryptophan catabolism:				
Tryptophan (umol/L)	53.1 (51.4–60.5)	46.2 (40.5–50.8)	66.1 (60.3–73.1)	<0.0001
Kynurenine (umol/L)	1.4 (1.3–1.6)	1.65 (1.3–2)	1.7 (1.4–1.9)	ns
Anthralinic acid (nmol/L)	12.2 (10.1–16.2)	21.6 (16.3–28.8)	15.3 (11.6–20.6)	0.0007
Kynurenic acid (nmol/L)	40.5 (36.2–50.6)	30.3 (17.6–48.9)	54.2 (48.9–70.2)	0.0015
3-Hydroxykunrenin (nmol/L)	45.9 (33.8–59.6)	35.1 (29.5–48.7)	41.9 (33.1–53.1)	ns
Xanthurenic acid (nmol/L)	11.8 (8.4–19.5)	9.2 (3.4–14.5)	19.9 (15–27.8)	0.0013
3-Hydroxyantralinic acid (nmol/L)	27.2 (22.5–35.5)	31.6 (20–47.7)	31.5 (23.4–41.1)	ns
Quinilonic acid (nmol/L)	349 (262.3–448.5)	474.4 (352.2–669.2)	359 (304–425)	0.039
K/T ratio	24.8 (21.2–30.8)	34.8 (31.2–46.9)	24.6 (20.8–28.9)	0.0001

Table 2. Soluble markers of inflammation and metabolites of kynurenine/tryptophan catabolism in plasma. *Kruskal-Wallis test was used for comparison between three groups, and Dunn's Multiple Comparison Test for post-hoc pairwise analyses. Two-tailed Mann-Whitney U-test was applied for comparisons between two groups. NA (not available). ns (non significant) indicates p-value > 0.05.

	NMDS1	NMDS2
	R^2	R^2
hs-CRP (pg/ml)	−0.22	−0.06
LBP (ng/ml)	0.18	0.27
sCD14 (pg/ml)	−0.09	−0.29
Tryptophan (umol/L)	−0.20	−0.46*
Kynurenine (umol/L)	0.12	0.07
Anthralinic acid (nmol/L)	0.08	0.39
Kynurenic.acid (nmol/L)	−0.11	−0.18
3-Hydroxykynurenin (nmol/L)	0.05	−0.10
Xanthurenic acid (nmol/L)	−0.11	−0.29**
3-Hydroxyantralinic acid (nmol/L)	−0.08	0.02
Quinilonic acid (nmol/L)	0.08	0.28
K/T ratio	0.17	0.43*

Table 3. Correlation strengths (R^2) for each NMDS axis/marker. *Indicates p-value < 0.01. **Indicates p-value < 0.05.

by microbial products, and higher proportions of mucosal adherent bacteria possessing IDO homologs have been found in HIV-infected individuals[10]. Several tryptophan catabolites, e.g. 3-hydroxyanthranilic acid, influence T-cell activation and contribute to the loss of gut resident Th17+ T-cells and to alterations in the ratio between Th17 and regulatory T-cells[13]. Additionally, kynurenine has been found to impair the survival of memory CD4+ T-cells by inhibition of IL-2 signaling[48]. In concordance with results from other studies[10, 13, 49], our data confirm that tryptophan metabolism in plasma is increased in progressive HIV-infection, but not in EC. Furthermore, IDO-1 activity (measured by K/T-ratio) correlated with NMDS2 axis, separating naive patients from the other groups, supporting that the gut microbiota composition affects systemic metabolism of tryptophan through kynurenine pathway. To our knowledge, we present the novel finding that IDO-1 activity is inversely correlated to the richness of gut microbiota, alike CD4/8 T-cell activation (data not shown). Up to now, most studies on probiotics supplementation have focused on suspension with single or very few bacterial species[50–54]. Our observation indicates that alternative therapeutic interventions modulating gut microbiota richness and not only composition are warranted in order to reduce HIV-related inflammation.

As illuminated by Moya and Ferrer[55], not only the bacterial composition is important in a given microbiota. Other factors like stability, resistance, resilience, and redundancy contribute to the functional properties of the microbiome. During HIV-infection, shifts in gut microbiota have been associated with alterations of metabolites involved in epithelial barrier integrity, hepatic function, viral infectivity and inflammation, influencing the recovery and activation of T-lymphocytes.

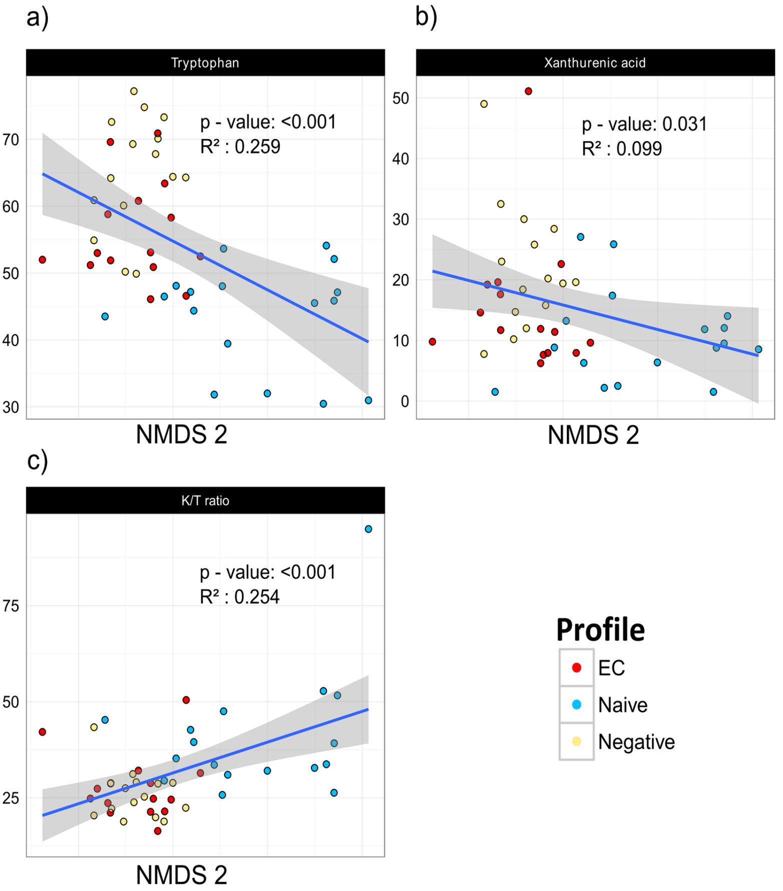


Figure 5. Correlations between tryptophan catabolism metabolites and NMDS 2 axis reveal clustering of naive patients. Significant correlations between NMDS 2 axis and tryptophan (**a**), xanthurenic acid (**b**) and K/T ratio (**c**) were observed, separating naive patients from EC and negative controls. The gray area defines the 95% confidence interval for the linear regression coefficients. The different groups are represented by different colors (EC-red, naive-blue, negative-yellow). Spearman's correlation was applied for testing correlations between metabolites and NMDS plot axis coordinates.

Up till now, only a few studies included functional analysis of gut microbiota in HIV patients[10,34,36]. In our cohort, inferred functional analysis of microbiota revealed interesting changes of gene abundance between the groups. We found lower abundance of genes involved in metabolism of carbohydrates; instead lipid metabolism related genes were enriched in EC. These differences were observed at both KEGG levels. Obviously, intracellular metabolic pathways involved in carbohydrate and lipid metabolism (like glycolysis, PPP, oxidation and synthesis of fatty acids, amino acid metabolism) are major players regulating both innate and adaptive immune cells[56]. Given that the vast majority of immune cells are located in the gut, the availability of nutrients for the gut-resident immune cells and the local metabolic milieu may influence the immunometabolism in gut compartment, subsequently tuning the immunological architecture and response to microbial stimuli. For instance, the short-chain fatty acid butyrate, derived from commensal microbiota, has been found to preferentially induce differentiation of colonic regulatory T-cells by expression of *Foxp3* gene, mediated by butyrate driven epigenetic modifications promoting inhibition of histone deacetylases (HDACs)[57]. Also long chain omega 3- polyunsaturated fatty acids (PUFA), e.g. alpha-lineolic acid which in our study correlated positively to BL CD4+ T-cell count and negatively to immune activation, have immunomodulatory properties involved in activation, differentiation and signaling of CD4+ T-cells[58]. Additionally, improved gut microbiota composition and positive immunomodulatory effects have been associated with oral supplementation of the nutritional mixture including several prebiotic oligosaccharides and omega-3/6 fatty acids in ART naive HIV-infected subjects[59,60]. Based upon these findings,

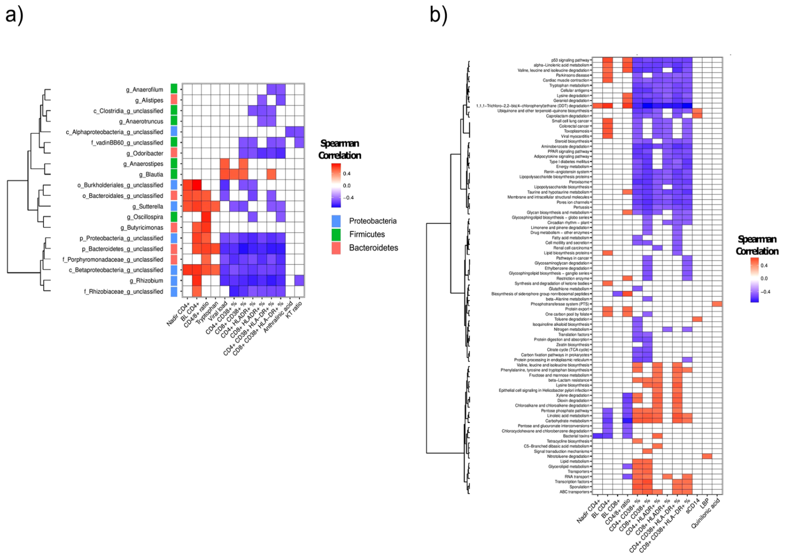


Figure 6. The composition and functionality of gut microbiota correlate with markers of immune activation and inflammation. Most cellular and some soluble markers of immune activation correlated to specific genera and functional pathways of gut microbiota. Correlations are presented by genus (**a**) and functional pathways (**b**). Spearman's correlation was used. Associations with a Benjamini – Hochberg adjusted p-value lower than 0.01 were considered relevant. Immune activation and inflammatory parameters associated with less than two bacteria were discarded when plotting the heatmap.

we hypothesize that the composition and functional capacity of gut microbiota in EC may be one of the factors contributing to virological and immunological control of the HIV-infection in absence of ART. Even if the EC group was very similar to negative subjects at both compositional and inferred functionality analyses, there were still significant differences present between the Elite controllers and negative subjects.

We acknowledge the lack of extensive dietary data, which could bias our analysis. Additionally, gene functional profiles were inferred from 16S sequences. While inferred function has shown to be robust, particularly for gut microbiome[23], they should be interpreted with caution. Our study was not designed to provide the answer about the association between the HIV progression and microbiota changes, which could be addressed in population studies with longitudinal design. On the other hand, our study was carefully designed regarding possible confounding and to our knowledge, we analyzed the microbiome of the largest cohort of EC described. Additionally, we cautiously report only correlations data which had a significance level <0.01, providing further strength to our results and conclusions.

In summary, we report that the microbiota of EC is different from individuals with progressive infection and more similar to HIV negative individuals. The differences are robust, present both in number of observed species, richness, composition and inferred functionality. Our data suggest the concept of microbiota related control of HIV infection in EC, presumably at metabolomics level. If confirmed by metabolomics studies, new intervention strategies to control HIV can be considered.

References

1. Younas, M., Psomas, C., Reynes, J. & Corbeau, P. Immune activation in the course of HIV-1 infection: Causes, phenotypes and persistence under therapy. *HIV medicine* **17**, 89–105, doi:10.1111/hiv.12310 (2016).
2. Brenchley, J. M. *et al.* Microbial translocation is a cause of systemic immune activation in chronic HIV infection. *Nat Med* **12**, 1365–1371, doi:10.1038/nm1511 (2006).
3. Dillon, S. M., Frank, D. N. & Wilson, C. C. The gut microbiome and HIV-1 pathogenesis: a two-way street. *AIDS* **30**, 2737–2751, doi:10.1097/QAD.0000000000001289 (2016).
4. Okulicz, J. F. & Lambotte, O. Epidemiology and clinical characteristics of elite controllers. *Current opinion in HIV and AIDS* **6**, 163–168, doi:10.1097/COH.0b013e328344f35e (2011).
5. Olson, A. D. *et al.* An evaluation of HIV elite controller definitions within a large seroconverter cohort collaboration. *PLoS One* **9**, e86719, doi:10.1371/journal.pone.0086719 (2014).
6. Saez-Cirion, A. *et al.* HIV controllers exhibit potent CD8 T cell capacity to suppress HIV infection *ex vivo* and peculiar cytotoxic T lymphocyte activation phenotype. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 6776–6781, doi:10.1073/pnas.0611244104 (2007).
7. Hunt, P. W. *et al.* Relationship between T cell activation and CD4+ T cell count in HIV-seropositive individuals with undetectable plasma HIV RNA levels in the absence of therapy. *J Infect Dis* **197**, 126–133, doi:10.1086/524143 (2008).

8. Noguera-Julian, M. *et al.* Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine* **5**, 135–146, doi:10.1016/j.ebiom.2016.01.032 (2016).
9. Nowak, P. *et al.* Gut microbiota diversity predicts immune status in HIV-1 infection. *AIDS* **29**, 2409–2418, doi:10.1097/QAD.0000000000000869 (2015).
10. Vujkovic-Cvijin, I. *et al.* Dysbiosis of the gut microbiota is associated with HIV disease progression and tryptophan catabolism. *Science translational medicine* **5**, 193ra191, doi:10.1126/scitranslmed.3006438 (2013).
11. Jenabian, M. A. *et al.* Distinct tryptophan catabolism and Th17/Treg balance in HIV progressors and elite controllers. *PLoS One* **8**, e78146, doi:10.1371/journal.pone.0078146 (2013).
12. Gilbert, J. A. *et al.* Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94–103, doi:10.1038/nature18850 (2016).
13. Favre, D. *et al.* Tryptophan catabolism by indoleamine 2,3-dioxygenase 1 alters the balance of TH17 to regulatory T cells in HIV disease. *Science translational medicine* **2**, 32ra36, doi:10.1126/scitranslmed.3000632 (2010).
14. Buggert, M. *et al.* Multiparametric bioinformatics distinguish the CD4/CD8 ratio as a suitable laboratory predictor of combined T cell pathogenesis in HIV infection. *Journal of immunology* **192**, 2099–2108, doi:10.4049/jimmunol.1302596 (2014).
15. StratecMolecular. User manual PSP® Spin Stool DNA Kit/PSP® Spin Stool DNAPlusKit. http://www.stratec.com/share/molecular/Manuals/Single/Pathogens/PSPSpinStool_StoolPlusKit.pdf (2016).
16. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75**, 7537–7541, doi:10.1128/AEM.01541-09 (2009).
17. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200, doi:10.1093/bioinformatics/btr381 (2011).
18. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**, 5261–5267, doi:10.1128/AEM.00062-07 (2007).
19. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* **41**, D590–596, doi:10.1093/nar/gks1219 (2013).
20. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
21. Jari Oksanen, F. G. B. *et al.* vegan: Community Ecology Package (2014).
22. Bray, J. R. C. J. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol Monogr* **27**, 325–349 (1957).
23. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology* **31**, 814–821, doi:10.1038/nbt.2676 (2013).
24. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal* **6**, 610–618, doi:10.1038/ismej.2011.139 (2012).
25. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Statistics in medicine* **9**, 811–818 (1990).
26. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology* **10**, 766, doi:10.15252/msb.20145645 (2014).
27. Wallet, M. A. *et al.* Microbial translocation induces persistent macrophage activation unrelated to HIV-1 levels or T-cell activation following therapy. *AIDS* **24**, 1281–1290, doi:10.1097/QAD.0b013e328339e228 (2010).
28. Vesterbacka, J. *et al.* Kinetics of microbial translocation markers in patients on efavirenz or lopinavir/r based antiretroviral therapy. *PLoS One* **8**, e55038, doi:10.1371/journal.pone.0055038 (2013).
29. Dinh, D. M. *et al.* Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *J Infect Dis* **211**, 19–27, doi:10.1093/infdis/jiu409 (2015).
30. Lozupone, C. A. *et al.* Alterations in the gut microbiota associated with HIV-1 infection. *Cell host & microbe* **14**, 329–339, doi:10.1016/j.chom.2013.08.006 (2013).
31. Dillon, S. M. *et al.* An altered intestinal mucosal microbiome in HIV-1 infection is associated with mucosal and systemic immune activation and endotoxemia. *Mucosal Immunol* **7**, 983–994, doi:10.1038/mi.2013.116 (2014).
32. Mutlu, E. A. *et al.* A compositional look at the human gastrointestinal microbiome and immune activation parameters in HIV infected subjects. *PLoS pathogens* **10**, e1003829, doi:10.1371/journal.ppat.1003829 (2014).
33. Cockerham, L. R. & Hatano, H. Elite control of HIV: is this the right model for a functional cure? *Trends in microbiology* **23**, 71–75, doi:10.1016/j.tim.2014.11.003 (2015).
34. McHardy, I. H. *et al.* HIV Infection is associated with compositional and functional shifts in the rectal mucosal microbiota. *Microbiome* **1**, 26, doi:10.1186/2049-2618-1-26 (2013).
35. Dillon, S. M. *et al.* Gut dendritic cell activation links an altered colonic microbiome to mucosal and systemic T-cell activation in untreated HIV-1 infection. *Mucosal Immunol* **9**, 24–37, doi:10.1038/mi.2015.33 (2016).
36. Serrano-Villar, S. *et al.* Gut Bacteria Metabolism Impacts Immune Recovery in HIV-infected Individuals. *EBioMedicine* **8**, 203–216, doi:10.1016/j.ebiom.2016.04.033 (2016).
37. Hiippala, K., Kainulainen, V., Kalliomaki, M., Arkkila, P. & Satokari, R. Mucosal Prevalence and Interactions with the Epithelium Indicate Commensalism of Sutterella spp. *Frontiers in microbiology* **7**, 1706, doi:10.3389/fmicb.2016.01706 (2016).
38. Jangi, S. *et al.* Alterations of the human gut microbiome in multiple sclerosis. *Nature communications* **7**, 12015, doi:10.1038/ncomms12015 (2016).
39. Montassier, E. *et al.* Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome medicine* **8**, 49, doi:10.1186/s13073-016-0301-4 (2016).
40. Konikoff, T. & Gophna, U. Oscillospira: a Central, Enigmatic Component of the Human Gut Microbiota. *Trends in microbiology* **24**, 523–524, doi:10.1016/j.tim.2016.02.015 (2016).
41. Fischer, H. M. Genetic regulation of nitrogen fixation in rhizobia. *Microbiological reviews* **58**, 352–386 (1994).
42. Vazquez-Castellanos, J. F. *et al.* Altered metabolism of gut microbiota contributes to chronic immune activation in HIV-infected individuals. *Mucosal Immunol* **8**, 760–772, doi:10.1038/mi.2014.107 (2015).
43. Zelante, T. *et al.* Tryptophan catabolites from microbiota engage aryl hydrocarbon receptor and balance mucosal reactivity via interleukin-22. *Immunity* **39**, 372–385, doi:10.1016/j.immuni.2013.08.003 (2013).
44. Williams, M. N. & Signer, E. R. Metabolism of Tryptophan and Tryptophan Analogs by Rhizobium meliloti. *Plant physiology* **92**, 1009–1013 (1990).
45. Zhang, L. S. & Davies, S. S. Microbial metabolism of dietary components to bioactive metabolites: opportunities for new therapeutic interventions. *Genome medicine* **8**, 46, doi:10.1186/s13073-016-0296-x (2016).
46. Romani, L. *et al.* Microbiota control of a tryptophan-AhR pathway in disease tolerance to fungi. *European journal of immunology* **44**, 3192–3200, doi:10.1002/eji.201344406 (2014).
47. Mellor, A. L. & Munn, D. H. IDO expression by dendritic cells: tolerance and tryptophan catabolism. *Nature reviews. Immunology* **4**, 762–774, doi:10.1038/nri1457 (2004).
48. Dagenais-Lussier, X. *et al.* Kynurenine Reduces Memory CD4 T-Cell Survival by Interfering with Interleukin-2 Signaling Early during HIV-1 Infection. *Journal of virology* **90**, 7967–7979, doi:10.1128/JVI.00994-16 (2016).
49. Gaardbo, J. C. *et al.* Increased Tryptophan Catabolism Is Associated With Increased Frequency of CD161+ Tc17/MAIT Cells and Lower CD4+ T-Cell Count in HIV-1 Infected Patients on cART After 2 Years of Follow-Up. *Journal of acquired immune deficiency syndromes* **70**, 228–235, doi:10.1097/QAI.0000000000000758 (2015).

50. Stiksrud, B. *et al*. Reduced Levels of D-dimer and Changes in Gut Microbiota Composition after Probiotic Intervention in HIV-infected Individuals on Stable ART. *Journal of acquired immune deficiency syndromes*. doi:10.1097/QAI.0000000000000784 (2015).
51. d'Ettorre, G. *et al*. Probiotics Reduce Inflammation in Antiretroviral Treated, HIV-Infected Individuals: Results of the "Probio-HIV" Clinical Trial. *PLoS One* **10**, e0137200, doi:10.1371/journal.pone.0137200 (2015).
52. Ortiz, A. M. *et al*. IL-21 and probiotic therapy improve Th17 frequencies, microbial translocation, and microbiome in ARV-treated, SIV-infected macaques. *Mucosal Immunol* **9**, 458–467, doi:10.1038/mi.2015.75 (2016).
53. Villar-Garcia, J. *et al*. Effect of probiotics (Saccharomyces boulardii) on microbial translocation and inflammation in HIV-treated patients: a double-blind, randomized, placebo-controlled trial. *Journal of acquired immune deficiency syndromes* **68**, 256–263, doi:10.1097/QAI.0000000000000468 (2015).
54. Yang, O. O., Kelesidis, T., Cordova, R. & Khanlou, H. Immunomodulation of antiretroviral drug-suppressed chronic HIV-1 infection in an oral probiotic double-blind placebo-controlled trial. *AIDS research and human retroviruses* **30**, 988–995, doi:10.1089/AID.2014.0181 (2014).
55. Moya, A. & Ferrer, M. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. *Trends in microbiology* **24**, 402–413, doi:10.1016/j.tim.2016.02.002 (2016).
56. O'Neill, L. A., Kishton, R. J. & Rathmell, J. A guide to immunometabolism for immunologists. *Nature reviews. Immunology* **16**, 553–565, doi:10.1038/nri.2016.70 (2016).
57. Furusawa, Y. *et al*. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446–450, doi:10.1038/nature12721 (2013).
58. Hou, T. Y., McMurray, D. N. & Chapkin, R. S. Omega-3 fatty acids, lipid rafts, and T cell signaling. *European journal of pharmacology* **785**, 2–9, doi:10.1016/j.ejphar.2015.03.091 (2016).
59. Gori, A. *et al*. Specific prebiotics modulate gut microbiota and immune activation in HAART-naive HIV-infected adults: results of the "COPA" pilot randomized trial. *Mucosal Immunol* **4**, 554–563, doi:10.1038/mi.2011.15 (2011).
60. Cahn, P. *et al*. The immunomodulatory nutritional intervention NR100157 reduced CD4+ T-cell decline and immune activation: a 1-year multicenter randomized controlled double-blind trial in HIV-infected persons not receiving antiretroviral therapy (The BITE Study). *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* **57**, 139–146, doi:10.1093/cid/cit171 (2013).

Acknowledgements

The study was supported by Stockholm's County Council (SLL-KI for P. N), Swedish Physician against AIDS Research Fund, Swedish Research Council, and Swedish Society of Medicine (SVLS) and by a research grant from the Fundació Glória Soler. M.N-J, J.R and R.P acknowledge funds from the Gala contra la SIDA 2013, 2014 and 2015 editions, the Nit per la Recerca a la Catalunya Central 2015 edition and the Red de investigación en SIDA, RD12/0017/0002 as part of the Plan Nacional R+D+I and cofinanced by the Instituto de Salud Carlos III (ISCIII)-Subdirección General de Evaluación y el Fondo Europeo de Desarrollo Regional (FEDER). J.R. is supported through a grant for pre-doctoral studies from Noel Alimentaria to the University of Vic (UVic-UCC) and M.C. through MTM2015-64465-C2-1-R, Ministerio de Economía y Competitividad, Spain.

Author Contributions

J.V. supervised collection of samples and clinical data, analyzed the data, and wrote the manuscript. J.R. analyzed the data, discussed the results and contributed to manuscript preparation. K.N. and M.P. performed the experiments. U.N. and M.C. supervised the experiments and data analysis. R.P. supervised the experiments, data analysis and discussed the data. A.S. planned the experiments, supervised patient inclusion and data analysis. M.N. analyzed and discussed the data, and contributed to manuscript preparation. P.N. planned the experiments, supervised patient inclusion and sample collection, discussed the data, verified the data analysis and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-06675-1

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CHAPTER D

Balances: A New Perspective for Microbiome Analysis

AMERICAN SOCIETY FOR MICROBIOLOGY mSystems®

METHODS AND PROTOCOLS
Host-Microbe Biology

Balances: a New Perspective for Microbiome Analysis

J. Rivera-Pinto,[a,b] J. J. Egozcue,[c] V. Pawlowsky-Glahn,[d] R. Paredes,[a,b,e,f] M. Noguera-Julian,[a,b,e] M. L. Calle[b]

[a]irsiCaixa AIDS Research Institute, Badalona, Spain
[b]Universitat de Vic—Universitat Central de Catalunya, Vic, Spain
[c]Universitat Politècnica de Catalunya, Barcelona, Spain
[d]Universitat de Girona, Girona, Spain
[e]Universitat Autónoma de Barcelona, Barcelona, Spain
[f]Infectious Diseases Service, Hospital Germans Trias i Pujol, Badalona, Spain

ABSTRACT High-throughput sequencing technologies have revolutionized microbiome research by allowing the relative quantification of microbiome composition and function in different environments. In this work we focus on the identification of microbial signatures, groups of microbial taxa that are predictive of a phenotype of interest. We do this by acknowledging the compositional nature of the microbiome and the fact that it carries relative information. Thus, instead of defining a microbial signature as a linear combination in real space corresponding to the abundances of a group of taxa, we consider microbial signatures given by the geometric means of data from two groups of taxa whose relative abundances, or balance, are associated with the response variable of interest. In this work we present *selbal*, a greedy stepwise algorithm for selection of balances or microbial signatures that preserves the principles of compositional data analysis. We illustrate the algorithm with 16S rRNA abundance data from a Crohn's microbiome study and an HIV microbiome study.

IMPORTANCE We propose a new algorithm for the identification of microbial signatures. These microbial signatures can be used for diagnosis, prognosis, or prediction of therapeutic response based on an individual's specific microbiota.

KEYWORDS balances, compositional data, microbiome

Human microbiome research, focused on the study of the microorganisms that live throughout the human body and their role in health and disease, has experienced significant growth in the last few years. High-throughput sequencing technologies have revolutionized this field by allowing the quantification of microbiome composition and function in different environments. Large-scale projects, such as the Human Microbiome Project (1, 2) or MetaHIT (metagenomics of the human intestinal tract), have established standardized protocols for creating, processing, and interpreting metagenomics data (3). However, analysis of microbiome data is still challenging due to, among other reasons, their inherently compositional nature.

High-throughput DNA sequencing generates thousands of sequence reads that, after bioinformatics preprocessing and quality control, are annotated to different microbial species or taxa. An abundance table of counts summarizes the number of sequences per sample of each taxon. The total number of counts per sample, also known as sequencing depth or library size, is highly variable and constrained by the maximum number of sequence reads of the instrument. This total count constraint induces strong dependencies among the abundances of the different taxa; an increase in the abundance of one taxon requires the decrease of the observed number of counts for some of the other taxa so that the total number of counts does not exceed the specified sequencing depth. Moreover, observed raw abundances and the total number

Received 13 April 2018 **Accepted** 21 June 2018 **Published** 17 July 2018

Citation Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. 2018. Balances: a new perspective for microbiome analysis. mSystems 3:e00053-18. https://doi.org/10.1128/mSystems.00053-18.

Editor Catherine Lozupone, University of Colorado Denver

Address correspondence to M. Noguera-Julian, mnoguera@irsicaixa.es.

M.N.-J. and M.L.C. contributed equally to this article.

Use of compositional balances as biomarkers in clinical microbiome studies.

of reads per sample are noninformative since they represent only a fraction or random sample of the original DNA content in the environment. These characteristics of microbiome abundance data clearly fall into the notion of compositional data. Compositional data are defined as a vector of strictly positive real numbers with an unknown or uninformative total. Compositional data carry relative information, i.e., information contained in the ratios between components, and the numerical value of each component by itself is irrelevant (4). Except for the fact that microbiome abundance tables contain many zeros, microbiome data fit the definition of compositional data and, as already acknowledged by many authors (5, 6), their analysis requires the use of a proper mathematical theory (7).

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 outcome. Multivariate methods such as PERMANOVA (8, 9), implemented in the *Adonis()* function of the R package *vegan* (10), and MiRKAT (11) address the first issue. The second issue is approached with univariate methods where each taxon is tested for association with the outcome. When the response variable is dichotomous, the testing method is known as differential abundance testing and methods specifically developed for transcriptome sequencing (RNA-Seq) data, such as *DESeq2* (12) and *edgeR* (13), are commonly used. Other methods, such as *ANCOM* (14) and *ALDEx2* (15), have been proposed that acknowledge the compositionality of microbiome data. See a previous report by Weiss et al. for a comprehensive comparison of methods for microbiome differential abundance testing (16).

In this paper we focus on a different issue. The goal of the proposed methodology is to identify microbial signatures, that is, groups of microbial taxa that are predictive of a phenotype of interest. These microbial signatures can be used for diagnosis, prognosis, or prediction of therapeutic response based on an individual's specific microbiota (17). The identification of microbial signatures involves both modeling and variable selection: modeling the response variable and identifying the smallest number of taxa with the highest prediction or classification accuracy. We present *selbal*, a model selection procedure that searches for a sparse model that adequately explains the response variable of interest. Similarly to forward stepwise linear regression, *selbal* performs multiple regressions a number of times, each time adding a new taxon to the model. Unlike linear regression, the raw variables in *selbal* are not included in a linear equation in real space but are added as part of what is called a "balance" in the compositional data analysis literature, i.e., as part of a particular type of log-contrast.

Mathematically, a compositional balance is defined as follows. Let $X = (X_1, X_2, \ldots, X_k)$ be a composition with k components or parts. Given two disjoint subsets of components in X, denoted by X_+ and X_-, indexed by I_+ and I_-, and composed of k_+ and k_- parts, respectively, the balance between X_+ and X_- is defined as the normalized log ratio of the geometric mean of the values for the two groups of components as follows:

$$B(X_+, X_-) = \sqrt{\frac{k_+ \cdot k_-}{k_+ + k_-}} \log \frac{\left(\Pi_{i \in I_+} X_i\right)^{1/k_+}}{\left(\Pi_{j \in I_-} X_j\right)^{1/k_-}}$$

Expanding the logarithm, we obtain a more usual expression of a balance that is proportional to the difference between the means of the log-transformed variables of the two groups of components as follows:

$$B(X_+, X_-) \propto \frac{1}{k_+} \sum_{i \in I_+} \log X_i - \frac{1}{k_-} \sum_{j \in I_-} \log X_j$$

A compositional balance is a special kind of a log-contrast, defined as a linear combination of the log-transformed components of a composition with the restriction that the coefficients of the linear function add up to zero (4). The importance of working with log-contrast functions or, in particular, with balances, in analyzing compositional data is that this kind of function preserves scale invariance, one of the principles that should be fulfilled in compositional data analysis (4, 7).

Our algorithm for balance selection, *selbal*, starts with a first thorough search of the two taxa whose balance, or log ratio, is most closely associated with the response. Once the first two-taxon balance is selected, the algorithm performs a forward selection process where, at each step, a new taxon is added to the existing balance such that the specified criterion is improved (area under the receiver operating characteristic [ROC] curve [AUC] or mean squared error [MSE]). The algorithm stops when there is no additional variable that improves the current optimization parameter or when the maximum number of components to be included in the balance is achieved. This number is established with a cross-validation (CV) procedure, which is also used to explore the robustness of the identified balance. A more detailed description of the algorithm is given in Materials and Methods.

selbal is different from other modeling approaches for microbiome analysis such as MiRKAT (11), which performs an overall association test comparing the microbiome and the phenotype but does not perform model selection.

Model selection for microbial signature identification can also be performed in two separate steps: first, variable selection; second, model building with the selected variables. When the outcome variable is dichotomous, variable selection can be obtained with methods for microbiome differential abundance testing mentioned before, such as *DESeq2* (12), *edgeR* (13), or, in the context of compositional data analysis, *ANCOM* (14) or *ALDEx2* (15). However, it is not clear how to combine the selected variables to obtain the best joint sparse model. This is specially challenging for microbiome analysis, where the compositional nature of microbiome data induces spurious correlations among the variables. We think that a joint procedure that involves both modeling and variable selection, as performed in *selbal*, is more appropriate in this context.

Other authors (18–20) have previously proposed the use of balances for microbiome analysis regarding the construction of an isometric log ratio (ILR) transformation (21), which allows compositional data to be represented in a real Euclidean space, where standard statistical methods can be applied. Silverman et al. (18) and Washburne et al. (19) proposed methods that use microbial phylogenetic information to guide the sequential binary partition in the construction of a particular ILR transformation. This phylogenetically driven ILR transformation would help to detect relevant evolutionary factors or phylogenetically related bacterial groups (clades) related to host-microbiome interactions (18, 19). In the method proposed by Morton et al. (20), instead of using phylogenetic information, they use the response variable to define the binary sequential partitions of the ILR transformation. *selbal* is different from these methods in the following ways: first, only one balance is considered and not a sequence of balances in *selbal*; second, the purpose of the selected balance is classification or prediction and not a new representation of the data.

As with any other compositional data method, one important issue that *selbal* addresses before the searching algorithm can be applied is that of how to deal with the large amount of zeros that are typically present in microbiome data sets. Their treatment is different depending on whether they represent essential or rounded zeros (22). In microbiome analysis, an essential or structural zero represents the absolute absence of a taxon in a sample, e.g., because the microbe is unable to live in that environment. In dealing with essential zeros, samples are considered to belong to two distinct subpopulations according to the presence or absence of a zero. On the other hand, rounded zeros arise because of insufficient sampling depth: they correspond to taxa in such a small proportion in the sample that they were not picked during the sequencing process. The common practice for the treatment of rounded zeros is their replacement by a small positive value. This replacement can be implemented in different ways, including replacing the zeros by a constant, adding a constant to all values in the data set, and using more-sophisticated methods for zero replacement that are designed to preserve as much as possible the covariance structure of the initial data. Though the user can perform other zero replacements before using *selbal* (18, 20), the default option in *selbal* is geometric Bayesian multiplicative replacement (GBM) (23) (described in more detail in Materials and Methods).

Another important issue in microbiome analysis is sampling variance. As discussed by Gloor et al. (24), microbiome data, as with any other kind of high-throughput sequencing data, are subject to high levels of variability or uncertainty that should be conveniently treated. The number of reads obtained in an experiment represents just an instance from a random sample of the true microbial composition in the environment of interest. The same protocol applied twice to the same sample would provide different microbial compositions. This imprecision, which arises during the library preparation and sequencing process, is larger for taxa of low abundance and should be taken into account when dealing with observed zeros. One way to handle this variability is by modeling the read counts per sample as multinomial or negative binomial, which implies that the vectors of relative abundances of the different taxa follow approximately a Dirichlet distribution. Then, using a Dirichlet Monte Carlo sampling approach, we can obtain multiple instances of the posterior distribution of the relative abundances determined for each sample (25). By multiplying the relative abundances of each sample by the observed total number of counts, we obtain multiple abundance tables which can be analyzed with *selbal*. The comparison of the microbial signatures obtained from these new Monte Carlo abundance tables to the microbial signature obtained with the initial table can be used to evaluate the robustness of the initial result.

The remainder of this paper is organized as follows. In the next section, we illustrate the proposed algorithm with a Crohn's disease (CD) microbiome study and an HIV microbiome study. Some discussions and suggestions for future work are provided in the Discussion. In Materials and Methods, we present the detailed description of the algorithm. *selbal* is accessible as an R package in GitHub (https://github.com/UVic-omics/selbal), where the data sets and scripts to reproduce this work are also available.

RESULTS

We illustrate the proposed methodology for use with microbiome compositions at the genus level for a Crohn's disease study (26) and an HIV microbiome study (27). We did not perform the bioinformatics processing of the sequences (with Mothur or Qiime) but took the processed operational taxonomic unit (OTU) tables available in Qiita and Bioproject repositories (accession numbers provided below). We performed additional filtering and agglomeration steps to obtain the abundance tables at the genus level. The scripts to obtain those genus-level abundance tables are available at GitHub.

Microbiome and Crohn's disease. Crohn's disease (CD) is an inflammatory bowel disease that has been linked to microbial alterations in the gut (26, 28). We use data from a large pediatric CD cohort study (26) to illustrate the proposed methodology for identification of microbial signatures. Microbiome data from 16S rRNA gene sequencing and QIIME 1.7.0 bioinformatics processing were downloaded from Qiita https://qiita.ucsd.edu (study identifier [ID]: 1939). Only patients with Crohn's disease ($n = 662$) and those without any symptoms ($n = 313$) were analyzed. The OTU table was agglomerated to the genus level, resulting in a matrix with 48 genera and 975 samples.

The goal of *selbal* analysis is to identify a microbial signature for Crohn's disease that is able to discriminate between CD and non-CD individuals. This microbial signature is defined by two groups of taxa whose relative abundances, or balance, are associated with Crohn's disease status.

As explained in Materials and Methods, we first ran a cross-validation (CV) process with the function selbal.cv*()* that helped us to determine the optimal number of taxa to be included in the balance. Figure 1 provides the mean AUC and standard error of the balances obtained in the CV process as a function of the number of taxa. In this case, and following the 1se rule, the optimal number of taxa is 12.

Once the number of taxa is determined, we apply the main function *selbal()* to the whole data set, with the number of taxa $C = 12$, and obtain what we call the "global balance."

The two groups of taxa defining the global balance, or microbial signature, for Crohn's disease are $X_+ = \{$*g_Roseburia*, *o_Clostridiales_g_*, *g_Bacteroides*, *f_Peptostreptococcaceae_g_*$\}$ and $X_- = \{$*g_Dialister*, *g_Dorea*, *o_Lactobacillales_g_*, *g_Eggerthella*,

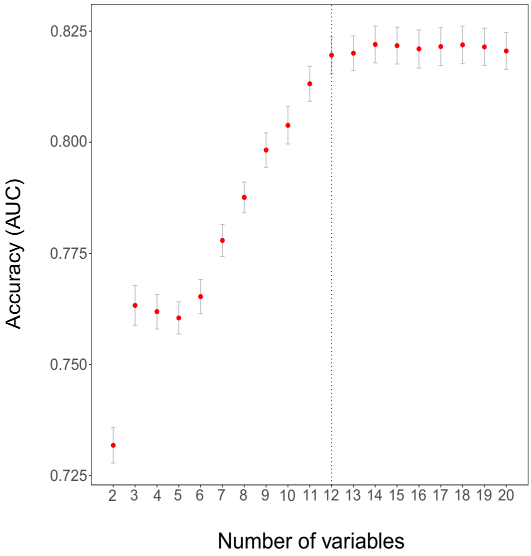


FIG 1 Mean area under the ROC curve (AUC) as a function of the number of components included in the balance in the cross-validation process for Crohn's disease. The optimal number of components according to the "1se rule" is highlighted with a vertical dashed line.

g_Aggregatibacter, *g_Adlercreutzia*, *g_Streptococcus*, *g_Oscillospira*}. Figure 2 presents the distribution of the microbial signature values for CD and non-CD individuals. Patients with Crohn's disease have lower balance scores than controls, which means that there are lower relative abundances of taxa in group X_+ than in group X_-. *Bacteroides* and *Clostridiales* have also been previously identified as less abundant in Crohn's disease individuals than in controls (26). The discrimination value of the identified balance is important, with an apparent AUC value of 0.838. However, this apparent AUC is known to overestimate the discrimination value of the microbial signature, since it has been measured on the same data set that was used to build the model. A more accurate estimation is obtained from the CV process, which provides a cv-AUC of 0.819, a very good discrimination value.

Robustness of the selected global balance. CV can also help us to assess the robustness of the proposed global balance. In Fig. 3, we summarize the different balances obtained with 12 taxa in the CV process. On the one hand, we have the frequency of the different CV balances and, on the other, the frequency of selection of each taxon. Rows represent the most frequent taxa, with their percentage of selection given in the second column; the third column represents the global balance, that is, the balance obtained using all the samples; and the last three columns represent the three most frequent balances selected in the CV procedure. Colored rectangles indicate whether the taxon is in the numerator of the balance (red) or in the denominator (blue) or not included (white). The last row indicates the proportion of times each balance has been selected as optimal in the CV procedure. From the data presented in Fig. 3, it follows that the identified global balance for Crohn's disease is very robust; the global balance coincides with the balance most frequently found in the CV process, which turned out to be the optimal balance 36% of the time. Moreover, the taxa which form the global balance are also those most frequently selected in the CV procedure.

An alternative approach for exploring the robustness of the selected global balance is by implementing Monte Carlo sampling from a Dirichlet distribution prior to microbiome signature identification with *selbal*. This process returns a set of different

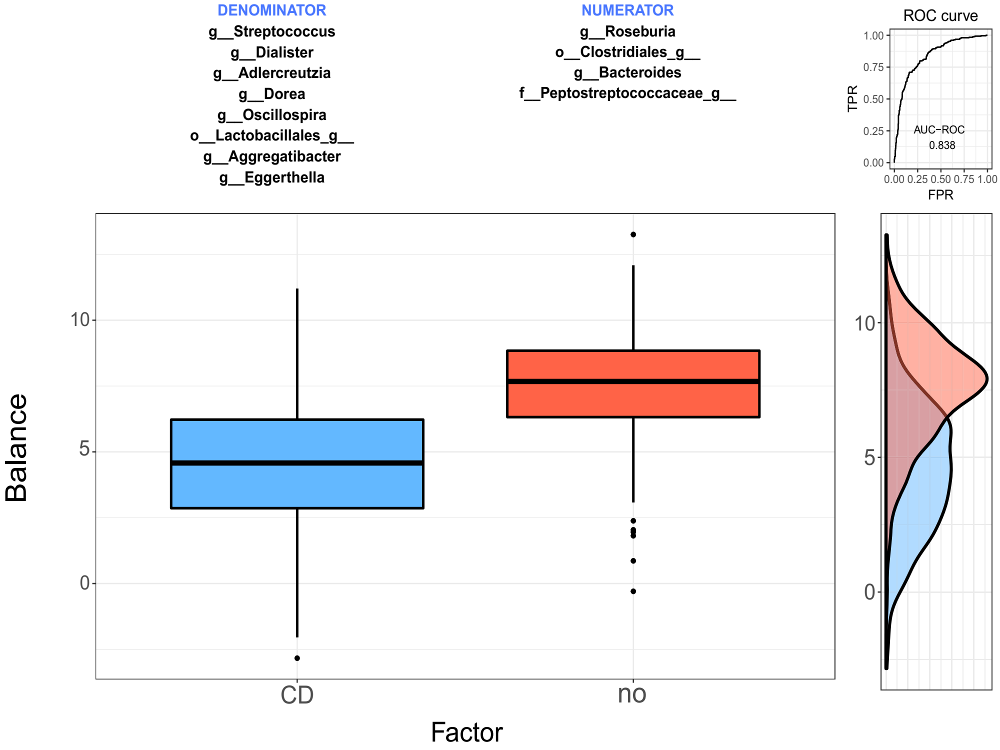
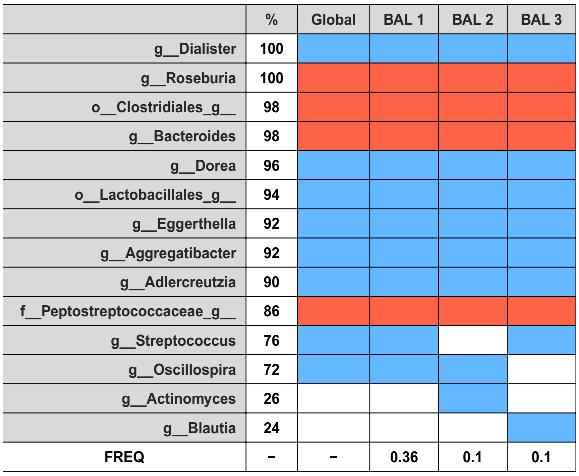


FIG 2 Description of the global balance for Crohn's disease. The two groups of taxa that form the global balance are specified at the top of the plot. The box plot represents the distribution of the balance scores for CD and non-CD individuals. The right part of the figure contains the ROC curve with its AUC value (0.838) and the density curve for each group.

microbiome signatures that can be compared with the global balance. We performed this strategy with Crohn's microbiome data with 100 sampling iterations and obtained balances highly concordant with the proposed global balance (see Fig. S1 in the supplemental material). Similar results were obtained with two zero replacement strategies, a constant of 1 added to all values, and geometric Bayesian multiplicative replacement (GBM) (23).

Comparison with other approaches. Using the Crohn's disease data set, we compared the classification accuracy of *selbal* with that of strategies employing two steps: first, variable selection; second, model building. For the variable-selection step, we considered *DESeq2*, *edgeR*, *ANCOM*, and *ALDEx2*, and then we built a model or microbial signature with the selected variables. The model is a linear combination of the selected variables for *DESeq2*, *edgeR*, and *ANCOM*, whereas for *ALDEx2* the model is defined as a linear combination of the selected variables, previously transformed according to the centered log-ratio transformation (15). *selbal* cannot be compared with these methods in terms of false-discovery rate (FDR) and power because the goal of *selbal* is not to identify all taxa that are associated with the response but to obtain the best sparse model to predict the response. In a cross-validation process, we measured the test prediction accuracy and sparsity of the models (microbial signatures) obtained with each method. The results are given in Table 1, and in Fig. S2 we can see the variability of cv-AUC for the different methods.

selbal and *ALDEx2* are the methods with the best classification accuracy, but *selbal* is more parsimonious, which is also a desirable feature of microbial signatures. *selbal*



	%	Global	BAL 1	BAL 2	BAL 3
g__Dialister	100	🟦	🟦	🟦	🟦
g__Roseburia	100	🟥	🟥	🟥	🟥
o__Clostridiales_g__	98	🟥	🟥	🟥	🟥
g__Bacteroides	98	🟥	🟥	🟥	🟥
g__Dorea	96	🟦	🟦	🟦	🟦
o__Lactobacillales_g__	94	🟦	🟦	🟦	🟦
g__Eggerthella	92	🟦	🟦	🟦	🟦
g__Aggregatibacter	92	🟦	🟦	🟦	🟦
g__Adlercreutzia	90	🟦	🟦	🟦	🟦
f__Peptostreptococcaceae_g__	86	🟥	🟥	🟥	🟥
g__Streptococcus	76	🟦	🟦		🟦
g__Oscillospira	72	🟦	🟦	🟦	
g__Actinomyces	26			🟦	
g__Blautia	24				🟦
FREQ	–	–	0.36	0.1	0.1

FIG 3 Cross-validation (CV) results for Crohn's disease study: most frequent taxa and most frequent balances selected in the CV procedure compared to the global balance obtained with the whole data set. Colored rectangles indicate if the component is in the numerator of the balance (BAL) (red), in the denominator (blue), or not included (white). FREQ, frequency.

obtains discrimination accuracy with only 12 taxa that is similar to that obtained with *ALDEx2* with 31 taxa. *DESeq2* and *edgeR* provide similar results: high numbers of selected taxa but lower classification accuracy. This suggests that among the variables selected by *DESeq2* and *edgeR*, there are some false positives. *ANCOM* is the best in terms of parsimony; it selects the smallest number of variables, with classification accuracy comparable to that of *DESeq2* and *edgeR*. This is in accordance with previous simulation studies (16) that indicated that *ANCOM* has very low FDR and comparable power to other methods. These results cannot be generalized since they reflect the behavior of the methods with only one specific data set. A more general conclusion would require a comprehensive simulation study.

Microbiome and HIV infection. Understanding the role of the gut microbiome in HIV-1 infection may help to design novel interventions to improve HIV-1-associated immune dysfunction. We considered a cross-sectional HIV microbiome study conducted in Barcelona, Spain, that included both HIV-infected subjects and HIV-negative controls (27). Microbiome data were obtained from a MiSeq 16S rRNA sequence and bioinformatically processed with Mothur (29) and are available at BioProject (PRJNA307231). After applying abundance filters and agglomerating taxa to the genus level, microbiome abundance data are summarized in a matrix of raw abundances for 155 samples and 60 different genera.

The main goal of this analysis is to find a microbial signature for HIV status, that is, two groups of taxa whose relative abundance or balance data are able to discriminate

TABLE 1 Comparison of model complexity and discrimination accuracy of microbial signatures for Crohn's disease status[a]

Method	Median no. of taxa	Mean cv-AUC
selbal	12	0.8196
DESeq2	33	0.7752
edgeR	34	0.7721
ANCOM	5	0.7125
ALDEx2	31	0.8156

[a] For each method, the table indicates the median number of taxa included in the model and the mean cv-AUC for 10 iterations of a 5-fold cross-validation process.

between HIV-positive and HIV-negative individuals. As reported by Noguera-Julian et al. (27), a possible confounder in HIV microbiome studies is the HIV risk factor MSM (men who have sex with men versus non-MSM). *selbal* algorithm implements a regression model which allows adjustment for other variables. Thus, we applied the algorithm to Y = HIV status and X = microbiome abundance at the genus level, adjusted by Z = MSM factor.

According to the cross-validation (CV) procedure implemented with function selbal.cv*()*, the optimal number of components to be included in the balance is 2 (Fig. S3). The microbiome signature for HIV status identified with *selbal* is given by the log ratio between the abundance of a taxon of the family *Erysipelotrichaceae* and of unknown genus and a taxon of the family *Ruminococcaceae* and of unknown genus (Fig. S4). HIV-negative individuals are associated with lower balance values, most of them negative, that is, with larger relative abundances of *Ruminococcaceae* than of *Erysipelotrichaceae*, while HIV-positive individuals have heterogeneous balance values. The discrimination accuracy of this balance is moderate, with an AUC of 0.786 on the whole sample and a mean cross-validation AUC of 0.674 measured on the test data sets given MSM status. Figure S5 shows the result of the cross-validation procedure. The balance identified with the whole data set is that most frequently identified in the cross-validation procedure, appearing 44% of the time, an indicator of robustness for the proposed global balance.

Microbiome and soluble CD14 inflammation marker. Acute inflammation and chronic inflammation typically occur after HIV infection. Even patients administered antiretroviral medications and with an undetectable viral load present with chronic inflammation, which may cause tissue damage and is associated with many chronic diseases (30). In this context, there is great interest in defining possible interventions involving modifications of the gut bacterial environment which may reduce inflammation in HIV patients (31, 32). This requires a good understanding of the association between gut microbial composition and several inflammation markers. In this case, we focus on immune markers related to chronic inflammation: the levels of soluble CD14 (sCD14), which was measured for a subset of samples (n = 151). We apply *selbal* to search for a microbial signature that is predictive of an sCD14 inflammation marker. According to the cv-MSE (Fig. S6), the optimal number of components to be included in the model is four. The balance that is identified as that most closely associated with sCD14 is composed of two taxa in the numerator, X_+ = {*g_Subdoligranulum*, *f_Lachnospiraceae_g_Incertae_Sedis*}, and two in the denominator, X_- = {*f_Lachnospiraceae_g_unclassified*, *g_Collinsella*}. The association is moderate, with R = 0.53. Since sCD14 data are continuous, we represent the result with a scatter plot of the balance values and sCD14 values. We observe that higher balance scores are associated with higher sCD14 values (Fig. S7). The robustness of the selected balance can be evaluated through the results of the CV procedure (Fig. S8). We see that the proposed global balance is also the one that has been the most frequently (34% of the time) selected in the CV. The four taxa defining the global balance correspond to the top 4 most frequently selected in the cross-validation. These results emphasize the robustness of the selected global balance.

DISCUSSION

The identification of microbial signatures that are predictive of a variable of interest is an essential step toward the translation of microbiome research to clinical practice. In this work, we present *selbal*, a greedy stepwise algorithm for the identification of microbial signatures consisting of two groups of taxa whose relative abundances, or balance, are predictive of the outcome. Working with balances and, in general, with log-contrast functions preserves the scale-invariant principle for compositional data analysis.

In the Crohn's disease study considered in this work, *selbal* outperformed methods commonly used in microbiome analysis, such as *DESeq2* and *edgeR*, in terms of discrimination accuracy. With respect to methods for compositional data, *selbal* performs much better than ANCOM and similarly to *ALDEx2* in terms of classification accuracy, but *selbal* is more parsimonious.

selbal overcomes the problem of differences in sample size that is usually accommodated with different methods based on count normalization, rarefaction, or transformation into proportions. These normalization techniques are controversial since they may have an important impact on the analysis (16, 33, 34). The only way in which data are altered in *selbal* is at the zero imputation stage, which is required because of the use of logarithms and ratios in the definition of balances. This replacement of zeros by positive numbers is performed under the assumption that the observed zeros represent rounded zeros, that is, that all taxa are present in all the samples but some of them are not detected because of low abundance and insufficient sampling depth. However, it is not clear how the imputation method and the presence of structural zeros (absence of the taxa in the sample) may influence the results. Future research will focus on the treatment of zeros, with the aim of more precisely evaluating if zeros are rounded or structural, and on selecting the best replacement method.

The technical variability of microbiome sequencing data should be taken in consideration. The effects of this uncertainty on the results of *selbal* can be explored through Monte Carlo sampling from a Dirichlet distribution (25) prior to microbiome signature identification with *selbal*.

A limitation of *selbal* is that the greedy algorithm does not guarantee that the global optimum will be found. Due to the computational cost, *selbal* does not explore the whole balance space; the method for selecting the optimal balance is suboptimal and may be improved. In this respect, the iterative CV process included in the *selbal* algorithm is useful for exploring the robustness of the result. The degree of concordance between the balances obtained in the CV process and the global balance can provide reasonable evidence to support the optimality of the global balance. Exploring possible alternative approaches in the search of the optimal balance is another topic of future research.

MATERIALS AND METHODS

Let $X = (X_1, X_2, \ldots, X_k)$ be a composition, that is, a vector of strictly positive real numbers that carry relative information. Given two disjoint subsets of components in X, denoted by X_+ and X_-, indexed by I_+ and I_-, and composed of k_+ and k_- features, respectively, the balance between X_+ and X_- is defined as the normalized log ratio of the geometric mean of the two groups of components as follows:

$$B\left(X_+, X_-\right) = \sqrt{\frac{k_+ \cdot k_-}{k_+ + k_-}} \log \frac{\left(\Pi_{i \in I_+} X_i\right)^{1/k_+}}{\left(\Pi_{j \in I_-} X_j\right)^{1/k_-}}$$

Expanding the logarithm, we obtain the result that a balance is proportional to the difference between the arithmetic means of the log-transformed variables of the two groups of components as follows:

$$B\left(X_+, X_-\right) \propto \frac{1}{k_+} \sum_{i \in I_+} \log X_i - \frac{1}{k_-} \sum_{j \in I_-} \log X_j$$

The second expression is preferable from a computational point of view and is the one implemented in the proposed algorithm.

Given Y, a response variable, which can be either numeric or dichotomous, a composition $X = (X_1, X_2, \ldots, X_k)$, and additional covariates $Z = (Z_1, Z_2, \ldots, Z_r)$, the goal of the algorithm is to determine two subcompositions of X, X_+ and X_-, indexed by $I_+ \subset \{1, 2, \ldots, k\}$ and $I_- \subset \{1, 2, \ldots, k\}$, respectively, so that the balance $B(X_+, X_-)$ between X_+ and X_- is highly associated with Y after adjustment for covariates Z. Depending on the nature of the dependent variable, the association can be defined in several ways.

For a continuous variable Y, the optimization criterion is defined as minimization of the MSE of the linear regression model as follows:

$$Y = \beta_0 + \beta_1 B\left(X_+, X_-\right) + \gamma' Z$$

For a dichotomous variable Y, we fit the logistic regression model as follows:

$$\operatorname{logit}\left(Y\right) = \beta_0 + \beta_1 B\left(X_+, X_-\right) + \gamma' Z$$

In this case, we consider three possible optimization criteria corresponding to the maximization of the area under the ROC curve (default option), the maximization of the explained variance (35), or the discrimination coefficient (36).

Main function: *selbal()*. The main function of the proposed algorithm to detect the most closely associated balance is called *selbal()* and employs the following three steps.

Step 0: zero replacement. The initial matrix of counts in a microbiome study, denoted $\widetilde{X}$, typically contains many zeros that must be treated prior to using *selbal* algorithm. Though the user can perform other zero replacements before using *selbal* (18, 20), the default option in *selbal* is geometric Bayesian multiplicative replacement (GBM) (23) as implemented in the *cmultRepl()* function of the R package *zCompositions* (37). GBM performs Bayesian estimation of the zero values, assuming a Dirichlet model and a multiplicative modification of the nonzero values, so that both the ratios between parts and the total sum of the initial vector before the replacement are preserved. GBM performs better than other Bayesian multiplicative replacements assuming a Dirichlet distribution (23). *selbal* also provides the option of adding a value of 1 to all values in the data set. The resulting matrix with zeros replaced by positive values is denoted by *X*.

Step 1: optimal balance between two components. The algorithm evaluates exhaustively all possible balances composed of only two components, that is, all balances of the following form:

$$B_{ij} = \sqrt{\frac{1}{2}}\left[\log(X_i) - \log(X_j)\right] \text{ for } i,\ j \in \{1, \ldots, k\},\ i \neq j$$

Each two-component balance (B_{ij}) is tested for association with the response variable Y with one of these regression models as follows:

$$Y = \beta_0 + \beta_1 B_{ij} + \gamma' Z, \text{ for a continuous response, or}$$

$$\text{logit}(Y) = \beta_0 + \beta_1 B_{ij} + \gamma' Z, \text{ for a dichotomous variable } Y.$$

The balance that maximizes the optimization criteria (MSE or AUC) is selected and denoted by B_1.

Note that in defining a balance for a pair of components (X_i, X_j), there are two options that differ only in their signs but provide the same association with the response:

$$B_{ij} = \sqrt{\frac{1}{2}}\left[\log(X_i) - \log(X_j)\right] \text{ and } B_{ji} = \sqrt{\frac{1}{2}}\left[\log(X_j) - \log(X_i)\right]$$

selbal returns the balance whose regression coefficient is positive.

Step *s*: optimal balance—adding a new component. For $s = >1$ and until the stop criterion is fulfilled, let $B^{(s-1)}$ be the balance defined in the previous step:

$$B^{(s-1)} \propto \frac{1}{k_+^{(s-1)}} \sum_{i \in I_+^{(s-1)}} \log(X_i) - \frac{1}{k_-^{(s-1)}} \sum_{j \in I_-^{(s-1)}} \log(X_j)$$

where $I_+^{(s-1)}$ and $I_-^{(s-1)}$ are two disjoint subsets of indices in (1, . . ., *k*), with $k_+^{(s-1)}$ and $k_-^{(s-1)}$ elements, respectively.

For each of the remaining variables, X_p, not yet included in the balance, $p \notin [I_+^{(s-1)} \cup I_-^{(s-1)}]$, the algorithm considers the balance that is obtained by adding $\log(X_p)$to the positive part of the previous balance $B^{(s-1)}$

$$B_p^{(s+)} \propto \left\{ \frac{1}{k_+^{(s-1)} + 1} \left[\sum_{i \in I_+^{(s-1)}} \log(X_i) + \log(X_p) \right] - \frac{1}{k_-^{(s-1)}} \sum_{j \in I_-^{(s-1)}} \log(X_j) \right\}$$

and the balance that is obtained by adding $\log(X_p)$ to the negative part of $B^{(s-1)}$

$$B_p^{(s-)} \propto \left\{ \frac{1}{k_+^{(s-1)}} \sum_{i \in I_+^{(s-1)}} \log(X_i) - \frac{1}{k_-^{(s-1)} + 1} \left[\sum_{j \in I_-^{(s-1)}} \log(X_j) + \log(X_p) \right] \right\}$$

Each of these pairs of balances, $B_p^{(s+)}$ and $B_p^{(s-)}$, for each of the remaining variables, X_p, is tested for association with the response variable through one of these two regression models, where *B* denotes the balance tested:

$$Y = \beta_0 + \beta_1 B + \gamma' Z, \text{ for a continuous response, or}$$

$$\text{logit}(Y) = \beta_0 + \beta_1 B + \gamma' Z, \text{ for a dichotomous variable } Y.$$

The balance that maximizes the optimization criterion defines the new balance $B^{(s)}$.

Stop criteria. There are two stopping rules: the iterative algorithm stops when value corresponding to the the improvement of the optimization parameter is lower than a specified threshold *th.imp* (default *th.imp* = 0) or when the specified maximum number of components, *C*, has been included in the balance (default *C* = 20).

Iterative cross-validation: selbal.cv*()*. An iterative cross-validation procedure is implemented in selbal.cv*()* function with two goals: (i) to identify the optimal number of components to be included in the balance and (ii) to explore the robustness of the global balance identified with the whole data set.

Let *M* be the number of iterations (default *M* = 10), *K* the number of folds in the cross-validation (default *K* = 5), and *C* the maximum number of variables or components included in a balance (default *C* = 20).

At each iteration of *m* ∈ {1, . . ., *M*}, the data are divided into *K* folds, $D_1^m, \ldots, D_K^m$.

For each *k* ∈ {1, . . ., *K*}, *selbal()* is applied to the training data set, $\text{U}_{j \neq k} \text{D}_j^m$, and the optimal balance with *C* = 20 variables is obtained, $B_k^m(20)$. Since *selbal()* is a forward selection process where variables are included sequentially at every step, we have a sequence of balances, including from *C* = 2 to *C* = 20 variables:

$$B_k^m(2),\ B_k^m(3), \ldots, B_k^m(20)$$

The classification accuracy (MSE or AUC) of these balances is measured on the test data set, D_k^m, giving a sequence of accuracy measures for each number of variables included in the balance:

$$MSE_k^m(2),\ MSE_k^m(3),\ \ldots, MSE_k^m(20)$$

and similarly for AUC.

Optimal number of components. For each number of components $c \in \{2, \ldots, C\}$ we have $K \times M$ measures of accuracy, MSE or AUC. The mean and the standard error are computed and are represented in a plot (Fig. 1).

Similarly to the cross-validation process in LASSO for finding the optimal penalization parameter lambda (38), we follow the "1se strategy" and define the optimal number of variables included in the balance (k_{opt}) as the lowest number whose mean MSE is within 1 standard error of the minimum mean MSE (or whose mean AUC is within 1 standard error of the maximum mean AUC). Usually, the 1se strategy provides sparser models than taking the minimum mean MSE (or maximum mean AUC), with very similar accuracy. This 1se strategy is the default option in *selbal*, but there is also the possibility of determining the optimal number of variables as the value reaching the optimum (minimum mean MSE or maximum mean AUC).

Global balance. Once the optimal number of components k_{opt} has been determined, we apply the main function *selbal()* to the whole dataset, with the number of taxa $C = k_{opt}$, and obtain what we call the global balance.

Robustness of the result. Any method that requires variable selection may result in overfitting. In order to explore the robustness of the global balance and the variables that form it, we retrieve all the balances with k_{opt} components obtained in the cross-validation process $B_k^m(k_{opt})$, $k \in \{1, \ldots, K\}$, $m \in \{1, \ldots, M\}$ and compare them with the global balance. We summarize these cross-validation balances in two different ways: per balance and per variable. We provide the relative frequencies of the different balances and the proportion of times that each taxon has been included into a balance. This information, available in the output of *selbal.cv()*, is summarized in a table such as that shown in Fig. 3.

This cross-validation process can also be used to obtain the cross-validation accuracy, defined as the mean MSE or mean AUC of the balances obtained in the CV process that have the same number of variables as the global balance: $\underset{k,m}{\text{mean}}[\text{MSE}_k^m(k_{opt})]$ or $\underset{k,m}{\text{mean}}[\text{AUC}_k^m(k_{opt})]$.

Data availability. *selbal* is accessible as an R package in Github (https://github.com/UVic-omics/selbal), where the data sets and scripts to reproduce this work are also available. Microbiome data were obtained from a MiSeq 16S rRNA sequence and bioinformatically processed with Mothur (29) and are available at BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) (accession number PRJNA307231; SRA accession number SRP068240).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00053-18.

FIG S1, PDF file, 0.1 MB.

FIG S2, PDF file, 0.04 MB.

FIG S3, PDF file, 0.03 MB.

FIG S4, PDF file, 0.1 MB.

FIG S5, PDF file, 0.1 MB.

FIG S6, PDF file, 0.03 MB.

FIG S7, PDF file, 0.1 MB.

FIG S8, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

The study was supported by a research grant from the Fundació Glória Soler. M.N.-J., J.R.-P., and R.P. acknowledge funds from the Gala contra la SIDA 2013, 2014, and 2015 editions, from the Nit per la Recerca a la Catalunya Central 2015 edition, and from the Red de Investigación en Sida (RD12/0017/0002) as part of the Plan Nacional R+D+I, cofinanced by the Instituto de Salud Carlos III (ISCIII)-Subdirección General de Evaluación y Fomento de la Investigación el Fondo Europeo de Desarrollo Regional (FEDER). J.R.-P. is supported through a grant for predoctoral studies from Noel Alimentaria to the University of Vic (UVic-UCC) and M.L.C. through MTM2015-64465-C2-1-R, Ministerio de Economía y Competitividad, Spain.

REFERENCES

1. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Doré J, MetaHIT Consortium, Antolín M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, et al. 2011. Enterotypes of the human gut microbiome. Nature 473:174–180. https://doi.org/10.1038/nature09944.

2. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The Human Microbiome Project. Nature 449:804–810. https://doi.org/10.1038/nature06244.
3. Santiago A, Panda S, Mengels G, Martinez X, Azpiroz F, Dore J, Guarner F, Manichanh C. 2014. Processing faecal samples: a step forward for standards in microbial community analysis. BMC Microbiol 14:112. https://doi.org/10.1186/1471-2180-14-112.
4. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. 2015. Modeling and analysis of compositional data. John Wiley & Sons, Inc, Hoboken, NJ.
5. Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. 2016. It's all relative: analyzing microbiome data as compositions. Ann Epidemiol 26:322–329. https://doi.org/10.1016/j.annepidem.2016.03.003.
6. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: and this is not optional. Front Microbiol 8:1–6. https://doi.org/10.3389/fmicb.2017.02224.
7. Aitchison J. 1986. The statistical analysis of compositional data. Chapman & Hall, London, United Kingdom.
8. Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol 26:32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x.
9. McArdle BH, Anderson MJ. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82: 290–297. https://doi.org/10.1890/0012-9658(2001)082[0290:FMMTCD]2.0.CO;2.
10. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin P, O'Hara RB, Simpson G, Solymos P, Stevens MHH, Wagner H. 2015. vegan: community ecology package. R package version 2.2-1. https://CRAN.R-project.org/package=vegan.
11. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC. 2015. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. Am J Hum Genet 96:797–807. https://doi.org/10.1016/j.ajhg.2015.04.003.
12. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8.
13. Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25. https://doi.org/10.1186/gb-2010-11-3-r25.
14. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis 26:27663. https://doi.org/10.3402/mehd.v26.27663.
15. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. 2013. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One 8:e67019. https://doi.org/10.1371/journal.pone.0067019.
16. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5:27. https://doi.org/10.1186/s40168-017-0237-y.
17. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. 2011. Human-associated microbial signatures: examining their predictive value. Cell Host Microbe 10:292–296. https://doi.org/10.1016/j.chom.2011.09.003.
18. Silverman JD, Washburne AD, Mukherjee S, David LA. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. Elife 6:1–20. https://doi.org/10.7554/eLife.21887.
19. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ 5:e2969. https://doi.org/10.7717/peerj.2969.
20. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, Navas-Molina JA, Song SJ, Metcalf JL, Hyde ER, Lladser M, Dorrestein PC, Knight R. 2017. Balance trees reveal microbial niche differentiation. mSystems 2:e00162-16. https://doi.org/10.1128/mSystems.00162-16.
21. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. 2003. Isometric Logratio transformations for compositional data analysis. Math Geol 35:279–300. https://doi.org/10.1023/A:1023818214614.
22. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Math Geol 35:253–278. https://doi.org/10.1023/A:1023866030544.
23. Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat Modelling 15:134–158. https://doi.org/10.1177/1471082X14535524.
24. Gloor GB, Macklaim JM, Vu M, Fernandes AD. 2016. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. Austrian J Stat 45:73. https://doi.org/10.17713/ajs.v45i4.122.
25. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2:15. https://doi.org/10.1186/2049-2618-2-15.
26. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, Xavier RJ. 2014. The treatment-naïve microbiome in new-onset Crohn's disease. Cell Host Microbe 15:382–392. https://doi.org/10.1016/j.chom.2014.02.005.
27. Noguera-Julian M, Rocafort M, Guillén Y, Rivera J, Casadellà M, Nowak P, Hildebrand F, Zeller G, Parera M, Bellido R, Rodríguez C, Carrillo J, Mothe B, Coll J, Bravo I, Estany C, Herrero C, Saz J, Sirera G, Torrela A, Navarro J, Crespo M, Brander C, Negredo E, Blanco J, Guarner F, Calle ML, Bork P, Sönnerborg A, Clotet B, Paredes R. 2016. Gut microbiota linked to sexual preference and HIV infection. EBioMedicine 5:135–146. https://doi.org/10.1016/j.ebiom.2016.01.032.
28. Øyri SF, Műzes G, Sipos F. 2015. Dysbiotic gut microbiome: a key element of Crohn's disease. Comp Immunol Microbiol Infect Dis 43:36–49. https://doi.org/10.1016/j.cimid.2015.10.005.
29. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541. https://doi.org/10.1128/AEM.01541-09.
30. Brenchley JM, Price DA, Schacker TW, Asher TE, Silvestri G, Rao S, Kazzaz Z, Bornstein E, Lambotte O, Altmann D, Blazar BR, Rodriguez B, Teixeira-Johnson L, Landay A, Martin JN, Hecht FM, Picker LJ, Lederman MM, Deeks SG, Douek DC. 2006. Microbial translocation is a cause of systemic immune activation in chronic HIV infection. Nat Med 12:1365–1371. https://doi.org/10.1038/nm1511.
31. Klatt NR, Canary LA, Sun X, Vinton CL, Funderburg NT, Morcock DR, Quiñones M, Deming CB, Perkins M, Hazuda DJ, Miller MD, Lederman MM, Segre JA, Lifson JD, Haddad EK, Estes JD, Brenchley JM. 2013. Probiotic/prebiotic supplementation of antiretrovirals improves gastrointestinal immunity in SIV-infected macaques. J Clin Invest 123:903–907. https://doi.org/10.1172/JCI66227.
32. d'Ettorre G, Ceccarelli G, Giustini N, Serafino S, Calantone N, De Girolamo G, Bianchi L, Bellelli V, Ascoli-Bartoli T, Marcellini S, Turriziani O, Brenchley JM, Vullo V. 2015. Probiotics reduce inflammation in antiretroviral treated, HIV-infected individuals: results of the 'probio-HIV' clinical trial. PLoS One 10:e0137200. https://doi.org/10.1371/journal.pone.0137200.
33. McMurdie PJ, Holmes S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 10:e1003531. https://doi.org/10.1371/journal.pcbi.1003531.
34. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; French StatOmique Consortium. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14:671–683. https://doi.org/10.1093/bib/bbs046.
35. Mittlböck M, Schemper M. 1996. Explained variation for logistic regression. Stat Med 15:1987–1997. https://doi.org/10.1002/(SICI)1097-0258(19961015)15:19<1987::AID-SIM318>3.0.CO;2-9.
36. Tjur T. 2009. Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. Am Stat 63: 366–372. https://doi.org/10.1198/tast.2009.08210.
37. Palarea-Albaladejo J, Martín-Fernández JA. 2015. ZCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemometr Intell Lab Syst 143:85–96. https://doi.org/10.1016/j.chemolab.2015.02.019.
38. Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning. Springer-Verlag, New York, NY.

Acknowledgements

Hasta este punto de la tesis he tratado de mantener la precisión que un escrito científico requiere. En los siguientes párrafos dejo este rigor de lado para expresar mi agradecimiento a toda la gente que me ha ayudado a hacer posible todo esto, independientemente del cómo. Así pues, todo el contenido matemático que se ha ido tratando hasta ahora se reduce únicamente a la propiedad conmutativa, encargada de otorgar la misma importancia a cada una de las partes a las que quiero mostrar mi cariño por un cuatrienio que tiene como desenlace esta tesis.

Igual que las fórmulas ayudan a describir el mundo y las leyes que lo rigen, las palabras pero sobre todo los gestos, son la herramienta más adecuada para expresar los sentimientos. Es por ello que dejo por un momento fórmulas y ecuaciones de lado, para expresar mi agradecimiento a un gran número de personas. Bueno, ... ojo que igual no me resulta tan fácil dejar las matemáticas de lado.

Me gustaría comenzar agradeciendo a M. Luz Calle y a Marc Noguera la confianza que depositaron en mí para realizar el doctorado bajo su dirección. Así mismo, agradezco a Noel Alimentaria su apuesta por la ciencia y más concretamente la beca que crearon junto con IrsiCaixa y que me ha permitido ser parte de todo esto. En este momento de agradecimiento a los directores de tesis, me gustaría resaltar todo el tiempo que tanto *Malu* como Marc han dedicado a la corrección de artículos, presentaciones, pósters, capítulos de esta tesis, ... cosa que agradezco enormemente ya que me ha permitido mejorar todo lo que he hecho. Además, tenerlos como tutores me ha servido para ver el trabajo desde dos puntos de vista diferentes: el más teórico y el práctico; lo cuál ha sido muy enriquecedor para mí.

Por otro lado quisiera agradecer a las dos instituciones de las que he formado parte estos cuatro años. Tanto IrsiCaixa como la UVic-UCC me han facilitado la asistencia a muchos congresos y seminarios a nivel nacional e internacional, así como la colaboración con científicos de otras instituciones. Poder asistir al CROI en Boston, participar en el CoDaWork o relizar la estancia en BioSS, han sido algunas de las cosas que sin su ayuda habría resultado difícil llevar a cabo. También me gustaría destacar la posibilidad que tanto la UVic-UCC como Malu me han dado de impartir clases en la universidad; lo cierto es que me ha permitido disfrutar mucho y quién sabe si también descubrir a lo que me quiero dedicar en los próximos años, ...

Como advertí al inicio, en estos agradecimientos el orden de los factores resulta irrelevante. Y lo cierto es que estos años me ha sido tan importante tener un apoyo a nivel profesional como a nivel personal. Y es aquí donde quisiera detenerme un poco más, porque hasta en los párrafos anteriores cuando hablaba de instituciones o empresas, en el fondo todo se reduce a las personas que las componen. Volviendo un poco a la parte más científica, aunque no existe una fórmula consenso para describir la felicidad, creo que de existir no se podría negar que en ella interviene una componente referente a la familia. Una famila que podríamos definir como una *función monótona creciente y dependiente del tiempo*; y es que, quien entra a formar parte de la familia, una vez lo hace, por mucho que esté lejos o no volvamos a verle más, nunca dejará de formar parte de ella.

El primer término de esta *función familia* es el más importante: el intercepto. Su importancia radica en que tiene un valor fijo desde el primer instante y nada ni nadie lo puede modificar. Queda determinado desde el mismo momento en que uno nace y, ... ¡qué suerte tengo de que el intercepto de mi ecuación sea positivo y taaaaan grande! Casi treinta años lleva establecido en mi fórmula manteniendo el valor que tenía el primer día. Entiendo que hacer una metáfora matemática pueda resultar un poco complejo, así que por si acaso, ... ¡mamá, papá, abuelos (estéis donde estéis haciendo de *guardianes invisibles*) y demás familia ... ¡muchas gracias por ser un intercepto tan valioso! Andoni, tú no eres parte de este intercepto, te recuerdo que es el valor que toma la función cuando la variable independiente (en este caso el tiempo t) es 0, y tú no apareces hasta $t = 6$. Por si lo habías olvidado, yo soy el hermano mayor, no el más alto, pero sí el mayor.

Y en cuanto a la parte dependiente del tiempo, quisiera hacer un *break* en $t = 25$, que es cuando vine a Badalona y comencé el doctorado. Previo a este tiempo quiero destacar a quienes a través de la enseñanza hicieron que mi interés por las matemáticas creciese hasta llegar al punto de decidirme por estudiar la licenciatura. También a quienes ya dentro de la universidad, de una forma u otra me animaron a hacer el máster en estadística que me permitió llegar hasta aquí. ¡Cómo no!, acordarme por supuesto de la *cuadrilla* de Ermua, con la que he compartido y espero compartir muchos momentos de diversión y alegría. Y, ... (ahora sí Andoni), mencionar por supuesto a mi hermano, con quien al llevar casi siete años lejos de Ermua, no he podido compartir tantos momentos como hubiese querido, pero al que agradezco que me haya hecho compañía desde la distancia todo este tiempo.

¡Y qué decir de los últimos cuatro años! Cuando pasas tanto tiempo en un sitio trabajando, de forma natural surgen vínculos afectivos con las personas con las que compartes *lloc de feina*. Cierto es que no empecé con buen pie, ocupando el sitio en el que se colocaba la comida de los cumpleaños, ... pero poco a poco eso ha ido quedando en el olvido (¿no?). Es muy gratificante venir a trabajar sabiendo que formas parte de un grupo en el que si tienes cualquier duda de biología siempre habrá alguien que te dedique de su tiempo para aclarar un concepto determinado, pese a que seas un poco *cansino*, y tengas que preguntarlo varias veces porque no lo has entendido a la primera. Día tras día *yo li*ándoles con preguntas y más preguntas, ... aunque también espero haber podido responder a algunas de las que ellos han tenido. También resulta divertido saber que llega la hora de comer y que vas a poder desconectar un poco y pasar un rato entretenido hablando sobre un tema aleatorio que alguien propone: los planes para el fin de semana, la última película vista, lo bueno que es el *jengibre* (*pa tó*), ... y sino, pues igual toca prepararse, armarse de fuerza y aguantar que te metan caña durante toda la comida; eso sí, siempre desde el cariño (¿no?).

Gratificante ha sido también poder exponer resultados o presentar artículos y ver cómo te atienden y preguntan independientemente de que el tema tratado no sea muy de su interés. Por supuesto que ha sido muy positivo también poder compartir momentos con algunos compañeros fuera del trabajo: destacar las quedadas con otros *PreDoc survivors* para tomar algo, los momentos en congresos fuera de la parte científica, alguna que otra *cursa* realizada en este tiempo, las cenas de Navidad (y lo que las sigue, ...), tardes ~~jugando~~ ganando al volley, ...

Pero sobre todo, ... ¡qué agradable es darse cuenta de que has conocido gente con la que conectas rápidamente!, personas con las que hay tan buena sintonía, que está *clara* su incorporación a la ecuación de la familia para quedarse por siempre.

Por todo lo dicho y lo que queda en el tintero ...

¡Muchas gracias! Moltes gràcies! Eskerrik asko!